HAL
open science

High-dimensional mixture regression models, application to functional data

Emilie Devijver

To cite this version:

Emilie Devijver. High-dimensional mixture regression models, application to functional data. Statistics [math.ST]. Université Paris Sud (Paris 11), 2015. English. NNT : . tel-01695257

HAL Id: tel-01695257

https://hal.science/tel-01695257

Submitted on 29 Jan 2018

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

École Doctorale de mathématiques de la Région Paris-Sud

Laboratoire de Mathématiques d'Orsay

THÈSE

présentée pour obtenir le grade de
docteur en sciences de l'Université Paris-Sud

Discipline : Mathématiques

par

Emilie Devijver

Modèles de mélange pour la régression en grande dimension, application aux données fonctionnelles.

Soutenue le 02 juillet 2015 devant la commission d'examen :

Francis	Bach	INRIA Paris-Rocquencourt	Examinateur
Christophe	Biernacki	Université de Lille 1	Rapporteur
Yannig	Goude	EDF R&D	Examinateur
Jean-Michel	Loubes	Université de Toulouse	Rapporteur
Pascal	Massart	Université Paris-Sud	Directeur de thèse
Jean-Michel	Poggi	Université Paris-Sud	Directeur de thèse

Thèse préparée sous la direction de Pascal Massart
et de Jean-Michel Poggi
au Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8625),
Bât. 425, Université Paris Sud
91405 Orsay Cedex

Modèles de mélange pour la régression en grande dimension, application aux données fonctionnelles.

Les modèles de mélange pour la régression sont utilisés pour modéliser la relation entre la réponse et les prédicteurs, pour des données issues de différentes sous-populations. Dans cette thèse, on étudie des prédicteurs de grande dimension et une réponse de grande dimension.
Tout d'abord, on obtient une inégalité oracle ℓ_1 satisfaite par l'estimateur du Lasso. On s'intéresse à cet estimateur pour ses propriétés de régularisation ℓ_1.
On propose aussi deux procédures pour pallier ce problème de classification en grande dimension.
La première procédure utilise l'estimateur du maximum de vraisemblance pour estimer la densité conditionnelle inconnue, en se restreignant aux variables actives sélectionnées par un estimateur de type Lasso.
La seconde procédure considère la sélection de variables et la réduction de rang pour diminuer la dimension.
Pour chaque procédure, on obtient une inégalité oracle, qui explicite la pénalité nécessaire pour sélectionner un modèle proche de l'oracle.
On étend ces procédures au cas des données fonctionnelles, où les prédicteurs et la réponse peuvent être des fonctions. Dans ce but, on utilise une approche par ondelettes. Pour chaque procédure, on fournit des algorithmes, et on applique et évalue nos méthodes sur des simulations et des données réelles. En particulier, on illustre la première méthode par des données de consommation électrique.

Mots-clés : modèles de mélange en régression, classification non supervisée, grande dimension, sélection de variables, sélection de modèles, inégalité oracle, données fonctionnelles, consommation électrique, ondelettes.

High-dimensional mixture regression models, application to functional data.

Finite mixture regression models are useful for modeling the relationship between a response and predictors, arising from different subpopulations. In this thesis, we focus on high-dimensional predictors and a high-dimensional response.
First of all, we provide an ℓ_1-oracle inequality satisfied by the Lasso estimator. We focus on this estimator for its ℓ_1-regularization properties rather than for the variable selection procedure.
We also propose two procedures to deal with this issue. The first procedure leads to estimate the unknown conditional mixture density by a maximum likelihood estimator, restricted to the relevant variables selected by an ℓ_1-penalized maximum likelihood estimator.
The second procedure considers jointly predictor selection and rank reduction for obtaining lower-dimensional approximations of parameters matrices.
For each procedure, we get an oracle inequality, which derives the penalty shape of the criterion, depending on the complexity of the random model collection. We extend these procedures to the functional case, where predictors and responses are functions. For this purpose, we use a wavelet-based approach. For each situation, we provide algorithms, apply and evaluate our methods both on simulations and real datasets. In particular, we illustrate the first procedure on an electricity load consumption dataset.

Keywords: mixture regression models, clustering, high dimension, variable selection, model selection, oracle inequality, functional data, electricity consumption, wavelets.

Contents

Résumé **6**

Introduction **10**

Notations **33**

1 Two procedures **38**
1.1 Introduction . . . 39
1.2 Gaussian mixture regression models . . . 41
1.3 Two procedures . . . 45
1.4 Illustrative example . . . 47
1.5 Functional datasets . . . 53
1.6 Conclusion . . . 60
1.7 Appendices . . . 60

2 An ℓ_1-oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models **67**
2.1 Introduction . . . 69
2.2 Notations and framework . . . 70
2.3 Oracle inequality . . . 72
2.4 Proof of the oracle inequality . . . 74
2.5 Proof of the theorem according to $\mathcal{T}$ or $\mathcal{T}^c$ . . . 78
2.6 Some details . . . 83

3 An oracle inequality for the Lasso-MLE procedure **90**
3.1 Introduction . . . 91
3.2 The Lasso-MLE procedure . . . 92
3.3 An oracle inequality for the Lasso-MLE model . . . 95
3.4 Numerical experiments . . . 98
3.5 Tools for proof . . . 101
3.6 Appendix: technical results . . . 109

4 An oracle inequality for the Lasso-Rank procedure **116**
4.1 Introduction . . . 117
4.2 The model and the model collection . . . 119
4.3 Oracle inequality . . . 122
4.4 Numerical studies . . . 125
4.5 Appendices . . . 126

5 Clustering electricity consumers using high-dimensional regression mixture models. 135
5.1 Introduction 136
5.2 Method 137
5.3 Typical workflow using the example of the aggregated consumption 138
5.4 Clustering consumers 142
5.5 Discussion and conclusion 149

List of Figures

1 Exemple de données simulées . . . 12
2 Illustration de l'estimateur du Lasso . . . 16
3 Saut de dimension . . . 23
4 Heuristique des pentes . . . 23

1.1 Number of FR and TR . . . 49
1.2 Zoom in on number of FR and Tr . . . 49
1.3 Slope graph for our Lasso-Rank procedure . . . 50
1.4 Slope graph for our Lasso-MLE procedure . . . 50
1.5 Boxplot of the Kullback-Leibler divergence . . . 51
1.6 Boxplot of the ARI . . . 51
1.7 Boxplot of the Kullback-Leibler divergence . . . 51
1.8 Boxplot of the ARI . . . 51
1.9 Boxplot of the Kullback-Leibler divergence . . . 52
1.10 Boxplot of the ARI . . . 52
1.11 Boxplot of the Kullback-Leibler divergence . . . 52
1.12 Boxplot of the ARI . . . 52
1.13 Boxplot of the ARI . . . 56
1.14 Plot of the 70-sample of half-hour load consumption, on the two days . . . 56
1.15 Plot of a week of load consumption . . . 56
1.16 Summarized results for the model 1 . . . 58
1.17 Summarized results for the model 1 . . . 59

3.1 Boxplot of the Kullback-Leibler divergence . . . 95
3.2 Boxplot of the Kullback-Leibler divergence . . . 99
3.3 Boxplot of the ARI . . . 99
3.4 Summarized results for the model 1 . . . 100
3.5 Summarized results for the model 1 . . . 100

5.1 Load consumption of a sample of 5 consumers over a week in winter . . . 137
5.2 Projection of a load consumption for one day into Haar basis, level 4. By construction, we get $s = A_4 + D_4 + D_3 + D_2 + D_1$. On the left side, the signal is considered with reconstruction of dataset, the dotted being preprocessing 1 and the dotted-dashed being the preprocessing 2 . . . 139
5.3 We select the model $\hat{m}$ using the slope heuristic . . . 140
5.4 Minimization of the penalized log-likelihood. Interesting models are branded by red squares, the selected one by green diamond . . . 140
5.5 Representation of the regression matrix β_k for the preprocessing 1. . . . 140
5.6 Representation of the regression matrix β_k for the preprocessing 2. . . . 140
5.7 For the selected model, we represent $\hat{\beta}$ in each cluster . . . 140

5.8 For the model selected, we represent Σ in each cluster 141
5.9 Assignment boxplots per cluster 141
5.10 Clustering representation. Each curve is the mean in each cluster 141
5.11 Clustering representation. Each curve is the mean in each cluster 142
5.12 Saturday and Sunday load consumption in each cluster. 143
5.13 Proportions in each cluster for models constructed by our procedure 144
5.14 Regression matrix in each cluster for the model with 2 clusters 145
5.15 Daily mean consumptions of the cluster centres along the year for 2 (top) and 5 clusters (bottom) 145
5.16 Daily mean consumptions of the cluster centres in function of the daily mean temperature for 2 (on the left) and 5 clusters (on the right) 146
5.17 Average (over time) week of consumption for each centre of the two classifications (2 clusters on the top and 5 on the bottom) 146
5.18 Out of bag error of the random forest classifiers in function of the number of trees 147
5.19 RMSE on Thursday prediction for each procedure over all consumers 148
5.20 Daily mean consumptions of the cluster centres in function of the daily mean temperature for 5 clusters, clustering done by observing Thursday and Wednesday in summer 149
5.21 Daily mean consumptions of the cluster centres along the year for 3 clusters, clusering done on weekend observation 149

Introduction

Faire des classes pour mieux modéliser un échantillon est une méthode classique en statistiques. L'exemple pratique développé dans cette thèse est la classification de consommateurs électriques en Irlande. Habituellement, on classe des individus dans un problème d'estimation, mais ici, pour souligner l'aspect prédictif, on classe les observations qui ont le même type de comportement d'un jour sur l'autre. C'est une classification plus fine, qui agit sur des régresseurs homoscédastiques. Dans notre exemple, on classifie les consommateurs sur un modèle de régression d'un jour par rapport à leur consommation de la veille. Dans cette thèse, on a développé ce contexte d'un point de vue méthodologique, en proposant deux procédures de classification en régression par modèles de mélange qui pallient le problème de grande dimension, mais aussi d'un point de vue théorique en justifiant la partie sélection de modèles de nos procédures, et d'un point de vue pratique en travaillant sur le jeu de données réelles de consommation électrique en Irlande. Dans cette introduction, nous présentons les notions développées dans cette thèse.

Régression linéaire

Les méthodes de régression consistent à chercher le lien qui existe entre deux variables aléatoires X et Y. La variable X représente les régresseurs, les variables explicatives, alors que Y décrit la réponse. Le modèle linéaire Gaussien, le modèle le plus simple en régression, suppose que Y dépend de X de façon linéaire, à un bruit Gaussien près. Plus formellement, si X et Y sont deux vecteurs aléatoires, $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}$, étudier un modèle linéaire Gaussien sur (X, Y) consiste à trouver $\beta \in \mathbb{R}^p$ tel que

$$Y = \beta X + \epsilon \tag{1}$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$, σ^2 étant connue ou à estimer suivant les cas. L'étude de ce modèle est, à ce jour, assez complète. Soit $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon. Si l'échantillon est de taille suffisante, on connaît un estimateur consistant (c'est-à-dire qui converge vers la vraie valeur), *l'estimateur des moindres carrés* (qui coïncide dans ce cas avec l'estimateur du maximum de vraisemblance). On notera $(\hat{\beta}, \hat{\sigma}^2)$ cet estimateur, où

$$\hat{\beta} = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{y}\,;$$
$$\hat{\sigma}^2 = \frac{||\mathbf{y} - \mathbf{x}\hat{\beta}||^2}{n-p}.$$

On connaît la loi de cet estimateur : $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{x}^t\mathbf{x})^{-1})$ et $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, ce qui nous permet de déduire des intervalles de confiance pour chaque paramètre.
On peut généraliser ce modèle à une variable $Y \in \mathbb{R}^q$ multivariée. Dans ce cas, on observe un échantillon $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$, et on construit un estimateur

$(\hat{\beta}, \hat{\Sigma}) \in \mathbb{R}^{q\times p} \times \mathbb{S}_q^{++}$, où

$$\hat{\beta} = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{y}\,; \tag{2}$$
$$\hat{\Sigma} = \left(\frac{\langle(\mathbf{y}-\mathbf{x}\hat{\beta})_{l_1}, (\mathbf{y}-\mathbf{x}\hat{\beta})_{l_2}\rangle}{n-p}\right)_{\substack{1\le l_1\le q\\ 1\le l_2\le q}}\,;$$

et $\mathbb{S}_q^{++}$ est l'ensemble des matrices symétriques définies positives de taille q.
Ce modèle de régression est utilisé, par exemple, pour prédire de nouvelles valeurs. Si on connaît le modèle sous-jacent à nos données, c'est-à-dire que l'on a déterminé $\hat{\beta}$ et $\hat{\Sigma}$ à partir d'un échantillon $((x_1, y_1), \ldots, (x_n, y_n))$, et que l'on observe un nouvel x_{n+1}, on peut calculer $\hat{y}_{n+1} = \hat{\beta}x_{n+1}$. Dans ce cas $\hat{y}_{n+1}$ est la valeur dite *prédite*.
Si on s'intéresse au couple des variables aléatoires (X, Y), on peut estimer la densité du couple, mais on peut aussi étudier la densité conditionnelle. C'est cette dernière quantité que l'on a décrite par un modèle linéaire dans (1). Les covariables peuvent aussi avoir différents statuts : soit elles sont fixées, déterministes, soit elles sont aléatoires, et on travaille alors conditionnellement à la loi sous-jacente. Dans cette thèse, on s'intéresse à la loi conditionnelle, pour des régresseurs fixes ou aléatoires.
Cependant, cette hypothèse de modèle linéaire est très contraignante : si on observe un échantillon $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$, on suppose que chaque y_i dépend de la même façon de x_i, à un bruit près, pour $i \in \{1, \ldots, n\}$. Si le modèle est adapté aux données, le bruit sera petit, ce qui signifie que les coefficients de la matrice de covariance Σ seront petits. Cependant, de nombreux jeux de données ne peuvent pas être bien résumés par un modèle linéaire.

Modèles de mélange en régression

Pour affiner ce modèle, on peut choisir de construire plusieurs classes, et de faire dépendre nos estimateurs $\hat{\beta}$ et $\hat{\Sigma}$ de la classe. Plus formellement, on étudie un *modèle de mélange en régression* de K Gaussiennes : si on observe un échantillon $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$, et si on suppose que l'observation i appartient à la classe k, alors il existe β_k et Σ_k tels que

$$y_i = \beta_k x_i + \epsilon_i$$

où $\epsilon_i \sim \mathcal{N}(0, \Sigma_k)$. Dans cette thèse, on considère un mélange avec un nombre de classes K fini.
Les modèles de mélange sont traditionnellement étudiés dans des problèmes d'estimation (citons par exemple McLachlan et Basford, [MB88], McLachlan et Peel, [MP04], qui sont deux livres fondateurs pour les modèles de mélange, Fraley et Raftery, [FR00], pour un état de l'art, ou encore Celeux et Govaert, [CG93], pour l'étude d'une densité de mélange dans un but de classification).
L'idée principale est d'estimer la densité inconnue s^* par un mélange de densités classiques $(s_{\theta_k})_{1\le k\le K}$: on peut alors écrire

$$s^* = \sum_{k=1}^{K} \pi_k s_{\theta_k}$$
$$\sum_{k=1}^{K} \pi_k = 1$$

Dans cette thèse, on s'intéresse à un mélange de modèles linéaires Gaussiens à poids constants, les densités s_{θ_k} sont donc des densités Gaussiennes conditionnelles.

Plusieurs idées émergent avec les modèles de mélange, en régression ou non. Si on connaît les paramètres de notre modèle, on peut calculer pour chaque observation la probabilité a posteriori d'appartenir à une classe. D'après le *principe du maximum a posteriori* (noté MAP), on peut alors affecter une classe à chaque observation. Pour ce faire, on calcule la probabilité a posteriori de chaque observation d'appartenir à une classe, et on affecte les observations aux classes les plus probables. On peut ainsi caractériser chaque observation par les paramètres de la densité conditionnelle associée à la classe.
Dans le cadre de la classification supervisée, on connaît l'affectation de chaque observation, et on cherche à comprendre la formation des classes pour classer une nouvelle observation ; quand on fait de la classification semi-supervisée, on connaît certaines affectations, et on cherche à comprendre le modèle (souvent, c'est très coûteux de connaître les affectations des observations) ; dans le cadre de la classification non supervisée, on ne connaît pas du tout les affectations.
Nous nous plaçons dans cette thèse dans le contexte de classification non supervisée, avec une approche en régression. Cela a déjà été envisagé, citons par exemple Städler et coauteurs ([SBG10]) ou Meynet ([Mey13]), qui travaillent avec ce modèle pour des réponses Y univariées.
On observe des couples $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$, et on veut faire des classes, pour regrouper les observations (x_i, y_i) pour $i \in \{1, \ldots, n\}$ qui ont la même relation entre y_i et x_i. Sur certains jeux de données, cette approche semble naturelle, et on connaît le nombre de classes.

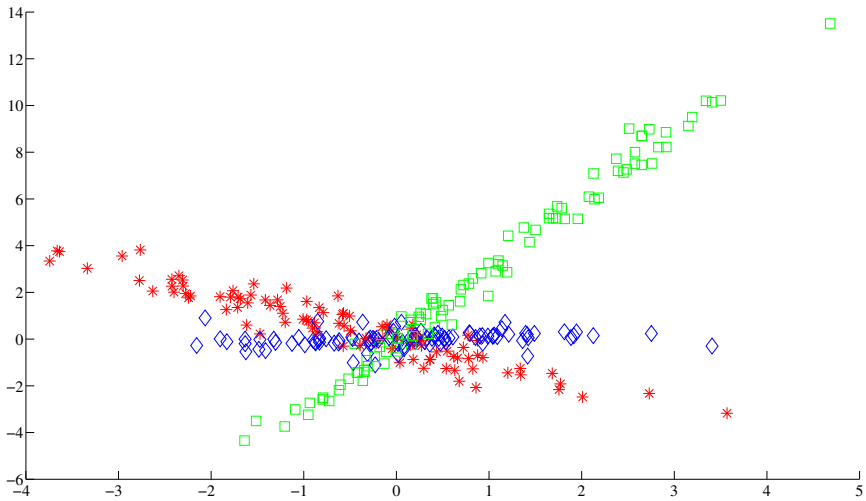

FIGURE 1 – Exemple de données simulées où les classes en régression se comprennent par la réalisation graphique. On a ici envie de construire 3 classes, représentées par les différents symboles : $*$ pour la classe 1, $\Diamond$ pour la classe 2, $\square$ pour la classe 3. Les données sont en dimension un, $(X, Y) \in \mathbb{R} \times \mathbb{R}$. Les observations $\mathbf{x}$ sont issues d'une loi Gaussienne centrée réduite, et $\mathbf{y}$ est simulée suivant un mélange de Gaussiennes de moyennes $\beta_k x$ et de variance 0.25, où $\boldsymbol{\beta} = [-1, 0.1, 3]$.

Néanmoins, sur un jeu de données quelconque, on va surtout chercher à mieux comprendre la relation entre les variables aléatoires Y et X en regroupant les observations qui ont la même dépendance entre Y et X. La structure de groupes n'est pas forcément clairement induite par les données, et on ne sait pas forcément combien de classes on a intérêt à former : dans certains cas, il va falloir estimer K.
Si on considère le modèle de mélange de Gaussiennes multivariées en régression à K classes, on

peut décrire ce modèle à l'aide des outils statistiques classiques.
Si on suppose que les vecteurs aléatoires Y sont indépendants conditionnellement à X, on considère la densité conditionnelle s_ξ^K, où

$$\begin{aligned} s_\xi^K &: \mathbb{R}^q \to \mathbb{R} \\ y &\mapsto s_\xi^K(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{1}{2}(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)\right); \end{aligned}$$

où les paramètres à estimer sont $\xi = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \Sigma_1, \ldots, \Sigma_K) \in \Xi_K$, avec

$$\begin{aligned} \Xi_K &= \Pi_K \times \left(\mathbb{R}^{q\times p}\right)^K \times \left(\mathbb{S}_q^{++}\right)^K ; \\ \Pi_K &= \left\{ (\pi_1, \ldots, \pi_K) \in [0,1]^K \,\middle|\, \sum_{k=1}^K \pi_k = 1 \right\}. \end{aligned}$$

A partir de cette densité conditionnelle, on peut définir l'*estimateur du maximum de vraisemblance* par

$$\hat{\xi}_K^{EMV} = \underset{\xi\in\Xi_K}{\operatorname{argmin}} \left\{ -\frac{1}{n} l(\xi, \mathbf{x}, \mathbf{y}) \right\}; \tag{3}$$

où la *log-vraisemblance* est définie par

$$l(\xi, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \log\left(s_\xi^K(y_i|x_i)\right).$$

Comme dans la plupart des modèles de mélange classiques, dans le cadre de la régression, l'estimateur du maximum de vraisemblance n'est pas explicite. Cela complique l'analyse théorique : on n'a pas accès à la loi des estimateurs, ni des affectations. D'un point de vue pratique, on a recours à l'algorithme EM pour approcher cet estimateur. Cet algorithme, introduit par Dempster et coauteurs dans [DLR77], permet d'estimer les paramètres d'un modèle de mélange. Il consiste à alterner deux étapes, jusqu'à convergence des paramètres ou d'une fonction des paramètres. Décrivons ce résultat.
On note $\mathbf{Z} = (Z_1, \ldots, Z_n)$ le vecteur aléatoire des affectations des observations, et $\mathcal{Z}$ l'ensemble de toutes les partitions possibles : $Z_{i,k} = 1$ si l'observation i est issue de la classe k, 0 sinon. On notera aussi, pour $i \in \{1, \ldots, n\}$, $f(z_i|x_i, y_i, \xi)$ la probabilité que $Z_i = z_i$ sachant (x_i, y_i) et connaissant le paramètre ξ. On définit la log-vraisemblance complétée par

$$l_c(\xi, (\mathbf{x}, \mathbf{y}, \mathbf{z})) = \sum_{i=1}^n \left(\log\left(f(z_i|x_i, y_i, \xi)\right) + \log\left(s_\xi^K(y_i|x_i)\right)\right).$$

Alors, on peut décomposer la log-vraisemblance comme suit :

$$l(\xi, \mathbf{x}, \mathbf{y}) = Q(\xi|\tilde{\xi}) - H(\xi|\tilde{\xi})$$

où $Q(\xi|\tilde{\xi})$ est l'espérance sur les variables latentes Z de la vraisemblance complétée,

$$Q(\xi|\tilde{\xi}) = \sum_{z\in\mathcal{Z}} \sum_{i=1}^n l_c(\xi, x_i, y_i, z_i) f(z_i|x_i, y_i, \tilde{\xi})$$

et

$$H(\xi|\tilde{\xi}) = \sum_{z\in\mathcal{Z}} \sum_{i=1}^{n} \log f(z|x_i, y_i, \xi) f(z_i|x_i, y_i, \tilde{\xi})$$

l'espérance étant prise sur les variables latentes Z. Dans l'algorithme EM, on itère le calcul jusqu'à la convergence et la maximisation de $Q(\xi|\xi^{(\text{ite})})$, où $\xi^{(\text{ite})}$ est l'estimation des paramètres à l'itération (ite) $\in \mathbb{N}^*$ de l'algorithme EM. En effet, si on fait croître $\xi \mapsto Q(\xi|\xi^{(ite)})$, on fait aussi croître la vraisemblance. Dempster, dans [DLR77], a donc proposé l'algorithme suivant.

Algorithme 1 : Algorithme EM

Data : $\mathbf{x}$,$\mathbf{y}$,K
Result : $\hat{\xi}_K^{EMV}$

1. Initialisation de $\xi^{(0)}$
2. Itération jusqu'à atteindre un critère d'arrêt
   — Étape E : calcul, pour tout ξ de
   $$Q(\xi|\xi^{(ite)})$$
   — Étape M : calcul de $\xi^{(ite+1)}$ tel que
   $$\xi^{(ite+1)} \in \operatorname{argmax} Q(\xi|\xi^{(ite)})$$

Dans la première étape, on affecte les observations à des classes, et dans la seconde étape on met à jour l'estimation des paramètres. Pour l'étude générale de cet algorithme, on peut citer le livre de McLachlan et Krishnan dans [MK97]. Un des points problématiques est l'initialisation de cet algorithme : même si on peut, dans de nombreux cas (le nôtre par exemple), montrer que l'algorithme converge vers la valeur voulue, il faut l'initialiser correctement. On peut citer Biernacki, Celeux et Govaert, [BCG03], qui décrivent diverses stratégies d'initialisation, ou encore Yang, Lai et Lin dans [YLL12] qui proposent un algorithme EM robuste pour l'initialisation. Tous ces problèmes, classiques dans l'étude des modèles de mélange, se retrouvent dans notre cadre. D'un point de vue théorique, un résultat important pour les modèles de mélange est l'identifiabilité. Rappelons qu'un modèle paramétrique est dit *identifiable* si différentes valeurs des paramètres génèrent différentes distributions de probabilité. Ici, les modèles de mélange ne sont pas identifiables, car il est possible d'intervertir les étiquettes des classes sans changer la densité (ce que l'on appelle le *label switching*). Pour un point de vue détaillé sur ces questions, citons par exemple Titterington, dans [TSM85].
A partir d'un échantillon $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$, on peut construire un modèle de mélange de régression, à K classes, en estimant les paramètres par exemple avec l'estimateur (3). Avec ce modèle, on peut obtenir une classification des données en calculant les probabilités a posteriori. Chaque observation (x_i, y_i) est alors affectée à une classe $\hat{k}_i$, et on a accès au lien $\hat{\beta}_{\hat{k}_i}$ qui existe entre x_i et y_i, et au bruit $\hat{\Sigma}_{\hat{k}_i}$ associé à cette classe grâce à l'estimateur (3). Tant qu'on a de bonnes propriétés sur l'estimateur du maximum de vraisemblance (par exemple quand la taille de l'échantillon n est grande), si on règle les soucis d'initialisation et de convergence de l'algorithme EM, on peut bien estimer les paramètres de notre modèle.
Cet algorithme a été généralisé par Städler et coauteurs dans [SBG10] pour l'estimation des paramètres d'un modèle de mélange de Gaussiennes en régression univariée. On l'a généralisé, dans cette thèse, pour les données multivariées.

Sélection de variables et estimateur du Lasso

Nous nous sommes intéressés à un problème de classification en régression non supervisée de données en *grande dimension*, c'est-à-dire que les vecteurs aléatoires $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}^q$ peuvent être des vecteurs de grande taille, éventuellement plus grande que le nombre d'observations n.
Ce problème est très étudié actuellement. En effet, avec l'amélioration de l'informatique, on a de plus en plus de données de grande dimension, et on a accès à de plus en plus de variables explicatives pour décrire un même objet.
Dans le cadre du modèle linéaire, si $p > n$, on ne peut plus calculer $\hat{\beta}$ avec la formule (2) : la matrice $\mathbf{x}^t\mathbf{x}$ n'est plus inversible. En fait, on cherche à estimer, dans le cas du modèle linéaire, $pq + q^2$ paramètres, ce qui peut être plus grand que le nombre d'observations n si p et q sont grands.
Dans un premier temps, restreignons-nous au modèle linéaire. L'estimateur du Lasso, introduit par Tibshirani dans [Tib96], et parallèlement par Chen et coauteurs en théorie du signal dans [CDS98], est un estimateur classique pour pallier le problème de la grande dimension. On peut aussi citer l'estimateur de Dantzig, introduit par Candès et Tao dans [CT07] ; l'estimateur Ridge, qui utilise une pénalité ℓ_2 plutôt que la pénalité ℓ_1 exploitée par l'estimateur du Lasso, introduit par Hoerl et Kennard dans [HK70] ; l'Elastic net, introduit par Zou et Hastie dans [ZH05], et qui est défini avec une double pénalisation ℓ_1 et ℓ_2, qui fait donc un compromis entre l'estimateur Ridge et l'estimateur Lasso ; le Fused Lasso, introduit par Tibshirani dans [TSR+05] ; le Lasso adaptatif, introduit par Zou dans [Zou06], l'estimateur du Group-Lasso, introduit par Yuan et Lin dans [YLL06], pour avoir de la parcimonie par groupes, ...
Dans cette thèse, nous nous concentrerons sur l'estimateur du Lasso (ou du Group-Lasso), même si certains résultats théoriques restent valables pour n'importe quelle méthode de sélection de variables.
L'idée introduite par [Tib96] est de supposer que la matrice β dans le modèle linéaire est creuse, ce qui réduit le nombre de paramètres à estimer. En effet, s'il y a peu de coefficients non nuls, en notant $J \subset \mathcal{P}(\{1, \dots, q\} \times \{1, \dots, p\})$ l'ensemble des indices des coefficients de la matrice de régression non nuls et $|J|$ son cardinal, la taille de l'échantillon pourra être plus grande que le nombre de paramètres à estimer, qui est alors $|J| + q^2$. Dans le cas du modèle linéaire (1), *l'estimateur du Lasso* est défini par

$$\hat{\beta}^{\text{Lasso}}(\lambda) = \underset{\beta \in \mathbb{R}^{q \times p}}{\operatorname{argmin}} \left\{ ||Y - \beta X||_2^2 + \lambda ||\beta||_1 \right\} ; \tag{4}$$

avec $\lambda > 0$ un *paramètre de régularisation* à préciser. On peut aussi le définir, de manière équivalente, par

$$\tilde{\beta}^{\text{Lasso}}(A) = \underset{\substack{\beta \in \mathbb{R}^{q \times p} \\ ||\beta||_1 \leq A}}{\operatorname{argmin}} \left\{ ||Y - \beta X||_2^2 \right\} .$$

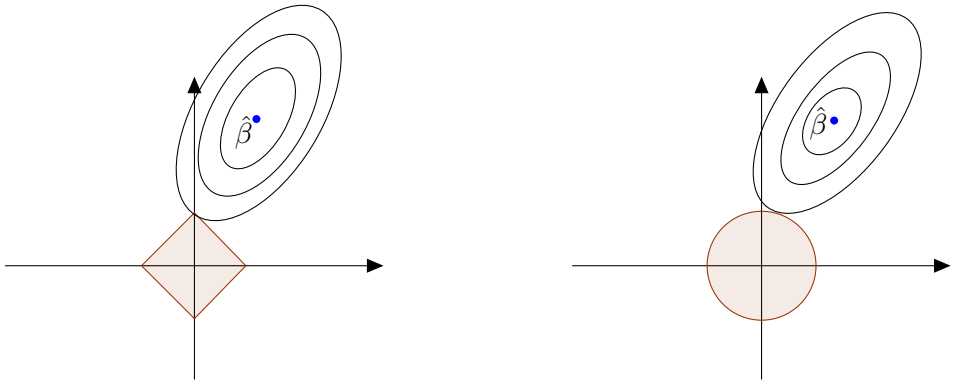

FIGURE 2 – Illustration de l'estimateur du Lasso (à gauche) et de l'estimateur Ridge (à droite). $\hat{\beta}$ représente l'estimateur des moindres carrés, les lignes correspondant à l'erreur des moindres carrés, la boule ℓ_1 correspond au problème de l'estimateur du Lasso et la boule ℓ_2 correspond au problème de l'estimateur Ridge. Cette figure est issue de [Tib96].

Cet estimateur a été beaucoup étudié ces dernières années. Citons Friedman, Hastie et Tibshirani, [HTF01] qui ont étudié le chemin de régularisation du Lasso ; Bickel, Ritov et Tsybakov, [BRT09], qui ont étudié cet estimateur en comparaison avec le sélecteur Dantzig. On peut aussi citer Osborne, [OPT99], qui étudie l'estimateur du Lasso et sa forme duale.
Notons que la norme ℓ_1 apparaît ici comme une relaxation convexe de la pénalité ℓ_0 (où $||\beta||_0 = \text{Card}(j|\beta_j \neq 0)$), qui vise à estimer les coefficients par 0.
Grâce à la géométrie de la boule ℓ_1, le Lasso a tendance à annuler des coefficients, comme on peut le voir sur la figure 2. Plus λ est grand, plus on pénalise, et plus on a de coefficients de $\hat{\beta}^{\text{Lasso}}(\lambda)$ qui sont nuls. Si $\lambda = 0$, on retrouve l'estimateur du maximum de vraisemblance.
Résumons les résultats principaux obtenus ces dernières années pour l'estimateur du Lasso. Cet estimateur peut facilement être approximé par l'algorithme LARS, introduit dans [EHJT04]. La relaxation convexe de la pénalité ℓ_0 par la pénalité ℓ_1 permet d'approcher numériquement plus facilement cet estimateur. On sait qu'il est linéaire par morceaux, et on peut expliciter ses valeurs grâce aux conditions de Karush-Kuhn-Tucker. Citons par exemple [EHJT04] ou [ZHT07].
D'un point de vue théorique, sous des hypothèses plus ou moins fortes, il existe des inégalités oracle pour l'erreur de prédiction ou l'erreur ℓ_q des coefficients estimés, avec un paramètre de régularisation de l'ordre de $\sqrt{\log(p)/n}$. On peut citer par exemple Bickel, Ritov et Tsybakov, [BRT09], qui obtiennent une inégalité oracle pour le risque en prédiction dans un modèle général non paramétrique en régression, et une inégalité oracle pour la perte ℓ_p en estimation ($1 \leq p \leq 2$) pour le modèle linéaire. Les hypothèses nécessaires et les résultats obtenus dans ce sens sont résumés par van de Geer et Bühlmann dans [vdGB09].
Il est aussi à noter que l'estimateur du Lasso a de bonnes propriétés de sélection de variables. Citons par exemple Meinshausen et Bühlmann [MB06], Zhang et Huang, dans [ZH08], Zhao et Yu, dans [ZY06], ou Meinshausen et Yu, [MY09], qui montrent que le Lasso est consistant en sélection de variables sous diverses hypothèses plus ou moins contraignantes.
Il paraît donc cohérent d'utiliser l'estimateur du Lasso pour sélectionner les variables importantes.
Dans les modèles de mélange de modèles linéaires Gaussiens, on peut étendre la définition de l'estimateur du Lasso par

$$\hat{\xi}^{\text{Lasso}}(\lambda) = \underset{\xi \in \Xi_K}{\text{argmin}} \left\{ -\frac{1}{n} l_\lambda(\xi, \mathbf{x}, \mathbf{y}) \right\} ; \tag{5}$$

où

$$l_\lambda(\xi, \mathbf{x}, \mathbf{y}) = l(\xi, \mathbf{x}, \mathbf{y}) - \lambda \sum_{k=1}^{K} \pi_k ||P_k \beta_k||_1 \,;$$
$$\Xi_K = \Pi_K \times (\mathbb{R}^{q\times p})^K \times \left(\mathbb{S}_q^{++}\right)^K \,;$$

où P_k est la racine de Cholesky de l'inverse de la matrice de covariance, i.e. $P_k^t P_k = \Sigma_k^{-1}$, pour tout $k \in \{1, \dots, K\}$.
Cette définition a été introduite par Städler et coauteurs dans [SBG10]. La pénalité est différente de celle de l'estimateur (4). Premièrement, on ne pénalise non pas par la moyenne conditionnelle dans chaque classe, mais par une version reparamétrisée par la variance. En effet, dans les modèles de mélange, il est important d'avoir un bon estimateur de la variance, et de l'estimer en même temps que la moyenne, pour ne pas favoriser les classes à variance trop élevée. De plus, pour avoir un estimateur invariant par changement d'échelle, il faut prendre en compte la variance dans la pénalité ℓ_1. Ensuite, l'étude de la racine de Cholesky de l'inverse de la matrice de covariance plutôt que de la matrice de covariance permet d'obtenir une optimisation convexe, ce qui facilitera la partie algorithmique. Enfin, on pénalise la vraisemblance par la somme de l'estimateur de la moyenne reparamétrisé par la variance, pondérée sur chaque classe, pour prendre en compte la différence de taille entre les différentes classes.
Notons que la sélection de variables dans les modèles de mélange a déjà été envisagée dans des problèmes d'estimation. Citons par exemple Raftery et Dean, [RD06], ou Maugis et Michel dans [MM11b].
L'estimateur (5) peut s'approcher algorithmiquement à l'aide d'une généralisation de l'algorithme EM, introduite pas Städler et coauteurs dans le cas univarié, et étendue dans cette thèse.
Dans le cadre des modèles de mélange, en régression, on connaît principalement deux résultats théoriques pour l'estimateur du Lasso, valables pour Y réel et à nombre de classes K fixé et connu. Pour $Y \in \mathbb{R}$, et $X \in \mathbb{R}^p$, Städler et coauteurs, dans [SBG10], ont montré que, sous la condition de valeurs propres restreintes (notée REC, citée ci-dessous), l'estimateur du Lasso vérifie une inégalité oracle pour des covariables fixes. Rappelons le contexte de mélange de Gaussiennes en régression univariées. Si Y, conditionnellement à X, appartient à la classe k, on note $Y = \beta_k X + \epsilon$, avec $\epsilon \sim \mathcal{N}(0, \sigma_k^2)$. On note de plus $\phi_k = \sigma_k^{-1}\beta_k$, et J l'ensemble des indices des coefficients non nuls de la matrice de régression.

Hypothèse REC. *Il existe $\kappa \geq 1$ tel que, pour tout $\phi \in (\mathbb{R}^p)^K$ vérifiant $||\phi_{J^c}|| \leq 6||\phi_J||_1$, on a*

$$||\phi_J||_2^2 \leq \kappa^2 \sum_{k=1}^{K} \phi_k^t \hat{\Sigma}_{\mathbf{x}} \phi_k$$

où $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n x_i^t x_i$.

Dans le même cadre, mais sans l'hypothèse REC, Meynet, dans [Mey13], a montré une inégalité oracle ℓ_1 pour l'estimateur du Lasso.
Dans cette thèse, nous nous sommes intéressés aux propriétés de régularisation ℓ_1 de l'estimateur du Lasso, dans notre cadre de modèles de mélange en régression multivariée. On fixe les variables explicatives $\mathbf{x} = (x_1, \dots, x_n)$. Sans restriction, on peut supposer que $x_i \in [0,1]^p$ pour tout $i \in \{1, \dots, n\}$. On suppose qu'il existe $(A_\beta, a_\Sigma, A_\Sigma, a_\pi)$ des réels positifs, qui définissent

l'ensemble des paramètres

$$\tilde{\Xi}_K = \left\{ \xi \in \Xi_K \,\middle|\, \text{pour tout } k \in \{1,\dots,K\}, \max_{z\in\{1,\dots,q\}} \sup_{x\in[0,1]^p} |[\beta_k x]_z| \leq A_\beta, \right.$$
$$\left. a_\Sigma \leq m(\Sigma_k^{-1}) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k \right\}; \quad (6)$$

où $m(A)$ et $M(A)$ désignent respectivement la valeur absolue de la plus petite et de la plus grande valeur propre de la matrice A. Soit, pour $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \Xi_K$,

$$N_1^{[2]}(s_\xi^K) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^K \sum_{j=1}^p \sum_{z=1}^q |[\beta_k]_{z,j}|. \quad (7)$$

la pénalité envisagée, et KL_n la divergence de Kullback-Leibler à design fixé :

$$\begin{aligned}\mathrm{KL}_n(s,t) &= \frac{1}{n}\sum_{i=1}^n \mathrm{KL}(s(.|x_i), t(.|x_i)) \\ &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}_s\left(\log\left(\frac{s(.|x_i)}{t(.|x_i)}\right)\right).\end{aligned}$$

Voici le théorème que nous obtenons.

Inégalité oracle ℓ_1 pour le Lasso. *Soit $(\boldsymbol{x}, \boldsymbol{y}) = (x_i, y_i)_{1\leq i\leq n}$ les observations, issues d'une densité conditionnelle inconnue $s^* = s_{\xi_0}$, où $\xi_0 \in \tilde{\Xi}_K$, cet ensemble étant défini par l'équation* (6)*, et le nombre de classes K étant fixé. On notera $a \vee b = \max(a,b)$. Soit $N_1^{[2]}(s_\xi^K)$ définie par* (7)*. Pour $\lambda \geq 0$, on définit l'estimateur du Lasso, noté $\hat{s}^{Lasso}(\lambda)$, par*

$$\hat{s}^{Lasso}(\lambda) = \underset{s_\xi^K \in S}{\operatorname{argmin}} \left(-\frac{1}{n}\sum_{i=1}^n \log(s_\xi^K(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi^K) \right); \quad (8)$$

avec

$$S = \left\{ s_\xi^K, \xi \in \tilde{\Xi}_K \right\}.$$

Si

$$\lambda \geq \kappa \left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right)\sqrt{\frac{K}{n}}\left(1 + q\log(n)\sqrt{K\log(2p+1)}\right)$$

avec κ une constante positive, alors l'estimateur (8) *vérifie l'inégalité suivante :*

$$\begin{aligned}\mathrm{E}[\mathrm{KL}_n(s^*, \hat{s}^{Lasso}(\lambda))] \leq &(1+\kappa^{-1}) \inf_{s_\xi^K \in S}\left(\mathrm{KL}_n(s^*, s_\xi^K) + \lambda N_1^{[2]}(s_\xi^K)\right) + \lambda \\ &+ \kappa' \sqrt{\frac{K}{n}} \frac{e^{-\frac{1}{2}}\pi^{q/2} a_\pi}{A_\Sigma^{q/2}} \sqrt{2q} \\ &+ \kappa' \frac{K^{3/2}}{\sqrt{n}}\left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right) \\ &\times \left(1 + A_\beta + \frac{q}{a_\Sigma}\right)^2;\end{aligned}$$

où κ' est une constante positive.

Ce théorème peut être vu comme une inégalité oracle ℓ_1, mais ce n'est pas l'approche que l'on souhaite développer ici. En effet, la démonstration de ce théorème passe par un théorème de sélection de modèles, mais ce thème sera abordé dans la partie correspondante. Ici, on voit plutôt ce théorème comme une assurance que l'estimateur du Lasso fonctionne bien pour la régularisation ℓ_1. La particularité de ce résultat est qu'il ne requiert que peu d'hypothèses : on travaille avec des prédicteurs fixés, qui sont supposés (sans restriction) être inclus dans $[0,1]^p$, et les paramètres de nos densités conditionnelles sont supposés bornés, au sens où ils appartiennent à $\tilde{\Xi}_K$. Cette hypothèse est nécessaire, pour assurer en particulier que la vraisemblance est finie. On la retrouvera dans les autres théorèmes démontrés dans cette thèse. On remarque aussi que la borne sur le paramètre de régularisation λ n'est pas celle, optimale, classique, obtenue dans d'autres cas plus généraux, mais cela est dû au fait que nous ne faisons pas d'hypothèses sur les régresseurs. L'article de van de Geer, [vdG13], permet d'obtenir cette borne optimale sous des hypothèses plus fortes sur le design. Il est à noter que le design est supposé fixe ici.
Comme l'estimateur du Lasso surestime les paramètres (citons par exemple Fan et Peng, [FP04], ou Zhang [Zha10]), nous proposons de l'utiliser pour la sélection de variables, et non pour l'estimation des paramètres. Ainsi, pour un paramètre de régularisation $\lambda \geq 0$ à définir, on sélectionnera les variables importantes pour expliquer Y en fonction de X. Plus formellement, définissons la notion de variable active pour la classification.

Définition. *Une variable est* active *pour la classification si elle est non nulle dans au moins une classe : la variable d'indice* $(z,j) \in \{1,\dots,q\} \times \{1,\dots,p\}$ *est* active *s'il existe* $k \in \{1,\dots,K\}$ *tel que* $[\beta_k]_{z,j} \neq 0$.

Ainsi, pour un certain $\lambda \geq 0$, on peut estimer $\hat{\xi}^{\text{Lasso}}(\lambda)$ et en déduire l'ensemble J_λ des variables actives pour la classification.

Ré-estimations

En se restreignant aux variables sélectionnées indicées par J_λ par l'estimateur du Lasso de paramètre de régularisation $\lambda \geq 0$, on travaille avec un modèle de dimension plus petite. En effet, plutôt que $(pq+q^2+1)K-1$ paramètres à estimer, on en a $(|J_\lambda|+q^2+1)K-1$. Si on suppose de plus que la matrice de covariance est diagonale (ce qui implique que les variables sont non corrélées), on obtient un modèle de dimension $(|J_\lambda|+q+1)K-1$, et la dimension du modèle peut devenir plus petite que le nombre d'observations, ou au moins être raisonnable. On peut alors utiliser un autre estimateur, restreint aux variables sélectionnées, qui aura de bonnes propriétés d'estimation en dimension raisonnable (meilleures que celles de l'estimateur du Lasso).
Ré-estimer les paramètres sur les variables sélectionnées n'est pas une idée nouvelle. On veut tirer parti des avantages de la sélection de variables par l'estimateur du Lasso (ou par une autre technique), mais on veut aussi diminuer le biais induit par cet estimateur. Citons par exemple Belloni et Chernozhukov, [BC11], qui obtiennent une inégalité oracle pour montrer que l'estimateur du maximum de vraisemblance calculé sur les variables sélectionnées par le Lasso fonctionne mieux que l'estimateur du Lasso, pour un modèle linéaire en grande dimension. On peut aussi citer Zhang et Sun, [SZ12], qui estiment le bruit et la matrice de régression dans un modèle linéaire en grande dimension par l'estimateur des moindres carrés après sélection de modèles.
On propose dans une première procédure, appelée *procédure Lasso-EMV*, d'estimer les paramètres, en se restreignant aux variables actives, par l'estimateur du maximum de vraisemblance, qui a de bonnes propriétés pour un échantillon suffisamment grand.

On propose aussi, dans une seconde procédure que l'on appellera *procédure Lasso-Rang*, d'utiliser le maximum de vraisemblance avec une contrainte de faible rang. En effet, jusqu'ici, nous n'avons pas tenu compte de la structure matricielle de $\boldsymbol{\beta}$. Comme les matrices de covariance $(\Sigma_k)_{1\leq k\leq K}$ sont supposées diagonales, on aurait pu travailler avec chaque coordonnée de Y comme q problèmes distincts et indépendants. En cherchant une structure de faible rang, on suppose que peu de combinaisons linéaires de prédicteurs suffisent à expliquer la réponse. C'est aussi une seconde méthode pour diminuer la dimension, au cas où la sélection de variables par pénalisation ℓ_1 ne soit pas suffisante. Certains jeux de données sont particulièrement propices à cette réduction dimensionnelle par faible rang. On peut citer par exemple l'analyse d'image fMRI (Harrison, Penny, Frishen, [FHP03]), l'analyse de décodage de données EEG (Anderson, Stolz, Shamsunder, [ASS98]), la modélisation de réponse de neurones (Brown, Kass, Mitra, [BKM04]), ou encore l'analyse de données génomiques (Bunea, She, Wegkamp, [BSW11]). D'un point de vue plus théorique, on peut citer Izenman ([Ize75]) qui a introduit cette méthode dans le cas du modèle linéaire, Giraud ([Gir11]) ou Bunea et coauteurs ([Bun08]) qui ont complété l'étude théorique et pratique de sélection de rang. Dans cette thèse, nous employons ces méthodes dans un cadre de mélange en régression. Il est à noter qu'on aura sélectionné, par l'estimateur du Lasso, des colonnes, ou des lignes, ou les deux, mais que les lignes ou les colonnes où seuls certains coefficients sont inactifs ne seront pas sélectionnées, la structure matricielle étant requise pour estimer un paramètre par faible rang. On n'impose pas que toutes les moyennes conditionnelles aient le même rang.
Notons $\hat{\xi}_J^{EMV}$ et $\hat{\xi}_J^{Rang}$ les estimateurs associés à chaque procédure, avec J comme ensemble des variables actives :

$$\hat{\xi}_J^{EMV} = \underset{\xi\in\Xi_{(K,J)}}{\operatorname{argmin}} \left\{ -\frac{1}{n} l(\xi^{[J]}, \mathbf{x}, \mathbf{y}) \right\} ; \tag{9}$$

$$\hat{\xi}_J^{Rang} = \underset{\xi\in\check{\Xi}_{(K,J,R)}}{\operatorname{argmin}} \left\{ -\frac{1}{n} l(\xi^{[J]}, \mathbf{x}, \mathbf{y}) \right\} ; \tag{10}$$

où $\check{\Xi}_{(K,J,R)} = \left\{ \xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \Xi_{(K,J)} \middle| \operatorname{rang}(\beta_k) = R_k \text{ pour tout } k \in \{1, \ldots, K\} \right\}$, où $\xi^{[J]}$ signifie que l'on a sélectionné les variables d'indice J, et où $\Xi_{(K,J)}$ est défini par

$$\Xi_{(K,J)} = \Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K$$
$$\Pi_K = \left\{ (\pi_1, \ldots, \pi_K) \in (0,1)^K \middle| \sum_{k=1}^K \pi_k = 1 \right\}$$

A noter que l'on ré-estime tous les paramètres de notre modèle : les moyennes conditionnelles, les variances et les poids.
D'un point de vue pratique, la généralisation de l'algorithme EM (Algorithme 1 page 14) permet de calculer l'estimateur du maximum de vraisemblance, sous contrainte de faible rang ou non, dans le cas de mélange de Gaussiennes en régression.

Sélection de modèles

Pour un paramètre de régularisation $\lambda \geq 0$ fixé, après avoir ré-estimé nos paramètres, nous obtenons un modèle associé à nos observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$ qui est de dimension raisonnable et qui est bien estimé, si λ a été bien choisi. Cependant, on a dû faire de nombreux choix pour construire ce modèle. Pour le modèle de mélange, il est possible qu'on ne connaisse pas au préalable le nombre de classes K, il faut donc le choisir ; pour la

sélection de variables, construire J_λ correspond à sélectionner un paramètre de régularisation $\lambda \geq 0$; dans les cas de ré-estimation par faible rang, on doit sélectionner le vecteur des rangs dans chaque composante.
Dans chacun de ces trois cas, différentes méthodes existent dans la littérature. Pour le paramètre λ de régularisation du Lasso, on peut citer par exemple le livre de van de Geer et Bühlmann, [BvdG11], où une valeur λ proportionnelle à $\sqrt{\log(p)/n}$ est considérée comme optimale.
Pour le choix du nombre de classes K et du vecteur de rangs R, de nombreux auteurs se ramènent à un problème de sélection de modèles.
Dans cette thèse, nous allons voir le problème de sélection de paramètres comme un problème de sélection de modèles, en construisant une collection de modèles, avec plus ou moins de classes, et plus ou moins de coefficients actifs. Il restera à choisir un modèle parmi cette collection.
En toute généralité, notons $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ la collection de modèles que l'on considère, indicée par $\mathcal{M}$. Il est à noter que, contrairement à l'idée que l'on pourrait s'en faire, avoir une collection de modèles trop grande peut porter préjudice, par exemple en sélectionnant des estimateurs non consistants (voir Bahadur, [Bah58]) ou sous-optimaux (voir Birgé et Massart, [BM93]). C'est ce qu'on appelle le paradigme du choix de modèle.
On considère une *fonction de contraste*, notée γ, telle que $s^* = \operatorname*{argmin}_{t\in\mathcal{S}} \mathrm{E}(\gamma(t))$. La *fonction de perte* associée, notée l, est définie par

$$\text{pour tout } t \in \mathcal{S}, \qquad l(s^*, t) = \mathrm{E}(\gamma(t)) - \mathrm{E}(\gamma(s^*)).$$

On définit aussi le *contraste empirique* γ_n par

$$\text{pour tout } t \in \mathcal{S}, \qquad \gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t, x_i, y_i)$$

pour un échantillon $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{1\leq i\leq n}$. Pour le modèle m, on considère $\hat{s}_m$ la densité qui minimise le contraste empirique γ_n. C'est cette densité que l'on va utiliser pour représenter ce modèle. Par exemple, on peut prendre la log-vraisemblance comme contraste, et la divergence de Kullback-Leibler comme fonction de perte.
Le but de la sélection de modèles est de sélectionner le meilleur estimateur parmi la collection $(\hat{s}_m)_{m\in\mathcal{M}}$. Le meilleur estimateur peut être défini comme l'estimateur qui minimise le risque avec la vraie densité notée s^*. Cet estimateur, et le modèle correspondant, seront appelés dans cette thèse *oracle* (voir Donoho et Johnstone par exemple, [DJ94]). On note

$$m_O = \operatorname*{argmin}_{m\in\mathcal{M}} [l(s^*, \hat{s}_m)]. \tag{11}$$

Malheureusement, on ne peut pas évaluer cette quantité, puisque l'on n'a pas accès à s^*. On va utiliser l'oracle en théorie pour évaluer notre sélection de modèles : on veut que le risque associé à l'estimateur du modèle sélectionné soit le plus près possible de celui de l'oracle. Une partie des résultats théoriques en sélection de modèles sont des *inégalités oracle*, qui permettent d'assurer la cohérence de la sélection de modèles. Ces inégalités sont de la forme, pour $\hat{m}$ l'indice du modèle sélectionné,

$$\mathrm{E}(l(s^*, \hat{s}_{\hat{m}})) \leq C_1\, \mathrm{E}(l(s^*, s_O)) + \frac{C_2}{n} \tag{12}$$

où (C_1, C_2) sont des constantes absolues, avec C_1 la plus proche possible de 1. L'inégalité oracle est dite *exacte* si $C_1 = 1$.

Désormais, décrivons comment sélectionner un modèle. On va minimiser un critère pénalisé, pour parvenir à un compromis biais/variance. En effet, on peut décomposer le risque de la façon suivante :

$$l(s^*, \hat{s}_m) = \underbrace{l(s^*, s_m)}_{\text{biais}_m} + \underbrace{\mathrm{E}\left(\gamma(s_m) - \gamma(\hat{s}_m)\right)}_{\text{variance}_m};$$

où $s_m \in \underset{t \in S_m}{\operatorname{argmin}} [\mathrm{E}(\gamma(t))]$ (c'est une des meilleures approximations de s^* dans S_m). Or, pour minimiser le biais, il faut un modèle complexe, qui colle de très près aux données ; et pour minimiser la variance, il ne faut pas considérer des modèles trop complexes, pour ne pas surapprendre des données.
Une méthode pour rendre compte de cette remarque est de considérer le contraste empirique L_n pénalisé par la dimension : on va pénaliser les modèles trop complexes, qui suraprennent de nos données. Soit pen : $\mathcal{M} \to \mathbb{R}^+$ une pénalité à construire ; on va sélectionner

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{-\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\right\}.$$

Akaike et Schwarz ont introduit cette méthode pour l'estimation avec la vraisemblance, voir respectivement [Aka74] et [Sch78]. Ils proposaient les critères désormais classiques AIC et BIC, où la pénalité vaut respectivement

$$\begin{aligned} \mathrm{pen}_{AIC}(m) &= D_m; \\ \mathrm{pen}_{BIC}(m) &= \frac{\log(n) D_m}{2}; \end{aligned}$$

où D_m est la dimension du modèle m, et n est la taille de l'échantillon considéré. Ces critères sont fortement utilisés aujourd'hui. Il est à noter qu'ils sont basés sur des approximations asymptotiques (AIC sur le théorème de Wilks, et BIC sur une approche bayésienne), et on peut se méfier de leur comportement en non asymptotique. Ils supposent de plus, en toute rigueur, une collection de modèles fixée.
Mallows, au même moment, dans [Mal73], a étudié cette méthode dans le cadre de la régression linéaire. Il a obtenu

$$\mathrm{pen}_{Mallows}(m) = \frac{2 D_m \sigma^2}{n}$$

où σ^2 est la variance des erreurs, supposée connue.
Birgé et Massart, dans [BM01], ont introduit l'*heuristique des pentes*, qui est une méthodologie non asymptotique pour sélectionner un modèle parmi une collection de modèles.
Décrivons les idées de cette heuristique. On cherche une pénalité proche de $m \in \mathcal{M} \mapsto l(s^*, \hat{s}_m) - \gamma_n(\hat{s}_m)$. Comme on ne connaît pas s^*, on va essayer d'approcher cette quantité :

$$\begin{aligned} l(s^*, \hat{s}_m) - \gamma_n(\hat{s}_m) = & \underbrace{\mathrm{E}(\gamma(\hat{s}_m)) - \mathrm{E}(\gamma(s_m))}_{v_m} + \underbrace{\mathrm{E}(\gamma(s_m)) - \mathrm{E}(\gamma(s^*))}_{(1)} \\ & - \underbrace{\left(\gamma_n(\hat{s}_m) - \gamma_n(s_m)\right)}_{\hat{v}_m} - \underbrace{\left(\gamma_n(s_m) - \gamma_n(s^*)\right)}_{(2)} - \gamma_n(s^*) \end{aligned}$$

où v_m peut être vue comme un terme de variance, et $\hat{v}_m$ comme une version empirique. On définit $\Delta_n(s_m) = (1) + (2)$, qui correspond à la différence entre le terme de biais et sa version

empirique. Si on choisit $\text{pen}(m) = \hat{v}_m$, on va choisir un modèle qui limite le biais mais pas la variance : on va sélectionner un modèle trop complexe. Cette pénalité est *minimale* : si on pose $\text{pen}(m) = \kappa\hat{v}_m$, si $\kappa < 1$ on va choisir un modèle trop complexe, et si $\kappa > 1$, la dimension du modèle sera plus raisonnable.
En fait, la pénalité *optimale* est le double de la pénalité minimale. Comme $\hat{v}_m$ est la version empirique de v_m, $v_m \approx \hat{v}_m$. Comme $\Delta_n(s_m)$ est d'espérance nulle, on peut contrôler ses fluctuations. On a donc envie de choisir $\text{pen}(m) = 2v_m$.
Ainsi, on peut trouver, sur un jeu de données, si on utilise l'heuristique des pentes, la pénalité qui nous permettra de sélectionner un modèle : soit on cherche le plus grand saut de dimension, soit on regarde la pente asymptotique de $\gamma_n(s_m)$, ce qui nous donne la pénalité minimale, et il suffit de la multiplier par deux pour obtenir la pénalité optimale.
Les figures 3 et 4 illustrent ces idées.

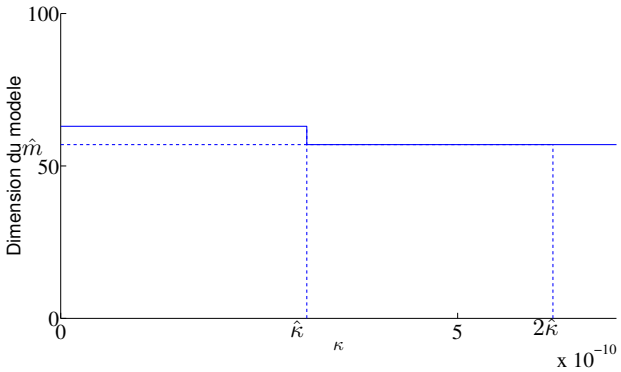

FIGURE 3 – Illustration de l'heuristique des pentes : on estime κ par $\hat{\kappa}$ le plus grand saut de dimension. On sélectionne alors le modèle qui minimise la log-vraisemblance pénalisée par $2\hat{\kappa}$.

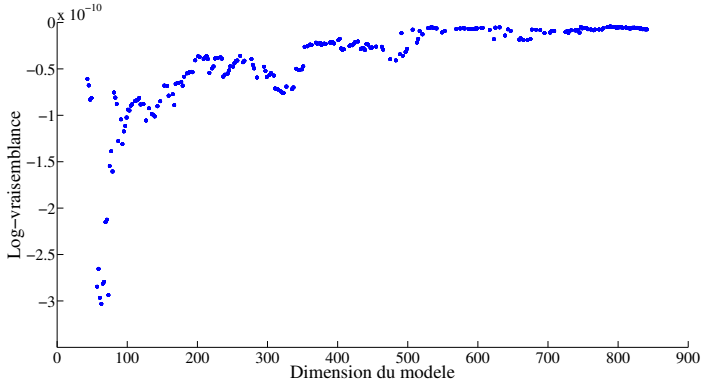

FIGURE 4 – Illustration de l'heuristique des pentes : on estime κ par $\hat{\kappa}$ la pente asymptotique de la log-vraisemblance.

D'un point de vue pratique, on utilise le package Capushe, développé par Baudry et coauteurs dans [BMM12] sur le logiciel Matlab.
D'un point de vue théorique, on a obtenu des inégalités oracle, qui justifient la sélection de modèles dans chacune de nos procédures.
Citons le théorème général issu de [Mas07] qui est à la base de nos résultats théoriques.
On travaille avec la log-vraisemblance comme contraste empirique. On note KL la divergence de

Kullback-Leibler, définie par

$$\mathrm{KL}(s,t) = \begin{cases} \mathrm{E}_s\left(\log\left(\frac{s}{t}\right)\right) & \text{si } s << t \\ +\infty \text{ sinon.} \end{cases}$$

D'abord, nous avons besoin d'une hypothèse structurelle. C'est une condition sur le crochet d'entropie du modèle S_m par rapport à la *distance de Hellinger*, définie par

$$(d_H(s,t))^2 = \frac{1}{2}\int (s-t)^2.$$

Un *crochet* $[l,u]$ est une paire de fonctions telles que pour tout $y, l(y) \leq s(y) \leq u(y)$. Pour $\epsilon > 0$, on définit l'*entropie à crochet* $\mathcal{H}_{[.]}(\epsilon, S, d_H)$ d'un ensemble S par le logarithme du nombre minimal de crochets $[l,u]$ de largeur $d_H(l,u)$ inférieure à ϵ telle que toutes les densités de S appartiennent à l'un de ces crochets.
Soit $m \in \mathcal{M}$.

Hypothèse $\mathbf{H}_m$. *Il existe une fonction croissante ϕ_m telle que $\varpi \mapsto \frac{1}{\varpi}\phi_m(\varpi)$ est décroissante sur $(0,+\infty)$ et telle que pour tout $\varpi \in \mathbb{R}^+$ et tout $s_m \in S_m$,*

$$\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, S_m(s_m,\varpi), d_H)}d\epsilon \leq \phi_m(\varpi);$$

où $S_m(s_m,\varpi) = \{t \in S_m, d_H(t,s_m) \leq \varpi\}$. La complexité du modèle *$\mathcal{D}_m$ est alors définie par $n\varpi_m^2$ avec ϖ_m l'unique solution de*

$$\frac{1}{\varpi}\phi_m(\varpi) = \sqrt{n}\varpi. \tag{13}$$

Notons que la complexité du modèle ne dépend pas du crochet d'entropie des modèles globaux S_m, mais des ensembles plus petits, localisés. C'est une hypothèse plus faible.
Pour des raisons techniques, une hypothèse de séparabilité est aussi nécessaire.

Hypothèse $\mathbf{Sep}_m$. *Il existe un ensemble dénombrable $S_m^{'}$ de S_m et un ensemble $\mathcal{Y}_m^{'}$ avec $\lambda(\mathbb{R}^q \setminus \mathcal{Y}_m^{'}) = 0$, pour λ la mesure de Lebesgue, tel que pour tout $t \in S_m$, il existe une suite $(t_l)_{l\geq 1}$ d'éléments de $S_m^{'}$ telle que pour tout $y \in \mathcal{Y}_m^{'}$, $\log(t_l(y))$ tend vers $\log(t(y))$ quand l tend vers l'infini.*

On a aussi besoin d'une hypothèse de théorie de l'information sur notre collection de modèles.

Hypothèse K. *La famille de nombres positifs $(w_m)_{m\in\mathcal{M}}$ vérifie*

$$\sum_{m\in\mathcal{M}} e^{-w_m} \leq \Omega < +\infty.$$

Alors, on peut écrire le théorème général de sélection de modèles.

Inégalité oracle pour une famille d'EMV. *Soient $(X_1,\ldots,X_n)$ des variables aléatoires de densité inconnue s^*. On en observe une réalisation $(x_1,\ldots,x_n)$. Soit $\{S_m\}_{m\in\mathcal{M}}$ une collection de modèles au plus dénombrable, où, pour tout $m \in \mathcal{M}$, les éléments de S_m sont des densités de probabilité, et S_m vérifie l'hypothèse Sep_m. On considère de plus la collection d'estimateurs de maxima de vraisemblance à ρ près notés $(\hat{s}_m)_{m\in\mathcal{M}}$: on a, pour tout $m \in \mathcal{M}$,*

$$-\frac{1}{n}\sum_{i=1}^n \ln(\hat{s}_m(x_i)) \leq \inf_{t\in S_m} -\frac{1}{n}\sum_{i=1}^n \ln(t(x_i)) + \rho.$$

Soient $\{w_m\}_{m\in\mathcal{M}}$ *une famille de nombres positifs vérifiant l'hypothèse K, et, pour tout* $m \in \mathcal{M}$*, on considère* ϕ_m *qui vérifie la condition* H_m*, avec* ϖ_m *l'unique solution positive de l'équation*

$$\phi_m(\varpi) = \sqrt{n}\varpi^2.$$

On suppose de plus que, pour tout $m \in \mathcal{M}$*, l'hypothèse* Sep_m *est vérifiée.*
Soit $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ *et soit le critère de log-vraisemblance pénalisé*

$$\mathrm{crit}(m) = -\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_m(x_i))) + \mathrm{pen}(m).$$

Alors, il existe des constantes κ *et* C *telles que, lorsque*

$$\mathrm{pen}(m) \geq \kappa\left(\varpi_m^2 + \frac{w_m}{n}\right)$$

pour tout $m \in \mathcal{M}$*, alors il existe* $\hat{m}$ *qui minimise le critère* crit *sur* $\mathcal{M}$*, et de plus,*

$$\mathrm{E}_s(d_H^2(s, \hat{s}_{\hat{m}})) \leq C\left(\inf_{m\in\mathcal{M}}\left(\inf_{t\in S_m} \mathrm{KL}(s,t) + \mathrm{pen}(m)\right) + \rho + \frac{\Omega}{n}\right)$$

où d_H *est la distance de Hellinger, et* KL *la divergence de Kullback-Leibler.*

Ce théorème nous indique que, si notre collection de modèles est bien construite (c'est-à-dire satisfait les hypothèses H_m, K et Sep_m), on peut trouver une pénalité telle que le modèle minimisant le critère pénalisé satisfasse une inégalité oracle.
Cette approche a déjà été envisagée pour sélectionner le nombre de classes d'un modèle de mélange. On peut citer par exemple Maugis et Michel, [MM11b], ou Maugis et Meynet, [MMR12]. Ces auteurs voient le problème de sélection du nombre de composantes et de sélection de variables comme un problème de sélection de modèles.
Pour le rang, Giraud, dans [Gir11], et Bunea, dans [Bun08], proposent une pénalité pour choisir de façon optimale le rang. Ils obtiennent, à variance connue et inconnue, des inégalités oracle qui permettent de sélectionner le rang, où la pénalité est proportionnelle au rang. Ma et Sun, dans [MS14], obtiennent pour ces modèles une borne minimax, ce qui revient à dire que la pénalité construite est optimale pour la sélection du rang.
Les procédures d'estimation que l'on a décrites dans la partie précédente sont sujettes à des choix de paramètres (le nombre de classes, le paramètre de régularisation du Lasso, et le rang dans la deuxième procédure). On peut réécrire la sélection de ces paramètres comme un problème de sélection de modèles : en faisant varier ces paramètres, on obtient une collection de modèles.
Commençons par définir les collections de modèles associés à chacune de nos procédures.
Pour la procédure Lasso-EMV, pour $(K, J) \in \mathcal{K} \times \mathcal{J}$,

$$\begin{aligned}
\mathcal{S}_{(K,J)} &= \left\{y \in \mathbb{R}^q \mapsto s_\xi^{(K,J)}(y|x)\right\} \qquad (14)\\
s_\xi^{(K,J)}(y|x) &= \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi\det(\Sigma_k)}} \exp\left(-\frac{1}{2}(y-\beta_k^{[J]}x)^t\Sigma_k^{-1}(y-\beta_k^{[J]}x)\right)\\
\xi &= (\pi_1, \ldots, \pi_K, \beta_1^{[J]}, \ldots, \beta_K^{[J]}, \Sigma_1, \ldots, \Sigma_K) \in \Xi_{(K,J)}\\
\Xi_{(K,J)} &= \Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K
\end{aligned}$$

Pour la procédure Lasso-Rang, pour $(K, J, R) \in \mathcal{K} \times \mathcal{J} \times \mathcal{R}$,

$$\mathcal{S}_{(K,J,R)} = \left\{ y \in \mathbb{R}^q \mapsto s_\xi^{(K,J,R)}(y|x) \right\} \tag{15}$$

$$s_\xi^{(K,J,R)}(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi \det(\Sigma_k)}} \exp\left(-\frac{1}{2}(y - (\beta_k^{R_k})^{[J]}x)^t \Sigma_k^{-1} (y - (\beta_k^{R_k})^{[J]}x) \right)$$

$$\xi = (\pi_1, \ldots, \pi_K, (\beta_1^{R_1})^{[J]}, \ldots, (\beta_K^{R_K})^{[J]}, \Sigma_1, \ldots, \Sigma_K) \in \Xi_{(K,J,R)}$$

$$\check{\Xi}_{(K,J,R)} = \Pi_K \times \Psi_{(K,J,R)} \times (\mathbb{S}_q^{++})^K$$

$$\Psi_{(K,J,R)} = \left\{ ((\beta_1^{R_1})^{[J]}, \ldots, (\beta_K^{R_K})^{[J]}) \in (\mathbb{R}^{q \times p})^K \middle| \text{ pour tout } k \in \{1, \ldots, K\}, \mathrm{Rang}(\beta_k) = R_k \right\}.$$

où $\mathcal{K}$ est l'ensemble des valeurs possibles pour le nombre de composantes, $\mathcal{J}$ est l'ensemble des ensembles d'indices de variables actives possibles, et $\mathcal{R}$ est l'ensemble des valeurs possibles pour les vecteurs de rang.
Pour obtenir des résultats théoriques, on a besoin de borner nos paramètres. On considère

$$\mathcal{S}_{(K,J)}^{\mathcal{B}} = \left\{ s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)} \middle| \xi \in \tilde{\Xi}_{(K,J)} \right\} \tag{16}$$

$$\tilde{\Xi}_{(K,J)} = \Pi_K \times ([-A_\beta, A_\beta]^{|J|})^K \times ([a_\Sigma, A_\Sigma]^q)^K \tag{17}$$

et

$$\mathcal{S}_{(K,J,R)}^{\mathcal{B}} = \left\{ s_\xi^{(K,J,R)} \in \mathcal{S}_{(K,J,R)} \middle| \xi \in \tilde{\Xi}_{(K,J,R)} \right\} \tag{18}$$

$$\tilde{\Xi}_{(K,J,R)} = \Pi_K \times \tilde{\Psi}_{(K,J,R)} \times ([a_\Sigma, A_\Sigma]^q)^K \tag{19}$$

$$\tilde{\Psi}_{(K,J,R)} = \left\{ (\beta_1^{R_1}, \ldots, \beta_K^{R_K}) \in \Psi_{(K,J,R)} \middle| \text{ pour tout } k \in \{1, \ldots, K\}, \beta_k^{R_k} = \sum_{l=1}^{R_k} \sigma_l u_l^t v_l, \sigma_l < A_\sigma \right\}.$$

Comme on travaille en régression, on définit une version tensorisée de la divergence de Kullback-Leibler

$$\mathrm{KL}^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(s(.|x_i), t(.|x_i)) \right] ;$$

une version tensorisée de la distance d'Hellinger,

$$(d_H^{\otimes_n})^2(s,t) = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} d_H^2(s(.|x_i), t(.|x_i)) \right].$$

On a aussi besoin de la divergence de Jensen-Kullback-Leibler, définie par, pour $\rho \in (0,1)$, par

$$\mathrm{JKL}_\rho(s,t) = \frac{1}{\rho} \mathrm{KL}(s, (1-\rho)s + \rho t);$$

et de sa version tensorisée

$$\mathrm{JKL}_\rho^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathrm{JKL}_\rho(s(.|x_i), t(.|x_i)) \right].$$

Le nombre de classes est supposé inconnu, on va l'estimer ici, contrairement à l'inégalité oracle ℓ_1 pour le Lasso. On suppose que les covariables, bien qu'aléatoires, sont bornées. Pour simplifier la lecture, on suppose que $X \in [0,1]^p$.
Dans cette thèse, on obtient alors les deux théorèmes suivants.

Inégalité oracle Lasso-EMV. *Soit $(x_i, y_i)_{i\in\{1,\dots,n\}}$ les observations, issues d'une densité conditionnelle inconnue s^*. Soit $\mathcal{S}_{(K,J)}$ définie par (14). On considère $\mathcal{J}^L \subset \mathcal{J}$ la sous-collection d'ensembles d'indices construite en suivant le chemin de régularisation de l'estimateur du Lasso. Pour $(K, J) \in \mathcal{K} \times \mathcal{J}^L$, notons $\mathcal{S}^{\mathcal{B}}_{(K,J)}$ le modèle défini par (16).*
On considère l'estimateur du maximum de vraisemblance

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi^{(K,J)}(y_i|x_i)) \right\}.$$

Notons $D_{(K,J)}$ la dimension du modèle $\mathcal{S}^{\mathcal{B}}_{(K,J)}$, $D_{(K,J)} = K(|J| + q + 1) - 1$. Soit $\bar{s}_\xi^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}$ telle que

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_\xi^{(K,J)}) \leq \inf_{t \in \mathcal{S}^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

et soit $\tau > 0$ tel que $\bar{s}_\xi^{(K,J)} \geq e^{-\tau} s^*$. Soit $\mathrm{pen} : \mathcal{K} \times \mathcal{J} \to \mathbb{R}_+$, et supposons qu'il existe une constante absolue $\kappa > 0$ telle que, pour tout $(K, J) \in \mathcal{K} \times \mathcal{J}$,

$$\begin{aligned}\mathrm{pen}(K, J) \geq \kappa \frac{D_{(K,J)}}{n} \Bigg[& B^2(A_\beta, A_\Sigma, a_\Sigma) - \log\left(\frac{D_{(K,J)}}{n} B^2(A_\beta, A_\Sigma, a_\Sigma) \wedge 1 \right) \\ & + (1 \vee \tau) \log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq} \right) \Bigg] \ ;\end{aligned}$$

où les constantes $A_\beta, A_\Sigma, a_\Sigma$ sont définies par (17). Si on sélectionne le modèle indicé par $(\hat{K}, \hat{J})$, où

$$(\hat{K}, \hat{J}) = \underset{(K,J) \in \mathcal{K} \times \mathcal{J}^L}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log(\hat{s}^{(K,J)}(y_i|x_i)) + \mathrm{pen}(K, J) \right\},$$

alors l'estimateur $\hat{s}_{(\hat{K},\hat{J})}$ vérifie, pour tout $\rho \in (0, 1)$,

$$\mathrm{E}\left[\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}})\right] \leq C\, \mathrm{E}\left(\inf_{(K,J) \in \mathcal{K} \times \mathcal{J}^L} \left(\inf_{t \in S_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \mathrm{pen}(K, J) \right) \right) + \frac{4}{n};$$

pour une constante C absolue.

Ce théorème nous donne une pénalité théorique pour laquelle le modèle minimisant le critère pénalisé a de bonnes propriétés d'estimation. Les constantes, bien que non optimales, sont explicites, en fonction des bornes de l'espace des paramètres. La pénalité est presque proportionnelle (à un terme logarithmique près) à la dimension du modèle. Le terme logarithmique a été étudié dans la thèse de Meynet, [Mey12]. Ici, en pratique, on prend la pénalité proportionnelle à la dimension.

Inégalité oracle Lasso-Rang. *Soit $(x_i, y_i)_{i\in\{1,\dots,n\}}$ les observations, issues d'une densité conditionnelle inconnue s^*. Pour $(K, J, R) \in \mathcal{K} \times \mathcal{J} \times \mathcal{R}$, soit $\mathcal{S}_{(K,J,R)}$ définie par (15), et $\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$ définie par (18). Soit $\mathcal{J}^L \subset \mathcal{J}$ une sous-collection construite en suivant le chemin de régularisation de l'estimateur du Lasso.*
Soit $\bar{s}^{(K,J,R)} \in S^{\mathcal{B}}_{(K,J,R)}$ telle que, pour $\delta_{\mathrm{KL}} > 0$,

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}^{(K,J,R)}) \leq \inf_{t \in S^{B}_{(K,J,R)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n}$$

et telle qu'il existe $\tau > 0$ tel que

$$\bar{s}^{(K,J,R)} \geq e^{-\tau} s^*. \tag{20}$$

On considère la collection d'estimateurs $\{\hat{s}^{(K,J,R)}\}_{(K,J,R)\in\mathcal{K}\times\mathcal{J}\times\mathcal{R}}$ de $\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$, vérifiant

$$\hat{s}^{(K,J,R)} = \underset{s_\xi^{(K,J,R)}\in\mathcal{S}^{\mathcal{B}}_{(K,J,R)}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log\left(s_\xi^{(K,J,R)}(y_i|x_i)\right)\right\}.$$

Notons $D_{(K,J,R)}$ la dimension du modèle $\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$. Soit $\mathrm{pen} : \mathcal{K}\times\mathcal{J}\times\mathcal{R} \to \mathbb{R}_+$ *définie par, pour tout $(K,J,R) \in \mathcal{K}\times\mathcal{J}\times\mathcal{R}$,*

$$\begin{aligned}\mathrm{pen}(K,J,R) \geq & \kappa\frac{D_{(K,J,R)}}{n}\left\{2B^2(A_\beta, A_\Sigma, a_\Sigma, A_\sigma) - \log\left(\frac{D_{(K,J,R)}}{n}B^2(A_\beta, A_\Sigma, a_\Sigma, A_\sigma)\wedge 1\right)\right.\\ &\left. + \log\left(\frac{4epq}{(D_{(K,J,R)}-q^2)\wedge pq} + \sum_{k=1}^{K} R_k\right)\right\},\end{aligned}$$

avec $\kappa > 0$ une constante absolue.
Alors, l'estimateur $\hat{s}_{(\hat{K},\hat{J},\hat{R})}$, avec

$$(\hat{K},\hat{J},\hat{R}) = \underset{(K,J,R)\in\mathcal{K}\times\mathcal{J}^L\times\mathcal{R}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{s}^{(K,J,R)}(y_i|x_i)) + \mathrm{pen}(K,J,R)\right\},$$

vérifie, pour tout $\rho \in (0,1)$,

$$\begin{aligned}&\mathrm{E}\left(\mathrm{JKL}_\rho^{\otimes n}\left(s^*, \hat{s}^{(\hat{K},\hat{J},\hat{R})}\right)\right)\\ &\leq C\,\mathrm{E}\left(\inf_{(K,J,R)\in\mathcal{K}\times\mathcal{J}^L\times\mathcal{R}}\left(\inf_{t\in\mathcal{S}_{(K,J,R)}} \mathrm{KL}^{\otimes n}(s^*,t) + \mathrm{pen}(K,J,R)\right)\right) + \frac{4}{n},\end{aligned} \tag{21}$$

pour une constante $C > 0$.

Ce théorème nous donne une pénalité théorique pour laquelle le modèle minimisant le critère pénalisé a de bonnes propriétés d'estimation.
Ces deux théorèmes ne sont pas des inégalités oracle exactes, à cause de la constante C, mais on contrôle cette constante. Ces résultats sont non-asymptotiques, ce qui nous permet de les utiliser dans notre cadre de grande dimension. Ils justifient l'utilisation de l'heuristique des pentes.

Données fonctionnelles

Ce travail de classification de données en régression a été développé en toute généralité, mais aussi plus particulièrement dans le cas des données fonctionnelles. L'analyse des données fonctionnelles s'est beaucoup développée, grâce notamment aux progrès techniques récents qui permettent d'enregistrer des données sur des grilles de plus en plus fines. Pour une analyse générale de ce type de données, on peut citer par exemple le livre de Ramsay et Silverman, [RS05] ou celui de Ferraty et Vieu, [FV06].
Dans cette thèse, on prend le parti de projeter les fonctions observées sur une base orthonormale. Cette approche a l'avantage de considérer l'aspect fonctionnel, par rapport à l'analyse multivariée de la discrétisation du signal, et de résumer le signal en peu de coefficients, si la base est bien choisie.

Plus précisément, on choisit de projeter les fonctions étudiées sur des bases d'ondelettes. On décompose de façon hiérarchique les signaux dans le domaine temps-échelle. On peut alors décrire une fonction à valeurs réelles par une approximation de cette fonction, et un ensemble de détails. Pour une étude générale des ondelettes, on peut citer Daubechies, [Dau92], ou Mallat, [Mal99].
Il est à noter que, par données fonctionnelles, dans notre cadre de modèles de mélange en régression, on entend soit des régresseurs fonctionnels et une réponse vectorielle (typiquement, l'analyse spectrophotométrique de données, comme le jeu de données classiques Tecator où la proportion de gras d'un morceau de viande est exprimée en fonction de la courbe de spectrophotométrie), soit des régresseurs vectoriels et une réponse fonctionnelle, soit des régresseurs et une réponse fonctionnels (comme la régression de la consommation électrique d'un jour sur la veille).
On considère un échantillon de signaux $(f_i)_{1\leq i\leq n}$ observés sur une grille de temps $\{t_1,\ldots,t_T\}$. On peut considérer l'échantillon $(f_i(t_j))_{\substack{1\leq i\leq n\\1\leq j\leq p}}$ pour faire notre analyse (ces observations sont soit les régresseurs, soit la réponse, soit les deux suivant la nature des données), mais on peut aussi considérer les projections de l'échantillon sur une base orthonormée $\mathcal{B}=\{\alpha_j\}_{j\in\mathbb{N}^*}$. Dans ce cas, il existe $(b_j)_{j\in\mathbb{N}^*}$ tels que, pour tout f, pour tout t,

$$f(t)=\sum_{j=1}^{\infty} b_j\alpha_j(t).$$

On peut choisir une base d'ondelettes $\mathcal{B}=\{\phi,\psi_{l,h}\}_{\substack{l\geq 0\\0\leq h\leq 2^l-1}}$, où

- ψ est une ondelette réelle, vérifiant $\psi\in L^1\cap L^2$, $t\psi\in L^1$, et $\int_{\mathbb{R}}\psi(t)dt=0$.
- pour tout t, $\psi_{l,h}(t)=2^{l/2}\psi(2^lt-h)$ pour $(l,h)\in\mathbb{Z}^2$.
- ϕ une fonction d'échelle associée à ψ.
- pour tout t, $\phi_{l,h}(t)=2^{l/2}\phi(2^lt-h)$ pour $(l,h)\in\mathbb{Z}^2$.

On peut alors écrire, pour tout f, pour tout t,

$$f(t)=\sum_{h\in\mathbb{Z}}\beta_{L,h}(f)\phi_{L,h}(t)+\sum_{h\in\mathbb{Z}}\sum_{l\leq L}d_{l,h}(f)\psi_{l,h}(t)$$

où

$$\begin{cases}\beta_{l,h}(f) &=<f,\phi_{l,h}> \text{ pour tout } (l,h)\in\mathbb{Z}^2,\\ d_{l,h}(f) &=<f,\psi_{l,h}> \text{ pour tout } (l,h)\in\mathbb{Z}^2.\end{cases}$$

Alors, plutôt que de considérer l'échantillon $(f_i)_{1\leq i\leq n}$, on peut travailler avec l'échantillon $(x_i)_{1\leq i\leq n}=(\beta_{L,h}(f_i),(d_{l,h}(f_i))_{l\geq L,h\in\mathbb{Z}})_{1\leq i\leq n}$. Comme la base est orthonormale, supposer que les $(f_i)_{1\leq i\leq n}$ suivent une loi Gaussienne revient à supposer que les $(x_i)_{1\leq i\leq n}$ suivent une loi Gaussienne.
D'un point de vue pratique, la décomposition d'un signal sur une base d'ondelettes est très efficace. Citons Mallat, [Mal99], pour une description détaillée des ondelettes et de leur utilisation. L'intérêt majeur est la décomposition temps-échelle du signal, qui permet d'analyser les coefficients. De plus, si on choisit bien la base, on peut résumer un signal en peu de coefficients, ce qui permet de réduire la dimension du problème. D'un point de vue plus pratique, on peut citer Misiti et coauteurs, [MMOP07] pour la mise en pratique de la décomposition d'un signal sur une base d'ondelettes.
L'application principale de cette thèse est la classification des consommateurs électriques dans un but de prédiction de la consommation agrégée. Si on prévoit la consommation de chaque consommateur, et qu'on somme ces prédictions, on va sommer les erreurs de prédiction, donc

on peut faire beaucoup d'erreurs. Si on prévoit la consommation totale, on risque de faire des erreurs en n'étudiant pas assez les variations de chaque consommation individuelle. Cela explique le besoin de faire de la classification, et la classification en régression est faite dans un but de prédiction. Cependant, on sait que la prédiction de la consommation électrique peut être améliorée avec des modèles beaucoup plus adaptés. L'objectif de cette procédure est de classer ensemble, dans une étape préliminaire, les consommateurs qui ont le même comportement d'un jour à l'autre, ces groupes seront alors construits dans un but de prédiction.

Plan de la thèse

Cette thèse est principalement centrée sur les modèles de mélange en régression, et les problèmes de classification avec des données de régression en grande dimension. Elle est découpée en 5 chapitres, qui peuvent tous être lus de manière indépendante.
Le premier chapitre est consacré à l'étude du modèle principal. On décrit le modèle de mélange de Gaussiennes en régression, où la réponse et les régresseurs sont multivariés. On propose plusieurs approches pour estimer les paramètres de ce modèle, entre autres en grande dimension (pour la réponse et pour les régresseurs). On définit, dans ce cadre, plusieurs estimateurs pour les paramètres inconnus : une extension de l'estimateur du Lasso, l'estimateur du maximum de vraisemblance et l'estimateur du maximum de vraisemblance sous contrainte de faible rang. Dans chaque cas, on a cherché à optimiser les définitions des estimateurs dans un but algorithmique. On décrit, de manière précise, les deux procédures proposées dans cette thèse pour classifier des variables dans un cadre de régression en grande dimension, et estimer le modèle sous-jacent. C'est une partie méthodologique, qui décrit précisément le fonctionnement de nos procédures, et qui explique comment les mettre en œuvre numériquement. Des illustrations numériques sont proposées, pour confirmer en pratique l'utilisation de nos procédures. On utilise pour cela des données simulées, où l'aspect de grande dimension, l'aspect fonctionnel, et l'aspect de classification sont surlignés, et on utilise aussi des données de référence, où la vraie densité (inconnue donc) n'appartient plus à la collection de modèles considérée.
Dans un deuxième chapitre, on obtient un résultat théorique pour l'estimateur du Lasso dans les modèles de mélange de Gaussiennes en régression, en tant que régularisateur ℓ_1. Remarquons qu'ici, la pénalité est différente de celle du chapitre 1, cet estimateur étant une extension directe de l'estimateur du Lasso introduit par Tibshirani pour le modèle linéaire. Nous établissons une inégalité oracle ℓ_1 qui compare le risque de prédiction de l'estimateur Lasso à l'oracle ℓ_1. Le point important de cette inégalité oracle est que, contrairement aux résultats habituels sur l'estimateur du Lasso, nous n'avons pas besoin d'hypothèses sur la non colinéarité entre les variables. En contrepartie, la borne sur le paramètre de régularisation n'est pas optimale, dans le sens où des résultats d'optimalité ont été démontrés pour une borne inférieure à celle que l'on obtient, mais sous des hypothèses plus contraignantes. Notons que les constantes sont toutes explicites, même si l'optimalité de ces quantités n'est pas garantie.
Dans les chapitres 3 et 4, on propose une étude théorique de nos procédures de classification. On justifie théoriquement l'étape de sélection de modèles en établissant une inégalité oracle dans chaque cas (correspondant respectivement à l'inégalité oracle pour la procédure Lasso-EMV et l'inégalité oracle pour la procédure Lasso-Rang). Dans un premier temps, on a obtenu un théorème général de sélection de modèles, qui permet de choisir un modèle parmi une sous-collection aléatoire, dans un cadre de régression. Ce résultat, démontré à l'aide d'inégalités de concentration et de contrôles par calcul d'entropie métrique, est une généralisation à une sous-collection aléatoire de modèles d'un résultat déjà existant. Cette amélioration nous permet d'obtenir une inégalité oracle pour chacune de nos procédures : en effet, nous considérons une sous-collection aléatoire de modèles, décrite par le chemin de régularisation de l'estimateur du Lasso, et cet effet

aléatoire nécessite de prendre des précautions dans les inégalités de concentration. Ce résultat fournit une forme de pénalité minimale garantissant que l'estimateur du maximum de vraisemblance pénalisé est proche de l'oracle ℓ_0. En appliquant une telle pénalité lors de nos procédures, nous sommes sûrs d'obtenir un estimateur avec un faible risque de prédiction. L'hypothèse majeure que l'on fait pour obtenir ce résultat est de borner les paramètres du modèle de mélange. Remarquons que la pénalité n'est pas proportionnelle à la dimension, il y a un terme logarithmique en plus. On peut alors s'interroger quant à la nécessité de ce terme. On illustre aussi cette étape dans chacun des chapitres sur des jeux de données simulées et des jeux de données de référence. Il est important de souligner que ces résultats, théoriques et pratiques, sont envisageables car nous avons réestimé les paramètres par l'estimateur du maximum de vraisemblance, en se restreignant aux variables actives pour ne plus avoir de problème de grande dimension.
Dans le chapitre 5, on s'intéresse à un jeu de données réelles. On met en pratique la procédure Lasso-EMV de classification des données en régression en grande dimension pour comprendre comment classer les consommateurs électriques, dans le but d'améliorer la prédiction. Ce travail a été effectué en collaboration avec Yannig Goude et Jean-Michel Poggi. Le jeu de données utilisé est un jeu de données irlandaises, publiques. Il s'agit de consommations électriques individuelles, relevées sur une année. Nous avons aussi accès à des données explicatives, telles que la température, et des données personnelles pour chaque consommateur. Nous avons utilisé la procédure Lasso-EMV de trois manières différentes. Un problème simple, qui nous a permis de calibrer la méthode, est de considérer la consommation agrégée sur les individus, et de classifier les transitions de jour. Les données sont alors assez stables, et les résultats interprétables (on veut classer les transitions de jours de semaine ensemble par exemple). Le deuxième schéma envisagé est de classifier les consommateurs, sur leur consommation moyenne. Pour ne pas perdre l'aspect temporel, on a considéré les jours moyens. Finalement, pour compléter l'analyse, on a classifié les consommateurs sur leur comportement sur deux jours fixés. Le problème majeur de ce schéma est l'instabilité des données. Cependant, l'analyse des résultats, par des critères classiques en consommation électrique, ou grâce aux variables explicatives disponibles avec ce jeu de données, permet de justifier l'intérêt de notre méthode pour ce jeu de données.
A travers ce manuscrit, nous illustrons donc l'utilisation des modèles de mélange de Gaussiennes en régression, d'un point de vue méthodologique, mis en oeuvre d'un point de vue pratique, et justifié d'un point de vue théorique.

Perspectives

Pour poursuivre l'exploration des résultats de nos méthodes sur des données réelles, on pourrait utiliser un modèle de prédiction dans chaque classe. L'idée serait alors de comparer la prédiction obtenue plus classiquement avec la prédiction agrégée obtenue à l'aide de notre classification.

D'un point de vue méthodologique, on pourrait développer des variantes de nos procédures.
Par exemple, on pourrait envisager de relaxer l'hypothèse d'indépendance des variables induite par la matrice de covariance diagonale. Il faudrait la supposer parcimonieuse, pour réduire la dimension sous-jacente, et ainsi considérer des variables possiblement corrélées.
On pourrait aussi améliorer le critère de sélection de modèles, en l'orientant plus pour la classification. Par exemple, le critère ICL, introduit dans [BCG00] et développé dans la thèse de Baudry, [Bau09], tient compte de cet objectif de classification en considérant l'entropie.

D'un point de vue théorique, d'autres résultats pourraient être envisagés.
Les inégalités oracles obtenues donnent une pénalité minimale conduisant à de bons résultats, mais on pourrait vouloir démontrer que l'ordre de grandeur est le bon, à l'aide d'une borne

minimax.
On pourrait aussi s'intéresser à des intervalles de confiance. Le théorème de van de Geer et al., dans [vdGBRD14], valable pour des pertes convexes, peut être généralisé à notre cas, et on obtient ainsi assez facilement un intervalle de confiance pour la matrice de régression. Cependant, dans un but de prédiction, il pourrait être plus intéressant d'obtenir un intervalle de confiance pour la réponse, mais c'est un problème bien plus difficile.
Dans un but de classification, on pourrait aussi vouloir obtenir des résultats similaires aux inégalités oracles pour un autre critère que la divergence de Kullback-Leibler, plus orienté classification.

Notations

In this thesis, we denote (unless otherwise stated) by capital letter random variables, by lower case observations, and in bold letters the observation vector. For a matrix A, we denote by $[A]_{i,.}$ its ith row, $[A]_{.,j}$ its jth column, and $[A]_{i,j}$ its coefficient indexed by (i, j). For a vector B, we denote by $[B]_j$ its jth component.

Usual notations

${}^cA$	complement of A
A^t	transpose of A
$Tr(A)$	trace of a square matrix A
$\mathrm{E}(X)$	esperance of the random variable X
$\mathrm{Var}(X)$	variance of the random variable X
$\mathcal{N}$	Gaussian distribution
$\mathcal{N}_q$	Gaussian multivariate distribution of size q
χ^2	chi-squared distribution
$\mathcal{B}$	orthonormal basis
$\mathbb{1}_A$	indicator function on a set A
$\lfloor a \rfloor$	floor function of a
$a \vee b$	notation for the maximum between a and b
$a \wedge b$	notation for the minimum between a and b
$f \asymp g$	f is asymptotically equivalent to g
Δ	discriminant for a quadratic polynomial function
I_q	identity matrix of size q
$< a, b >$	scalar product between a and b
$\mathcal{P}(\{1, \ldots, p\})$	set of parts of $\{1, \ldots, p\}$
$\mathbb{S}_q^{++}$	set of positive-definite matrix of size q

Acronym

AIC	Akaike Information Criterion
ARI	Adjusted Rand Index
BIC	Bayesian Information Criterion
EM	Expectation-Maximization (algorithm)
EMV	Estimateur du Maximum de Vraisemblance
FR	False Relevant (variables)
LMLE	Lasso-Maximum Likelihood Estimator procedure
LR	Lasso-Rank procedure
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimator
REC	Restricted Eigenvalue Condition
SNR	Signal-to-Noise Ratio
TR	True Relevant (variables)

Variables and observations

X	regressors: random variable of size p
x_i	i^{th} observation of the variable X
$\mathbf{x}$	vector of the observations
Y	response: random variable of size q
y_i	i^{th} observation of the variable Y
$\mathbf{y}$	vector of the observations
Z	random variable for the affectation: vector of size K, $Z_k = 1$ if the variable Y, conditionally to X, belongs to the cluster k, 0 otherwise
z_i	observation of the variable Z for the observation y_i conditionally to $X_i = x_i$
F	functional regressor: random functional variable
f_i	i^{th} observation of the variable F
G	functional response: random functional response
g_i	i^{th} observation of the variable G
$\tilde{y}$	reparametrization of observation y, matrix of size $n \times K \times q$
$\tilde{x}$	reparametrization of observation x, matrix of size $n \times K \times p$
ϵ	Gaussian variable
ϵ_i	i^{th} observation of the variable ϵ
$\hat{y}_i$	prediction of the value of y_i from observation x_i

Parameters

$\boldsymbol{\beta}$	conditional mean, of size $q \times p \times K$
$\boldsymbol{\sigma}$	variance, in univariate models, of size K
$\boldsymbol{\Sigma}$	covariance matrix, in multivariate models, of size $q \times q \times K$
$\boldsymbol{\Phi}$	reparametrized conditional mean, of size $q \times p \times K$
P	reparametrized covariance matrix, of size $q \times q \times K$
$\boldsymbol{\pi}$	proportions coefficients, of size K
$\hat{\tau}$	A Posteriori Probability, matrix of size $n \times K$
ξ	vector of all parameters before reparametrization: $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$
θ	vector of all parameters after reparametrization: $(\boldsymbol{\pi}, \boldsymbol{\phi}, \mathbf{P})$
λ	regularization parameter for the Lasso estimator
$\lambda_{k,j,z}$	Lasso regularization parameter to cancel coefficient $[\phi_k]_{z,j}$ in mixture models
Ω	parameter for the assumption K
w_m	weights for the assumption K
$\tau_{i,k}(\theta)$	probability for the observation i to belong to the cluster k, according to the parameter θ
κ	parameter for the slope heuristic
$\mathbf{R}$	vector of ranks for conditional mean, of size K
R_k	rank value of the conditional mean ϕ_k in the component k
ξ^0	true parameter (Chapter 2)

Estimators

$\hat{\beta}$	estimator of the conditional mean in linear model
$\hat{\sigma}^2$	estimator of the variance in linear model
$\hat{\Sigma}$	estimator of the covariance matrix in multivariate linear model
$\hat{\theta}^{Lasso}(\lambda)$	estimator of θ by the Lasso estimator, with regularization parameter λ
$\hat{\beta}^{Lasso}(\lambda)$	estimator of β by the Lasso estimator, with regularization parameter λ
$\tilde{\beta}^{Lasso}(A)$	estimator of β by the Lasso estimator, with regularization parameter A according to the dual formulae
$\hat{\xi}_K^{Lasso}(\lambda)$	estimator of ξ_K by the Lasso estimator, with regularization parameter λ
$\hat{\xi}_K^{EMV}$	estimator of ξ_K by the maximum likelihood estimator
$\hat{\xi}_J^{EMV}$	estimator of ξ_K by the maximum likelihood estimator, restricted to J for the relevant variables
$\hat{\xi}_J^{Rank}$	estimator of ξ_K by the low rank estimator, restricted to J for the relevant variables
$\hat{\Sigma}_{\mathbf{x}}$	Gram matrix, according to the sample $\mathbf{x}$
$\hat{\theta}^{Group-Lasso}(\lambda)$	estimator of θ by the Group-Lasso estimator with regularization parameter λ
$\hat{\beta}^{LR}(\lambda)$	estimator of β by the low-rank estimator, restricted to variables detected by $\hat{\beta}^{Lasso}(\lambda)$
$\hat{P}^{LR}(\lambda)$	estimator of P by the low-rank estimator, restricted to variables detected by $\hat{\beta}^{Lasso}(\lambda)$

Sets of densities

$\mathcal{H}_{(K,J)}$	set of conditional densities, with parameters θ, with K clusters, and J for relevant variable set
$\check{\mathcal{H}}_{(K,J)}$	set of conditional densities, with parameters θ, with K clusters, and J for relevant variable set, and with vector of ranks R
$\mathcal{S}_{(K,J)}$	set of conditional densities, with parameters ξ, with K clusters, and J for relevant variable set
$\mathcal{S}_{(K,J,R)}$	set of conditional densities, with parameters ξ, with K clusters, and J for relevant variable set, and with vector of ranks R
$\mathcal{S}^{\mathcal{B}}_{(K,J)}$	subset of $\mathcal{S}_{(K,J)}$ with bounded parameters
$\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$	subset of $\mathcal{S}_{(K,J,R)}$ with bounded parameters

Dimensions

p	number of regressors
q	response size
K	number of components
n	sample size
D_m	dimension of the model $\mathcal{S}_m$
$D_{(K,J)}$	dimension of the model $\mathcal{S}_{(K,J)}$
$D_{(K,J,R)}$	dimension of the model $\mathcal{S}_{(K,J,R)}$

Sets of parameters

Θ_K	set of θ with K components
$\Theta_{(K,J)}$	set of θ with K components and J for relevant variables set
$\Theta_{(K,J,R)}$	set of θ with K components and J for relevant variables set, and R for vector of ranks for the conditional mean
Ξ_K	set of ξ with K components
$\Xi_{(K,J)}$	set of ξ with K components and J for relevant variables set
$\Xi_{(K,J,R)}$	set of ξ with K components and J for relevant variables set, and R for vector of ranks for the conditional mean
$\tilde{\Xi}_{(K,J,R)}$	subset of $\Xi_{(K,J,R)}$ with bounded parameters
$\mathcal{K}$	set of possible number of components
$\mathcal{J}$	set of set of relevant variables
$\mathcal{J}^L$	set of set of relevant variables, determined by the Lasso estimator
$\widetilde{\mathcal{J}}$	set of set of relevant variables, determined by the Group-Lasso estimator
$\mathcal{R}$	set of possible rank vectors
$\mathcal{S}$	set of densities
$\mathcal{M}$	model collection indices for the model collection constructed by our procedure
$\mathcal{M}^L$	random model collection indices for the model collection constructed by our procedure, according to the Lasso estimator
$\check{\mathcal{M}}$	random model collection indices
$\tilde{\mathcal{M}}$	model collection indices for the Group-Lasso-MLE model collection
$\tilde{\mathcal{M}}^L$	random model collection indices for the Group-Lasso-MLE model collection
Π_K	simplex of proportion coefficients
T_q	upper triangular matrices, of size q
J	set of relevant variables
J_λ	set of relevant variables dected by the Lasso estimator with regularization parameter λ
$\tilde{J}$	set of relevant variables detected by the Group-Lasso estimator
$\Psi_{(K,J,R)}$	set of conditional means, with J for relevant columns
$\tilde{\Psi}_{(K,J,R)}$	subset of $\Psi_{(K,J,R)}$ with bounded parameters and R for vector of ranks, in a mixture with K components
$\mathcal{F}_J$	set of conditional Gaussian density, with bounded conditional means and bounded covariance coefficients, and relevant variables set defined by J
$\mathcal{F}_{(J,R)}$	set of conditional Gaussian density, with relevant variables set defined by J and vector of ranks defined by R, and bounded covariance coefficients and bounded singular values
G_K	grid of regularization parameters, for model with K clusters

Functions

KL	Kullback-Leibler divergence
KL_n	Kullback-Leibler divergence for fixed covariates
$\mathrm{KL}^{\otimes_n}$	tensorized Kullback-Leibler divergence
d_H	Hellinger distance
$d_H^{\otimes_n}$	tensorized Hellinger distance
JKL_ρ	Jensen-Kullback-Leibler divergence, with parameter $\rho \in (0,1)$
$\mathrm{JKL}_\rho^{\otimes_n}$	tensorized Jensen-Kullback-Leibler divergence, with parameter $\rho \in (0,1)$
s_ξ	conditional density, with parameter ξ
s_ξ^K	conditional density, with parameter ξ and with K components
$s_\xi^{(K,J)}$	conditional density, with parameter ξ and with K components and J for relevant variables set
$s_\xi^{(K,J,R)}$	conditional density, with parameter ξ and with K components and J for relevant variables set and R for vector of ranks
s^*	true density
s_O	oracle conditional density
s_m	density for the model m
l	log-likelihood function
l_λ	penalized log-likelihood function for the Lasso estimator
$\tilde{l}_\lambda$	penalized log-likelihood function for the Group-Lasso estimator
l_c	complete log-likelihood function
γ	constrast function
l	loss function
γ_n	empirical contrast function
pen	penalty
l	lower function in a bracket
u	upper function in a bracket
φ	Gaussian density
ψ	wavelet function
ϕ	scaling function in wavelet decomposition
$\xi(x)$	parameters in mixture regressions, defining from regressors x
$m(A)$	smallest eigenvalue of the matrix A
$M(A)$	biggest eigenvalue of the matrix A
$\mathcal{H}_{[.]}(\epsilon, S, \|\|.\|\|)$	bracketing entropy of a set S, with brackets of width $\|\|.\|\|$ smaller than ϵ

Indices

j	varying from 1 to p
z	varying from 1 to q
k	varying from 1 to K
i	varying from 1 to n
m	varying in $\mathcal{M}$
m_O	index of the oracle
$\hat{m}$	selected index

Chapter 1

Two procedures

In this chapter, we describe two procedures to cluster data in a regression context. Following Maugis and Meynet [MMR12], we propose two global model selection procedures to simultaneously select the number of clusters and the set of relevant variables for the clustering. It is especially suited to deal with high-dimension and low sample size settings.
We take advantage of regression datasets to underline reliance between the regressors and the responses, and cluster data from this approach. This idea could be interesting for prediction, because observations sharing the same reliance will be considered in the same cluster.
In addition, we focus on functional dataset, for which the projection onto a wavelet basis leads to sparse representation. We also illustrate those procedures on simulated and benchmark dataset.

1.1 Introduction

Owing to the increasing of high-dimensional datasets, regression models for multivariate response and high-dimensional predictors have become important tools.
The goal of this chapter is to describe two procedures which cluster regression datasets. We focus on the model-based clustering. Each cluster is represented by a parametric conditional distribution, the entire dataset being modeled by a mixture of these distributions. It provides a rigorous statistical framework, and allows to understand the role of each variable in the clustering process. The model considered is then, for $i \in \{1, \dots, n\}$, if $(y_i, x_i) \in \mathbb{R}^q \times \mathbb{R}^p$ belongs to the component k, there exists an unknown $q \times p$ matrix of coefficients β_k and an unknown covariance matrix Σ_k such that

$$y_i = \beta_k x_i + \epsilon_i \tag{1.1}$$

where $\epsilon_i \sim \mathcal{N}_q(0, \Sigma_k)$. We will work with high-dimensional datasets, that is to say $q \times p$ could be larger than the sample size n, then we have to reduce the dimension. Two ways will be considered here, coefficients sparsity and ranks sparsity.
We could work with a sparse model if the matrix β could be estimated by a matrix with few nonzero coefficients. The well-known Lasso estimator, introduced by Tibshirani in 1996 in [Tib96] for linear models, is the solution chosen here. Indeed, the Lasso estimator is used for variable selection, cite for example Meinshausen and Bühlmann in [MB10] for stability selection results. We could also cite the book of Bühlmann and van de Geer [BvdG11] for an overview of the Lasso estimator.
If we look for rank sparsity for β, we have to assume that a lot of regressors are linearly dependent. This approach date's back to the 1950's, and was initiated by Anderson in [And51] for the linear model. Izenman, in [Ize75], introduced the term of reduced-rank regression for this class of models. A number of important works followed, cite for example Giraud in [Gir11] and Bunea et al. in [BSW12] for recent results. Nevertheless, the linear regression model used in those methods is appropriate for modeling the relationship between response and predictors when the reliance is the same for all observations, and it is inadequate for settings in which the regression coefficients differ across subgroups of the observations, then we will consider here mixture models.
An important example of high-dimensional datasets is functional datasets (functional predictors and/or functional response). They have been studied for example in the book of Ramsay and Silverman, [RS05]. A lot of recent works have been done on regression models for functional datasets: for example, cite the article of Ciarleglio ([CO14]) which deals with scalar response

and functional regressors. One way to consider functional datasets is to project it onto a basis well-suited. We can cite for example Fourier basis, splines, or, the one we consider here, wavelet basis. Indeed, wavelets are particularly well suited to handle many types of functional data, because they represent global and local attributes of functions, and can deal with discontinuities. Moreover, to deal with the sparsity previously mentioned, a large class of functions can be well represented with few non-zero coefficients, for any suitable wavelet.
We propose here two procedures which cluster high-dimensional data or data described by a functional variable, explained by high-dimensional predictors or by predictor variables arising from sampling continuous curves. Note that we estimate the number of components, parameters of each model, and the proportions. We assume we do not have any knowledge about the model, except that it could be well approximated by sparse mixture Gaussian regression model. The high-dimensional problem is solved by using variable selection to detect relevant variables. Since the structure of interest may often be contained into a subset of available variables and many attributes may be useless or even harmful to detect a reasonable clustering structure, it is important to select the relevant clustering variables. Moreover, removing irrelevant variables enables to get simpler modeling and can largely enhance comprehension.
Our two procedures are mainly based on three recent works. Firstly, we could cite the article of Städler et al. [SBG10], which studies finite mixture regression model. Even if we work on a multivariate version of it, the model considered in the article [SBG10] is adopted here. The second, Meynet and Maugis article [MMR12], deals with model-based clustering in density estimation. They propose a procedure, called Lasso-MLE procedure, which determines the number of clusters, the set of relevant variables for the clustering, and a clustering of the observations, with high-dimensional data. We extend this procedure with conditional densities. Finally, we could cite the article [Gir11] of Giraud. It suggests a low-rank estimator for the linear model. To take into account the matrix structure, we will consider this approach in our mixture models.
We consider finite mixture of Gaussian regression model. We propose two different procedures, considering more or less the matrix structure. Both of them have the same frame. Firstly, an ℓ_1-penalized likelihood approach is considered to determine potential sets of relevant variables. Introduced by Tibshirani in [Tib96], the Lasso is used to select variables. This allows one to efficiently construct a data-driven model subcollection with reasonable complexity, even for high-dimensional situations, with different sparsities, varying the regularization parameter in the ℓ_1-penalized likelihood function. The second step of the procedures consists to estimate parameters in a better way than by the Lasso estimator. Then, we select a model among the collection using the slope heuristic, which is developed by Birgé and Massart in [BM07]. Differences between the both procedures are the estimation of parameters in each model. The first one, later called Lasso-MLE procedure, uses the maximum likelihood estimator rather than the ℓ_1-penalized maximum likelihood estimator. It avoids estimation problems due to the ℓ_1-penalization shrinkage. The second one, called Lasso-Rank procedure, deals with low rank estimation. For each model in the collection, we construct a subcollection of models with conditional means estimated by various low ranks matrices. It leads to sparsity and for the coefficients, and for the rank, and consider the conditional mean with its matrix structure.
The chapter is organized as follows. Section 1.2 deals with Gaussian mixture regression models. It describes the model collection that we will consider. In Section 1.3, we describe both procedures that we propose to solve the problem of clustering high-dimensional regression data. Section 1.4 presents an illustrative example, to highlight each choice involved by both procedures. Section 1.5 states the functional data case, with a description of the projection proposed to convert these functions into coefficients data. We end this section by study of simulated and benchmark data. Finally, a conclusion section ends this chapter.

1.2 Gaussian mixture regression models

We have to construct a statistical framework on the observations. Because we estimate the conditional densities by multivariate Gaussian in each cluster, the model used is a finite Gaussian mixture regression model. Städler et al in [SBG10] describe this model, when X is multidimensional, and Y is scalar. We generalize it in the multivariate response case in this section. Moreover, we will describe a model collection of Gaussian mixture regression models, with several sparsities.

1.2.1 Gaussian mixture regression

We observe n independent couples $(x_i, y_i)_{1\leq i\leq n}$, realizations of random variables $(X_i, Y_i)_{1\leq i\leq n}$, with $Y_i \in \mathbb{R}^q$ and $X_i \in \mathbb{R}^p$ for all $i \in \{1, \dots, n\}$, coming from a probability distribution with unknown conditional density denoted by s^*. We want to perform model-based clustering, then we assume that data could be well approximated by a mixture conditional density $s(y|x) = \sum_{k=1}^K \pi_k s_k(y|x)$, with K unknown. To get a Gaussian mixture regression model, we suppose that, if Y conditionally to X belongs to the cluster k,

$$Y = \beta_k X + \epsilon;$$

where $\epsilon \sim \mathcal{N}_q(0, \Sigma_k)$. We then assume that s_k is a multivariate Gaussian conditional density. Thus, the random response variable $Y \in \mathbb{R}^q$ depends on a set of explanatory variables, written $X \in \mathbb{R}^p$, through a regression-type model. By considering multivariate response, we could work with more general datasets. Indeed, we could for example explain a functional response by a functional regressor, as done with the electricity dataset in Section 1.5.3. Some assumptions are in order, for a mixture of K Gaussian regression models.

— the variables Y_i conditionally to X_i are independent, for all $i \in \{1, \dots, n\}$;

— we let $Y_i|X_i = x_i \sim s_\xi^K(y|x_i)dy$, with

$$s_\xi^K(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right)$$

$$\xi = (\pi_1, \dots, \pi_K, \beta_1, \dots, \beta_K, \Sigma_1, \dots, \Sigma_K) \in \Xi_K = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right)$$

$$\Pi_K = \left\{(\pi_1, \dots, \pi_k); \pi_k > 0 \text{ for } k \in \{1, \dots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$$

$\mathbb{S}_q^{++}$ is the set of symmetric positive definite matrices on $\mathbb{R}^q$.

Then, we want to estimate the conditional density function s_ξ^K from the observations. For all $k \in \{1, \dots, K\}, \beta_k$ is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k. The π_ks are the mixture proportions. Actually, for all $k \in \{1, \dots, K\}$, for all $z \in \{1, \dots, q\}$, $[\beta_k x]_z = \sum_{j=1}^p [\beta_k]_{z,j} x_j$ is the zth component of the conditional mean of the mixture component k for the conditional density $s_\xi^K(.|x)$.

In order to have a scale-invariant maximum likelihood estimator, and to have a convex optimization problem, we reparametrize the model described above by generalizing the reparametrization described in [SBG10].

For all $k \in \{1, \dots, K\}$, we then define $\Phi_k = P_k \beta_k$, in which ${}^tP_k P_k = \Sigma_k^{-1}$ (it is the Cholesky decomposition of the positive definite matrix Σ_k^{-1}). Our hypotheses could now be rewritten:

— the variables Y_i conditionally to X_i are independent, for all $i \in \{1, \dots, n\}$;

— we let $Y_i|X_i = x_i \sim h_\theta^K(y|x_i)dy$, for $i \in \{1, \dots, n\}$, with

$$h_\theta^K(y|x) = \sum_{k=1}^K \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y - \Phi_k x)^t (P_k y - \Phi_k x)}{2}\right)$$

$$\theta = (\pi_1, \dots, \pi_K, \Phi_1, \dots, \Phi_K, P_1, \dots, P_K) \in \Theta_K = \left(\Pi_K \times (\mathbb{R}^{p\times q})^K \times (T_q)^K\right)$$

$$\Pi_K = \left\{(\pi_1, \dots, \pi_K); \pi_k > 0 \text{ for } k \in \{1, \dots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$$

T_q is the set of lower triangular matrix with non-negative diagonal entries.

The log-likelihood of this model is equal to, according to the sample $(x_i, y_i)_{1\le i\le n}$,

$$l(\theta, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y_i - \Phi_k x_i)^t (P_k y_i - \Phi_k x_i)}{2}\right)\right);$$

and the maximum log-likelihood estimator (later denoted by MLE) is

$$\hat\theta^{MLE} := \underset{\theta\in\Theta_K}{\operatorname{argmin}} \left\{-\frac{1}{n} l(\theta, \mathbf{x}, \mathbf{y})\right\}.$$

This estimator is scale-invariant, and the optimization is convex in each cluster.
Since we deal with the $p \times q >> n$ case, this estimator has to be regularized to obtain accurate estimates. As a result, we propose the ℓ_1-norm penalized MLE

$$\hat\theta^{\text{Lasso}}(\lambda) := \underset{\theta\in\Theta_K}{\operatorname{argmin}} \left\{-\frac{1}{n} l_\lambda(\theta, \mathbf{x}, \mathbf{y})\right\}; \tag{1.2}$$

where

$$-\frac{1}{n} l_\lambda(\theta, \mathbf{x}, \mathbf{y}) = -\frac{1}{n} l(\theta, \mathbf{x}, \mathbf{y}) + \lambda \sum_{k=1}^K \pi_k ||\Phi_k||_1;$$

where $||\Phi_k||_1 = \sum_{j=1}^p \sum_{z=1}^q |[\Phi_k]_{z,j}|$, and with $\lambda > 0$ to specify. This estimator is not the usual ℓ_1-estimator, called the Lasso estimator, introduced by Tibshirani in [Tib96]. It penalizes the ℓ_1-norm of the coefficients and small variances simultaneously, which has some close relations to the Bayesian Lasso (see Park and Casella [PC08]). Moreover, the reparametrization allows us to consider non-standardized data.
Notice that we restrict ourselves in this chapter to diagonal covariance matrices which are dependent of the clusters, that is to say for all $k \in \{1, \dots, K\}$, $\Sigma_k = \text{Diag}\left([\Sigma_k]_{1,1}, \dots, [\Sigma_k]_{q,q}\right)$. Then, with the renormalization described above, the restriction becomes, for all $k \in \{1, \dots, K\}$, $P_k = \text{Diag}\left([P_k]_{1,1}, \dots, [P_k]_{q,q}\right)$. In that case, we assume that variables are not correlated, which is a strong assumption, but it allows to reduce easily the dimension.

1.2.2 Clustering with Gaussian mixture regression

Suppose we know how many clusters there are, denoted by K, and assume that we get, from the observations, an estimator $\hat\theta$ such that $h_{\hat\theta}^K$ well approximate the unknown conditional density s^*. Then, we want to group the data into clusters between observations which seem similar. From a different point of view, we can look at this problem as a missing data problem. Indeed, the complete data are $((x_1, y_1, z_1), \dots, (x_n, y_n, z_n))$ in which the latent random variables are $Z = (Z_1, \dots, Z_n)$, $Z_i = ([Z_i]_1, \dots, [Z_i]_K)$ for $i \in \{1, \dots, n\}$ being defined by

$$[Z_i]_k = \begin{cases} 1 & \text{if } Y_i \text{ arises from the } k^{th} \text{ subpopulation ;} \\ 0 & \text{otherwise.} \end{cases}$$

Thanks to the estimation $\hat{\theta}$, we could use the Maximum A Posteriori principle (later denoted MAP principle) to cluster data. Specifically, for all $i \in \{1, \dots, n\}$, for all $k \in \{1, \dots, K\}$, consider

$$\tau_{i,k}(\theta) = \frac{\pi_k \det(P_k) \exp\left(-\frac{1}{2}(P_k y_i - \Phi_k x_i)^t (P_k y_i - \Phi_k x_i)\right)}{\sum_{r=1}^{K} \pi_k \det(P_k) \exp\left(-\frac{1}{2}(P_k y_i - \Phi_k x_i)^t (P_k y_i - \Phi_k x_i)\right)}$$

the posterior probability of y_i coming from the component number k, where $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \mathbf{P})$. Then, the data are partitioned by the following rule:

$$[Z_i]_k = \begin{cases} 1 & \text{if } \tau_{i,k}(\hat{\theta}) > \tau_{i,l}(\hat{\theta}) \text{ for all } l \neq k \text{ ;} \\ 0 & \text{otherwise.} \end{cases}$$

1.2.3 EM algorithm

From an algorithmic point of view, we will use a generalization of the EM algorithm to compute the MLE and the ℓ_1-norm penalized MLE. The EM algorithm was introduced by Dempster et al. in [DLR77] to approximate the maximum likelihood estimator of parameters of mixture model. It is an iterative process based on the minimization of the expectation of the empirical contrast for the complete data conditionally to the observations and the current estimation of the parameter $\theta^{(\text{ite})}$ at each iteration (ite) $\in \mathbb{N}^*$. Thanks to the Karush-Kuhn-Tucker conditions, we could extend the second step to compute the maximum likelihood estimators, penalized or not, under rank constraint or not, as it was done in the scalar case in [SBG10]. All those calculus are available in Appendix 1.7.1. We therefore obtain the next updating formulae for the Lasso estimator defined by (1.2). Remark that it includes maximum likelihood estimator, and the rank constraint could be easily computed according to a singular value decomposition.

$$\pi_k^{(\text{ite}+1)} = \pi_k^{(\text{ite})} + t^{(\text{ite})} \left(\frac{n_k}{n} - \pi_k^{(\text{ite})} \right); \tag{1.3}$$

$$[P_k]_{z,z}^{(\text{ite}+1)} = \frac{n_k \langle [\tilde{\mathbf{y}}]_{k,z}^{(\text{ite})}, [\Phi_k]_{z,.}^{(\text{ite})} [\tilde{\mathbf{x}}]_{k,.}^{(\text{ite})} \rangle + \sqrt{\Delta}}{2 n_k ||[\tilde{\mathbf{y}}]_{k,z}^{(\text{ite})}||_2^2}; \tag{1.4}$$

$$[\Phi_k]_{z,j}^{(\text{ite}+1)} = \begin{cases} \frac{-[S_k]_{j,z}^{(\text{ite})} + n\lambda\pi_k^{(\text{ite})}}{||[\tilde{\mathbf{x}}]_{k,j}^{(\text{ite})}||_2^2} & \text{if } [S_k]_{j,z}^{(\text{ite})} > n\lambda\pi_k^{(\text{ite})}; \\ -\frac{[S_k]_{j,z}^{(\text{ite})} + n\lambda\pi_k^{(\text{ite})}}{||[\tilde{\mathbf{x}}]_{k,j}^{(\text{ite})}||_2^2} & \text{if } [S_k]_{j,z}^{(\text{ite})} < -n\lambda\pi_k^{(\text{ite})}; \\ 0 & \text{else ;} \end{cases} \tag{1.5}$$

with, for $j \in \{1, \ldots, p\}, k \in \{1, \ldots, K\}, z \in \{1, \ldots, q\}$,

$$[S_k]_{j,z}^{(\text{ite})} = -\sum_{i=1}^{n} [\tilde{x}_i]_{k,j}^{(\text{ite})} [P_k]_{z,z}^{(\text{ite})} [\tilde{y}_i]_{k,z}^{(\text{ite})} + \sum_{\substack{j_2=1 \\ j_2 \neq j}}^{p} [\tilde{x}_i]_{k,j}^{(\text{ite})} [\tilde{x}_i]_{k,j_2}^{(\text{ite})} [\Phi_k]_{z,j_2}^{(\text{ite})}; \tag{1.6}$$

$$n_k = \sum_{i=1}^{n} \tau_{i,k}^{(\text{ite})};$$

$$([\tilde{y}_i]_{k,z}^{(\text{ite})}, [\tilde{x}_i]_{k,j}^{(\text{ite})}) = \sqrt{\tau_{i,k}^{(\text{ite})}}([y_i]_z, [x_i]_j);$$

$$\Delta = \left(-n_k \langle [\tilde{\mathbf{y}}]_{k,z}^{(\text{ite})}, [\Phi_k]_{z,\cdot}^{(\text{ite})} [\tilde{\mathbf{x}}]_{k,\cdot}^{(\text{ite})} \rangle\right)^2 - 4||[\tilde{\mathbf{y}}]_{k,z}^{(\text{ite})}||_2^2;$$

$$\tau_{i,k}^{(\text{ite})} = \frac{\pi_k^{(\text{ite})} \left(\det P_k^{(\text{ite})}\right) \exp\left(-1/2 \left(P_k^{(\text{ite})} y_i - \Phi_k^{(\text{ite})} x_i\right)^t \left(P_k^{(\text{ite})} y_i - \Phi_k^{(\text{ite})} x_i\right)\right)}{\sum_{r=1}^{K} \pi_k^{(\text{ite})} \left(\det P_k^{(\text{ite})}\right) \exp\left(-1/2 \left(P_k^{(\text{ite})} y_i - \Phi_k^{(\text{ite})} x_i\right)^t \left(P_k^{(\text{ite})} y_i - \Phi_k^{(\text{ite})} x_i\right)\right)}; \tag{1.7}$$

and $t^{(ite)} \in (0, 1]$, the largest value in the grid $\{\delta^l, l \in \mathbb{N}\}$, $0 < \delta < 1$, such that the function is not increasing.
In our case, the EM algorithm corresponds to switch between the E-step which corresponds to the calculus of (1.3), (1.4) and (1.5), and the M-step, which corresponds to the calculus of (1.7). To avoid convergence to local maximum rather than global maximum, we need to precise the initialization and the stopping rules. We initialize the clustering with the k-means algorithm on the couples $(x_i, y_i)_{1 \leq i \leq n}$. According to this clustering, we compute the linear regression estimators in each class. Then, we run a small number of times the EM-algorithm, repeat this initialization many times, and keep the one which maximizes the log-likelihood function: how the computation will start is important. Finally, to stop the algorithm, we could wait for any convergence, but the EM algorithm is known to check the convergence hypothesis, without converging, because of local maximum. Consequently, we choose to fix a minimum number of iterations to ensure non-local maximum, and to specify a maximum number of iterations to ensure stopping. Between these two bounds, we stop if there is convergence of the log-likelihood and of the parameters (with a relative criteria), adapted from [SBG10].

1.2.4 The model collection

We want to deal with high-dimensional data, that is why we have to determine which variables are relevant for the Gaussian regression mixture clustering. Indeed, we observe a small sample and we have to estimate many coefficients: we have a problem of identifiability. The sample size n is smaller than $K(pq + q + 1) - 1$, the size of parameters to estimate. A way to solve this problem is to select few variables to describe the problem. We then assume that we could estimate s^* by a sparse model.
To reduce the dimension, we want to determine which variables are useful for the clustering, and which are not. It leads to the definition of an irrelevant variable.

Definition 1.2.1. *A variable indexed by $(z, j) \in \{1, \ldots, q\} \times \{1, \ldots, p\}$ is* irrelevant *for the clustering if $[\Phi_1]_{z,j} = \ldots = [\Phi_K]_{z,j} = 0$. A* relevant *variable is a variable which is not irrelevant: at least in one cluster, this variable is not equal to zero. We denote by J the relevant variables set.*

We denote by $\Phi_k^{[J]}$ the matrix with 0 on the set ${}^cJ$, for all $k \in \{1, \ldots, K\}$, and $\mathcal{H}_{(K,J)}$ the model with K components and with J for relevant variables set:

$$\mathcal{H}_{(K,J)} = \left\{ y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto h_\theta^{(K,J)}(y|x) \right\}; \tag{1.8}$$

where

$$h_\theta^{(K,J)}(y|x) = \sum_{k=1}^{K} \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y - \Phi_k^{[J]} x)^t (P_k y - \Phi_k^{[J]} x)}{2} \right),$$

and

$$\theta = (\pi_1, \ldots, \pi_K, \Phi_1^{[J]}, \ldots, \Phi_K^{[J]}, P_1, \ldots, P_K) \in \Theta_{(K,J)} = \Pi_K \times \left(\mathbb{R}^{q\times p}\right)^K \times \left(\mathbb{R}_+^q\right)^K;$$

where the notation $A^{[J]}$ means that J is the relevant set variable for the matrix A.
We will construct a model collection, by varying the number of components K and the relevant variables subset J.

1.3 Two procedures

The goal of our procedures is, given a sample $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^p \times \mathbb{R}^q)^n$, to discover the structure of the variable Y according to X. Thus, we have to estimate, according to the representation of $\mathcal{H}_{(K,J)}$, the number of clusters K, the relevant variables set J, and the parameters θ. To overcome this difficulty, we want to take advantage of the sparsity property of the ℓ_1-penalization to perform automatic variable selection in clustering high-dimensional data. Then, we could compute another estimator restricted on relevant variables, which will work better because it is no longer an high-dimensional issue. Thus, we avoid shrinkage problems due to the Lasso estimator. The first procedure takes advantage of the maximum likelihood estimator, whereas the second one takes into account the matrix structure of $\mathbf{\Phi}$ with a low rank estimation.

1.3.1 Lasso-MLE procedure

This procedure is decomposed into three main steps: we construct a model collection, then in each we compute the maximum likelihood estimator, and we choose the best one among the model collection.
The first step consists of constructing a model collection $\{\mathcal{H}_{(K,J)}\}_{(K,J)\in\mathcal{M}}$ in which $\mathcal{H}_{(K,J)}$ is defined by equation (1.8), and the model collection is indexed by $\mathcal{M} = \mathcal{K} \times \mathcal{J}$. We denote by $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components. We assume we could bound $\mathcal{K}$ without loss of estimation. We also note $\mathcal{J} \subset \mathcal{P}(\{1, \ldots, q\} \times \{1, \ldots, p\})$.
To detect the relevant variables, and construct the set $J \in \mathcal{J}$, we penalize the empirical contrast by an ℓ_1-penalty on the mean parameters proportional to $||\Phi_k||_1 = \sum_{j=1}^p \sum_{z=1}^q |[\Phi_k]_{z,j}|$. In the ℓ_1-procedures, the choice of the regularization parameters is often difficult: fixing the number of components $K \in \mathcal{K}$, we propose to construct a data-driven grid G_K of regularization parameters by using the updating formulae of the mixture parameters in the EM algorithm. We can give a formula for λ, the regularization parameter, depending on which coefficient we want to cancel, for all $k \in \{1, \ldots, K\}, j \in \{1, \ldots, p\}, z \in \{1, \ldots, q\}$:

$$[\Phi_k]_{z,j} = 0 \quad \Leftrightarrow \quad \lambda_{k,j,z} = \frac{|[S_k]_{j,z}|}{n\pi_k};$$

with $[S_k]_{j,z}$ defined by (1.6). Then, we define the data-driven grid by

$$G_K = \{\lambda_{k,j,z}, k \in \{1, \ldots, K\}, j \in \{1, \ldots, p\}, z \in \{1, \ldots, q\}\}.$$

We could compute it from maximum likelihood estimations.
Then, for each $\lambda \in G_K$, we could compute the Lasso estimator defined by

$$\hat{\theta}^{\text{Lasso}}(\lambda) = \underset{\theta \in \Theta_{(K,J)}}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(h_\theta^{(K,J)}(y_i|x_i)) + \lambda \sum_{k=1}^{K} \pi_k ||\Phi_k||_1 \right\}.$$

For a fixed number of mixture components $K \in \mathcal{K}$ and a regularization parameter $\lambda \in G_K$, we could use an EM algorithm, recalled in Appendix 1.7.1, to approximate this estimator. Then, for each $K \in \mathcal{K}$, and for each $\lambda \in G_K$, we could construct the relevant variables set J_λ. We denote by $\mathcal{J}$ the collection of all these sets.
The second step consists of approximating the MLE

$$\hat{h}^{(K,J)} = \underset{h \in \mathcal{H}_{(K,J)}}{\text{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(h(y_i|x_i)) \right\};$$

using the EM algorithm for each model $(K, J) \in \mathcal{M}$.
The third step is devoted to model selection. Rather than select the regularization parameter, we select the refitted model. We use the slope heuristic described in [BM07]. Explain briefly how it works. Firstly, models are grouping according to their dimension D, to obtain a model collection $\{\mathcal{H}_D\}_{D \in \mathcal{D}}$. The dimension of a model is the number of parameters estimated in the model. For each dimension D, let $\hat{h}_D$ be the estimator maximizing the likelihood among the estimators associated to a model of dimension D. Also, the function $D/n \mapsto 1/n \sum_{i=1}^{n} \log(\hat{h}_D(y_i|x_i))$ has a linear behavior for large dimensions. We estimate the slope, denoted by $\hat{\kappa}$, which will be used to calibrate the penalty. The minimizer $\hat{D}$ of the penalized criterion $-1/n \sum_{i=1}^{n} \log(\hat{h}_D(y_i|x_i)) + 2\hat{\kappa} D/n$ is determined, and the model selected is $(K_{\hat{D}}, J_{\hat{D}})$. Remark that $D = K(|J| + q + 1) - 1$. Note that the model is selected after parameters refitting, which avoids issue of regularization parameter selection. For an oracle inequality to justify the slope heuristic used here, see [Dev14b].

1.3.2 Lasso-Rank procedure

Whereas the previous procedure does not take into account the multivariate structure, we propose a second procedure to perform this point. For each model belonging to the collection $\mathcal{H}_{(K,J)}$, a subcollection is constructed, varying the rank of Φ. Let us describe this procedure.
As in the Lasso-MLE procedure, we first construct a collection of models, thanks to the ℓ_1-approach. For $\lambda \geq 0$, we obtain an estimator for θ, denoted by $\hat{\theta}^{\text{Lasso}}(\lambda)$, for each model belonging to the collection. We could deduce the set of relevant columns, denoted by J_λ, and this for all $K \in \mathcal{K}$: we deduce $\mathcal{J}$ the collection of relevant variables set.
The second step consists to construct a subcollection of models with rank sparsity, denoted by

$$\{\check{\mathcal{H}}_{(K,J,R)}\}_{(K,J,R) \in \check{\mathcal{M}}}.$$

The model $\check{\mathcal{H}}_{(K,J,R)}$ has K components, the set J for active variables, and R is the vector of the ranks of the matrix of regression coefficients in each group:

$$\check{\mathcal{H}}_{(K,J,R)} = \left\{ y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto h_\theta^{(K,J,R)}(y|x) \right\} \tag{1.9}$$

where

$$h_\theta^{(K,J,R)}(y|x) = \sum_{k=1}^{K} \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y - (\Phi_k^{R_k})^{[J]} x)^t (P_k y - (\Phi_k^{R_k})^{[J]} x)}{2} \right);$$

$$\theta = (\pi_1, \ldots, \pi_K, (\Phi_1^{R_1})^{[J]}, \ldots, (\Phi_K^{R_K}))^{[J]}, P_1, \ldots, P_K) \in \Theta_{(K,J,R)} = \Pi_K \times \Psi_{(K,J,R)} \times \left(\mathbb{R}_+^q\right)^K;$$

$$\Psi_{(K,J,R)} = \left\{ ((\Phi_1^{R_1})^{[J]}, \ldots, (\Phi_K^{R_K})^{[J]}) \in \left(\mathbb{R}^{q \times p}\right)^K \middle| \text{Rank}(\Phi_k) = R_k \text{ for all } k \in \{1, \ldots, K\} \right\};$$

and $\mathcal{M}^R = \mathcal{K} \times \mathcal{J} \times \mathcal{R}$. We denote by $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components, $\mathcal{J}$ a collection of subsets of $\{1, \ldots, p\}$, and $\mathcal{R}$ the set of vectors of size $K \in \mathcal{K}$ with ranks values for each mean matrix. We could compute the MLE under the rank constraint thanks to an EM algorithm. Indeed, we could constrain the estimation of Φ_k, for the cluster k, to have a rank equal to R_k, in keeping only the R_k largest singular values. More details are given in Section 1.7.1. It leads to an estimator of the mean with row sparsity and low rank for each model. As described in the above section, a model is selected using the slope heuristic. This step is justified theoretically in [Dev15].

1.4 Illustrative example

We illustrate our procedures on four different simulated datasets, adapted from [SBG10], belonging to the model collection. They have been both implemented in Matlab, with the help of Benjamin Auder, and the Matlab code is available. Firstly, we will present models used in these simulations. Then, we validate numerically each step, and we finally compare results of our procedures with others. Remark that we propose here some examples to illustrate our methods, but not a complete analysis. We highlight some issues which seems important. Moreover, we do not illustrate the one-component case, focusing on the clustering. If, on some dataset, we are not convinced by the clustering, we could add to the model collection models with one component, more or less sparse, using the same pattern (computing the Lasso estimator to get the relevant variables for various regularization parameters, and refit parameters with the maximum likelihood estimator, under rank constraint or not), and select a model among this collection of linear and mixture models.

1.4.1 The model

Let $\mathbf{x}$ be a sample of size n distributed according to multivariate standard Gaussian. We consider a mixture of two components. Besides, we fix the number of active variables to 4 in each cluster. More precisely, the first four variables of Y are explained respectively by the four first variables of X. Fix $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$ and $P_k = I_q$ for all $k \in \{1, 2\}$.
The difficulty of the clustering is partially controlled by the signal-to-noise ratio. In this context, we could extend the natural idea of the SNR with the following definition, where $\mathrm{Tr}(A)$ denotes the trace of the matrix A.

$$\mathrm{SNR} = \frac{\mathrm{Tr}(\mathrm{Var}(Y))}{\mathrm{Tr}(\mathrm{Var}(Y|\beta_k = 0 \text{ for all } k \in \{1, \ldots, K\}))}.$$

Remark that it only controls the distance between the signal with or without the noise, and not the distance between the both signals.
We compute four different models, varying n, the SNR, and the distance between the clusters. Details are available in the Table 1.1.

	Model 1	Model 2	Model 3	Model 4	Model 5
n	2000	100	100	100	50
k	2	2	2	2	2
p	10	10	10	10	30
q	10	10	10	10	5
$\beta_{1\|J}$	3	3	3	5	3
$\beta_{2\|J}$	-2	-2	-2	3	-2
σ	1	1	3	1	1
SNR	3.6	3.6	1.88	7.8	3.6

Table 1.1: Description of the different models

Take a sample of Y according to a Gaussian mixture, meaning in $\beta_k X$ and with covariance matrix $\Sigma_k = (P_k^t P_k)^{-1} = \sigma I_q$, for the cluster k. We run our procedures with the number of components varying in $\mathcal{K} = \{2, \ldots, 5\}$.
The model 1 illustrates our procedures in low dimension models. Moreover, it is chosen in the next section to illustrate well each step of the procedure (variable selection, models construction and model selection). Model 5 is considered high-dimensional, because $p \times K > n$. The model 2 is easier than the others, because clusters are not so close to each other according to the noise. Model 3 is constructed as the models 1 and 2, but n is small and the noise is more important. We will see that it gives difficulty for the clustering. Model 4 has a larger SNR, nevertheless, the problem of clustering is difficult, because each β_k is closer to the others.
Our procedures are run 20 times, and we compute statistics on our results over those 20 simulations: it is a small number, but the whole procedure is time-consuming, and results are convincing enough.
For the initialization, we repeat 50 times the initialization, and keep the one which maximizes the log-likelihood function after 10 iterations. Those choices are size-dependent, a numerical study not reported here concludes that it is enough in that case.

1.4.2 Sparsity and model selection

To illustrate the both procedures, all the analyses made in this section are done from the model 1, since the choice of each step is clear.
Firstly, we compute the grid of regularization parameters. More precisely, each regularization parameter is computed from maximum likelihood estimations (using EM algorithm), and give an associated sparsity (computed by the Lasso estimator, using again the EM algorithm). In Figure (1.1) and Figure (1.2), the collection of relevant variables selected by this grid are plotted. Firstly, we could notice that the number of relevant variables selected by the Lasso decreases with the regularization parameter. We could analyze more precisely which variables are selected, that is to say if we select true relevant or false relevant variables. If the regularization parameter is not too large, the true relevant variables are selected. Even more, if the regularization parameter is well-chosen, we select only the true relevant variables. In our example, we remark that if $\lambda = 0.09$, we have selected exactly the true relevant variables. This grid construction seems to be well-chosen according to these simulations.
From this variable selection, each procedure (Lasso-MLE or Lasso-Rank) leads to a model collection, varying the sparsity thanks to the regularization parameters grid, and the number of components.
Among this collection, we select a model with the slope heuristic.
We want to select the best model by improving a penalized criterion. This penalty is computed

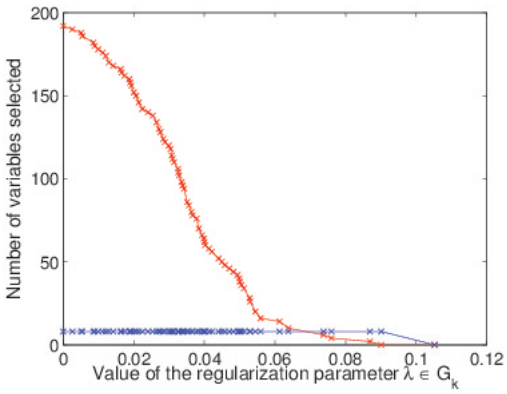

Figure 1.1: For one simulation, number of false relevant (in red color) and true relevant (in blue color) variables generated by the Lasso, by varying the regularization parameter λ in the grid G_2

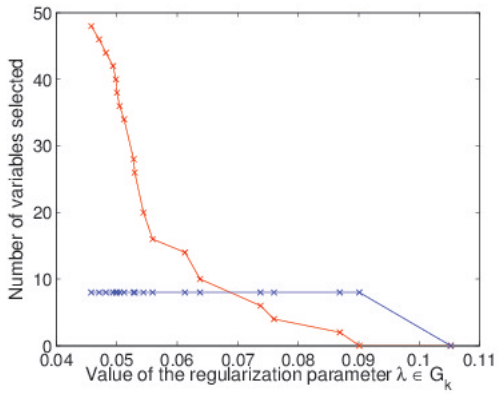

Figure 1.2: For one simulation, zoom in on number of false relevant (in red color) and true relevant (in blue color) variables generated by the Lasso, by varying the regularization parameter λ around the interesting values

by performing a linear regression on the couples of points $\{(D/n; -1/n \sum_{i=1}^{n} \log(\hat{h}_D(y_i|x_i)))\}$, D varying. The slope $\hat{\kappa}$ allows to have access to the best model, the one with dimension $\hat{D}$ minimizing $-1/n \sum_{i=1}^{n} \log(\hat{h}_D(y_i|x_i)) + 2\hat{\kappa}D/n$. In practice, we have to look if couples of points have a linear comportment . For each procedure, we construct the different model collection, and we have to justify this behavior. The Figures (1.3) and (1.4) represent the log-likelihood in function of the dimension of the models, for model collections constructed respectively by the Lasso-MLE procedure and by the Lasso-Rank procedure. The couples are plotted by points, whereas the estimated slope is specified by a dotted line. We could observe more than a line (4 for the Lasso-MLE procedure, more for the Lasso-Rank procedure). This phenomenon could be explained by a linear behavior for each mixture, fixing the number of classes, and for each rank. Nevertheless, slopes are almost the same, and select the same model. In practice, we estimate the slope with the Capushe package [BMM12].

1.4.3 Assessment

We compare our procedures to three other procedures on simulated models 1, 2, 3 and 4.
Firstly, let us give some remarks about the model 1. For each procedure, we get a good clustering and a very low Kullback-Leibler divergence. Indeed, the sample size is large, and the estimations are good. That is the reason why we focus in this section on models 2, 3 and 4.
To compare our procedures with others, the Kullback-Leibler divergence with the true density and the ARI (the Adjusted Rand Index, measuring the similarity between two data clusterings, knowing that the closer to 1 the ARI, the more similar the two partitions) are computed, and we note which variables are selected, and how many clusters are selected. For more details on the ARI, see [Ran71].
From the Lasso-MLE model collection, we construct two models, to compare our procedures with. We compute the oracle (the model which minimizes the Kullback-Leibler divergence with the true density), and the model which is selected by the BIC criterion instead of the slope heuristic. Thanks to the oracle, we know how good we could get from this model collection for the Kullback-Leibler divergence, and how this model, as good it is possible for the log-likelihood, performs the clustering.

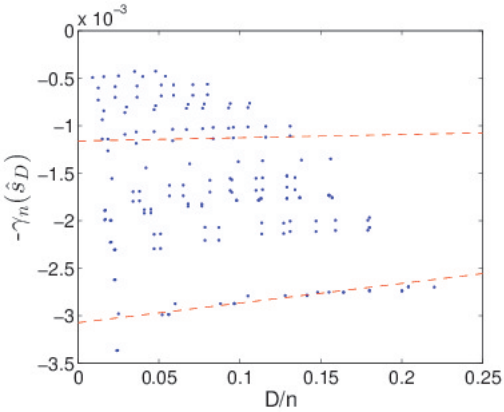

Figure 1.3: For one simulation, slope graph obtain by our Lasso-Rank procedure. For large dimensions, we observe a linear part

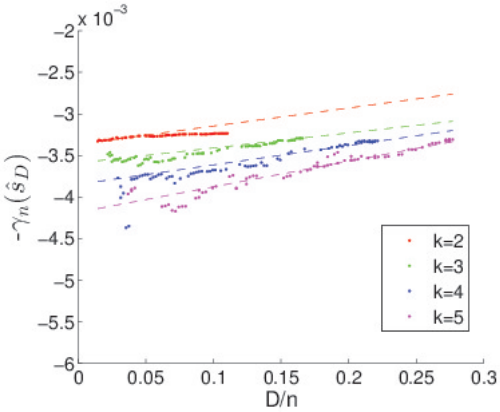

Figure 1.4: For one simulation, slope graph obtain by our Lasso-MLE procedure. For large dimensions, we observe a linear part

The third procedure we compare with is the maximum likelihood estimator, assuming that we know how many clusters there are, fixed to 2. We use this procedure to show that variable selection is needed.
In each case, we apply the MAP principle, to compare clusterings.
We do not plot the Kullback-Leibler divergence for the MLE procedure, because values are too high, and make the boxplots unreadable.
For the model 2, according to the Figure (1.5) for the Kullback-Leibler divergence, and Figure (1.6) for the ARI, the Kullback-Leibler divergence is small and the ARI is close to 1, except for the MLE procedure. Boxplots are still readable with those values, but it is important to highlight that variable selection is needed, even in a model with reasonable dimension. The model collections are then well constructed. The model 3 is more difficult, because the noise is higher. That is why results, summarized in Figures (1.7) and (1.8), are not as good as for the model 2. Nevertheless, our procedures lead to the best ARI, and the Kullback-Leibler divergences are close to the one of the oracle. We could make the same remarks for the model 4. In this study, means are closer, according to the noise. Results are summarized in Figures (1.9) and (1.10). The model 5 is in high-dimension. Models selected by the BIC criterion are bad, in comparison with models selected by our procedures, or oracles. They are bad for estimation, according to the Kullback-Leibler divergence boxplots in Figure (1.11), but also for clustering, according to Figure (1.12). Our models are not as well as constructed as previously, but it is explained by the high-dimensional context. It is explained by the high Kullback-Leibler divergence. Nevertheless, our performances in clustering are really good.
Note that the Kullback-Leibler divergence is smaller for the Lasso-MLE procedure, thanks to the maximum likelihood refitting. Moreover, the true model has not any matrix structure. If we look after the MLE, where we do not use the sparsity hypothesis, we could conclude that estimations are not satisfactory, which could be explained by an high-dimensional issue. The Lasso-MLE procedure, the Lasso-Rank procedure and the BIC model work almost as well as the oracle, which mind that the models are well selected.

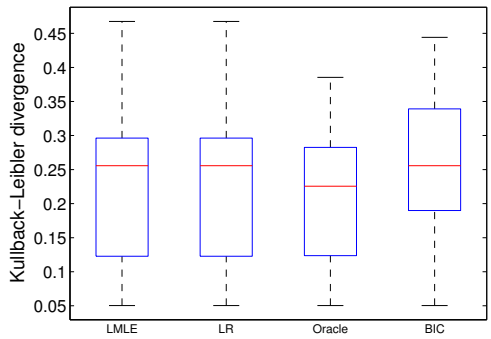

Figure 1.5: Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) for the model 2

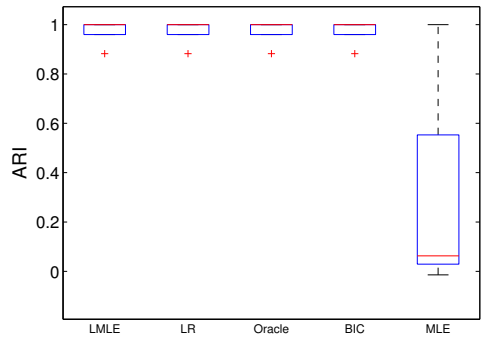

Figure 1.6: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE) for the model 2

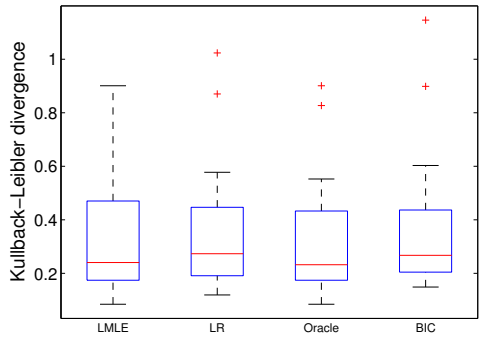

Figure 1.7: Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) for the model 3

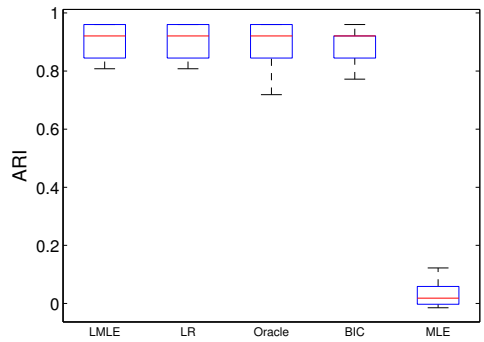

Figure 1.8: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE) for the model 3

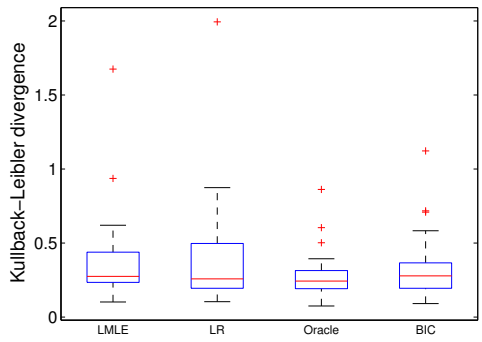

Figure 1.9: Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) for the model 4

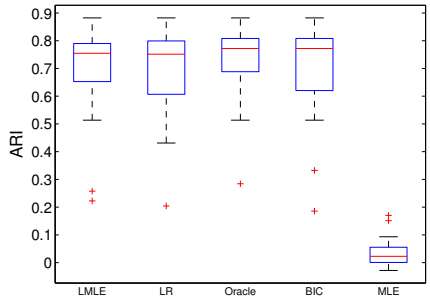

Figure 1.10: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE) for the model 4

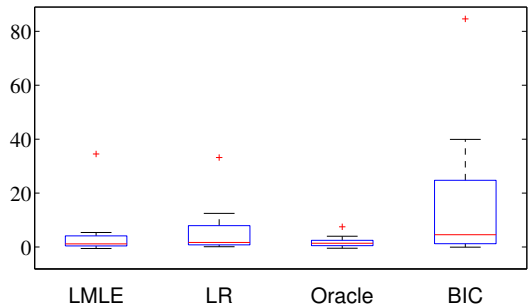

Figure 1.11: Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) for the model 5

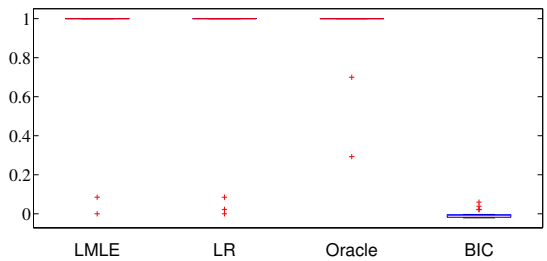

Figure 1.12: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE) for the model 5

	Model 2		Model 3		Model 4	
Procedure	TR	FR	TR	FR	TR	FR
Lasso-MLE	8(0)	2.2(6.9)	8(0)	4.3(28.8)	8(0)	2(13, 5)
Lasso-Rank	8(0)	24(0)	8(0)	24(0)	8(0)	24(0)
Oracle	8(0)	1.5(3.3)	7.8(0.2)	2.2(11.7)	8(0)	0.8(2.6)
BIC estimator	8(0)	2.6(15.8)	7.8(0.2)	5.7(64.8)	8(0)	2.6(11.8)

Table 1.2: Mean number {TR, FR} of true relevant and false relevant variables over the 20 simulations for each procedure, for models 2, 3 and 4. The standard deviations are put into parenthesis

In Table (1.2), we summarize results about variable selection. For each model, for each procedure, we compute how many true relevant and false relevant variables are selected.
The true model has 8 relevant variables, which are always recognized. The Lasso-MLE has less false relevant variables than the others, which means that the true structure was found. The Lasso-Rank has 24 false relevant variables, because of the matrix structure: the true rank in each component was 4, then the estimator restricted on relevant variables is a 4×4 matrix, and we get 12 false relevant variables in each component. Nevertheless, we do not have more variables, that is to say the model constructed is the best possible. The BIC estimator and the oracle have a large variability for the false relevant variables.
For the number of components, we find that all the procedures have selected the true number 2. Thanks to the MLE, the first procedure has good estimations (better than the second one). Nevertheless, depending on the data, the second procedure could be more attractive. If there is a matrix structure, for example if most of the variation of the response Y is caught by a small number of linear combinations of the predictors, the second procedure will work better.
We could conclude that the model collection is well constructed, and that the clustering is well-done.

1.5 Functional datasets

One of the main interest of our methods is to be applied to functional datasets. Indeed, in different fields of applications, considered data are functions. The functional data analysis has been popularized first by Ramsay and Silverman in their book [RS05]. It gives a description of the main tools to analyze functional datasets. Another book is the Ferraty and Vieu one's [FV06]. However, the main part of the existing literature about functional regression is concentrated on the case Y scalar and X functional. For example, we can cite Zhao et al., in [ZOR12] for using wavelet basis in linear model, Yao et al. ([YFL11]) for functional mixture regression, or Ciarleglio et al. ([CO14]) for using wavelet basis in functional mixture regression. In this section, we concentrate on Y and X both functional. In this regression context, we could be interested in clustering: it leads to identify the individuals involved in the same reliance between Y and X. Denote that, with functional datasets, we have to denoise and smooth signals to remove the noise and capture only the important patterns in the data. Here, we explain how our procedures can be applied in this context. Note that we could apply our procedures with scalar response and functional regressor, or, on the contrary, with functional response for scalar regressor. We explain how our procedures are generalized in the more difficult case, the other cases resulting of that. Remark that we focus here on the wavelet basis, to take advantage of the time-scale decomposition, but the same analysis is available on Fourier basis or splines.

1.5.1 Functional regression model

Suppose we observe a centered sample of functions $(f_i, g_i)_{1\leq i\leq n}$, associated with the random variables (F, G), coming from a probability distribution with unknown conditional density s^*. We want to estimate this model by a functional mixture model: if the variables (F, G) come from the component k, there exists a function β_k such that

$$G(t) = \int_{I_x} F(u)\beta_k(u,t)du + \epsilon(t), \tag{1.10}$$

where ϵ is a residual function. This linear model is introduced in Ramsay and Silverman's book [RS05]. They propose to project onto basis and the response, and the regressors. We extend their model in mixture model, to consider several subgroups for a sample.
If we assume that, for all t, for all $i \in \{1, \ldots, n\}$, $\epsilon_i(t) \sim \mathcal{N}(0, \Sigma_k)$, the model (1.10) is an integrated version of the model (1.1). Depending on the cluster k, the linear reliance of G with respect to F is described by the function β_k.

1.5.2 Two procedures to deal with functional datasets

Projection onto a wavelet basis

To deal with functional data, we project them onto some basis, to obtain data as described in the Gaussian mixture regression models (1.1). In this chapter, we choose to deal with wavelet basis, given that they represent localized features of functions in a sparse way. If the coefficients matrix $\mathbf{x}$ and $\mathbf{y}$ are sparse, the regression matrix β has more chance to be sparse. Moreover, we could represent a signal with a few coefficients dataset, then it is a way to reduce the dimension. For details about the wavelet theory, see the Mallat's book [Mal99].
Begin by an overview of some important aspects of wavelet basis.
Let ψ a real wavelet function, satisfying

$$\psi \in L^1 \cap L^2, t\psi \in L^1, \text{ and } \int_{\mathbb{R}} \psi(t)dt = 0.$$

We denote by $\psi_{l,h}$ the function defined from ψ by dyadic dilation and translation:

$$\psi_{l,h}(t) = 2^{l/2}\psi(2^l t - h) \text{ for } (l,h) \in \mathbb{Z}^2.$$

We could define wavelet coefficients of a signal f by

$$d_{l,h}(f) = \int_{\mathbb{R}} f(t)\psi_{l,h}(t)dt \text{ for } (l,h) \in \mathbb{Z}^2.$$

Let φ be a scaling function related to ψ, and $\varphi_{l,h}$ the dilatation and translation of φ for $(l,h) \in \mathbb{Z}^2$. We also define, for $(l,h) \in \mathbb{Z}^2$, $\beta_{l,h}(f) = \int_{\mathbb{R}} f(t)\varphi_{l,h}(t)dt$.
Note that scaling functions serve to construct approximations of the function of interest, while the wavelet functions serve to provide the details not captured by successive approximations.
We denote by V_l the space generated by $\{\varphi_{l,h}\}_{h\in\mathbb{Z}}$, and by W_l the space egenrated by $\{\psi_{l,h}\}_{h\in\mathbb{Z}}$ for all $l \in \mathbb{Z}$. Remark that

$$V_{l-1} = V_l \oplus W_l \text{ for all } l \in \mathbb{Z}$$
$$L^2 = \oplus_{l\in\mathbb{Z}} W_l$$

Let $L \in \mathbb{N}^*$. For a signal f, we could define the approximation at the level L by

$$A_L = \sum_{l>L}\sum_{h\in\mathbb{Z}} d_{l,h}\psi_{l,h};$$

and f could be decomposed by the approximation at the level L and the details $(d_{l,h})_{l<L}$.
The decomposition of the basis between scaling function and wavelet function emphasizes on the local nature of the wavelets, and that is an important aspect in our procedures, because we want to know which details allow us to cluster two observations together.
Consider the sample $(f_i, g_i)_{1\leq i\leq n}$, and introduce the wavelet expansion of f_i in the basis $\mathcal{B}$: for all $t \in [0,1]$,

$$f_i(t) = \underbrace{\sum_{h\in\mathbb{Z}} \beta_{L,h}(f_i)\varphi_{L,h}(t)}_{A_L} + \sum_{l\leq L}\sum_{h\in\mathbb{Z}} d_{l,h}(f_i)\psi_{l,h}(t).$$

The collection $\{\beta_{L,h}(f_i), d_{l,h}(f_i)\}_{l\leq L, h\in\mathbb{Z}}$ is the Discrete Wavelet Transform (DWT) of f in the basis $\mathcal{B}$.
Because we project onto an orthonormal basis, this leads to a n-sample $(x_1, \ldots, x_n)$ of wavelet coefficient decomposition vectors, with

$$f_i = Wx_i;$$

in which x_i is the vector of the discretized values of the signal, x_i the matrix of coefficients in the basis $\mathcal{B}$, and W a $p \times p$ matrix defined by φ and ψ. The DWT can be performed by a computationally fast pyramid algorithm (see Mallat, [Mal89]). In the same way, there exists $W^{'}$ such that $g_i = W^{'}y_i$, with $\mathbf{y} = (y_1, \ldots, y_n)$ a n sample of wavelet coefficient decomposition vectors. Because the matrices W and $W^{'}$ are orthogonal, we keep the mixture structure, and the noise is also Gaussian. We could consider the wavelet coefficient dataset $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$, which defines of n observations whose probability distribution could be modeled by the finite Gaussian mixture regression model (1.1).

Our procedures

We could apply our both procedures to this dataset, and obtain a clustering of the data. Indeed, rather than considering $(\mathbf{f}, \mathbf{g})$, we run our procedures on the sample $(\mathbf{x}, \mathbf{y})$, varying the number of clusters in $\mathcal{K}$.
The notion of relevant variable is natural: the function $\varphi_{l,h}$ or $\psi_{l,h}$ is irrelevant if it appears in none of the wavelet coefficient decomposition of the functions in each cluster.

1.5.3 Numerical experiments

We will illustrate our procedures on functional datasets by using the Matlab wavelet toolbox (see Misiti et al. in [MMOP04] for details). Firstly, we simulate functional datasets, where the true model belongs to the model collection. Then, we run our procedure on an electricity dataset, to cluster successive days. We have access to time series, measured every half-hour, of a load consumption, on 70 days. We extract the signal of each day, and construct couples by each day and its eve, and we aim at clustering these couples. To finish, we test our procedures on the well-known Tecator dataset. This benchmark dataset corresponds to the spectrometric curves and fat contents of meat. These experiments illustrate different aspects of our procedures. Indeed, the simulated example proves that our procedures work in a functional context. The second example is a toy example used to validate the classification, on real data already studied, and in which we clearly understand the clusters. The last example illustrates the use of the classification to perform prediction, and the description given by our procedures to the model constructed.

Simulated functional data

Firstly, we simulate a mixture regression model. Let $\mathbf{f}$ be a sample of the noised cosine function, discretized on a 15 points grid. Let $\mathbf{g}$ be, depending on the cluster, either $\mathbf{f}$, or the function $-\mathbf{f}$, computed by a white-noise.
We use the Daubechies-2 basis at level 2 to decompose the signal.
Our procedures are run 20 times, and the number of clusters are fixed to $\mathcal{K} = 2$. Then our procedures run on the projection are compared with the oracle among the collection constructed by the Lasso-MLE procedure, and with the model selected by the BIC criterion among this collection. The MLE is also computed.

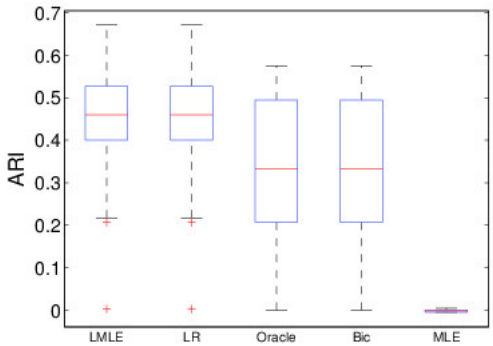

Figure 1.13: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the Lasso-Rank procedure (LR), the oracle (Oracle), the BIC estimator (Bic) and the MLE (MLE)

This simulated dataset proves that our procedures also perform clustering functional data, considering the projection dataset.

Electricity dataset

We also study the clustering on electricity dataset. This example is studied in [MMOP07]. We work on a sample of size 70 of couples of days, which is plotted in Figure 5.1. For each couple, we have access to the half-hour load consumption.

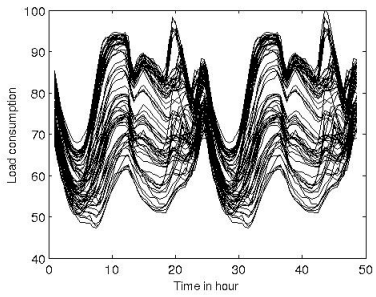

Figure 1.14: Plot of the 70-sample of half-hour load consumption, on the two days

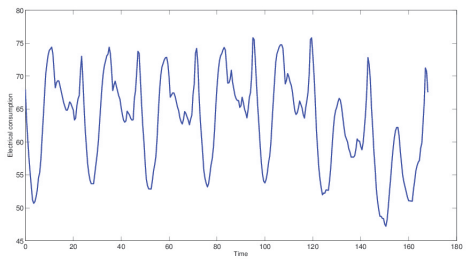

Figure 1.15: Plot of a week of load consumption

As we said previously, we want to cluster the relationship between two successive days. In Figure 1.15, we plot a week of consumption.

The regression model taken is F for the first day, and G for the second day of each couple. Besides, discretization of each day on 48 points, every half-hour, is made available. In our opinion, a linear model is not appropriate, as the behavior from the eve to the day depends on which day we consider: there is a difference between working days and weekend days, as involved in Figure 1.15.
To apply our procedures, we project $\mathbf{f}$ and $\mathbf{g}$ onto wavelet basis. The symmlet-4 basis, at level 5, is used.
We run our procedures with the number of clusters varying from 2 to 6. Our both procedures select a model with 4 components. The first one considers couples of weekdays, the second Friday-Saturday, the third component is Saturday-Sunday and the fourth considers Sunday-Monday. This result is faithful with the knowledge we have about these data. Indeed, working days have the same behavior, depending on the eve, whereas days off have not the same behavior, depending on working days, and conversely. Moreover, in the article [MMOP07], which also studied this example, they get the same classification.

Tecator dataset

This example deals with spectrometric data. More precisely, a food sample has been considered, which contained finely chopped pure meat with different fat contents. The data consist of a 100-channel spectrum of absorbances in the wavelength range $850 - 1050$ nm, and of the percentage of fat. We observe a sample of size 215. Those data have been studied in a lot of articles, cite for example Ferraty and Vieu's book [FV06]. They work on different approaches. They test prediction, and classification, supervised (where the fat content become a class, larger or smaller than 20%), or not (ignoring the response variable). In this work, we focus on clustering data according to the reliance between the fat content and the absorbance spectrum. We could not predict the response variable, because we do not know the class of a new observation. Estimate it is a difficult problem, in which we are not involved in this chapter.
We will take advantage of our procedures to know which coefficients, in the wavelet basis decomposition of the spectrum, are useful to describe the fat content.
The sample will be split into two subsamples, 165 observations for the learning set, and 50 observations for the test set. We split it to have the same marginal distribution for the response in each sample.
The spectrum is a function, which we decompose into the Haar basis, at level 6. Nevertheless, our model did not take into account a constant coefficient to describe the response. Thereby, before run our procedure, we center and the y according to the learning sample, and each function x_i for all observations in the whole sample. Then, we could estimate the mean of the response by the mean $\hat{\mu}$ over the learning sample.
We construct models on the training set by our procedure Lasso-MLE. Thanks to the estimations, we have access to relevant variables, and we could reconstruct signals keeping only relevant variables. We have also access to the a posteriori probability, which leads to know which observation is with high probability in which cluster. However, for some observations, the a posteriori probability do not ensure the clustering, being almost the same for different clusters. The procedure selects two models, which we describe here. In Figures 1.16 and 1.17, we represent clusters done on the training set for the different models. The graph on the left is a candidate for representing each cluster, constructed by the mean of spectrum over an a posteriori probability greater than 0.6. We plot the curve reconstruction, keeping only relevant variables in the wavelet decomposition. On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.
The first model has two classes, which could be distinguish in the absorbance spectrum by the

bump on wavelength around 940 nm. The first cluster is dominating, with $\hat{\pi}_1 = 0.95$. The fat content is smaller in the first cluster than in the second cluster. According to the signal reconstruction, we could see that almost all variables have been selected. This model seems consistent according to the classification goal.
The second model has 3 classes, and we could remark different important wavelength. Around 940 nm, there is some difference between classes, corresponding to the bump underline in the model 1, but also around 970 nm, with higher or smaller values. The first class is dominating, with $\hat{\pi}_1 = 0.89$. Just a few of variables have been selected, which give to this model the understanding property of which coefficients are discriminating.
We could discuss about those models. The first one selects only two classes, but almost all variables, whereas the second model has more classes, and less variables: there is a trade-off between clusters and variable selection for the dimension reduction.

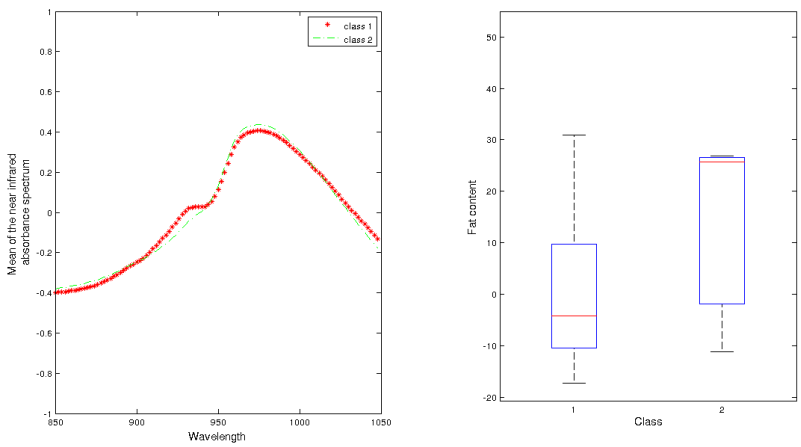

Figure 1.16: Summarized results for the model 1. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than 0.6 On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.

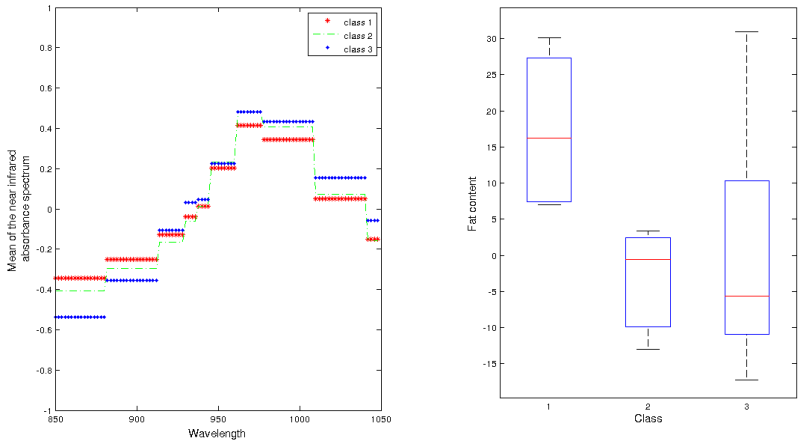

Figure 1.17: Summarized results for the model 2. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than 0.6 On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.

According to those classifications, we could compute the response according to the linear model. We use two ways to compute $\hat{y}$: either consider the linear model in the cluster selected by the MAP principle, or mix estimations in each cluster thanks to these a posteriori probabilities. We compute the Mean Absolute Percentage Error, MAPE $= \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - y_i)/y_i|$. Results are summarized in Table 1.3.

	Linear model in the class with higher probability	Mixing estimation
Model 1	0.200	0.198
Model 2	0.055	0.056

Table 1.3: Mean absolute percentage of error of the predicted value, for each model, for the learning sample

Thus, we work on the test sample. We use the response and the regressors to know the a posteriori of each observation. Then, using our models, we could compute the predicted fat values from the spectrometric curve, as before according to two ways, mixing or choosing the classes.

	Linear model in the class with higher probability	Mixing estimation
Model 1	0.22196	0.21926
Model 2	0.20492	0.20662

Table 1.4: Mean absolute percentage of error of the predicted value, for each model, for the test sample

Because the models are constructed on the learning sample, MAPE are lower than for the test sample. Nevertheless, results are similar, saying that models are well constructed. This is particularly the case for the model 1, which is more consistent over a new sample.

To conclude this study, we could highlight the advantages of our procedure on these data. It provides a clustering of data, similar to the one done with supervised clustering in [FV06], but we could explain how this clustering is done.
This work has been done with the Lasso-MLE procedure. Nevertheless, the same kind of results have been get with the Lasso-Rank procedure.

1.6 Conclusion

In this chapter, two procedures are proposed to cluster regression data. Detecting the relevant clustering variables, they are especially designed for high-dimensional datasets. We use an ℓ_1-regularization procedure to select variables, and then deduce a reasonable random model collection. Thus, we recast estimations of parameters of these models into a general model selection problem. These procedures are compared with usual criteria on simulated data: the BIC criterion used to select a model, the maximum-likelihood estimator, and to the oracle when we know it. In addition, we compare our procedures to others on benchmark data.
One main asset of those procedures is that it can be applied to functional datasets. We also develop this point of view.

1.7 Appendices

In those appendices, we develop calculus for EM algorithm updating formulae in Section 1.7.1, for Lasso and maximum likelihood estimators, and for low ranks estimators. In Section 1.7.2, we extend our procedures with the Group-Lasso estimator to select relevant variables, rather than use the Lasso estimator.

1.7.1 EM algorithms

EM algorithm for the Lasso estimator

Introduced by Dempster et al. in [DLR77], the EM (Expectation-Maximization) algorithm is used to compute maximum likelihood estimators, penalized or not.
The expected complete negative log-likelihood is denoted by

$$Q(\theta|\theta') = -\frac{1}{n} E_{\theta'}(l_c(\theta, \mathbf{X}, \mathbf{Y}, \mathbf{Z})|\mathbf{X}, \mathbf{Y})$$

in which

$$\begin{aligned} l_c(\theta, \mathbf{X}, \mathbf{Y}, \mathbf{Z}) = & \sum_{i=1}^{n} \sum_{k=1}^{K} [Z_i]_k \log\left(\frac{\det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{1}{2}(P_k Y_i - X_i \Phi_k)^t (P_k Y_i - X_i \Phi_k)\right)\right) \\ & + [Z_i]_k \log(\pi_k); \end{aligned}$$

with $[Z_i]_k$ are independent and identically distributed unobserved multinomial variables, showing the component-membership of the ith observation in the finite mixture regression model.
The expected complete penalized negative log-likelihood is

$$Q_{\text{pen}}(\theta|\theta') = Q(\theta|\theta') + \lambda \sum_{k=1}^{K} \pi_k ||\Phi_k||_1.$$

Calculus for updating formula

— E-step: compute $Q(\theta|\theta^{(\text{ite})})$, or, equivalently, compute for $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$,

$$\begin{aligned}\tau_{i,k}^{(\text{ite})} &= E_{\theta^{(\text{ite})}}([Z_i]_k|\mathbf{Y}) \\ &= \frac{\pi_k^{(\text{ite})} \det P_k^{(\text{ite})} \exp\left(-\frac{1}{2}\left(P_k^{(\text{ite})}Y_i - X_i\Phi_k^{(\text{ite})}\right)^t \left(P_k^{(\text{ite})}Y_i - X_i\Phi_k^{(\text{ite})}\right)\right)}{\sum_{r=1}^K \pi_k^{(\text{ite})} \det P_k^{(\text{ite})} \exp\left(-\frac{1}{2}\left(P_k^{(\text{ite})}Y_i - X_i\Phi_k^{(\text{ite})}\right)^t \left(P_k^{(\text{ite})}Y_i - X_i\Phi_k^{(\text{ite})}\right)\right)}\end{aligned}$$

This formula updates the clustering, thanks to the MAP principle.

— M-step: improve $Q_{\text{pen}}(\theta|\theta^{(\text{ite})})$.

For this, rewrite the Karush-Kuhn-Tucker conditions. We have

$$\begin{aligned}&Q_{\text{pen}}(\theta|\theta^{(\text{ite})}) \\ &= -\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K E_{\theta^{(\text{ite})}}\left([Z_i]_k \log\left(\frac{\det(P_k)}{(2\pi)^{q/2}}\exp\left(-\frac{1}{2}(P_kY_i - X_i\Phi_k)^t(P_kY_i - X_i\Phi_k)\right)\right)|\mathbf{Y}\right) \\ &\quad -\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K E_{\theta^{(\text{ite})}}([Z_i]_k \log\pi_k|\mathbf{Y}) + \lambda\sum_{k=1}^K \pi_k||\Phi_k||_1 \\ &= -\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K -\frac{1}{2}(P_kY_i - X_i\Phi_k)^t(P_kY_i - X_i\Phi_k)E_{\theta^{(\text{ite})}}([Z_i]_k|\mathbf{Y}) \\ &\quad -\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K\sum_{z=1}^q \log\left(\frac{[P_k]_{z,z}}{\sqrt{2\pi}}\right)E_{\theta^{(\text{ite})}}[[Z_i]_k|\mathbf{Y}] \\ &\quad -\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K E_{\theta^{(\text{ite})}}([Z_i]_k|\mathbf{Y})\log\pi_k + \lambda\sum_{k=1}^K \pi_k||\Phi_k||_1. \end{aligned} \tag{1.11}$$

Firstly, we optimize this formula with respect to $\boldsymbol{\pi}$: it is equivalent to optimize

$$-\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \tau_{i,k}\log(\pi_k) + \lambda\sum_{k=1}^K \pi_k||\Phi_k||_1.$$

We obtain

$$\pi_k^{(\text{ite}+1)} = \pi_k^{(\text{ite})} + t^{(\text{ite})}\left(\frac{\sum_{i=1}^n \tau_{i,k}}{n} - \pi_k^{(\text{ite})}\right);$$

with $t^{(\text{ite})} \in (0, 1]$, the largest value in the grid $\{\delta^l, l \in \mathbb{N}\}$, with $0 < \delta < 1$, such that the function is not increasing.

To optimize (1.11) with respect to $(\mathbf{\Phi}, \mathbf{P})$, we could rewrite the expression: it is similar to the optimization of

$$-\frac{1}{n}\sum_{i=1}^n\left(\tau_{i,k}\sum_{z=1}^q \log([P_k]_{z,z}) - \frac{1}{2}(P_k\tilde{Y}_i - \tilde{X}_i\Phi_k)^t(P_k\tilde{Y}_i - \tilde{X}_i\Phi_k)\right) + \lambda\pi_k||\Phi_k||_1$$

for all $k \in \{1, \dots, K\}$, which is equivalent to the optimization of

$$-\frac{1}{n}n_k\sum_{z=1}^q \log([P_k]_{z,z}) + \frac{1}{2n}\sum_{i=1}^n\sum_{z=1}^q\left([P_k]_{z,z}[\tilde{Y}_i]_{k,z} - [\Phi_k]_{z,.}[\tilde{X}_i]_{k,.}\right)^2 + \lambda\pi_k||\Phi_k||_1;$$

where $n_k = \sum_{i=1}^n \tau_{i,k}$. The minimum in $[P_k]_{z,z}$ is the function which cancel its partial derivative with respect to $[P_k]_{z,z}$:

$$-\frac{n_k}{n}\frac{1}{[P_k]_{z,z}} + \frac{1}{2n}\sum_{i=1}^n 2[\tilde{Y}_i]_{k,z}\left([P_k]_{z,z}[\tilde{Y}_i]_{k,z} - [\Phi_k]_{z,.}[\tilde{X}_i]_{k,.}\right) = 0$$

for all $k \in \{1,\dots,K\}$, for all $z \in \{1,\dots,q\}$, which is equivalent to

$$\begin{aligned}
&-1 + \frac{1}{n_k}[P_k]_{z,z}^2\sum_{i=1}^n[\tilde{Y}_i]_{k,z}^2 - \frac{1}{n_k}[P_k]_{z,z}\sum_{i=1}^n[\tilde{Y}_i]_{k,z}[\Phi_k]_{z,.}[\tilde{X}_i]_{k,.} = 0\\
\Leftrightarrow\ &-1 + [P_k]_{z,z}^2\frac{1}{n_k}||[\tilde{\mathbf{Y}}]_{k,z}||_2^2 - [P_k]_{z,z}\frac{1}{n_k}\langle[\tilde{\mathbf{Y}}]_{k,z},[\Phi_k]_{z,.}[\tilde{\mathbf{X}}]_{k,.}\rangle = 0.
\end{aligned}$$

The discriminant is

$$\Delta = \left(-\frac{1}{n_k}\langle[\tilde{\mathbf{Y}}]_{k,z},[\Phi_k]_{z,.}[\tilde{\mathbf{X}}]_{k,.}\rangle\right)^2 - \frac{4}{n_k}||[\tilde{\mathbf{Y}}]_{k,z}||_2^2.$$

Then, for all $k \in \{1,\dots,K\}$, for all $z \in \{1,\dots,q\}$,

$$[P_k]_{z,z} = \frac{n_k\langle[\tilde{\mathbf{Y}}]_{k,z},[\Phi_k]_{z,.}[\tilde{\mathbf{X}}]_{k,.}\rangle + \sqrt{\Delta}}{2n_k||[\tilde{\mathbf{Y}}]_{k,z}||_2^2}.$$

We could also look at the equation (1.11) as a function of the variable $\mathbf{\Phi}$: according to the partial derivative with respect to $[\Phi_k]_{z,j}$, we obtain for all $z \in \{1,\dots,q\}$, for all $k \in \{1,\dots,K\}$, for all $j \in \{1,\dots,p\}$,

$$\sum_{i=1}^n[\tilde{X}_i]_{k,j}\left([P_k]_{z,z}[\tilde{Y}_i]_{k,z} - \sum_{j_2=1}^p[\tilde{X}_i]_{k,j_2}[\Phi_k]_{z,j_2}\right) - n\lambda\pi_k\mathrm{sgn}([\Phi_k]_{z,j}) = 0.$$

Then, for all $k \in \{1,\dots,K\}, j \in \{1,\dots,p\}, z \in \{1,\dots,q\}$,

$$[\Phi_k]_{z,j} = \frac{\sum_{i=1}^n[\tilde{X}_i]_{k,j}[P_k]_{z,z}[\tilde{Y}_i]_{k,z} - \sum_{\substack{j_2=1\\ j_2\neq j}}^p[\tilde{X}_i]_{k,j}[\tilde{X}_i]_{k,j_2}[\Phi_k]_{z,j_2} - n\lambda\pi_k\mathrm{sgn}([\Phi_k]_{j,z})}{||[\tilde{\mathbf{X}}]_{k,j}||_2^2}.$$

To reduce notations, let, for all $k \in \{1,\dots,K\}, j \in \{1,\dots,p\}, z \in \{1,\dots,q\}$,

$$[S_k]_{j,z} = -\sum_{i=1}^n[\tilde{X}_i]_{k,j}[P_k]_{z,z}[\tilde{Y}_i]_{k,z} + \sum_{\substack{j_2=1\\ j_2\neq j}}^p[\tilde{X}_i]_{k,j}[\tilde{X}_i]_{k,j_2}[\Phi_k]_{z,j_2}.$$

Then

$$\begin{aligned}
[\Phi_k]_{z,j} &= \frac{-[S_k]_{j,z} - n\lambda\pi_k\mathrm{sgn}([\Phi_k]_{z,j})}{||[\tilde{\mathbf{X}}]_{k,j}||_2^2}\\
&= \begin{cases}
\frac{-[S_k]_{j,z}+n\lambda\pi_k}{||[\tilde{\mathbf{X}}]_{k,j}||_2^2} & \text{if } [S_k]_{j,z} > n\lambda\pi_k\\
-\frac{[S_k]_{j,z}+n\lambda\pi_k}{||[\tilde{\mathbf{X}}]_{k,j}||_2^2} & \text{if } [S_k]_{j,z} < -n\lambda\pi_k\\
0 & \text{elsewhere.}
\end{cases}
\end{aligned}$$

From these equalities, we could write the updating formulae. For $j \in \{1, \dots, p\}, k \in \{1, \dots, K\}$, $z \in \{1, \dots, q\}$,

$$[S_k]_{j,z}^{(\text{ite})} = -\sum_{i=1}^{n} [\tilde{X}_i]_{k,j} [P_k]_{z,z}^{(\text{ite})} [\tilde{Y}_i]_{k,z} + \sum_{\substack{j_2=1 \\ j_2 \neq j}}^{p} [\tilde{X}_i]_{k,j} [\tilde{X}_i]_{k,j_2} [\Phi_k]_{z,j_2}^{(\text{ite})};$$

$$n_k = \sum_{i=1}^{n} \tau_{i,k};$$

$$([\tilde{Y}_i]_{k,.}, [\tilde{X}_i]_{k,.}) = \sqrt{\tau_{i,k}} (Y_i, X_i).$$

and $t^{(ite)} \in (0, 1]$, the largest value in the grid $\{\delta^l, l \in \mathbb{N}\}$, $0 < \delta < 1$, such that the function is not increasing.

EM algorithm for the rank procedure

To take into account the matrix structure, we want to make a dimension reduction on the rank of the mean matrix. If we know to which cluster each sample belongs, we could compute the low rank estimator for linear model in each component.
Indeed, an estimator of fixed rank r is known in the linear regression case: denoting A^+ the Moore-Penrose pseudo-inverse of A, and $[A]_r = UD_rV^t$ in which D_r is obtained from D by setting $(D_r)_{i,i} = 0$ for $i \geq r+1$, with UDV^t the singular decomposition of A, if $Y = \beta X + \Sigma$, an estimator of β with rank r is $\hat{\beta}_r = [(\mathbf{x}^t\mathbf{x})^+\mathbf{x}^t\mathbf{y}]_r$.
We do not know the clustering of the sample, but the E-step in the EM algorithm computes it. We suppose in this case that Σ_k and π_k are known, for all $k \in \{1, \dots, K\}$. We use this algorithm to determine Φ_k, for all $k \in \{1, \dots, K\}$, with ranks fixed to $\mathbf{R} = (R_1, \dots, R_K)$.

— E-step: compute for $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$,

$$\begin{aligned}\tau_{i,k} &= E_{\theta^{(\text{ite})}}([Z_i]_k | Y) \\ &= \frac{\pi_k^{(\text{ite})} \det P_k^{(\text{ite})} \exp\left(-\frac{1}{2}\left(P_k^{(\text{ite})} y_i - x_i \Phi_k^{(\text{ite})}\right)^t \left(P_k^{(\text{ite})} y_i - x_i \Phi_k^{(\text{ite})}\right)\right)}{\sum_{r=1}^{K} \pi_k^{(\text{ite})} \det P_k^{(\text{ite})} \exp\left(-\frac{1}{2}\left(P_k^{(\text{ite})} y_i - x_i \Phi_k^{(\text{ite})}\right)^t \left(P_k^{(\text{ite})} y_i - x_i \Phi_k^{(\text{ite})}\right)\right)}\end{aligned}$$

— M-step: assign each observation in its estimated cluster, by the MAP principle applied thanks to the E-step. We say that Y_i comes from component number $\underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \tau_{i,k}^{(\text{ite})}$.
Then, we can define $\tilde{\beta}_k^{(\text{ite})} = (\mathbf{x}_{|k}^t \mathbf{x}_{|k})^{-1} \mathbf{x}_{|k}^t \mathbf{y}_{|k}$, in which $\mathbf{x}_{|k}$ and $\mathbf{y}_{|k}$ are a restriction of the sample to the cluster k, which we decompose in singular value with $\tilde{\beta}_k^{(\text{ite})} = USV^t$. Using the singular value decomposition described before, we obtain the estimator.

1.7.2 Group-Lasso MLE and Group-Lasso Rank procedures

One way to perform those procedures is to consider the Group-Lasso estimator rather than the Lasso estimator to select relevant variables. Indeed, this estimator is more natural, according to the relevant variable definition. Nevertheless, results are very similar, because we select grouped variables in both case, selected by the Lasso or by the Group-Lasso estimator. In this section, we describe our procedures with the Group-Lasso estimator, which could be understood as an improvement of our procedures.

Context - definitions

Our both procedures take advantage of the Lasso estimator to select relevant variables, to reduce the dimension in case of high-dimensional datasets. First, recall what is a relevant variable.

Definition 1.7.1. *A variable indexed by* $(z, j) \in \{1, \ldots, q\} \times \{1, \ldots, p\}$ *is* irrelevant *for the clustering if* $[\Phi_1]_{z,j} = \ldots = [\Phi_K]_{z,j} = 0$. *A* relevant *variable is a variable which is not irrelevant. We denote by* J *the relevant variables set.*

According to this definition, we could introduce the Group-Lasso estimator.

Definition 1.7.2. *The* Lasso estimator *for mixture regression models with regularization parameter* $\lambda \geq 0$ *is defined by*

$$\hat{\theta}^{Lasso}(\lambda) := \underset{\theta \in \Theta_K}{\operatorname{argmin}} \left\{ -\frac{1}{n} l_\lambda(\theta) \right\};$$

where

$$-\frac{1}{n} l_\lambda(\theta) = -\frac{1}{n} l(\theta) + \lambda \sum_{k=1}^{K} \pi_k ||\Phi_k||_1;$$

where $||\Phi_k||_1 = \sum_{j=1}^{p} \sum_{z=1}^{q} |[\Phi_k]_{z,j}|$, *and with* λ *to specify.*

It is the estimator used in the both procedures described in previous parts.

Definition 1.7.3. *The* Group-Lasso estimator *for mixture regression models with regularization parameter* $\lambda \geq 0$ *is defined by*

$$\hat{\theta}^{Group\text{-}Lasso}(\lambda) := \underset{\theta \in \Theta_K}{\operatorname{argmin}} \left\{ -\frac{1}{n} \tilde{l}_\lambda(\theta) \right\};$$

where

$$-\frac{1}{n} \tilde{l}_\lambda(\theta) = -\frac{1}{n} l(\theta) + \lambda \sum_{j=1}^{p} \sum_{z=1}^{q} \sqrt{k} ||[\mathbf{\Phi}]_{z,j}||_2;$$

where $||[\mathbf{\Phi}]_{z,j}||_2^2 = \sum_{k=1}^{K} |[\Phi_k]_{z,j}|^2$, *and with* λ *to specify.*

This Group-Lasso estimator has the advantage to cancel grouped variables rather than variables one by one. It is consistent with the relevant variable definition.
Nevertheless, depending on datasets, it could be interesting to look after which variables are canceled first. One way could be to extend this work with Lasso-Group-Lasso estimator, described for the example for the linear model in [SFHT13].
Let describe two additional procedures, which will use the Group-Lasso estimator rather than the Lasso estimator to detect relevant variables.

Group-Lasso-MLE procedure

This procedure is decomposed into three main steps: we construct a model collection, then in each model we compute the maximum likelihood estimator, and we select the best one among all the models.
The first step consists of constructing a collection of models $\{\mathcal{H}_{(K,\tilde{J})}\}_{(K,\tilde{J}) \in \mathcal{M}}$ in which $\mathcal{H}_{(K,\tilde{J})}$ is defined by

$$\mathcal{H}_{(K,\tilde{J})} = \{y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto h_\theta(y|x)\}; \tag{1.12}$$

where

$$h_\theta(y|x) = \sum_{k=1}^{K} \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y - \Phi_k^{[\tilde{J}]} x)^t (P_k y - \Phi_k^{[\tilde{J}]} x)}{2}\right),$$

and

$$\theta = (\pi_1, \ldots, \pi_K, \Phi_1, \ldots, \Phi_K, \rho_1, \ldots, \rho_K) \in \Pi_K \times \left(\mathbb{R}^{q\times p}\right)^K \times \left(\mathbb{R}_+^q\right)^K.$$

The model collection is indexed by $\mathcal{M} = \mathcal{K} \times \tilde{\mathcal{J}}$. Denote $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components. We could bound $\mathcal{K}$ without loss of estimation. Denote also $\tilde{\mathcal{J}}$ a collection of subsets of $\{1, \ldots, q\} \times \{1, \ldots, p\}$, constructed by the Group-Lasso estimator.
To detect the relevant variables, and construct the set $\tilde{J} \in \tilde{\mathcal{J}}$, we will use the Group-Lasso estimator defined by (1.7.3). In the ℓ_1-procedures, the choice of the regularization parameters is often difficult: fixing the number of components $K \in \mathcal{K}$, we propose to construct a data-driven grid G_K of regularization parameters by using the updating formulae of the mixture parameters in the EM algorithm.
Then, for each $\lambda \in G_K$, we could compute the Group-Lasso estimator defined by

$$\hat{\theta}^{\text{Group-Lasso}} = \underset{\theta\in\Theta_K}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(h_\theta(y_i|x_i)) + \lambda \sum_{j=1}^{p}\sum_{z=1}^{q} \sqrt{K} ||[\mathbf{\Phi}]_{z,j}||_2 \right\}.$$

For a fixed number of mixture components $K \in \mathcal{K}$ and a regularization parameter λ, we could use a generalized EM algorithm to approximate this estimator. Then, for each $K \in \mathcal{K}$, and for each $\lambda \in G_K$, we have constructed the relevant variables set $\tilde{J}_\lambda$. We denote by $\tilde{\mathcal{J}}$ the collection of all these sets.
The second step consists of approximating the MLE

$$\hat{h}^{(K,\tilde{J})} = \underset{t\in\mathcal{H}_{(K,\tilde{J})}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(t(y_i|x_i)) \right\};$$

using the EM algorithm for each model $(K, \tilde{J}) \in \mathcal{M}$.
The third step is devoted to model selection. We use the slope heuristic described in [BM07].

Group-Lasso-Rank procedure

As when the relevant variables were selected by the Lasso estimator, whereas the previous procedure does not take into account the multivariate structure, we propose a second procedure to perform this point. For each model belonging to the collection $\mathcal{H}_{(K,\tilde{J})}$, a subcollection is constructed, varying the rank of $\mathbf{\Phi}$. Let us describe this procedure.
As in the Group-Lasso-MLE procedure, we first construct a collection of models, thanks to the ℓ_1-approach. We obtain an estimator for θ, denoted by $\hat{\theta}^{\text{Group-Lasso}}$, for each model belonging to the collection. We could deduce the set of relevant variables, denoted by $\tilde{J}$, and this for all $K \in \mathcal{K}$: we deduce $\tilde{\mathcal{J}}$ the collection of set of relevant variables.
The second step consists to construct a subcollection of models with rank sparsity, denoted by

$$\{\check{\mathcal{H}}_{(K,\tilde{J},R)}\}_{(K,\tilde{J},R)\in\check{\mathcal{M}}}.$$

The model $\{\check{\mathcal{H}}_{(K,\tilde{J},R)}\}$ has K components, the set $\tilde{J}$ for active variables, and R is the vector of the ranks of the matrix of regression coefficients in each group:

$$\check{\mathcal{H}}_{(K,\tilde{J},R)} = \left\{ y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto h_\theta^{(K,\tilde{J},R)}(y|x) \right\} \tag{1.13}$$

where

$$
\begin{aligned}
h_\theta^{(K,\tilde{J},R)}(y|x) =& \sum_{k=1}^{K} \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{(P_k y - (\Phi_k^{R_k})^{[\tilde{J}]}x)^t (P_k y - (\Phi_k^{R_k})^{[\tilde{J}]}x)}{2}\right); \\
\theta =& (\pi_1, \ldots, \pi_K, \Phi_1^{R_1}, \ldots, \Phi_K^{R_K}, P_1, \ldots, P_K) \in \Pi_K \times \Psi_K^R \times \left(\mathbb{R}_+^q\right)^K; \\
\Psi_K^R =& \left\{(\Phi_1^{R_1}, \ldots, \Phi_K^{R_K}) \in \left(\mathbb{R}^{q\times p}\right)^K | \mathrm{Rank}(\Phi_1) = R_1, \ldots, \mathrm{Rank}(\Phi_K) = R_K\right\};
\end{aligned}
$$

and $\tilde{\mathcal{M}}^R = \mathcal{K} \times \tilde{\mathcal{J}} \times \mathcal{R}$. Denote $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components, $\tilde{\mathcal{J}}$ a collection of subsets of $\{1, \ldots, q\} \times \{1, \ldots, p\}$, and $\mathcal{R}$ the set of vectors of size $K \in \mathcal{K}$ with ranks values for each mean matrix. We could compute the MLE under the rank constraint thanks to an EM algorithm. Indeed, we could constrain the estimation of Φ_k, for all k, to have a rank equal to R_k, in keeping only the R_k largest singular values. More details are given in Section 1.7.1. It leads to an estimator of the mean with row sparsity and low rank for each model.

Chapter 2

An ℓ_1-oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models

Contents

1.1 Introduction **40**
1.2 Gaussian mixture regression models **42**
1.2.1 Gaussian mixture regression 42
1.2.2 Clustering with Gaussian mixture regression 43
1.2.3 EM algorithm 44
1.2.4 The model collection 45
1.3 Two procedures **46**
1.3.1 Lasso-MLE procedure 46
1.3.2 Lasso-Rank procedure 47
1.4 Illustrative example **48**
1.4.1 The model 48
1.4.2 Sparsity and model selection 49
1.4.3 Assessment 50
1.5 Functional datasets **54**
1.5.1 Functional regression model 55
1.5.2 Two procedures to deal with functional datasets 55
Projection onto a wavelet basis 55
Our procedures 56
1.5.3 Numerical experiments 56
Simulated functional data 57
Electricity dataset 57
Tecator dataset 58
1.6 Conclusion **61**
1.7 Appendices **61**
1.7.1 EM algorithms 61
EM algorithm for the Lasso estimator 61
EM algorithm for the rank procedure 64
1.7.2 Group-Lasso MLE and Group-Lasso Rank procedures 64
Context - definitions 65
Group-Lasso-MLE procedure 65

Group-Lasso-Rank procedure 66

This chapter focuses on the Lasso estimator for its regularization properties. We consider a finite mixture of Gaussian regressions for high-dimensional heterogeneous data, where the number of covariates and the dimension of the response may be much larger than the sample size. We estimate the unknown conditional density by an ℓ_1-penalized maximum likelihood estimator. We provide an ℓ_1-oracle inequality satisfied by this Lasso estimator according to the Kullback-Leibler loss. This result is an extension of the ℓ_1-oracle inequality established by Meynet in [Mey13] in the multivariate case. It is deduced from a model selection theorem, the Lasso being viewed as the solution of a penalized maximum likelihood model selection procedure over a collection of ℓ_1-ball models.

2.1 Introduction

Finite mixture regression models are useful for modeling the relationship between response and predictors, arising from different subpopulations. Due to recent improvements, we are faced with high-dimensional data where the number of variables can be much larger than the sample size. We have to reduce the dimension to avoid identifiability problems. Considering a mixture of linear models, an assumption widely used is to say that only a few covariates explain the response. Among various methods, we focus on the ℓ_1-penalized least squares estimator of parameters to lead to sparse regression matrix. Indeed, it is a convex surrogate for the non-convex ℓ_0-penalization, and produces sparse solutions. First introduced by Tibshirani in [Tib96] in a linear model $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and $\epsilon \sim \mathcal{N}(0, \Sigma)$, the Lasso estimator is defined in the linear model by

$$\hat{\beta}^{\text{Lasso}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}, \qquad \lambda > 0.$$

Many results have been proved to study the performance of this estimator. For example, cite [BRT09] and [EHJT04], for studying this estimator as a variable selection procedure in this linear model case. Note that those results need strong assumptions on the Gram matrix X^tX, as the restrictive eigenvalue condition, that can be not fulfilled in practice. A summary of assumptions and results is given by Bühlmann and van de Geer in [vdGB09]. One can also cite van de Geer in [vdGBZ11] and discussions, who precises a chaining argument to perform rate, even in a non linear case.

If we assume that $(x_i, y_i)_{1 \leq i \leq n}$ arise from different subpopulations, we could work with finite mixture regression models. Indeed, the homogeneity assumption of the linear model is often inadequate and restrictive. This model was introduced by Städler et al., in [SBG10]. They assume that, for $i \in \{1, \ldots, n\}$, the observation y_i, conditionally to $X_i = x_i$, comes from a conditional density $s_{\xi^0}(.|x_i)$ which is a finite mixture of K Gaussian conditional densities with proportion vector $\boldsymbol{\pi}$, where

$$Y_i | X_i = x_i \sim s_{\xi^0}(y_i|x_i) = \sum_{k=1}^{K} \frac{\pi_k^0}{\sqrt{2\pi}\sigma_k^0} \exp\left(-\frac{(y_i - \beta_k^0 x_i)^2}{2(\sigma_k^0)^2} \right)$$

for some parameters $\xi^0 = (\pi_k^0, \beta_k^0, \sigma_k^0)_{1 \leq k \leq K}$. They extend the Lasso estimator by

$$\hat{s}^{\text{Lasso}}(\lambda) = \underset{s_\xi^K}{\text{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} |\sigma_k^{-1} [\beta_k]_j| \right\}, \qquad \lambda > 0 \tag{2.1}$$

For this estimator, they provide an ℓ_0-oracle inequality satisfied by $\hat{s}^{Lasso}(\lambda)$, according to the restricted eigenvalue condition also, and margin conditions, which lead to link the Kullback-Leibler loss function to the ℓ_2-norm of the parameters.
Another way to study this estimator is to look after the Lasso for its ℓ_1-regularization properties. For example, cite [MM11a], [Mey13], and [RT11]. Contrary to the ℓ_0-results, some ℓ_1-results are valid with no assumptions, neither on the Gram matrix, nor on the margin. This can be achieved due to the fact that they are looking for rate of convergence of order $1/\sqrt{n}$ rather than $1/n$. For finite mixture Gaussian regression models, we could cite Meynet in [Mey13] who give an ℓ_1-oracle inequality for the Lasso estimator (2.1).
In this chapter, we extend this result to finite mixture of multivariate Gaussian regression models. We will work with random multivariate variables $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$. As in [Mey13], we shall restrict to the fixed design case, that is to say non-random regressors. We observe $(x_i)_{1\leq i\leq n}$. Without any restriction, we could assume that the regressors $x_i \in [0, 1]^p$ for all $i \in \{1, \ldots, n\}$. Under only bounded parameters assumption, we provide a lower bound on the Lasso regularization parameter λ which guarantees such an oracle inequality.
This result is non-asymptotic: the number of observations is fixed, and the number p of covariates can grow. Remark that the number K of clusters in the mixture is supposed to be known. Our result is deduced from a finite mixture multivariate Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation. We establish the general theorem following the one of Meynet in [Mey13], which combines Vapnik's structural risk minimization method (see Vapnik in [Vap82]) and theory around model selection (see Le Pennec and Cohen in [CLP11] and Massart in [Mas07]). As in Massart and Meynet in [MM11a], our oracle inequality is deduced from this general theorem, the Lasso being viewed 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 model and the framework are introduced in Section 2.2. In Section 2.3, we state the main result of the chapter, which is an ℓ_1-oracle inequality satisfied by the Lasso in finite mixture of multivariate Gaussian regression models. Section 2.4 is devoted to the proof of this result and of the general theorem, deriving from two easier propositions. Those propositions are proved in Section 2.5, whereas details of lemma states in Section 2.6.

2.2 Notations and framework

2.2.1 Finite mixture regression model

We observe n independent couples $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{1\leq i\leq n} \in ([0, 1]^p \times \mathbb{R}^q)^n$, with $y_i \in \mathbb{R}^q$ a random observation, realization of variable Y_i, and $x_i \in [0, 1]^p$ fixed for all $i \in \{1, \ldots, n\}$. We assume that, conditionally to the x_is, the Y_is are independent identically distributed with conditional density $s_{\xi^0}(.|x_i)$ which is a finite mixture of K Gaussian regressions with unknown parameters ξ^0. In this chapter, K is fixed, then we do not precise it with unknown parameters. We will estimate the unknown conditional density by a finite mixture of K Gaussian regressions. Each subpopulation is then estimated by a multivariate linear model. Detail the conditional density.

For all $y \in \mathbb{R}^q$, for all $x \in [0,1]^p$,

$$s_\xi(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right) \tag{2.2}$$

$$\xi = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \Sigma_1, \ldots, \Sigma_K) \in \Xi = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right)$$

$$\Pi_K = \left\{(\pi_1, \ldots, \pi_K); \pi_k > 0 \text{ for } k \in \{1, \ldots, K\} \text{ and } \sum_{k=1}^{K} \pi_k = 1\right\}$$

$\mathbb{S}_q^{++}$ is the set of symmetric positive definite matrices on $\mathbb{R}^q$.

We want to estimate ξ^0 from the observations. For all $k \in \{1, \ldots, K\}, \beta_k$ is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k, whereas the π_ks are the mixture proportions. For $x \in [0,1]^p$, we define the parameter $\xi(x)$ of the conditional density $s_\xi(.|x)$ by

$$\xi(x) = (\pi_1, \ldots, \pi_K, \beta_1 x, \ldots, \beta_K x, \Sigma_1, \ldots, \Sigma_K) \in \mathbb{R}^K \times (\mathbb{R}^q)^K \times (\mathbb{S}_q^{++})^K.$$

For all $k \in \{1, \ldots, K\}$, for all $x \in [0,1]^p$, for all $z \in \{1, \ldots, q\}$, $[\beta_k x]_z = \sum_{j=1}^{p} [\beta_k]_{z,j}[x]_j$, and then $\beta_k x \in \mathbb{R}^q$ is the mean vector of the mixture component k for the conditional density $s_\xi(.|x)$.

2.2.2 Boundedness assumption on the mixture and component parameters

Denote, for a matrix A, $m(A)$ the modulus of the smallest eigenvalue of A, and $M(A)$ the modulus of the largest eigenvalue of A. We shall restrict our study to bounded parameters vector $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_K), \boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_K)$. Specifically, we assume that there exists deterministic positive constants $A_\beta, a_\Sigma, A_\Sigma, a_\pi$ such that ξ belongs to $\tilde{\Xi}$, with

$$\tilde{\Xi} = \Big\{\xi \in \Xi : \text{ for all } k \in \{1, \ldots, K\}, \max_{z\in\{1,\ldots,q\}} \sup_{x\in[0,1]^p} |[\beta_k x]_z| \le A_\beta, \\ a_\Sigma \le m(\Sigma_k^{-1}) \le M(\Sigma_k^{-1}) \le A_\Sigma, a_\pi \le \pi_k\Big\}. \tag{2.3}$$

Let S the set of conditional densities s_ξ,

$$S = \left\{s_\xi, \xi \in \tilde{\Xi}\right\}.$$

2.2.3 Maximum likelihood estimator and penalization

In a maximum likelihood approach, we consider the Kullback-Leibler information as the loss function, which is defined for two densities s and t by

$$\mathrm{KL}(s,t) = \begin{cases} \displaystyle\int_{\mathbb{R}^q} \log\left(\frac{s(y)}{t(y)}\right) s(y)dy \text{ if } sdy << tdy; \\ +\infty \text{ otherwise.} \end{cases} \tag{2.4}$$

In a regression framework, we have to adapt this definition to take into account the structure of conditional densities. For the fixed covariates $(x_1, \ldots, x_n)$, we consider the average loss function

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

Using the maximum likelihood approach, we want to estimate s_{ξ^0} by the conditional density s_ξ which maximizes the likelihood conditionally to $(x_i)_{1\leq i\leq n}$. Nevertheless, because we work with high-dimensional data, we have to regularize the maximum likelihood estimator. We consider the ℓ_1-regularization, and a generalization of the estimator associated, the Lasso estimator, which we define by

$$\hat{s}^{\text{Lasso}}(\lambda) := \underset{s_\xi\in S}{\text{argmin}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log(s_\xi(y_i|x_i)) + \lambda\sum_{k=1}^{K}\sum_{z=1}^{q}\sum_{j=1}^{p}|[\beta_k]_{z,j}|\right\};$$

where $\lambda > 0$ is a regularization parameter, for $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$.
We define also, for s_ξ defined as in (2.2), and with parameters $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$,

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q}|[\beta_k]_{z,j}|. \tag{2.5}$$

2.3 Oracle inequality

In this section, we provide an ℓ_1-oracle inequality satisfied by the Lasso estimator in finite mixture multivariate Gaussian regression models, which is the main result of this chapter.

Theorem 2.3.1. *We observe n couples $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in ([0,1]^p \times \mathbb{R}^q)^n$ coming from the conditional density s_{ξ^0}, where $\xi^0 \in \tilde{\Xi}$, where*

$$\tilde{\Xi} = \Bigg\{\xi \in \Xi : \text{ for all } k \in \{1, \ldots, K\}, \max_{z\in\{1,\ldots,q\}}\sup_{x\in[0,1]^p}|[\beta_k x]_z| \leq A_\beta, \\ a_\Sigma \leq m(\Sigma_k^{-1}) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k\Bigg\}.$$

Denote by $a \vee b = \max(a, b)$.
We define the Lasso estimator, denoted by $\hat{s}^{Lasso}(\lambda)$, for $\lambda \geq 0$, by

$$\hat{s}^{Lasso}(\lambda) = \underset{s_\xi\in S}{\text{argmin}}\left(-\frac{1}{n}\sum_{i=1}^{n}\log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi)\right); \tag{2.6}$$

with

$$S = \left\{s_\xi, \xi \in \tilde{\Xi}\right\}$$

and where, for $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$,

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q}|[\beta_k]_{z,j}|.$$

Then, if

$$\lambda \geq \kappa\left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right)\sqrt{\frac{K}{n}}\left(1 + q\log(n)\sqrt{K\log(2p+1)}\right)$$

with κ an absolute positive constant, the estimator (2.6) satisfies the following ℓ_1-oracle inequality.

$$\begin{aligned}
\mathrm{E}[\mathrm{KL}_n(s_{\xi^0}, \hat{s}^{Lasso}(\lambda))] \leq \quad & (1+\kappa^{-1}) \quad \inf_{s_\xi \in S} \left(\mathrm{KL}_n(s_{\xi^0}, s_\xi) + \lambda N_1^{[2]}(s_\xi) \right) + \lambda \\
+ \quad & \kappa' \sqrt{\tfrac{K}{n}} \quad \frac{e^{-\frac{1}{2}} \pi^{q/2} a_\pi}{A_\Sigma^{q/2}} \sqrt{2q} \\
+ \quad & \kappa' \sqrt{\tfrac{K}{n}} \quad \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1) A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \\
& \qquad \times K \left(1 + A_\beta + \frac{q}{a_\Sigma} \right)^2
\end{aligned}$$

where κ' is a positive constant.

This theorem provides information about the performance of the Lasso as an ℓ_1-regularization algorithm. If 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, which is the solution of the ℓ_1-penalized true risk minimization problem, up to an error term of order λ.
Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates and the size q of he response can grow with respect to n and can be much larger than n. The number K of clusters in the mixture is fixed.
There is no assumption neither on the Gram matrix, nor on the margin, which are classical assumptions for oracle inequality for the Lasso estimator. Moreover, this kind of assumptions involve unknown constants, whereas here, every constants are explicit. We could compare this result with the ℓ_0-oracle inequality established in [SBG10], which need those assumptions, and is therefore difficult to interpret. Nevertheless, they get faster rate, the error term in the oracle inequality being of order $1/n$ rather than $1/\sqrt{n}$.
The main assumption we make to establish the Theorem 2.3.1 is the boundedness of the parameters, which is also assumed in [SBG10]. It is needed, to tackle the problem of the unboundedness of the parameter space (see [MP04] for example).
Moreover, we let regressors to belong to $[0,1]^p$. Because we work with fixed covariates, they are finite. To simplify the reading, we choose to rescale $\mathbf{x}$ to get $||\mathbf{x}||_\infty \leq 1$. Nevertheless, if we not rescale the covariates, and the regularization parameter λ bound and the error term of the oracle inequality depend linearly of $||\mathbf{x}||_\infty$.
The regularization parameter λ bound is of order $(q^2+q)/\sqrt{n}\log(n)^2\sqrt{\log(2p+1)}$. For $q=1$, we recognize the same order, as regards to the sample size n and the number of covariates p, to the ℓ_1-oracle inequality in [Mey13].
Van de Geer, in [vdGBZ11], gives some tools to improve the bound of the regularization parameter to $\sqrt{\frac{\log(p)}{n}}$. Nevertheless, we have to control eigenvalues of the Gram matrix of some functions $(\psi_j(x_i))_{\substack{1\leq j\leq D\\ 1\leq i\leq n}}$, D being the number of parameters to estimate, where $\psi_j(x_i)$ satisfies

$$|\log(s_\xi(y_i|x_i)) - \log(s_{\tilde{\xi}}(y_i|x_i))| \leq \sum_{j=1}^{D} |\xi_j - \tilde{\xi}_j| \psi_j(x_i).$$

In our case of mixture of regression models, control eigenvalues of the Gram matrix of functions $(\psi_j(x_i))_{\substack{1\leq j\leq D\\ 1\leq i\leq n}}$ corresponds to make some assumptions, as REC, to avoid dimension reliance on n, K and p. Without this kind of assumptions, we could not guarantee that our bound is of order $\sqrt{\frac{\log(p)}{n}}$, because we could not guarantee that eigenvalues does not depend on dimensions. In

order to get a result with smaller assumptions, we do not use the chaining argument developed in [vdGBZ11]. Nevertheless, one can easily compute that, under restricted eigenvalue condition, we could perform the order of the regularization parameter to $\lambda \asymp \sqrt{\frac{\log(p)}{n}}\log(n)$.

2.4 Proof of the oracle inequality

2.4.1 Main propositions used in this proof

The first result we will prove is the next theorem, which is an ℓ_1-ball mixture multivariate regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation in the Gaussian framework.

Theorem 2.4.1. *We observe $(x_i, y_i)_{1\leq i\leq n}$ with unknown conditional Gaussian mixture density s_{ξ^0}.*
For all $m \in \mathbb{N}^*$, we consider the ℓ_1-ball $S_m = \{s_\xi \in S, N_1^{[2]}(s_\xi) \leq m\}$ for $S = \{s_\xi, \xi \in \tilde{\Xi}\}$, and $\tilde{\Xi}$ defined by

$$\tilde{\Xi} = \Bigg\{\xi \in \Xi : \text{ for all } k \in \{1,\dots,K\}, \max_{z\in\{1,\dots,q\}} \sup_{x\in[0,1]^p} |[\beta_k x]_z| \leq A_\beta, \\ a_\Sigma \leq m(||\Sigma_k^{-1}||) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k\Bigg\}.$$

For $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, let

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^K \sum_{j=1}^p \sum_{z=1}^q |[\beta_k]_{z,j}|.$$

Let $\hat{s}_m$ an η_m-log-likelihood minimizer in S_m, for $\eta_m \geq 0$:

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

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

$$\lambda \geq \kappa\left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right)\sqrt{\frac{K}{n}}\left(1 + q\log(n)\sqrt{K\log(2p+1)}\right)$$

for a constant κ. Then, if $\hat{m}$ is such that

$$-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_m(y_i|x_i)) + pen(m)\right) + \eta$$

for $\eta \geq 0$, the estimator $\hat{s}_{\hat{m}}$ satisfies

$$\begin{aligned}\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) \leq &(1+\kappa^{-1})\inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \eta_m\right) + \eta \\ &+ \kappa'\sqrt{\frac{K}{n}}\frac{e^{-\frac{1}{2}}\pi^{q/2}}{A_\Sigma^{q/2}}\sqrt{2qa_\pi} \\ &+ \kappa'\sqrt{\frac{K}{n}}K\left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + \frac{4(q+1)}{2}A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\sigma}\right)\right) \\ &\times\left(1 + A_\beta + \frac{q}{a_\Sigma}\right)^2;\end{aligned}$$

where κ' is a positive constant.

It is an ℓ_1-ball mixture regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation in the Gaussian framework. Its proof could be deduced from the two following propositions, which split the result if the variable Y is large enough or not.

Proposition 2.4.2. *We observe $(x_i, y_i)_{1\leq i\leq n}$, with unknown conditional density denoted by s_{ξ^0}. Let $M_n > 0$, and consider the event*

$$\mathcal{T} := \left\{ \max_{i\in\{1,\dots,n\}} \max_{z\in\{1,\dots,q\}} |[Y_i]_z| \leq M_n \right\}.$$

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

$$S_m = \{s_\xi \in S, N_1^{[2]}(s_\xi) \leq m\}$$

where $S = \{s_\xi, \xi \in \tilde{\Xi}\}$ and

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q} |[\beta_k]_{z,j}|$$

for $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$.
Let $\hat{s}_m$ an η_m-log-likelihood minimizer in S_m, for $\eta_m \geq 0$:

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

Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n| + A_\beta)^2 A_\Sigma^2, \frac{q(|M_n|+A_\beta)A_\Sigma}{2}\right)$. Assume that for all $m \in \mathbb{N}^*$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \kappa\frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\left(1 + 9q\log(n)\sqrt{K\log(2p+1)}\right)$$

for some absolute constant κ. Then, any estimate $\hat{s}_{\hat{m}}$ with $\hat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m)\right) + \eta$$

for $\eta \geq 0$, satisfies

$$\begin{aligned}\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) \leq &(1+\kappa^{-1})\inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \eta_m\right)\\ &+ \frac{\kappa' K^{3/2} q C_{M_n}}{\sqrt{n}}\left(1 + \left(A_\beta + \frac{q}{a_\Sigma}\right)^2\right);\end{aligned}$$

where κ' is an absolute positive constant.

Proposition 2.4.3. *Let $s_{\xi^0}, \mathcal{T}$ and $\hat{s}_{\hat{m}}$ defined as in the previous proposition. Then,*

$$\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}\sqrt{2Knqa_\pi}e^{-1/4(M_n^2 - 2M_nA_\beta)a_\Sigma}.$$

2.4.2 Notations

To prove those two propositions, and the theorem, begin with some notations.
For any measurable function $g : \mathbb{R}^q \mapsto \mathbb{R}$, we consider the empirical norm

$$g_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} g^2(y_i|x_i)};$$

its conditional expectation

$$\mathrm{E}(g(Y|X)) = \int_{\mathbb{R}^q} g(y|x)s_{\xi^0}(y|x)dy;$$

its empirical process

$$P_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(y_i|x_i);$$

and its normalized process

$$\nu_n(g) := P_n(g) - \mathrm{E}_X(P_n(g)) = \frac{1}{n}\sum_{i=1}^{n}\left[g(y_i|x_i) - \int_{\mathbb{R}^q} g(y|x_i)s_{\xi^0}(y|x_i)dy\right].$$

For all $m \in \mathbb{N}^*$, for all model S_m, we define F_m by

$$F_m = \left\{f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right), s_m \in S_m\right\}.$$

Let $\delta_{\mathrm{KL}} > 0$. For all $m \in \mathbb{N}^*$, let $\eta_m \geq 0$. There exist two functions, denoted by $\hat{s}_{\hat{m}}$ and $\bar{s}_m$, belonging to S_m, such that

$$\begin{aligned} P_n(-\log(\hat{s}_{\hat{m}})) &\leq \inf_{s_m \in S_m} P_n(-\log(s_m)) + \eta_m; \\ \mathrm{KL}_n(s_{\xi^0}, \bar{s}_m) &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \delta_{\mathrm{KL}}. \end{aligned} \tag{2.7}$$

Denote by $\hat{f}_m := -\log\left(\frac{\hat{s}_m}{s_{\xi^0}}\right)$ and $\bar{f}_m := -\log\left(\frac{\bar{s}_m}{s_{\xi^0}}\right)$. Let $\eta \geq 0$ and fix $m \in \mathbb{N}^*$. We define the set $M(m)$ by

$$M(m) = \left\{m' \in \mathbb{N}^* | P_n(-\log(\hat{s}_{m'})) + \mathrm{pen}(m') \leq P_n(-\log(\hat{s}_m)) + \mathrm{pen}(m) + \eta\right\}. \tag{2.8}$$

2.4.3 Proof of the Theorem 2.4.1 thanks to the Propositions 2.4.2 and 2.4.3

Let $M_n > 0$ and $\kappa \geq 36$. Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n| + A_\beta)^2 A_\Sigma^2, q(|M_n| + A_\beta)A_\Sigma/2\right)$. Assume that, for all $m \in \mathbb{N}^*$, $\mathrm{pen}(m) = \lambda m$, with

$$\lambda \geq \kappa C_{M_n}\sqrt{\frac{K}{n}}\left(1 + q\log(n)\sqrt{K\log(2p+1)}\right).$$

We derive from the two propositions that there exists κ' such that, if $\hat{m}$ satisfies

$$-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m)\right) + \eta;$$

then $\hat{s}_{\hat{m}}$ satisfies

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) =& \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) + \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \\
\leq& (1+\kappa^{-1}) \inf_{m\in\mathbb{N}^*} \left(\inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \eta_m \right) \\
&+ \kappa' \frac{C_{M_n}}{\sqrt{n}} K^{3/2} q \left(1 + (A_\beta + \frac{q}{a_\Sigma})^2 \right) + \eta \\
&+ \kappa' \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}} \sqrt{2Knqa_\pi} e^{-1/4(M_n^2 - 2M_nA_\beta)a_\Sigma}.
\end{aligned}$$

In order to optimize this equation with respect to M_n, we consider M_n the positive solution of the polynomial

$$\log(n) - \frac{1}{4}(X^2 - 2XA_\beta)a_\Sigma = 0;$$

we obtain $M_n = A_\beta + \sqrt{A_\beta^2 + \frac{4\log(n)}{a_\Sigma}}$ and $\sqrt{n}e^{-1/4(M_n^2-2M_nA_\beta)a_\Sigma} = \frac{1}{\sqrt{n}}$.
On the other hand,

$$\begin{aligned}
C_{M_n} &\leq \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left[1 + \frac{q+1}{2} A_\Sigma (M_n + A_\beta)^2 \right] \\
&\leq \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left[1 + 4(q+1)A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right].
\end{aligned}$$

We get

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) =& \quad \mathrm{E}(\mathrm{KL}_n \quad (s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) + \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \\
\leq& \quad (1+\kappa^{-1}) \quad \inf_{m\in\mathbb{N}^*} \left(\inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \eta_m \right) + \eta \\
&+\kappa' \sqrt{\tfrac{K}{n}} \quad \frac{e^{-\frac{1}{2}}\pi^{q/2}}{(qA_\Sigma)^{q/2}} \sqrt{2qa_\pi} \\
&+\kappa' \sqrt{\tfrac{K}{n}} \quad \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1)A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \\
&\qquad \times K \left(1 + \left(A_\beta + \frac{q}{a_\Sigma} \right)^2 \right).
\end{aligned}$$

2.4.4 Proof of the Theorem 2.3.1

We will show that there exists $\eta_m \geq 0$, and $\eta \geq 0$ such that $\hat{s}^{\mathrm{Lasso}}(\lambda)$ satisfies the hypothesis of the Theorem 2.4.1, which will lead to Theorem 2.3.1.
First, let show that there exists $m \in \mathbb{N}^*$ and $\eta_m \geq 0$ such that the Lasso estimator is an η_m-log-likelihood minimizer in S_m.
For all $\lambda \geq 0$, if $m_\lambda = \lceil N_1^{[2]}(\hat{s}(\lambda)) \rceil$,

$$\hat{s}^{\mathrm{Lasso}}(\lambda) = \underset{\substack{s\in S \\ N_1^{[2]}(s)\leq m_\lambda}}{\mathrm{argmin}} \left(-\frac{1}{n} \sum_{i=1}^n \log(s(y_i|x_i)) \right).$$

We could take $\eta_m = 0$.
Secondly, let show that there exists $\eta \geq 0$ such that

$$-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \text{pen}(m_\lambda) \leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_m(y_i|x_i)) + \text{pen}(m)\right) + \eta.$$

Taking $\text{pen}(m_\lambda) = \lambda m_\lambda$,

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \text{pen}(m_\lambda) &= -\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \lambda m_\lambda \\
&\leq -\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \lambda N_1^{[2]}(\hat{s}^{\text{Lasso}}(\lambda)) + \lambda \\
&\leq \inf_{s_\xi\in S}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi)\right\} + \lambda \\
&\leq \inf_{m\in\mathbb{N}^*}\inf_{s_\xi\in S_m}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi)\right\} + \lambda \\
&\leq \inf_{m\in\mathbb{N}^*}\left(\inf_{s_\xi\in S_m}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log(s_\xi(y_i|x_i))\right\} + \lambda m\right) + \lambda \\
&\leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{s}_m(y_i|x_i)) + \lambda m\right) + \lambda.
\end{aligned}$$

which is exactly the goal, with $\eta = \lambda$. Then, according to the Theorem 2.4.1, with $\hat{m} = m_\lambda$, and $\hat{s}_{\hat{m}} = \hat{s}^{\text{Lasso}}(\lambda)$, for

$$\lambda \geq \kappa\left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right)\sqrt{\frac{K}{n}}\left(1 + q\log(n)\sqrt{K\log(2p+1)}\right),$$

we get the oracle inequality.

2.5 Proof of the theorem according to $\mathcal{T}$ or $\mathcal{T}^c$

2.5.1 Proof of the Proposition 2.4.2

This proposition corresponds to the main theorem according to the event $\mathcal{T}$. To prove it, we need some preliminary results.
From our notations, reminded in Section 2.4.2, we have, for all $m \in \mathbb{N}^*$ for all $m' \in M(m)$,

$$\begin{aligned}
P_n(\hat{f}_{m'}) + \text{pen}(m') &\leq P_n(\hat{f}_m) + \text{pen}(m) + \eta \leq P_n(\bar{f}_m) + \text{pen}(m) + \eta_m + \eta; \\
\text{E}(P_n(\hat{f}_{m'})) + \text{pen}(m') &\leq \text{E}(P_n(\bar{f}_m)) + \text{pen}(m) + \eta_m + \eta + \nu_n(\bar{f}_m) - \nu_n(\hat{f}_{m'}); \\
\text{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \text{pen}(m') &\leq \inf_{s_m\in S_m}\text{KL}_n(s_{\xi^0}, s_m) + \delta_{\text{KL}} + \text{pen}(m) + \eta_m + \eta + \nu_n(\bar{f}_m) - \nu_n(\hat{f}_{m'});
\end{aligned} \tag{2.9}$$

thanks to the inequality (2.7).
The goal is to bound $-\nu_n(\hat{f}_{m'}) = \nu_n(-\hat{f}_{m'})$.
To control this term, we use the following lemma.

Lemme 2.5.1. *Let $M_n > 0$. Let*

$$\mathcal{T} = \left\{ \max_{i\in\{1,\dots,n\}} \left(\max_{z\in\{1,\dots,q\}} |[Y_i]_z| \right) \leq M_n \right\}.$$

Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n| + A_\beta)^2 A_\Sigma^2, \frac{q(|M_n|+A_\beta)A_\Sigma}{2}\right)$ *and*

$$\Delta_{m'} = m' \log(n)\sqrt{K\log(2p+1)} + 6\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right).$$

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

$$\sup_{f_{m'}\in\mathcal{F}_{m'}} |\nu_n(-f_{m'})| \leq \frac{4C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \sqrt{2}\sqrt{t}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\right)$$

Proof. Page 83 ▲

From (2.9), on the event $\mathcal{T}$, for all $m \in \mathbb{N}^*$, for all $m' \in M(m)$, for all $t > 0$, with probability greater than $1 - e^{-t}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \delta_{\mathrm{KL}} + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \sqrt{2}\sqrt{t}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\right) + \eta_m + \eta \\
&\leq \inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}t\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}},
\end{aligned}$$

the last inequality being true because $2ab \leq \frac{1}{\sqrt{K}}a^2 + \sqrt{K}b^2$. Let $z > 0$ such that $t = z + m + m'$. On the event $\mathcal{T}$, for all $m \in \mathbb{N}$, for all $m' \in M(m)$, with probability greater than $1 - e^{-(z+m+m')}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2\right) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(\sqrt{K}(z + m + m')\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) - \nu_n(\bar{f}_m) &\leq \inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + 4\frac{C_{M_n}}{\sqrt{n}}\sqrt{K}m \\
&\quad + \left[\frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}(m' + 9q\Delta_{m'}) - \mathrm{pen}(m')\right] \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

Let $\kappa \geq 1$, and assume that $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\left(1 + 9q\log(n)\sqrt{K\log(2p+1)}\right)$$

Then, as

$$\Delta_{m'} = m' \log(n)\sqrt{K\log(2p+1)} + 6\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right),$$

with

$$\kappa^{-1} = \frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\frac{1}{\lambda} \leq \frac{1}{1 + 9q\log(n)\sqrt{K\log(2p+1)}},$$

we get that

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) - \nu_n(\bar{f}_m) &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(54\sqrt{K}q\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right) + \sqrt{K}z\right) \\
&\quad + \eta + \delta_{\mathrm{KL}} + \eta_m \\
&\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(27K^{3/2} + \frac{27+1/2}{\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

Let $\hat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m)\right) + \eta;$$

and $M(m) = \{m' \in \mathbb{N}^* | P_n(-\log(\hat{s}_{m'})) + \mathrm{pen}(m') \leq P_n(-\log(\hat{s}_m)) + \mathrm{pen}(m) + \eta\}$. By definition, $\hat{m} \in M(m)$. Because for all $m \in \mathbb{N}^*$, for all $m' \in M(m)$,

$$1 - \sum_{\substack{m \in \mathbb{N}^* \\ m' \in M(m)}} e^{-(z+m+m')} \geq 1 - e^{-z} \sum_{(m,m') \in (\mathbb{N}^*)^2} e^{-m-m'} \geq 1 - e^{-z},$$

we could sum up over all models.
On the event $\mathcal{T}$, for all $z > 0$, with probability greater than $1 - e^{-z}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}}) - \nu_n(\bar{f}_m) &\leq \inf_{m \in \mathbb{N}^*}\left(\inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(27K^{3/2} + \frac{55q}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) \\
&\quad + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

By integrating over $z > 0$, and noticing that $\mathrm{E}(\nu_n(\bar{f}_m)) = 0$ and that δ_{KL} can be chosen arbitrary

small, we get

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) &\leq \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(27K^{3/2} + \frac{q}{\sqrt{K}}\frac{55}{2}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}\right) + \eta \\
&\leq \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{332K^{3/2}qC_{M_n}}{\sqrt{n}}\left(1 + \left(A_\beta + \frac{q}{a_\Sigma}\right)^2\right) + \eta.
\end{aligned}$$

2.5.2 Proof of the Proposition 2.4.3

We want an upper bound of $\mathrm{E}\left(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}\right)$. Thanks to the Cauchy-Schwarz inequality,

$$\mathrm{E}\left(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}\right) \leq \sqrt{\mathrm{E}(\mathrm{KL}_n^2(s_{\xi^0}, \hat{s}_{\hat{m}}))}\sqrt{P(\mathcal{T}^c)}.$$

However, for all $s_\xi \in S$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, s_\xi) &= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^q} \log\left(\frac{s_{\xi^0}(y|x_i)}{s_\xi(y|x_i)}\right) s_{\xi^0}(y|x_i)dy \\
&= \frac{1}{n}\sum_{i=1}^n \left(\int_{\mathbb{R}^q} \log(s_{\xi^0}(y|x_i))s_{\xi^0}(y|x_i)dy - \int_{\mathbb{R}^q} \log(s_\xi(y|x_i))s_{\xi^0}(y|x_i)dy\right) \\
&\leq -\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^q} \log(s_\xi(y|x_i))s_{\xi^0}(y|x_i)dy.
\end{aligned}$$

Because parameters are assumed to be bounded, according to the set (2.3), we get, with $(\boldsymbol{\beta^0}, \boldsymbol{\Sigma^0}, \boldsymbol{\pi^0})$ the parameters of s_{ξ^0} and $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ the parameters of s_ξ,

$$\begin{aligned}
\log(s_\xi(y|x_i))s_{\xi^0}(y|x_i) &= \log\left(\sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}}\exp\left(-\frac{(y-\beta_k x_i)^t\Sigma_k^{-1}(y-\beta_k x_i)}{2}\right)\right) \\
&\quad \times \sum_{k=1}^K \frac{\pi_k^0}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k^0)}}\exp\left(-\frac{(y-\beta_k^0 x_i)^t(\Sigma_k^0)^{-1}(y-\beta_k^0 x_i)}{2}\right) \\
&\geq \log\left(K\frac{a_\pi\sqrt{\det(\Sigma_1^{-1})}}{(2\pi)^{q/2}}\exp\left(-(y^t\Sigma_1^{-1}y + x_i^t\beta_1^t\Sigma_1^{-1}\beta_1 x_i)\right)\right) \\
&\quad \times K\frac{a_\pi\sqrt{\det((\Sigma_1^0)^{-1})}}{(2\pi)^{q/2}}\exp\left(-(y^t\Sigma_1^{-1}y + x_i^t\beta_1^t\Sigma_1^{-1}\beta_1 x_i)\right) \\
&\geq \log\left(K\frac{a_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\exp\left(-(y^ty + A_\beta^2)A_\Sigma\right)\right) \\
&\quad \times K\frac{a_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\exp\left(-(y^ty + A_\beta^2)A_\Sigma\right).
\end{aligned}$$

Indeed, for $u \in \mathbb{R}^q$, if we use the eigenvalue decomposition of $\Sigma = P^t DP$,

$$\begin{aligned}|u^t \Sigma u| &= |u^t P^t DPu| \leq ||Pu||_2 ||DPU||_2 \leq M(D)||Pu||_2^2 \\ &\leq M(D)||u||_2^2 \leq A_\Sigma ||u||_2^2.\end{aligned}$$

To recognize the expectation of a Gaussian standardized variables, we put $u = \sqrt{2A_\Sigma}y$:

$$\begin{aligned}\mathrm{KL}(s_{\xi^0}(.|x_i), s_\xi(.|x_i)) &\leq -\frac{Ka_\pi e^{-A_\beta^2 A_\Sigma} a_\Sigma^{q/2}}{(2A_\Sigma)^{q/2}} \int_{\mathbb{R}^q} \left[\log\left(\frac{Ka_\Sigma^{q/2} a_\pi}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{u^t u}{2}\right] \frac{e^{\frac{-u^t u}{2}}}{(2\pi)^{q/2}} du \\ &\leq -\frac{a_\Sigma^{q/2} Ka_\pi e^{-A_\beta^2 A_\Sigma}}{(2A_\Sigma)^{q/2}} \mathrm{E}\left[\log\left(\frac{Ka_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{U^2}{2}\right] \\ &\leq -\frac{Ka_\Sigma^{q/2} a_\pi e^{-A_\beta^2 A_\Sigma}}{(2A_\Sigma)^{q/2}} \left[\log\left(\frac{Ka_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{1}{2}\right] \\ &\leq -\frac{Ka_\Sigma^{q/2} a_\pi e^{-A_\beta^2 A_\Sigma - 1/2}}{(2\pi)^{q/2} A_\Sigma^{q/2}} e^{1/2} \pi^{q/2} \log\left(\frac{Ka_\pi e^{-A_\beta^2 A_\Sigma - 1/2} a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) \\ &\leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}};\end{aligned}$$

where $U \sim \mathcal{N}_q(0,1)$. We have used that for all $t \in \mathbb{R}$, $t\log(t) \geq -e^{-1}$. Then, we get, for all $s_\xi \in S$,

$$\mathrm{KL}_n(s_{\xi^0}, s_\xi) \leq \frac{1}{n}\sum_{i=1}^n \mathrm{KL}(s_{\xi^0}(.|x_i), s_\xi(.|x_i)) \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}.$$

As it is true for all $s_\xi \in S$, it is true for $\hat{s}_{\hat{m}}$, then

$$\sqrt{\mathrm{E}(\mathrm{KL}_n^2(s_{\xi^0}, \hat{s}_{\hat{m}}))} \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}.$$

For the last step, we need to bound $P(\mathcal{T}^c)$.

$$P(\mathcal{T}^c) = \mathrm{E}(\mathbb{1}_{\mathcal{T}^c}) = \mathrm{E}(\mathrm{E}_X(\mathbb{1}_{\mathcal{T}^c})) = \mathrm{E}(P_X(\mathcal{T}^c)) \leq \mathrm{E}\left(\sum_{i=1}^n P_X(||Y_i||_\infty > M_n)\right).$$

Nevertheless, let $Y_x \sim \sum_{k=1}^K \pi_k \mathcal{N}_q(\beta_k x, \Sigma_k)$, then,

$$\begin{aligned}P(||Y_x||_\infty > M_n) &= \int_{\mathbb{R}^q} \mathbb{1}_{\{||Y_x||_\infty \geq M_n\}} \sum_{k=1}^K \pi_k \frac{1}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}} e^{\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right)} dy \\ &= \sum_{k=1}^K \pi_k \int_{\mathbb{R}^q} \mathbb{1}_{\{||Y_x||_\infty \geq M_n\}} \frac{1}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}} e^{\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right)} dy \\ &= \sum_{k=1}^K \pi_k P_X(||Y_x^k||_\infty > M_n) \leq \sum_{k=1}^K \sum_{z=1}^q \pi_k P_X(|[Y_x^k]_z| > M_n)\end{aligned}$$

with $Y_x^k \sim \mathcal{N}(\beta_k x, \Sigma_k)$ and $[Y_x^k]_z \sim \mathcal{N}([\beta_k x]_z, [\Sigma_k]_{z,z})$.

We need to control $P_X(|[Y_x^k]_z| > M_n)$, for all $z \in \{1, \dots, q\}$.

$$\begin{aligned}
P_X(|[Y_x^k]_z| > M_n) &= P_X([Y_x^k]_z > M_n) + P_X([Y_{x,k}]_z < -M_n) \\
&= P_X\left(U > \frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) + P_X\left(U < \frac{-M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) \\
&= P_X\left(U > \frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) + P_X\left(U > \frac{M_n + [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) \\
&\leq e^{-\frac{1}{2}\left(\frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} + e^{-\frac{1}{2}\left(\frac{M_n + [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} \\
&\leq 2e^{-\frac{1}{2}\left(\frac{M_n - |[\beta_k x]_z|}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} \\
&\leq 2e^{-\frac{1}{2}\frac{M_n^2 - 2M_n|[\beta_k x]_z| + |[\beta_k x]_z|^2}{[\Sigma_k]_{z,z}}}.
\end{aligned}$$

where $U \sim \mathcal{N}(0,1)$. Then,

$$P(||Y_x||_\infty > M_n) \leq 2Kqe^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma},$$

and we get $P(\mathcal{T}^c) \leq \mathrm{E}\left(\sum_{i=1}^n 2Kqa_\pi e^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma}\right) \leq 2Kna_\pi qe^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma}$. We have obtained the wanted bound for $\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c})$.

2.6 Some details

2.6.1 Proof of the Lemma 2.5.1

First, give some tools to prove the Lemma 2.5.1.
We define $||g||_n = \sqrt{\frac{1}{n}\sum_{i=1}^n g^2(y_i|x_i)}$ for any measurable function g.
Let $m \in \mathbb{N}^*$. We have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left| \frac{1}{n}\sum_{i=1}^n (f_m(y_i|x_i) - \mathrm{E}(f_m(Y_i|x_i))) \right|.$$

To control the deviation of such a quantity, we shall combine concentration with symmetrization arguments. We first use the following concentration inequality which can be found in [BLM13].

Lemme 2.6.1. *Let $(Z_1, \dots, Z_n)$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Assume that there exists R_n a non-random constant such that $\sup_{\gamma \in \Gamma} ||\gamma||_n \leq R_n$. Then, for all $t > 0$,*

$$P\left(\sup_{\gamma \in \Gamma}\left|\frac{1}{n}\sum_{i=1}^n \gamma(Z_i) - \mathrm{E}(\gamma(Z_i))\right| > \mathrm{E}\left[\sup_{\gamma \in \Gamma}\left|\frac{1}{n}\sum_{i=1}^n \gamma(Z_i) - \mathrm{E}(\gamma(Z_i))\right|\right] + 2\sqrt{2}R_n\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

Proof. See [BLM13]. ▲

Then, we propose to bound $\mathrm{E}\left[\sup_{\gamma \in \Gamma}\left|\frac{1}{n}\sum_{i=1}^n \gamma(Z_i) - \mathrm{E}(\gamma(Z_i))\right|\right]$ thanks to the following symmetrization argument. The proof of this result can be found in [vdVW96].

Lemme 2.6.2. *Let $(Z_1, \dots, Z_n)$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Let $(\epsilon_1, \dots, \epsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Then,*

$$\mathrm{E}\left[\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i) - \mathrm{E}(\gamma(Z_i))\right|\right] \leq 2\,\mathrm{E}\left[\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\gamma(Z_i)\right|\right].$$

Proof. See [vdVW96]. ▲

Then, we have to control $\mathrm{E}(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\gamma(Z_i)\right|)$.

Lemme 2.6.3. *Let $(Z_1, \dots, Z_n)$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Let $(\epsilon_1, \dots, \epsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Define R_n a non-random constant such that*

$$\sup_{\gamma\in\Gamma}||\gamma||_n \leq R_n.$$

Then, for all $S \in \mathbb{N}^*$,

$$\mathrm{E}\left[\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\gamma(Z_i)\right|\right] \leq R_n\left(\frac{6}{\sqrt{n}}\sum_{s=1}^{S}2^{-s}\left(\sqrt{\log(1+N(2^{-s}R_n,\Gamma,||.||_n))}+2^{-S}\right)\right)$$

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

Proof. See [Mas07]. ▲

In our case, we get the following lemma.

Lemme 2.6.4. *Let $m \in \mathbb{N}^*$. Consider $(\epsilon_1, \dots, \epsilon_n)$ a Rademacher sequence independent of $(Y_1, \dots, Y_n)$. Then, on the event T,*

$$\mathrm{E}\left(\sup_{f_m\in F_m}\left|\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right) \leq 18\sqrt{K}\frac{C_{M_n}q}{\sqrt{n}}\Delta_m;$$

where $\Delta_m := m\log(n)\sqrt{K\log(2p+1)} + 6(1+K(A_\beta + \frac{q}{a_\Sigma}))$.

Proof. Let $m \in \mathbb{N}^*$. According to Lemma 2.6.5, we get that on the event $\mathcal{T}$,

$$\sup_{f_m\in F_m}||f_m||_n \leq R_n := 2C_{M_n}(1+K(A_\beta + \frac{q}{a_\Sigma})).$$

Besides, on the event $\mathcal{T}$, for all $S \in \mathbb{N}^*$,

$$\begin{aligned}
&\sum_{s=1}^{S} 2^{-s}\sqrt{\log[1+N(2^{-s}R_n, F_m, ||.||_n)]} \leq \sum_{s=1}^{S} 2^{-s}\sqrt{\log(2N(2^{-s}R_n, F_m, ||.||_n))} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left(\sqrt{\log(2)}+\sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n}\right) \\
&+\sum_{s=1}^{S} 2^{-s}\sqrt{K\log\left(1+\frac{2^{s+3}C_{M_n}qK}{R_n a_\Sigma}\right)\left(1+\frac{2^{s+3}C_{M_n}}{R_n}\right)} \text{ according to Lemma 4.15} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left(\sqrt{\log(2)}+\sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n}\right) \\
&+\sum_{s=1}^{S} 2^{-s}\sqrt{K\log\left(1+2^{s+3}\frac{C_{M_n}}{R_n}\max(1, qK/a_\Sigma)\right)^2} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left[\sqrt{\log(2)}+\sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n}+\sqrt{2(s+3)K\log(2)q/a_\Sigma}\right] \\
\leq &\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)}+\sqrt{\log(2)}\left(1+\frac{\sqrt{q}}{a_\Sigma}\left(\sqrt{6K}+2\sum_{s=1}^{S}2^{-s}\sqrt{s}\right)\right) \\
\leq &\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)}+\sqrt{\log(2)}\left(1+\frac{\sqrt{q}}{a_\Sigma}\sqrt{6K}+\sqrt{q}\sqrt{K}\frac{\sqrt{2e}}{2-\sqrt{e}}\right)
\end{aligned}$$

because $2^{-s}\sqrt{s} \leq \left(\frac{\sqrt{e}}{2}\right)^s$ for all $s \in \mathbb{N}^*$. Then, thanks to the Lemma 2.6.3,

$$\begin{aligned}
\mathrm{E}\left(\sup_{f_m \in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right) \leq \quad & R_n \quad \left(\frac{6}{\sqrt{n}}\sum_{s=1}^{S}2^{-s}\sqrt{\log[1+N(2^{-s}R_n, F_m, ||.||_n)]}+2^{-S}\right) \\
\leq \quad & R_n \quad \left[\frac{6}{\sqrt{n}}\left(\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)}\right.\right. \\
& \left.\left.+\sqrt{\log(2)}\left(1+\frac{q}{a_\Sigma}\sqrt{6K}+\frac{q}{a_\Sigma}\sqrt{K}\frac{2e}{2-\sqrt{e}}\right)\right)+2^{-S}\right].
\end{aligned}$$

Taking $S = \frac{\log(n)}{\log(2)}$ to obtain the same order in the both terms depending on S, we could deduce

that

$$
\begin{aligned}
&\mathrm{E}\left(\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right)\\
\leq&\frac{12C_{M_n}Kmq}{\sqrt{n}}\sqrt{\log(2p+1)}\frac{\log(n)}{\log(2)}\\
&+2C_{M_n}\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right)\left[\frac{\sqrt{\log(2)}}{\sqrt{n}}\left(1+\sqrt{6K}+\frac{\sqrt{2e}}{2-\sqrt{2e}}\right)+\frac{1}{n}\right]\\
\leq&\frac{18C_{M_n}Kmq}{\sqrt{n}}\sqrt{\log(2p+1)}\log(n)\\
&+2\frac{\sqrt{K}}{\sqrt{n}}C_{M_n}\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right)\left[\sqrt{\log(2)}\left(1+\sqrt{6}+\frac{\sqrt{2e}}{2-\sqrt{2e}}\right)+1\right]\\
\leq&18\frac{\sqrt{K}}{\sqrt{n}}C_{M_n}\left[mq\sqrt{K\log(2p+1)}\log(n)+6\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right)\right].
\end{aligned}
$$

It completes the proof. ▲

We are now able to prove the Lemma 2.5.1.

$$
\begin{aligned}
\sup_{f_m\in F_m}|\nu_n(-f_m)| &= \sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}(f_m(y_i|x_i)-\mathrm{E}_X(f_m(Y_i|x_i)))\right|\\
&\leq \mathrm{E}\left(\sup_{f_m\in F_m}\left|\sum_{i=1}^{n}f_m(Y_i|x_i)-\mathrm{E}(f_m(Y_i|x_i))\right|\right)+2\sqrt{2}R_n\sqrt{\frac{t}{n}}\\
&\text{with probability greater than } 1-e^{-t} \text{ and where } R_n\\
&\text{is a constant computed from the Lemma 2.6.5}\\
&\leq 2\,\mathrm{E}\left(\sup_{f_m\in F_m}\left|\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right)+2\sqrt{2}R_n\sqrt{\frac{t}{n}}\\
&\text{with } \epsilon_i \text{ a Rademacher sequence,}\\
&\text{independent of } Z_i\\
&\leq 2\left(18\sqrt{K}\frac{C_{M_n}q}{\sqrt{n}}\Delta_m\right)+2\sqrt{2}R_n\sqrt{\frac{t}{n}}\\
&\leq 4C_{M_n}\left(9\frac{\sqrt{K}q}{\sqrt{n}}\Delta_m+\sqrt{2}\sqrt{\frac{t}{n}}\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right)\right).
\end{aligned}
$$

2.6.2 Lemma 2.6.5 and Lemma 4.15

Lemme 2.6.5. *On the event*

$$
\mathcal{T}=\left\{\max_{i\in\{1,\dots,n\}}\max_{z\in\{1,\dots,q\}}|[Y_i]_z|\leq M_n\right\},
$$

for all $m\in\mathbb{N}^*$,

$$
\sup_{f_m\in F_m}||f_m||_n\leq 2C_{M_n}\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right):=R_n.
$$

Proof. Let $m \in \mathbb{N}^*$. Because $f_m \in \mathcal{F}_m = \left\{ f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right), s_m \in S_m \right\}$, there exists $s_m \in S_m$ such that $f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right)$. For all $x \in [0,1]^p$, denote $\xi(x) = (\boldsymbol{\pi}, \beta_1 x, \ldots, \beta_K x, \boldsymbol{\Sigma})$ the parameters of $s_m(.|x)$. For all $i \in \{1, \ldots, n\}$,

$$\begin{aligned}|f_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} &= |\log(s_m(y_i|x_i)) - \log(s_{\xi^0}(y_i|x_i))|\mathbb{1}_{\mathcal{T}} \\ &\leq \sup_{x\in[0,1]^p} \sup_{\xi\in\Xi} \left| \frac{\partial \log(s_\xi(y_i|x))}{\partial \xi} \right| ||\xi(x_i) - \xi^0(x_i)||_1 \mathbb{1}_{\mathcal{T}},\end{aligned}$$

thanks to the Taylor formula. Then, we need an upper bound of the partial derivate.
For all $x \in [0,1]^p$, for all $y \in \mathbb{R}^q$, we could write

$$\log(s_\xi(y|x)) = \log\left(\sum_{k=1}^K h_k(x,y)\right)$$

where, for all $k \in \{1, \ldots, K\}$,

$$\begin{aligned}h_k(x,y) =& \frac{\pi_k}{(2\pi)^{q/2}\det\Sigma_k} \\ &\times \exp\left[-\frac{1}{2}\left(\sum_{z_2=1}^q \left(\sum_{z_1=1}^q y_{z_1} - \sum_{j=1}^p x_j[\beta_k]_{z_1,j}\right)[\Sigma_k]^{-1}_{z_1,z_2}\right)\left(y_{z_2} - \sum_{j=1}^p [\beta_k]_{z_2,j}x_j\right)\right].\end{aligned}$$

Then, for all $l \in \{1, \ldots, K\}$, for all $z_1 \in \{1, \ldots, q\}$, for all $z_2 \in \{1, \ldots, q\}$, for all $y \in \mathbb{R}^q$, for all $x \in [0,1]^p$,

$$\begin{aligned}\left|\frac{\partial \log(s_\xi(y|x))}{\partial([\beta_l x]_{z_1})}\right| =& \left|\frac{h_l(x,y)}{\sum_{k=1}^K h_k(x,y)}\right| \left(-\frac{1}{2}\sum_{z_2=1}^q [\Sigma_l]^{-1}_{z_1,z_2}([\beta_l x]_{z_2} - y_{z_2})\right) \leq \frac{q(|y| + A_\beta)A_\Sigma}{2}; \\ \left|\frac{\partial \log(s_\xi(y|x))}{\partial([\Sigma_l]_{z_1,z_2})}\right| =& \frac{1}{\sum_{k=1}^K h_k(x,y)} \\ &\times \left|\frac{-h_l \mathrm{Cof}_{z_1,z_2}(\Sigma_l)}{\det(\Sigma_l)} - \frac{h_l(x,y)(y_{z_1} - [\beta_l x]_{z_1})(y_{z_2} - [\beta_l x]_{z_2})[\Sigma_l]^{-2}_{z_1,z_2}}{2}\right| \\ \leq& \left|\frac{-\mathrm{Cof}_{z_1,z_2}(\Sigma_l)}{\det(\Sigma_l)} + \frac{(y_{z_1} - [\beta_l x]_{z_1})(y_{z_2} - [\beta_l x]_{z_2})[\Sigma_l]^{-2}_{z_1,z_2}}{2}\right| \\ \leq& A_\Sigma + \frac{1}{2}(|y| + A_\beta)^2 A_\Sigma^2,\end{aligned}$$

where $\mathrm{Cof}_{z_1,z_2}(\Sigma_k)$ is the (z_1, z_2)-cofactor of Σ_k. We also have, for all $l \in \{1, \ldots, K\}$, for all $x \in [0,1]^p$, for all $y \in \mathbb{R}^q$,

$$\left|\frac{\partial \log(s_\xi(y,x))}{\partial \pi_l}\right| = \left|\frac{h_l(x,y)}{\pi_l \sum_{k=1}^K h_k(x,y)}\right| \leq \frac{1}{a_\pi}.$$

Thus, for all $y \in \mathbb{R}^q$,

$$\sup_{x\in[0,1]^p} \sup_{\xi\in\tilde{\Xi}} \left|\frac{\partial \log(s_\xi(y|x))}{\partial \xi}\right| \leq \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|y| + A_\beta)^2 A_\Sigma^2, \frac{q(|y| + A_\beta)A_\Sigma}{2}\right) = C_y.$$

We have $C_y \leq \left(A_\Sigma \wedge \frac{1}{a_\pi}\right)\left[1 + \frac{q+1}{2}A_\Sigma(|y| + A_\beta)^2\right]$. For all $m \in \mathbb{N}^*$,

$$
\begin{aligned}
|f_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} &\leq C_{y_i}||\xi(x_i) - \xi^0(x_i)||_1 \mathbb{1}_{\mathcal{T}} \\
&\leq C_{M_n} \sum_{k=1}^{K} (||\beta_k x_i - \beta_k^0 x_i||_1 + ||\Sigma_k - \Sigma_k^0||_1 + |\pi_k - \pi_k^0|).
\end{aligned}
$$

Since f_m and f_m^0 belong to $\tilde{\Xi}$, we obtain

$$
|f_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}(KA_\beta + K\frac{q}{a_\Sigma} + 1)
$$

and then

$$
\sup_{f_m \in F_m} ||f_m||_n \mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}(KA_\beta + K\frac{q}{a_\Sigma} + 1).
$$

▲

For the next results, we need the following lemma, proved in [Mey13].

Lemme 2.6.6. *Let $\delta > 0$ and $(A_{i,j})_{\substack{i\in\{1,\dots,n\}\\ j\in\{1,\dots,p\}}} \in [0,1]^{n\times p}$. There exists a family B of $(2p+1)^{1/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $\mu \in \mathbb{R}^p$ in the ℓ_1-ball, there exists $\mu' \in B$ such that*

$$
\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}(\mu_j - \mu'_j)A_{i,j}\right)^2 \leq \delta^2.
$$

Proof. See [Mey13]. ▲

With this lemma, we can prove the following one:

Lemme 2.6.7. *Let $\delta > 0$ and $m \in \mathbb{N}^*$. On the event $\mathcal{T}$, we have the upper bound of the δ-packing number of the set of functions F_m equipped with the metric induced by the norm $||.||_n$:*

$$
N(\delta, F_m, ||.||_n) \leq (2p+1)^{4C_{M_n}^2 K^2 q^2 m^2/\delta^2}\left(1 + \frac{8C_{M_n}qK}{a_\Sigma \delta}\right)^K \left(1 + \frac{8C_{M_n}}{\delta}\right)^K.
$$

Proof. Let $m \in \mathbb{N}^*$ and $f_m \in F_m$. There exists $s_m \in S_m$ such that $f_m = -\log(s_m/s_{\xi^0})$. Introduce s'_m in S and put $f'_m = -\log(s'_m/s_{\xi^0})$. Denote by $(\beta_k, \Sigma_k, \pi_k)_{1\leq k\leq K}$ and $(\beta'_k, \Sigma'_k, \pi'_k)_{1\leq k\leq K}$ the parameters of the densities s_m and s'_m respectively. First, applying Taylor's inequality, on the event

$$
\mathcal{T} = \left\{\max_{i\in\{1,\dots,n\}} \max_{z\in\{1,\dots,q\}} |[Y_i]_z| \leq M_n\right\},
$$

we get, for all $i \in \{1,\dots,n\}$,

$$
\begin{aligned}
|f_m(y_i|x_i) - f'_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} &= |\log(s_m(y_i|x_i)) - \log(s'_m(y_i|x_i))|\mathbb{1}_{\mathcal{T}} \\
&\leq \sup_{x\in[0,1]^p} \sup_{\xi\in\tilde{\Xi}} \left|\frac{\partial \log(s_\xi(y_i|x))}{\partial \xi}\right| ||\xi(x_i) - \xi'(x_i)||_1 \mathbb{1}_{\mathcal{T}} \\
&\leq C_{M_n} \sum_{k=1}^{K}\left(\sum_{z=1}^{q}\left|[\beta_k x_i]_z - [\beta'_k x_i]_z\right| + ||\Sigma_k - \Sigma'_k||_1 + |\pi_k - \pi'_k|\right).
\end{aligned}
$$

Thanks to the Cauchy-Schwarz inequality, we get that

$$(f_m(y_i|x_i) - f'_m(y_i|x_i))^2 \mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}^2 \left[\left(\sum_{k=1}^{K} \sum_{z=1}^{q} \left| \beta_k x_i - \beta'_k x_i \right| \right)^2 + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right]$$

$$\leq 2C_{M_n}^2 \left[Kq \sum_{k=1}^{K} \sum_{z=1}^{q} \left(\sum_{j=1}^{p} [\beta_k]_{z,j} [x_i]_j - \sum_{j=1}^{p} [\beta'_k]_{z,j} [x_i]_j \right)^2 + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right],$$

and

$$||f_m - f'_m||_n^2 \mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}^2 \left[Kq \sum_{k=1}^{K} \sum_{z=1}^{q} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} [\beta_k]_{z,j} [x_i]_j - \sum_{j=1}^{p} [\beta'_k]_{z,j} [x_i]_j \right)^2 \right.$$
$$\left. + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right].$$

Denote by

$$a = Kq \sum_{k=1}^{K} \sum_{z=1}^{q} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} [\beta_k]_{z,j} [x_i]_j - \sum_{j=1}^{p} [\beta'_k]_{z,j} [x_i]_j \right)^2.$$

Then, for all $\delta > 0$, if

$$a \leq \delta^2/(4C_{M_n}^2)$$
$$||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 \leq \delta/(4C_{M_n})$$
$$||\boldsymbol{\pi} - \boldsymbol{\pi'}|| \leq \delta/(4C_{M_n})$$

then $||f_m - f'_m||_n^2 \leq \delta^2$. To bound a, we write

$$a = Kqm^2 \sum_{=1}^{K} \sum_{z=1}^{q} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{[\beta_k]_{z,j}}{m} [x_i]_j - \sum_{j=1}^{p} \frac{[\beta'_k]_{z,j}}{m} [x_i]_j \right)^2$$

and we apply Lemma 2.6.6 to $[\beta_k]_{z,.}/m$ for all $k \in \{1, \ldots, K\}$, and for all $z \in \{1, \ldots, q\}$. Since $s_m \in S_m$, we have $\sum_{z=1}^{q} \sum_{j=1}^{p} \left| \frac{[\beta_k]_{z,j}}{m} \right| \leq 1$, thus there exists a family $\mathcal{B}$ of $(2p+1)^{4C_{M_n}^2 q^2 K^2 m^2/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $k \in \{1, \ldots, K\}$, for all $z \in \{1, \ldots, q\}$, for all $[\beta_k]_{z,.}$, there exists $[\beta'_k]_{z,.} \in \mathcal{B}$ such that $a \leq \delta^2/(4C_{M_n}^2)$. Moreover, since $||\boldsymbol{\Sigma}||_1 \leq \frac{qK}{a_\Sigma}$ and $||\boldsymbol{\pi}||_1 \leq 1$, we get that, on the event $\mathcal{T}$,

$$N(\delta, F_m, ||.||_n) \leq \text{card}(\mathcal{B}) N\left(\frac{\delta}{4C_{M_n}}, B_1^K(\frac{qK}{A_\Sigma}), ||.||_1 \right) N\left(\frac{\delta}{4C_{M_n}}, B_1^K(1), ||.||_1 \right)$$
$$\leq (2p+1)^{4C_{M_n}^2 q^2 K^2 m^2/\delta^2} \left(1 + \frac{8C_{M_n} qK}{a_\Sigma \delta} \right)^K \left(1 + \frac{8C_{M_n}}{\delta} \right)^K$$

▲

Chapter 3

An oracle inequality for the Lasso-MLE procedure

Contents

2.1 Introduction **69**
2.2 Notations and framework **70**
2.2.1 Finite mixture regression model 70
2.2.2 Boundedness assumption on the mixture and component parameters . . 71
2.2.3 Maximum likelihood estimator and penalization 71
2.3 Oracle inequality **72**
2.4 Proof of the oracle inequality **74**
2.4.1 Main propositions used in this proof 74
2.4.2 Notations 76
2.4.3 Proof of the Theorem 2.4.1 thanks to the Propositions 2.4.2 and 2.4.3 . 76
2.4.4 Proof of the Theorem 2.3.1 77
2.5 Proof of the theorem according to $\mathcal{T}$ or $\mathcal{T}^c$ **78**
2.5.1 Proof of the Proposition 2.4.2 78
2.5.2 Proof of the Proposition 2.4.3 81
2.6 Some details **83**
2.6.1 Proof of the Lemma 2.5.1 83
2.6.2 Lemma 2.6.5 and Lemma 4.15 86

In this chapter, we focus on a theoretical result for the Lasso-MLE procedure. We will get a penalty which depends on the model complexity for which the model selected by the penalized criterion among the collection constructed satisfies an oracle inequality. This result is non-asymptotic. We derive it from a general model selection theorem, also detailed here, which is a generalization of Cohen and Le Pennec Theorem, [CLP11], for a model collection constructed randomly.

3.1 Introduction

The goal of clustering methods is to discover a structure among individuals described by several variables. Specifically, in regression case, given n observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$ which are realizations of random variables (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, one aims at grouping the data into clusters such that the observations Y conditionally to X in the same cluster are more similar to each other than those from the other clusters. Different methods could be considered, more geometric or more statistical. We are dealing with model-based clustering, in order to have a rigorous statistical framework to assess the number of clusters and the role of each variable. This method is known to have good empirical performance relative to its competitors, see in [TMZT06].

Datasets are described by a lot of explicative variables, sometimes much more than the sample size. All the information should not be relevant for the clustering. To solve this problem, we propose a procedure which provide a data clustering from variable selection. In a density estimation way, we could also cite Pan and Shen, in [PS07], who focus on mean variable selection, Zhou and Pan, in [ZPS09], who use the Lasso estimator to regularize Gaussian mixture model with general covariance matrices, Sun and Wand, in [SWF12], who propose to regularize the k-means algorithm to deal with high-dimensional data, Guo et al. in [GLMZ10] who propose a pairwise variable selection method. All of them deal with penalized model-based clustering.

In a regression framework, the Lasso estimator, introduced by Tibshirani in [Tib96], is a classical tool in this context. Working well in practice, many efforts have been made recently on this estimator to get some theoretical results. Under a variety of different assumptions on the design matrix, we could have oracle inequalities for the Lasso estimator. For example, we can state the restricted eigenvalue condition, introduced by Bickel, Ritov and Tsybakov in [BRT09], who get an oracle inequality with this assumption. For an overview of existing results, cite for example [vdGB09].

Whereas focus on the estimation, the Lasso estimator could be used to select variables, and, for this goal, many results without strong assumptions are proved. The first result in this way is from Meinshausen and Bühlmann, in [MB10], who show that, for neighborhood selection in Gaussian graphical models, under a neighborhood stability condition, the Lasso estimator is consistent. Under different assumptions, as the irrepresentable condition, described in [ZY06], one get the same kind of result: true variables are selected consistently.

Thanks to those results, one could refit the estimation, after the variable selection, with an estimator with better properties. In this chapter, we focus on the maximum likelihood estimator on the estimated active set. In a linear regression framework, we could cite Massart and Meynet, [MM11a], or Belloni and Chernozhukov, [BC13], or also Sun and Zhang, [SZ12] for using this idea.

In our case of finite mixture regression, we propose a procedure which is based on a modeling that recasts variable selection and clustering problems into a model selection problem. This procedure is developed in [Dev14c], with methodology, computational issues, simulations and

data analysis. First, for some data-driven regularization parameters, we construct a relevant variables set. Then, restricted on those sets, we compute the maximum likelihood estimator. Considering the model collection with various number of components and various sparsities, we select a model thanks to the slope heuristic. Then, we get a clustering of the data thanks to the Maximum A Posteriori principle. This procedure could be used to cluster heterogeneous multivariate regression data and understand which variables explain the clustering, in high-dimension. Considering a regression clustering could refine a clustering, and it could be more adapted for instance for prediction. In this chapter, we focus on the theoretical point of view. We define a penalized criterion which allows to select a model (defined by the number of clusters and the set of relevant variables) from a non-asymptotic point of view. Penalizing the empirical contrast is an idea emerging from the seventies. Akaike, in [Aka74], proposed the Akaike's Information Criterion (AIC) in 1973, and Schwarz in 1978 in [Sch78] suggested the Bayesian Information Criterion (BIC). Those criteria are based on asymptotic heuristics. To deal with non-asymptotic observations, Birgé and Massart in [BM07] and Barron et al. in [YB99], define a penalized data-driven criterion, which leads to oracle inequalities for model selection. The aim of our approach is to define penalized data-driven criterion which leads to an oracle inequality for our procedure. In our context of regression, Cohen and Le Pennec, in [CLP11], proposed a general model selection theorem for maximum likelihood estimation, adapted from Massart's Theorem in [Mas07]. Nevertheless, we can not apply it directly, because it is stated for a deterministic model collection, whereas our data-driven model collection is random, constructed by the Lasso estimator. As Maugis and Meynet have done in [MMR12] to generalize Massart's Theorem, we extend the theorem to cope with the randomness of our model collection. By applying this general theorem to the finite mixture regression random model collection constructed by our procedure, we derive a convenient theoretical penalty as well as an associated non-asymptotic penalized criteria and an oracle inequality fulfilled by our Lasso-MLE estimator. The advantage of this procedure is that it does not need any restrictive assumption.
To obtain the oracle inequality, we use a general theorem proposed by Massart in [Mas07], which gives the form of the penalty and associated oracle inequality in term of the Kullback-Leibler and Hellinger loss. In our case of regression, Cohen and Le Pennec, in [CLP11], generalize this theorem in term of Kullback-Leibler and Jensen-Kullback-Leibler loss. Those theorems are based on the centred process control with the bracketing entropy, allowing to evaluate the size of the models. Our setting is more general, because we work with a random family denoted by $\check{\mathcal{M}}$. We have to control the centred process thanks to Bernstein's inequality.
The rest of this chapter is organized as follows. In the Section 3.2, we define the multivariate Gaussian mixture regression model, and we describe the main steps of the procedure we propose. We also illustrate the requirement of refitting by some simulations. We present our oracle inequality in the Section 3.3.2. In Section 3.4, we illustrate the procedure on simulated dataset and benchmark dataset. Finally, in Section 3.5, we give some tools to understand the proof of the oracle inequality, with a global theorem of model selection with a random collection in Section 3.5.1 and sketch of proofs after. All the details are given in Appendix.

3.2 The Lasso-MLE procedure

In order to cluster high-dimensional regression data, we will work with the multivariate Gaussian mixture regression model. This model is developed in [SBG10] in the scalar response case. We generalize it in Section 3.2.1. Moreover, we want to construct a model collection. We propose, in Section 3.2.2, a procedure called Lasso-MLE which constructs a model collection, with various sparsity and various number of components, of Gaussian mixture regression models. The different sparsities solve the high-dimensional problem. We conclude this section with

simulations, which illustrate the advantage of refitting.

3.2.1 Gaussian mixture regression model

We observe n independent couples $(x_i, y_i)_{1\leq i\leq n}$ realizing random variables (X, Y), where $X \in \mathbb{R}^p$, and $Y \in \mathbb{R}^q$ comes from a probability distribution with unknown conditional density denoted by s^*. To solve a clustering problem, we use a finite mixture model in regression. In particular, we will approximate the density of Y conditionally to X with a mixture of K multivariate Gaussian regression models. If the observation i belongs to the cluster k, we are looking for $\beta_k \in \mathbb{R}^{q\times p}$ such that $y_i = \beta_k x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \Sigma_k)$. Remark that we also have to estimate the number of clusters K.
Thus, the random response variable $Y \in \mathbb{R}^q$ depends on a set of random explanatory variables, written $X \in \mathbb{R}^p$, through a regression-type model. Give more precisions on the assumptions on the model we use.

— The variables Y_i, conditionally to X_i, are independent, for all $i \in \{1, \dots, n\}$;
— $Y_i|X_i = x_i \sim s_\xi^K(y|x_i)dy$, with

$$s_\xi^K(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t\Sigma_k^{-1}(y-\beta_k x)}{2}\right) \tag{3.1}$$

$$\xi = (\pi_1, \dots, \pi_K, \beta_1, \dots, \beta_K, \Sigma_1, \dots, \Sigma_K) \in \Xi_K = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right)$$

$$\Pi_K = \left\{(\pi_1, \dots, \pi_K); \pi_k > 0 \text{ for } k \in \{1, \dots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$$

$\mathbb{S}_q^{++}$ is the set of symmetric positive definite matrices on $\mathbb{R}^q$.

We want to estimate the conditional density function s_ξ^K from the observations. For all $k \in \{1, \dots, K\}, \beta_k$ is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k. The π_ks are the mixture proportions. In fact, for a regressor x, for all $k \in \{1, \dots, K\}$, for all $z \in \{1, \dots, q\}$, $[\beta_k x]_z = \sum_{j=1}^p [\beta_k]_{z,j} x_j$ is the zth component of the mean of the mixture component k. To deal with high-dimensional data, we select variables.

Definition 3.2.1. *A variable $(z, j) \in \{1, \dots, q\} \times \{1, \dots, p\}$ is said to be* irrelevant *if, for all $k \in \{1, \dots, K\}$, $[\beta_k]_{z,j} = 0$. A variable is* relevant *if it is not irrelevant.*
A model is said to be sparse *if there are a few of relevant variables.*

We denote by $\Phi^{[J]}$ the matrix with 0 on the set ${}^cJ$, and $\mathcal{S}_{(K,J)}$ the model with K components and with J for relevant variables set:

$$\mathcal{S}_{(K,J)} = \left\{y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto s_\xi^{(K,J)}(y|x)\right\} \tag{3.2}$$

where

$$s_\xi^{(K,J)}(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k^{[J]}x)^t\Sigma_k^{-1}(y-\beta_k^{[J]}x)}{2}\right).$$

This is the main model used in this chapter. To construct the set of relevant variables J, we use the Lasso estimator. Rather than select a regularization parameter, we consider a collection, which leads to a model collection. Detail the procedure.

3.2.2 The Lasso-MLE procedure

The procedure we propose, which is particularly interesting in high-dimension, could be decomposed into three main steps. First, we construct a model collection, with models more or less sparse and with more or less components. Then, we refit estimations with the maximum likelihood estimator. Finally, we select a model thanks to the slope heuristic. It leads to a clustering according to the MAP principle on the selected model.

Model collection construction The first step consists of constructing a collection of models $\{\mathcal{S}_{(K,J)}\}_{(K,J)\in\mathcal{M}}$ in which the model $\mathcal{S}_{(K,J)}$ is defined by equation (3.2), and the model collection is indexed by $\mathcal{M} = \mathcal{K} \times \mathcal{J}$. Denote by $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components, and denote by $\mathcal{J}$ a collection of subsets of $\{1, \dots, q\} \times \{1, \dots, p\}$.
To detect the relevant variables, and construct the set J in each model, we generalize the Lasso estimator. Indeed, we penalize the empirical contrast by an ℓ_1-penalty on the mean parameters proportional to $||P_k\beta_k||_1 = \sum_{j=1}^{p}\sum_{z=1}^{q} |(P_k\beta_k)_{z,j}|$, where $P_k^t P_k = \Sigma_k^{-1}$ for all $k \in \{1, \dots, K\}$. Then, we will consider

$$\hat{\xi}_K^{\text{Lasso}}(\lambda) = \underset{\xi=(\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\Sigma})\in\Xi_K}{\text{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^{K} \pi_k ||P_k\beta_k||_1 \right\}.$$

This leads to penalize simultaneously the ℓ_1-norm of the mean coefficients and small variances. Computing those estimators lead to construct the relevant variables set. For a fixed number of mixture components $K \in \mathcal{K}$, denote by G_K a candidate of regularization parameters. Fix a parameter $\lambda \in G_K$, we could then use an EM algorithm to compute the Lasso estimator, and construct the set of relevant variables $J_{(\lambda,K)}$, saying the non-zero coefficients. We denote by $\mathcal{J}$ the random collection of all these sets, $\mathcal{J} = \bigcup_{K\in\mathcal{K}} \bigcup_{\lambda\in G_K} J_{(\lambda,K)}$.

Refitting The second step consists of approximating the maximum likelihood estimator

$$\hat{s}^{(K,J)} = \underset{t\in\mathcal{S}_{(K,J)}}{\text{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(t(y_i|x_i)) \right\}$$

using an EM algorithm for each model $(K, J) \in \mathcal{K}\times\mathcal{J}$. Remark that we estimate all parameters, to reduce bias induced by the Lasso estimator.

Model selection The third step is devoted to model selection. We get a model collection, and we need to select the best one. Because we do not have access to s^*, we can not take the one which minimizes the risk. The Theorem 3.5.1 solves this problem: we get a penalty achieving to an oracle inequality. Then, even if we do not have access to s^*, we know that we can do almost like the oracle.

3.2.3 Why refit the Lasso estimator?

In order to illustrate the refitting, we compute multivariate data, the restricted eigenvalue condition being not satisfied, and run our procedure. We consider an extension of the model studied in Giraud et al. article [BGH09] in the Section 6.3. Indeed, this model is a linear regression with a scalar response which does not satisfy the restricted eigenvalues condition. Then, we define different classes, to get a finite mixture regression model, which does not satisfied the restricted eigenvalues condition, and extend the dimension for multivariate response. We could compare

the result of our procedure with the Lasso estimator, to illustrate the oracle inequality we have get. Let precise the model.
Let $[\mathbf{x}]_1, [\mathbf{x}]_2, [\mathbf{x}]_3$ be three vectors of $\mathbb{R}^n$ defined by

$$\begin{aligned}
[\mathbf{x}]_1 &= (1, -1, 0, \dots, 0)^t/\sqrt{2} \\
[\mathbf{x}]_2 &= (-1, 1.001, 0, \dots, 0)^t/\sqrt{1+0.001^2} \\
[\mathbf{x}]_3 &= (1/\sqrt{2}, 1/\sqrt{2}, 1/n, \dots, 1/n)^t/\sqrt{1+(n-2)/n^2}
\end{aligned}$$

and for $4 \leq j \leq n$, let $[\mathbf{x}]_j$ be the jth vector of the canonical basis of $\mathbb{R}^n$. We take a sample of size $n = 20$, and vector of size $p = q = 10$. We consider two classes, each of them defined by $[\beta_1]_{z,j} = 10$ and $[\beta_2]_{z,j} = -10$ for $j \in \{1, \dots, 2\}$, $z \in \{1, \dots, 10\}$. Moreover, we define the covariance matrix of the noise by a diagonal matrix with 0.01 for diagonal coefficient in each class.
We run our procedure on this model, and compare it with the Lasso estimator, without refitting. We compute the model selected by the slope heuristic over the model collection constructed by the Lasso estimator. In Figure 3.1 are the boxplots of each procedure, running 20 times. The Kullback-Leibler divergence is computed over a sample of size 5000.

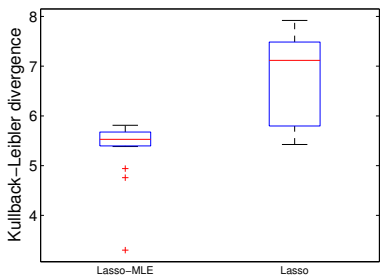

Figure 3.1: Boxplot of the Kullback-Leibler divergence between the true model and the one constructed by each procedure, the Lasso-MLE procedure and the Lasso estimator

We could see that a refitting after variable selection by the Lasso estimator leads to a better estimation, according to the Kullback-Leibler loss.

3.3 An oracle inequality for the Lasso-MLE model

Before state the main theorem of this chapter, we need to precise some definitions and notations.

3.3.1 Notations and framework

We assume that the observations $(x_i, y_i)_{1 \leq i \leq n}$ are realizations of random variables (X, Y) where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$.
For $(K, J) \in \mathcal{K} \times \mathcal{J}$, for a model $\mathcal{S}_{(K,J)}$, we denote by $\hat{s}^{(K,J)}$ the maximum likelihood estimator

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)}}{\operatorname{argmin}} \left(-\sum_{i=1}^n \log s_\xi^{(K,J)}(y_i|x_i) \right).$$

To avoid existence issue, we could work with almost minimizer of this quantity and define an η-log-likelihood minimizer:

$$\sum_{i=1}^n -\log(\hat{s}^{(K,J)}(y_i|x_i)) \leq \inf_{s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)}} \left(\sum_{i=1}^n -\log s_\xi^{(K,J)}(y_i|x_i) \right) + \eta.$$

The best model in this collection is the one with the smallest risk. However, because we do not have access to the true density s^*, we can not select the best model, which we call the oracle. Thereby, there is a trade-off between a bias term measuring the closeness of s^* to the set $\mathcal{S}_{(K,J)}$ and a variance term depending on the complexity of the set $\mathcal{S}_{(K,J)}$ and on the sample size. A good set $\mathcal{S}_{(K,J)}$ will be one for which this trade-off leads to a small risk bound. Because we are working with a maximum likelihood approach, the most natural quality measure is thus the Kullback-Leibler divergence denoted by KL.

$$\mathrm{KL}(s,t) = \begin{cases} \displaystyle\int_{\mathbb{R}} \log\left(\frac{s(y)}{t(y)}\right) s(y)dy \text{ if } sdy << tdy; \\ +\infty \text{ otherwise;} \end{cases} \tag{3.3}$$

for s and t two densities.
As we deal with conditional densities and not classical densities, the previous divergence should be adapted. We define the tensorized Kullback-Leibler divergence by

$$\mathrm{KL}^{\otimes n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(s(.|x_i), t(.|x_i))\right].$$

This divergence, defined in [CLP11] appears as the natural one in this regression setting.
Namely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0,1)$, which is defined by

$$\mathrm{JKL}_\rho(s,t) = \frac{1}{\rho}\,\mathrm{KL}(s,(1-\rho)s+\rho t);$$

and the tensorized one

$$\mathrm{JKL}_\rho^{\otimes n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{JKL}_\rho(s(.|x_i), t(.|x_i))\right].$$

This divergence is studied in [CLP11]. We use this divergence rather than the Kullback-Leibler one because we need a boundedness assumption on the controlled functions that is not satisfied by the log-likelihood differences $-\log\left(s_\xi^{(K,J)}/s^*\right)$. When considering the Jensen-Kullback-Leibler divergence, those ratios are replaced by ratios

$$-\frac{1}{\rho}\log\left(\frac{(1-\rho)s^* + \rho s_\xi^{(K,J)}}{s^*}\right)$$

that are close to the log-likelihood differences when s_m are close to s^* and always upper bounded by $-\log(1-\rho)/\rho$. Indeed, this bound is needed to use deviation inequalities for sums of random variables and their suprema, which is the key of the proof of oracle type inequality.

3.3.2 Oracle inequality

We denote by $(\mathcal{S}_{(K,J)})_{(K,J)\in\mathcal{K}\times\mathcal{J}^L}$ the model collection constructed by the Lasso-MLE procedure, with $\mathcal{J}^L$ a random subcollection of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$ constructed by the Lasso estimator. The grid of regularization parameter considered is data-driven, then random. Because we work in high-dimension, we could not look at all subsets of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$. Considering the Lasso estimator through its regularization path is the solution chosen here, but it needs more control because of the random family. To get theoretical results, we need to work with restricted

parameters. Assume Σ_k diagonal, with $\Sigma_k = \mathrm{diag}([\Sigma_k]_{1,1}, \dots, [\Sigma_k]_{q,q})$, for all $k \in \{1, \dots, K\}$. We define

$$\mathcal{S}^{\mathcal{B}}_{(K,J)} = \Bigg\{ s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)} \Big| \text{ for all } k \in \{1, \dots, K\}, [\beta_k]^{[J]} \in [-A_\beta, A_\beta]^{q \times p}, \\ a_\Sigma \leq [\Sigma_k]_{z,z} \leq A_\Sigma \text{ for all } z \in \{1, \dots, q\}, \text{ for all } k \in \{1, \dots, K\} \Bigg\}. \tag{3.4}$$

Moreover, we assume that the covariates X belong to an hypercube. Without any restriction, we could assume that $X \in [0, 1]^p$.

Remark 3.3.1. *We have to denote that in this chapter, the relevant variables set is designed by the Lasso estimator. Nevertheless, any tool could be used to construct this set, and we obtain analog results. We could work with any random subcollection of $\mathcal{P}(\{1, \dots, q\} \times \{1, \dots, p\})$, the controlled size being required in high-dimensional case.*

Theorem 3.3.2. *Let $(x_i, y_i)_{1 \leq i \leq n}$ the observations, with unknown conditional density s^*. Let $\mathcal{S}_{(K,J)}$ defined by (3.2). For $(K, J) \in \mathcal{K} \times \mathcal{J}^L$, $\mathcal{J}^L$ being a random subcollection of $\mathcal{P}(\{1, \dots, q\} \times \{1, \dots, p\})$ constructed by the Lasso estimator, denote $\mathcal{S}^{\mathcal{B}}_{(K,J)}$ the model defined by (3.4). Consider the maximum likelihood estimator*

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}}{\mathrm{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^n \log s_\xi^{(K,J)}(y_i|x_i) \right\}.$$

Denote by $D_{(K,J)}$ the dimension of the model $\mathcal{S}^{\mathcal{B}}_{(K,J)}$, $D_{(K,J)} = K(|J| + q + 1) - 1$. Let $\bar{s}^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}$ such that

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}^{(K,J)}) \leq \inf_{t \in \mathcal{S}^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

and let $\tau > 0$ such that $\bar{s}^{(K,J)} \geq e^{-\tau} s^*$. Let $\mathrm{pen} : \mathcal{K} \times \mathcal{J} \to \mathbb{R}_+$, and suppose that there exists an absolute constant $\kappa > 0$ and an absolute constant $B(A_\beta, A_\Sigma, a_\Sigma)$ such that, for all $(K, J) \in \mathcal{K} \times \mathcal{J}$,

$$\mathrm{pen}(K, J) \geq \kappa \frac{D_{(K,J)}}{n} \Bigg[B^2(A_\beta, A_\Sigma, a_\Sigma) - \log \left(\frac{D_{(K,J)}}{n} B^2(A_\beta, A_\Sigma, a_\Sigma) \wedge 1 \right) \\ + (1 \vee \tau) \log \left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq} \right) \Bigg].$$

Then, the estimator $\hat{s}^{(\hat{K}, \hat{J})}$, with

$$(\hat{K}, \hat{J}) = \underset{(K,J) \in \mathcal{K} \times \mathcal{J}^L}{\mathrm{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^n \log(\hat{s}^{(K,J)}(y_i|x_i)) + \mathrm{pen}(K, J) \right\}$$

satisfies

$$\mathrm{E}\left[\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}^{(\hat{K}, \hat{J})}) \right] \leq C_1 \, \mathrm{E} \left(\inf_{(K,J) \in \mathcal{K} \times \mathcal{J}^L} \left(\inf_{t \in S^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \mathrm{pen}(K, J) \right) \right) \\ + C_2 \frac{(1 \vee \tau)}{n};$$

for some absolute positive constants C_1 and C_2.

This oracle inequality compares performances of our estimator with the best model in the collection. Nevertheless, as we consider mixture of Gaussian, if we take enough clusters, we could approximate well a lot of densities. This result could be compared with the oracle inequality get in [SBG10], Theorem 4. Indeed, under restricted eigenvalues condition and fixed design, they get an oracle inequality for the Lasso estimator in finite mixture regression model, with scalar response and high-dimension regressors. Note that they control the divergence with the true parameters. We get a similar result for the Lasso-MLE estimator. Moreover, our procedure work in a more general framework, the only assumption needed is to be bounded.

3.4 Numerical experiments

To illustrate this procedure, we study some simulations and real data. The main algorithm is a generalized version of the EM algorithm, which is used many times for the procedure. We first use it to compute maximum likelihood estimator, to construct the regularization parameter grid. Then, we use it to compute the Lasso estimator for each regularization parameter belonging to the grid, and we are able to construct the relevant variables set. Finally, we could compute the maximum likelihood estimator, restricted to those relevant variables in each model. Among this model collection, we select one using the Capushe package. More details, as initialization rule, stopping rule, and more numerical experiments, are available in [Dev14c].

3.4.1 Simulation illustration

We illustrate the procedure on a simulated dataset, adapted from [SBG10].
Let $\mathbf{x}$ be a sample of size $n = 100$ distributed according to multivariate standard Gaussian. We consider a mixture of two components, and we fix the dimension of the regressor and of the response variables to $p = q = 10$. Besides, we fix the number of relevant variables to 4 in each cluster. More precisely, the first four variables of Y are explained respectively by the four first variables of X. Fix $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$, $[\beta_1]^{[J]} = 3$, $[\beta_2]^{[J]} = -2$ and $P_k = 3I_q$ for all $k \in \{1, 2\}$.
The difficulty of the clustering is partially controlled by the signal-to-noise ratio. In this context, we could extend the natural idea of the SNR with the following definition, where $\mathrm{Tr}(A)$ denotes the trace of the matrix A.

$$\mathrm{SNR} = \frac{\mathrm{Tr}(\mathrm{Var}(Y))}{\mathrm{Tr}(\mathrm{Var}(Y|\beta_k = 0 \text{ for all } k \in \{1, \dots, K\}))} = 1.88.$$

We take a sample of Y knowing $X = x$ according to a Gaussian mixture, meaning in $\beta_k x$ and with covariance matrix $\Sigma_k = (P_k^t P_k)^{-1} = \sigma I_q$, for the cluster $k \in \{1, 2\}$. We run our procedures with the number of components varying in $\mathcal{K} = \{2, \dots, 5\}$.
To compare our procedure with others, we compute the Kullback-Leibler divergence with the true density, the ARI (the Adjusted Rand Index measures the similarity between two data clusterings, knowing that the closer to 1 the ARI, the more similar the two partitions), and how many clusters are selected.
From the Lasso-MLE model collection, we construct two models, to compare our procedures with. We compute the oracle (the model which minimizes the Kullback-Leibler divergence with the true density), and the model which is selected by the BIC criterion instead of the slope heuristic. Thanks to the oracle, we know how good we could be from this model collection for the Kullback-Leibler divergence, and how this model, as good it is possible for the contrast, performs the clustering.
The third procedure we compare with is the maximum likelihood estimator, assuming that we know how many clusters there are, fixed to 2. We use this procedure to show that variable selection is necessary.

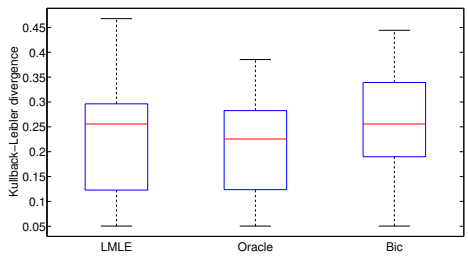

Figure 3.2: Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the 20 simulations, for the Lasso-MLE procedure (LMLE), the oracle (Oracle), the BIC estimator (BIC)

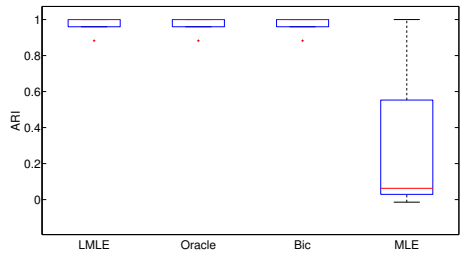

Figure 3.3: Boxplots of the ARI over the 20 simulations, for the Lasso-MLE procedure (LMLE), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE)

Results are summarized in Figure 3.2 and in Figure 3.3. The Kullback-Leibler divergence is smaller for models selected in our model collection (else by BIC, or by slope heuristic, or the oracle) than for the model constructed by the MLE. The ARI is closer to 1 in those case, and, moreover, is better for the model selected by the slope heuristic. We could conclude that the model collection is well constructed, selecting relevant variables, and also that the model is well selected among this collection, near the oracle.

3.4.2 Real data

We also illustrate the procedure on the Tecator dataset, which deal with spectrometric data. We summarize here results which are described in [Dev14c]. Those data have been studied in a lot of articles, cite for example Ferraty and Vieu's book [FV06]. The data consist of a 100-channel spectrum of absorbances in the wavelength range $850 - 1050$ nm, and of the percentage of fat. We observe a sample of size 215. In this work, we focus on clustering data according to the reliance between the fat content and the absorbance spectrum. The sample will be split into two subsamples, 165 observations for the learning set, and 50 observations for the test set. We split it to have the same marginal distribution of the response in each sample.
The spectrum is a function, which we decompose into the Haar basis, at level 6.
The procedure selects two models, which we describe here. In Figures (3.4) and (3.5), we represent clusters done on the training set for the different models.
The graph on the left is a candidate for representing each cluster, constructed by the mean of spectrum over an a posteriori probability greater than 0.6. We plot the curve reconstruction, keeping only active variables in the wavelet decomposition. On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.
The first model has two clusters, which could be distinguish in the absorbance spectrum by the bump on wavelength around 940 nm. The first class is dominating, with $\hat{\pi}_1 = 0.95$. The fat content is smaller in the first class than in the second class. According to the signal reconstruction, we could see that almost all variables have been selected. This model seems consistent according to the classification goal.
The second model has 3 clusters, and we could remark different important wavelength. Around 940 nm, there is some differences between clusters, corresponding to the bump underline in the model 1, but also around 970 nm, with higher or smaller values. The first class is dominating,

with $\hat{\pi}_1 = 0.89$. Just a few of variables have been selected, which give to this model the understanding property of which coefficient are discriminating.

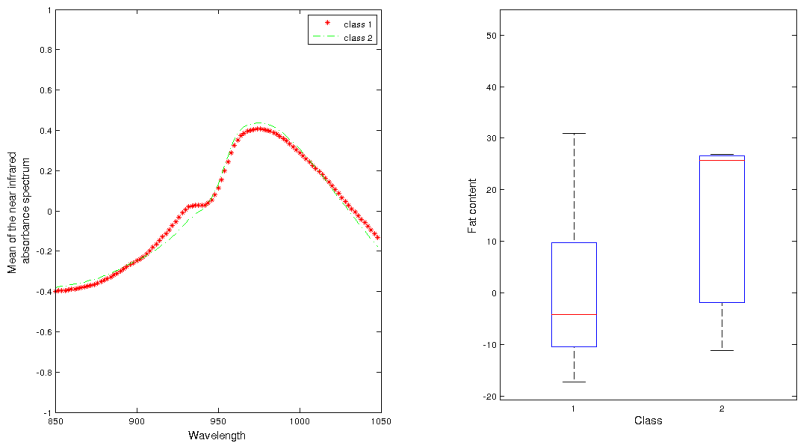

Figure 3.4: Summarized results for the model 1. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than 0.6 On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.

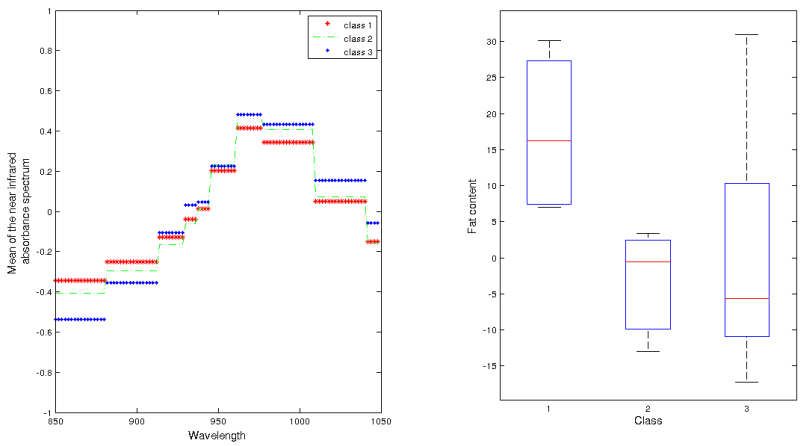

Figure 3.5: Summarized results for the model 2. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than 0.6 On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.

We could discuss about those models. The first select only two clusters, but almost all variables, whereas the second model has more clusters, and less variables: there is a trade-off between clusters and variable selection for the dimension reduction.

3.5 Tools for proof

In this section, we present the tools needed to understand the proof. First, we present a general theorem for model selection in regression among a random collection. Then, in subsection 3.5.2, we present the proof of this theorem, and in the next subsection we explain how we could use the main theorem to get the oracle inequality. All details are available in Appendix.

3.5.1 General theory of model selection with the maximum likelihood estimator.

To get an oracle inequality for our clustering procedure, we have to use a general model selection theorem. Because the model collection constructed by our procedure is random, because of the Lasso estimator which select variables randomly, we have to generalize Cohen and Le Pennec Theorem. Begin by some general model selection theory.
Before state the general theorem, begin by talking about the assumptions. We work here in a more general context, $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, and $(S_m)_{m\in\mathcal{M}}$ defining a model collection indexed by $\mathcal{M}$. First, we impose a structural assumption on each model indexed by $m \in \mathcal{M}$. It is a bracketing entropy condition on the model S_m with respect to the Hellinger divergence, defined by

$$(d_H^{\otimes n})^2(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} d_H^2(s(.|x_i), t(.|x_i))\right].$$

A bracket $[l, u]$ is a pair of functions such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}, l(y, x) \leq s(y|x) \leq u(y, x)$. The bracketing entropy $\mathcal{H}_{[.]}(\epsilon, S, d_H^{\otimes n})$ of a set S is defined as the logarithm of the minimum number of brackets $[l, u]$ of width $d_H^{\otimes n}(l, u)$ smaller than ϵ such that every functions of S belong to one of these brackets. It leads to the Assumption H_m.

Assumption $\mathbf{H}_m$. *There is a non-decreasing function ϕ_m such that $\varpi \mapsto \frac{1}{\varpi}\phi_m(\varpi)$ is non-increasing on $(0, +\infty)$ and for every $\varpi \in \mathbb{R}^+$ and every $s_m \in S_m$,*

$$\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, S_m(s_m, \varpi), d_H^{\otimes n})}d\epsilon \leq \phi_m(\varpi);$$

where $S_m(s_m, \varpi) = \{t \in S_m, d_H^{\otimes n}(t, s_m) \leq \varpi\}$. The model complexity $\mathcal{D}_m$ is then defined as $n\varpi_m^2$ with ϖ_m the unique root of

$$\frac{1}{\varpi}\phi_m(\varpi) = \sqrt{n}\varpi. \tag{3.5}$$

Remark that the model complexity depends on the bracketing entropies not of the global models S_m but of the ones of smaller localized sets. This is a weaker assumption.
For technical reason, a separability assumption is also required.

Assumption $\mathbf{Sep}_m$. *There exists a countable subset $S_m^{'}$ of S_m and a set $\mathcal{Y}_m^{'}$ with $\lambda(\mathcal{Y}\setminus\mathcal{Y}_m^{'}) = 0$ such that for every $t \in S_m$, there exists a sequence $(t_l)_{l\geq 1}$ of elements of $S_m^{'}$ such that for every x and every $y \in \mathcal{Y}_m^{'}$, $\log(t_l(y|x))$ goes to $\log(t(y|x))$ as l goes to infinity.*

This assumption leads to work with a countable family, which allows to cope with the randomness of $\hat{s}_m$. We also need an information theory type assumption on our collection. We assume the existence of a Kraft-type inequality for the collection.

Assumption K. *There is a family $(w_m)_{m\in\mathcal{M}}$ of non-negative numbers such that*

$$\sum_{m\in\mathcal{M}} e^{-w_m} \leq \Omega < +\infty.$$

The difference with Cohen and Le Pennec's Theorem is that we consider a random collection of models $\check{\mathcal{M}}$, included in the whole collection $\mathcal{M}$. In our procedure, we deal with high-dimensional models, and we cannot look after all the models: we have to restrict ourselves to a smaller subcollection of models, which is then random. In the proof of the theorem, we have to be careful with the recentred process of $-\log(\bar{s}_m/s^*)$. Because we conclude by taking the expectation, if $\mathcal{M}$ is fixed, this term is non-interesting, but if we consider a random family, we have to use the Bernstein inequality to control this quantity, and then we have to make the assumption (3.6). Let state our main global theorem.

Theorem 3.5.1. *Assume we observe $(x_i, y_i)_{1\leq i\leq n}$ with unknown conditional density s^*. Let the model collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ be at most countable collection of conditional density sets. Assume Assumption K holds, while assumptions H_m and Sep_m hold for every $m \in \mathcal{M}$. Let $\delta_{\mathrm{KL}} > 0$, and $\bar{s}_m \in S_m$ such that*

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) \leq \inf_{t\in S_m} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

and let $\tau > 0$ such that

$$\bar{s}_m \geq e^{-\tau} s^*. \tag{3.6}$$

Introduce $(S_m)_{m\in\check{\mathcal{M}}}$ some random subcollection of $(S_m)_{m\in\mathcal{M}}$. Consider the collection $(\hat{s}_m)_{m\in\check{\mathcal{M}}}$ of η-log-likelihood minimizer in S_m, satisfying, for all $m \in \check{\mathcal{M}}$,

$$\sum_{i=1}^n -\log(\hat{s}_m(y_i|x_i)) \leq \inf_{s_m\in S_m} \left(\sum_{i=1}^n -\log(s_m(y_i|x_i)) \right) + \eta.$$

Then, for any $\rho \in (0, 1)$ and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq \kappa(\mathcal{D}_m + (1 \vee \tau)w_m) \tag{3.7}$$

with $\kappa > \kappa_0$, and where the model complexity $\mathcal{D}_m$ is defined in (3.5), the penalized likelihood estimate $\hat{s}_{\hat{m}}$ with $\hat{m} \in \check{\mathcal{M}}$ such that

$$-\sum_{i=1}^n \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\check{\mathcal{M}}} \left(-\sum_{i=1}^n \log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m) \right) + \eta'$$

satisfies

$$\begin{aligned} \mathrm{E}(\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}})) \leq & C_1 \,\mathrm{E}\left(\inf_{m\in\check{\mathcal{M}}} \left(\inf_{t\in S_m} \mathrm{KL}^{\otimes_n}(s^*, t) \right) + 2\frac{\mathrm{pen}(m)}{n} \right) \\ & + C_2(1 \vee \tau)\frac{\Omega^2}{n} + \frac{\eta' + \eta}{n}. \end{aligned} \tag{3.8}$$

Obviously, one of the models minimizes the right hand side. Unfortunately, there is no way to know which one without knowing s^*. Hence, this oracle model can not be used to estimate s^*. We nevertheless propose a data-driven strategy to select an estimate among the collection of estimates $\{\hat{s}_m\}_{m\in\check{\mathcal{M}}}$ according to a selection rule that performs almost as well as if we had known this oracle, according to the absolute constant C_1. Using simply the log-likelihood of the estimate in each model as a criterion is not sufficient. It is an underestimation of the true risk of the estimate and this leads to select models that are too complex. By adding an adapted penalty $\mathrm{pen}(m)$, one hopes to compensate for both the variance term and the bias term between

$-1/n\sum_{i=1}^{n}\log\left(\hat{s}_{\hat{m}}(y_i|x_i)/s^*(y_i|x_i)\right)$ and $\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s^*,s_m)$. For a given choice of $\mathrm{pen}(m)$, the best model $S_{\hat{m}}$ is chosen as the one whose index is an almost minimizer of the penalized η-log-likelihood.
Talk about the assumption (3.6). If s is bounded, with a compact support, this assumption is satisfied. It is also satisfied in other cases, more particular. Then it is not a strong assumption, but it is needed to control the random family.
This theorem is available for whatever model collection constructed, whereas assumptions H_m, K and Sep_m are satisfied. In the following, we will use this theorem for the procedure we propose to cluster high-dimensional data. Nevertheless, this theorem is not specific for our context, and could be used whatever the problem.
Remark that the constant associated to the Assumption K appears squared in the bound. It is due to the random subcollection $\check{\mathcal{M}}$ of $\mathcal{M}$, if the model collection is fixed, we get a linear bound. Moreover, the weights w_m appears linearly in the penalty bound.

3.5.2 Proof of the general theorem

For any model S_m, we have denoted by $\bar{s}_m$ a function such that

$$\mathrm{KL}^{\otimes n}(s^*,\bar{s}_m)\leq\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s^*,s_m)+\frac{\delta_{\mathrm{KL}}}{n}.$$

Fix $m\in\mathcal{M}$ such that $\mathrm{KL}^{\otimes n}(s^*,\bar{s}_m)<+\infty$. Introduce

$$\mathcal{M}(m)=\left\{m'\in\mathcal{M}\,\middle|\,P_n(-\log\hat{s}_{m'})+\frac{\mathrm{pen}(m')}{n}\right.$$
$$\left.\leq P_n(-\log\hat{s}_m)+\frac{\mathrm{pen}(m)}{n}+\frac{\eta'}{n}\right\};$$

where $P_n(g)=1/n\sum_{i=1}^{n}g(y_i|x_i)$. We define the functions $kl(\bar{s}_m),kl(\hat{s}_m)$ and $jkl_\rho(\hat{s}_m)$ by

$$kl(\bar{s}_m)=-\log\left(\frac{\bar{s}_m}{s^*}\right);\qquad kl(\hat{s}_m)=-\log\left(\frac{\hat{s}_m}{s^*}\right);$$
$$jkl_\rho(\hat{s}_m)=-\frac{1}{\rho}\log\left(\frac{(1-\rho)s^*+\rho\hat{s}_m}{s^*}\right).$$

For every $m'\in\mathcal{M}(m)$, by definition,

$$\begin{aligned}P_n(kl(\hat{s}_{m'}))+\frac{\mathrm{pen}(m')}{n}\leq&P_n(kl(\hat{s}_m))+\frac{\mathrm{pen}(m)+\eta'}{n}\\ \leq&P_n(kl(\bar{s}_m))+\frac{\mathrm{pen}(m)+\eta'+\eta}{n}.\end{aligned}$$

Let $\nu_n^{\otimes n}(g)$ denote the recentred process $P_n(g)-P^{\otimes n}(g)$. By concavity of the logarithm,

$$kl(\hat{s}_{m'})\ \geq\ jkl_\rho(\hat{s}_{m'}),$$

and then

$$\begin{aligned}&P^{\otimes n}(jkl_\rho(\hat{s}_{m'}))-\nu_n^{\otimes n}(kl(\bar{s}_m))\\ \leq&P^{\otimes n}(kl(\bar{s}_m))+\frac{\mathrm{pen}(m)}{n}-\nu_n^{\otimes n}(jkl_\rho(\hat{s}_{m'}))+\frac{\eta'+\eta}{n}-\frac{\mathrm{pen}(m')}{n},\end{aligned}$$

which is equivalent to

$$\mathrm{JKL}_{\rho}^{\otimes_n}(s^*,\hat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq \mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl_{\rho}(\hat{s}_{m'})) + \frac{\eta^{'}+\eta}{n} - \frac{\mathrm{pen}(m^{'})}{n}. \tag{3.9}$$

Mimic the proof as done in Cohen and Le Pennec [CLP11], we could obtain that except on a set of probability less than $e^{-w_{m'}-w}$, for all w, for all $\mathfrak{z}_{m'} > \sigma_{m'}$, there exist absolute constants $\kappa_0^{'}, \kappa_1^{'}, \kappa_2^{'}$ such that

$$\frac{-\nu_n^{\otimes_n}(jkl_{\rho}(\hat{s}_{m'}))}{\mathfrak{z}_{m'}^2 + \kappa_0^{'}(d_H^{\otimes_n})^2(s^*,\hat{s}_{m'})} \leq \frac{\kappa_1^{'}\sigma_{m'}}{\mathfrak{z}_{m'}} + \kappa_2^{'}\sqrt{\frac{w_{m'}+w}{n\mathfrak{z}_{m'}^2}} + \frac{18}{\rho}\frac{w_{m'}+w}{n\mathfrak{z}_{m'}^2}. \tag{3.10}$$

To obtain this inequality we use the hypothesis Sep_m and H_m. This control is derived from maximal inequalities, described in [Mas07].
Our purpose is now to control $\nu_n^{\otimes_n}(kl(\bar{s}_m))$. This is the difference with the Theorem of Cohen and Le Pennec: we work with a random subcollection $\mathcal{M}^L$ of $\mathcal{M}$.
By definition of kl and $\nu_n^{\otimes_n}$,

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) = -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{\bar{s}_m(y_i|x_i)}{s^*(y_i|x_i)}\right) + \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \log\left(\frac{\bar{s}_m(Y_i|X_i)}{s^*(Y_i|X_i)}\right)\right].$$

We want to apply Bernstein's inequality, which is recalled in Appendix.
If we denote by Z_i the random variable $Z_i = -\frac{1}{n}\log\left(\frac{\bar{s}_m(Y_i|X_i)}{s^*(Y_i|X_i)}\right)$, we get

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) = \sum_{i=1}^n (Z_i - \mathrm{E}(Z_i)).$$

We need to control the moments of Z_i to apply Bernstein's inequality.

Lemme 3.5.2. *Let s^* and $\bar{s}_m$ two conditional densities with respect to the Lebesgue measure. Assume that there exists $\tau > 0$ such that $\log\left(\left|\left|\frac{s^*}{\bar{s}_m}\right|\right|_\infty\right) \leq \tau$. Then,*

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^q}\left(\log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)\right)^2 s^*(y|x_i)dy\right) \leq \frac{\tau^2}{e^{-\tau}+\tau-1}\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m).$$

We prove this lemma in Appendix 3.6.2.
Because $\frac{\tau^2}{e^{-\tau}+\tau-1} \underset{\tau\to\infty}{\sim} \tau$, there exists A such that $\frac{\tau^2}{e^{-\tau}+\tau-1} \leq 2\tau$ for all $\tau \geq A$. For $\tau \in (0,A]$, because this function is continuous and equivalent to 2 in 0, there exists $B > 0$ such that $\frac{\tau^2}{e^{-\tau}+\tau-1} \leq B$. We obtain that $\sum_{i=1}^n \mathrm{E}(Z_i^2) \leq \frac{1}{n}\delta(1\vee\tau)\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)$, where $\delta = 2\vee B$.
Moreover, for all integers $K \geq 3$,

$$\begin{aligned}
\sum_{i=1}^n \mathrm{E}((Z_i)_+^K) &\leq \sum_{i=1}^n \frac{1}{n^K}\int_{\mathbb{R}^q}\left(\log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)\right)_+^K s^*(y|x_i)dy \\
&\leq \frac{n}{n^K}\int_{\mathbb{R}^q}\log\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right)^{K-2}\log\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right)^2 \mathbb{1}_{s^*\geq\bar{s}_m(y|x)}s^*(y|x)dy \\
&\leq \frac{n}{n^K}\tau^{K-2}\delta(1\vee\tau)\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m).
\end{aligned}$$

Assumptions of Bernstein's inequality are then satisfied, with

$$v = \frac{\delta(1 \vee \tau)\,\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)}{n}, \qquad c = \frac{\tau}{n},$$

thus, for all $u > 0$, except on a set with probability less than e^{-u},

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq \sqrt{2vu} + cu.$$

Thus, for all $z > 0$, for all $u > 0$, except on a set with probability less than e^{-u},

$$\frac{\nu_n^{\otimes_n}(kl(\bar{s}_m))}{z^2 + \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)} \leq \frac{\sqrt{2vu} + cu}{z^2 + \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)} \leq \frac{\sqrt{vu}}{z\sqrt{2\,\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)}} + \frac{cu}{z^2}. \tag{3.11}$$

We apply this bound to $u = w + w_m + w_{m'}$. We get that, except on a set with probability less than $e^{-(w+w_m+w_{m'})}$, using that $a^2 + b^2 \geq a^2$, from the inequality (3.10),

$$-\nu_n^{\otimes_n}(jkl_\rho(\hat{s}_{m'})) \leq \left(\mathfrak{z}_{m'}^2 + \kappa_0'(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'})\right)\left(\frac{\kappa_1' + \kappa_2'}{\theta} + \frac{18}{\theta^2\rho}\right),$$

and, from the inequality (3.11),

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq (\beta + \beta^2)\left(z_{m,m'}^2 + \mathrm{KL}^{\otimes_n}(s, s_m)\right),$$

where we have chosen

$$\mathfrak{z}_{m'} = \theta\sqrt{\sigma_{m'}^2 + \frac{w_{m'} + w}{n}},$$

with $\theta > 1$ to fix later, and

$$z_{m,m'} = \beta^{-1}\sqrt{\left(\frac{v}{2\,\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)} + c\right)(w + w_m + w_{m'})},$$

with $\beta > 0$ to fix later.
Coming back to the inequality (3.9),

$$\begin{aligned}\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq\, & \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) + \frac{\mathrm{pen}(m)}{n} \\ & + (\mathfrak{z}_{m'}^2 + \kappa_0'(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'}))\left(\frac{\kappa_1' + \kappa_2'}{\theta} + \frac{18}{\theta^2\rho}\right) \\ & + \frac{\eta' + \eta}{n} - \frac{\mathrm{pen}(m')}{n} + (\beta + \beta^2)(z_{m,m'}^2 + \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)).\end{aligned}$$

Recall that $\bar{s}_m$ is chosen such that

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) \leq \inf_{s_m \in S_m} \mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\delta_{\mathrm{KL}}}{n}.$$

Put $\kappa(\beta) = 1 + (\beta + \beta^2)$, and let $\epsilon_1 > 0$, we define θ_1 by $\kappa_0'\left(\frac{\kappa_1' + \kappa_2'}{\theta_1} + \frac{18}{\theta_1^2\rho}\right) = C_\rho\epsilon_1$ where C_ρ is defined by $C_\rho(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'}) \leq \mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'})$, and put $\kappa_2 = \frac{C_\rho\epsilon_1}{\kappa_0}$. We get that

$$\begin{aligned}(1 - \epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq\, & \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n} - \frac{\mathrm{pen}(m')}{n} \\ & + \kappa(\beta)\frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n} + \mathfrak{z}_{m'}^2\kappa_2 + (\beta + \beta^2)z_{m,m'}^2.\end{aligned}$$

Since $\tau \leq 1 \vee \tau$, if we choose β such that $(\beta + \beta^2)(\delta/2 + 1) = \alpha\theta_1^{-2}\beta^{-2}$, and if we put $\kappa_1 = \alpha\gamma^{-2}(\beta^{-2} + 1)$, since $1 \leq 1 \vee \tau$, using the expressions of $\mathfrak{z}_{m'}$ and $z_{m,m'}$, we get that

$$\begin{aligned}(1-\epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq & \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n} - \frac{\mathrm{pen}(m')}{n} \\ & + \kappa(\beta)\frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n} \\ & + \kappa_2\theta_1^2\left(\sigma_{m'}^2 + \frac{w + w_{m'}}{n}\right) + \kappa_1(1 \vee \tau)\frac{w + w_m + w_{m'}}{n} \\ \leq & \kappa(\beta)\,\mathrm{KL}^{\otimes n}(s^*, s_m) + \left(\frac{\mathrm{pen}(m)}{n} + \kappa_1(1 \vee \tau)\frac{w_m}{n}\right) \\ & + \left(-\frac{\mathrm{pen}(m')}{n} + \kappa_2\theta_1^2\left(\sigma_{m'}^2 + \frac{w_{m'}}{n}\right) + \kappa_1(1 \vee \tau)\frac{w_{m'}}{n}\right) \\ & + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n} + (\kappa_2\theta_1^2 + \kappa_1(1 \vee \tau))\frac{w}{n}.\end{aligned}$$

Now, assume that $\kappa_1 \geq \kappa$ in inequality (3.7), we get

$$\begin{aligned}(1-\epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq & \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta + \eta'}{n} \\ & + (\kappa_2\theta_1^2 + \kappa_1(1 \vee \tau))\frac{w}{n}.\end{aligned}$$

It only remains to sum up the tail bounds 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 $e^{-(w+w_m+w_{m'})}$,

$$\begin{aligned}\sum_{\substack{m\in\mathcal{M} \\ m'\in\mathcal{M}(m)}} e^{-(w+w_m+w_{m'})} &\leq e^{-w} \sum_{(m,m')\in\mathcal{M}\times\mathcal{M}} e^{-(w_m+w_{m'})} \\ &= e^{-w}\left(\sum_{m\in\mathcal{M}} e^{-w_m}\right)^2 = \Omega^2 e^{-w}\end{aligned}$$

from the Assumption K.
We then have simultaneously for all $m \in \mathcal{M}$, for all $m' \in \mathcal{M}(m)$, except on a set with probability less than $\Omega^2 e^{-w}$,

$$\begin{aligned}(1-\epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq & \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} + \frac{\delta_{\mathrm{KL}}}{n} \\ & + \frac{\eta + \eta'}{n} + \left(\kappa_2\theta_1^2 + \kappa_1(1 \vee \tau)\right)\frac{w}{n}.\end{aligned}$$

It is in particular satisfied for all $m \in \check{\mathcal{M}}$ and $m' \in \check{\mathcal{M}}(m)$, and, since $\hat{m} \in \check{\mathcal{M}}(m)$ for all $m \in \check{\mathcal{M}}$, we deduce that except on a set with probability less than $\Omega^2 e^{-w}$,

$$\begin{aligned}\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}}) \leq \frac{1}{(1-\epsilon_1)} \times \Bigg(& \inf_{m\in\check{\mathcal{M}}}\left\{\kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n}\right\} \\ & + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta + \eta'}{n} + \left(\kappa_2\theta_1^2 + \kappa_1(1 \vee \tau)\right)\frac{w}{n}\Bigg).\end{aligned}$$

By integrating over all $w > 0$, because for any non negative random variable Z and any $a > 0$, $\mathrm{E}(Z) = a \int_{z\geq 0} P(Z > az)dz$, we obtain that

$$\mathrm{E}\left(\mathrm{JKL}_{\rho}^{\otimes_n}(s^*, \hat{s}_{\hat{m}}) - \frac{1}{(1-\epsilon_1)}\left(\inf_{m\in\check{\mathcal{M}}}\left(\kappa(\beta)\,\mathrm{KL}^{\otimes n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n}\right) + \frac{\delta_{\mathrm{KL}} + \eta + \eta'}{n}\kappa_0\theta^2\right)\right)$$
$$\leq \left(\kappa_2\theta_1^2 + \kappa_1(1\vee\tau)\right)\frac{\Omega^2}{n}.$$

As δ_{KL} can be chosen arbitrary small, this implies that

$$\begin{aligned}\mathrm{E}(\mathrm{JKL}^{\otimes_n}(s^*, \hat{s}_{\hat{m}})) \leq & \frac{1}{1-\epsilon_1}\,\mathrm{E}\left(\inf_{m\in\check{\mathcal{M}}}\kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n}\right)\\ & + \frac{\eta+\eta'}{n} + (\kappa_2\theta_1^2 + \kappa_1(1\vee\tau))\frac{\Omega^2}{n}\\ \leq & C_1\,\mathrm{E}\left(\inf_{m\in\check{\mathcal{M}}}\left(\inf_{t\in S_m}\mathrm{KL}^{\otimes_n}(s^*, t)\right) + \frac{\mathrm{pen}(m)}{n}\right)\\ & + C_2(1\vee\tau)\frac{\Omega^2}{n} + \frac{\eta'+\eta}{n}\end{aligned}$$

with $C_1 = \frac{2}{1-\epsilon_1}$ and $C_2 = \kappa_2\theta_1^2 + \kappa_1$.

3.5.3 Sketch of the proof of the oracle inequality 3.3.2

To prove the Theorem 3.3.2, we have to apply the Theorem 3.5.1. Then, our model collection has to satisfy all the assumptions. Here, $m = (K, J)$. The Assumption Sep_m is true when we consider Gaussian densities. If s^* is bounded, with compact support, the assumption (3.6) is satisfied. It is also true in other particular cases. We have to look after assumption H_m and Assumption K. Here we present only the main step to prove these assumptions. All the details are in Appendix.

Assumption $\mathbf{H}_m$

We could take $\phi_m(\varpi) = \int_0^{\varpi}\sqrt{\mathcal{H}_{[.]}(\epsilon, S_m, d_H^{\otimes_n})}d\epsilon$ for all $\varpi > 0$. It could be better to consider more local version of the integrated square root entropy, but the global one is enough in this case to define the penalty. As done in Cohen and Le Pennec [CLP11], we could decompose the entropy by

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{S}_{(K,J)}^{\mathcal{B}}, d_H^{\otimes_n}) \leq \mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes_n}) + K\mathcal{H}_{[.]}(\epsilon, \mathcal{F}_J, d_H^{\otimes_n})$$

where

$$\mathcal{S}_{(K,J)}^{\mathcal{B}} = \left\{\begin{array}{l} y\in\mathbb{R}^q | x\in\mathbb{R}^p \mapsto s_{\xi}^{(K,J)}(y|x) = \sum_{k=1}^K \pi_k\varphi(y|\beta_k^{[J]}x, \Sigma_k)\\ \xi = \left\{\pi_1, \ldots, \pi_K, \beta_1^{[J]}, \ldots, \beta_K^{[J]}, \Sigma_1, \ldots, \Sigma_K\right\} \in \tilde{\Xi}_{(K,J)}\\ \tilde{\Xi}_{(K,J)} = \Pi_K \times ([-A_\beta, A_\beta]^{q\times p})^K \times ([a_\Sigma, A_\Sigma]^q)^K \end{array}\right\}$$

$$\Pi_K = \left\{(\pi_1, \ldots, \pi_K) \in (0,1)^K; \sum_{k=1}^K \pi_k = 1\right\}$$

$$\mathcal{F}_J = \left\{\varphi(.|\beta^{[J]}X, \Sigma); \beta\in[-A_\beta, A_\beta]^{q\times p}, \Sigma = \mathrm{diag}([\Sigma]_{1,1}, \ldots, [\Sigma]_{q,q}) \in [a_\Sigma, A_\Sigma]^q\right\}$$

where φ denote the Gaussian density, and $A_\beta, a_\Sigma, A_\Sigma$ are absolute constants.

Calculus for the proportions We could apply a result proved by Wasserman and Genovese in [GW00] to bound the entropy for the proportions. We get that

$$\mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes n}) \leq \log\left(K(2\pi e)^{K/2}\left(\frac{3}{\epsilon}\right)^{K-1}\right).$$

Calculus for the Gaussian The family

$$B_\epsilon(\mathcal{F}_J) = \left\{\begin{array}{l} l(y,x) = (1+\delta)^{-p^2q-3q/4}\varphi(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]}) \\ u(y,x) = (1+\delta)^{p^2q+3q/4}\varphi(y|\nu_J x, (1+\delta)B^{[2]}) \\ B^{[a]} = \mathrm{diag}(b_{i(1)}, \ldots, b_{i(q)}), \text{ with } i \text{ a permutation, for } a \in \{1,2\}, \\ \text{and} \left\{\begin{array}{l} b_l = (1+\delta)^{1-l/2}A_\Sigma, l \in \{2,\ldots,N\} \\ \forall (z,j) \in J^c, \nu_{z,j} = 0 \\ \forall (z,j) \in J, \nu_{z,j} = \sqrt{c}\delta A_\Sigma u_{z,j} \end{array}\right. \end{array}\right\} \tag{3.12}$$

is an ϵ-bracket covering for $\mathcal{F}_J$, where $u_{z,j}$ is a net for the mean, N is the number of parameters needed to recover all the variance set, $\delta = \frac{1}{\sqrt{2}(p^2q+3/4q)}\epsilon$, and $c = \frac{5(1-2^{-1/4})}{8}$.
We obtain that

$$|B_\epsilon(\mathcal{F}_J)| \leq 2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\delta^{-1-|J|};$$

and then we get

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{F}_J, d_H^{\otimes n}) \leq \log\left(2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\delta^{-1-|J|}\right).$$

Proposition 3.5.3. *Put $D_{(K,J)} = K(1+|J|)$. For all $\epsilon \in (0,1)$,*

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes n}) \leq \log(C) + D_{(K,J)}\log\left(\frac{1}{\epsilon}\right);$$

with

$$C = 2K(2\pi e)^{K/2}\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{K|J|} 3^{K-1}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)^K.$$

Determination of a function ϕ We could take

$$\phi_{(K,J)}(\varpi) = \sqrt{D_{(K,J)}}\varpi\left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right].$$

This function is non-decreasing, and $\varpi \mapsto \frac{\phi_{(K,J)}(\varpi)}{\varpi}$ is non-increasing.
The root $\varpi_{(K,J)}$ is the solution of $\phi_{(K,J)}(\varpi_{(K,J)}) = \sqrt{n}\varpi^2_{(K,J)}$. With the expression of $\phi_{(K,J)}$, we get

$$\varpi^2_{(K,J)} = \sqrt{\frac{D_{(K,J)}}{n}}\varpi\left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi_{(K,J)} \wedge 1}\right)}\right].$$

Nevertheless, we know that $\varpi^* = \sqrt{\frac{D_{(K,J)}}{n}}B(A_\beta, A_\Sigma, a_\Sigma)$ minimizes $\varpi_{(K,J)}$: we get

$$\varpi^2_{(K,J)} \leq \frac{D_{(K,J)}}{n}\left[2B^2(A_\beta, A_\Sigma, a_\Sigma) + \log\left(\frac{1}{\frac{D_{(K,J)}}{n}B^2(A_\beta, A_\Sigma, a_\Sigma) \wedge 1}\right)\right].$$

Assumption K

We want to group models by their dimension.

Lemme 3.5.4. *The quantity card*$\{(K,J) \in \mathbb{N}^* \times \mathcal{P}(\{1,\dots,q\} \times \{1,\dots,p\}), D(K,J) = D\}$ *is upper bounded by*

$$\begin{cases} 2^{pq} \text{ if } pq \leq D - q^2 \\ \left(\frac{epq}{D-q^2}\right)^{D-q^2} \text{ otherwise.} \end{cases}$$

Proposition 3.5.5. *Consider the weight family* $\{w_{(K,J)}\}_{(K,J)\in\mathcal{K}\times\mathcal{J}}$ *defined by*

$$w_{(K,J)} = D_{(K,J)} \log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq}\right).$$

Then we have $\sum_{(K,J)\in\mathcal{K}\times\mathcal{J}} e^{-w_{(K,J)}} \leq 2$.

3.6 Appendix: technical results

In this appendix, we give more details for the proofs.

3.6.1 Bernstein's Lemma

Lemme 3.6.1 (Bernstein's inequality)**.** *Let* $(X_1,\dots,X_n)$ *be independent real valued random variables. Assume that there exists some positive numbers* v *and* c *such that* $\sum_{i=1}^n \mathrm{E}(X_i^2) \leq v$, *and, for all integers* $K \geq 3$, $\sum_{i=1}^n \mathrm{E}((X_i)_+^K) \leq \frac{K!}{2} v c^{K-2}$. *Let* $S = \sum_{i=1}^n (X_i - \mathrm{E}(X_i))$. *Then, for every positive* x,

$$P(S \geq \sqrt{2vx} + cx) \leq \exp(-x).$$

3.6.2 Proof of Lemma 3.5.2

This proof is adapted from Maugis and Meynet, [MMR12]. First, let give some bounds.

Lemme 3.6.2. *Let* $\tau > 0$. *For all* $x > 0$, *consider*

$$f(x) = x\log(x)^2, \qquad h(x) = x\log(x) - x + 1, \qquad \phi(x) = e^x - x - 1.$$

Then, for all $0 < x < e^\tau$, *we get*

$$f(x) \leq \frac{\tau^2}{\phi(-\tau)} h(x).$$

To prove this, we have to show that $y \mapsto \frac{\phi(y)}{y^2}$ is non-decreasing. We omit the proof here.
We want to apply this inequality, in order to derive the Lemma 3.5.2. As $\log\left(\left|\left|\frac{s^*}{\bar{s}_m}\right|\right|_\infty\right) \leq \tau$,

$$\left|\left|\frac{s^*}{\bar{s}_m}\right|\right|_\infty \leq e^\tau;$$

and we could apply the previous inequality to $s^*/\bar{s}_m$. Indeed, for all x, for all y,

$$f\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right) \leq \frac{\tau^2}{\phi(-\tau)} h\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right).$$

Integrating with respect to the density $\bar{s}_m$, we get that

$$\begin{aligned}
&\int_{\mathbb{R}^q} \frac{s^*(y|.)}{\bar{s}_m(y|.)} \log\left(\frac{s^*(y|.)}{\bar{s}_m(y|.)}\right)^2 \bar{s}_m(y|.)dy \\
\leq &\int_{\mathbb{R}^q} \frac{\tau^2}{e^{-\tau}-\tau-1}\left(\frac{s^*(y|.)}{\bar{s}_m(y|.)}\log\frac{s^*(y|.)}{\bar{s}_m(y|.)} - \frac{s^*(y|.)}{\bar{s}_m(y|.)} + 1\right)\bar{s}_m(y|.)dy \\
\Longleftrightarrow &\frac{1}{n}\sum_{i=1}^n \int s^*(y|x_i)\log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)^2 dy \\
\leq &\frac{\tau^2}{e^{-\tau}-\tau-1}\frac{1}{n}\sum_{i=1}^n \int s^*(y|x_i)\log\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}dy.
\end{aligned}$$

It concludes the proof.

3.6.3 Determination of a net for the mean and the variance

In this subsection, we work with a Gaussian density, then $\beta \in \mathbb{R}^{q\times p}$ and $\Sigma \in \mathbb{S}_q^{++}$.

— **Step 1: construction of a net for the variance**
Let $\epsilon \in (0,1]$, and $\delta = \frac{1}{\sqrt{2}(p^2q+\frac{3}{4}q)}\epsilon$. Let $b_j = (1+\delta)^{1-\frac{j}{2}}A_\Sigma$. For $2 \leq j \leq N$, we have $[a_\Sigma, A_\Sigma] = [b_N, b_{N-1}]\bigcup\ldots\bigcup[b_3, b_2]$, where N is chosen to recover everything. We want that

$$\begin{aligned}
& a_\Sigma = (1+\delta)^{1-N/2}A_\Sigma \\
\Leftrightarrow \quad & \log\frac{a_\Sigma}{A_\Sigma} = \left(1-\frac{N}{2}\right)\log(1+\delta) \\
\Leftrightarrow \quad & N = \frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)}.
\end{aligned}$$

We want N to be an integer, then $N = \left\lceil \frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)} \right\rceil$. We get a net for the variance. We could let $B = \mathrm{diag}(b_{i(1)}, \ldots, b_{i(q)})$, close to Σ (and deterministic, independent of the values of Σ), where i is a permutation such that $b_{i(z)+1} \leq [\Sigma]_{z,z} \leq b_{i(z)}$ for all $z \in \{1,\ldots,q\}$. Remember that $\frac{b_{j+1}}{b_j} = \frac{1}{\sqrt{1+\delta}}$.

— **Step 2: construction of a net for the mean vectors**
We select only the relevant variables detected by the Lasso estimator. For $\lambda \geq 0$,

$$J_\lambda = \left\{(z,j) \in \{1,\ldots,q\}\times\{1,\ldots,p\} | \hat{\beta}^{\text{Lasso}}_{z,j}(\lambda) \neq 0\right\}.$$

Let $f = \varphi(.|\beta x, \Sigma) \in \mathcal{F}_J$.

— **Definition of the brackets**
Define the bracket by the functions l and u:

$$\begin{aligned}
l(y,x) &= (1+\delta)^{-p^2q-3q/4}\varphi\left(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]}\right); \\
u(y,x) &= (1+\delta)^{p^2q+3q/4}\varphi\left(y|\nu_J x, (1+\delta)B^{[2]}\right).
\end{aligned}$$

We have chosen i such that $[B^{[1]}]_{z,z} \leq \Sigma_{z,z} \leq [B^{[2]}]_{z,z}$ for all $z \in \{1,\ldots,q\}$.
We need to define ν such that $[l,u]$ is an ϵ-bracket for f.

— **Proof that $[l, u]$ is a bracket for f**
We are looking for a condition on ν_J to have $\frac{f}{u} \leq 1$ and $\frac{l}{f} \leq 1$.
We will use the following lemma to compute these ratios.
Lemme 3.6.3. *Let $\varphi(.|\mu_1, \Sigma_1)$ and $\varphi(.|\mu_2, \Sigma_2)$ be two Gaussian densities. If their variance matrices are assumed to be diagonal, with $\Sigma_a = \mathrm{diag}([\Sigma_a]_{1,1}, \ldots, [\Sigma_a]_{q,q})$ for $a \in \{1, 2\}$, such that $[\Sigma_2]_{z,z} > [\Sigma_1]_{z,z} > 0$ for all $z \in \{1, \ldots, q\}$, then, for all $y \in \mathbb{R}^q$,*

$$\frac{\varphi(y|\mu_1, \Sigma_1)}{\varphi(y|\mu_2, \Sigma_2)} \leq \prod_{z=1}^{q} \frac{\sqrt{[\Sigma_2]_{z,z}}}{\sqrt{[\Sigma_1]_{z,z}}} e^{\frac{1}{2}(\mu_1-\mu_2)^t \mathrm{diag}\left(\frac{1}{[\Sigma_2]_{1,1}-[\Sigma_1]_{1,1}}, \ldots, \frac{1}{[\Sigma_2]_{q,q}-[\Sigma_1]_{q,q}}\right)(\mu_1-\mu_2)}.$$

For the ratio $\frac{f}{u}$ we get:

$$\begin{aligned}
\frac{f(y|x)}{u(y,x)} =& \frac{1}{(1+\delta)^{p^2q+3q/4}} \frac{\varphi(y|\beta x, \Sigma)}{\varphi(y|\nu_J x, (1+\delta)B^{[2]})} \\
\leq& \frac{1}{(1+\delta)^{p^2q+3q/4}} \prod_{z=1}^{q} \frac{b_z}{[\Sigma]_{z,z}} (1+\delta)^{q/2} \times e^{\frac{1}{2}(\beta x-\nu_J x)^t((1+\delta)B^{[2]}-\Sigma)^{-1}(\beta x-\nu_J x)} \\
\leq& (1+\delta)^{p^2q-q/4}(1+\delta)^{q/4} e^{\frac{1}{2}(\beta x-\nu_J x)^t(\delta B^{[2]})^{-1}(\beta x-\nu_J x)} \\
\leq& (1+\delta)^{p^2q} e^{\frac{1}{2\delta}(\beta x-\nu_J x)^t[B^{[2]}]^{-1}(\beta x-\nu_J x)}.
\end{aligned} \tag{3.13}$$

For the ratio $\frac{l}{f}$ we get:

$$\begin{aligned}
\frac{l(y,x)}{f(y|x)} =& \frac{1}{(1+\delta)^{p^2q+3q/4}} \frac{\varphi(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]})}{\varphi(y|\beta x, \Sigma)} \\
\leq& \frac{1}{(1+\delta)^{p^2q+3q/4}} \prod_{z=1}^{q} \frac{\Sigma_{z,z}}{b_z} (1+\delta)^{q/8} \times e^{\frac{1}{2}(\beta x-\nu_J x)^t(\Sigma-B^{[1]})^{-1}(\beta x-\nu_J x)} \\
\leq& (1+\delta)^{-p^2q-3q/8}(1+\delta)^{q/4} \times e^{\frac{1}{2}(\beta x-\nu_J x)^t((1-(1+\delta)^{-1/4})B^{[1]})^{-1}(\beta x-\nu_J x)} \\
\leq& (1+\delta)^{-p^2q-3q/8} e^{\frac{1}{2(1-(1+\delta)^{-1/4})}(\beta x-\nu_J x)^t[B^{[1]}]^{-1}(\beta x-\nu_J x)}.
\end{aligned} \tag{3.14}$$

We want to bound the ratios (3.13) and (3.14) by 1. Put $c = \frac{5(1-2^{-1/4})}{8}$, and develop these calculus. A necessary condition to obtain this bound is

$$||\beta x - \nu_J x||_2^2 \leq pq\delta^2(1-2^{-1/4})A_\Sigma^2.$$

Indeed, we want

$$\begin{aligned}
(1+\delta)^{-p^2q-3q/8} e^{\frac{1}{2(1-(1+\delta)^{-1/4})}(\beta x-\nu_J x)^t[B^{[2]}]^{-1}(\beta x-\nu_J x)} &\leq 1 \\
(1+\delta)^{-p^2q} e^{\frac{1}{2\delta A_\Sigma}(\beta x-\nu_J x)^t[B^{[1]}]^{-1}(\beta x-\nu_J x)} &\leq 1;
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
||\beta x - \nu_J x||_2^2 &\leq p^2q\frac{\delta^2}{2}A_\Sigma^2; \\
||\beta x - \nu_J x||_2^2 &\leq \left(p^2q + \frac{3}{4}q\right)\delta^2(1-2^{-1/4})A_\Sigma.
\end{aligned}$$

As $||\beta x - \nu_J x||_2^2 \leq p||\beta - \nu_J||_2^2||x||_\infty$, and $x \in [0,1]^p$, we need to get $||\beta - \nu_J||_2^2 \leq pq\delta^2(1-2^{-1/4})A_\Sigma^2$ to have the wanted bound. Put

$$U := \mathbb{Z} \cap \left[\left\lfloor \frac{-A_\beta}{\sqrt{c}\delta A_\Sigma} \right\rfloor, \left\lfloor \frac{A_\beta}{\sqrt{c}\delta A_\Sigma} \right\rfloor\right].$$

For all $(z,j) \in J$, choose

$$u_{z,j} = \underset{v_{z,j} \in U}{\operatorname{argmin}} \left| \beta_{z,j} - \sqrt{c}\delta A_\Sigma v_{z,j} \right|. \tag{3.15}$$

Define ν by

$$\begin{aligned} &\text{for all } (z,j) \in J^c, \nu_{z,j} = 0; \\ &\text{for all } (z,j) \in J\ , \nu_{z,j} = \sqrt{c}\delta A_\Sigma u_{z,j}. \end{aligned}$$

Then, we get a net for the mean vectors.

— **Proof that** $d_H(l,u) \leq \epsilon$

We will work with the Hellinger distance.

$$\begin{aligned} d_H^2(l,u) &= \frac{1}{2}\int_{\mathbb{R}^q} (\sqrt{l} - \sqrt{u})^2 \\ &= \frac{1}{2}\int_{\mathbb{R}^q} l + u - 2\sqrt{lu} \\ &= \frac{1}{2}\left[(1+\delta)^{-p^2q-3q/4} + (1+\delta)^{p^2q+3q/4}\right] - \int_{\mathbb{R}^q} \sqrt{\varphi_l \varphi_u} \\ &= \frac{1}{2}\left[(1+\delta)^{-p^2q-3q/4} + (1+\delta)^{p^2q+3q/4}\right] \\ &- \left(\prod_{z=1}^{q} \frac{\sqrt{2b_{i(z)+1}b_{i(z)}}(1+\delta)^{1/2}(1+\delta)^{-1/8}\,{}^2}{(1+\delta)b_{i(z)+1} + (1+\delta)^{-1/4}b_{i(z)}}\right)^{1/2} * 1. \end{aligned}$$

We have used the following lemma:

Lemme 3.6.4. *The Hellinger distance of two Gaussian densities with diagonal variance matrices is given by the following expression:*

$$\begin{aligned} &d_H^2(\varphi(.|\mu_1,\Sigma_1), \varphi(.|\mu_2,\Sigma_2)) \\ =&2 - 2\left(\prod_{z=1}^{q} \frac{2\sqrt{[\Sigma_1]_{z,z}[\Sigma_2]_{z,z}}}{[\Sigma_1]_{z,z} + [\Sigma_2]_{z,z}}\right)^{1/2} \\ &\times \exp\left\{-\frac{1}{4}(\mu_1-\mu_2)^t \operatorname{diag}\left(\left(\frac{1}{[\Sigma_1]_{z,z}^2 + [\Sigma_2]_{z,z}^2}\right)_{z\in\{1,\dots,q\}}\right)(\mu_1-\mu_2)\right\} \end{aligned}$$

As $b_{i(z)+1} = (1+\delta)^{-1/2}b_{i(z)}$, we get that

$$\begin{aligned} 2\frac{(1+\delta)^{3/8}b_{i(z)}}{b_{i(z)+1}\left[(1+\delta)^{-1/4} + (1+\delta)^{1/2}(1+\delta)\right]} &= 2\frac{(1+\delta)^{5/8}}{(1+\delta)^{-1/4} + (1+\delta)^{3/2}} \\ &= \frac{2}{(1+\delta)^{-7/8} + (1+\delta)^{7/8}}. \end{aligned}$$

Then

$$\begin{aligned} d_H^2(l,u) =& \frac{1}{2}\left[(1+\delta)^{-(p^2q+3q/4)} + (1+\delta)^{p^2q+3q/4}\right] - \left(\frac{2}{(1+\delta)^{-7/8} + (1+\delta)^{7/8}}\right)^{q/2} \\ =& \cosh((p^2q+3q/4)\log(1+\delta)) - 2\cosh(7/8\log(1+\delta))^{-q/2} \\ =& \cosh((p^2q+3q/4)\log(1+\delta)) - 1 + 1 - 2^{-q/2}\cosh(7/8\log(1+\delta))^{-q/2}. \end{aligned}$$

We want to apply the Taylor formula to $f(x) = \cosh(x) - 1$ to obtain an upper bound, and to $g(x) = 1 - 2^{-q/2}\cosh(x)^{-q/2}$. Indeed, there exists c such that, on the good interval, $f(x) \leq \cosh(c)\frac{x^2}{2}$ and $g(x) \leq q^2\frac{x^2}{2}$. Then, and because $\log(1+\delta) \leq \delta$,

$$\begin{aligned} d_H^2(l,u) &\leq \cosh((p^2q + 3q/4)\log(1+\delta)) - 2\cosh(7/8\log(1+\delta))^{-q/2} \\ &\leq (p^2q + 3q/4)^2\delta^2\left(\cosh(\alpha) + \frac{49}{128}\right) \\ &\leq 2(p^2q + 3q/4)^2\delta^2 \leq \epsilon^2; \end{aligned}$$

where $\epsilon \geq \sqrt{2}(p^2q + \frac{3}{4}q)\delta$.

— **Step 3: Upper bound of the number of ϵ-brackets for $\mathcal{F}_J$.**
From Step 1 and Step 2, the family

$$B_\epsilon(\mathcal{F}_J) = \left\{ \begin{array}{l} l(y,x) = (1+\delta)^{-(p^2q+3q/4)}\varphi(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]}) \\ u(y,x) = (1+\delta)^{p^2q+3q/4}\varphi(y|\nu_J x, (1+\delta)B^{[2]}) \\ B^{[a]} = \mathrm{diag}(b^{[a]}_{i(1)}, \ldots, b^{[a]}_{i(q)}) \text{ where } i_a \text{ is a permutation, for } a \in \{1,2\}, \\ \text{with} \left\{ \begin{array}{l} b^{[a]}_{i(z)} = (1+\delta)^{1-i_a(z)/2}A_\Sigma \text{ for all } z \in \{1,\ldots,q\} \\ \forall (z,j) \in J^c, \nu_{z,j} = 0 \\ \forall (z,j) \in J, \nu_{z,j} = \sqrt{c}\delta A_\Sigma u_{z,j} \end{array} \right. \end{array} \right\} \tag{3.16}$$

is an ϵ-bracket for $\mathcal{F}_J$, for $u_{z,j}$ defined by (3.15). Therefore, an upper bound of the number of ϵ-brackets necessary to cover $\mathcal{F}_J$ is deduced from an upper bound of the cardinal of $B_\epsilon(\mathcal{F}_J)$.

$$|B_\epsilon(\mathcal{F}_J)| \leq \sum_{l=2}^{N} \prod_{(z,j)\in J} \left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right) \leq \left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right)^{|J|} \sum_{l=2}^{N} 1 \leq \left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right)^{|J|}(N-1).$$

As $N \leq \frac{2(A_\Sigma/a_\Sigma + 1/2)}{\delta}$, we get

$$|B_\epsilon(\mathcal{F}_J)| \leq 2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\delta^{-1-|J|}.$$

3.6.4 Calculus for the function ϕ

From the Proposition 3.5.3, we obtain, for all $\varpi > 0$,

$$\int_0^\varpi \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes n})}d\epsilon \leq \varpi\sqrt{\log(C)} + \sqrt{D_{(K,J)}}\int_0^{\varpi\wedge 1}\sqrt{\log\left(\frac{1}{\epsilon}\right)}d\epsilon \tag{3.17}$$

We need to control $\int_0^\varpi \sqrt{\log\left(\frac{1}{\epsilon}\right)}d\epsilon$, which is done in Maugis-Rabusseau and Meynet ([MMR12]).

Lemme 3.6.5. *For all $\varpi > 0$,*

$$\int_0^\varpi \sqrt{\log\left(\frac{1}{\epsilon}\right)}d\epsilon \leq \varpi\left[\sqrt{\pi} + \sqrt{\log\left(\frac{1}{\varpi}\right)}\right].$$

Then, according to (3.17),

$$\begin{aligned}\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes n})}d\epsilon \leq & \varpi\sqrt{\log(C)} + \sqrt{D_{(K,J)}}(\varpi \wedge 1)\left[\sqrt{\pi} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right] \\ \leq & \varpi\sqrt{D_{(K,J)}}\left[\sqrt{\frac{\log(C)}{D_{(K,J)}}} + \sqrt{\pi} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right]\end{aligned}$$

Nevertheless,

$$\begin{aligned}\log(C) \leq & \log(2) + \log(K) + \frac{K}{2}\log(2\pi e) \\ & + K|J|\log\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right) + K\log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right) + (K-1)\log(3) \\ \leq & D_{(K,J)}\Bigg[\log(2) + \log(\sqrt{2\pi e}) + 1 + \log(3) \\ & + \log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right) + \log\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)\Bigg] \\ \leq & D_{(K,J)}\left[1 + \log\left(\frac{A_\beta}{A_\Sigma}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\right) + \log\left(2^{5/2}3\sqrt{\frac{\pi e}{c}}\right)\right].\end{aligned}$$

Then

$$\begin{aligned}& \int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes n})}d\epsilon \\ \leq & \varpi\sqrt{D_{(K,J)}}\Bigg[\sqrt{1 + \log\left(\frac{A_\beta}{A_\Sigma}\left(\frac{A_\Sigma}{A_\Sigma} + \frac{1}{2}\right)\right) + \log\left(2^{5/2}3\sqrt{\frac{\pi e}{c}}\right)} \\ & + \sqrt{\pi} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\Bigg] \\ \leq & \varpi\sqrt{D_{(K,J)}}\left[1 + \sqrt{\log\left(\frac{A_\beta}{A_\Sigma}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\right)} + a + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right] \\ \leq & \varpi\sqrt{D_{(K,J)}}\left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right];\end{aligned}$$

with

$$B(A_\beta, A_\Sigma, a_\Sigma) = 1 + \sqrt{\log\left(\frac{A_\beta}{A_\Sigma}\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\right)} + a;$$

and $a = \sqrt{\pi} + \sqrt{\log(2^{5/2}3\sqrt{\frac{\pi e}{c}})}$.

3.6.5 Proof of the Proposition 3.5.5

We are interested in $\sum_{(K,J)\in\mathcal{K}\times\mathcal{J}} e^{-w_{(K,J)}}$. Considering

$$w_{(K,J)} = D_{(K,J)}\log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq}\right),$$

we could group models by their dimensions to compute this sum. Denote by C_D the cardinal of models of dimension D.

$$\begin{aligned}
&\sum_{\substack{K\in\mathbb{N}^*\\ J\in\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})}} e^{-D_{(K,J)}\log\left(\frac{4epq}{(D_{(K,J)}-q^2)\wedge pq}\right)} = \sum_{D\geq 1} C_D e^{-D\log\left(\frac{4epq}{(D-q^2)\wedge pq}\right)}\\
&= \sum_{D=1}^{pq+q^2} e^{-D\log\left(\frac{4epq}{(D-q^2)}\right)}\left(\frac{epq}{D-q^2}\right)^{D-q^2} + \sum_{D=pq+q^2+1}^{+\infty} e^{-D\log\left(\frac{4epq}{pq}\right)}2^{pq}\\
&= \sum_{D=1}^{pq+q^2} 4^{-D}\left(\frac{epq}{D-q^2}\right)^{-q^2} + \sum_{D=pq+q^2+1}^{+\infty} e^{-D(\log(4)+1)+pq\log(2)}\\
&\leq \sum_{D=1}^{pq+q^2} 2^{-D} + \sum_{D=pq+q^2+1}^{+\infty} 2^{-D} = 2.
\end{aligned}$$

3.6.6 Proof of the Lemma 3.5.4

We know that $D_{(K,J)} = K-1+|J|K+Kq^2$. Then,

$$\begin{aligned}
C_D &= \mathrm{card}\{(K,J)\in\mathbb{N}^*\times\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\}), D(K,J)=D\}\\
&\leq \sum_{K\in\mathbb{N}^*}\sum_{\substack{1\leq z\leq q\\ 1\leq j\leq p}}\binom{pq}{|J|}\mathbb{1}_{K(|J|+q^2+1)-1=D}\\
&\leq \sum_{|J|\in\mathbb{N}^*}\binom{pq}{|J|}\mathbb{1}_{|J|\leq pq\wedge(D-q^2)}.
\end{aligned}$$

If $pq < D-q^2$,

$$\sum_{|J|>0}\binom{pq}{|J|}\mathbb{1}_{|J|\leq pq\wedge(D-q^2)} = 2^{pq}.$$

Otherwise, according to the Proposition 2.5 in Massart, [Mas07],

$$\sum_{|J|>0}\binom{pq}{|J|}\mathbb{1}_{|J|\leq pq\wedge(D-q^2)} \leq f(D-q^2)$$

where $f(x) = (epq/x)^x$ is an increasing function on $\{1,\dots,pq\}$. As pq is an integer, we get the result.

Chapter 4

An oracle inequality for the Lasso-Rank procedure

Contents

3.1 Introduction **91**
3.2 The Lasso-MLE procedure **92**
3.2.1 Gaussian mixture regression model 93
3.2.2 The Lasso-MLE procedure 94
3.2.3 Why refit the Lasso estimator? 94
3.3 An oracle inequality for the Lasso-MLE model **95**
3.3.1 Notations and framework 95
3.3.2 Oracle inequality 96
3.4 Numerical experiments **98**
3.4.1 Simulation illustration 98
3.4.2 Real data 99
3.5 Tools for proof **101**
3.5.1 General theory of model selection with the maximum likelihood estimator. 101
3.5.2 Proof of the general theorem 103
3.5.3 Sketch of the proof of the oracle inequality 3.3.2 107
Assumption H_m 107
Assumption K 109
3.6 Appendix: technical results **109**
3.6.1 Bernstein's Lemma 109
3.6.2 Proof of Lemma 3.5.2 109
3.6.3 Determination of a net for the mean and the variance 110
3.6.4 Calculus for the function ϕ 113
3.6.5 Proof of the Proposition 3.5.5 114
3.6.6 Proof of the Lemma 3.5.4 115

In this chapter, we focus on a theoretical result for the Lasso-Rank procedure. Indeed, we get the same kind of results as in the previous chapter, with rank constraint on the estimator. We get a theoretical penalty for which the model selected by a penalized criterion, among the collection, satisfies an oracle inequality. We also illustrate in more details benefits of this procedure with simulated and benchmark dataset including rank structure.

4.1 Introduction

The multivariate response regression model

$$Y = \beta X + \epsilon$$

postulates a linear relationship between Y, the $q \times n$ matrix containing q responses for n subjects, and X, the $p \times n$ matrix on p predictor variables. The term ϵ is an $q \times n$ matrix with independent columns, $\epsilon_i \sim \mathcal{N}_q(0, \Sigma)$ for all $i \in \{1, \dots, n\}$. The unknown $q \times p$ coefficient matrix β needs to be estimate.
In a more general way, we could use finite mixture of linear model, which models the relationship between response and predictors, arising from different subpopulations: if the variable Y, conditionally to X, belongs to the cluster k, there exists $\beta_k \in \mathbb{R}^{q \times p}$ and $\Sigma_k \in \mathbb{S}_q^{++}$ such that $Y = \beta_k X + \epsilon$, with $\epsilon \sim \mathcal{N}_q(0, \Sigma_k)$.
If we use this model to deal with high-dimensional data, the number of variables could be quickly much larger than the sample size, because and predictors and response variables could be high-dimensional. To solve this problem, we have to reduce the parameter set dimension.
One way to cope the dimension problem is to select relevant variables, in order to reduce the number of unknowns. Indeed, all the information should not be interesting for the clustering, and could even be harmful. In a density estimation way, we could cite Pan and Shen, in [PS07], who focus on mean variable selection, Meynet and Maugis in [MMR12] who extend their procedure in high-dimension, Zhou et al., in [ZPS09], who use the Lasso estimator to regularize Gaussian mixture model with general covariance matrices, Sun et al., in [SWF12], who propose to regularize the k-means algorithm to deal with high-dimensional data, Guo et al, in [GLMZ10], who propose a pairwise variable selection method, among others.
In a regression framework, we could use the Lasso estimator, introduced by Tibshirani in [Tib96], which is a sparse estimator. It penalizes the maximum likelihood estimator by the ℓ_1-norm, which achieves the sparsity, as the ℓ_0-penalty, but leads also to a convex optimization. Because we work with the multivariate linear model, to deal with the matrix structure, we could prefer the group-Lasso estimator, variables grouped by columns, which selects columns rather than coefficients. This estimator was introduced by Zhou and Zhu in [ZZ10] in the general case. If we select $|J|$ columns among the p possible, we have to estimate $|J|q$ coefficients rather than pq for A, which could be smaller than nq if $|J|$ is smaller enough.
Another estimator which reduces the dimension, is the low rank estimator. Introduced by Izenman in [Ize75] in the linear model, and more used the last decades, with among others Bunea et al. in [BSW12] and Giraud in [Gir11], the regression matrix is estimated by a matrix of rank R, $R < p \wedge q$. Then, we have to estimate $R(p + q - R)$ coefficients, which could be smaller than nq.
In this chapter, we have chosen to mix these two estimators to provide a sparse and low rank estimator in mixture models. This method was introduced by Bunea et al. in [BSW12], in the case of linear model and known noise covariance matrix. They present different ways, more or

less computational, with more or less good results in theory. They get an oracle inequality, which say that, among a model collection, they are able to choose an estimator with true rank and true relevant variables. For this model, Ma and Sun in [MS14] get a minimax lower bound, which precise that they attain nearly optimal rates of convergence adaptively for square Schatten norm losses.

In this chapter, we consider finite mixture of K linear models in high-dimension. This model is studied in details by Städler et al. for real response variable in [SBG10], and by Devijver for multivariate response variable in [Dev14c]. We will estimate β_k for all $k \in \{1, \dots, K\}$ by a column sparse and low rank estimator. The Lasso estimator is used to select variables, whereas we refit the estimation by a low rank estimator, restricted on relevant variables. The procedure we propose is based on a modeling that recasts variable selection, rank selection, and clustering problems into a model selection problem. This procedure is developed in [Dev14c], with methodology, computational issues, simulations and data analysis. In this chapter, we focus on theoretical point of view, and developed simulations and data analysis for the low rank issue. We construct a model collection, with models more or less sparse, and with vector of ranks varying with values more or less small. Among this collection, we have to select a model. We use the slope heuristic, which is a non-asymptotic criterion. In a theoretical way, in this chapter, we get an oracle inequality for the collection constructed by our procedure, which makes a performance comparison between our selected model and the oracle for a specified penalty.

This result is an extension of the work of Bunea et al. in [BSW12], to mixture models and with unknown covariance matrices $(\Sigma_k)_{1 \leq k \leq K}$. They ensure that mixing sparse estimator and low rank matrix could be interesting. Indeed, whereas we have to estimate $q \times p$ coefficients in each cluster for the regression matrix, we get only $R(|J| + q - R)$ unknown variables, which could be smaller than the number of observations nq if $|J|$ and R are small. Even if the oracle inequality we get in this chapter is an extension of Bunea et al. result, we use a really different way to prove it. Considering the model collection constructed, we want to select a model as good as possible. For that, we use the slope heuristic, which leads to construct a penalty, proportional to the dimension, and we select the model minimizing the penalized log-likelihood. Theoretically, we construct also a penalty, proportional to the dimension (up to a logarithm term). We provide an oracle inequality which compares, up to a constant, the Jensen-Kullback-Leibler divergence between our model and the true model to the Kullback-Leibler divergence between the oracle and the true model. Then, in estimation term, we do as well as possible. This oracle inequality is deduced from a general model selection theorem for maximum likelihood estimator of Massart developed in [Mas07]. Controlling the bracketing entropy of models, we could prove the result. Remark that we work in a regression framework, then we rather use an extension of this theorem proved in Cohen and Le Pennec article [CLP11]. As our model collection is random, constructed by the Lasso estimator, we rather use an extension of this theorem proved in [Dev14b]. To illustrate this procedure, in a computational way, we validate it on simulated dataset, and benchmark dataset. If the data have a low rank structure, we could easily find it with our methodology.

This chapter is organized as follows. In the Section 4.2, we describe the finite mixture regression model used in this procedure, and the main step of the procedure. In the Section 4.3, we present the main result of this chapter, which is an oracle inequality for the procedure proposed. Finally, in Section 4.4, we illustrate the procedure on simulated and benchmark dataset. Proof details of the oracle inequality are given in Appendix.

4.2 The model and the model collection

We introduce our procedure of estimation by sparse and low rank matrix in the linear model, in Section 4.2.1, and extend it in Section 4.2.2 for mixture models. We also present the main algorithm used in this context, and we describe the procedure we propose in Section 4.2.4.

4.2.1 Linear model

We consider the observations $(x_i, y_i)_{1\leq i\leq n}$ which realized random variables (X, Y), satisfying the linear model

$$Y = \beta X + \epsilon;$$

where $Y \in \mathbb{R}^q$ are the responses, $X \in \mathbb{R}^p$ are the regressors, $\beta \in \mathbb{R}^{q\times p}$ is an unknown matrix, and $\epsilon \in \mathbb{R}^q$ are random errors, $\epsilon \sim \mathcal{N}_q(0, \Sigma)$, with $\Sigma \in \mathbb{S}_q^{++}$ a symmetric positive definite matrix. We will work in high-dimension, then $q \times p$ could be larger than the number of observations nq. We will construct an estimator which is sparse and low rank for β to cope with the high-dimension issue. Moreover, to reduce the covariance matrix dimension, we compute a diagonal estimator of Σ. The procedure we propose could be explained into two steps. First, we estimate the relevant columns of β thanks to the Lasso estimator, for $\lambda > 0$, using the estimator

$$\hat{\beta}^{\text{Lasso}}(\lambda) = \underset{\beta\in\mathbb{R}^{q\times p}}{\text{argmin}} \left\{ ||Y - \beta X||_2^2 + \lambda ||\beta||_1 \right\}; \tag{4.1}$$

where $||\beta||_1 = \sum_{j=1}^p \sum_{z=1}^q |\beta_{z,j}|$. We assume that the covariance matrix is unknown.
For $\lambda > 0$, computing the Lasso estimator of $\hat{\beta}^{\text{Lasso}}(\lambda)$, we could deduce the relevant columns. Restricted to these relevant columns, in the second step of the procedure, we compute a low rank estimator of β, saying of rank at most R. Indeed, as explained in Giraud in [Gir11], we restrict the maximum likelihood estimator to have a rank at most R, keeping only the R biggest singular values in the corresponding decomposition. We get an explicit formula.
This two steps procedure leads to an estimator of β which is sparse and has a low rank. We have also reduced the dimension into two ways. We refit the covariance matrix estimator by the maximum likelihood estimator.
This estimator is studied in Bunea et al. in [BSW12], in method 3. Let extend it in mixture models.

4.2.2 Mixture model

We observe n independent couples $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{1\leq i\leq n}$ of random variables (X, Y), with $Y \in \mathbb{R}^q$ and $X \in \mathbb{R}^p$. We will estimate the unknown conditional density s^* by a multivariate Gaussian mixture regression model. In our model, if the observation i belongs to the cluster k, we assume that there exists $\beta_k \in \mathbb{R}^{q\times p}$, and $\Sigma_k \in \mathbb{S}_q^{++}$ such that $y_i = \beta_k x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}_q(0, \Sigma_k)$.
Thus, the random response variable $Y \in \mathbb{R}^q$ will be explained by a set of explanatory variables, written $X \in \mathbb{R}^p$, through a mixture of linear regression-type model. Give more precisions on the assumptions.

— The variables Y_i are independent conditionally to X_i, for all $i \in \{1, \ldots, n\}$;

— we let $Y_i|X_i = x_i \sim s_\xi^K(y|x_i)dy$, with

$$s_\xi^K(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right) \tag{4.2}$$

$$\xi = (\pi_1, \dots, \pi_k, \beta_1, \dots, \beta_k, \Sigma_1, \dots, \Sigma_K) \in \Xi_K = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right)$$

$$\Pi_K = \left\{(\pi_1, \dots, \pi_K); \pi_k > 0 \text{ for } k \in \{1, \dots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$$

$\mathbb{S}_q^{++}$ is the set of symmetric positive definite matrices on $\mathbb{R}^q$.

For all $k \in \{1, \dots, K\}, \beta_k$ is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k. The π_ks are the mixture proportions. For all $k \in \{1, \dots, K\}$, for all $z \in \{1, \dots, q\}$, $[\beta_k^t x]_z = \sum_{j=1}^p [\beta_k]_{z,j} x_j$ is the zth component of the mean of the mixture component k for the conditional density $s_\xi^K(.|x)$.
To detect the relevant variables, we generalize the Lasso estimator defined by (4.1) for mixture models. Indeed, we penalize the empirical contrast by an ℓ_1-penalty on the mean parameters proportional to

$$||P_k\beta_k||_1 = \sum_{j=1}^p \sum_{z=1}^q |(P_k\beta_k)_{z,j}|,$$

where the Cholesky decomposition $P_k^t P_k = \Sigma_k^{-1}$ defines P_k for all $k \in \{1, \dots, K\}$. Then, we will consider

$$\hat{\xi}_K^{\text{Lasso}}(\lambda) = \underset{\xi=(\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\Sigma})\in\Xi_K}{\text{argmin}} \left\{-\frac{1}{n}\sum_{i=1}^n \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^K \pi_k ||P_k\beta_k||_1\right\}. \tag{4.3}$$

Remark that the penalty take into account the mixture weight. To reduce the dimension and simplify computations, we will estimate Σ_k by a diagonal matrix, thus P_k will be also estimated by a diagonal matrix, for all $k \in \{1, \dots, K\}$.
As in Section 4.2.1, we refit the estimator, restricted on releant columns, with low rank estimator. In Section 4.2.3, we will extend the EM algorithm to compute those two estimators.

4.2.3 Generalized EM algorithm

In a computational way, we will use two generalized EM algorithms, in order to deal with high-dimensional data and get a sparse and low rank estimator. Give some details about those algorithms.
Initially, the EM algorithm was introduced by Dempster et al. in [DLR77]. It alternates two steps until convergence, an expectation step to cluster data, and a maximization step to update estimation.
In our procedure, we want to know which columns are relevant, then we have to compute (4.3), and we want to refit the estimators by a maximum likelihood under low rank constraint estimator.
From the Lasso estimator (4.3), we could use a generalization of the EM algorithm described in [Dev14c]. From the estimate of β, we could deduce which columns are relevant.
The second algorithm we use leads to determine β_k restricted on relevant columns, for all $k \in \{1, \dots, K\}$, with rank R_k. We alternate two steps, E-step and M-step, until relative convergence of the parameters and of the likelihood. We restrict the dataset to relevant columns, and construct an estimator of size $q \times |J|$ rather than $q \times p$, where β_k has for rank R_k, for all $k \in \{1, \dots, K\}$. Explain the both steps at iteration $(ite) \in \mathbb{N}^*$.

— E-step: compute for $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$, the expected value of the log-likelihood function,

$$\tau_{i,k} = E_{\theta^{(\text{ite})}}([Z_i]_k|Y) = \frac{\gamma_k}{\sum_{l=1}^{K} \gamma_l}$$

where

$$\gamma_l = \frac{\pi_l^{(\text{ite})}}{\det \Sigma_l^{(\text{ite})}} \exp^{-\frac{1}{2}\left(y_i - \beta_l^{(\text{ite})} x_i\right)^t (\Sigma_l^{-1})^{(\text{ite})} \left(y_i - \beta_l^{(\text{ite})} x_i\right)}$$

for $l \in \{1, \dots, K\}$, and Z_i is the component-membership variable for an observation i.

— M-step:

— To get estimation in linear model, we assign each observation in its estimated cluster, by the MAP principle. We could compute this thanks to the E-step. Indeed,y_i is assigned to the component number $\underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \tau_{i,k}$.

— Then, we could define $\tilde{\beta}_k^{(\text{ite})} = (\mathbf{x}_{|k}^t \mathbf{x}_{|k})^{-1} \mathbf{x}_{|k}^t \mathbf{y}_{|k}$, in which $\mathbf{x}_{|k}$ and $\mathbf{y}_{|k}$ are the sample restriction to the cluster k. We decompose $\tilde{\beta}_k^{(\text{ite})}$ in singular values such that $\tilde{\beta}_k^{(\text{ite})} = USV^t$ with $S = \text{diag}(s_1, \dots, s_q)$ and $s_1 \geq s_2 \geq \dots \geq s_q$ the singular values. Then, the estimator $\hat{\beta}_k^{(\text{ite})}$ is defined by $\hat{\beta}_k^{(\text{ite})} = US_{R_k}V^t$ with $S_{R_k} = \text{diag}(s_1, \dots, s_{R_k}, 0, \dots, 0)$. We do it for all $k \in \{1, \dots, K\}$.

4.2.4 The Lasso-Rank procedure

The procedure we propose, which is particularly interesting in high-dimension, could be decomposed into three main steps. First, we construct a model collection, with models more or less sparse and with more or less components. Then, we refit estimations with the maximum likelihood estimator, under rank constraint. Finally, we select a model with the slope heuristic.

Model collection construction Fix $K \in \mathcal{K}$. To get various relevant columns, we construct a data-driven grid of regularization parameters G_K, coming from EM algorithm formula. See [Dev14c] for more details. For each $\lambda \in G_K$, we could compute the Lasso estimator (4.3), and deduce relevant variables set, denoted by $J_{(K,\lambda)}$. Then, varying $\lambda \in G_K$, and $K \in \mathcal{K}$, we construct $\mathcal{J} = \cup_{K \in \mathcal{K}} \cup_{\lambda \in G_K} J_{(K,\lambda)}$ the set of relevant variables sets.

Refitting We could also define a low rank estimator $\hat{s}^{(K,J,R)}$ restricted to relevant variables detected by the Lasso estimator, indewed by J.
From this procedure, we construct a model with K clusters, $|J|$ relevant columns and matrix of regression coefficients of ranks $R \in \mathbb{N}^K$, as described by the next model $\mathcal{S}_{(K,J,R)}$.

$$\mathcal{S}_{(K,J,R)} = \{y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto s_\xi^{(K,J,R)}(y|x)\} \tag{4.4}$$

where

$$\begin{aligned}
s_\xi^{(K,J,R)}(y|x) &= \sum_{k=1}^{K} \frac{\pi_k \det(P_k)}{(2\pi)^{q/2}} \exp\left(-\frac{1}{2}(y - (\beta_k^{R_k})^{[J]}x)^t \Sigma_k^{-1} (y - (\beta_k^{R_k})^{[J]}x)\right); \\
\xi &= (\pi_1, \dots, \pi_K, \beta_1^{R_1}, \dots, \beta_K^{R_K}, \Sigma_1, \dots, \Sigma_K) \in \Xi_{(K,J,R)}; \\
\Xi_{(K,J,R)} &= \Pi_K \times \Psi_{(K,J,R)} \times (\mathbb{S}_q^{++})^K; \\
\Psi_{(K,J,R)} &= \left\{((\beta_1^{R_1})^{[J]}, \dots, (\beta_K^{R_K})^{[J]}) \in (\mathbb{R}^{q \times p})^K | \text{ for all } k \in \{1, \dots, K\}, \text{Rank}(\beta_k) = R_k\right\}.
\end{aligned}$$

Varying $K \in \mathcal{K} \subset \mathbb{N}^*, J \in \mathcal{J} \subset \mathcal{P}(\{1,\dots,p\})$, and $R \in \mathcal{R} \subset \{1,\dots,|J| \wedge q\}^K$, we get a model collection with various number of components, relevant columns and matrix of regression coefficients.

Model selection Among this model collection, during the last step, a model has to be selected. As in Maugis and Michel in [MM11b], and in Maugis and Meynet in [MMR12], among others, a non asymptotic penalized criterion is used. The slope heuristic was introduced by Birgé and Massart in [BM07], and developed in practice by Baudry et al. in [BMM12] with the Capushe package. To use it in our context, we have to extend theoretical result to determine the penalty shape in the high-dimensional context, with a random model collection, in a regression framework. The main result is described in the next section, whereas proof details are given in Appendix.

4.3 Oracle inequality

In a theoretical point of view, we want to ensure that the slope heuristic which penalizes the log-likelihood with the model dimension will select a good model. We follow the approach developed by Massart in [Mas07] which consists of defining a non-asymptotic penalized criterion, leading to an oracle inequality. In the context of regression, Cohen and Le Pennec, in [CLP11], and Devijver in [Dev14b], propose a general model selection theorem for maximum likelihood estimation. The result we get is a theoretical penalty, for which the model selected is as good as the best one, according to the Kullback-Leibler loss.

4.3.1 Framework and model collection

Among the model collection constructed by the procedure developed in Section 4.2.2, with various rank and various sparsity, we want to select an estimator which is close to the best one. The oracle is by definition the model belonging to the collection which minimizes the contrast with the true model. In practice, we do not have access to the true model, then we could not know the oracle. Nevertheless, the goal of the model selection step of our procedure is to be nearest to the oracle. In this section, we present an oracle inequality, which means that if we have penalized the log-likelihood in a good way, we will select a model which is as good as the oracle, according to the Kullback-Leibler loss.
We consider the model collection defined by (4.4).
Because we work in high-dimension, p could be big, and it will be time-consuming to test all the parts of $\{1,\dots,p\}$. We construct a sub-collection denoted by $\mathcal{J}^L$, which is constructed by the Lasso estimator, which is also random. This step is explained in more details in [Dev14c].
Moreover, to get the oracle inequality, we assume that the parameters are bounded:

$$\begin{aligned}\mathcal{S}^{\mathcal{B}}_{(K,J,R)} = \Bigg\{ s_\xi^{(K,J,R)} \in \mathcal{S}_{(K,J,R)} \Bigg| \; & \xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}), \text{ for all } k \in \{1,\dots,K\}, \\ & \Sigma_k = \operatorname{diag}([\Sigma_k]_{1,1},\dots,[\Sigma_k]_{q,q}), \\ & \text{for all } z \in \{1,\dots,q\}, a_\Sigma \le [\Sigma_k]_{z,z} \le A_\Sigma, \\ & \text{for all } k \in \{1,\dots,K\}, \beta_k^{R_k} = \sum_{r=1}^{R_k} [\sigma_k]_r [u_k^t]_{.,r} [v_k]_{r,.}, \\ & \text{for all } r \in \{1,\dots,R_k\}, [\sigma_k]_r < A_\sigma \Bigg\}.\end{aligned} \tag{4.5}$$

Remark that it is the singular value decomposition of β_k is the singular value decomposition, with $([\sigma_k]_r)_{1\leq r\leq R_k}$ the singular values, and u_k and v_k unit vectors, for $k \in \{1, \dots, K\}$.
We also assume that covariates belong to an hypercube: without restrictions, we could assume that $X \in [0,1]^p$.
Fixing $\mathcal{K}$ the possible number of components, $\mathcal{J}^L$ the relevant columns set constructed by the Lasso estimator, and $\mathcal{R}$ the possible vector of ranks, we get a model collection

$$\bigcup_{K\in\mathcal{K}} \bigcup_{J\in\mathcal{J}^L} \bigcup_{R\in\mathcal{R}} \mathcal{S}^{\mathcal{B}}_{(K,J,R)}. \tag{4.6}$$

4.3.2 Notations

Before state the main theorem which leads to the oracle inequality for the model collection (4.6), we need to define some metrics used to compare the conditional densities. First, the Kullback-Leibler divergence is defined by

$$\mathrm{KL}(s,t) = \begin{cases} \displaystyle\int_{\mathbb{R}} \log\left(\frac{s(y)}{t(y)}\right) s(y)dy \text{ if } sdy << tdy; \\ +\infty \text{ otherwise;} \end{cases} \tag{4.7}$$

for s and t two densities. To deal with regression data, for observed covariates $(x_1, \dots, x_n)$, we define

$$\mathrm{KL}^{\otimes_n}(s,t) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(s(.|x_i), t(.|x_i))\right) \tag{4.8}$$

for s and t two densities.
We also define the Jensen-Kullback-Leibler divergence, first introduced in Cohen and Le Pennec in [CLP11], by

$$\mathrm{JKL}_\rho(s,t) = \frac{1}{\rho}\mathrm{KL}(s, (1-\rho)s + \rho t)$$

for $\rho \in (0,1)$, s and t two densities. The tensorized one is defined by

$$\mathrm{JKL}_\rho^{\otimes_n}(s,t) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{JKL}_\rho(s(.|x_i), t(.|x_i))\right).$$

Note that those divergences are not metrics, they do not satisfy the triangular inequality and they are not symmetric, but they are also wildly used in statistics to compare two densities.

4.3.3 Oracle inequality

Let state the main theorem.

Theorem 4.3.1. *Assume that we observe $(x_i, y_i)_{1\leq i\leq n} \in ([0,1]^p \times \mathbb{R}^q)^n$ with unknown conditional density s^*. Let $\mathcal{M} = \mathcal{K} \times \mathcal{J} \times \mathcal{R}$ and $\mathcal{M}^L = \mathcal{K} \times \mathcal{J}^L \times \mathcal{R}$, where $\mathcal{J}^L$ is constructed by the Lasso estimator. For $(K,J,R) \in \mathcal{M}$, let $\bar{s}^{(K,J,R)} \in \mathcal{S}^{\mathcal{B}}_{(K,J,R)}$, where $\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$ is defined by (4.5), such that, for $\delta_{\mathrm{KL}} > 0$,*

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}^{(K,J,R)}) \leq \inf_{t\in\mathcal{S}^{\mathcal{B}}_{(K,J,R)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n}$$

and there exists $\tau > 0$ such that

$$\bar{s}^{(K,J,R)} \geq e^{-\tau} s^*. \tag{4.9}$$

For $(K,J,R) \in \mathcal{M}$, consider the rank constraint log-likelihood minimizer $\hat{s}^{(K,J,R)}$ in $\mathcal{S}_{(K,J,R)}$, satisfying

$$\hat{s}^{(K,J,R)} = \underset{s_\xi^{(K,J,R)} \in \mathcal{S}^{\mathcal{B}}_{(K,J,R)}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log\left(s_\xi^{(K,J,R)}(y_i|x_i) \right) \right\}.$$

Denote by $D_{(K,J,R)}$ the dimension of the model $\mathcal{S}^{\mathcal{B}}_{(K,J,R)}$. Let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ *defined by, for all $(K,J,R) \in \mathcal{M}$,*

$$\begin{aligned}\mathrm{pen}(K,J,R) \geq \kappa \frac{D_{(K,J,R)}}{n} \Bigg\{ & 2B^2(A_\sigma, A_\Sigma, a_\Sigma) - \log\left(\frac{D_{(K,J,R)}}{n} B^2(A_\sigma, A_\Sigma, a_\Sigma) \wedge 1 \right) \\ & + (1 \vee \tau) \log\left(\frac{4epq}{D_{(K,J,R)-q^2} \wedge pq} + \sum_{k=1}^{K} R_k \right) \Bigg\}\end{aligned}$$

where $\kappa > 0$ is an absolute constant, and $B(A_\sigma, A_\Sigma, a_\Sigma)$ is an absolute constant, depending on parameter bounds.
Then, the estimator $\hat{s}^{(\hat{K},\hat{J},\hat{R})}$, with

$$(\hat{K}, \hat{J}, \hat{R}) = \underset{(K,J,R) \in \mathcal{M}^L}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{(K,J,R)}(y_i|x_i)) + \mathrm{pen}(K,J,R) \right\},$$

satisfies

$$\begin{aligned}&\mathrm{E}\left(\mathrm{JKL}_\rho^{\otimes n}(s^*, \hat{s}^{(\hat{K},\hat{J},\hat{R})}) \right) \\ &\leq C\, \mathrm{E}\left(\inf_{(K,J,R) \in \mathcal{M}^L} \left(\inf_{t \in \mathcal{S}_{(K,J,R)}} \mathrm{KL}^{\otimes n}(s^*, t) + \mathrm{pen}(K,J,R) \right) + \frac{(1 \vee \tau)}{n} \right) \qquad (4.10)\end{aligned}$$

some absolute positive constant C.

The proof of the Theorem 4.3.1 is given in Section 4.5. Note that condition (4.9) leads to control the random model collection. The mixture parameters are bounded in order to construct brackets over $\mathcal{S}_{(K,J,R)}$, and thus to upper bound the entropy. The inequality (4.10) not exactly an oracle inequality, since the Jensen-Kullback-Leibler risk is upper bounded by the Kullback-Leibler divergence. Note that we use the Jensen-Kullback-Leibler divergence rather than the Kullback-Leibler divergence, because it is bounded. This boundedness turns out to be crucial to control the loss of the penalized maximum likelihood estimator under mild assumptions on the complexity of the model and on parameters.
Because we are looking at a random sub-collection of models which is small enough, our estimator $\hat{s}^{(K,J,R)}$ is attainable in practice. Moreover, it is a non-asymptotic result, which allows us to study cases for which p increases with n.
We could compare our inequality with the bound of Bunea et al, in [BSW12], who computed a procedure similar to ours, for a linear model. According to consistent group selection for the group-Lasso estimator, they get adaptivity of the estimator to an optimal rate, and their estimators perform the bias variance trade-off among all reduced rank estimators. Nevertheless, their results are obtained according to some assumptions, for instance the mutual coherence on X^tX, which postulates that the off-diagonal elements have to be small. Some assumptions on the design are required, whereas our result just needs to deal with bounded parameters and bounded covariates.

4.4 Numerical studies

We will illustrate our procedure with simulated and benchmark datasets, to highlight the advantages of our method. We adapt the simulations part of Bunea et al. article, [BSW12]. Indeed, we work in the same way, to get a sparse and low rank estimator. Nevertheless, we add a mixture framework to be consistent with our clustering method, and to have more flexibility.

4.4.1 Simulations

To illustrate our procedure, we use simulations adapted from the article of Bunea [BSW12], extended to mixture models.
The design matrix X has independent and identically distributed rows X_i, distributed from a multivariate Gaussian $\mathcal{N}_q(0, \Sigma)$ with $\Sigma = \rho I$, $\rho > 0$. We consider a mixture of 2 components. According to the cluster, the coefficient matrix β_k has the form

$$\beta_k = \begin{bmatrix} b_k B^0 & b_k B^1 \\ 0 & 0 \end{bmatrix}$$

for $k \in \{1, 2\}$, with B^0 a $J \times R_k$ matrix and B^1 a $R_k \times q$ matrix. All entries in B^0 and B^1 are independent and identically distributed according to $\mathcal{N}(0, 1)$. The noise matrix ϵ has independent $\mathcal{N}(0, 1)$ entries. Let ϵ_i denotes its ith row.
The proportion vector π is defined by $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$, i.e. all clusters have the same probability.
Each row Y_i in Y is then generated as, if the observation i belongs to the cluster k, $Y_i = \beta_k X_i + \epsilon_i$, for all $i \in \{1, \ldots, n\}$. This setup contains many noise features, but the relevant ones lie in a low-dimensional subspace. We report two settings:

— $p > n$: $n = 50, |J| = 6, p = 100, q = 10, R = (3, 3), \rho = 0.1, b = (3, -3)$.
— $p < n$: $n = 200, |J| = 6, p = 10, q = 10, R = (3, 3), \rho = 0.01, b = (3, -3)$.

The current setups show that variable selection, without taking the rank information into consideration, may be suboptimal, even if the correlations between predictors are low. Each model are simulated 20 times, and Table 4.1 summarizes our findings. We evaluate the prediction accuracy of each estimator $\hat{\beta}$ by the Kullback-Leibler divergence (KL) using a test sample at each run. We also report the median rank estimate (denoted by $\hat{R}$) over all runs, rates of non included true variables (denoted by M for misses) and the rates of incorrectly included variables (FA for false actives). Ideally, we are looking for a model with low KL, low M and low FA.

	KL	$\hat{R}$	M	FA	ARI
$p > n$	19.03	[2.8,3]	0	20	0.95
$p < n$	3.28	[3,3]	0	0.6	0.99

Table 4.1: Performances of our procedure. Mean number $\{\text{KL}, \hat{R}, M, FA, ARI\}$ of the Kullback-Leibler divergence between the model selected and the true model, the estimated rank of the model selected in each cluster, the missed variables, the false relevant variables, and the ARI, over 20 simulations

We could draw the following conclusions from Table 4.1. When we work in low-dimensional framework, we get very good results. Even if we could use any estimator, because we do not have dimension problem, with our procedure we get the matrix structure involved by the model. Over 20 simulations, we get almost exact clustering, and the Kullback-Leibler divergence between the model we construct and the true model is really low. In case of high-dimensional data, when p is larger than n, we get also good results. We find the good structure, selecting the relevant variables (our model will have false relevant variables, but no missed variables), and selecting

the true ranks. We could remark that false relevant variables have low values. Comparing to another procedure which will not reduce the rank, we will perform the dimension reduction.

4.4.2 Illustration on a real dataset

In this section, we apply our procedure to real data set. The Norwegian paper quality data were obtained from a controlled experiment that was carried out at a paper factory in Norway to uncover the effect of three control variables X_1, X_2, X_3 on the quality of the paper which was measured by 13 response variables. Each of the control variables X_i takes values in $\{-1, 0, 1\}$. To account for possible interactions and nonlinear effects, second order terms were added to the set of predictors, yielding $X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_2X_3$, and the intercept term. There were 29 observations with no missing values made on all response and predictor variables. The Box Behnken design of the experiment and the resulting data are described in Aldrin [Ald96] and Izenman [Ize75]. Moreover, Bunea et al. in [BSW12] also study this dataset. We always center the responses and the predictors. The dataset clearly indicates that dimension reduction is possible, making it a typical application for reduced rank regression methods. Moreover, our method will exhibit different clusters among this sample.
We construct a model collection varying the number of clusters in $\mathcal{K} = \{2, \dots, 5\}$. We select a model with 2 clusters. We select all variables except X_1X_2 and X_2X_3, which is consistent with comments of Bunea et al. In the two clusters, we get two mean matrices, with ranks equal to 2 and 4. One cluster describes the mean comportment (with rank equals to 2), whereas the other cluster contains values more different.

4.5 Appendices

In those appendices, we present the details of the proof of the Theorem 4.3.1. It derives from a general model selection theorem, stated in Section 4.5.1, and proved in the Chapter 3. Then, the proof of the Theorem 4.3.1 could be summarized by satisfying assumptions H_m, Sep_m and K described in Section 4.5.1.

4.5.1 A general oracle inequality for model selection

Model selection appears with the AIC criterion and BIC criterion. A non-asymptotic theory was developed by Birgé and Massart in [BM07]. With some assumptions detailed here, we get an oracle inequality for the maximum likelihood estimator among a model collection. Cohen and Le Pennec, in [CLP11], generalize this theorem in regression framework. We have to use a generalization of this theorem detailed in [Dev14b] because we consider a random collection of models. Let state the main theorem. We consider a model collection $(S_m)_{m\in\mathcal{M}}$, indexed by $\mathcal{M}$. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$.
Begin by describe the assumptions. First, we impose a structural assumption. It is a bracketing entropy condition on the model with respect to the tensorized Hellinger divergence

$$(d_H^{\otimes n})^2(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} d_H^2(s(.|x_i), t(.|x_i))\right];$$

for two densities s and t. A bracket $[l, u]$ is a pair of functions such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$l(y,x) \le s(y|x) \le u(y,x).$$

For $\epsilon > 0$, the bracketing entropy $\mathcal{H}_{[.]}(\epsilon, S, d_H^{\otimes n})$ of a set S is defined as the logarithm of the minimum number of brackets $[l, u]$ of width $d_H^{\otimes n}(l, u)$ smaller than ϵ such that every densities of S belong to one of these brackets.
Let $m \in \mathcal{M}$.

Assumption $\mathbf{H}_m$. *There is a non-decreasing function ϕ_m such that $\varpi \mapsto 1/\varpi\phi_m(\varpi)$ is non-increasing on $(0, +\infty)$ and for every $\varpi \in \mathbb{R}^+$ and every $s_m \in S_m$,*

$$\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, S_m(s_m, \varpi), d_H^{\otimes n})} d\epsilon \leq \phi_m(\varpi);$$

where $S_m(s_m, \varpi) = \{t \in S_m, d_H^{\otimes n}(t, s_m) \leq \varpi\}$. The model complexity $\mathcal{D}_m$ is then defined as $n\varpi_m^2$ with ϖ_m the unique solution of

$$\frac{1}{\varpi}\phi_m(\varpi) = \sqrt{n}\varpi. \tag{4.11}$$

Remark that the model complexity depends on the bracketing entropies not of the global models S_m but of the ones of smaller localized sets. It is a weaker assumption.
For technical reasons, a separability assumption is also required.

Assumption $\mathbf{Sep}_m$. *There exists a countable subset $S_m^{'}$ of S_m and a set $\mathcal{Y}_m^{'}$ with $\lambda(\mathcal{Y} \backslash \mathcal{Y}_m^{'}) = 0$, for λ the Lebesgue measure, such that for every $t \in S_m$, there exists a sequence $(t_l)_{l \geq 1}$ of elements of $S_m^{'}$ such that for every x and every $y \in \mathcal{Y}_m^{'}$, $\log(t_l(y|x))$ goes to $\log(t(y|x))$ as l goes to infinity.*

According to this assumption, we could work with a countable subset.
We also need an information theory type assumption on our model collection. We assume the existence of a Kraft-type inequality for the collection.

Assumption K. *There is a family $(w_m)_{m \in \mathcal{M}}$ of non-negative numbers such that*

$$\sum_{m \in \mathcal{M}} e^{-w_m} \leq \Omega < +\infty.$$

Then, we could write our main global theorem to get an oracle inequality in regression framework, with a random collection of models.

Theorem 4.5.1. *Assume we observe $(x_i, y_i)_{1 \leq i \leq n} \in ([0, 1]^p \times \mathbb{R}^q)^n$ with unknown conditional density s^*. Let $\mathcal{S} = (S_m)_{m \in \mathcal{M}}$ be at most countable collection of conditional density sets. Let assumption K holds while assumptions H_m and Sep_m hold for every models $S_m \in \mathcal{S}$. Let $\delta_{\mathrm{KL}} > 0$, and $\bar{s}_m \in S_m$ such that*

$$\mathrm{KL}^{\otimes n}(s^*, \bar{s}_m) \leq \inf_{t \in S_m} \mathrm{KL}^{\otimes n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

and let $\tau > 0$ such that

$$\bar{s}_m \geq e^{-\tau} s^*. \tag{4.12}$$

Introduce $(S_m)_{m \in \check{\mathcal{M}}}$ some random sub-collection of $(S_m)_{m \in \mathcal{M}}$. Consider the collection $(\hat{s}_m)_{m \in \check{\mathcal{M}}}$ of η-log-likelihood minimizer in S_m, satisfying

$$\sum_{i=1}^n -\log(\hat{s}_m(y_i|x_i)) \leq \inf_{s_m \in S_m} \left(\sum_{i=1}^n -\log(s_m(y_i|x_i)) \right) + \eta.$$

Then, for any $\rho \in (0,1)$ and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq \kappa(\mathcal{D}_m + (1 \vee \tau) w_m) \tag{4.13}$$

with $\kappa > \kappa_0$, and where the model complexity $\mathcal{D}_m$ is defined in (4.11), the penalized likelihood estimate $\hat{s}_{\hat{m}}$ with $\hat{m} \in \check{\mathcal{M}}$ such that

$$\sum_{i=1}^{n} -\log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \check{\mathcal{M}}} \left(\sum_{i=1}^{n} -\log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m) \right) + \eta'$$

satisfies

$$\begin{aligned} \mathrm{E}(\mathrm{JKL}_{\rho}^{\otimes n}(s^*, \hat{s}_{\hat{m}})) \leq C_1 \, \mathrm{E} \left(\inf_{m \in \check{\mathcal{M}}} \inf_{t \in S_m} \mathrm{KL}^{\otimes n}(s^*, t) + 2 \frac{\mathrm{pen}(m)}{n} \right) \\ + C_2 (1 \vee \tau) \frac{\Omega^2}{n} + \frac{\eta' + \eta}{n}. \end{aligned} \tag{4.14}$$

Remark 4.5.2. *We get that, among a random model collection, we are able to choose a model which is as good as the oracle, up to a constant C_1, and some additive terms being around $1/n$. This result is non-asymptotic, and gives a theoretical penalty to select this model.*

Remark 4.5.3. *The proof of this theorem is detailed in [Dev14b]. Nevertheless, we could give the main ideas to understand the assumptions. From assumptions H_m and Sep_m, we could use maximal inequalities which lead to, except on a set of probability less than $e^{-w_{m'}-w}$, for all w, a control of the ratio of the centered empirical process of $\log(\hat{s}_{m'})$ over the Hellinger distance between s^* and $\hat{s}_{m'}$, this control being around $1/n$. Thanks to Bernstein inequality, satisfied according to the inequality (4.12), and thanks to the assumption K, we get the oracle inequality.*

Now, to prove Theorem 4.3.1, we have to satisfy assumptions H_m and K, assumption Sep_m being true for our conditional densities.

4.5.2 Assumption H_m

Decomposition

As done in Cohen and Le Pennec [CLP11], we could decompose the entropy by

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J,R)}, d_H^{\otimes n}) \leq \mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes n}) + \sum_{k=1}^{K} \mathcal{H}_{[.]}(\epsilon, \mathcal{F}_{(J,R_k)}, d_H^{\otimes n}) \tag{4.15}$$

where

$$\mathcal{S}^{\mathcal{B}}_{(K,J,R)} = \left\{ \begin{array}{l} y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto s_\xi^{(K,J,R)}(y|x) = \sum_{k=1}^K \pi_k \varphi\left(y|(\beta_k^{R_k})^{[J]}x, \Sigma_k\right) \\ \xi = \left(\pi_1, \ldots, \pi_K, (\beta_1^{R_1})^{[J]}, \ldots, (\beta_K^{R_K})^{[J]}, \Sigma_1, \ldots, \Sigma_K\right) \in \Xi_{(K,J,R)} \\ \Xi_{(K,J,R)} = \Pi_K \times \tilde{\Psi}_{(K,J,R)} \times ((a_\Sigma, A_\Sigma]^q)^K \end{array} \right\}$$

$$\Psi_{(K,J,R)} = \left\{ ((\beta_1^{R_1})^{[J]}, \ldots, (\beta_K^{R_K})^{[J]}) \in (\mathbb{R}^{q\times p})^K | \mathrm{Rank}(\beta_k) = R_k \right\};$$

$$\tilde{\Psi}_{(K,J,R)} = \left\{ ((\beta_1^{R_1})^{[J]}, \ldots, (\beta_K^{R_K})^{[J]}) \in \Psi_{(K,J,R)} \middle| \text{ for all } k \in \{1, \ldots, K\}, \right.$$
$$\left. \beta_k^{R_k} = \sum_{r=1}^{R_k} \sigma_r u_r^t v_r, \text{ with } \sigma_r < A_\sigma \text{ for all } r \in \{1, \ldots, R_k\} \right\}$$

$$\Pi_K = \left\{ (\pi_1, \ldots, \pi_K) \in (0,1)^K; \sum_{k=1}^K \pi_k = 1 \right\};$$

$$\mathcal{F}_{(J,R)} = \left\{ \varphi(.|(\beta^R)^{[J]}X, \Sigma); \beta^R = \sum_{r=1}^R \sigma_r u_r^t v_r, \text{ with } \sigma_r < A_\sigma, \right.$$
$$\left. \Sigma = \mathrm{diag}(\Sigma_{1,1}, \ldots, \Sigma_{q,q}) \in [a_\Sigma, A_\Sigma]^q \right\}$$

φ the Gaussian density.

For the proportions, it is known that (see Wasserman and Genovese in [GW00])

$$\mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes n}) \leq \log\left(K(2\pi e)^{K/2} \left(\frac{3}{\epsilon}\right)^{K-1} \right).$$

Look after the Gaussian entropy.

For the Gaussian

We want to bound the integrated entropy. For that, first we have to construct some brackets to recover S_m. Fix $f \in S_m$. We are looking for functions l and u such that $l \leq f \leq u$. Because f is a Gaussian, l and u are dilatations of Gaussian. We then have to determine the mean, the variance and the dilatation coefficient of l and u. We need the both following lemmas to construct these brackets.

Lemme 4.5.4. *Let $\varphi(.|\mu_1, \Sigma_1)$ and $\varphi(.|\mu_2, \Sigma_2)$ be two Gaussian densities. If their variance matrices are assumed to be diagonal, with $\Sigma_a = \mathrm{diag}([\Sigma_a]_{1,1}, \ldots, [\Sigma_a]_{q,q})$ for $a \in \{1, 2\}$, such that $[\Sigma_2]_{z,z} > [\Sigma_1]_{z,z} > 0$ for all $z \in \{1, \ldots, q\}$, then, for all $x \in \mathbb{R}^q$,*

$$\frac{\varphi(x|\mu_1, \Sigma_1)}{\varphi(x|\mu_2, \Sigma_2)} \leq \prod_{z=1}^q \frac{\sqrt{[\Sigma_2]_{z,z}}}{\sqrt{[\Sigma_1]_{z,z}}} e^{\frac{1}{2}(\mu_1-\mu_2)^t \mathrm{diag}\left(\frac{1}{[\Sigma_2]_{1,1}-[\Sigma_1]_{1,1}}, \ldots, \frac{1}{[\Sigma_2]_{q,q}-[\Sigma_1]_{q,q}}\right)(\mu_1-\mu_2)}.$$

Lemme 4.5.5. *The Hellinger distance of two Gaussian densities with diagonal variance matrices*

is given by the following expression:

$$
\begin{aligned}
&d_H^2(\varphi(.|\mu_1,\Sigma_1),\varphi(.|\mu_2,\Sigma_2))\\
=&2-2\left(\prod_{z=1}^{q}\frac{2\sqrt{[\Sigma_1]_{z,z}[\Sigma_2]_{z,z}}}{[\Sigma_1]_{z,z}+[\Sigma_2]_{z,z}}\right)^{1/2}\\
&\times\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)^t\operatorname{diag}\left(\left(\frac{1}{[\Sigma_1]_{z,z}+[\Sigma_2]_{z,z}}\right)_{z\in\{1,\dots,q\}}\right)(\mu_1-\mu_2)\right\}
\end{aligned}
$$

To get an ϵ-bracket for the densities, we have to construct a δ-net for the variance and the mean, δ to be specify later.

— **Step 1: construction of a net for the variance**
Let $\epsilon\in(0,1]$, and $\delta=\frac{\epsilon}{\sqrt{2q}}$. Let $b_j=(1+\delta)^{1-\frac{j}{2}}A_\Sigma$. For $2\le j\le N$, we have $[a_\Sigma,A_\Sigma]=[b_N,b_{N-1}]\bigcup\dots\bigcup[b_3,b_2]$, where N is chosen to recover everything. We want that

$$
\begin{aligned}
& a_\Sigma=(1+\delta)^{1-N/2}A_\Sigma\\
\Leftrightarrow\quad & \log\frac{a_\Sigma}{A_\Sigma}=\left(1-\frac{N}{2}\right)\log(1+\delta)\\
\Leftrightarrow\quad & N=\frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)}.
\end{aligned}
$$

We want N to be an integer, then $N=\left\lceil\frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)}\right\rceil$. We get a regular net for the variance. We could let $B=\operatorname{diag}(b_{i(1)},\dots,b_{i(q)})$, close to Σ (and deterministic, independent of the values of Σ), where i is a permutation such that $b_{i(z)+1}\le\Sigma_{z,z}\le b_{i(z)}$ for all $z\in\{1,\dots,q\}$.

— **Step 2: construction of a net for the mean vectors**
We use the singular decomposition of β, $\beta=\sum_{r=1}^{R}\sigma_r u_r^t v_r$, with $(\sigma_r)_{1\le r\le R}$ the singular values, and $(u_r)_{1\le r\le R}$ and $(v_r)_{1\le r\le R}$ unit vectors. Those vectors are also bounded.
We are looking for l and u such that $d_H(l,u)\le\epsilon$, and $l\le f\le u$. We will use a dilatation of a Gaussian to construct such an ϵ-bracket of φ. We let

$$
\begin{aligned}
l(x,y)&=(1+\delta)^{-(p^2qR+3q/4)}\varphi(y|\nu_{J,R}x,(1+\delta)^{-1/4}B^1)\\
u(x,y)&=(1+\delta)^{p^2qR+3q/4}\varphi(y|\nu_{J,R}x,(1+\delta)B^2)
\end{aligned}
$$

where B^1 and B^2 are constructed such that, for all $z\in\{1,\dots,q\}$, $[B^1]_{z,z}\le\Sigma_{z,z}\le[B^2]_{z,z}$ (see step 1).
The means $\nu_{J,R}\in\mathbb{R}^{q\times p}$ will be specified later. Just remark that J is the set of the relevant columns, and R the rank of $\nu_{J,R}$: we will decompose $\nu_{J,R}=\sum_{r=1}^{R}\tilde\sigma_r\tilde u_r^t\tilde v_r$, $\tilde u\in\mathbb{R}^{|J|\times R}$, and $\tilde v\in\mathbb{R}^{q\times R}$.
We get

$$l(x,y)\le f(y|x)\le u(x,y)$$

if we have

$$||\beta x-\nu_{J,R}x||_2^2\le p^2qR\frac{\delta^2}{2}a_\Sigma^2(1-2^{-1/4}).$$

Remark that $||\beta x - \nu_{J,R} x||_2^2 \leq p||\beta - \nu_{J,R}||_2^2 ||x||_\infty$ We need then

$$||\beta - \nu_{J,R}||_2^2 \leq pqR\frac{\delta^2}{2}a_\Sigma^2(1-2^{-1/4}) \tag{4.16}$$

According to [Dev14b], $d_H(l,u) \leq 2(p^2qR + 3q/4)^2\delta^2$, then, with

$$\delta = \frac{\epsilon}{\sqrt{2}(pqR + 3/4q)}$$

we get the wanted bound.
Now, explain how to construct $\nu_{J,R}$ to get (4.16).

$$\begin{aligned}
||\beta - \nu_{J,R}||_2^2 &= \sum_{j=1}^{p}\sum_{z=1}^{q}\left|\sum_{r=1}^{R}\sigma_r u_{r,j}v_{r,z} - \tilde{\sigma}_r\tilde{u}_{r,j}\tilde{v}_{r,z}\right|^2 \\
&= \sum_{j=1}^{p}\sum_{z=1}^{q}\left|\sum_{r=1}^{R}|\sigma_r - \tilde{\sigma}_r||u_{j,r}v_{z,r}| - \tilde{\sigma}_r|\tilde{u}_{r,j} - u_{r,j}||\tilde{v}_{z,r}| - \tilde{\sigma}_r u_{r,j}|v_{r,z} - \tilde{v}_{r,z}|\right|^2 \\
&\leq \sum_{j=1}^{p}\sum_{z=1}^{q}\left|\sum_{r=1}^{R}|\sigma_r - \tilde{\sigma}_r| + A_\sigma|\tilde{u}_{r,j} - u_{r,j}| + A_\sigma|v_{r,z} - \tilde{v}_{r,z}|\right|^2 \\
&\leq 2pqR\left(\max_r|\sigma_r - \tilde{\sigma}_r|^2 + A_\sigma\max_{r,j}|\tilde{u}_{r,j} - u_{r,j}|^2 + A_\sigma\max_{r,z}|\tilde{v}_{r,z} - v_{r,z}|^2\right)
\end{aligned}$$

We need $||\beta - \nu_{J,R}||_2^2 \leq pqR\frac{\delta^2}{2}a_\Sigma^2(1-2^{-1/4})$.
If we choose $\tilde{\sigma}_r$, $\tilde{u}_{r,j}$ and $\tilde{v}_{r,z}$ such that

$$\begin{aligned}
|\sigma_r - \tilde{\sigma}_r| &\leq \frac{\delta}{\sqrt{12}}a_\Sigma\sqrt{1-2^{-1/4}}, \\
|u_{r,j} - \tilde{u}_{r,j}| &\leq \frac{\delta}{\sqrt{12}A_\sigma}a_\Sigma\sqrt{1-2^{-1/4}}, \\
|v_{r,z} - \tilde{v}_{r,z}| &\leq \frac{\delta}{\sqrt{12}A_\sigma}a_\Sigma\sqrt{1-2^{-1/4}},
\end{aligned}$$

then it works.

To get this, we let, for $\lfloor . \rfloor$ the floor function,

$$
\begin{aligned}
S &= \mathbb{Z} \cap \left[0, \left\lfloor \frac{A_\sigma}{\frac{\delta}{\sqrt{12}} a_\Sigma \sqrt{1-2^{-1/4}}} \right\rfloor \right] \\
\tilde{\sigma}_r &= \underset{\varsigma \in S}{\operatorname{argmin}} \left| \sigma_r - \frac{\delta}{\sqrt{12}} a_\Sigma \sqrt{1-2^{-1/4}} \varsigma \right|, \\
U &= \mathbb{Z} \cap \left[0, \left\lfloor \frac{A_\sigma}{\frac{\delta}{\sqrt{12} A_\sigma} a_\Sigma \sqrt{1-2^{-1/4}}} \right\rfloor \right] \\
\tilde{u}_{r,j} &= \underset{\mu \in U}{\operatorname{argmin}} \left| u_{r,j} - \frac{\delta}{\sqrt{12} A_\sigma} a_\Sigma \sqrt{1-2^{-1/4}} \mu \right|, \\
V &= \mathbb{Z} \cap \left[0, \left\lfloor \frac{A_\sigma}{\frac{\delta}{\sqrt{12} A_\sigma} a_\Sigma \sqrt{1-2^{-1/4}}} \right\rfloor \right] \\
\tilde{v}_{z,r} &= \underset{\nu \in V}{\operatorname{argmin}} \left| v_{z,r} - \frac{\delta}{\sqrt{12} A_\sigma} a_\Sigma \sqrt{1-2^{-1/4}} \nu \right|
\end{aligned}
$$

for all $r \in \{1, \dots, R\}, j \in \{1, \dots, p\}, z \in \{1, \dots, q\}$.
Remark that we just need to determine the vectors $((\tilde{u}_{r,j})_{1 \le j \le J-r})_{1 \le r \le R}$ and $((\tilde{v}_{z,r})_{1 \le z \le q-r})_{1 \le r \le R}$ because those vectors are unit.
Then, we let

$$
\begin{aligned}
&\text{for all } j \in J^c, \text{ for all } z \in \{1, \dots, p\}, (\nu_{J,R})_{z,j} = 0 \\
&\text{for all } j \in J, \text{ for all } z \in \{1, \dots, p\}, (\nu_{J,R})_{z,j} = \sum_{r=1}^{R} \tilde{\sigma}_r \tilde{u}_{r,j} \tilde{v}_{z,r}
\end{aligned}
$$

— **Step 3: Upper bound of the number of ϵ-brackets for $\mathcal{F}_{(J,R)}$**
We have defined our bracket. Let $c = \frac{1-2^{-1/4}}{12}$. We want to control the entropy.

$$
\begin{aligned}
|\mathcal{B}_\epsilon(\mathcal{F}_{(J,R)})| \le & \sum_{l=2}^{N} \left(\frac{A_\sigma}{\delta a_\Sigma \sqrt{c}} \right)^R \left(\frac{A_\sigma^2}{\delta a_\Sigma \sqrt{c}} \right)^{R\left(\frac{2J-R-1}{2} + \frac{2q-R-1}{2}\right)} \\
\le & (N-1) \left(\frac{A_\sigma^2}{\delta a_\Sigma \sqrt{c}} \right)^{R(J+q-R)} A_\sigma^{-R} \\
\le & C(a_\Sigma, A_\Sigma, A_\sigma, J, R) \delta^{-D_{(J,R)}-1}
\end{aligned}
$$

with $C(a_\Sigma, A_\Sigma, A_\sigma, J, R) = 2 \left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2} \right) \left(\frac{A_\sigma^2}{a_\Sigma \sqrt{c}} \right)^{R(J+q-R)} A_\sigma^{-R}$.

For the mixture

Begin by computing the bracketing entropy: according to (4.15),

$$
\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J,R)}, d_H) \le \log \left(C \left(\frac{1}{\epsilon} \right)^{D_{(K,J,R)}} \right)
$$

where

$$C = 2^K K(2\pi e)^{K/2} 3^{K-1} \left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)^K \left(\frac{A_\sigma \sqrt{12}}{a_\Sigma^2 \sqrt{1-2^{-1/4}}}\right)^{D_{(K,J,R)}} A_\sigma^{-\sum_{k=1}^K R_k} \tag{4.17}$$

and $D_{(K,J,R)} = \sum_{k=1}^K R_k(|J| + q - R_k)$.
We have to determine $\phi_{(K,J,R)}$ such that

$$\int_0^\varpi \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J,R)}(s^{(K,J,R)}, \varpi), d_H^{\otimes_n})} d\epsilon \leq \phi_{(K,J,R)}(\varpi). \tag{4.18}$$

Let compute the integral.

$$\begin{aligned}
&\int_0^\varpi \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J,R)}(s^{(K,J,R)}, \varpi), d_H^{\otimes_n})} d\epsilon \\
&\leq \varpi \sqrt{\log(C)} + \sqrt{D_{(K,J,R)}} \int_0^\varpi \sqrt{\log\left(\frac{1}{\epsilon}\right)} d\epsilon \\
&\leq \sqrt{D_{(K,J,R)}}\varpi \left[\sqrt{\pi} + \sqrt{\frac{\log(C)}{D_{(K,J,R)}}} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right]
\end{aligned}$$

with, according to (4.17),

$$\begin{aligned}
\log(C) =& K\log(2) + \frac{K}{2}\log(2\pi e) + (K-1)\log(3) + K\log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right) \\
&+ D_{(K,J,R)} \log\left(\frac{A_\sigma^2\sqrt{12}}{a_\Sigma\sqrt{1-2^{-1/4}}}\right) + \log(K) + \sum_{k=1}^K R_k \log(\frac{1}{A_\sigma}) \\
\leq& D_{(K,J,R)}\left(\log(2) + \log(2\pi e) + \log 3 + 1 + \log\left(\frac{A_\sigma^2\sqrt{12}}{a_\Sigma\sqrt{1-2^{-1/4}}}\right)\right) \\
&+ D_{(K,J,R)}\left(\log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\right) \\
\leq& D_{(K,J,R)}\left(\log\left(\frac{12\sqrt{12}\pi e}{\sqrt{1-2^{-1/4}}}\right) + \log\left(\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\left(\frac{A_\sigma^2}{a_\Sigma}\right)\right)\right)
\end{aligned}$$

Then,

$$\begin{aligned}
&\int_0^\varpi \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J,R)}(s^{(K,J,R)}, \varpi), d_H^{\otimes_n})} d\epsilon \\
&\leq \sqrt{D_{(K,J,R)}}\left(2 + \sqrt{\log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\left(\frac{A_\sigma^2}{a_\Sigma}\right)} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right)
\end{aligned}$$

Consequently, by putting

$$B = 2 + \sqrt{\log\left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2}\right)\left(\frac{A_\sigma^2}{a_\Sigma}\right)};$$

we get that the function $\phi_{(K,J,R)}$ defined on $\mathbb{R}_+^*$ by

$$\phi_{(K,J,R)}(\varpi) = \sqrt{D_{(K,J,R)}}\varpi\left(B + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right)$$

satisfies (4.18). Besides, $\phi_{(K,J,R)}$ is nondecreasing and $\varpi \mapsto \phi_{(K,J,R)}(\varpi)/\varpi$ is non-increasing, then $\phi_{(K,J,R)}$ is convenient.
Finally, we need to find an upper bound of ϖ_* satisfying

$$\phi_{(K,J,R)}(\varpi_*) = \sqrt{n}\varpi_*^2.$$

Consider ϖ_* such that $\phi_{(K,J,R)}(\varpi_*) = \sqrt{\varpi_*^2}$.
This is equivalent to solve

$$\varpi_* = \sqrt{\frac{D_{(K,J,R)}}{n}}\left(B + \sqrt{\log\left(\frac{1}{\varpi_* \wedge 1}\right)}\right)$$

and then we could choose

$$\varpi_*^2 \leq \frac{D_{(K,J,R)}}{n}\left(2B^2 + \log\left(\frac{1}{1 \wedge \frac{D_{(K,J,R)}}{n}B^2}\right)\right).$$

4.5.3 Assumption K

Let $w_{(K,J,R)} = \tilde{D}_{(K,J)} \log\left(\frac{4epq}{(\tilde{D}_{(K,J)}-q^2)\wedge pq}\right) + \sum_{k\in\{1,\dots,K\}} R_k$, where $\tilde{D}_{(K,J)} = K(1+|J|)$. Then, we could compute the sum

$$\begin{aligned}
\sum_{(K,J,R)} e^{-w_{(K,J,R)}} &\leq \sum_{K\geq 1}\left[\left(\sum_{1\leq|J|\leq pq} e^{-\tilde{D}_{(K,J)}\log\left(\frac{4epq}{(\tilde{D}_{(K,J)}-q^2)\wedge pq}\right)}\right)\left(\sum_{R\geq 1} e^{-R}\right)^K\right] \\
&\leq \sum_{K\geq 1}\left[\left(\sum_{1\leq|J|\leq pq} e^{-\tilde{D}_{(K,J)}\log\left(\frac{4epq}{(\tilde{D}_{(K,J)}-q^2)\wedge pq}\right)}\right) 1^K\right] \\
&\leq 2
\end{aligned}$$

The last inequality is computed in Proposition 4.5 in [Dev14b] by 2. Then,

$$\sum_{(K,J,R)} e^{-w_{(K,J,R)}} \leq 2.$$

Chapter 5

Clustering electricity consumers using high-dimensional regression mixture models.

Contents

4.1 Introduction **117**
4.2 The model and the model collection **119**
4.2.1 Linear model 119
4.2.2 Mixture model 119
4.2.3 Generalized EM algorithm 120
4.2.4 The Lasso-Rank procedure 121
4.3 Oracle inequality **122**
4.3.1 Framework and model collection 122
4.3.2 Notations 123
4.3.3 Oracle inequality 123
4.4 Numerical studies **125**
4.4.1 Simulations 125
4.4.2 Illustration on a real dataset 126
4.5 Appendices **126**
4.5.1 A general oracle inequality for model selection 126
4.5.2 Assumption H_m 128
Decomposition 128
For the Gaussian 129
For the mixture 132
4.5.3 Assumption K 134

5.1 Introduction

New metering infrastructures as smart meters provide new and potentially massive informations about individual (household, small and medium enterprise) consumption. As an example, in France, ERDF (Electricite Reseau Distribution de France the French manager of the public electricity distribution network) deployed 250 000 smart meters, covering a rural and an urban territory and providing half-hourly household energy used each day. ERDF plans to install 35 millions of them over the French territory by the end of 2020 and exploiting such an amount of data is an exciting but challenging task (see `http://www.erdf.fr/Linky`).
Many applications coming from individual data analysis can be found in the literature. The first and most popular one is load profiling. Understanding consumers time of use, seasonal patterns and the different features that drive their consumption is a fundamental task for electricity providers to design their offer and more generally for marketing studies (see e.g. [Irw86]). Energy agencies and states can also benefit from profiling for efficiency programs and improve recommendation policies. Customer segmentation based on load classification is a natural approach for that and [ZYS13] proposes a nice review of the most popular methods, concluding that classification for smart grids is a hard task due to the complexity, massiveness, high dimension and heterogeneity of the data. Another problem pointed out is the dynamic structure of smart meters data and particularly the issue of portfolio variations (losses and gains of customers), the update of a previous classification when new customers arrive or the clustering of a new customer with very few observations on its load. In [KFR14], the authors propose a segmentation algorithm based on K-means to uncover shape dictionaries that help to summarize information and cluster a large population of 250 000 households in California. However, the proposed solution exploits a quite long historic of data of at least 1 year.
Recently, other important questions were raised by smart meters and the new possibility to send potentially complex signal to consumers (incentive payments, time varying prices...) and demand response program tailoring attracts a lot of attention (see [US 06], [H$^+$13]). Local optimization of electricity production and real time management of individual demand thus play an important role in the smart grid landscape. It induces a need for local electricity load forecasting at different levels of the grid and favorites bottom-up approaches based on a two stage process. First, it consists in building classes in a population such that each class could be sufficiently well forecast but corresponds to different load shapes or reacts differently to exogenous variables like temperature or prices (see e.g. [LSD15] in the context of demand response). The second stage consists in aggregating forecasts to forecast the total or any subtotal of the population consumption. For example, identify and forecast the consumption of a sub-population reactive to an incentive is an important need to optimize a demand response program. Surprisingly, few papers consider the problem of clustering individual consumption for forecasting and specially for forecasting at a disaggregated level (e. g. in each class). In [AS13], clustering procedures are compared according to the forecasting performances of their corresponding bottom-up forecasts of the total consumption of 6 000 residential customers and small-to-medium enterprises in Ireland. Even if they achieve nice performances at the end, the proposed clustering methods are quite independent to the VAR model used for forecasting. In [MMOP10], a clustering algorithm is proposed that couples hierarchical clustering and multi-linear regression models to improve the forecast of the total consumption of a French industrial subset. They obtain a real forecasting gain but need a sufficiently long dataset (2-3 years) and the algorithm is computationally intensive.
We propose here a new methodology based on high dimensional regression models. Our main contribution is that we focus on uncovering classes corresponding to different regression models. As a consequence, these classes could then be exploited for profiling as well as forecasting in

each classes or for bottom-up forecasts in an unified view. More precisely, we consider regression models where $Y_d = X_d\beta + \varepsilon_d$ is typically an individual -high dimension- load curve for day d and X_d could be alternatively Y_{d-1} or any other exogenous covariates.
We consider a real dataset of Irish individual consumers of 4 225 meters, each with 48 half-hourly meter reads per day over 1 year: from 1st January 2010 up to 31st December 2010. These data have already been studied in [AS13] and [Cha14] and we refer to those papers for a presentation of the data. For computational and time reasons, we draw a random sample of around 500 residential consumers among the 90% closest to the mean, to demonstrate the feasibility of our methods. We show that, considering only 2 days of consumption, we obtain physically interpretable clusters of consumers.
According to the Fig. 5.1, deal with individual consumption curves is an hard task, because of the high variability.

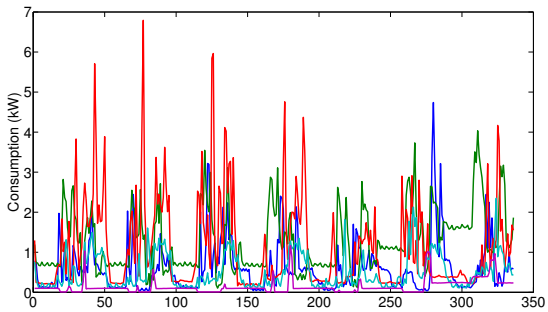

Figure 5.1: Load consumption of a sample of 5 consumers over a week in winter

5.2 Method

We propose to use model-based clustering and adopt the model selection paradigm. Indeed, we consider the model collection of conditional mixture densities,

$$\mathcal{S} = \left\{ s_\xi^{(K,J)}, K \in \mathcal{K}, J \in \mathcal{J} \right\},$$

where K denotes the number of clusters, J the set of relevant variables for clustering, and $\mathcal{K}$ and $\mathcal{J}$ being respectively the set of possible values of K and J.
The basic model we propose to use is a finite mixture regression of K multivariate Gaussian densities (see [SBvdGR10] for a recent and fruitful reference), the conditional density being, for $x \in \mathbb{R}^p, y \in \mathbb{R}^q$, φ denoting the Gaussian density,

$$s_\xi^{(K,J)}(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k^{[J]} x, \Sigma_k)$$

Such a model can be interpreted and used from two different viewpoints.
First, from a clustering perspective, given the estimation $\hat{\xi}$ of the parameters $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, we could deduce data clustering from the Maximum A Posteriori principle: for each observation i, we compute the posterior probability $\tau_{i,k}(\hat{\xi})$ of each cluster k from the estimation $\hat{\xi}$, and we assign the observation i to the cluster $\hat{k}_i = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\, \tau_{i,k}(\hat{\xi})$. Proportions of each cluster are estimated by $\hat{\pi}$.

Second, in each cluster, the corresponding model is meaningful and its interpretation allows to understand the relationship between variables Y and X since it is of the form

$$Y = X\beta_k + \epsilon,$$

the noise intensity being measured by Σ_k.
Parameters are estimated from the Lasso-MLE procedure, which is described in details in [Dev14c], and theoretically approved in [Dev14b]. To overcome the high-dimension issue, we use the Lasso estimator on the regression parameters and we restrict the covariance matrix to be diagonal. To avoid shrinkage, we estimate parameters by Maximum Likelihood Estimator on relevant variables selected by the Lasso estimator. Rather than selecting a regularization parameter, we present this issue at a model selection problem, considering a grid of regularization parameters. Indices of relevant variables for this grid of regularization parameters are denoted by $\mathcal{J}$. Since we also have to estimate the number of components, we compute those models for different number of components, belonging to $\mathcal{K}$. In this paper, $\mathcal{K} = \{1, \ldots, 8\}$.
Among this collection, we could focus on a few models which seem interesting for clustering, depending on which characteristics we want to highlight. We propose to use the slope heuristics to extract potentially interesting models. The selected model minimizes the log-likelihood penalized by $2\hat{\kappa}D_{(K,J)}/n$, where D_m denotes the dimension of the model m, and where $\hat{\kappa}$ is constructed from a completely data-driven procedure. In practice, we use the Capushe package, see [BMM12].
In addition to this family of models, we need to have powerful tools to translate curves into variables. Rather than dealing with the discretization of the load consumption, we project it onto a functional basis to take into account the functional structure. Since we are interested in not only representing the curves into a functional basis, but also to benefit from a time-scale interpretation of coefficients, we propose to use wavelet basis, see [Mal99] for a theoretical approach, and [MMOP07] for a practical purpose. To simplify our presentation, we will focus on the Haar basis.

5.3 Typical workflow using the example of the aggregated consumption

5.3.1 General framework

The goal is to cluster electricity consumers using a regression mixture model. We will consider the consumption of the eve for the regressors, to explain the consumption of the day. Consider the daily consumption, where we observe 48 points. We project the signal onto the Haar basis at level 4. The signal could be decomposed in approximation, denoted by A_4, and several details, denoted by D_4, D_3, D_2, and D_1. We illustrate it in Fig. 5.2, where in addition the decomposition in sum of orthogonal signals on the left, one can find a colored representation of the corresponding wavelet coefficients in the time scale plane. For an automatic denoising, we remove details of level 1 and 2, which correspond to high-frequency components. Two centerings will be considered:

- preprocessing 1: before projecting, we center each signal individually.
- preprocessing 2: we consider details coefficients of level 4 and 3. Here, we remove also a low-frequency approximation.

Depending on the preprocessing, we will get different clusterings
We observe the load consumption of n residentials over a year, denoted by $(z_{i,t})_{1\leq i\leq n, t\in T}$. We consider

- $Z_t = \sum_{i=1}^{n} z_{i,t}$ the aggregated signal,

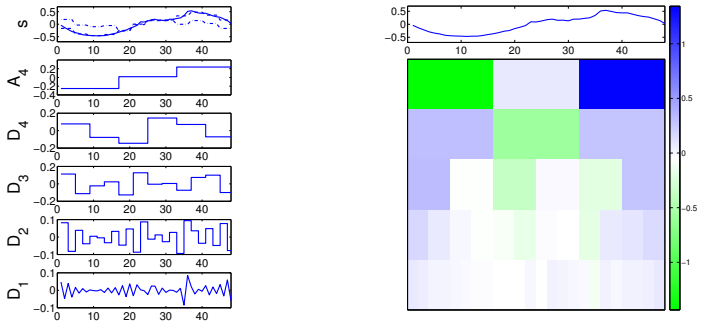

Figure 5.2: Projection of a load consumption for one day into Haar basis, level 4. By construction, we get $s = A_4 + D_4 + D_3 + D_2 + D_1$. On the left side, the signal is considered with reconstruction of dataset, the dotted being preprocessing 1 and the dotted-dashed being the preprocessing 2

— $\mathfrak{Z}_d = (Z_t)_{48(d-1)\leq t\leq 48d}$ the aggregated signal for the day d,
— $\mathfrak{z}_{i,d} = (z_{i,t})_{48(d-1)\leq t\leq 48d}$ the signal for the residential i for the day d.

We consider three different ways to analyze this dataset.
The first one consider $(\mathfrak{Z}_d, \mathfrak{Z}_{d+1})_{1\leq d\leq 338}$ over time, and the results are easy to interpret. We take this opportunity to develop in details the steps of the method we propose from the model to the clusters via model visualization and interpretation. In the second one, we want to cluster consumers on mean days. Working with mean days leads to some stability. The last one is the most difficult, since we consider individuals curves $(\mathfrak{z}_{i,d_0}, \mathfrak{z}_{i,d_0+1})_{1\leq i\leq n}$ and we classify these individuals for the days (d_0, d_0+1).

5.3.2 Cluster days on the residential synchronous curve

In this Section, we focus on the residential synchronous $(Z_t)_{t\in T}$. We will illustrate our procedure step by step, and highlight some features of data. The whole analysis will be done for the preprocessing 2.

Model selection

Our procedure leads to a model collection, with various number of components and various sparsities. Let us explain how to select some interesting models, thanks to the slope heuristic. We define

$$(K(\kappa), J(\kappa)) = \underset{(K,J)}{\operatorname{argmin}} \left(-\gamma_n(\hat{s}^{(K,J)}) + 2\kappa D_{(K,J)}/n \right),$$

where γ_n is the log-likelihood function and $\hat{s}^{(K,J)}$ is the log-likelihood minimizer among the collection $S^{(K,J)}$. We consider the step function $\kappa \mapsto D_{K(\kappa),J(\kappa)}$, $\hat{\kappa}$ being the abscissa which leads to the biggest dimension jump. We select the model $(\hat{K}, \hat{J}) = (K(2\hat{\kappa}), J(2\hat{\kappa}))$. To improve that, we could consider the points $(D_{(K,J)}, -\gamma_n(s^{(K,J)}) + 2\hat{\kappa}D_{(K,J)}/n)_{(K,J)}$, and select some models minimizing this criterion. According to Figs. 5.3 and 5.4, we could consider some $\hat{\kappa}$ which seem to create big jumps, and several models which seem to minimize the penalized log-likelihood.

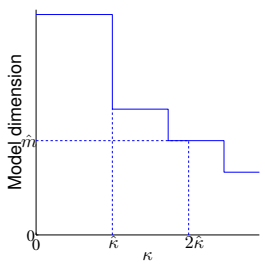

Figure 5.3: We select the model $\hat{m}$ using the slope heuristic

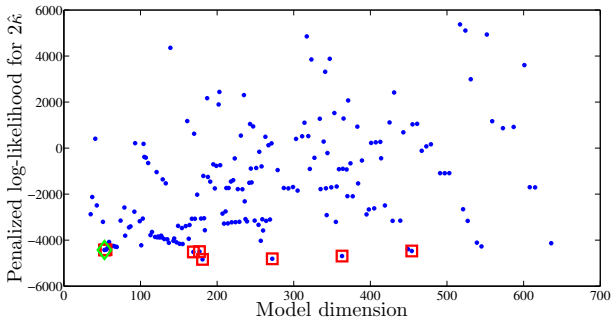

Figure 5.4: Minimization of the penalized log-likelihood. Interesting models are branded by red squares, the selected one by green diamond

Model visualization

Thanks to the model-based clustering, we have constructed a model in each cluster. Then, we could understand differences between clusters from $\hat{\beta}$ and $\hat{\Sigma}$ estimations. We represent it with an image, each coefficient being represented by a pixel. As we consider the linear model $Y = X\beta_k$ for each cluster, rows correspond to regressors coefficients and columns to response coefficients. Diagonal coefficients will explain the main part. The Figs. 5.5 and 5.6 explain the image construction, whereas we compute it for the model selected by the previous step in Fig. 5.7.

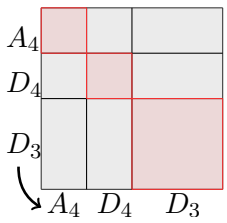

Figure 5.5: Representation of the regression matrix β_k for the preprocessing 1.

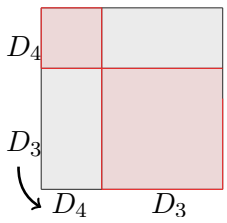

Figure 5.6: Representation of the regression matrix β_k for the preprocessing 2.

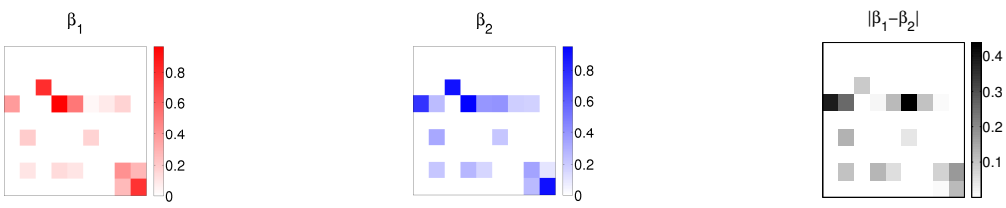

Figure 5.7: For the selected model, we represent $\hat{\beta}$ in each cluster. Absolute values of coefficients are represented by different colormaps, white for 0. Each color represents a cluster

To highlight differences between clusters, we also plot $\hat{\beta}_1 - \hat{\beta}_2$. First, we remark that $\hat{\beta}_1$ and $\hat{\beta}_2$ are sparse, thanks to the Lasso estimator. Moreover, the main difference between $\hat{\beta}_1$ and $\hat{\beta}_2$ is row 4, columns $1, 2$ and 6. We could say that the procedure uses, depending on cluster, more or less the first coefficient of D_3 of X to describe coefficients 1 and 2 of D_3 and coefficient 3 of D_4 of Y. The Fig. 5.11 enlightens those differences between clusters.

We represent the covariance matrix in Fig. 5.8. Because we estimate it by a diagonal matrix in each cluster, we just display the diagonal coefficients. We keep the same scale for all the clusters, to highlight which clusters are noisier.

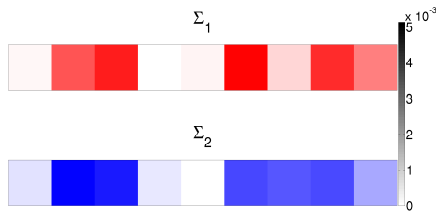

Figure 5.8: For the model selected, we represent Σ in each cluster

Model-based clustering

We could compute the posterior probability for each observation. In Fig. 5.9, we represent it by boxplots. Closer there are to 1, more different are the clusters.

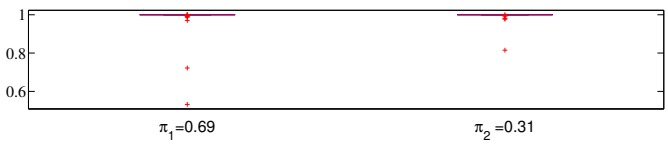

Figure 5.9: Assignment boxplots per cluster

In the present case, the two clusters are well defined and the clustering problem is quite easy, but see for example Fig. 5.13, in a different clustering issue, which presents a model with affectations less well separated.

Clustering

Now, we are able to try to interpret each cluster. In Fig. 5.10, we represent the mean curves for each cluster. We can also use functional descriptive statistics (see [Sha11]). Because clusters are done on the reliance between a day and its eve, we represent the both days.

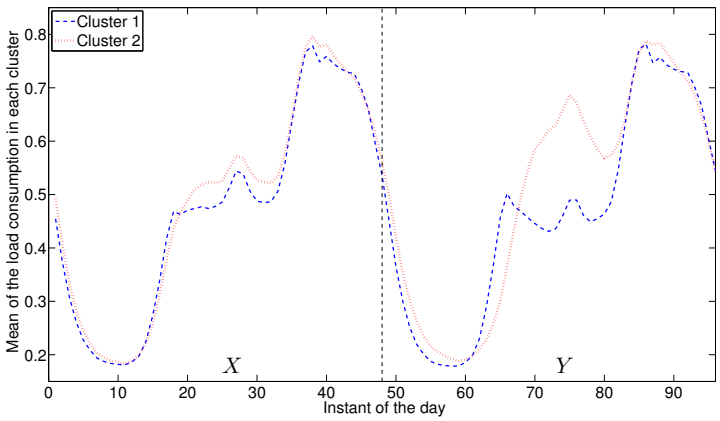

Figure 5.10: Clustering representation. Each curve is the mean in each cluster

Discussion

According to the preprocessing 2, we cluster weekdays and weekend days. The same procedure done with the preprocessing 1 shows the temperature influence. We construct four clusters, two of them being weekend days, and the two others are weekdays, differences made according to the temperature. In Fig. 5.11, we summarize clusters by the mean curves for this second model.

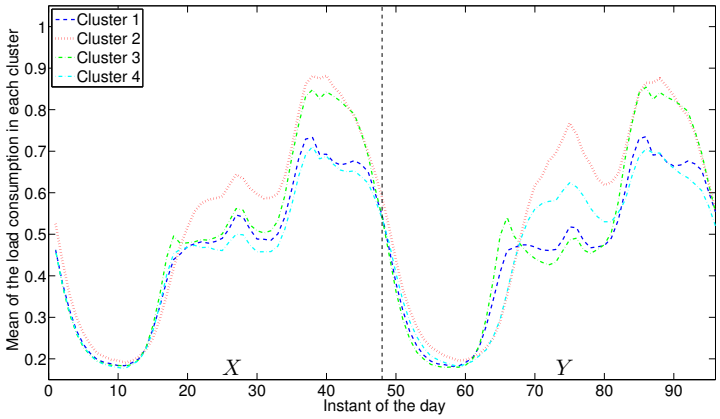

Figure 5.11: Clustering representation. Each curve is the mean in each cluster

In Table 5.1, we summarize the both models considered according to the day type.

Interpretation	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.
week	0.88	0.96	0.94	0.98	0.96	0	0
weekend	0.12	0.04	0.06	0.02	0.04	1	1
week, low T.	0.26	0.46	0.46	0.47	0.51	0	0
weekend, low T.	0.1	0.02	0.03	0	0	0.2	0.65
week, high T.	0.64	0.52	0.5	0.53	0.45	0	0
weekend, high T.	0	0	0	0	0.04	0.79	0.35

Table 5.1: For each model selected, we summarize the proportion of day type in each cluster, and interpret it, T corresponding to the temperature.

According to Table 5.1, difference between both preprocessing is the temperature influence: when we center curves before projecting, we translate the whole curves, but when we remove the low frequency approximation, we skip the main features. Depending on the goal, each preprocessing could be interesting.

5.4 Clustering consumers

Another important approach considered in this paper is to cluster consumers. Before dealing with their daily consumption, in Section 5.4.2, which is an hard problem because of the irregularity of signals, we cluster consumers on mean days in Section 5.4.1.

5.4.1 Cluster consumers on mean days

Define $\tilde{\mathfrak{z}}_{i,d}$ the mean signal over all days d for the residential i, over all days $d \in \{1, \ldots, 7\}$. Then we consider couples $(\tilde{\mathfrak{z}}_{i,d}, \tilde{\mathfrak{z}}_{i,d+1})_{1 \leq i \leq n}$.
If we look on the model collection constructed by our procedure for $\mathcal{K} = \{1, \ldots, 8\}$, we always select models with only one component, for every days d. Nevertheless, if we force the model to have several clusters, restricting $\mathcal{K}$ to $\mathcal{K}' = \{2, \ldots, 8\}$, we get some interesting informations. All results get here are done for preprocessing 2.
For weekdays couples, Monday/Tuesday, Tuesdays/Wednesday, Wednesday/Thursday, Thursday/Friday, we select models with two clusters, with same means and same covariance matrices: the model with one component is needed. The only difference on load consumption is on the mean comportment. It is relevant with clusterings obtained in Section 5.3.2.
We focus here on Saturday/Sunday, for which there are different interesting clusters, see Fig. 5.12. Remark that we cannot summarize a cluster to its mean because of the high variability. The main differences between those two clusters are on differences between lunch time and afternoon, and on the Sunday morning. Notice that the big variability over the two days is not explained by our model, for which the variability is small, explaining the noise for the reliance between a day and its eve.

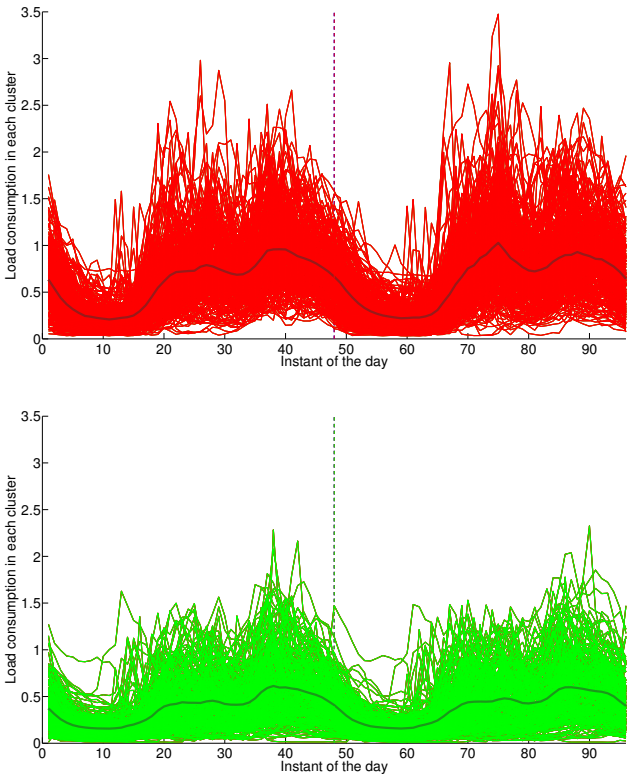

Figure 5.12: Saturday and Sunday load consumption in each cluster.

On Sunday/Monday, we get also three different clusters. Even if we identify differences on the shape, the main difference is still on the mean consumption. On Friday/Saturday, we see differences between people who have the same consumptions, and people who have a really different comportment.
However, because the selected model is again with one component, we think that consider the mean over days of each consumer cancel interesting effects, as the temperature and seasonal

influence.

5.4.2 Cluster consumers on individual curves

One major objective of this work is individual curves clustering. As already pointed in the introduction, this is a very challenging task that has various applications for smart grid management, going from demand response programs to energy reduction recommendations or household segmentation. We consider the complex situation where an electricity provider as access to very few recent measurements -2 days here- of each individual customers but need to classify them anyway. That could happen in a competitive electricity market where customers can change their electricity provider at any time without providing their past consumption curves.
First, we focus on two successive days of electricity consumption measurements - Tuesday and Wednesday in winter: January 5th and 6th 2010- for our 487 residential customers. Note that we choose week days in winter following our experience on electricity data, as those days often bring important information about residential electricity usage (electrical heating, intra-day cycle, tariff effect...).

Selected mixture models

We apply the model-based clustering approach presented in Section 5.2 for preprocessing 1 and obtain two models minimizing the penalized log-likelihood corresponding to 2 and 5 clusters. Although these two classifications are based on auto-regression mixture model, we are able to analyze and interpret clusters in terms of consumption profiles and provide bellow a physical interpretation for the classes. In Fig. 5.13, we plot the proportions in each cluster for both models constructed by our procedure. The first remark is that this issue is harder than the one in Section 5.3.2. Nevertheless, even if there are a lot of outliers, for the model M1, a lot of affectations are well-separated. It is obviously less clear with the model M2, with 5 components.

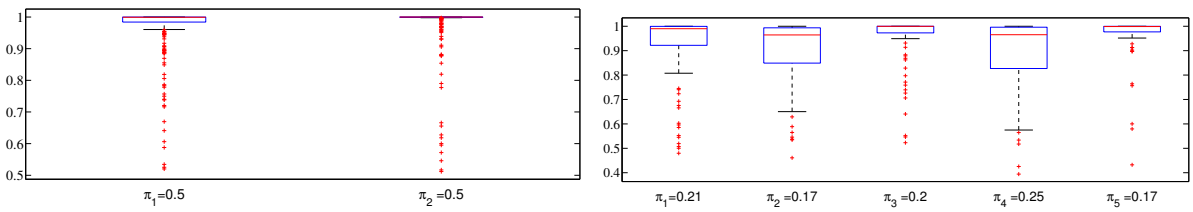

Figure 5.13: Proportions in each cluster for models constructed by our procedure

In Fig. 5.14, we plot the regression matrix to highlight differences between clusters. Indeed, those two matrices are different, for example more variables are needed to describe the cluster 2.

Analyzing clusters using classical profiling features

We first represent the main features of the cluster centres (the mean of all individual curves in a cluster). Fig. 5.15 represents daily mean consumptions of these centres along the year. We clearly see that the two classifications separate customers that have different mean level of consumption (small, -middle- and big residential consumers) and different ratio winter/summer consumption probably due to the amount of electrical heating among house usage. Let recall that the model based clustering is done on centered data and that the mean level discrimination is

Figure 5.14: Regression matrix in each cluster for the model with 2 clusters

not straightforward. Schematically, the 2 clusters classification seems to separate big customers with electrical heating from small customers with few electrical heating. Whereas the 5 clusters classification separates the small customers with few electrical heating but peaks in consumption (flat curve with peaks, probably due to auxiliary heating when temperature is very low) and middle customers with electrical heating. The two clusters in the middle customers population don't present any visible differences on this figure. The two big customers clusters have a different ratio winter/summer probably due to a difference in terms of electrical heating usage.

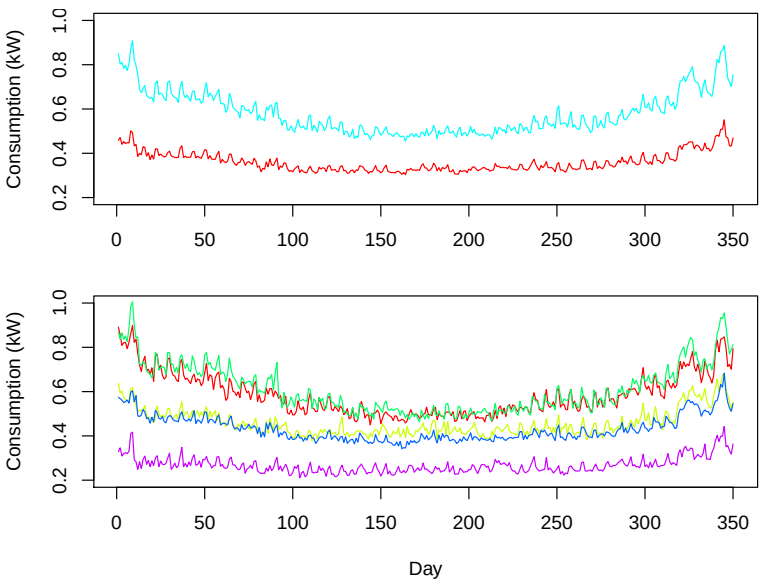

Figure 5.15: Daily mean consumptions of the cluster centres along the year for 2 (top) and 5 clusters (bottom)

This analysis is confirmed in Fig. 5.16 where we represent the daily mean consumptions of the cluster centres in function of the daily mean temperature for the two classifications. Points correspond to the mean consumption of one day and smooth curves are obtained with P-spline regression. We observe that for all classes, the temperature effect due to electrical heating starts at around 12°C and that the different clusters have various heating profiles. The 2 clusters classification profiles confirm the observation of Fig. 5.15 that the population is divided into small and big customers with electrical heating. Concerning the 5 clusters classification, we clearly see on the small customer profile (purple points/curves) an inflexion at around 0°C -this inflexion is also observed in the small customer cluster of the 2 clusters classification- corresponding to e.g. an auxiliary heating device effect or at least an increase of consumption

of the house for very low temperature. This is also what distinguishes the 2 middle customers classes (blue and green points/curves). The two big customers' clusters have similar heating profile, except that the green cluster correspond to higher electrical heating usage.

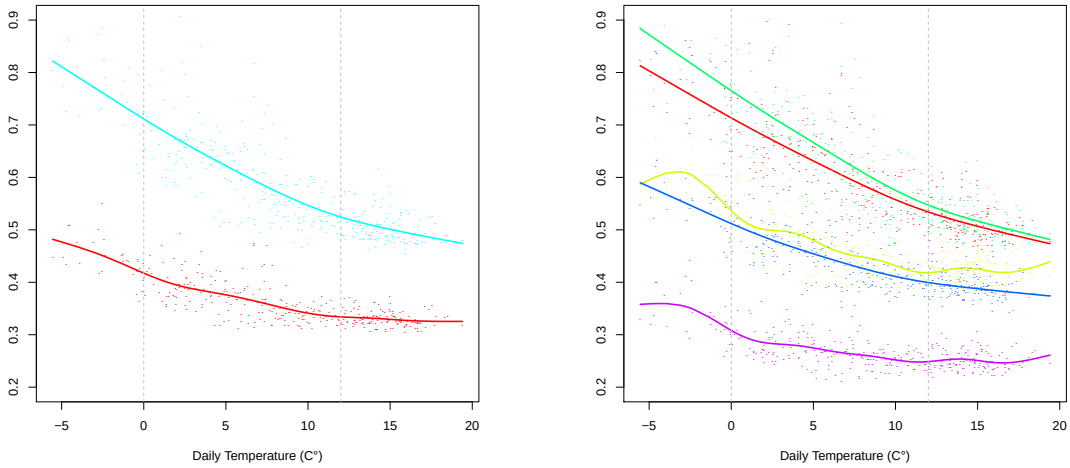

Figure 5.16: Daily mean consumptions of the cluster centres in function of the daily mean temperature for 2 (on the left) and 5 clusters (on the right)

Another interesting observation concern the weekly and daily profiles of the centres. We represent on Fig. 5.17 an average (over time) week of consumption for each centre of the two classifications. For the 2 clusters classification, we see again the difference in average consumption between the big customer cluster profile and the small customer one. They have similar shapes but the difference day/night and peak loads (at around 8h, 13h and 18h for weekdays), are more marked. For the 5 clusters curves, even if the weekly profiles are quite similar (no major difference in the week days/week ends profiles in each clusters), the daily shapes exhibit more differences. The day/night ratio could be very different as well as small variation along the day, probably related to different tariff options (see [Com11a] for a description of the tariffs).

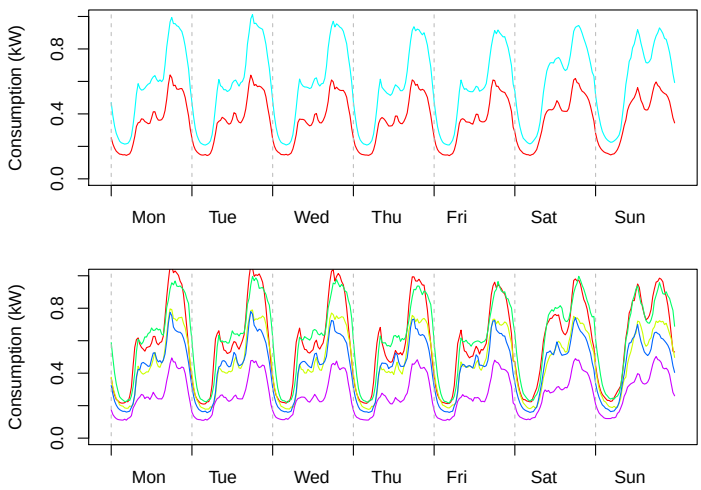

Figure 5.17: Average (over time) week of consumption for each centre of the two classifications (2 clusters on the top and 5 on the bottom)

Cross analysis using survey data

To enrich this analysis, we also consider extra information providing in a survey realized by the Irish Commission for Energy Regulation. We summarize this large amount of information into 10 variables: *ResidentialTariffallocation* corresponds to the tariff option (see [Com11a], [Com11b]), *Socialclass*: AB, C1, C2, F in UK demographic classifications, *Ownership* whether or not a customer owns his/her house, *ResidentialStimulusallocation* the stimulus sends to the customer (see [Com11a], [Com11b]), *Built.Year* the year of construction of the building, *Heat.Home* and *Heat.Water* electrical or not, *Windows.Doubleglazed* the proportion of double gazed window in the house (none, quarter, half, 3 quarters, all), *Home.Appliance.White.goods* the number of white goods/major appliances of the household. To relate our clusters to those variables we consider the classification problem consisting in recovering model based classes with a random forest classifier. Random forest introduced in [Bre01] is a well known and tested non-parametric classification method that has the advantage to be quite automatic and easy to calibrate. In addition, it provides a nice summary of the previous covariate in terms of their importance for classification. On the Fig. 5.18 we represent the out of bag error of the random forest classifiers in function of the number of trees (one major parameter of the random forest classifier) for the two clusterings. That corresponds to a good estimate of what could be the classification error on a independent dataset. We have observed that, choosing a sufficiently large number of trees for the forest (300), the classification error rate attains 40% in the 2 clusters case and 75% in the 5 classes case which has to be compared to a random classifier who has respectively a 50% and 80% error rate. That means that the 10 previous covariates provide few but some information about the clusters.

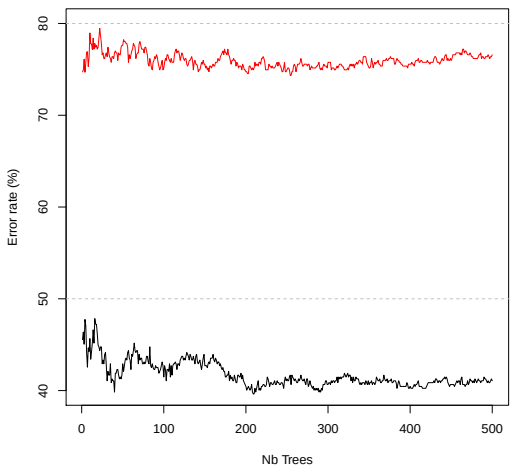

Figure 5.18: Out of bag error of the random forest classifiers in function of the number of trees

Quantifying the importance of each survey covariate in the classifications we observe that in both cases, *Home.Appliance.White.goods* and *Socialclass* play a major role. That could be explained as those covariates could discriminate small and big customers. Another interesting point is that in the 5 clusters classification, the variable *Built.Year* plays an important role which is probably related to different heating profiles explained by different isolation standards. That could explain the two big customers clusters. Then come the tariffs options which, in the 5 clusters case, could explain the different daily shapes of the Fig. 5.17.
It is noteworthy that these two classifications provide clusters that exhibit those very nice

physical interpretation, considering that we only use two days of consumption in winter for each customer.

Using model on one-day shifted data

To highlight advantages of our procedure, we compare prediction for the consumption of Thursday, 7th January, 2010. Indeed, even if the method is not designed for forecasting purposes, we want to show that model-based clustering is an interesting tool also for prediction. We will compare linear models, estimated on couples Tuesday, 5th January and Wednesday, 6th January, and we will predict Thursday from Wednesday. This is suggested by the clustering get in Section 5.3.2, showing that transitions between weekdays are similar. We then compare the following models. First, the most common is the linear model, without clustering. The second model is the first constructed by our procedure, described before, with 2 components. Moreover, we could use the clustering get by the models constructed by our procedure, but estimate parameters without variable selection, using full linear model in each component. We restrict here our study to one model to narrow the analysis, but everything is also computable with the models with 5 clusters.
For each consumer i, for each prediction procedure, we compute two prediction errors: the RMSE on Thursday prediction, and the RMSE of Wednesday prediction. Remind that RMSE, for a consumer i, is defined by

$$RMSE(i) = \sqrt{\frac{1}{48}\sum_{t=1}^{48}(\hat{z}_{i,t} - z_{i,t})^2}.$$

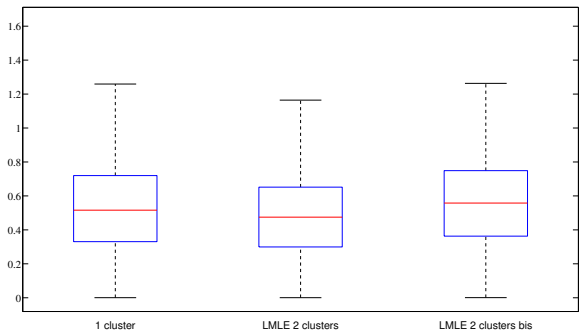

Figure 5.19: RMSE on Thursday prediction for each procedure over all consumers

We remark that if RMSE are almost the same for the three different models, the one estimated by our procedure leads to smaller median and interquartile range. For the three considered models, the median of the RMSE on Wednesday prediction (learning sample) and the RMSE on Thursday prediction (test sample) are close to each other, which means that the clustering remains good, even for one-day shifted data, of course as long as we remain in the class of working days, according to Section 5.3.2. To highlight this remark, we also compute our procedure on couple Wednesday/Thursday. Then, we select three different models, and involved clusterings are quite similar to clusterings given by models in Section 5.4.2.

Remarks on similar analyses

Alternatively, we make the same analysis on two successive weekdays of electricity consumption measurement in summer. We obtain three models, corresponding to 2, 3 and 5 clusters respectively. We compute, as in the Subsection 5.4.2, daily mean consumptions of the cluster center along the year, and in function of the daily mean temperature. The main difference is about the inflexion at around 0°C. Because clustering is done for summer days, we do not distinguish cold effects. Moreover there are no cooling effects. We could remark again that clusterings are hierarchical, but different from those get in the winter study, as we expected.

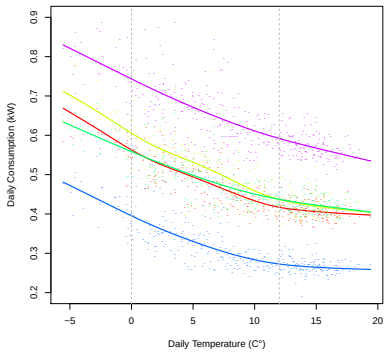

Figure 5.20: Daily mean consumptions of the cluster centres in function of the daily mean temperature for 5 clusters, clustering done by observing Thursday and Wednesday in summer

We also study two successive weekend days of electricity consumption, in winter and in summer. We recognize different clusters, depending on behavior of consumers. We work with Friday/Saturday couples. The main thing we observe in summer is a cluster with no-consumption, consumers who leave their home. It could be useful to predict the Sunday consumption, but no more general for other weekend.

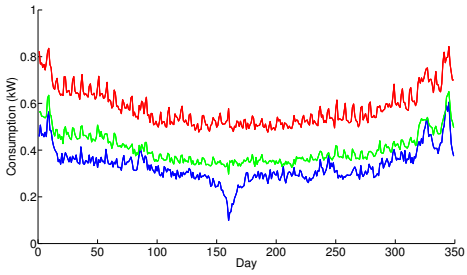

Figure 5.21: Daily mean consumptions of the cluster centres along the year for 3 clusters, clusering done on weekend observation

5.5 Discussion and conclusion

Massive information about individual (household, small and medium enterprise) consumption are now provided with new metering technologies and smart grids. Two major exploitations of individual data are load profiling and forecasting at different scales on the grid. Customer segmentation based on load classification is a natural approach for that and is a prolific way of research. We propose here a new methodology based on high-dimensional regression models.

The novelty of our approach is that we focus on uncovering clusters corresponding to different regression models that could then be exploited for profiling as well as forecasting. We focus on profiling and show how, exploiting few temporal measurements of 500 residential customers consumption, we can derive informative clusters. We provide some encouraging elements about how to exploit these models and clusters for bottom up forecasting.

Bibliography

[Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[Ald96] M. Aldrin. Moderate projection pursuit regression for multivariate response data. *Computational Statistics & Data Analysis*, 21(5):501–531, 1996.

[And51] T. W. Anderson. Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

[AS13] C. Alzate and M. Sinn. Improved electricity load forecasting via kernel spectral clustering of smart meters. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 943–948, 2013.

[ASS98] C.W. Anderson, E.A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, 1998.

[Bah58] R. R. Bahadur. Examples of inconsistency of maximum likelihood estimates. *Sankhya: The Indian Journal of Statistics (1933-1960)*, 20(3/4):pp. 207–210, 1958.

[Bau09] J-P Baudry. Model Selection for Clustering. Choosing the Number of Classes. *Ph.D. thesis, Université Paris-Sud 11*, 2009.

[BC11] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

[BC13] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[BCG00] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, July 2000.

[BCG03] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

[BGH09] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *The Annals of Statistics*, 37(2):630–672, 2009.

[BKM04] E. Brown, R. Kass, and P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, 2004.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BM93] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2), 2007.

[BMM12] J-P Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[Bre01] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[BSW11] F. Bunea, Y. She, and M. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[BSW12] F. Bunea, Y. She, and M. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388, 2012.

[Bun08] F. Bunea. Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH, 2008.

[BvdG11] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

[CDS98] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *Siam Journal On Scientific Computing*, 20:33–61, 1998.

[CG93] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. Technical Report RR-2028, INRIA, 1993.

[Cha14] M. Chaouch. Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves. *IEEE Transactions on Smart Grid*, 5(1):411–419, 2014.

[CLP11] S. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. Research Report RR-7596, 2011.

[CO14] A. Ciarleglio and T. Ogden. Wavelet-based scalar-on-function finite mixture regression models. *Computational Statistics & Data Analysis*, 2014.

[Com11a] Commission for energy regulation, Dublin. Electricity smart metering customer behaviour trials findings report. 2011.

[Com11b] Commission for energy regulation, Dublin. Results of electricity coast-benefit analysis, customer behaviour trials and technology trials commission for energy regulation. 2011.

[CT07] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[Dev14a] E. Devijver. An ℓ_1-oracle inequality for the lasso in finite mixture of multivariate gaussian regression models, 2014. arXiv:1410.4682.

[Dev14b] E. Devijver. Finite mixture regression: A sparse variable selection by model selection for clustering, 2014. arXiv:1409.1331.

[Dev14c] E. Devijver. Model-based clustering for high-dimensional data. application to functional data, 2014. arXiv:1409.1333.

[Dev15] E. Devijver. Joint rank and variable selection for parsimonious estimation in high-dimension finite mixture regression model, 2015. arXiv:1501.00442.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Discussion. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[FHP03] K.J. Friston, L. Harrison, and W.D. Penny. Dynamic Causal Modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[FP04] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[FR00] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2000.

[FV06] F. Ferraty and P. Vieu. *Nonparametric functional data analysis : theory and practice*. Springer series in statistics. Springer, New York, 2006.

[Gir11] C. Giraud. Low rank multivariate regression. *Electronic Journal of Statistics*, 5:775–799, 2011.

[GLMZ10] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

[GW00] C. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Annals of Statistics*, 28(4):1105–1127, 2000.

[H+13] L. Hancher et al. Think topic 11: 'Shift, not drift: Towards active demand response and beyond'. 2013.

[HK70] A. Hoerl and R. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[Irw86] G.W. Irwin. Statistical electricity demand modelling from consumer billing data. *IEE Proceedings C (Generation, Transmission and Distribution)*, 133:328–335(7), September 1986.

[Ize75] A. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

[KFR14] J. Kwac, J. Flora, and R. Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1):420–430, 2014.

[LSD15] W. Labeeuw, J. Stragier, and G. Deconinck. Potential of active demand reduction with residential wet appliances: A case study for belgium. *Smart Grid, IEEE Transactions on*, 6(1):315–323, Jan 2015.

[Mal73] C. L. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973.

[Mal89] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[Mal99] S. Mallat. *A wavelet tour of signal processing.* Academic Press, 1999.

[Mas07] P. Massart. *Concentration inequalities and model selection.* Lecture Notes in Mathematics. Springer, 33, 2003, Saint-Flour, Cantal, 2007.

[MB88] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York, 1988.

[MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[MB10] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[Mey12] C. Meynet. Sélection de variables pour la classification non supervisée en grande dimension. *Ph.D. thesis, Université Paris-Sud 11*, 2012.

[Mey13] C. Meynet. An ℓ_1-oracle inequality for the lasso in finite mixture gaussian regression models. *ESAIM: Probability and Statistics*, 17:650–671, 2013.

[MK97] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley & Sons, New York, 1997.

[MM11a] P. Massart and C. Meynet. The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5:669–687, 2011.

[MM11b] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68, 2011.

[MMOP04] M. Misiti, Y. Misiti, G. Oppenheim, and J-M. Poggi. *Matlab Wavelet Toolbox User's Guide. Version 3.* The Mathworks, Inc., Natick, MA., 2004.

[MMOP07] M. Misiti, Y. Misiti, G. Oppenheim, and J-M Poggi. Clustering signals using wavelets. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Computational and Ambient Intelligence*, volume 4507 of *Lecture Notes in Computer Science*, pages 514–521. Springer Berlin Heidelberg, 2007.

[MMOP10] M. Misiti, Y. Misiti, G. Oppenheim, and J-M Poggi. Optimized clusters for disaggregated electricity load forecasting. *REVSTAT*, 8(2):105–124, 2010.

[MMR12] C. Meynet and C. Maugis-Rabusseau. A sparse variable selection procedure in model-based clustering. Research report, September 2012.

[MP04] G. McLachlan and D. Peel. *Finite Mixture Models.* Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2004.

[MS14] Z. Ma and T. Sun. Adaptive sparse reduced-rank regression, 2014. arXiv:1403.1922.

[MY09] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[OPT99] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999.

[PC08] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[PS07] W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007.

[Ran71] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[RD06] A. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101:168–178, 2006.

[RS05] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer series in statistics. Springer, New York, 2005.

[RT11] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[SBG10] N. Städler, P. Bühlmann, and S. Van de Geer. ℓ_1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[SBvdGR10] N. Städler, P. Bühlmann, S. van de Geer, and Rejoinder. Comments on ℓ_1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[Sch78] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

[Sha11] H-L Shang. rainbow: An R Package for Visualizing Functional Time Series . *The R Journal*, 3(2):54–59, dec 2011.

[SWF12] W. Sun, J. Wang, and Y. Fang. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012.

[SZ12] T. Sun and C-H Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, 58(1):267–288, 1996.

[TMZT06] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

[TSM85] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1985.

[TSR+05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

[US 06] US Department of Energy . Benefits of demand response in electricity markets and recommendations for achieving them - a report to the united states congress pursant to section 1252 of the energy policy act of 2005. page 122, 02/2006 2006.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[vdG13] S. van de Geer. Generic chaining and the ℓ_1-penalty. *Journal of Statistical Planning and Inference*, 143(6):1001 – 1012, 2013.

[vdGB09] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[vdGBRD14] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014.

[vdGBZ11] S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

[vdVW96] AW van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

[YB99] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[YFL11] F. Yao, Y. Fu, and T. Lee. Functional mixture regression. *Biostatistics*, (2):341–353, 2011.

[YLL06] M. Yuan, Y. Lin, and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[YLL12] M-S Yang, C-Y Lai, and C-Y Lin. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.

[ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[ZH08] C-H Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[Zha10] C-H Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[ZHT07] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

[ZOR12] Y. Zhao, T. Ogden, and P. Reiss. Wavelet-Based LASSO in Functional Linear Regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617, 2012.

[Zou06] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[ZPS09] H. Zhou, W. Pan, and X. Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009.

[ZY06] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[ZYS13] K-L Zhou, S-L Yang, and C. Shen. A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews*, 24(0):103 – 110, 2013.

[ZZ10] N. Zhou and J. Zhu. Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, 3:557–574, 2010.