N° d'ordre

Ibrahim SIDI ZAKARI SELECTION DE VARIABLES ET REGRESSION SUR LES QUANTILES 2013 M

UNIVERSITÉ CADI AYYAD
FACULTÉ DES SCIENCES
SEMLALIA – MARRAKECH

UNIVERSITÉ LILLE1
SCIENCES ET
TECHNOLOGIES

N° d'ordre :

****

THÈSE

présentée à la Faculté pour obtenir le grade de :

Docteur

CED : Sciences et Techniques (Université Cadi Ayyad)

ED : Sciences Pour l'Ingénieur (Université Lille 1)

Spécialité : Statistiques

SELECTION DE VARIABLES ET REGRESSION SUR LES QUANTILES

par :

Ibrahim SIDI ZAKARI

(Dernier diplôme: Master Spécialité : Mathématiques)

Soutenue le **10 Juillet 2013** devant la commission d'examen :

Président :	**Youssef OUKNINE**	PES	Université Cadi Ayyad (Maroc)
Examinateurs :	**Abdallah MKHADRI**	PES	Université Cadi Ayyad (Maroc)
	Abdelaziz NASROALLAH	PES	Université Cadi Ayyad (Maroc)
	Assi N'GUESSAN	PH	Université Lille 1 (France)
	Célestin KOKONENDJI	PES	Université de Franche-Comté (France)
	Sophie DABO-NIANG	PES	Université Lille 3 (France)

Thèse

Présentée pour obtenir le grade de

Docteur De L'Université Cadi Ayyad et L'Université Lille 1

Discipline :

Mathématiques appliquées

Spécialité :

Statistique

par

SIDI ZAKARI Ibrahim

Sélection de variables et régression sur les quantiles.

Travail effectué sous la direction de :

MKHADRI Abdallah, Professeur, Université Cadi Ayyad (FSSM).
N'GUESSAN Assi Maître de Conférence HDR, Université Lille 1 (Polytech'lille).

Année universitaire 2012-2013.

FICHE PRÉSENTATIVE DE LA THÈSE

-Nom et Prénom de l'auteur : SIDI ZAKARI Ibrahim

-Intitulé du travail : Sélection de variables et régression sur les quantiles

-Encadrant :

MKHADRI Abdallah, PES
LIB-MA, Université Cadi Ayyad

-Co-Encadrant :

N'GUESSAN Assi, PH
Paul Painlevé, Université Lille 1

-Lieux de réalisation des travaux (laboratoires, institution,...) :

LIB-MA et Paul Painlevé

-Période de réalisation du travail de thèse : Décembre 2008-Mai 2013

-Rapporteurs autres que l'encadrant (nom, prénom, grade, institution)

(Pour l'Université Cadi Ayyad)
-El Mostapha QANNARI PES Université de Nantes (France)
-Gilles CELEUX PES Université Orsay (France)
-Sophie DABO-NIANG PES Université Lille 3 (France)

(Pour l'Université Lille 1)
-Célestin KOKONENDJI PES Université de Franche-Comté (France)
-Gilles CELEUX PES Université Orsay (France)

-Cadres de coopération (ou de soutien) :

Cotutelle de thèse dans le cadre des mobilités internationales de l'Agence Universitaire de la Francophonie (AUF).

-Ce travail a donné lieu aux résultats suivants (communications, publications,...) :

-"A mixture of local and quadratic approximation variable selection algorithm in nonconcave penalized regression."Sidi Zakari I., Mkhadri A., N'Guessan A."ARIMA Journal, Vol. 16, 29-46 (2013).

-"Stability selection and randomization in L_1 quantile regression."Sidi Zakari I., Mkhadri A., N'Guessan A. Contributed paper to appear in the Proceedings of the 15th Applied Stochastic Models and Data Analysis International Conference, 25 - 28 June 2013, Mataró (Barcelona, Spain).

-"Variables Selection and Quantile regression : Application to Merlan fish data". Communication à la 1ère édition des journées de Probabilités et Statistique, 15-17 décembre 2011, (ENSA Marrakech).

-"Smoothly Clipped Absolute Deviation for Correlated variables". Communication à la 5 ème Conférence COMPUTATIONAL STATISTICS (COMPSTAT2010) organisée par l'Institut National de Recherche en Informatique et Automatique(INRIA) et le Conservatoire National des Arts et Métiers (CNAM).Paris (France), du 22 au 27 Août 2010.

-"Sélection de modèles de régression par augmentation de dimension". Communication à la 2ème conférence de la Société Marocaine de Mathématiques Appliquées (SM2A). Rabat (Maroc), du 28 -30 Juin 2010.

-"Sélection de variables pour modèles de régression en grande dimension". Communication à la Vème Conférence Internationale en Recherche Opérationnelle (CIRO'10). Marrakech (Maroc), 24-27 Mai 2010.

REMERCIEMENTS

Je tiens tout d'abord à remercier mes parents **M. Zakari SIDI** et **Mme Fatouma NA-ANY** pour m'avoir soutenu durant l'élaboration de ce travail. Ils ont beaucoup contribué dans mes prises de décision ainsi que mes orientations académiques. Je leur dédie cette thèse ainsi qu'à tous ceux qui me sont chers.
Le présent travail n'aurait pu être réalisé sans l'implication de mes directeurs de thèse, les Professeurs **Abdallah MKHADRI** du Laboratoire Ibn al Banna de Mathématiques et Applications (LIB-MA, Université Cadi Ayyad) et **Assi N'GUESSAN** du Laboratoire Paul Painlevé (Université Lille 1). Leurs conseils, leur vision ainsi que les discussions directement ou indirectement liées à ce travail m'ont été d'une grande utilité. Je ne saurai les remercier assez pour tout ce qu'ils ont fait et continuent de faire malgré leurs multiples engagements. Les mots ne sauraient suffir en guise de remerciements.
Mes remerciements vont également à l'endroit des Professeurs **Célestin KOKONENDJI** du Laboratoire de Mathématiques de Besançon (Université de Franche-Comté), **El Mostafa QANNARI** de l'Unité Sensométrie et Chimiométrie à Oniris-Nantes (Université de Nantes), **Gilles CELEUX** Directeur de recherche à L'Inria (Université Paris Orsay) et **Sophie DABO-NIANG** du Laboratoire Economie QUantitative Intégration Politiques Publiques Econométrie (EQUIPPE, Université Lille 3) pour avoir accepté d'être rapporteurs de cette thèse. De même, le Professeur **Youssef OUKNINE** du LIB-MA et de l'Académie Hassan II des Sciences et Techniques pour avoir accepté de présider le jury de soutenance ainsi que le Professeur **Abdelaziz NASROALLAH** du LIB-MA pour avoir accepté d'en faire partie. A eux tous, je réitère mes sincères remerciements.
Une mention spéciale aux partenaires institutionnels ayant contribué à la réalisation du présent travail notamment l'Agence Nigérienne des Allocations et

Bourses (ANAB), l'Agence Marocaine à la Coopération Internationale (AMCI), l'Agence Universitaire de la Francophonie (AUF) et le Laboratoire des produits de la pêche de Boulogne-sur-Mer à travers M. **Alexandre DEHAUT** de l'Agence nationale de sécurité sanitaire, de l'alimentation, de l'environnement et du travail (Anses) pour nous avoir permis d'utiliser les données sur le merlan ainsi que pour les informations fournies à ce sujet.
Je ne saurai terminer sans remercier le Professeur **Abderrahim MAKKI NACIRI** pour les échanges fructueux durant mon cursus universitaire, les collègues des laboratoires LIB-MA et Paul Painlevé ainsi que tous ceux qui de près ou de loin ont contribué à l'aboutissement de ce travail.

Résumé

Ce travail est une contribution à la sélection de modèles statistiques et plus précisément à la sélection de variables dans le cadre de régression linéaire sur les quantiles pénalisés lorsque la dimension est grande. On se focalise sur deux points lors de la procédure de sélection : la stabilité de sélection et la prise en compte de variables présentant un effet de groupe. Dans une première contribution, on propose une transition des moindres carrés pénalisés vers la régression sur les quantiles (QR). Une approche de type bootstrap fondée sur la fréquence de sélection de chaque variable est proposée pour la construction de modèles linéaires (LM). Dans la majorité des cas, l'approche QR fournit plus de coefficients significatifs. Une deuxième contribution consiste à adapter certains algorithmes de la famille "Random" LASSO (Least Absolute Solution and Shrinkage Operator) au cadre de la QR et à proposer des méthodes de stabilité de sélection. Des exemples provenant de la sécurité alimentaire illustrent les résultats obtenus. Dans le cadre de la QR pénalisée en grande dimension, on établit la propriété d'effet groupement sous des conditions plus faibles ainsi que les propriétés oracles. Deux exemples de données réelles et simulées illustrent les chemins de régularisation des algorithmes proposés. La dernière contribution traite la sélection de variables pour les modèles linéaires généralisés (GLM) via la vraisemblance non concave pénalisée. On propose un algorithme pour maximiser la vraisemblance pénalisée pour une large classe de fonctions de pénalité non convexes. La propriété de convergence de l'algorithme ainsi que la propriété oracle de l'estimateur obtenu après une itération ont été établies. Des simulations ainsi qu'une application sur données réelles sont également présentées.

Summary

This work is a contribution to the selection of statistical models and more specifically in the selection of variables in penalized linear quantile regression when the dimension is high. It focuses on two points in the selection process : the stability of selection and the inclusion of variables by grouping effect. As a first contribution, we propose a transition from the penalized least squares regression to quantiles regression (QR). A bootstrap approach based on frequency of selection of each variable is proposed for the construction of linear models (LM). In most cases, the QR approach provides more significant coefficients. A second contribution is to adapt some algorithms of "Random" LASSO (Least Absolute Shrinkage and Solution Operator) family in connection with the QR and to propose methods of selection stability. Examples from food security illustrate the obtained results. As part of the penalized QR in high dimension, the grouping effect property is established under weak conditions and the oracle ones. Two examples of real and simulated data illustrate the regularization paths of the proposed algorithms. The last contribution deals with variable selection for generalized linear models (GLM) using the nonconcave penalized likelihood. We propose an algorithm to maximize the penalized likelihood for a broad class of non-convex penalty functions. The convergence property of the algorithm and the oracle one of the estimator obtained after an iteration have been established. Simulations and an application to real data are also presented.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

La présente thèse s'inscrit dans le cadre général de la sélection de variables pour la régression linéaire pénalisée sur les quantiles. Par ailleurs une partie est consacrée à la sélection de variables dans le cadre de la vraisemblance pénalisée pour les modèles linéaires généralisés. Elle est composée de cinq chapitres complétés par une annexe.

S'agissant de la régression sur les quantiles, nous pouvons dire qu'elle étend la notion de régression ordinaire aux quantiles de la variable à expliquer. Ce qui nous donne plus d'information sur la distribution de cette variable, car les quantiles sont des points particuliers pris à des intervalles réguliers à partir de la fonction de répartition d'une variable aléatoire. Nous rappelons que la régression ordinaire est un modèle pour la moyenne conditionnelle, où le conditionnement se fait par rapport aux variables explicatives. De manière similaire la régression sur les quantiles est un modèle pour les quantiles conditionnels. Cette approche peut être considérée comme semi-paramétrique car elle est basée sur le quantile (non-paramétrique), mais utilise des paramètres afin de fournir la relation entre le quantile et les variables explicatives. D'autre part, le modèle paramétrique requiert parfois des hypothèses restrictives et est plus sensible aux valeurs aberrantes. D'un point de vue méthodologique, cette approche peut être utilisée aussi bien pour les modèles linéaires que non linéaires (cf. Koenker[86]). Nous précisons par ailleurs que nous nous sommes principalement intéressés aux modèles linéaires. Les quantiles de régression sont plus efficaces que les estimations utilisant les moindres carrés notamment dans le cas des erreurs non gaussiennes. La représentation sous forme de programmation linéaire facilite l'estimation des quantiles de régression.

Le premier chapitre traite des généralités sur le concept de la régression quantile. Dans le cas non pénalisé, nous présentons les aspects méthodologiques, les propriétés d'équivariance, d'unicité ou non de la solution ainsi que la distribution asymptotique de l'estimateur obtenu. Cette partie est complétée par un exemple illustratif du domaine de la microéconomie. Les notions de scores de rang de régression, le concept de régression quantile composite ainsi que les conditions d'optimalité lors de l'estimation des paramètres sont également abordés. S'agissant de la sélection de variables qui correspond au cas pénalisé, nous avons fait une présentation de l'état de l'art incluant également les modèles de vraisemblance. D'autre part, le choix des paramètres de pénalité est d'une importance capitale non seulement pour le modèle final mais aussi pour les chemins de solution ou de régularisation. L'impact et la réduction de la dimension, qui sont des sujets d'actualité sont sans nul doute les motivations de la sélection de variables. Les aspects numériques ont été également abordés ; toutefois afin de ne pas alourdir ce chapitre, les détails algorithmiques sont laissés en annexe.

Le deuxième chapitre constitue une transition des moindres carrés pénalisés vers la régression sur les quantiles à travers une application réelle sur des données issues du domaine de la pêche (cf. Duflos et al.[38]). En effet, l'intérêt de la régression sur les quantiles par rapport à la regression linéaire classique n'est plus à présenter vu les innombrables contributions sur ce sujet[[86],[79],[16]]. L'objectif de ce chapitre est d'identifier les composés volatiles permettant de juger de la bonne ou mauvaise qualité du poisson. Afin de sélectionner les variables à inclure dans le modèle final nous avons évalué les fréquences de sélection de chaque variable en utilisant une approche de type bootstrap. La particularité est que nous considérons deux indicateurs qui représentent les variables à expliquer, ceci du fait que lors de l'évaluation de la qualité du merlan deux méthodes de notations ont été utilisées. Un modèle de régression linéaire est proposé pour chaque indicateur et les coefficients de régression correspondants sont comparés avec ceux obtenus avec la régression quantile. Dans la majorité des cas, la régression sur les quantiles fournit plus de coefficients significatifs. L'une des difficultés rencontrées dans cette partie est le fait que beaucoup de variables sont fortement corrélées ce qui a nécessité l'utilisation de méthodes de sélection de variables prenant en compte cet aspect.

Le troisième chapitre regroupe les principales méthodes de sélection de variables combinant la randomisation, le bootstrap tant au niveau de l'échantillon qu'au niveau des variables. A notre connaissance une telle approche n'a pas encore été traitée dans le cas de la régression sur les quantiles. La principale contribution de ce chapitre est l'adaptation des algorithmes Random Lasso (Wang et al.[145]), BoLasso (Bach[9]) au cadre de la régression sur les quantiles ainsi que les méthodes de stabilité de sélection (Meinshausen and Bühlmann[108]) dans le cas randomisé et non randomisé. L'utilité de la stabilité de sélection réside dans le fait de sélectionner les variables pertinentes tout en contrôlant le nombre de variables bruitées. Les aspects théoriques développés par Meinshausen and Bühlmann[108] ne peuvent être étendus au cas de la régression quantile du fait de la non différentiabilité de la fonction de perte correspondante ainsi que l'absence de solution explicite contrairement au cas des moindres carrés. Une version courte de ce chapitre a été acceptée comme contributed paper dans les proceedings de la conférence internationale ASMDA 2013[133] (`http://www.asmda.es/`).

Le quatrième chapitre regroupe l'aspect d'effet groupement et les propriétés d'oracle en considérant les pénalités de type Elastic net (Zou and Hastie[160]), Adaptive Elastic net (Zou and Zhang[165]), Berhu (Owen[114]) et Berhu adaptatif (Lambert-Lacroix and Zwald[100]). Une première approche a été proposée par Slawski[134] dans le cas de l'Elastic net avec la fonction de perte de la régression quantile (QR) et des SVM (Support Vector Machine) mais sans développement théorique. Dans le cas de l'Elastic net (QR), nous avons pu établir la propriété d'effet groupement avec de plus faibles conditions que celles utilisées par Zou and Hastie[160]. Dans le cas de l'Elastic net adaptatif (QR) les conditions utilisées sont les mêmes que celles de Ghosh[58], El Anbari and Mkhadri[43] dans le cas des moindres carrés. Nous avons également établi les propriétés d'oracle de l'estimateur de la régression quantile combinée avec l'Elastic net adaptatif. Enfin, nous avons pu établir les propriétés d'effet groupement dans le cas de l'estimateur de la régression quantile pénalisée avec la fonction du Berhu adaptatif. Les poids adaptatifs au niveau des termes de pénalité permettent d'avoir les propriétés d'oracle. Quand la taille de l'échantillon augmente, ces poids obtenus à partir d'un estimateur consistant deviennent très grands pour les coefficients nuls tandis qu'ils convergent vers une constante finie pour les coefficients non nuls de telle sorte qu'on

puisse asymptotiquement estimer, de manière non biaisée, les coefficients importants et d'exclure les coefficients nuls. Nous précisons également que les résultats asymptotiques sur l'estimateur de la régression quantile pénalisée ou non sont généralement basés sur la notion d'épi-convergence. Cette notion est également très utilisée dans le cadre des moindres carrés pénalisés.

Le cinquième chapitre traite la sélection de variables pour les modèles linéaires généralisés via la vraisemblance non concave pénalisée ainsi que les différentes approches d'approximations locales. L'optimisation se fait en général par des algorithmes itératifs ; cependant l'estimateur obtenu à la première itération peut s'avérer très utile dans certaines situations. Toutes ces techniques d'optimisation peuvent se classer dans la très grande classe des E-M algorithmes ou MM algorithmes. Les approximations locales ont été traitées aussi bien dans le cas de la régression linéaire classique que dans le cas de la régression sur les quantiles. Dans ce chapitre, nous avons considéré le problème de sélection de variables via la vraisemblance pénalisée en utilisant des fonctions de pénalité non convexes. Afin de maximiser la fonction objectif, qui est non différentiable et non concave, un algorithme basé sur une approximation linéaire locale et fournissant un estimateur épars a été récemment proposé. Cependant, il hérite de certains inconvénients du Lasso (Tibshirani[138]) en grande dimension. Afin d'y remédier, nous proposons un algorithme (MLLQA) pour maximiser la vraisemblance pénalisée pour une large classe de fonctions de pénalité non convexes. La propriété de convergence du MLLQA ainsi que la propriété d'oracle de l'estimateur obtenu après une itération ont été établies. Les résultats de convergence du MLLQA sont basés sur les travaux récents de Schifano et al.[128] tandis que les propriétés statistiques de l'estimateur obtenu sont inspirés des travaux de Fan and Li[48] et Zou and Li[161] avec d'autres conditions supplémentaires. Une grande partie de ce chapitre a fait l'objet de la publication Sidi Zakari et al.[132] (`http://intranet.inria.fr/international/arima/016/016002.html`).

L'annexe regroupe les détails des principaux algorithmes utilisés dans la partie pratique de notre travail. Ces détails interviennent soit au niveau du calcul de la solution du problème pénalisé (ou non) ou lors du calcul des chemins de régularisation. Cette annexe englobe également les algorithmes adaptés au cadre de la régres-

sion quantile ainsi que l'algorithme one step LLA (Zou and Li[161]) qui a été l'algorithme de base pour le cinquième chapitre. Dans le cadre de la régression quantile, les algorithmes présentés sont issus des travaux de Rosset and Zhu[125] notamment la version actualisée accessible à l'adresse `http://www.tau.ac.il/~saharon/`. Par soucis de concision nous n'avons pas présenté les détails de l'algorithme pour le problème de l'Elastic net QR. Plus de détails peuvent être trouvés en consultant la référence Slawski[134].

Première partie

Régression Quantile et fonctions de pénalité

Chapitre 1

Généralités

1.1 Régression sur les quantiles : définitions, propriétés et illustrations

L'Introduction de la régression sur les quantiles peut se faire sous plusieurs angles et nécessite beaucoup de notions statistiques. En effet, nous pouvons soit démarrer d'un problème de minimisation, soit à partir de la fonction de distribution (réelle ou empirique) d'une variable aléatoire réelle ou à partir des notions basiques sur les quantiles d'un échantillon aléatoire. De très bonnes propriétés ont été obtenues par plusieurs auteurs[[16, 19, 137]], chacun se basant sur une approche spécifique. D'autre part, la régression sur les quantiles peut être vue comme une alternative aux restrictions liées aux spécifications de modèles. Par exemple dans la majorité des études empiriques, des suppositions parfois très fortes, sont faites sur les distributions des termes d'erreur où l'on se contente d'estimer des effets moyens, ce qui n'est pas, par exemple, adapté au cas des distributions continues. A titre illustratif nous savons par exemple que les revenus ou salaires peuvent varier fortement en fonction du rôle du salarié. On peut ainsi se retrouver à modéliser des revenus moyens alors qu'un tel profil n'est pas majoritaire dans l'entreprise. De même les politiques publiques s'intéressent le plus aux cas extrêmes, par exemple

aux couches les plus vulnérables (pauvreté) ou aux personnes les plus riches dans le cas des impôts et taxes. La moyenne a certes ses avantages mais nous verrons dans la suite dans quel cas on peut utiliser l'une ou l'autre des méthodes de régression. Une mise au point importante est qu'il n'y a pas de terminologie ou notations fixes dans ce cadre. A cet effet, on rencontre souvent la terminologie régression quantile ou régression sur les quantiles. Les estimateurs statistiques obtenus sont souvent appelés quantiles de régression.

Nous précisons également que dans le présent travail, nous ne nous sommes pas intéressés au cas des quantiles extrêmes, localisés au niveau des queues des distributions. Asymptotiquement les quantiles de régression extrêmes convergent faiblement vers les minimas de fonctionnelles d'intégrales stochastiques de processus de poisson dépendant des prédicteurs. Les quantiles de régression intermédiaires et leurs fonctionnelles convergent vers des vecteurs de distribution normale dont les matrices de variance dépendent des paramètres de queues et de la structure des régresseurs. Plus de détails concernant ce point peuvent être trouvés dans Chernozhukov (2005).
Le présent chapitre introductif est structuré comme suit :

1. En premier lieu nous donnerons certaines définitions et propriétés liées à ce cadre ainsi qu'un exemple illustratif issu du domaine micro économique.
2. Nous évoquerons ensuite certains aspects relatifs à l'estimation des paramètres de régression (régression quantile non pénalisée).
3. L'aspect sélection de variables sera également introduit dans le cas classique et aussi dans notre cadre.
4. Les aspects numériques seront mis en exergue à titre de conclusion.

1.1.1 Aspects méthodologiques

Concernant l'approche méthodologique tantôt évoquée, plusieurs définitions ont été proposées chacune en rapport à un domaine particulier (optimisation, approche bayésienne, approche fréquentiste,...).Globalement, ces définitions sont basées sur la fonction de quantile conditionnelle, le modèle de régression quantile

(Bailar[10]), la fonction de contrôle (Koenker and Bassett[88]) ou enfin la densité asymétrique de Laplace (Yu and Moyeed[151]). Le présent travail se focalisera plus sur la deuxième et troisième formulations.

S'agissant de la définition basée sur la fonction de quantile conditionnelle, elle s'énonce comme suit pour $0 < \tau < 1$, $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: soit $q_\tau(\mathbf{x})$ le quantile d'ordre τ de la variable dépendante Y sachant $X = \mathbf{x}$, où $\mathbf{x} \in \mathbb{R}^p$. Nous pouvons avoir $q_\tau(\mathbf{x})$ en résolvant :

$$F_Y(q_\tau(X) \mid X = \mathbf{x}) = P(Y \leq q_\tau(X) \mid X = \mathbf{x}) = \tau,$$

où F_Y est la fonction de répartition de $Y \mid X = x$. D'une manière générale à $F_Y(y)$ correspond une fonction de quantile $Q_Y(\tau)$.
Dans le cas des modèles linéaires, on pose généralement

$$Y = \mathbf{x}^T\beta + \epsilon,$$

où ϵ est supposé vérifier $q_\tau(\epsilon) = 0$ et $\beta \in \mathbb{R}^p$ est un paramètre inconnu.
D'un point de vue optimisation, on considère le problème suivant :

$$min_{\beta \in \mathbb{R}^p} E\{\rho_\tau(Y - X^T\beta) \mid X = \mathbf{x}\},$$

où la fonction $\rho_\tau(.)$ définie sur $\mathbb{R}$ par

$$\rho_\tau(u) = \tau u 1_{[0,+\infty)}(u) - (1-\tau)1_{(-\infty,0)}(u) = (\tau - 1\{u < 0\})u = \frac{\mid u \mid +(2\tau - 1)u}{2}$$

est appelée la fonction de contrôle ou "check function". En considérant la représentation graphique de la fonction de contrôle $\rho_\tau(.)$ (voir figure 1.1), on peut remarquer qu'elle forme un angle $\alpha = tan^{-1}(\tau)$ avec le côté droit de l'axe des abscisses et un angle $\theta = tan^{-1}(1-\tau)$ du côté gauche.

En considérant un échantillon $(\mathbf{x}_i, y_i)_{1 \leq i \leq n}$ i.i.d, l'estimateur $\hat{\beta}(\tau) = argmin_\beta \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\beta)$ vérifie approximativement l'équation d'estimation

$$\sum_{i=1}^n \mathbf{x}_i\{1_{\{y_i - \mathbf{x}_i^T\beta(\tau) < 0\}} - \tau\} = 0$$

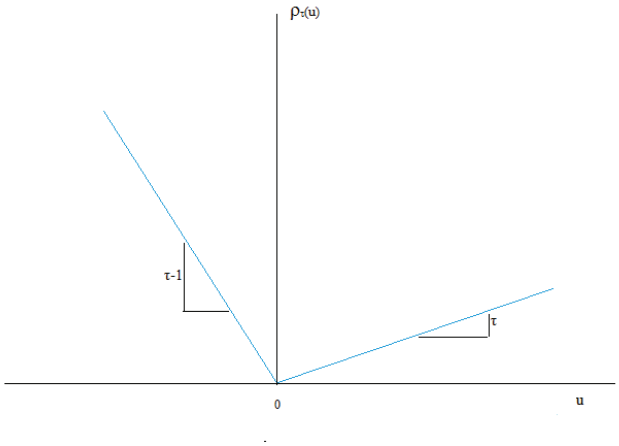

FIGURE 1.1 – Graphe de la fonction de perte(check loss) de la régression quantile ρ_τ.

où le terme $1_{\{y_i - \mathbf{x}_i^T \beta(\tau) < 0\}} - \tau$ peut être également perçu comme le résidu de la régression quantile.
Pour tout $\tau \in (0,1)$, la fonction de perte ρ_τ est convexe et vérifie

$$\rho_\tau(0) = 0, \quad lim_{|u| \to \infty} \rho_\tau(u) = \infty.$$

Elle est également Lipschitz continue de constante de Lipschitz

$$\mid L_\tau \mid = max\{\tau, 1-\tau\}.$$

De plus,

$$min\{\tau, 1-\tau\} \mid u \mid \leq \rho_\tau(u) \leq \mid L_\tau \mid\mid u \mid,$$

pour tout $u \in \mathbb{R}$. Contrairement aux quantiles d'un échantillon qui sont équidistants sur l'intervalle $[0,1]$ avec chaque statistique d'ordre distincte occupant exactement un intervalle de longueur $1/n$, les longueurs des intervalles dans le cadre de la régression sur les quantiles sont irrégulières et dépendent aussi bien de la configuration des prédicteurs que des valeurs observées de la variable réponse. Les paires de points jouent maintenant le rôle de statistiques d'ordre et servent à définir l'estimée de la fonction de quantile conditionnel linéaire de Y sachant $\mathbf{x}$ $Q_{Y|\mathbf{x}}(\tau) \equiv \mathbf{x}^T \beta(\tau)$, dans le cas des modèles linéaires.

Cependant, même dans le cas où $Q_{Y|\mathbf{x}}(\tau)$ est une fonction monotone croissante, cette propriété n'est pas vérifiée pour l'estimation dans le cadre de la régression quantile. En effet, l'estimé $\mathbf{x}^T\hat{\beta}(\tau)$ n'est pas monotone du fait de l'erreur d'estimation ; du moins la monotonie n'est assurée que pour $\mathbf{x} = \bar{\mathbf{x}}$ comme relaté par plusieurs auteurs, notamment Neocleous and Portnoy[109] qui ont étudié la monotonie des fonctions de quantiles de régression. Toutefois, Chernozhukov et al.[30] ont récemment proposé la méthode du bootstrap de quantiles ou réarrangement afin d'assurer la monotonie. La régression sur les quantiles conserve l'aspect important de la robustesse des quantiles ordinaires d'un échantillon. Ainsi, une perturbation des statistiques d'ordres avant ou après la médiane de telle sorte qu'elles restent toujours avant ou après la médiane assure que la position de la médiane reste inchangée. A titre de rappel pour un vecteur aléatoire (X,Y), la distribution conditionnelle de Y sachant X peut être représentée par deux approches équivalentes soit par la fonction de répartition conditionnelle ou la fonction de quantile conditionnelle. Afin d'estimer un quantile conditionnel, on peut utiliser une approche paramétrique, non paramétrique ou semi paramétrique. Nous nous focaliserons uniquement sur le cas de l'approche paramétrique.

Un point remarquable et très important dans les études empiriques est l'analyse du comportement d'une variable dépendante sachant l'information contenue dans un ensemble de variables explicatives. Bien que la moyenne et la médiane sont deux caractéristiques importantes de position, elles ne fournissent pas assez d'information sur les queues de distribution. Chaque quantile de régression caractérise un point particulier de la distribution (centre ou queue). Les aspects relatifs aux moments d'ordre, les mesures de dispersion basées sur les quantiles, les notions de point de rupture d'un estimateur ainsi que les mesures de la qualité d'ajustement du modèle de régression quantile ne seront pas développés dans cette thèse. Il en sera de même pour les notions de fonctions d'influence et de régression quantile tronquée qui s'apparente à la notion bien connue de moyenne tronquée.

1.1.2 Propriétés d'invariance de $\hat{\beta}$

La caractérisation de la ou des solutions du problème de minimisation de la régression sur les quantiles se base sur le théorème 3.1 établi par Koenker et Basset[88]). Ainsi, si X est la matrice des prédicteurs d'un certain rang K, l'ensemble des solutions contient au moins une solution basique de la forme $X^{-1}(h)y(h)$ pour un certain $h \in H$ où $H = \{h \in \{1, 2, ..., n\} | rang(X(h)) = K\}$. L'ensemble H représente l'ensemble des sous ensembles à K éléments de T $= \{1, 2, ..., n\}$. De plus, le même théorème stipule que cet ensemble de solutions représente l'enveloppe convexe de toutes les solutions de cette forme. Une remarque très importante relative à ce théorème est que dans les modèles de translation, les quantiles de l'échantillon sont identifiés soit à partir d'une statistique d'ordre unique issue de l'échantillon observé ou par exemple dans le cas de la médiane issue d'un échantillon de taille paire, nous avons l'identification à partir d'un intervalle fermé compris entre deux statistiques d'ordre adjacentes. Le théorème 3.1 évoqué précédemment illustre une généralisation de cet aspect aux quantiles de régression où les normales aux hyperplans définis par les sous ensembles de K observations jouent le rôle des statistiques d'ordre. Le théorème 3.2 (Koenker and Basset[88]) caractérise les propriétés d'équivariance des quantiles de régression relatifs à certaines transformations comme explicité ci dessous en partant d'une solution $\hat{\beta}(\tau; Y, X)$. Pour tout $a \geq 0$ et $\tau \in [0, 1]$,

$$\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X),$$

$$\hat{\beta}(1 - \tau; -ay, X) = -a\hat{\beta}(\tau; y, X).$$

De plus pour tout $\gamma \in I\!R^p$

$$\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma.$$

Enfin pour toute matrice A de dimension $p \times p$ inversible

$$\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X)$$

De ces propriétés d'équivariance nous constatons par exemple qu'à partir des deux premiers points précédents que $\hat{\beta}(1/2)$ est équivariant par transformation d'échelle de y.
D'autre part, si g est une fonction croissante et continue à gauche sur $\mathbb{R}$, nous avons la propriété dite d'équivariance par transformation monotone :

$$g(Q_\tau(Y|X)) \equiv Q_\tau(g(Y)|X)$$

alors que

$$E(g(Y)|X) \neq g(E(Y|X))$$

sauf dans le cas où la fonction g est affine. Ceci, car $\mathbb{P}(y \leq t) = \mathbb{P}(g(y) \leq g(t))$; ce qui implique que le quantile d'ordre τ de $g(y)$ est l'image par la fonction g du quantile d'ordre τ de y. De manière plus explicite pour une variable aléatoire réelle nous avons $q_\tau(aU + b) = aq_\tau(U) + b$, $q_\tau(a(X)U + b(X) \mid X) = a(X)q_\tau(U) + b(X)$, $q_\tau(max(0, U)) = max(0, q_\tau(U))$, $q_\tau(1_{\{U>0\}}) = 1_{\{q_\tau(U)>0\}}$. Enfin, en général $q_\tau(U_1 + U_2) \neq q_\tau(U_1) + q_\tau(U_2)$ contrairement à l'espérance qui est linéaire.

Example 1.1.1 *Considérons le modèle suivant :*

$$log(y) = \mathbf{x}^\mathbf{T}\beta + \epsilon.$$

Nous pouvons vérifier que si $\mathbf{x}^\mathbf{T}\beta$ est une spécification correcte de $\mathbb{E}(log(y) \mid \mathbf{x})$, $exp(\mathbf{x}^\mathbf{T}\beta)$ n'est pas correctement spécifié pour $\mathbb{E}(y \mid \mathbf{x})$. Ainsi, la propriété d'équivariance par transformation monotone nous assure que, quand $\mathbf{x}^\mathbf{T}\beta_\tau$ est une spécification correcte du quantile conditionnel d'ordre τ de $log(y)$, $\mathbf{x}^\mathbf{T}\beta_\tau = Q_{log(y)|\mathbf{x}}(\tau) = log(Q_{y|\mathbf{x}}(\tau))$, de telle sorte $exp(\mathbf{x}^\mathbf{T}\beta_\tau)$ est le quantile conditionnel de y. L'effet marginal de la variable x_j sur le quantile conditionnel de y est $exp(\mathbf{x}^\mathbf{T}\beta_\tau)\beta_{\tau,j}$.
Nous avons également cette propriété pour les fonctions suivantes : $g(y) = min\{0, y\}$ et $g(y) = sign\{y\}$.

1.1.3 Unicité de la solution $\hat{\beta}$

La condition nécessaire et suffisante de l'unicité de la solution $\hat{\beta} = X^{-1}(h)y(h)$ a été évoquée dans le théorème 3.3 (Koenker and Basset[88]) sous l'hypothèse de continuité de F, où F est la fonction de répartition du terme d'erreur. Cette même condition de continuité de F entraîne le fait que le nombre de composantes nulles du vecteur des résidus est égal à K avec une probabilité 1, de telle sorte que nous avons au moins $n\tau$ observations au dessous du τ ième hyperplan de la régression quantile et au plus $n\tau + K$ observations au dessus. L'unicité de la solution peut être illustrée par le fait que les moindres carrés fournissent toujours une droite de régression unique tandis que les écarts absolus peuvent conduire à plusieurs droites de régression.

D'autre part, pour un couple de données $(y, \mathbf{x})$, la droite de régression via les écarts absolus passera toujours par au moins deux points, ceci à moins qu'il n'y ait des solutions multiples. En présence de solutions multiples, le domaine des solutions valides sera borné par au moins deux droites passant chacune par au moins deux des points représentant les observations. D'une manière générale, si nous avons p prédicteurs ou régresseurs (constante comprise), alors au moins une surface de régression optimale passera par p points observés.

Le fait que la droite de régression passera toujours par au moins deux observations peut aider à comprendre l'instabilité du fait que la droite effectuera des sauts entre différents ensembles de points, ceci en réponse à la perturbation des données. Cette fixation de points, peut également aider à comprendre la propriété de robustesse. Par exemple si nous sommes en présence d'une observation aberrante et que la droite doit passer par deux observations, le point aberrant ne sera probablement pas l'un de ces deux points car la somme des écarts absolus ne sera pas minimisée dans la plupart des cas. Un cas particulier connu dans lequel des solutions multiples existent est obtenu pour un ensemble de points symétriques par rapport à une ligne horizontale comme dans la figure 1.2.

Afin de comprendre pourquoi il y a des solutions multiples dans l'exemple précédent, considérons la droite rose localisée dans le domaine de couleur verte

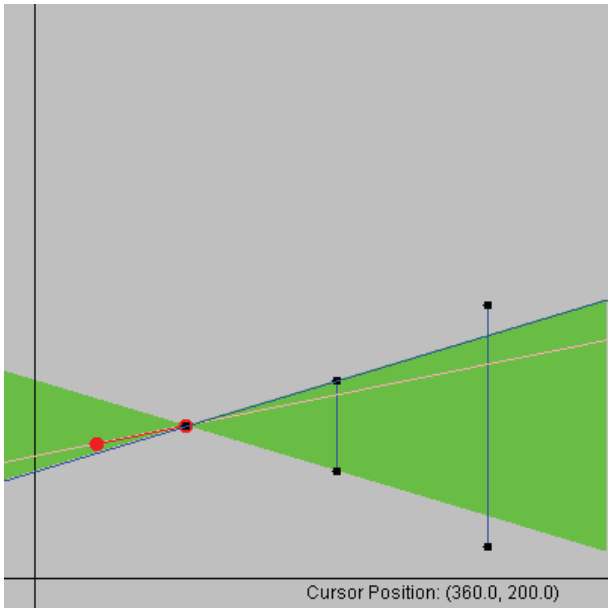

FIGURE 1.2 – Exemple de solutions multiples dans le cas des moindres écarts absolus.

(voir figure 1.2). La somme des erreurs absolues correspondant à cette droite est une certaine valeur S. Cette valeur restera inchangée si l'on devait légèrement déplacer la droite vers le haut ou vers le bas tout en la maintenant dans la région en vert. Cette valeur ne change pas car la distance de chaque point à la droite augmente d'un côté de la droite tandis que la distance de chaque point diminuera de la même quantité sur le côté opposé de la droite. De même, comme la droite peut pivoter suivant des incréments infiniment petits, ceci montre que s'il y a plus d'une solution, il y aura infinité de solutions. Pour les moindres carrés, nous rappelons que l'unique droite de régression passera toujours par le point moyen représenté par le couple $(\bar{y}, \bar{x})$. Ceci pourrait être le cas aussi de la régression quantile en considérant la formulation minimisant

$$(\tau - 1/2)(\bar{y} - \bar{x}\beta) + \frac{1}{2n}\sum_{i=1}^{n} \mid y_i - x_i\beta \mid .$$

Toutefois, dans beaucoup de situations l'unicité de la solution est imposée comme évoqué par Bera et al.[19] avec plus de détails au sujet de l'existence et l'unicité de solutions optimales.
La propriété très connue de robustesse de la médiane par rapport à la moyenne

en présence de valeurs aberrantes dans la variable réponse a été formalisée au cadre de la régression quantile dans la formulation du théorème 3.5 (Koenker and Basset[88]) qui affirme que si $\hat{\beta}(\tau; y, X)$ est solution, alors elle est également solution du problème obtenu en remplaçant y par $X\hat{\beta} + D\hat{u}$, où $\hat{u} = y - X\hat{\beta}$ et D une matrice $p \times p$ diagonale à composantes positives. Une remarque très importante est que la régression quantile jusque là évoquée est une régression par moindres écarts absolus asymétriques où les poids sont attribués en fonction du signe des résidus. Ce même concept a été utilisé par Efron[41] dans l'estimation des percentiles de la régression via les moindres carrés asymétriques. Une excellente comparaison des méthodes robustes ainsi qu'une présentation des différents types de valeurs inhabituelles (points aberrants, points de levier, résidus abérrants) en régression a été récemment présentée par Alma[5].

1.1.4 Distribution asymptotique

La théorie de la distribution asymptotique des quantiles de régression est similaire aux résultats classiques connus dans la littérature. A titre illustratif nous donnons les deux théorèmes ci dessous :

Théorème 1.1.1 *(Theorem 4.1 [Koenker and Basset[88]])*
Soit $\{\hat{q}_n(\tau_1), ..., \hat{q}_n(\tau_M)\}$ *avec* $0 < \tau_1 < \tau_2 < ... < \tau_M < 1$, *une séquence de quantiles uniques issus d'échantillons aléatoires de taille* n *d'une population ayant une fonction de distribution d'inverse définie par* $q(\tau) = F^{-1}(\tau)$. *Si* F *est continue et admet une densité continue et positive,* f, *au point* $q(\tau_i)$, $i = 1, ..., M$, *alors,*

$$\sqrt{n}[\hat{q}_n(\tau_1) - q(\tau_1), ..., \hat{q}_n(\tau_M) - q(\tau_M)]$$

converge en distribution vers un vecteur aléatoire gaussien (M)*-varié de moyenne, 0, et de matrice de covariance* $\Omega(\tau_1, ..., \tau_M; F)$ *d'éléments définis par*

$$\omega_{ij} = \frac{\tau_i(1 - \tau_j)}{f(q(\tau_i))f(q(\tau_j))}, i \leq j.$$

La médiane a une variance asymptotique égale à $[2f(q(1/2))]^{-2}$, valeur qui sera inférieure à la variance de la moyenne pour une large classe de distributions à queues épaisses.
Le théorème précédent a son analogue pour les quantiles de régression et est énoncé sous la formulation ci dessous.

Théorème 1.1.2 *(Theorem 4.2 [Koenker and Basset[88]])*
Soit $\{\hat{\beta}_n(\tau_1), \hat{\beta}_n(\tau_2), ..., \hat{\beta}_n(\tau_M)\}$ *avec* $0 < \tau_1 < \tau_2 < ... < \tau_M < 1$, *une séquence de quantiles de régression uniques spécifiés dans le cadre d'un modèle de régression linéaire. Notons* $q(\tau) = F^{-1}(\tau)$, $\mathbf{q}(\tau) = (q(\tau), 0, ..., 0) \in I\!R^K$ *et* $\hat{\mathbf{q}}_\mathbf{n}(\tau) = \hat{\beta}_n(\tau) - \beta$.
Supposons que :
(i) F *est continue et admet une densité continue et positive,* f, *au point* $q(\tau_i)$, $i = 1, ..., M$, *et*
(ii) $lim_{n\to\infty} n^{-1} X'X = Q$, *une matrice définie positive où* X *la matrice des prédicteurs inclue la colonne de l'intercept. Alors,*

$$\sqrt{n}[\hat{\mathbf{q}}_\mathbf{n}(\tau_\mathbf{1}) - \mathbf{q}(\tau_\mathbf{1}), ..., \hat{\mathbf{q}}_\mathbf{n}(\tau_\mathbf{M}) - \mathbf{q}(\tau_\mathbf{M})]$$

converge en distribution vers un vecteur aléatoire gaussien (MK)*-varié de moyenne 0, et de matrice de covariance* $\Omega(\tau_1, ..., \tau_M; F) \bigotimes Q^{-1}$ *où* Ω *est la matrice de covariance des* M *quantiles ordinaires de l'échantillon obtenus à partir d'échantillons aléatoires d'une distribution* F.

Comme évoqué dans (Koenker and Basset[88]) à titre de remarque pour ce théorème, la médiane de régression ou l'estimateur des moindres écarts absolus $\hat{\beta}(1/2)$ représente un cas particulier très important. Sans perte de généralité, le paramètre à estimer β peut être trouvé de telle sorte que $F(0) = 1/2$, donc $q(1/2) = 0$. Ainsi la distribution asymptotique de la variable aléatoire $\sqrt{n}(\hat{\beta}(1/2) - \beta)$ est gaussienne K-variée de moyenne zero et de matrice de covariance $[2f(0)]^{-2}Q^{-1}$. D'une manière générale nous avons la variance asymptotique de l'estimateur de régression quantile égale à $\frac{\tau(1-\tau)}{f^2(0)}(X^TX)^{-1}$ tandis que nous avons $\sigma^2(X^TX)^{-1}$ pour le cas des moindres carrés. Plus de détails peuvent être consultés dans les références Crépon and Jacquemet[34], D'haultfuille and Givord[35] où la variance asymptotique de l'estimateur de la régression quantile est comparée à celle des MCO. S'agissant

de la vitesse de convergence des processus de régression quantile, plus de détails peuvent être trouvés dans Portnoy[117].

1.1.5 Exemple illustratif de la régression quantile

Considérons maintenant l'illustration de la régression quantile sur un exemple de données réelles. Il s'agit de représenter les dépenses alimentaires en fonction du revenu du ménage de $n = 235$ ménages belges issus de la classe ouvrière au 19^{me} siècle. La base de données nommée Engel est accessible sous le logiciel R via le package "quantreg" fourni par R. Koenker. Plus de détails peuvent être trouvés dans la documentation de ce package. La figure1.3 représente le nuage de points ainsi que différentes droites de régression. Une version log-linéaire est donnée dans la figure 1.4. Une comparaison de ces deux graphiques nous montre le comportement de la fonction de quantiles conditionnels pour les cas hétéroscédastique (figure 1.3) et le cas homoscédastique (figure 1.4) obtenu après une transformation logarithmique. L'hétéroscédasticité peut être justifiée par le fait que nous avons une forte tendance d'augmentation de la dispersion des dépenses alimentaires en fonction de l'augmentation du revenu. La transformation logarithmique nous permet d'avoir un modèle log-linéaire qui peut être assimilé au cas classique où les erreurs sont i.i.d (homoscédasticité). Dans le cas homoscédastique, les droites des quantiles sont parallèles (homogénéité des pentes) ce qui n'est plus le cas si nous sommes en présence d'hétéroscédasticité. D'haultfuille and Givord[35] ont évoqué le fait qu'en général dans le cas où les coefficients correspondant à une régression linéaire et à une régression quantile sont les mêmes, la régression quantile fournira des résultats plus précis lorsque nous sommes en présence de résidus très dispersés. Dans le cas extrême où le terme d'erreur ϵ n'admet pas d'espérance (cas de certaines lois de Pareto), l'estimateur MCO n'est pas convergent, contrairement à l'estimateur de régression quantile qui sera convergent. Enfin, les estimées des fonctions de quantiles conditionnels empiriques (cf. Basset and Koenker[12]) sont également représentées pour les pauvres ($\tau = 0.1$) et les riches ($\tau = 0.9$) dans la figure 1.5.

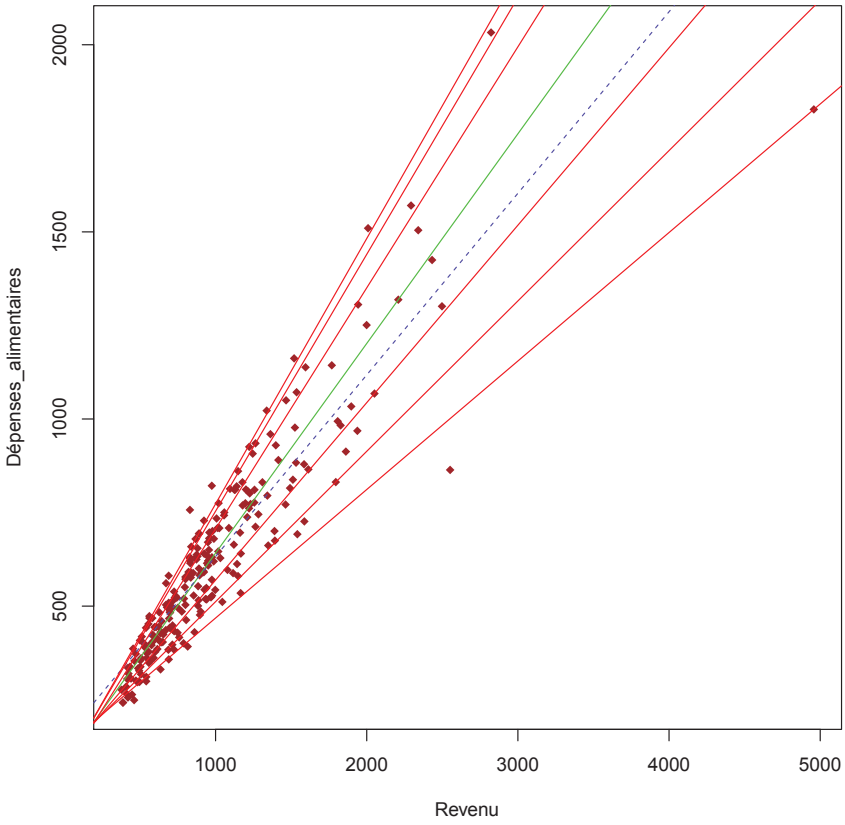

FIGURE 1.3 – Représentation du nuage de points ainsi que différentes droites de régression des données Engel. Les droites en rouge représentent les quantiles de régression pour $\tau \in \{0.05; 0.1; 0.25; 0.75; 0.90; 0.95\}$, la droite en vert représente la régression médiane($\tau = 0.5$) et la droite(traits discontinus) en bleu, la régression via les moindres carrés ordinaires (moyenne conditionnelle).

1.1.6 Scores de rang de régression

Gutenbrunner and Jureckova[61] ont montré que les quantiles de régression qui peuvent être calculés comme solutions d'un problème de programmation linéaire et les solutions du problème dual correspondant (scores de rang de régression) généralisent la notion de dualité entre statistiques d'ordre et rangs d'un modèle de translation au cas du modèle linéaire. Ces auteurs soulignent également le fait que les procédures non paramétriques dans le cas de modèles de translation sont

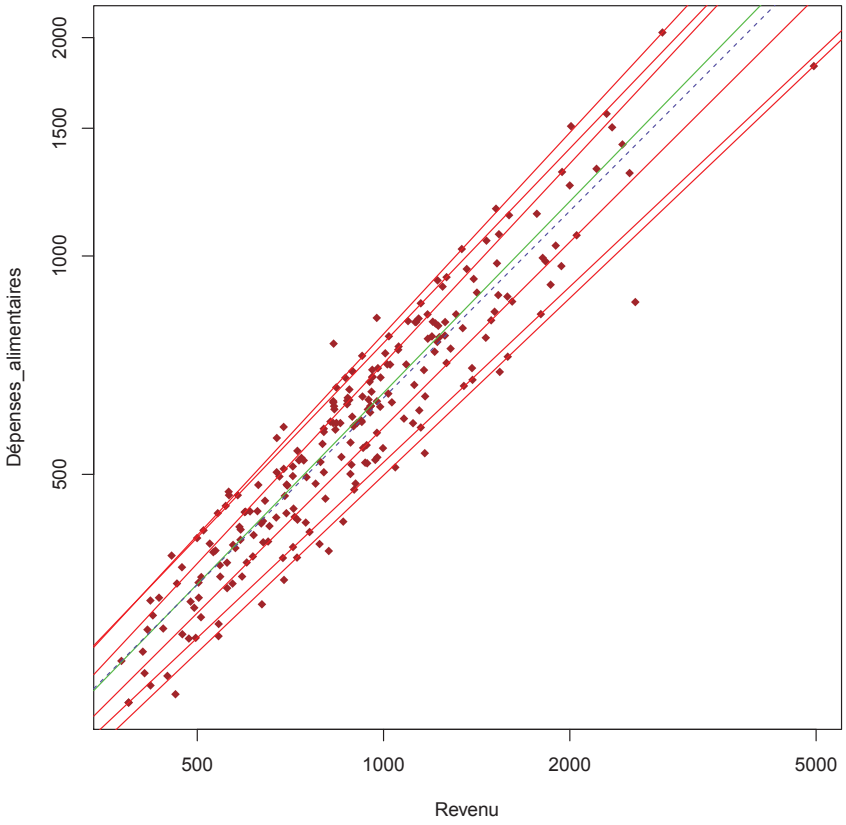

FIGURE 1.4 – Estimation par régression quantile pour une version log-linéaire du modèle d'Engel. Les droites en rouge représentent les quantiles de régression pour $\tau \in \{0.05; 0.1; 0.25; 0.75; 0.90; 0.95\}$, la droite en vert représente la régression médiane($\tau = 0.5$) et la droite(traits discontinus) en bleu, la régression via les moindres carrés ordinaires (moyenne conditionnelle).

généralement basées soit sur les statistiques d'ordre ou quantiles de l'échantillon $y_1, ..., y_n$ ou sur le vecteur $R_1, ..., R_n$ des rangs de l'observation. Nous venons de voir précédemment que les quantiles de régression dans le cadre du modèle de régression linéaire peuvent être perçus comme l'analogue des quantiles de l'échantillon. Cependant l'analogue des rangs était jusque là inconnu. La caractérisation de ces statistiques appelées scores de rang de régression a été effectuée par ces mêmes auteurs.

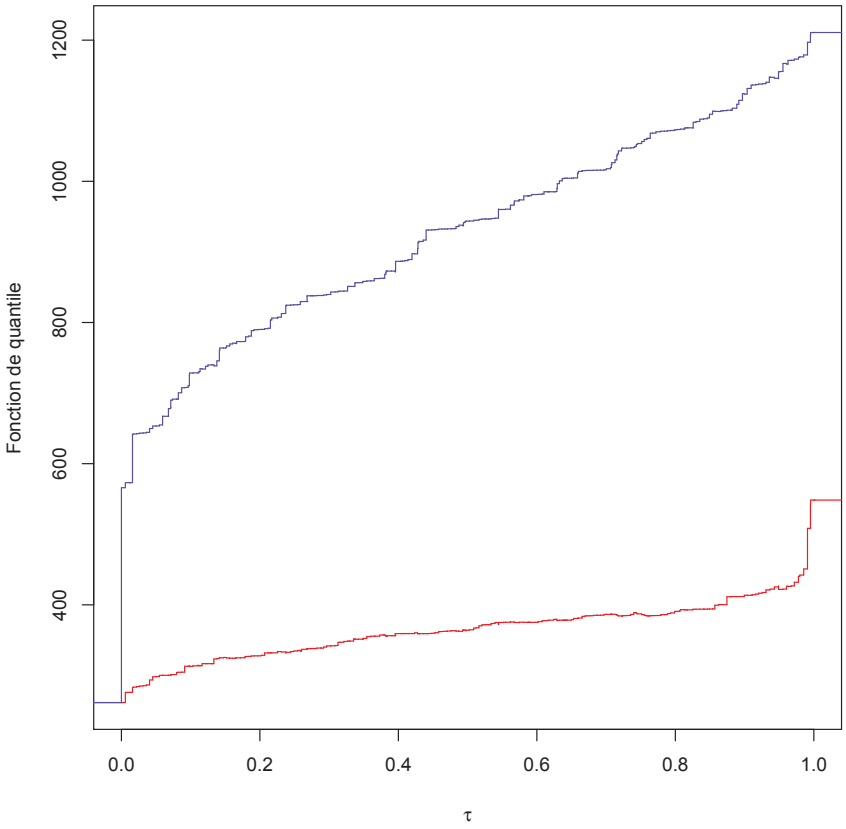

FIGURE 1.5 – Estimation du quantile conditionnel pour les dépenses alimentaires basées sur les données Engel : Deux estimations sont présentées l'une pour les ménages relativement pauvres(en rouge, $\tau = 0.1$) ayant un revenu de $504,5$ francs belges, et l'autre pour les ménages relativement riches(en bleu, $\tau = 0.9$) avec $1538,99$ francs belges.

En considérant par exemple un modèle linéaire hétéroscédastique

$$y = X\beta + \Gamma\epsilon,$$

où $\Gamma = diag(X\gamma)$, $y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ est le vecteur des observations, X est une matrice $(n \times p)$ des prédicteurs, β et $\gamma \in \mathbb{R}^p$ sont les paramètres inconnus et $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ est le vecteur de erreurs qui sont indépendantes et identiquement distribuées de fonction de distribution F inconnue. Des conditions supplémentaires sur X, Γ et F sont nécessaires pour une étude approfondie. A notre niveau nous

nous contenterons uniquement d'introduire le concept. D'autre part, nous rappelons que le τ-quantile de l'échantillon peut être caractérisé comme une solution du problème de minimisation (voir Koenker and Basset[88]) ci dessous :

$$min_{b\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - b), \tag{1.1}$$

tandis que le τ-quantile de régression $\hat{\beta}(\tau)$ peut être caractérisé comme solution de :

$$min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\mathbf{b}), \tag{1.2}$$

avec $\mathbf{x}_i^T$ la $i$ème ligne de X. La contre partie empirique du τ-quantile de régression est, du fait de l'hétéroscédasticité linéaire, le τ-quantile de régression de la population

$$\beta(\tau) = \beta + F^{-1}(\tau)\gamma, \qquad 0 < \tau < 1. \tag{1.3}$$

Le quantile de régression $\hat{\beta}(\tau)$ a été caractérisé (voir Koenker and Basset[88]) comme la composante $\hat{\beta}$ de la solution optimale $(\hat{\beta}, \mathbf{r}^+, \mathbf{r}^-)$ du programme linéaire

$$minimiser \left[\tau \mathbf{1}_n^T \mathbf{r}^+ + (1-\tau)\mathbf{1}_n^T \mathbf{r}^-\right] \tag{P}$$

sous contrainte :

$y = X\hat{\beta} + \mathbf{r}^+ - \mathbf{r}^-$
$(\hat{\beta}, \mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}, \mathbf{1}_n = (1, ..., 1)^T \in \mathbb{R}^n.$

Le programme dual correspondant peut être formulé comme suit :

$$maximiser \left[y^T \hat{\Delta}\right] \tag{D}$$

sous contrainte :

$X^T\hat{\Delta} = 0$
$\hat{\Delta} \in [\tau - 1, \tau]^n,$

ou de manière équivalente

$$maximiser \left[y^T \hat{\Delta} \hat{a} \right] \qquad (\tilde{D})$$

sous contrainte :

$X^T \hat{a} = (1-\tau) X^T \mathbf{1}_n$
$\hat{a} \in [0,1]^n$.

La solution optimale du problème dual $(\tilde{D})$ est $\hat{a}(\tau) = (\hat{a}_1(\tau), ..., \hat{a}_n(\tau))^T \in \mathbb{R}^n$. Les solutions de ce type sont appelées les scores de rang de régression.
Nous précisons aussi le fait que si $\{y_i : i \in M_{n\tau}\}$ sont les observations ajustées exactement par $\hat{\beta}(\tau)$, alors les variables duales $\hat{\mathbf{a}}(\tau)$ peuvent être caractérisées par les inégalités

$$\hat{a}_i(\tau) \in \begin{cases} 1 & \text{si } y_i > \mathbf{x}_i^T \hat{\beta}(\tau) \\ (0,1) & \text{si } y_i = \mathbf{x}_i^T \hat{\beta}(\tau) \\ 0 & \text{si } y_i < \mathbf{x}_i^T \hat{\beta}(\tau) \end{cases}$$

$i = 1, ..., n$ et aussi par les équations linéaires

$$\sum_{i \in M_{n\tau}} \hat{a}_i(\tau) \mathbf{x}_i = (1-\tau) \sum_{i=1}^{n} \mathbf{x}_i - \sum_{i=1}^{n} 1_{y_i > \mathbf{x}_i^T \hat{\beta}(\tau)} \mathbf{x}_i. \qquad (1.4)$$

Sous l'hypothèse de continuité de F, avec une probabilité 1, les p équations linéaires précédentes ont une unique solution pour tout $\tau \in (0,1)$, ce qui est en correspondance avec l'unicité des rangs dans le modèle de translation.
En résumé, le terme $x^T \hat{\beta}(\tau)$ estime $Q_Y(\tau \mid x)$, est constant par morceaux sur l'intervalle $[0,1]$ et enfin pour $X = 1_n$, $\hat{\beta}(\tau) = \hat{F}^{-1}(\tau)$. D'autre part, les éléments $\{\hat{a}_i(\tau)\}_{i=1}^n$ sont les fonctions de scores de rang de régression, elles sont linéaires par morceaux sur $[0,1]$ et pour $X = 1_n$ ils constituent les fonctions génératrices de rang de Hájek[63].

1.1.7 Régression quantile composite

Considérons le modèle linéaire suivant :

$$y = X\beta + \epsilon \tag{1.5}$$

où $X = (x_1, ..., x_n)^T$, $y = (y_1, ..., y_n)^T$ et ϵ est le vecteur des erreurs de dimension n.
Tant que la distribution du terme d'erreur est homoscédastique, $\mathbf{x}_i^T\beta$, est moyennant une constante additive, le τ quantile conditionnel de y_i sachant $\mathbf{x}_i$ (voir Fan and Lv[51]). A cet effet, β peut être estimé en considérant la régression quantile basée sur

$$\sum_{i=1}^{n} \rho_\tau(y_i - b_\tau - x_i^T\beta). \tag{1.6}$$

Koenker (1984) proposa de résoudre le problème de régression quantile composite pondérée en utilisant différents quantiles afin d'améliorer l'efficacité en minimisant par rapport à $b_1, ..., b_K$ et β

$$\sum_{k=1}^{K} w_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - b_k - x_i^T\beta), \tag{1.7}$$

où $\{\tau_k\}$ est une séquence donnée de quantiles et $\{w_k\}$ est une séquence donnée de poids. Zou and Yuan[162] ont ainsi proposé l'approche par la régression quantile composite pénalisée avec des poids égaux afin d'améliorer l'efficacité des moindres carrés pénalisés. L'intérêt de la régression quantile composite se trouve principalement dans la situation où la variance du terme d'erreur est infinie. Dans ce cas, l'estimateur obtenu possède les propriétés d'oracle, ce qui n'est pas le cas de l'estimateur des moindres carrés (Zou and Yuan[162]).

1.2 Estimation des paramètres.

1.2.1 Conditions d'optimalité.

Soit l'échantillon $(y_i, \mathbf{x}_i^T)^T$ pour $i = 1, ..., n$ et pour chaque indice i, $\mathbf{x}_i$ est un vecteur de dimension p. Considérons le modèle linéaire suivant :

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i. \tag{1.8}$$

L'estimateur du quantile de régression de β peut être obtenu en minimisant la fonction suivante (voir Koenker[86]) :

$$V_n(\beta;\tau) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \mathbf{x}_i^T\beta\right) = \frac{1}{n}\left[\tau \sum_{i:y_i \geq \mathbf{x}_i^T\beta} \mid y_i - \mathbf{x}_i^T\beta \mid + (1-\tau) \sum_{i:y_i < \mathbf{x}_i^T\beta} \mid y_i - \mathbf{x}_i^T\beta \mid\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\tau - 1\{y_i - \mathbf{x}_i^T\beta < 0\}\right)\left(y_i - \mathbf{x}_i^T\beta\right)$$

Du fait de la non-différentiabilité en $y_i = \mathbf{x}_i^T\beta$, la solution ne peut être calculée par les méthodes numériques classiques. A cet effet, nous considérons les dérivées directionnelles de $V_n(\beta;\tau)$ dans la direction w par :

$$\frac{d}{d\delta} V_n(\beta + \delta w;\tau) \mid_{\delta=0}$$

$$= \frac{1}{n}\left(\frac{d}{d\delta}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta - \delta\mathbf{x}_i^T w\})(\tau - 1\{y_i - \mathbf{x}_i^T\beta - \delta\mathbf{x}_i^T w < 0\})\right) \mid_{\delta=0}$$

$$= \frac{-1}{n}\sum_{i=1}^{n}\psi_\tau^*\left(y_i - \mathbf{x}_i^T\beta, -\mathbf{x}_i^T w\right)\mathbf{x}_i^T w,$$

où,

$$\psi_\tau^*(a,b) = \begin{cases} \tau - 1_{\{a<0\}} & \text{if } a \neq 0 \\ \tau - 1_{\{b<0\}} & \text{if } a = 0. \end{cases}$$

Un point est un minimum de $V_n(\beta;\tau)$ si les dérivées directionnelles en ce point sont toutes positives dans toutes les directions w. L'estimateur du quantile de régres-

sion d'ordre τ de β réalise le minimum et est noté $\hat{\beta}_\tau$. Le terme $\mathbf{x}_i^T\hat{\beta}_\tau$ représente l'hyperplan estimé de la régression quantile et les résidus de la régression quantile sont $\hat{e}_i(\tau) = y_i - x_i^T\hat{\beta}_\tau$.
Considérons le cas où $n = p$, le nombre de paramètres. Le point $\mathbf{b}$ pour lequel $y_i - x_i^T\mathbf{b}, (i = 1, ...p)$ doit être un minimum de V_p car la dérivée directionnelle en $\mathbf{b}$ de la forme

$$\frac{-1}{p}\sum_{i=1}^{p}(\tau - 1\{-\mathbf{x}_i^T w < 0\})\mathbf{x}_i^T w \geq 0,$$

pour tout $w \in \mathbb{R}^p$.
Cette solution conduit à un ajustement parfait des premières p observations et est connue comme une solution basique du problème de minimisation de $V_n(\beta;\tau)$.

Nous considérons κ un sous ensemble de k nombres de $\{1, 2, ..., n\}$, K l'ensemble de tous les κ, $X(\kappa)$ la matrice de dimension $p \times p$ de lignes $\mathbf{x}_i$, $i \in \kappa$, et $y(\kappa)$ de dimension $p \times 1$ dont les éléments sont y_i, $i \in \kappa$. D'autres solutions basiques sont $\mathbf{b}(\kappa) = X^{-1}(\kappa)y(\kappa)$ pour $\kappa \in K$ où chaque solution conduit à un ajustement parfait de p observations. Le point $\mathbf{b}$ mentionné précédemment correspond à une solution pour $\kappa = \{1, ..., k\}$. En d'autres termes l'hyperplan estimé de la régression sur les quantiles doit interpoler p observations dans l'échantillon. L'estimateur $\hat{\beta}_\tau$ peut être obtenu en cherchant parmi les solutions de base. A titre de complément d'information, Koenker and Portnoy[96] ont évoqué le fait que pour un modèle à K régresseurs, il y a exactement K résidus nuls en cas de non dégénérescence, sinon éventuellement plus. Ce fait est susceptible de se produire pour des valeurs discrètes de y. En cas de non dégénerescence le nombre de résidus négatifs N^- vérifie $N^- \leq n\tau \leq N^- + K$ et le nombre de résidus positifs N^+ vérifie $N^+ \leq n(1-\tau) \leq N^+ + K$ avec inégalités strictes en cas d'unicité de la solution.
Enfin, s'intéressant à la fonction de risque dans le cas des erreurs $\epsilon \sim N(0, \sigma^2)$, nous avons déjà évoqué le fait que la fonction théorique du τ^{eme} quantile conditionnel est donnée par $m_\tau(\mathbf{x}) = \sigma\Phi^{-1}(\tau) + \beta_0 + \mathbf{x}\beta$ et que la fonction estimée du τ^{eme} quantile conditionnel est $f(\mathbf{x}) = \hat{\beta}_0 + \mathbf{x}\hat{\beta}$. En utilisant la fonction de perte de la régression quantile, on peut calculer le risque de f, défini par (Yao and Lee[150])

$$R(f;\beta_0,\beta) := E\{\tau(Y - \hat{\beta}_0 - X\hat{\beta})_+ + (1-\tau)(Y - \hat{\beta}_0 - X\hat{\beta})_-\}$$

$$= \left\{ \tau - \Phi \left(\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 + \parallel \beta - \hat{\beta} \parallel_2^2}} \right) (\beta_0 - \hat{\beta}_0) \right\}$$

$$+ \sqrt{\frac{\sigma^2 + \parallel \beta - \hat{\beta} \parallel_2^2}{2\pi}} exp \left\{ - \frac{(\hat{\beta}_0 - \beta_0)^2}{2(\sigma^2 + \parallel \beta - \hat{\beta} \parallel_2^2)} \right\}.$$

Pour chaque valeur τ, le vrai risque de $m_\tau(\mathbf{x})$ est $(\sigma/\sqrt{2\pi})exp\{-\Phi^{-1}(\tau)^2\}$, ce qui représente le risque minimal atteignable. Le maximum des risques minimaux est obtenu quand $\tau = 0.5$ et dans ce cas la vraie fonction de médiane conditionnelle est $m_{0.5}(\mathbf{x}) = \beta_0 + \mathbf{x}\beta$ avec le risque de $\sigma/\sqrt{2\pi} \approx 2.821$.

1.3 Sélection de variables : régression pénalisée.

1.3.1 Méthodes classiques.

D'une manière générale, l'intérêt de la sélection de variables peut être considéré d'un point de vue explicatif et/ou prédictif. Intuitivement on est généralement amené à choisir des modèles faisant intervenir un nombre limité de variables. Dans notre contexte, s'agissant du problème d'estimation, nous supposerons toujours que le vrai modèle (inconnu) est parcimonieux, c'est à dire ne contient qu'un nombre limité de variables parmi un ensemble de variables initiales. En un mot, nous effectuons une réduction de dimension. Les variables les plus pertinentes relativement à un critère spécifique seront considérées dans le modèle final.

Plusieurs approches ont été fournies dans ce cadre là, soit pour palier à l'aspect algorithmique, au temps de calcul, à la gestion de la dimension du modèle, soit carrément dans certaines situations, au fondement méthodologique lui même. La réduction de dimension peut être perçue sous deux aspects, l'un classique où la taille de l'échantillon dépasse le nombre de variables, l'autre relativement récent où l'avancée technologique ainsi que l'innovation méthodologique permet de collecter de gros volumes de données en temps réel. Toutefois, malgré la disponibilité de ces données, l'exploitation de ces gigantesques *cimetières* d'information pose problème,

ce qui a valu l'utilisation des termes *fléau de la dimension* ou *malédiction de la dimension*. Nous évoquerons quelques approches proposées dans ce cadre.

A titre de rappel, Fan and Lv[51] ont évoqué le fait que toutes les procédures statistiques reposent sur trois piliers fondamentaux qui sont la précision statistique, l'interprétabilité du modèle et la complexité des calculs. Dans le cas classique où la taille n de l'échantillon dépasse le nombre de variables (paramètres) p, aucun de ces trois piliers ne doit être sacrifié au profit des autres. De nouveaux défis se posent alors quand le nombre de paramètres est comparable ou dépasse la taille de l'échantillon. Il se pose alors les problèmes de conception de procédures statistiques qui sont plus efficaces en termes d'inférence, de théorie asymptotique ou non asymptotique, d'estimation et d'interprétabilité de modèles, d'efficacité et de robustesse (en termes de calculs) des procédures statistiques. A cela s'ajoute la difficulté de sélection due à la multicolinéarité entre les variables. La colinéarité peut facilement être faussée en géométrie dimensionnelle élevée (Fan and Lv[49]). Une telle situation peut nous faire choisir un modèle inadapté et l'intuition classique peut s'avérer fausse. Quand p est très grand comparé à n il peut y avoir de fortes (fausses) corrélations même pour des prédicteurs indépendants et identiquement distribués [Fan and Lv[49], Fan, Guo and Hao[47]].

1.3.2 Régression pénalisée.

Dans le contexte actuel, les méthodes classiques de sélection de variables ou de modèles pas à pas ne sont pas utilisables ou non recommandées. A la limite elles peuvent intervenir dans une étape intermédiaire. Le processus de sélection repose sur l'utilisation d'un critère pénalisé où la fonction de pénalité est généralement convexe. La procédure la plus connue dans ce cas est la pénalité du Least absolute solution and shrinkage operator(Lasso, Tibshirani[138]) qui a fait l'objet de plusieurs travaux. En effet, en considérant le modèle (1.8), l'estimation du paramètre β requiert dans certains cas l'utilisation d'un paramètre de régularisation. Le cas de la régression ridge qui donne un estimateur légèrement biaisé mais avec une variance contrôlée comparé à celui des moindres carrés ordinaires. Le Lasso qui

utilise une pénalité de type

$$\| \beta \|_1 = \sum_{j=1}^{p} | \beta_j |$$

induit également un biais mais à l'avantage d'une faible variance. Dans le cas des moindres carrés pénalisés le problème suivant est considéré :

$$\hat{\beta}^{Lasso} = argmin_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right\} \tag{1.9}$$

où $\lambda > 0$ est le paramètre de régularisation qui permet de rétrécir les grandes composantes (coefficients) de β, et aussi d'annuler les plus petites en prenant une forte valeur de λ. Les problèmes d'optimisation utilisant cette pénalité ou une de ses variantes restent convexes. La résolution est ainsi facilitée même dans les situations où la solution ne peut être obtenue de manière explicite comme c'est le cas de la régression quantile pénalisée. Le cas non convexe qui a l'avantage de réduire le biais d'estimation nécessite généralement une transformation intermédiaire(approximation) afin de se ramener au cas convexe lors de l'estimation. On évite ainsi le problème de minimum local.Ce qui n'est pas le cas des problèmes convexes où le minimum local, quand il existe, est global. Une bonne fonction de pénalité doit permettre d'avoir un estimateur qui est non biaisé, parcimonieux et continu par rapport aux données afin de réduire l'instabilité dans la prédiction du modèle. Pour plus de détails par rapport à ces trois aspects, (Fan and Li[48]) et (Antoniadis and Fan[7]) sont d'excellentes références. Les fonctions de pénalité les plus connues sont les pénalités de type norme $L_q, q \geq 0$ dans la régression bridge (Frank and Friedman[56]) ou une combinaison de ces pénalités. La singularité à l'origine de la fonction de pénalité est une condition nécessaire de parcimonie. A titre d'exemple, la pénalité concave L_q avec $0 \leq q < 1$ ne conduit pas à un estimateur continu. Le cas convexe où $q > 1$ ne satisfait pas la condition de parcimonie et enfin le cas convexe $q = 1$ ou pénalité de type L_1 appelée également pénalité de type Lasso (Tibshirani[138]) à l'inconvénient de fournir des estimations biaisées du fait que c'est le même paramètre de régularisation qui est utilisé pour tous les coefficients de régression (Fan and Li[48], Zou[159]). En un mot, aucune des pénalités de type L_q ne satisfait simultanément ces trois conditions. Une des pénalités les satisfaisant simultanément est la pénalité Smoothly Clipped Absolute Deviation (SCAD, Fan[46], Fan and Li[48]) dans le cadre de la vraisemblance pénalisée au

même titre que le Lasso adaptatif (Zou[159]). La pénalité du SCAD est définie par :

$$P_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda \mid \beta_j \mid & \text{si } \mid \beta_j \mid \leq \lambda \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{si } \lambda < \mid \beta_j \mid \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{si } \mid \beta_j \mid > a\lambda. \end{cases}$$

Elle est continument différentiable sur $(-\infty, 0) \cup (0, +\infty)$ et sa dérivée première pour $a > 2$ et $\beta_j > 0$ à l'expression suivante

$$P_\lambda'(\beta_j) = \lambda \left\{ \mathbf{1}(\beta_j \leq \lambda) + \frac{(a\lambda - \beta_j)_+}{(a-1)\lambda} \mathbf{1}(\beta_j > \lambda) \right\}. \tag{1.10}$$

Cette pénalité comporte une singularité en zero et ses dérivées sont nulles sur $[-a\lambda, a\lambda]$. De plus, elle combine les avantages de la méthode de sélection du meilleur sous ensemble (Breiman[21]) ainsi que la régression ridge ($q = 2$). En effet, elle donne un estimateur parcimonieux, assure la stabilité de la sélection de modèle et fournit des estimations non biaisées pour les grands coefficients.

Du fait de la non différentiabilité en zero de la fonction de perte de la régression quantile, les propriétés d'oracle générales proposées par Fan and Li[48] dans le cadre de la vraisemblance pénalisée nonconcave ne s'appliquent pas directement. Afin de palier à la difficulté causée par cette singularité à l'origine, Wu and Liu[148] ont fait appel au lemme de convexité précédemment utilisé par Pollard[116].

La pénalité du Lasso adaptatif utilise une pénalité de type

$$\parallel \tilde{w}^T \beta \parallel_1 = \sum_{j=1}^{p} \mid \tilde{w}_j \beta_j \mid,$$

où $\tilde{w}$ représente le vecteur des poids généralement obtenu à partir d'un estimateur consistant de β.

1.3.3 Régression quantile pénalisée

Considérons le problème général de régression quantile pénalisée suivant

$$min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta_\tau) + \lambda P(\beta_\tau), \tag{1.11}$$

où $\lambda \geq 0$ est la paramètre de régularisation et $P(.)$ est le terme de pénalité.

Dans ce cadre là, le travail précurseur a été entamé par Wu and Liu[148] dans le cas du Lasso adaptatif et la pénalité de type SCAD, en montrant notamment leurs propriétés d'oracle. Le cas de l'Elastic net structuré a été proposé par Slawski[134] mais sans développement théorique. L'aspect théorique dans le cas du Lasso a été étudié en détails par Belloni and Chernozhukov[16] ainsi que Kato[87] dans le cas du Lasso adapté à la sélection de groupes de variables (Grouped Lasso). Wang, Li and Jiang[142] ont étudié le cas particulier des moindres écarts absolus avec la pénalité du Lasso adaptatif.

Afin de palier aux problèmes de sélection de variables en grande dimension ou pour certaines considérations asymptotiques, d'autres types de pénalités ont été proposées. S'agissant du présent travail, les plus utilisées seront l'Elastic net (Zou and Hastie[160])qui est une combinaison convexe des nomes L_1 et L_2, le Lasso adaptatif utilisant des pondérations (Zou[159]) et l'Elastic net adaptatif (Zou and Zhang[165]). Ces pénalités ont fait l'objet de beaucoup de développements dans le cadre des moindres carrés et ensuite des modèles linéaires généralisés.

1.3.4 Choix des paramètres de pénalité.

Le choix des paramètres de régularisation est d'une importance capitale dans le cadre de la régression pénalisée. Quand le paramètre de régularisation λ vaut 0, toutes les variables sont sélectionnées, de plus quand $p > n$ le modèle n'est pas identifiable. Dans le cas où $\lambda = \infty$, si la pénalité satisfait $lim_{\lambda \to \infty} P_\lambda(\mid \beta \mid) = \infty$ pour $\beta \neq 0$, aucune variable n'est alors sélectionnée. Les cas les plus intéressants

sont alors entre ces deux choix extrêmes. Ainsi, la complexité du modèle sélectionné est gouvernée par le paramètre λ. Une grande valeur de ce paramètre tend à sélectionner un modèle simple avec une variance d'estimation plus petite tandis qu'une valeur faible conduit à un modèle complexe ayant un biais plus petit. Dans la majorité des cas, le chemin des solutions du vecteur β comme fonction de λ est linéaire par morceaux. De ce fait, on peut obtenir une suite croissante $\lambda_0 = 0 < \lambda_1 < ... < \lambda_p$ de telle sorte que $\hat{\beta}_{\tau,\lambda_0}$ corresponde à l'estimateur de la régression quantile non pénalisée, $\hat{\beta}_{\tau,\lambda_1}$ ait exactement un seul coefficient nul, $\hat{\beta}_{\tau,\lambda_2}$ deux coefficients nuls, ainsi de suite jusqu'à $\hat{\beta}_{\tau,\lambda_p}$ qui correspond au vecteur nul.
Enfin, nous soulignons le fait que le compromis entre le biais et la variance est généralement mis en avant lors du choix du λ optimal (exemple de la validation croisée). La figure 1.6 synthétise les notions abordées dans cette partie ; l'illustration a été effectuée sur la base de données du merlan.

1.3.5 Choix de l'estimateur initial pour le Lasso adaptatif.

Considérons le problème de régression quantile avec la pénalité du Lasso adaptatif qui peut être perçue comme une généralisation de la pénalité du Lasso. Contrairement au Lasso qui pénalise les coefficients de manière uniforme, on utilise des poids adaptatifs afin de pénaliser les coefficients des différentes variables à des degrés différents. D'où se pose la question du choix des poids adaptatifs. Ainsi dans le cadre des moindres carrés, Zou[159] propose d'utiliser comme poids, l'inverse des estimations par moindres carrés élevées à une certaine puissance. Wu and Liu[148] ont considéré l'estimateur (non pénalisé) de la régression quantile en considérant

$$\tilde{\beta}_\tau = argmin_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta_\tau) \tag{1.12}$$

qui est un estimateur consistant de β_τ.
Ainsi l'estimateur Lasso adaptatif dans le cadre de la régression quantile est obtenu en minimisant, par rapport à β_τ, le critère pénalisé

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta_\tau) + \lambda \sum_{j=1}^{p} \tilde{w}_j \mid \beta_{\tau,j} \mid, \tag{1.13}$$

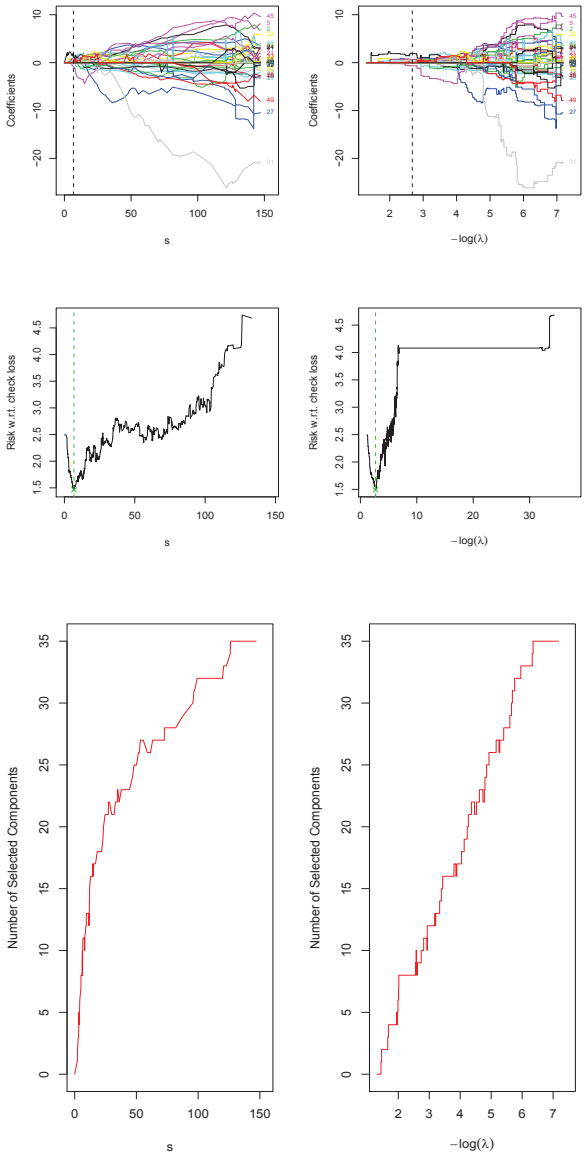

FIGURE 1.6 – Régression médiane L_1 utilisant les paramètres de pénalité λ et son analogue (en programmation linéaire) s pour l'indice de fraîcheur du merlan. Les lignes verticales brisées (noires et vertes) représentent respectivement les paramètres de régularisation optimaux obtenus par validation croisée (5-fold CV) ainsi que la valeur minimale du risque estimé. Une dizaine de variables ont été sélectionnées dans cet exemple.

où $\tilde{w}_j = \mid \tilde{\beta}_{\tau,j} \mid^{-\gamma}$, $j = 1, ..., p$ pour un certain choix judicieux de $\gamma > 0$. Le cas particulier où $\tilde{w}_j = 1$ pour $j = 1, ..., p$ correspond au cas de la régression quantile pénalisée de type Lasso.
Dans le cas où les erreurs sont i.i.d, nous pouvons considérer le critère précédent en β au lieu de β_τ. Pour des considérations asymptotiques, Wu and Liu[148] ont considéré la formulation suivante

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + n\lambda_n \sum_{j=1}^{p} \tilde{w}_j \mid \beta_j \mid . \tag{1.14}$$

Nous rappelons que le Lasso adaptatif sous sa forme précédente nécessite un estimateur initial consistant afin d'avoir les poids adaptatifs. D'une manière générale, dans le cas classique où la taille de l'échantillon n dépasse le nombre de paramètres p, c'est à dire $n > p$, ce problème ne se pose pas. L'estimateur de la régression quantile non pénalisée en (1.12) reste consistant même quand p augmente avec n à une vitesse $O(n^\alpha), 0 < \alpha < 1$ (He and Shao[67]). Quand $p > n$, se pose alors le problème d'obtention d'un estimateur initial consistant. Dans ce cas, on utilise généralement l'estimateur pénalisé (en norme L_2) comme estimateur initial pour générer les poids.

1.3.6 Sélection de variables et dimensionnalité.

Ce champ n'a pas encore été développé dans le cadre de la régression sur les quantiles, à cet effet nous ferons une synthèse du cadre existant.
Comme évoqué par Fan and Lv[51], les méthodes classiques de sélection du meilleur sous ensemble de variables sont trop coûteuses en termes de coût de calcul pour les applications actuelles qui couvrent presque tous les domaines d'activités. Malgré le fait que beaucoup de méthodes soient adaptées à ce cadre, nous passerons également en revue leurs limites en termes de dimension. Non seulement le nombre de variables est très grand, à cela s'ajoute la prise en compte dans certaines situations des interactions entre ces variables. Les termes grande dimensionnalité et ultra grande dimensionnalité se réfèrent respectivement au cas où la dimensionnalité croît à un taux non-polynomial quand la taille de l'échantillon augmente.

Alors que le terme grande dimensionnalité se rapporte au cas classique d'augmentation de dimensionnalité. Une définition plus détaillée a été donnée par Fan, Lv and Qi[53] : la grande dimension relative se réfère au cadre asymptotique où la dimension p est croissante, mais est d'un ordre plus petit que la taille de l'échantillon n ($p = o(n)$). La grande dimension modérée se réfère au cadre asymptotique dans lequel p croît proportionnellement à n ($p \sim cn, c > 0$). La grande dimension se réfère au cadre asymptotique dans lequel p peut croître polynomialement avec n ($p = O(n^{\alpha}), \alpha > 1$) et l'ultra grande dimensionnalité se réfère au cadre asymptotique où p peut augmenter de manière non polynomiale avec n ($logp = O(n^{\alpha}), a > 0$). Fan, Lv and Qi[53] affirment que l'inférence et la prédiction sont basées sur l'espace des paramètres en grande dimension.

Ce qui rend l'inférence statistique possible en grande dimension est l'hypothèse que la fonction de régression appartient à une variété de dimension réduite. Dans de tels cas, les paramètres de régression p-dimensionnels sont supposés être clairsemés avec de nombreuses composantes nulles ou négligeables. Les composantes non nulles indiquent les variables importantes. Comme évoqué par Fan, Lv and Qi[53], cette supposition est nécessaire pour l'identifiabilité du modèle notamment pour les échantillons de taille relativement faible. Cet aspect peut être également utilisé afin d'améliorer les temps de calcul comme c'est le cas pour les matrices "creuses".

Dans le cas des données génétiques issues des micro puces, le nombre d'observations est généralement de l'ordre des dizaines tandis que le nombre de profils d'expression génétique sont de l'ordre des dizaines de milliers. Dans le cas des interactions protéine-protéine, le nombre de caractères peut être de l'ordre des millions tandis que la taille de l'échantillon peut être de l'ordre des milliers. Ainsi les méthodes de sélection de variables peuvent être très coûteuses en termes de coût de calcul. A cet effet, il est tout à fait naturel de réduire la dimension p d'une très grande échelle ($logp = O(n^{a}), a > 0$) à une autre relativement grande ($d = O(n^{b}), b > 0$) via une méthode rapide, efficace et sûre. Ensuite, les méthodes de sélection de variables peuvent être utilisées sur l'espace paramètre réduit. Cette technique est appelée "screening" dans la littérature.

Cependant, l'approche par screening résulte dans la plupart des cas à l'intro-

duction de variables indésirables (false positive variables). Afin de réduire le taux de false positive, Fan, Samworth and Wu[55] ont proposé la technique de rééchantillonnage. Plus de détails sur cette technique peuvent être consultés dans Fan and Lv[51].

1.4 Aspects numériques.

1.4.1 Régression sur les quantiles non pénalisée

Concernant le domaine de la programmation linéaire dans notre cadre, Barrodale and Roberts[11] ont utilisé un algorithme basé sur le simplex pour l'estimation des moindres écarts absolus. Cette approche a été généralisée par Koenker et d'Orey[90] dans le cadre de l'estimation des quantiles de régression. Koenker et Park[95] ont aussi étendu la méthode du point intérieur de Karmakar[77] au cadre de la régression quantile. Portnoy et Koenker[118] ont combiné la méthode du point intérieur avec une approche de préprocessing pour faire l'estimation dans le cadre de la régression médiane. Hunter et Lange[73] ont introduit l'algorithme de maximisation-minimisation ; Chernozhukov et Hong[31] ont proposé l'estimateur généralisé de Laplace.

A titre de remarque sur le travail de Portnoy and Koenker[118], les auteurs soulignent le fait qu'en terme de temps de calcul les méthodes minimisant la norme quadratique des erreurs sont plus efficaces que celles utilisant la norme L_1, notamment dans le cas des grandes bases de données. La méthode du simplex qui est une référence dans le domaine de la programmation linéaire a l'inconvénient d'avoir un temps de calcul non négligeable. Ce qui n'est pas le cas des problèmes modérés où cet algorithme est très performant. Par exemple dans la version S-PLUS sur laquelle se base beaucoup de codes R, les méthodes L_1 utilisant le simplex sont plus rapides que les méthodes L_2 pour des observations de l'ordre de quelques centaines. Les méthodes du point intérieur sont également utilisées pour la résolution des problèmes de programmation linéaire. La mise en oeuvre des méthodes d'estimation et d'inférence en régression quantile non pénalisée est essentiellement

basée sur le travail de Chen et Wei[25]. S'agissant des méthodes d'estimation elles sont principalement basées sur trois procédures. La méthode du simplex utilisant l'approche de Barrodale et Roberts[11], celle du point intérieur (Karmakar[77]) utilisant l'algorithme primal-dual avec prédicteur-correcteur ou l'approximation de la fonction objective non différentiable par une fonction régulière.L'approximation est combinée à un algorithme de type Newton-Raphson et est connue sous le nom de méthode de lissage fini (Clark and Osborne[32], Madsen and Nielsen[105]). S'agissant du calcul des intervalles de confiance dans le cadre de la régression quantile, nous avons la méthode directe qui calcule les intervalles de confiance en utilisant la normalité asymptotique de l'estimateur ; la méthode du score de rang qui est basée sur l'inversion du test de score de rang et enfin les méthodes de reéchantillonnage basées sur le bootstrap. Pour les détails de ces méthodes, les références suivantes sont recommandées (Koenker[84], Kocherginsky et al.[81]). L'approche par les MM algorithms est également utilisée (voir Hunter and Lange[73]) mais elle ne sera pas développée dans cette partie. Chen et Wei[25] ainsi que les références incluses ont donné les détails des méthodes utilisées dans les différents logiciels statistiques pour ce qui est de la régression quantile non pénalisée.

1.4.2 Régression quantile pénalisée : Algorithmes et méthodes

Chemins de solutions régularisés linéaires par morceaux

Rosset et Zhu[125] ont donné une généralisation des chemins de régularisation pour divers problèmes de régression pénalisée. En effet en considérant un problème de type :

$$\hat{\beta}(\lambda) = argmin_{\beta \in \mathbb{R}^p} l(y, X\beta) + \lambda J(\beta), \tag{1.15}$$

où $\lambda \geq 0$ est le paramètre de régularisation. Ils ont souligné le fait que dans le cas du Lasso (Efron et al.[42]), ce qui correspond à une fonction de perte quadratique et $J(\beta) = \| \beta \|_1$, la norme L_1 de β, le chemin de régularisation du coefficient optimal est linéaire par morceaux, c'est à dire $\partial\hat{\beta}(\lambda)/\partial\lambda$ est constante par morceaux. Une généralisation en termes de caractérisation des propriétés de la fonction de perte et

de J ayant des chemins de régularisation linéaires par morceaux a ainsi été établie, en considérant $l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ et $J : \mathbb{R}^p \to \mathbb{R}$, avec $J(0) = 0$.

Dans le cas où le chemin optimal $\hat{\beta}(\lambda)$ est linéaire par morceaux, nous avons l'existence de $\lambda_0 = 0 < \lambda_1 < ... < \lambda_m = \infty$ et $\gamma_0, \gamma_1, ..., \gamma_{m-1} \in \mathbb{R}^p$ de telle sorte que $\hat{\beta}(\lambda) = \hat{\beta}(\lambda_k) + (\lambda - \lambda_k)\gamma_k$ pour $\lambda_k \leq \lambda \leq \lambda_{k+1}$. Cette formulation est très intéressante dans la mesure où nous avons l'intégralité du chemin de régularisation $\hat{\beta}(\lambda), 0 \leq \lambda \leq \infty$ en calculant simplement le pas entre deux valeurs consécutives de λ ainsi que les directions $\gamma_1, ..., \gamma_{m-1}$. Un exemple typique est le Lasso dont l'algorithme typique de résolution est le LAR-Lasso. Les problèmes ayant la propriété de linéarité par morceaux sont ceux pour lesquels la fonction de perte l est quadratique par morceaux et J est une fonction linéaire par morceaux. Afin d'avoir la linéarité par morceaux du chemin de régularisation, cela nécessite que $\frac{\partial \hat{\beta}(\lambda)}{\partial \lambda} / \parallel \frac{\partial \hat{\beta}(\lambda)}{\partial \lambda} \parallel$ soit un vecteur constant par morceaux comme fonction de λ. En utilisant le développement de Taylor des équations normales pour le problème de minimisation 1.15, (Rosset and Zhu[125]) ont montré que si l et J sont deux fois différentiables dans le voisinage d'une solution $\hat{\beta}(\lambda)$, alors

$$\frac{\partial \hat{\beta}(\lambda)}{\partial \lambda} = -[\nabla^2 l(\hat{\beta}(\lambda)) + \lambda \nabla^2 J(\hat{\beta}(\lambda))]^{-1} \nabla J(\hat{\beta}(\lambda)) \tag{1.16}$$

$l(\hat{\beta}(\lambda))$ dépend bien sûr de X et y qui sont supposés constants.

Fonctions de perte linéaires avec pénalité de type L_1

Nous nous intéresserons aux fonctions de perte linéaires par morceaux et non différentiables rencontrées aussi bien en régression (exemple de la régression sur les quantiles) qu'en classification (cas du SVM). Par exemple dans le cas de la régression sur les quantiles la fonction de perte a la forme :

$$l(y, x^T\beta) = \begin{cases} \tau. \mid y - x^T\beta \mid & \text{si } y - x^T\beta \geq 0 \\ (1-\tau). \mid y - x^T\beta \mid & \text{si } y - x^T\beta < 0 \end{cases}$$

et dans le cas des SVM la fonction de perte considérée (hinge loss) a l'expression $l(\tilde{y}, x^T\beta) = (1 - \tilde{y}x^T\beta)_+$, $\tilde{y} = 1$ ou -1. Les deux fonctions précédentes peuvent être généralisées dans l'écriture suivante :

$$l(r) = \begin{cases} b_1. \mid a + r \mid & \text{si } a + r \geq 0 \\ b_2. \mid a + r \mid & \text{si } a + r < 0 \end{cases}$$

avec le résidu r ayant la forme générale suivante :

$$r = \begin{cases} y - x^T\beta & \text{cas de la régression} \\ \tilde{y}.x^T\beta & \text{cas de la classification.} \end{cases}$$

Cette formulation nous donne le résultat suivant (voir Proposition 3 dans Rosset and Zhu[126]) :

Proposition 1.4.1 *Pour les fonctions de perte linéaires de forme générale précédente, il existe un ensemble de valeurs du paramètre de régularisation* $\lambda_0 = 0 < \lambda_1 < ... < \lambda_m = \infty$ *de telle sorte que :*

- *La solution* $\hat{\beta}(\lambda_k)$ *n'est pas définie de manière unique et l'ensemble des solutions optimales pour chaque valeur de* λ_k *est une ligne droite dans* $\mathbb{R}^p$
- $\forall\lambda \in (\lambda_k, \lambda_{k+1})$ *la solution* $\hat{\beta}(\lambda)$ *est fixe et égale à la solution de norme* L_1 *minimale pour* λ_k *et la solution de norme* L_1 *maximale pour* λ_{k+1}.

Le chemin de régularisation précédent est constant par morceaux comme fonction du paramètre de régularisation λ avec des sauts aux points $\lambda_1, ..., \lambda_m$. La linéarité par morceaux est ainsi toujours conservée (linéarité par rapport à la norme L_1 du chemin de solution $\| \hat{\beta}(\lambda) \|_1$).

Régression quantile et pénalités de type Lasso adaptatif et SCAD

La non convexité de la pénalité SCAD complique le problème d'optimisation. A cet effet, l'algorithme de différence convexe (An and Tao[6]) est utilisé. C'est un

algorithme local qui fait décroître la valeur de la fonction objectif à chaque itération et converge en un nombre fini d'itérations. Il fait partie de la classe des MM algorithmes avec la particularité que ce sont des approximations linéaires qui sont utilisées contrairement à l'approche de Hunter and Li[74] basée sur une approximation quadratique à chaque itération. D'autre part, Kwon et al.[98] ont proposé une approximation quadratique de la fonction de perte combinée avec la pénalité SCAD. Cela nécessite la décomposition de la pénalité en une somme de fonctions convexe et concave afin d'utiliser l'algorithme CCCP proposé par Kim et al.[80] dans le cas de grande dimension. Récemment, nous avons proposé (voir Sidi zakari et al.[132]) une méthode de sélection de variables basée sur une approximation du SCAD dans le cadre de la vraisemblance pénalisée. Cette contribution constitue la majeure partie du chapitre 5.

Une approche unifiée d'approximation quadratique de type moindres carrés (incluant plusieurs fonctions de perte non dérivables comme la fonction de régression quantile) a été récemment proposée par Kwon et al.[98] pour le cas de la pénalité SCAD comme prolongement du travail de Wang and Leng[141] pour le Lasso (incluant le Lasso adaptatif). Cette approche repose sur un estimateur consistant ainsi qu'une approximation de la matrice de variance covariance asymptotique. L'estimateur obtenu a les propriétés d'oracle dans le cas classique $p < n$. L'adaptation en grande dimension $p > n$ est un champ encore ouvert malgré une proposition de Wang et Leng[141] pour le cas du Lasso.

1.4.3 Cas de plusieurs pénalités

En considérant le problème ci dessous :
$\hat{\beta}(\lambda_1, \lambda_2, ..., \lambda_q) = min_\beta l(y, X\beta) + \lambda_1 J_1(\beta) + \lambda_2 J_2(\beta) + ... + \lambda_q J_q(\beta)$.
Comme évoqué par Rosset and Zhu (2004), tant que tous les termes de pénalités et la fonction de perte vérifient les conditions pour la linéarité par morceaux, la solution $\hat{\beta}(\lambda_1, \lambda_2, ..., \lambda_q)$ sera une surface affine par morceaux dans $\mathbb{R}^p$. En s'intéressant au cas particulier de la droite unidimensionnelle dans l'espace $(\lambda_1, \lambda_2, ..., \lambda_q)$, i.e $\lambda_k = b_k \lambda_1$, l'on obtient un chemin de solution linéaire par morceaux. Deux cas particuliers très importants sont les pénalités locales ainsi que la combinaison de

deux termes de pénalité. Notre travail a essentiellement porté sur ces deux aspects. S'agissant des pénalités locales, on utilise les fonctions de pénalité $J_k(\beta) =\mid \beta_k \mid$ et les différents λ_k correspondent à différents facteurs d'échelle (normalisation) des prédicteurs.

Récemment Slawski[134] a proposé une généralisation de l'Elastic net pour SVM et la régression quantile. Des algorithmes ont été proposés afin d'avoir les chemins de régularisation. Les solutions sont évaluées en faisant varier un paramètre de régularisation tout en fixant le second. Les chemins de solutions $\hat{\beta}_{\lambda_2}(\lambda_1)$ et $\hat{\beta}_{\lambda_1}(1/\lambda_2)$ sont respectivement des fonctions linéaires par morceaux comme illustrés dans Slawski[134] et Wang and al.[143]. Dans le cas des SVM avec pénalité de type Elastic net, le coût de calculs est de l'ordre de $O(pmin^2(n,p) + min^3(n,p))$ comme spécifié par Wang et al.[143].

D'autres types de pénalités ont également été proposées dans le cadre de la régression quantile multiple. L'objectif est d'estimer les quantiles conditionnels simultanément à partir d'un sous ensemble de variables communes en utilisant notamment une pénalité de type norme matricielle ou une pénalité à effet groupant (voir par exemple Yuan and Lin[153]) comme évoqué par Zou and Yuan[164]. L'approche d'estimation multiple (régression quantile composite) a également été traitée par Zou and Yuan[162]. Les propriétés d'oracle ainsi que le cas de la variance finie et/ou infinie y ont été également abordés.

Le concept d'estimation et/ou de prédiction multiple à partir d'un sous ensemble de prédicteurs a été developpé dans Turlach et al.[140] et Breiman and Friedman[22]. Nous avons effectué une application dans ce sens en considérant les indices de fraîcheur et de qualité du merlan tout en effectuant la sélection de variables. Cela nous a permis d'avoir les prédicteurs communs à ces deux indicateurs. Les détails sur cette base de données sont présentés dans les chapitres 2 et 3. S'agissant de la norme sup ($\| \, . \, \|_\infty$), Zou and Yuan[163] ont traité le cas des SVM avec la norme F_∞ tandis que Zhang and al.[156] ont traité le cas de la sélection de variables pour SVM multi-catégories avec régularisation de type norme sup adaptative. Liu and Wu[104] ont proposé la combinaison des pénalités L_0 et L_1 dans le cas général de la classification, de la régression quadratique et linéaire. Cette ap-

proche nécessite une association de la programmation entière à la programmation linéaire ou quadratique (mixed-integer linear programming or mixed-integer quadratic programming). La notion de sélection de groupes de variables a été traitée par Kato[79] dans le cas de la régression quantile pénalisée en grande dimension. Une borne non asymptotique de l'erreur quadratique de l'estimateur a été proposée et cette dernière permet d'expliquer dans quels cas l'estimateur obtenu avec la pénalité de groupe est potentiellement supérieur ou non à celui obtenu via la régularisation L_1 en termes d'erreur estimation. Une extension du choix du paramètre de régularisation initialement proposé par Belloni and Chernozhukov[16] a également été proposée.

D'autre part, la sélection de variables peut être également vue sous l'angle de la programmation linéaire comme détaillé par Yao and Lee[150]. Pour des raisons de simplicité nous ne développerons pas cet aspect.

Deuxième partie

Study of Merlan data set

Chapitre 2

Variable selection and quantile regression on freshness characterization of whiting (*Merlangius merlangus*)

The freshness and spoilage indices of whiting influenced by a large number of chemical volatile compounds are here analyzed. By means of recent statistical variable selection methods based on penalized linear regression, a small number of volatile compounds, fewer than sample size, that characterize the two indices are identified. Then, two separated linear regression approaches are conducted on the two indices and the selected volatile compounds. However, since the two response variables (indices) present many variations, linear regression based on the mean estimation is not appropriate, which means there is more than a single slope (rate of change) describing the relationship between response variables and selected volatile compounds. We show in this chapter that the quantile regression approach which estimates multiple rates of change (slopes) from minimum to maximum responses provides a more complete picture of the relationship between variables missed by linear regression.

2.1 Introduction

Fish freshness is a key attribute of the quality of fish, a highly perishable product. The industry relative to this sector is an important contributor to many economies in the world. One of the senses consumers use to assess the freshness of fish is smell. However, the smell of fish changes rapidly according to the product's degree of freshness, and this is why sensory analysis is used by consumers and industrialists to assess the fish quality. Then, the key volatile compounds that contribute to this characteristic odor are measured and used as quality indicators ([38],[37],[40]). These volatile aromatic compounds that characterize smell are generated by the action of bacterial, tissue enzymes and lipid autoxidation.

Sustaining business competitiveness by reducing cost and maintaining product quality will be essential for this industry. One of the challenges facing this industry is to provide more information on the relationship between the freshness (and/or spoilage) index and the chemical volatile compounds that contribute to this sensory analysis.

Many studies have been carried out to identify and quantify these compounds in various fish (mackerel, herring, cod, sardine and sea bream) using either static headspace analysis (SHA), dynamic headspace analysis (DHA) or an electronic nose in various phases ([8],[112],[3],[75],[120]). Bene et al.([17],[18]) used a solid phase microextraction (SPME) technique which is a new method to extract volatiles from matrixes. They first used it to assess fish quality when studying substances in salmon and whiting. Recently, Duflos et al.[38] studied the freshness of whiting at five stages of ice storage by comparing the analysis of volatile compounds obtained through the combination of gaz chromatography/mass spectrometry (GC/MS) and SPME with two sensory methods. They used two separate steps of statistical multidimensional approaches to identify volatile compounds and characterize fish freshness. In the first step, control charts were used to control the daily progression of freshness and/or spoilage indices. The second step begins by reducing the large dimension of the data set (excluding the two indices variables) to two principal components via the application of principal component analysis (PCA) method. Then, a hierarchical clustering approach and a heuristic variable

selection were used for clustering the fish samples on three classes and to identify the volatile compounds that characterize their classes, respectively. However, the freshness indices (or response variables) are not taken into account directly in the later procedure.

This chapter studies an initial data set on *Merlangius merlangus*, analyzed by Duflos et al.[38], composed of $n = 42$ fish samples (or observations) described by $p = 60$ descriptors (or variables). After deleting some missing observations and running selection methods, the final data set is composed with $n = 37$ observations described by $p = 49$ variables. Two variables measure the quality and freshness indices and the others are the chemical volatile compounds. The goal of any fishery industry is to efficiently make a high quality product. To this end, it is imperative that the manufacturer has an advanced knowledge of the process and causality. A common goal for statistical research is to investigate and quantify the relationship between independent variables (X volatile compounds) and a response variable (y freshness index or quality index). This example is in the class of $p > n$ data sets which recently have gained considerable importance in several areas of statistics. It poses numerous challenges for statistical theory, method and implementation in those problems. For example, in multiple linear regression (MLR) when dimension p is comparable or exceeds sample size n, the ordinary least squares (OLS) estimator is not appropriate. In this setting, we would like to achieve both variable reduction and prediction accuracy. Variable selection for high dimensional data has received a lot of attention recently. In the last decade interest has focused on penalized regression methods which implement both variable selection and coefficient shrinkage in a single procedure. The best known of these procedures is Lasso.

The objective of this chapter is to explore causality of freshness process variation beyond the mean of the conditional distribution. We will show that information contained in the percentiles is one of a key measure of quality and safety concerns. First, we use MLR to modelize the relationship between the two freshness indices and volatile compounds. However, since dimension p of data set of whiting exceeds sample size n, we introduce recent statistical variable selection methods based on penalized least squares to select a small number of volatile compounds. However, even in this case with a small number of selected volatile

compounds, the MLR is not appropriate because the response variable presents many variations. It is well known that MLR develops models based on the mean of the response variable (ex. freshness index), while quantile regression (QR) approach develops models for any percentile τ ($0 < \tau < 1$) of the response variable (ex. $\tau = 0.5$ corresponds to the median). Since response variable (freshness index or quality index) presents many variations, we show in this study that QR is more adapted than MLR to interpret the relationship between freshness and quality indices and the small number of selected volatile compounds, which means modeling beyond the mean of freshness index may greatly improve a related manufacturer's understanding of the process.

In section 2, we briefly present the experimental procedure and whiting data. In section 3, we present some recent and popular variable selection Lasso-type methods, based on penalized least squares, which are adapted to MLR. In the same section, we apply these variable selection methods to select the volatile compounds which separately affect the freshness and quality indices during the conservation. Since response variables present many variations, MLR on the subset of selected volatile compounds will not result in amelioration of prediction accuracy. We provide in section 4 quantile regression (QR) statistical approach, developed on the subset of selected volatile compounds, which can improve fish manufacturer's understanding of the process. Finally, we end by a brief conclusion in section 5.

2.2 Experimental procedure and data

We describe below the experimental procedure, quoted from Duflos et al.[38], used to extract data on freshness of whiting.

Chemicals

Carboxen (CAR)/polydimethylsiloxane (PDMS) StableFlex fibre (65 μm) came from Supelco (Bellefonte, PA, USA). Before the first use, each SPME fibre was conditioned as recommended by the manufacturer. NaCl came from Oxoid Ltd

(Basingstoke, UK), Milli-Q water (high-performance liquid chromatographic water) from Fisher Scientific Labosi (Elancourt, France) and 3-methyl-3-buten-1-ol from Sigma-Aldrich (Saint Quentin Fallavier, France).

Sensory evaluation

Two methods were used for the sensory evaluation of fish. The first method is the European Union's grading system presented in European Directive 2406/96. This system distinguishes between three freshness categories, E, A and B, corresponding to various levels of spoilage. Category E corresponds to the highest quality level, followed by categories A and B, while fish graded below B is considered to be non-edible. Another category corresponds to the product's discard level. In order to rate this evaluation, these letters have been replaced with numbers : 0 = E, 1 = A, 2 = B and 3 = unacceptable. The lower the number, the fresher the fish ; conversely, the higher the number, the greater the spoilage (below 18 the fish is acceptable). The second method is the quality index method (QIM) evaluation system adapted to whiting. It is based on changes in the sensory characteristics of raw fish during spoilage. Scores of $0-1$, $0-2$ or $0-3$ demerit (or index) points are attributed according to changes observed in the smell, texture, eye appearance, skin and gills. The points are added up to obtain an overall sensory score or quality index (QI).

Sample preparation

Whiting (*Merlangius merlangus*), caught the night before the start of the study, was acquired from Cooperative Maritime Etaploise (Boulogne-sur-Mer, France). Two different catches (20 and 22 fish respectively) were analyzed. The fish were stored in crushed ice at 4°C in self-draining polystyrene boxes for 10 days. Fresh crushed ice was added daily. Sensory evaluation and volatile analysis were performed on days $1, 2, 3, 4, 7$ and 10 (on seven different fish each day). Sensory evaluation was performed by two panelists familiar with the sensory evaluation of fish. Following sensory evaluation, each fish was filleted. The next step was made according to previous work (Duflos et al.[36]). The fillets were cut into 1 cm cubes, then 50 g of flesh was introduced into a stomacher bag with 100 mL of ultrapure

water saturated with NaCl. The contents were mixed for 2 min in a Stomacher Lab-Blender 400 (Seward, Thetford, UK). The aqueous phase was removed and centrifuged at 12000 $\times$ g for 10 min at 4°C (Multifuge 3 S-R Heraeus, Kendro Laboratory Products, Courtaboeuf, France).

SPME procedure

According to previous studies (cf. Duflos et al.[36]), 11 mL of supernatant of sample preparation was introduced into a hermetically sealed 20 mL vial. The vial was placed in the sample tray of a CombiPAL (CTCAnalytics, Zwingen, Switzerland)and then transferred to the mixer, where it was heated at $50^{\circ}C$ and mixed at 500 rpm for 10 min. After this equilibrium time the CAR/PDMS SPME fibre was inserted into the headspace of the sample and held there for 40 min at $50^{\circ}C$. The fibre was then removed from the headspace and inserted into the Merlin Microseal injector ($250^{\circ}C$) of a GC-17A gas chromatograph (GC) equipped with an MS-QP5000 mass spectrometer (MS) (Shimadzu, Kyoto, Japan) for desorption. The fibre was maintained 10 seconds.

GC/MS procedure

The GC was equipped with a BPX5 capillary column (60 m$\times$0.25 mm$\times$0.25 μm) (SGE, Courtaboeuf, France). The GC conditions were as follows : oven temperature set initially at $35^{\circ}C$ (5 min hold), increased to $100^{\circ}C$ at $10^{\circ}Cmin^{-1}$, then increased to $280^{\circ}C$ at $20^{\circ}Cmin^{-1}$ and maintained at $280^{\circ}C$ for 5 min ; the splitless mode was used for injection, with a purge time of 2 min. The fibre was maintained in the injection port for 10 s. The electron impact MS conditions were as follows : temperature of interface, $260^{\circ}C$; ionization voltage, 70 eV ; mass range, m/z $33 - 200$; scan speed, 250 per 0.5 s. After each injection the fibre was heated to $300^{\circ}C$ for 10 min in the SPME fibre conditioner. Volatile compounds were identified by matching their mass spectra with Mass Spectral Libraries 21 and 107 of the National Institute of Standards and Technology (NIST) (developed for Shimadzu by NIST, July 2002). Semi-quantification of the components was based on arbitrary units of total current ion peak area counts.

2.3 Variable selection in linear regression

In this section we consider the multiple linear regression model for modeling the relation between the response variable (ex. freshness index) and the predictors (ex. volatile compounds). Since we are facing with a high dimensional problem where p is greater than n, so the reduction of the dimension is necessary to use classical statistical methods. We present the principal variable selection methods based on penalized least squares and their performances on the whiting data set.

2.3.1 Statistical variable selection methods

Let's consider a linear regression model

$$y = X\beta + \varepsilon, \tag{2.1}$$

where y is the variable response, $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is a nxp matrix of n subjects described by p predictors, $\boldsymbol{\varepsilon}$ is a random error vector with $\mathbb{E}(\varepsilon) = 0$ and β is a vector of unknown parameters which are to be estimated. If the rank of X is equal to p, the ordinary least squares estimator (OLS) can be written as $\hat{\beta} = (X^T X)^{-1} X^T y$, where T denotes the transpose.

However, since all predictors are introduced in the model, the OLS estimator can lead to poor performance in prediction accuracy or/and interpretation of the model in the presence of large numbers of predictors. Indeed, the OLS estimates often have low bias but large variance and prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. Moreover, with a large number of predictors, we often would like to select a smaller subset that exhibits strongest effects, so leading to good interpretation of the coefficients.

Subset selection and dimension reduction approaches

The first classical approach used to reduce the dimension of the model is the subset selection which retains a subset of the predictors and eliminates the rest from the model. Least squares regression is used to estimate the coefficients of retained predictors. The resulting reduced model is interpretable and has possibly lower prediction error than the full model. However, since variable selection based on subset selection is a discrete process (predictors are either retained or deleted) it often leads to high variance, and so doesn't reduce the prediction error of the full model. Different strategies and algorithms are proposed to choose the subset and are implemented in different statistical software. However, implementing these procedures is computationally infeasible if the number of potentially predictors is even moderately large.

On the other hand, principal component regression (PCR) and partial least squares (PLS) are the principal approaches considered to reduce the high dimension to a small number of linear combinations Z_ℓ (called factors), $\ell = 1, \ldots, L$ of the original predictors $\mathbf{x}_j, j = 1, \ldots, p$. PCR and PLS differ in how the factors Z_ℓ are constructed. However, PCR doesn't take into account the information contained in the response variable and the interpretation of the regression coefficients associated to the resulting factors (of PCR or PLS) is not easy since all predictors are used in each factor. An excellent discussion and a detailed comparison of these approaches can be found in Hastie et al. (2009).

Variable selection and shrinkage procedures

During the latest few years a great deal of attention has been focused on the penalized least squares methods which perform estimation and variable selection in a continuous way by shrinking regression coefficients towards zero and by also setting some coefficients exactly equal to zero.

a) ***Lasso procedure***

The most popular regularization approach is the **L**east **a**bsolute **s**hrinkage and **s**election **o**perator [Tibshirani[138]] which is based on the penalized least squares by the L_1 penalty on the vector parameter.

$$\hat{\beta}_{Lasso}(\lambda) = argmin_{\beta} \parallel y - X\beta \parallel_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.2}$$

where λ is a positive tuning parameter. The L_1 penalty promotes sparsity in the OLS solution vector $\hat{\beta}$, which means to shrink $\hat{\beta}$ and a tuning parameter λ controls the degree of this shrinkage. This method was made particularly appealing by the advent of the Lars algorithm (Efron et al.[42]) which provided a highly efficient means of simultaneously producing the set of Lasso fits for all values of the tuning parameter.

b) ***Elastic net procedure***

Although Lasso is a highly successful procedure, it has two drawbacks. First, Lasso can select at most n predictors when $p > n$ which can be a limiting feature of a variable selection method in many situations including those involving microarray data where $p >> n$. Secondly, if there is a group of highly correlated predictors, Lasso tends to arbitrarily select only one from the group. Many alternatives have been proposed to deal with these two drawbacks. The most popular among is the Elastic net (Enet), proposed by Zou and Hastie[160], which uses the weighted combination of Ridge and Lasso penalties to alleviate these two issues. The coefficient estimates of Enet can be written as

$$\hat{\beta}_{Enet}(\lambda_1, \lambda_2) = argmin_{\beta} \parallel y - X\beta \parallel_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \tag{2.3}$$

where λ_1 and λ_2 are two positive tuning parameters. The addition of the L_2 norm is motivated by the property of Enet to include groups of predictors which are highly correlated. That is, if the predictors are highly correlated, as for example in gene expression in microarray data, the Lasso selects only one element of the group whereas Enet selects the whole group. Zou and Hastie[160] showed that the Lasso estimates are instable when predictors are highly correlated.

c) ***Rules for discarding predictors***

In genetic microarray studies, sample size n is measured in hundreds whereas the number of predictors p per sample can be large compared to n (a million!). Shrinkage regression techniques such as Lasso or Enet are useful for dealing with this type of high-dimensional predictors but their usefulness diminishes when p exceeds largely n where time and memory are of great importance. A natural idea is to reduce the size of p from huge to a relatively large dimension $d < n$, so that well-developed variable selection procedures can be used to select the important predictors.

Independent screening procedures based on all marginal regression models lead to computationally reasonable variable selection methods [Fan and Lv[50], Fan et al.[55]]. An example of independence screening is the correlation ranking proposed by Fan and Lv[49] which ranks the predictors according to the magnitude of its sample correlation with the response variable. Another is the p-values ranking procedure which ranks the coefficient p-values of p simple linear regression of each predictor with the response variable. However, independent screening strategies are not from an optimization point of view and may put set of coefficients to zero that are nonzero in the solution.

For computational efficiency, El Ghaoui et al.[45] propose "SAFE" rules for discarding predictors in Lasso regression problems. It is based on univariate inner products between each predictor and the response variable, that guarantee a coefficient of (4.3) will be zero in the solution vector. More precisely, fitting at λ the basic SAFE rule for Lasso discards predictor $\mathbf{x}_j$ if

$$|\mathbf{x}_j^T y| < \lambda - \| \mathbf{x}_j \|_2 \| y \|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}} \tag{2.4}$$

where $\lambda_{\max} = \max_\ell |\mathbf{x}_\ell^T y|$ is the smallest λ for which all coefficients are zero, $\| \mathbf{v} \|_2 = (\sum_j^p v_j^2)^{1/2}$. This bound is derived from a dual of the Lasso problem (4.3). Tibshirani et al.[139] claimed that the SAFE will never discard a predictor when its coefficient is truly nonzero. However, they can delete fewer predictors than strong sequential rule of Tibshirani et al.[139] which we will present below.

When the predictors are standardized ($\| \mathbf{x}_j \|_2 = 1$ for each j), then $\lambda_{\max} < \| \mathbf{y} \|_2$, so that the right bound of (4.5) is always smaller than $2\lambda - \lambda_{\max}$. Then, the basic strong rule in Tibshirani et al.[139] discards predictor j if

$$|\mathbf{x}_j^T y| < 2\lambda - \lambda_{\max}. \tag{2.5}$$

When the predictors are not standardized, the strong rule seems to discard more predictors unless the predictors have widely different marginal variances. To overcome this problem, they proposed instead the strong sequential rules (SSR) which delete a predictor j if

$$|\mathbf{x}_j^T \mathbf{r}| < 2\lambda - \lambda_0, \tag{2.6}$$

where $\mathbf{r} = y - X\hat{\beta}(\lambda_0)$ is the residual computed for the solution $\hat{\beta}(\lambda_0)$ at λ_0. In their empirical studies, Tibshirani et al.[139] claimed that SSR have a tendency to discard almost all predictors that have regression coefficients of zero.

All these approaches are implemented in the R software package "glmnet" where the latter authors have used the parametrization $(\alpha\lambda, (1-\alpha)\lambda)$ instead of (λ_1, λ_2) for Elastic net. So, the strong sequential rules associated to Elastic net is simply

$$|\mathbf{x}_j^T \mathbf{r}| < \alpha(2\lambda - \lambda_0), \tag{2.7}$$

where $\mathbf{r}$ is the residual computed for the Elastic net solution at $\alpha\lambda_0$. Lasso and Enet methods based on strong rules are noted by SLasso and SEnet.

2.3.2 Results and Discussion

We present below the empirical results of the analysis of freshness and spoilage indices described by a great number of volatile compounds. We begin with a brief introduction analysis of the data set and we describe two principal models chosen by principal variable selection methods. Computations and plots for this section have been made by R software and related packages. Specially, we use packages "lars","elasticnet" and "glmnet" for variable selection.

Analysis of freshness and quality indices by sampling day

The study consists of 42 measurements taken during six periods, D1 (day 1), D2 (day 2), D3 (day 3), D4 (day 4), D7 (day 7) and D10 (day 10), at a rate of seven measurements per period and some 60 volatile compounds observed per sample (cf. Table 2.3). Figure 2.1 shows the respective freshness indices and quality scores

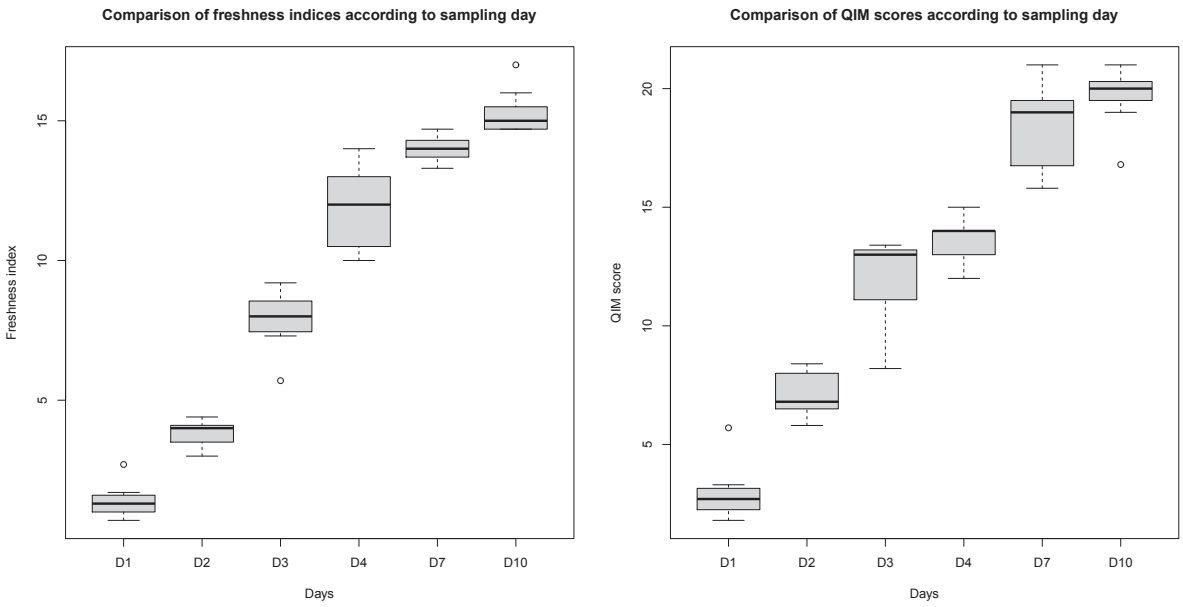

FIGURE 2.1 – Freshness and Quality indices boxplots.

associated with freshness and spoilage of the analyzed whiting catches. They illustrate both the average trend and the variability of the two indices (freshness and quality) by sampling day. We note that these two freshness and/or spoilage indicators tend to rise over sampling time. The daily variability of the two indicators is illustrated by the height of each box plot. This variation is greater on D3 and D4 for the freshness index and on D3 and D7 for the quality score. Three fishes on D1, D3 and D10, have freshness indices that are relatively far from the average, while two fishes on D1 and D10 have a quality score that is relatively far from the average. Concerning the freshness index, the estimate of means and standard deviations clearly shows a relatively significant difference between D2 and D3 on the one hand and between D4, D7 and D10 on the other hand. It thus appears that D3 represent a breaking point between high and low quality of the analyzed fish samples freshness.

According to Figure 2.2, the two indices therefore decrease between D2 and D3.

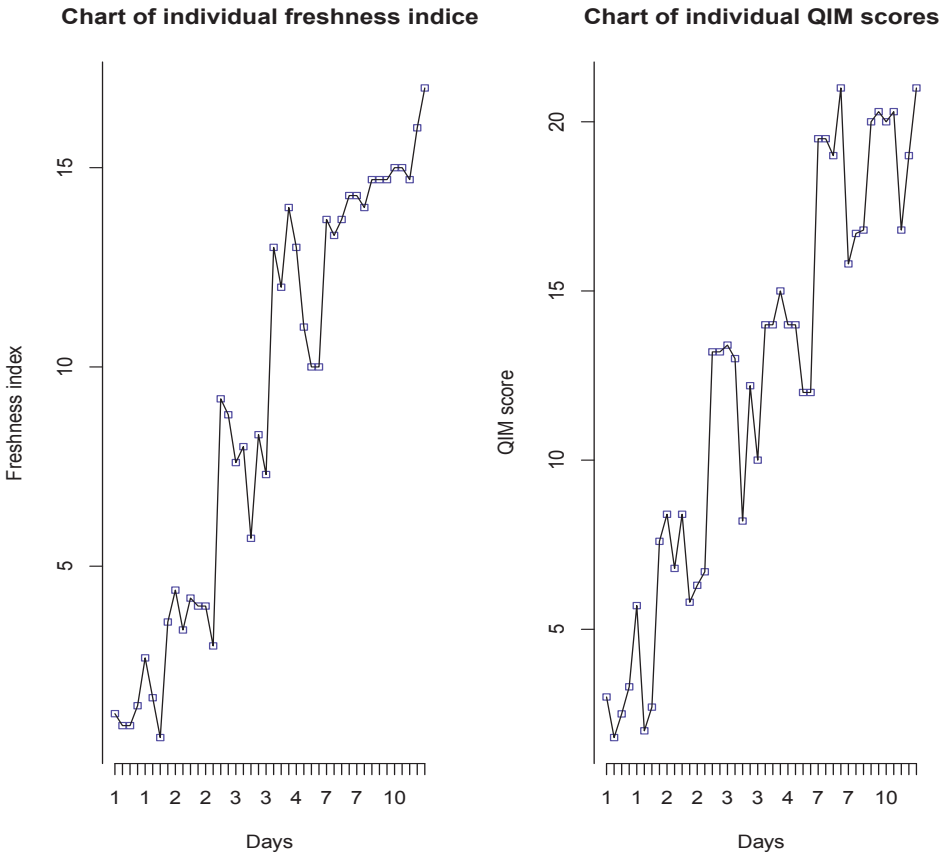

FIGURE 2.2 – Charts representation for Freshness and Quality indices.

Then, there is a transition phase from D3 to D4, leading to a confirmed spoilage zone starting on D4. These two graphs explain the upward trend observed above and specify the quite linear nature of this trend in relation to the sampling days. In the lines below since we are interested in variable selection for linear regression we will not focus on days for freshness and quality indices.

Selection of Compounds

In order to perform variable selection, we run methods on data sets 25 times using bootstrap with resampling.

Table 2.1 presents the most selected compounds related to freshness index. It can be seen that compounds *Trimethylamine*, *Ethanol*, *Ethyl acetate*, *1-Butanol*, *3-Methyl-1-butanol*, *1-Pentanol*, *2-Hexanone*, *Limonene* and *Nonanal* have been

Variables	Methods			
	Lasso	SLasso	Enet	SEnet
Trimethylamine	24	24	25	25
Ethanol	25	25	25	25
Dimethylsulfide	22	24	19	23
Carbondisulfide	22	24	19	23
Ethyl acetate	25	25	25	25
2-Methyl-1-propanol	4	22	25	22
3-Methyl butanal	0	0	11	0
1-Butanol	25	25	25	25
3-Hydroxy-2-butanone	18	23	21	23
3-Methyl-1-butanol	23	24	25	24
2-Methyl-1-butanol	3	1	4	7
1-Pentanol	22	24	23	24
2-Hexanone	23	24	25	25
Cyclohexanone	0	0	11	0
Limonene	25	25	25	25
Nonanal	20	23	23	24

TABLE 2.1 – Frequencies of variables selected for freshness index.

selected more than 19 times by all methods. On the other hand, compounds *3-Methyl-butanal* and *Cyclohexanone* are only selected by Enet (11 times), *2-Methyl-1-propanol* has been selected 4 times by Lasso and *2-Methyl-1-butanol* has the lowest rates of selection. We also remark that compounds *Ethanol*, *Ethyl acetate*, *1-Butanol* and *Limonene* have been selected all of the 25 times by selection methods.

Furthermore, the analysis of correlations between selected compounds reveals that the following three pairs (*3-Methyl-1-butanol*, *2-Methyl-1-butanol*), (*1-Pentanol*, *2-Hexanone*) and (*1-Pentanol*, *Nonanal*) are highly correlated with linear correlation coefficient $\rho = 0.942$, $\rho = 0.917$ and $\rho = 0.797$, respectively. Consequently, according to selection frequencies of each compound, the Variance Inflation Factor (VIF) which is related to the degree of multi-collinearity and the p-values of regression coefficients ; we finally choose the Model-F ={*Ethanol*, *Dimethyl sulfide*, *Ethyl acetate*, *2-Methyl-1-propanol*, *1-Butanol*, *3-Methyl-1-butanol*, *2-Hexanone*, *Nonanal*} composed of eight(8) compounds. Its R-squared and Adjusted R-squared are about 0.806 and 0.751, respectively. For this model, the maximum VIF value of predictors is 2.858 and the corresponding VIF mean value is 1.934. This does not

Variables	Methods			
	Lasso	SLasso	Enet	SEnet
Trimethylamine	1	0	1	2
Ethanol	25	25	25	25
Dimethylsulfide	7	0	1	3
Carbondisulfide	15	1	2	6
Ethyl acetate	25	24	25	25
1-Butanol	25	24	25	25
3-Hydroxy-2-butanone	3	0	1	3
3-Methyl-1-butanol	6	0	17	25
2-Methyl-1-butanol	21	25	25	25
(E)-2-pentenal	21	24	24	24
1-Pentanol	25	24	25	25
2-Hexanone	20	23	25	23
Cyclohexanone	2	0	0	0
Limonene	6	0	2	0
Nonanal	8	1	2	8

TABLE 2.2 – Frequencies of variables selected for quality index.

show that the regression suffers from multi-collinearity. Finally, Model-F's OLS estimates in Table 2.4 show that compounds *Ethyl acetate*, *2-Methyl-1-propanol*, *Nonanal* are significant at level 5% and *2-Hexanone* at level 0.1%.

Table 2.2 presents the most selected compounds by the four variable selection methods related to quality index. It can be seen that eight compounds *Ethanol*, *Ethyl acetate*, *1-Butanol*, *2-Methyl-1-butanol E2pentenal*, *1-Pentanol* and *2-Hexanone* are selected more than 19 times by all methods. The compounds *Trimethylamine* and *Cyclohexanone* have low selection frequencies followed by *Dimethyl sulfide*, *3-Hydroxy-2-butanone*, *Limonene*, *Nonanal*, *3-Methyl-1-butanol* and *Carbon disulfide*. *Ethyl acetate*, *1-Butanol* and *1-Pentanol* have the same selection frequencies for all variable selection methods and *Ethanol* is selected 25 times by all methods. In terms of correlations, in addition to the three previous correlations, we also found that *(E)-2-pentenal* is highly correlated with *1-Pentanol*, *2-Hexanone* and *Nonanal* (with $\rho = 0.937$, $\rho = 0.833$ and $\rho = 0.854$, respectively). The compound *(E)-2-pentenal* is not selected for freshness index while *2-Methyl-1-propanol* and *3-Methyl butanal* are discarded for quality index.

As for freshness index, according to selection frequencies of each compound, the Variance Inflation Factor (VIF) and p-values of regression coefficients, we finally choose the Model-Q={ *Ethanol*, *Ethyl acetate*, *1-Butanol*, *3-Methyl-1-butanol*, *2-Hexanone*}. Its R-squared and Adjusted R-squared are equal to 0.717 and 0.671, respectively. For this model, the maximum VIF value of predictors is 2.598 and the corresponding VIF mean value is 1.852. This does not show that the regression suffers from multi-collinearity. Finally, its OLS estimates in Table 2.5 show that compounds *Ethanol* and *2-Hexanone* are significant at level 5% and *Ethyl acetate* at level 10%.

Variable name	Compound name	Variable name	Compound name
a	Acetaldehyde	ag	3-Methyl-1-butanol
b	Methanethiol	ah	2-Methyl-1-butanol
c	Trimethylamine	aj	(E)-2-pentenal
d	Ethanol	ak	1-Pentanol
e	Pentane	al	(Z)-2-Penten-1-ol
f	Propanal	am	2-Hexanone
g	Dimethyl sulfide	an	Hexanal
h	Methylene chloride	ao	4,4-Dimethyl-1,3-dioxane
i	Carbon disulfide	ap	(E)-2-Hexenal
j	2,3-Butanedione	aq	1-Hexanol
k	Butanal	ar	3-Heptanone
l	2-Butanone	as	(Z)-4-Heptenal
m	2-Butanol	at	Heptanal
o	Ethyl acetate	au	Cyclohexanone
p	Acetic acid, ethyl ester	av	1-Heptanol
q	2-Methyl-1-propanol	aw	3,5,5-Trimethyl-2-hexene
r	3-Methyl butanal	ax	Benzaldehyde
s	1-Butanol	ay	1-Octen-3-ol
t	2-Methyl butanal	az	2,3-Octanedione
v	1-Penten-3-ol	ba	2,4-Heptadienal
w	1-Penten-3-one	bb	Octanal
x	Heptane	bc	(E,E)-2,4-Heptadienal
y	2-Ethyl furan	bd	2-Ethyl-1-hexanol
z	3-Pentanone	be	Limonene
aa	2,3-Pentanedione	bf	1-Octanol
ab	Pentanal	bg	3,5-Octadien-2-one
ac	3-Pentanol	bh	Nonanal
ae	3-Hydroxy-2-butanone		

TABLE 2.3 – Volatile compounds identified during spoilage analysis of whiting.

2.4 Quantile regression on selected volatile compounds

The latter multiple linear regression study of freshness and quality indices on reduced number of volatile compounds suggests that some of the observed variation in the estimated freshness or quality index of some volatile compounds may reflect the fact that characteristics are not estimated the same way across a given distribution of freshness (or quality) index. To examine this issue, we use quantile regression in the following section to identify the coefficients of large set of diverse predictors across different quantiles.

2.4.1 Results and Discussion

Models and estimations

In this subsection, we present results of quantile regression models for freshness and quality indices. For quantile regression, estimates are taken at the following quantiles :

$\tau \in \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 0.95\}$ but in Table 2.4 and Table 2.5 we only present results for quantiles : $0.1, 0.25, 0.5, 0.75$ and 0.9. OLS and quantile regression estimates are given with their corresponding estimated standard errors reported between brackets. We recall that standard errors and confidence intervals for the quantile regression estimates can be obtained by direct (asymptotic) and bootstrapping methods. As advocated in Koenker and Hallock[92], the two approaches lead to robust results and estimated standard errors presented here are based on classical independent and identically distributed (i.i.d) errors, which leads to direct estimation of the variance covariance matrix of the estimates. In Figures 2.3 and 2.4, each OLS estimate is represented by a continuous red line, while its lower and upper limits confidence interval are in red dashed lines. Quantile regression estimates are in a black broken line with 90% confidence band region in grey. Slopes and intercept of each estimated linear quantile regression model are plotted as a function of quantile τ.

	OLS	Quantiles				
variables		0.1	0.25	0.5	0.75	0.9
(Intercept)	4.405**	0.443	0.852	5.249**	8.208***	7.325****
	(1.325)	(0.619)	(1.005)	(1.887)	(1.726)	(1.136)
Ethanol	1.308	2.904****	2.959***	1.786	1.312	1.023
	(1.001)	(0.468)	(0.759)	(1.425)	(1.304)	(0.858)
Dimethyl	0.771	1.633***	1.064+	0.867	-0.108	-0.016
sulfide	(0.817)	(0.382)	(0.620)	(1.164)	(1.064)	(0.700)
Ethyl	2.941*	1.547**	2.191*	1.742	2.199	3.102**
acetate	(1.066)	(0.498)	(0.808)	(1.518)	(1.389)	(0.914)
2-Methyl-	1.696*	2.607****	2.613***	1.407	0.767	0.663
1-propanol	(0.742)	(0.347)	(0.563)	(1.058)	(0.967)	(0.637)
1-Butanol	-1.289	-1.295**	-1.334+	-0.871	-0.834	0.621
	(0.966)	(0.451)	(0.732)	(1.376)	(1.258)	(0.828)
3-Methyl-	1.137	1.446**	1.241+	0.836	0.343	1.732*
1-butanol	(0.916)	(0.428)	(0.695)	(1.305)	(1.193)	(0.785)
2-Hexanone	-3.128***	-2.067***	-1.626*	-3.626**	-4.313***	-3.725***
	(0.806)	(0.377)	(0.611)	(1.148)	(1.050)	(0.691)
Nonanal	2.265*	1.700***	1.353+	2.631+	3.135*	1.785*
	(0.968)	(0.452)	(0.734)	(1.379)	(1.261)	(0.830)

TABLE 2.4 – Model-F for freshness index with significance levels : 10%+,5%∗,1%∗∗,0.1%∗∗∗,0%∗∗∗∗

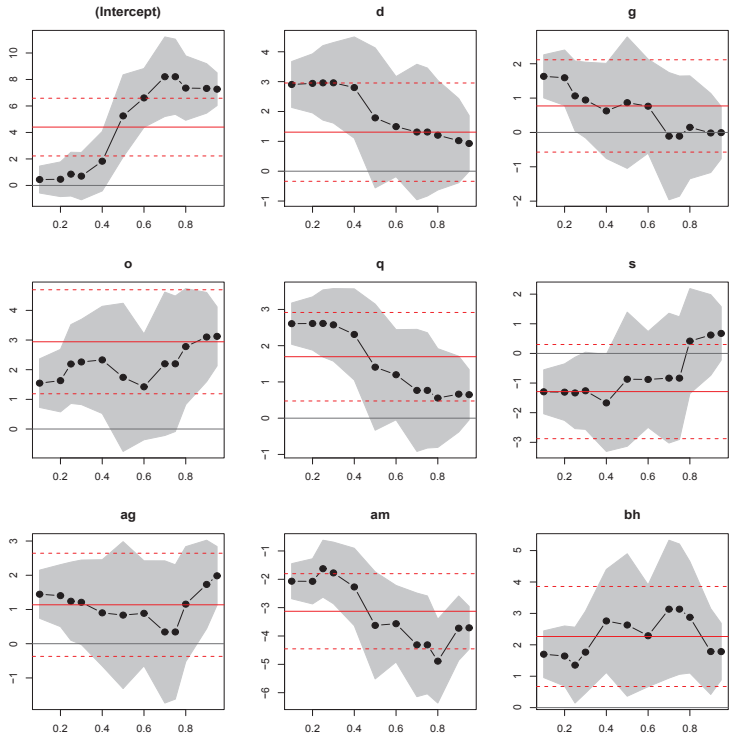

FIGURE 2.3 – Quantile plots for freshness index.

Freshness index Model-F

It can be seen from Table 2.4 that OLS regression coefficients of only four compounds *Ethyl acetate*, *2-Methyl-1-propanol*, *Nonanal* and *2-Hexanone* are significantly different from zero at level of 5% for the first three compounds and 0.1% for *2-Hexanone*. While the QR coefficient estimates of all compounds are significant at least at level of 1%. For quantile $\tau = 0.25$, compounds *Ethanol* and *2-Methyl-1-propanol* are significant at level 0.1% and *Ethyl acetate* and *2-Hexanone* are significant at 5%. Compounds *Dimethyl sulfide*, *1-Butanol*, *3-Methyl-1butanol* and *Nonanal* are significant at level of 10%. We mention that the intercept is not significant for $\tau = 0.1$ and $\tau = 0.25$. Moving from $\tau = 0.25$ to $\tau = 0.75$, all coefficients become insignificant except the coefficients estimates of *2-Hexanone* and *Nonanal*. For $\tau = 0.9$, only *Ethyl acetate*,*3-Methyl-1butanol*, *2-Hexanone* and *Nonanal* are significant. Finally, the coefficient compound *2-Hexanone* is negatively significant for all considered quantiles ; compound *Nonanal* is positively significant for all quantiles when considering 10% levels at $\tau = 0.25$ and $\tau = 0.5$.

Considering estimates size and direction, we can see for example that OLS coefficient of *Ethyl acetate*(2.941) overestimates quantile regression ones except for $\tau = 0.9$, where the estimation is about 3.102. On the other hand, *2-Methyl-1-propanol's* OLS coefficient(1.696) underestimates lower quantile effects with values 2.607 and 2.613, respectively for $\tau = 0.1$ and $\tau = 0.25$. For compound *2-Hexanone*, the corresponding OLS estimate(-3.128) underestimates regression quantiles for $\tau = 0.1$ and $\tau = 0.25$ which corresponding values are -2.067 and -1.626 ; but for the median and upper quantiles ($\tau = 0.75$ and $\tau = 0.9$) this effect is overestimated with values -3.626,-4.313 and -3.725. Finally, for compound *Nonanal*, OLS estimate(2.265) overestimates regression quantiles at $\tau = 0.1$, 0.25, 0.9 with respectively values about 1.700, 1.353 and 1.785. Median ($\tau = 0.5$) and $\tau = 0.75$ effects are underestimated with values 2.631 and 3.135. For this model, except compounds *2-Hexanone* and *Nonanal*, the remaining six(6) compounds have little effect on freshness index ranges from the median to $\tau = 0.9$ because for those quantiles, the mentioned volatiles are not significant except for *Ethyl acetate* and *3-Methyl-1butanol* at $\tau = 0.9$. For compound *Ethanol*, regression quantile estimates at $\tau = 0.1$ (2.904) and $\tau = 0.25$ (2.959) are three(3) times as great as for $\tau = 0.9$ (1.023). Compound *Dimethyl sulfide* estimate at $\tau = 0.1$ (1.633) is around one hundred(100) times greater than for absolute value of $\tau = 0.9$ (-0.016). *Ethyl acetate* estimate at $\tau = 0.9$ (3.102) is twice as great as for $\tau = 0.1$ (1.547). As partial conclusion for model-F, we can say that all of the eight(8) considered volatile compounds effects are remarked for fresh fishes($\tau = 0.1$ and $\tau = 0.25$). Volatiles *2-Hexanone* and *Nonanal* affect fresh fishes as well as fishes with great spoilage ($\tau = 0.9$). *Ethyl acetate* and *3-Methyl-1butanol* affects also fishes with great spoilage.

All previous remarks can be seen clearly on Figure 2.3, where in the first panel, the intercept of the model can be interpreted as the conditional quantile estimate of the freshness distribution for a fish with zero proportion of *Ethanol(d)*, *Dimethyl sulfide(g)*, *Ethyl acetate(o)*, *2-Methyl-1-propanol(q)*, *1-Butanol(s)*, *3-Methyl-1butanol(ag)*, *2-Hexanone(am)* and *Nonanal(bh)*. In each plot, the regression coefficient at a specific quantile indicates the effect on freshness of a unit change in the corresponding variable, assuming that the other variables are fixed, for example here with 90% confidence interval bands. The quantile regression plots give us on

overview and help us understand how variable these effects can be. These plots also highlight the fact that a classical linear regression might not be a suitable approach to explain this relationship. We highlight twelve (12) specific quantiles in our plots which correspond to $\tau \in \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 0.95\}$, so, our analysis will be based on lower ($\tau \in [0.1, 0.3]$), middle ($\tau \in [0.3, 0.6]$) and upper quantiles ($\tau \in [0.6, 0.95]$). A regression coefficient is said to be significantly or statistically different from zero if its related confidence interval does not include zero.

As a general tendency, the coefficients of compounds, *Ethanol(d)*, *Dimethyl sulfide(g)*, *2-Methyl-1-propanol(q)* and *2-Hexanone(am)* have a decreasing slopes. *Ethyl acetate(o)* and *1-Butanol(s)* slopes are characterized by an increasing trend.

Finally, compounds *3-Methyl-1butanol(ag)* and *Nonanal(bh)* are characterized by a decreasing-increasing or increasing-decreasing behavior generally around the median ($\tau = 0.5$). For compound *Ethanol(d)*, regression coefficients are only significant for lower and middle quantiles with practically no variation for lower quantiles ; since the lower the freshness index, the better the fish, we can say that this compound has a great impact during first days storage and that this effect decreases during whiting conservation. For this compound, the OLS estimate is not significantly different from zero and very close to quantile regression results for $\tau \in \{0.7, 0.75, 0.8\}$. Compound *Dimethyl sulfide(g)* estimates are also only significant for lower quantiles with more effect than middle and upper quantiles. Middle quantiles effects are very close to OLS estimate and upper quantiles estimates seem to vanish ; so, this compound seems to have no effect on fishes with relatively high level of spoilage, but acceptable for consummation.

Ethyl acetate(o) compound effect is only significantly different from zero for lower, first part of middle and last part of upper quantiles with an increasing trend. A decreasing-increasing trend is observed for the last part of middle quantiles and finally increasing trend for upper quantiles whose last part estimates are close to the significantly different from zero OLS estimate. Globally, we can say that the effect of this compound is more pronounced at last days of conservation so that *Ethyl acetate(o)* can be seen as spoilage indicator. *2-Methyl-1-propanol(q)* is glo-

bally characterized by a decreasing slope with significant estimates for lower(with practically no variation) and first part of middle quantiles. The significant OLS estimate underestimates lower quantile regression but highly overestimates upper quantile regression with an underestimate-overestimate in the middle. Compound *1-Butanol(s)* effect's plot shows us that this compound's significant influence is only for lower quantiles and we remark a negative effect very close to the non significant OLS estimate. For middle quantiles (except at $\tau = 0.4$) we relatively have no variation for corresponding estimates which are underestimated by OLS. Upper quantiles estimates are highly underestimated by OLS estimate (except at $\tau = 0.7$ and $\tau = 0.75$) and we have positive effects. Considering *3-Methyl-1butanol(ag)*, the corresponding effects are only significant for lower and last upper quantiles. This compound is characterized by a decreasing increasing trend around the non significant OLS estimate with little difference between extreme (lower and upper) estimates. This compound's effect seems to be stable during whiting conservation. Compound *2-Hexanone(am)* is characterized by negative significant effects on freshness index conditional distribution with a global decreasing trend. The significant OLS estimate underestimates lower and first part of middle quantiles effects but overestimates last middle and upper quantiles effects. Last middle part effects seem to be very similar to last upper effects with a decreasing effect between the two parts. Finally, *Nonanal(bh)* quantile regressions and OLS estimates are significantly different from zero ; this compound also seems to have stable positive effects around the OLS estimate. Last upper quantile regressions seem to be very close to lower quantile regressions.

	OLS	Quantiles				
variables		0.1	0.25	0.5	0.75	0.9
(Intercept)	9.638***	2.445***	2.757**	7.475***	5.988**	10.611****
	(1.603)	(0.453)	(0.791)	(1.667)	(1.772)	(1.100)
Ethanol	3.117*	2.983****	2.787***	1.762	1.564	0.800
	(1.229)	(0.347)	(0.607)	(1.278)	(1.359)	(0.844)
Ethyl	2.581+	1.926***	2.007**	3.159*	4.009*	1.761+
acetate	(1.372)	(0.388)	(0.677)	(1.426)	(1.517)	(0.942)
1-Butanol	-1.622	-1.345***	-1.519*	-0.536	1.879	2.048*
	(1.250)	(0.353)	(0.617)	(1.299)	(1.382)	(0.858)
3-Methyl-	1.639	2.656****	2.803***	0.789	2.190+	0.853
1-butanol	(1.164)	(0.329)	(0.575)	(1.210)	(1.287)	(0.799)
2-Hexanone	-2.002*	-1.159***	-0.415	-3.174**	-2.852**	-3.245***
	(0.886)	(0.250)	(0.438)	(0.922)	(0.980)	(0.608)

TABLE 2.5 – Model-Q for quality index with significance levels : 10%+,5%∗,1% ∗ ∗,0.1% ∗ ∗∗,0% ∗ ∗ ∗ ∗

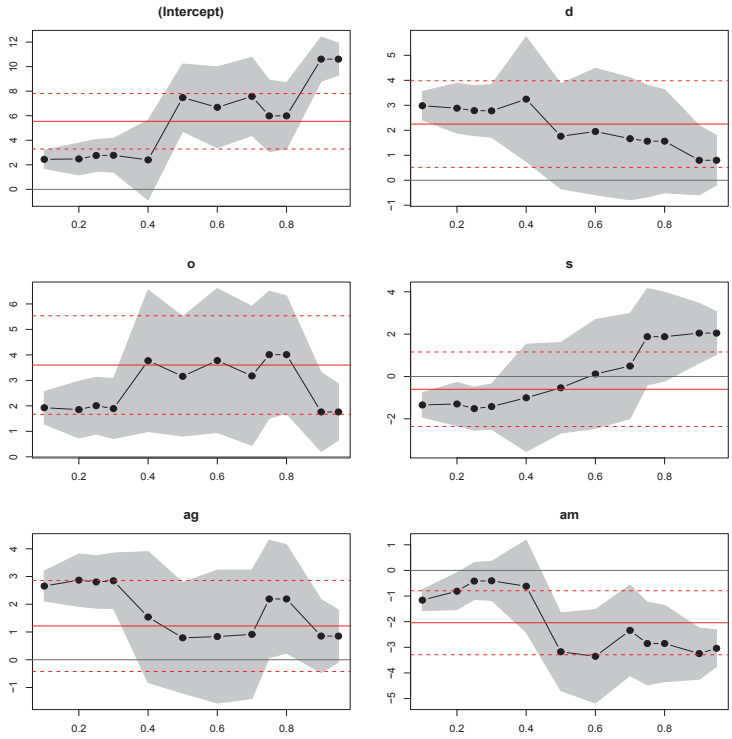

FIGURE 2.4 – Quantile plots for quality index.

Quality index Model-Q

According to Table 2.5 we can see that only three regression coefficients can be considered as significant for OLS model. *Ethanol* and *2-Hexanone* estimates are significant for OLS model at level 5% and *Ethyl acetate* at level 10%. Considering Quantile regression estimates, for $\tau \in \{0.1, 0.25\}$, all regression estimates are significant at least at 5% level except *2-Hexanone*. *Ethyl acetate* is also significant for all quantiles but at 10% level for $\tau = 0.9$. *3-Methyl-1-butanol* slope's is significant at 10% level for $\tau = 0.75$. Finally, we can surprisingly see that *1-Butanol* is positively significant for $\tau = 0.9$ while it is negatively significant for $\tau \in \{0.1, 0.25\}$. The intercept is significant for all considered quantiles.

The comparison between OLS and QR estimates from Model-Q, reveals that OLS overestimates *Ethanol* coefficient(3.117) compared to its QR estimates. Moreover, it is clear that *Ethanol*'s QR estimate for $\tau = 0.1$ (2.983) is more than three times as great as its corresponding for $\tau = 0.9$ (0.800). *Ethyl acetate*'s OLS coefficient (2.581) also overestimates its lower and upper QR slopes ($\tau = 0.1$ (1.926), $\tau = 0.25$ (2.007), $\tau = 0.9$ (1.761)). While for the median and $\tau = 0.75$, the corresponding QR estimates (3.159 and 4.009) are greater than OLS estimates. On the other hand, *2-Hexanone's* OLS estimate (-2.002) overestimates all the upper quantiles and the median QR estimate ; but it underestimates lower quantiles regression coefficients.

For this model, *Ethanol* and *3-Methyl-1butanol* seem to have low effects on quality index for median and upper quantiles. Compounds *Ethyl acetate* and *2-Hexanone* have an effect across all quality index conditional distribution quantiles except for *2-Hexanone* (*Ethyl acetate* is significant at level 10% for $\tau = 0.9$). Finally, the fact that all of the five compounds quantile regression estimates are significant for lower quantiles (except for *2-Hexanone*) means that the considered compounds effects are more important for fresh fishes (lower quantiles) than for fishes in great state of spoilage (upper quantiles). *Ethyl acetate*, *1-Butanol* and *2-Hexanone* effects have been identified also for fishes with great spoilage ($\tau = 0.9$).

On the other hand, quantiles plots in Figure 2.4 for Model-Q leads to the

following comments. First, the intercept term in the first left upper plot can be interpreted as the conditional quantile estimate of the quality index conditional distribution for a fish with zero proportion of *Ethanol(d)*, *Ethyl acetate(o)*, *1-Butanol(s)*, *3-Methyl-1butanol(ag)* and *2-Hexanone(am)*. More specifically, compound *Ethanol(d)* is characterized by a general decreasing trend with significant estimates for lower and first part of middle quantiles. Significant OLS estimate underestimates effects for those quantiles, but overestimates last middle and upper quantiles effects with a more pronounced effect for upper quantiles. Compound *Ethyl acetate(o)* seems to have stable significant positive effects on quality index conditional distribution with similar effects for extreme upper and lower quantiles. Little fluctuations around OLS estimate are remarked for middle and first part of upper quantiles. For quality index model, compound *1-Butanol(s)* is characterized by increasing slopes across quantiles with the particularity of opposite significant effects for lower and upper quantiles. Although effects at lower and upper quantiles seem to have very little variation, more pronounced variations are remarked for middle quantiles effects. Analyzing *3-Methyl-1butanol(ag)* slopes shows that only lower quantiles and a little region in upper quantiles seem to have significant positive effects. This compound has decreasing-increasing effects with extreme upper quantiles effects similar to middle quantiles effects. The particularity for this compound is that effects seem to remain constant on the considered quantiles set. Finally, compound *2-Hexanone(am)* has an increasing-decreasing trend with non significant effects for some lower and middle quantiles. Negative effects variations for this compound are more pronounced after quantile $\tau = 0.4$; OLS significant effect overestimates the median ($\tau = 0.5$) and remaining effects seem to have little variation with a jump at $\tau = 0.7$.

The relation between spoilage and regression estimates

Since our aim is to identify volatile compounds which are related to indicators of spoilage, using penalized regression and quantile regression, a natural question is : what is the relationship between regression estimates (slopes) and spoilage ? The most identified compounds can be considered as spoilage markers and OLS estimates cannot tell us how slopes behave as spoilage increases during fish conservation. Quantile regression gives answers to the previous points as shown in Table

	Regression coefficient decreases as spoilage increases	Regression coefficient increases as spoilage increases	Regression coefficient shows no definite pattern as spoilage increases
Models			
Freshness index			
Model-F	*Ethanol(d)* *Dimethyl sulfide(g)* *2-Methyl-1-propanol(q)* *2-Hexanone(am)*	*Ethyl acetate(o)* *1-Butanol(s)*	*3-Methyl-1-butanol(ag)* *Nonanal(bh)*
Quality index			
Model-Q	*Ethanol(d)*	*Ethyl acetate(o)* *1-Butanol(s)*	*3-Methyl-1-butanol(ag)* *2-Hexanone(am)*

TABLE 2.6 – The relationship between selected volatile compounds and spoilage as shown by the quantile regressions.

2.6.

2.5 Conclusions

According to selection results we can say that some interesting volatile compounds like *Ethanol(d)*, *Dimethyl sulfide(g)*, *Ethyl acetate(o)*, *2-Methyl-1-propanol(q)*, *1-Butanol(s)*, *3-Methyl-1-butanol(ag)*, *2-Hexanone(am)* and *Nonanal(bh)* have been identified as relevant to explain spoilage indicators of freshness and/or quality index.

Quantile regression approach clearly shows that some compounds are very significant while their corresponding OLS estimates are not significant. Results in Table 2.6 show that some compounds have a great impact on lower quantiles of freshness index but with less impact as spoilage increases, those compounds include *Ethanol(d)*, *Dimethyl sulfide(g)*, *2-Methyl-1-propanol(q)* and *2-Hexanone(am)*. The same remark takes place for compound *Ethanol(d)* when considering quality index

model. For the two selected models, *Ethyl acetate(o)* and *1-Butanol(s)* effects increase as spoilage increases. So, their QR coefficients are greater for upper quantiles of freshness and quality indices conditional distributions. While, the coefficients of *3-Methyl-1-butanol(ag)*,*2-Hexanone(am)* and *Nonanal(bh)* do not present a specific pattern for different quantiles.

Globally, quantile regression provides the trend of the regression coefficients across quantiles ; this fact is not possible with classical linear models. In our knowledge it is the first time that approaches of variable selection and QR were adopted in fishery. Our results show that the effect of volatile compounds on fish freshness and quality indices can be better explained by estimating quantile regressions. Quantile regression also provides more information about estimated effects level and the direction of change. Those differences may be explained by differences in freshness and quality indices during fish conservation. Even though variations in the quantity of volatile compounds across different freshness and/or quality index can be intuitive, quantile regression allows us to confirm these facts.

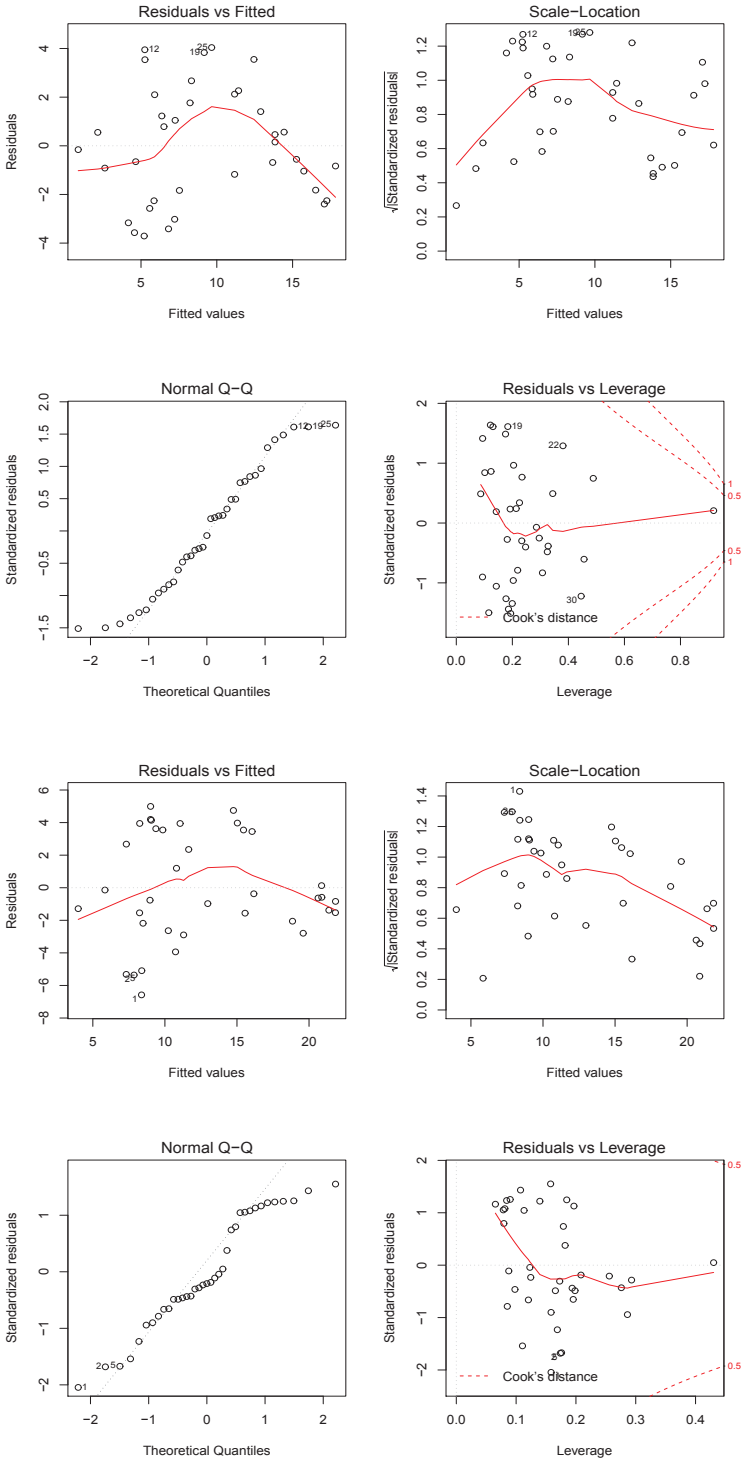

FIGURE 2.5 – Linear model (OLS) validation for freshness (four top plots) and quality (four bottom plots) indexes.

We also give regression diagnostic for ordinary least squares in figure 2.5. Considering the first four plots, the plot in the upper left shows the residual errors plotted versus their fitted values. The residuals seems to be randomly distributed around the horizontal line representing a residual error of zero ; that is, we have not a distinct trend in the distribution of observations. The plot in the lower left is a standard Q-Q plot (Henry line), which suggests that the residual errors are normally distributed for freshness (which is not the case for quality) index with small bias for extreme values. The scale-location plot in the upper right which is a sort of a square root of relative error, shows the square root of the standardized residuals as a function of the fitted values. Again, it seems that there is no obvious trend in this plot.

Finally, the plot in the lower right shows each points leverage, which is a measure of its importance in determining the regression result. Red broken lines are contour lines for the Cook's distance, which is another measure of the importance of each observation to the regression. Smaller distances mean that removing the observation has little affect on the regression results. Distances larger than 1 are suspicious and suggest the presence of a possible outlier or a poor model. According to those plots it seems that for both freshness and quality models, observations are in the good range.

Troisième partie

Bootstrap and Randomization Approaches

Chapitre 3

Stability Selection and Randomization in L_1 Quantile Regression

Statistical models in linear regression generally focus on estimation and interpretation of conditional mean effects. However, in some situations considering mean effects could be not appropriate when for example we have great variations in response variable percentiles or when we have outliers. We here propose the Stability Selection method for variables selection in high dimension penalized linear Quantile Regression. This approach combines subsampling and variable selection algorithms adapted to the case of high dimension. Particularly, we apply Stability Selection with Lasso and Randomized Lasso Quantile Regression. Finally, the proposed method is compared with its competitors on simulated and real data sets.

3.1 Introduction

Meinshausen and Bühlmann[108] advocate that subsampling can be used for Stability Selection in penalized linear regression models to determine the amount of regularization such that a certain family type I error rate in multiple testing can be conservatively controlled for finite sample size. Particularly for complex and high dimensional problems, a finite sample control is much more valuable than an asymptotic statement with the number of observations tending to ∞. Moreover, the previous authors also prove that subsampling in conjunction with L_1-penalized estimation requires much weaker assumptions on the design matrix for asymptotically consistent variable selection than what is needed for the non-subsampled L_1-penalty scheme. Furthermore they show that additional improvements can be achieved by randomizing not only via subsampling but also in the selection process for the variables. Recently a variant approach called Complementary Pairs Stability Selection (CPSS) has been also proposed by Shah and Samworth[131]. Beinrucker et al.[15] also propose a simple extension of the original stability feature selection approach used in Meinshausen and Bühlmann[108]. We can mention here that variable selection approaches based on bootstrap, Random Lasso[145] and BoLasso[9] have been proposed in linear regression case.

In the lines below, we will focus on the adaptation of Stability Selection (Meinshausen and Bühlmann[108]) and bootstrap based approaches to Quantile Regression. As a motivation of this work we have a dataset on volatile compounds previously used in Duflos et al.[38]. The classical approach based on penalized mean regression is not adapted for this setting since the response variable presents many variations. On the other hand, we want to select the set of stable variables among the volatile compounds which represent the predictors.

The second section of this paper is devoted to the adaptation of Stability Selection to Quantile Regression (QR). Moreover, an illustration on a real dataset is presented and the selection of the tuning parameter based on Belloni and Chernozhukov[16] idea is discussed in the same section. Finally, last section is devoted to numerical experiments including simulations and real data set applications.

3.2 Linear Quantile Regression Stability Selection

3.2.1 Variable selection in Quantile Regression

We consider a size n i.i.d sample $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ from some unknown population, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Linear quantile regression solves the following optimization problem for $0 < \tau < 1$:

The following penalized problem is considered :

$$(\hat{\beta}_0(\tau), \hat{\beta}(\tau)) = \text{argmin}_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^T\beta) + \lambda P(\beta) \right\}, \tag{3.1}$$

where $P(.)$ is the penalty function and the tuning parameter $\lambda > 0$ controls the sparsity of the model. As a survey on frequently used penalty functions in the field of quantile regression we can cite the excellent references of Zou and Yuan[164], Wu and Liu[148] and Slawski[134]. Since the idea of our proposed methods are based on Lasso penalty, in all the lines below, we will only focus on the following problem defined by :

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^T\beta) + \lambda \parallel \beta \parallel_1 \right\}. \tag{3.2}$$

All of the methods in this paper are based on the previous formulation except the Randomized Lasso (Meinshausen and Bühlmann[108]) for quantile regression with weakness $\alpha \in (0, 1]$ which takes the following form :

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^T\beta) + \lambda \sum_{k=1}^{p} \frac{\mid \beta_k \mid}{W_k} \right\}, \tag{3.3}$$

where W_k are iid random variables in $[\alpha, 1]$ for $k = 1, ...p$.

3.2.2 Stability Selection and pointwise control

This part is a slightly modified part of Meinshausen and Bühlmann[108] approach. Since stability paths are derived from the concept of regularization paths, we recall that for each quantile τ, $0 < \tau < 1$ a regularization path is given by the coefficient value of each variable over all regularization parameters

$$\{\hat{\beta}_k^\lambda(\tau); \lambda \in \Lambda, k = 1, ..., p\}.$$

Stability paths are the probability for each variable to be selected when randomly resampling from the data. For any given regularization parameter $\lambda \in \Lambda$, the selected set $\hat{S}_\tau^\lambda$ is implicitly a function of the samples $I = \{1, ..., n\}$.

Definition 3.2.1 *(selection probabilities)*

Let I be a random subsample of $\{1, ..., n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement. For every set $K \subseteq \{1, ..., p\}$, the probability of being in the selected set $\hat{S}_\tau^\lambda(I)$ is

$$\hat{\Pi}_K^\lambda(\tau) = P^*\{K \subseteq \hat{S}_\tau^\lambda(I)\}. \tag{3.4}$$

For every variable $k = 1, ..., p$, the stability path is given by the selection probabilities $\hat{\Pi}_K^\lambda(\tau)$, $\lambda \in \Lambda$.

In the remainder of the paper, we look at the selection probabilities of individual variables. Generally, for each quantile of interest variable selection is concerned by the choice of one element in the set of models

$$\{\hat{S}_\tau^\lambda; \lambda \in \Lambda\}, \tag{3.5}$$

where Λ is again the set of regularization parameters considered, which can be either continuous or discrete. There are typically two problems : first, the correct model S_τ might not be a member of set (4.12). Second, even if it is a member it

is typically very difficult for high dimensional data to determine the right amount of regularization λ to select exactly S_τ, or at least a close approximation. With Stability Selection, we do not simply select one model in the list (4.12), instead the data are perturbed (e.g by subsampling) many times and we choose all variables that occur in a large fraction of the resulting selection sets.

Definition 3.2.2 *(stable variables)*

For a cut-off π_{thr} with $0 < \pi_{thr} < 1$ and a set of regularization parameters Λ, the set of stable variables for a quantile τ is defined as

$$\hat{S}_\tau^{stable} = \{k : max_{\lambda \in \Lambda}(\hat{\Pi}_k^\lambda(\tau)) \geq \pi_{thr}\}. \tag{3.6}$$

We keep variables with a high selection probability and discard those with low selection probabilities.

3.2.3 Illustration on PAC dataset

In Figure3.1, we illustrate the advantage of using Stability Selection with or without Randomization. We use PAC data available under R software about GC-retention indices of polycyclic aromatic compounds(y) which have been modeled by molecular descriptors(X). The data set contains n=209 observations and p=467 predictors. We first take the 50 variables with the highest marginal correlation with log(y) and randomly select five predictors. These five predictors are kept unpermuted and the remaining 462 are permuted across the samples, using the same permutation that keeps the dependence structure between the permuted observations intact. The left plot, corresponding to L_1 penalized median regression shows that it is very difficult to isolate the five unpermuted variables paths with noise variables paths. For Stability Selection, a threshold of $\pi_{thr} = 0.6$ includes all of the five unpermuted variables with some noise variables. When using Stability Selection with randomization parameter $\alpha = 0.1$, we can see that with the same threshold $\pi_{thr} = 0.6$ all of the five unpermuted variables are selected without any

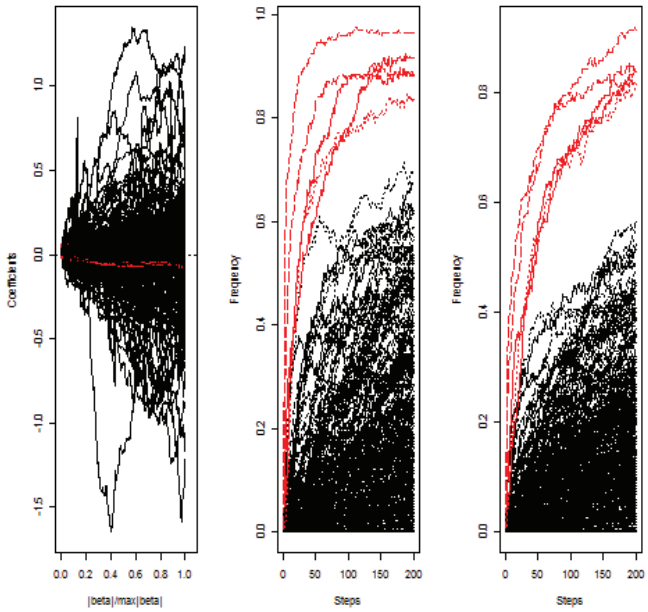

FIGURE 3.1 – From left to right : comparison between L_1 median regression regularization paths, QR Stability Selection without randomization and QR Randomized Stability Selection with $\alpha = 0.1$ on PAC dataset for log(y).

noise variable.

3.2.4 Tuning parameter selection

Among many alternatives on the choice of the tuning parameter we can cite (see Li and Zhu[103]) the Schwarz Information Criterion (SIC, Schwarz[129] ; Koenker, Ng, and Portnoy[94]) and the generalized approximate cross-validation criterion (Yuan[152]). The SIC is defined by

$$SIC(\lambda) = log(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - f(\mathbf{x_i}))) + \frac{log(n)}{2n}df,$$

where df is a measure of the effective dimensionality of the fitted model. In our case, $f(\mathbf{x_i}) = \beta_0 + \mathbf{x}_i^T\beta$. However, recently Koenker [87] claims that using the SIC optimization method often produced insufficient shrinkage and the optimiza-

tion process was quite slow. He also claims that when simulating realizations of the random vector $S_n = \sum_{i=1}^{n}(\tau - I(U_i \leq \tau))\mathbf{x_i}$, one can assert that the event $\|S_n\|_\infty \leq \lambda$ should hold with high probability, provided of course that λ is chosen sensibly so that $\hat{\beta}$ is close to the true parameter $\beta(\tau)$. Following this idea, Belloni and Chernozukov[16] suggested choosing $\hat{\lambda}$ as a $(1-\alpha)$ quantile of the simulated distribution of $\|S_n\|_\infty$, or perhaps a constant multiple of such a quantile for some $c \in (1, 2]$. This extremely simple approach was used in our simulations for $\alpha = 0.1$ and $c = 1$.

3.3 Other approaches

In this section we introduce other bootstrap based approaches which have been already studied in the case of classical linear regression models. We here propose their quantile regression versions.

3.3.1 Random Lasso quantile regression

We recall that in the case of linear regression models, the original Random Lasso[145] method consists of two major steps. More information and details about each step of this algorithm can be found in Wang et al.[145].

ALGORITHM : "GENERATE" AND "SELECT".

Step 1. Generating importance measures for all coefficients.

1a. Draw B bootstrap samples with size n by sampling with replacement from the original training data set.

1b. For the b_1th bootstrap sample, $b_1 \in \{1, ..., B\}$, randomly select q_1 candidate variables, and apply quantile regression Lasso to obtain estimators $\hat{\beta}_j^{(b_1)}$ for β_j, j=1,...,p. Estimators are zero for coefficients of those unselected variables, either

outside the subset of q_1 variables, or excluded by quantile regression Lasso.

1c. Compute the importance measure of x_j by $I_j =\mid B^{-1}\sum_{b_1=1}^{B}\hat{\beta}_j^{(b_1)} \mid$.

Step 2. Selecting variables.

2a. Draw another set of B bootstrap samples with size n by sampling with replacement from the original training data set.

2b. For the b_2th bootstrap sample, $b_2 \in \{1, ..., B\}$, randomly select q_2 candidate variables with selection probability of x_j proportional to its importance I_j obtained in Step 1c, and apply Lasso (or adaptive Lasso) to obtain estimators $\hat{\beta}_j^{(b_2)}$ for β_j, j=1,...,p. Estimators are zero for coefficients of those unselected variables, either outside the subset of q_2 variables, or excluded by quantile regression Lasso.

2c. Compute the final estimator $\hat{\beta}_j$ of β_j by $\hat{\beta}_j = B^{-1}\sum_{b_2=1}^{B}\hat{\beta}_j^{(b_2)}$.

3.3.2 BoLasso quantile regression

As related in Meinhaussen and Bühlam[108], Bach[9] has proposed the 'BoLasso'(for Bootstrapped enhanced Lasso) and shown that using a finite number of subsamples of the original Lasso procedure and applying basically stability selection with $\prod_{thr} = 1$ yield consistent variables selection under the condition that the penalty parameter vanishes faster than typically assumed, at rate $n^{-1/2}$, and that the model dimension p is fixed. We report the BoLasso algorithm as in Bach[9], adapted to quantile regression.

ALGORITHM : "Quantile Regression BoLasso"

Input : data $(\overline{X}, \overline{Y}) \in \mathbb{R}^{n\times(p+1)}$
number of bootstrap replicates m
regularization parameter μ

for $k = 1$ **to** m **do**

Generate bootstrap samples $(\overline{X}^k, \overline{Y}^k) \in \mathbb{R}^{n\times(p+1)}$

Compute Quantile regression Lasso estimate $\hat{w}^k$ from $(\overline{X}^k, \overline{Y}^k)$

Compute support $J_k = \{j, \hat{w}_j^k \neq 0\}$
end for
Compute $J = \bigcap_{k=1}^m J_k$
Compute $\hat{w}_J$ from $(\overline{X}_J, \overline{Y})$

3.4 Numerical results

3.4.1 Simulations settings

For our simulations, we consider the following model $Y = X\beta(0.5)+\epsilon$, where Y is the n dimensional response vector, X is the $n \times p$ predictors matrix, $\beta(\tau)$ is the true p dimensional parameter vector and ϵ is the n dimensional vector of errors. In the lines below, we consider $p = 200$-dimensional predictor variables follow an $N(0, \Sigma)$ distribution, where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho \in \{0, 0.5, 0.75, 0.9\}$. The sample size is fixed to $n = 100$ and the Signal to Noise Ratio $SNR = Var(X\beta(\tau))/Var(\epsilon)$ considered takes values in {0.5,2,4}. The error vector $\epsilon \sim N(0, 1)$ and $\beta(\tau)$ vector has s nonzero components chosen as uniforms on $[0, 1]$. The s value considered here is $s \in \{4, 8, 12, 20\}$. For the random Lasso we choose $q_1 = q_2 = (2, 4, 6, 8, 10)$ for $s \in \{4, 8\}$ and $q_1 = q_2 = (10, 12, 14, 16, 20)$ for $s \in \{12, 20\}$.

Simulations are performed 100 times and median of False positive, False negative and the mean probability to select $0.1s$ and $0.4s$ correct variables without any noise variable are given on Figures 3.2,3.3,3.6,3.7 and Figures 3.4,3.5,3.8,3.9.
Each plot gives the performances of the following five methods "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

Results for $n = 50$ seems to be similar. For Stability Selection based approaches, the threshold is fixed to $\pi_{thr} = 0.6$ and $\alpha = 0.5$ for Randomized Stability Selection. We use fixed value of the tuning parameter (pointwise control) as previously advocated. Nevertheless we can use R software "lpRegPath" package which generates the entire solution path as a function of the tuning parameter.

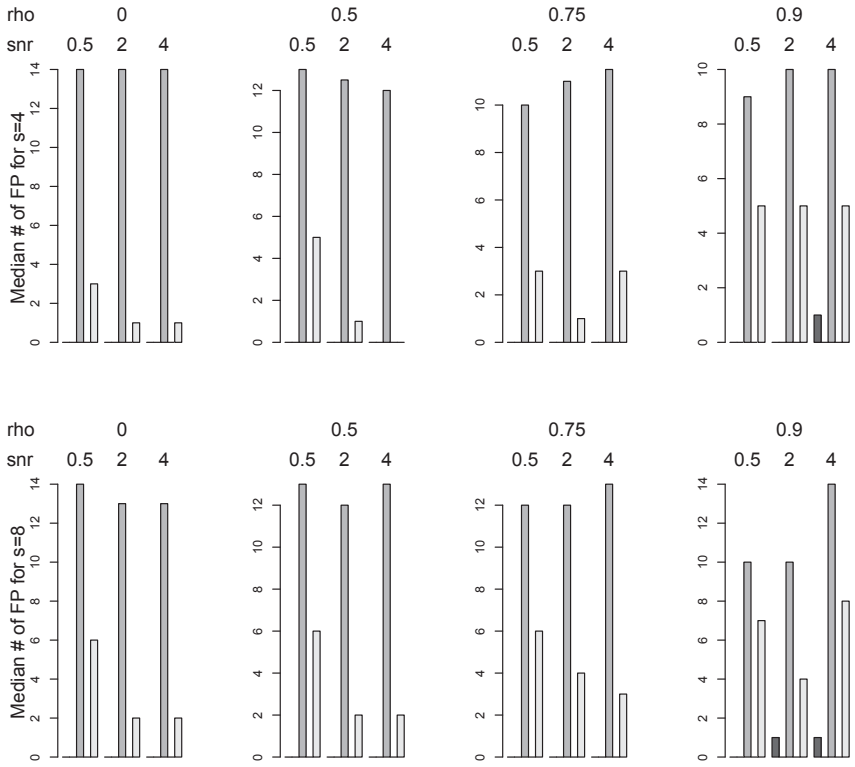

FIGURE 3.2 – Median number of False Positive selection among 100 replications for s=4(top row) and s=8(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

For $s = 4$ and $\rho = 0$, stability selection based methods and QR BoLasso do not introduce FP ($SNR \in \{0.5, 2, 4\}$). QR Lasso FP selection is around 14 FP which represents the median about 100 bootstrap ($SNR \in \{0.5, 2, 4\}$). QR Random Lasso seems to introduce more FP variables for $SNR = 0.5$ than for $SNR \in \{2, 4\}$. In terms of FN, QR Randomized Stability Selection and QR BoLasso seems to delete more true regression coefficients than other methods with decreasing number of median FN when SNR increases. In terms of P($0.1s$ correct),

QR Randomized Stability Selection and QR BoLasso give very good result (probability=1), QR Stability Selection corresponding probability increases when the SNR increases (0.73,0.77,0.78). Since QR Lasso introduces too many variables, the corresponding probability is zero. QR Random Lasso also have increasing probability with increasing SNR (0.02,0.27,0.41). For P($0.4s$ correct) and $SNR = 0.5$, QR Stability Selection gives great probability(0.68) followed by QR Randomized Stability Selection (probability=0.46) and QR BoLasso (0.18). QR Lasso and QR Random Lasso give zero probability. For $SNR \in \{2, 4\}$, QR Randomized Lasso and QR BoLasso give great similar probability (0.99 and 1 respectively) and we have probability values of 0.01 and 0.1 for the QR Random Lasso. QR Stability selection gives probability of 0.89 and 0.86. The previous comments and for other values of ρ can be seen on Figures 3.2,3.3 and Figures 3.4,3.5.

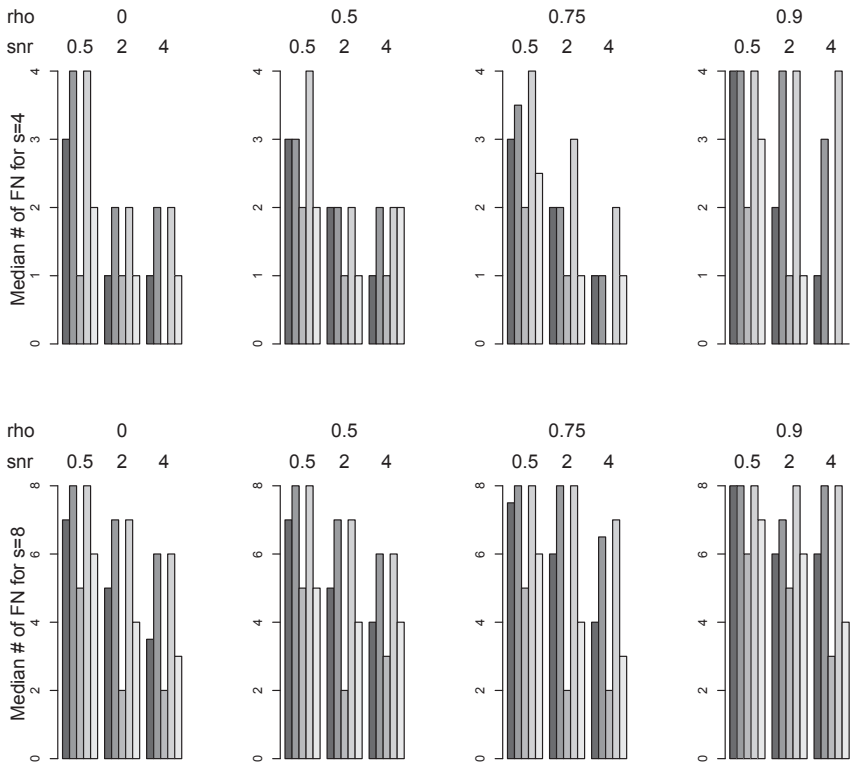

FIGURE 3.3 – Median number of False Negative selection among 100 replications for s=4(top row) and s=8(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

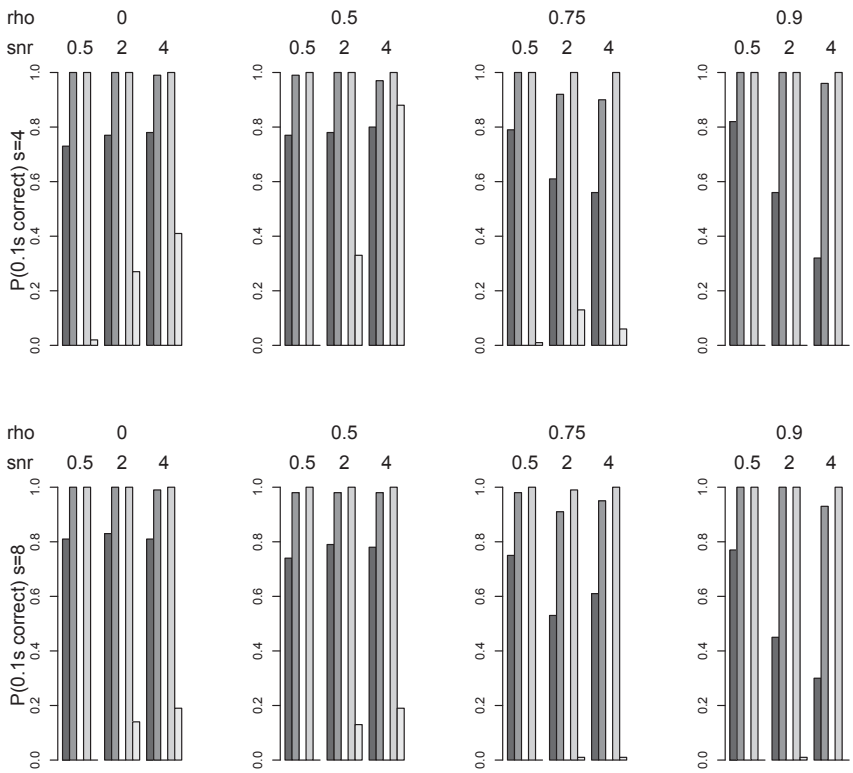

FIGURE 3.4 – Probability of selection of 0.1s of relevant variables without selection any noise variables among 100 replications for s=4(top row) and s=8(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

When $s = 8$ and $\rho \in \{0, 0.5\}$ we have the same remarks as previous (s=4) for the median number of FP and FN. For P($0.1s$ correct), QR Randomized Stability and QR BoLasso have higher performances followed by QR Stability Selection and QR Random Lasso. For P($0.4s$ correct), QR Stability Selection has the highest performances due to the fact that it has the best trade off between FP and FN, followed by Randomized Stability Selection. We remark that QR BoLasso has the higher performance for P($0.1s$ correct) when $\rho = 0.75$ followed by QR Randomized Stability Selection and QR Stability Selection without Randomization. QR Random Lasso has very low selection probability.
For P($0.4s$ correct) and SNR=0.5, all methods fail to select 40% of correct variables without introducing any noise variables, due to the fact that QR Lasso and QR

Random Lasso introduce noise variables, and other methods delete some correct variables. Results for other values of SNR and for $\rho = 0.9$ can be seen on Figures 3.2,3.3 and Figures 3.4,3.5.

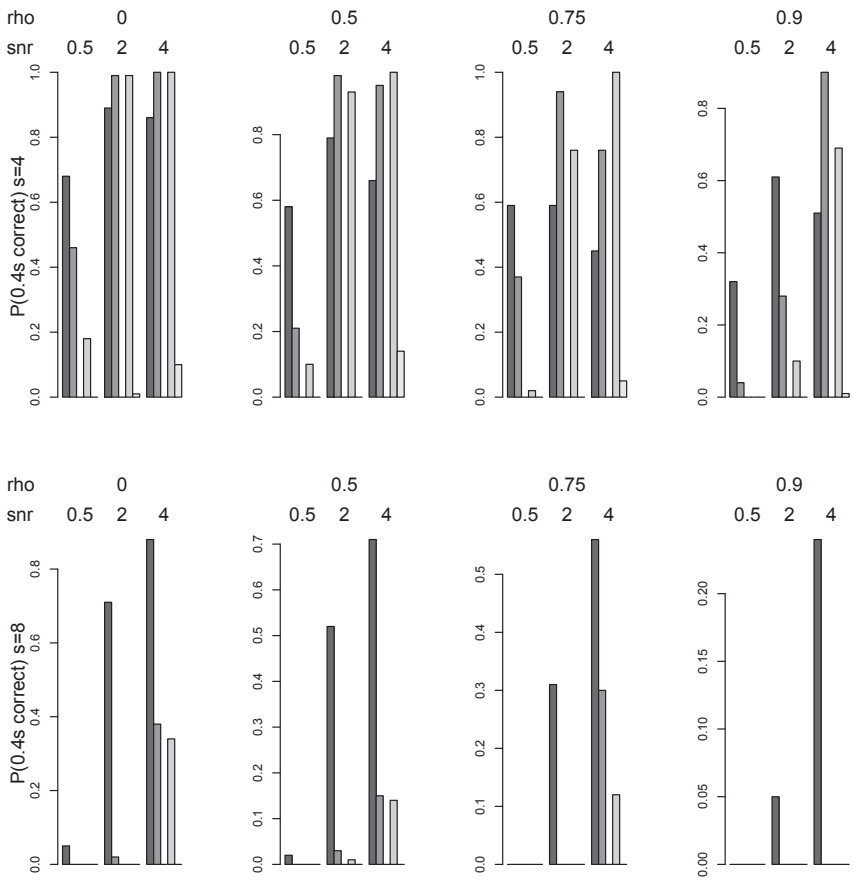

FIGURE 3.5 – Probability of selection of 0.4s of relevant variables without selection any noise variables among 100 replications for s=4(top row) and s=8(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

In the case that $s = 12$, for $\rho = 0$ we have the same remark for FP and FN for the case $s = 8$ but for P($0.1s$ correct), Stability selection has higher performances with increasing selection probability when SNR increases. Randomized Lasso and BoLasso also have this increasing trend when SNR increases. For P($0.4s$ correct) all of methods have zero probability for $SNR = 0.5$ and for $SNR \in \{2, 4\}$, Stability selection has selection probability of respectively 0.24 and 0.57. The case $SNR = 4$, leads to very low median selection of 0.02 for Randomized Stability selection. When $\rho = 0.5$, only Stability selection, Randomized stability selection and BoLasso have nonzero values for P($0.1s$ correct) with best performances obtained for Randomized Lasso except for $SNR = 0.5$ where stability selection is the winner. For P($0.4s$ correct) all methods have zero selection probability except Stability selection method which has selection probability of 0.18 and 0.42 for respectively $SNR = 2$ and 4. When considering FN indicator for $\rho = 0.75$, Random Lasso seems to have best performances. For P($0.1s$ correct), good performances are globally achieved for Random Lasso except for $SNR = 4$ where Randomized version is the winner. For P($0.4s$ correct) only Stability selection has nonzero median selection probability for $SNR \in \{2, 4\}$. When $\rho = 0.9$, Stability selection seems to include noise variables for $SNR = 4$ and has highest values for P($0.1s$ correct). Other methods have zero median selection probability values except for Randomized stability selection ($SNR \in \{2, 4\}$) and BoLasso ($SNR = 4$). For P($0.4s$ correct), we have the same remark as previous.

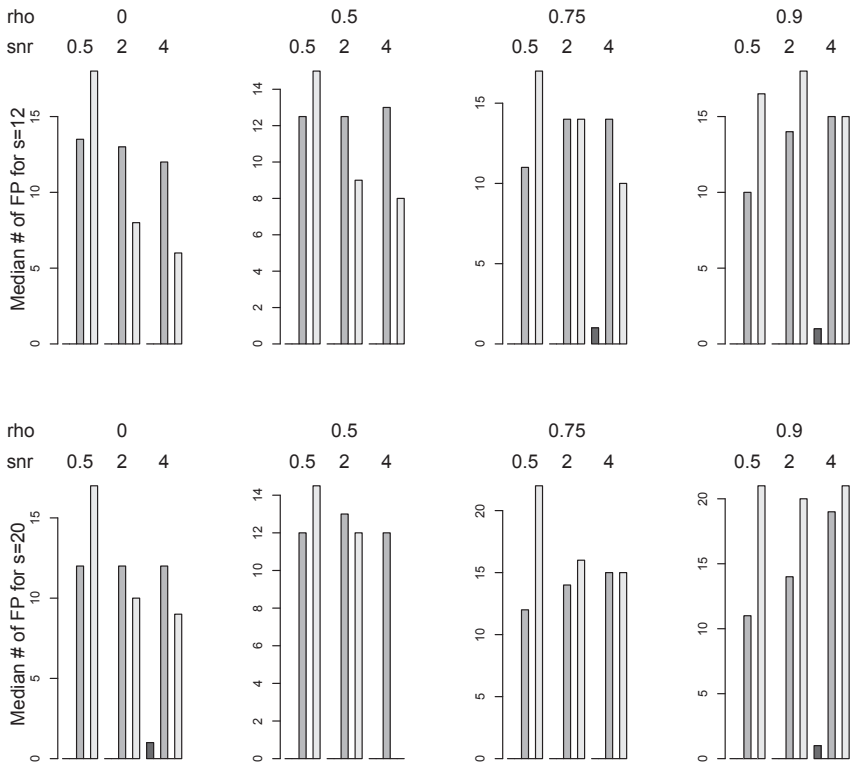

FIGURE 3.6 – Median number of False Positive selection among 100 replications for s=12(top row) and s=20(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

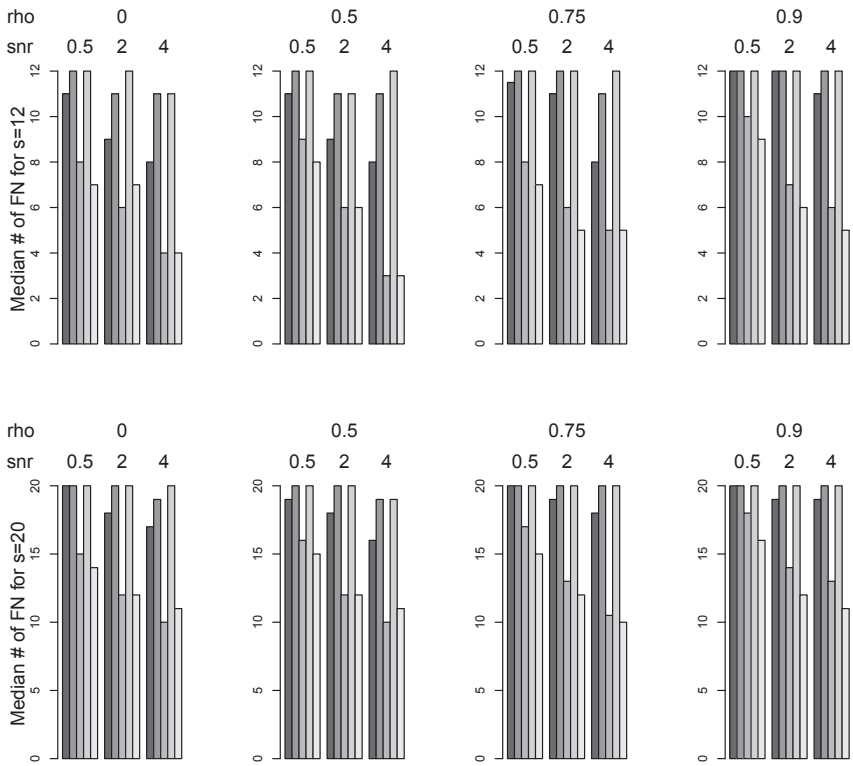

FIGURE 3.7 – Median number of False Negative selection among 100 replications for s=12(top row) and s=20(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

Considering $s = 20$ and $\rho = 0$, for P($0.1s$ correct), only Stability selection has nonzero increasing median values for all values of SNR. Median probability selection values for increasing SNR are respectively 0.07,0.56 and 0.68. For Randomized Stability Selection we have nonzero median probability selection values 0.04 and 0.16 for respectively $SNR = 2$ and $SNR = 4$. For this indicator, BoLasso has respectively 0.01 and 0.03 median values. Finally, for P($0.4s$ correct) all methods give zero median value for all SNR and ρ values. For $\rho = 0.5$, Stability Selection has increasing median selection probability with nonzero value. Randomized Stability selection has increasing median values for this indicator for $SNR \in \{2, 4\}$. BoLasso has a very low median selection probability of 0.04 for $SNR = 4$. Finally, when $\rho = 0.75$, we have the same remarks as previous for Stability selection. Randomized Stability selection has very low median of 0.01 for P($0.1s$ correct) and ($SNR \in \{2, 4\}$). When $\rho = 0.9$, only Stability selection has nonzero value for median of P($0.1s$ correct) with respectively values of 0.01 ($SNR = 5$) and 0.11 ($SNR \in \{2, 4\}$). For $SNR = 4$, Random Lasso seems to introduce noise variables (median $FP = 1$). As a general remark for $s = 20$, Random Lasso has the highest FP but the lowest FN ($\rho \in \{0.75, 0.9\}$).

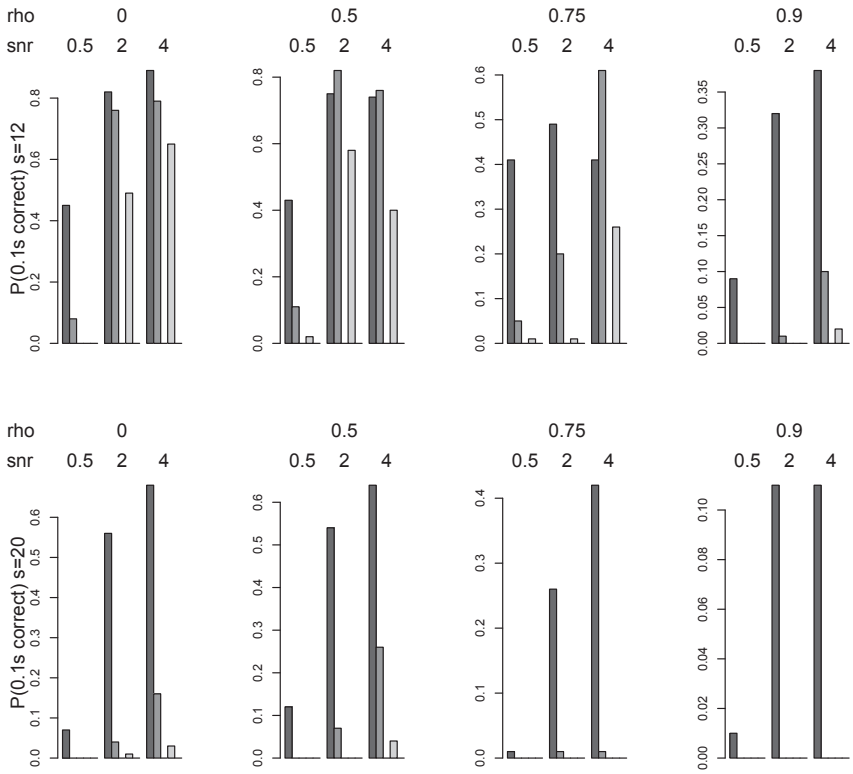

FIGURE 3.8 – Probability of selection of 0.1s of relevant variables without selection any noise variables among 100 replications for s=12(top row) and s=20(bottom row). For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

3.4.2 Real data application

We consider the data set on volatile compounds previously used in Duflos et al.[38]. The sample considered consists of $n = 37$ observations and $p = 49$ predictors(X) used to model Freshness index and Quality scores (y). A direct use of methods on the full data gives results only for L_1 median regression and QR BoLasso. For Stability Selection and bootstrap based methods, we have no results due to the fact that subsampling of size $n/2$ leads in some cases to singular design sub matrixes and we cannot compute the tuning parameter $\lambda = c\Lambda(1-\alpha|X)$ since the matrix of predictors X is very sparse with many zero entries. On another hand, QR BoLasso with 100 bootstrap selects no variables for many values of

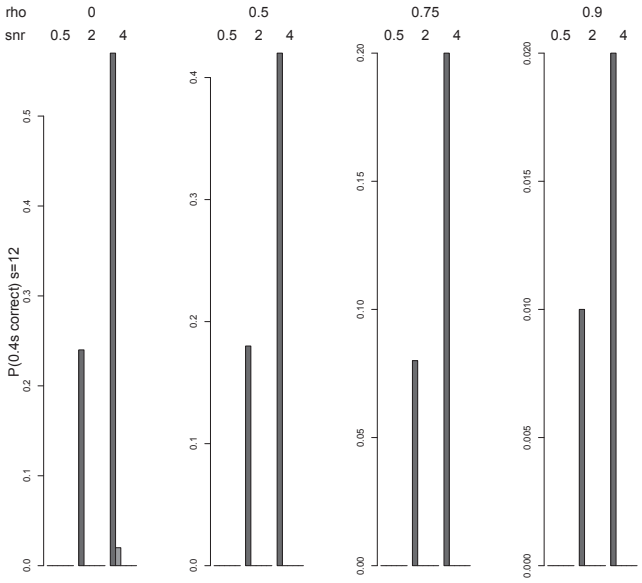

FIGURE 3.9 – Probability of selection of 0.4s of relevant variables without selection any noise variables among 100 replications for s=12. For each SNR value we have from left to right "QR Stability Selection","QR Randomized Stability Selection", "QR Lasso", "QR BoLasso" and "QR Random Lasso".

λ, so we consider QR Soft BoLasso (QR SBoLasso) which selects at least 60% of variables which are selected for all bootstrap samples. In the lines below we only consider compounds with at least $n/2$ nonzero observations which lead to a predictors matrix with sizes $n = 37$ and $p = 37$. Selection results are summarized below where for the tuning parameter λ, $c = 2$ and $\alpha = 0.1$. For stability selection methods, the threshold is fixed to $\pi_{thr} = 0.6$. According to selection results in Table 3.1, we see that for $\lambda = 25.253$ no variable is selected by all of the methods. Selection results for $\lambda/5$ and $\lambda/10$ are given in Table 3.1 with selected volatiles names given in Table 3.2. As expected, subsampling improves important variables selection and additional randomization prevents against noise variables selection. The selected compounds by Stability Selection approach also have been found by Duflos et al.[38] to be related to fish spoilage during storage. This important compounds are Ethanol(d), Ethyl acetate(o), 3-Methyl butanal(r) and 3-Methyl-1-butanol(ag) where more stable compounds seem to be Ethanol(d) and 3-Methyl-1-butanol(ag).

Methods	Selected Compounds	
λ=25.253	Freshness index	Quality scores
All methods	-	-
$\lambda/5$		
QR Lasso	a,d,l,o,q,r,aa,ag	b,d,r,aa,ag,ah
QR SBoLasso	d,r,ag	d,ag,ah
QR Stability Selection	d,r,ag	d,ag
QR Randomized Stability Selection $\alpha = 0.2$	-	-
QR Randomized Stability Selection $\alpha = 0.5$	d,ag	ag
QR Randomized Stability Selection $\alpha = 0.8$	d,r,ag	d,ag
$\lambda/10$		
QR Lasso	b,d,j,l,o,r,aa,ag,aq,ar,bh	b,d,l,aa,ag,ah,aq,bh
QR SBoLasso	d,o,r,ag,bh	d,r,ag
QR Stability Selection	d,o,r,ag	d,ag
QR Randomized Stability Selection $\alpha = 0.2$	d	d,ag
QR Randomized Stability Selection $\alpha = 0.5$	d,r,ag	d,ag
QR Randomized Stability Selection $\alpha = 0.8$	d,r,ag	d,ag

TABLE 3.1 – Selected volatiles for penalized median regression.

Compounds	Names	Compounds	Names
a	Acetaldehyde	r	3-Methyl butanal
b	Methanethiol	aa	2,3-Pentanedione
d	Ethanol	ag	3-Methyl-1-butanol
j	2,3-Butanedione	ah	2-Methyl-1-butanol
l	2-Butanone	aq	1-Hexanol
o	Ethyl acetate	ar	3-Heptanone
q	2-Methyl-1-propanol	bh	Nonanal

TABLE 3.2 – Selected volatiles names.

Quatrième partie

Penalized Quantile Regression : grouping effect and oracle properties

Chapitre 4

Grouping effect and oracle properties with correlated variables

We here consider the problem of variable selection in his more recent development. The advantage of using quantile regression is well known and has been used in many studies (Yuan and Yin[154], Kato[79], Burgette et al.[24]). Among the family of robust regression, we have the LAD estimator and the Huber estimator. These two regressions can be viewed as particular cases of quantile regression (median regression) or very close to this regression in some situations. The most previously used penalties are Lasso, Adaptive Lasso and SCAD. In this contribution we consider Berhu (Owen[114]), Elastic net (Zou and Hastie[160], Slawski[134]) and adaptive Elastic net (Zou and Zhang[165]) penalties in quantile regression. Their theoretical properties including the grouping effect and oracle ones are studied in details. Illustrations on simulated and real data sets are also presented.

4.1 Introduction

In penalized quantile regression, Wu and Liu[148] proposed SCAD and adaptive Lasso penalties. They showed the oracle properties in the case of both independent

and non independent errors. They also mention the fact that for the SCAD penalty computations are based on the Difference Convex Algorithm (DCA). Oracle properties with adaptive lasso penalties are established by Zou[159] with quadratic loss and Wang et al.[142] with LAD loss. Oracle properties with quadratic loss have also been showed by Zou and Zhang[165] for the adaptive Elastic net when the number of parameters diverges.

The SCAD penalty has been previously used by Fan and Li[48] in their general work including robust and non robust loss functions. Particularly they used Huber [71] loss function with SCAD penalty. Moreover, Fan and Peng[54] study oracle properties of SCAD penalty when the number of parameters diverges but are not large compared with the sample size ; finally Kim and al.[80] study the oracle properties of SCAD penalty estimator on high dimensions ($p >> n$). Since the Lasso is very popular we here only interested on its adaptive version previously used by Zou[159] for quadratic loss function and by Wu and Liu[148] for quantile regression. We just advocate that Belloni and Chernozhukov[16] recently gave very important contribution in the field of L_1 penalized quantile regression, both theoretically and computationally. Another very interesting penalty is the Elastic net [Zou and Hastie[160]] which is useful for high dimensional data sets with highly correlated predictors. As an extension of this method, Zou and Zhang[165] have proposed the adaptive Elastic net with a diverging number of parameters, where the authors highlight oracle properties of their approach for penalized least squares.

More general approach has been considered by Slawsky[134] for the structured Elastic net in the field of quantile regression and support vector classification. This author gave a very important contribution about computational aspect and regularity paths. This regularization path has been well developed by Rosset and Zou[125] for SVM with L_1 penalty and Elastic net for SVM has been proposed by Wang et al.[144] and Wang et al.[143] for high dimensional data sets ($p >> n$). Another penalty type entitled Berhu penalty which is a robust hybrid of L_1 and L_2 has been used by Owen[114]. More recently, Lambert-Lacroix and Zwald[100] combine Huber loss function with Berhu penalty and give the grouping effect and oracle properties of their approach.

We recall that in situations with non heavy tailed errors and without outliers, penalized OLS estimator is excepted to be more efficient. But in some situations, it is better to have loss function like Huber's one which combines squared error's for relatively small residuals and absolute residuals for relatively large ones. However this hybrid approach is also limited since we only have a loss function which is a combination of conditional median and/or conditional mean. Quantile regression offers more possibilities since we now deal with a family of robust loss functions. Therefore, since the quantile regression loss function is based on L_1 criterion we expect that it will be not adapted for small errors due to strong penalization of small residuals. This fact has been previously discussed by Lambert-Lacroix and Zwald[99] who proposed the Huber loss as an alternative to the LAD loss.

In this chapter we will study grouping effect for quantile regression Elastic net. We expect that this method does not possess oracle properties (even for the usual case $n > p$) as it is the case for least squares, so we consider its adaptive version and give its grouping effect and oracle properties. To our knowledge, these theoretical facts had not yet been studied in details for quantile regression. Practically, another motivation is to verify if the Elastic net had the grouping effect for quantile regression. For Berhu penalties we first recall its methodology and its interest for quantile regression and gives some theoretical results and an issue for computations.

4.2 Doubly regularized Quantile regression

We consider a size n i.i.d sample $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ from some unknown population, where each observation predictors vector $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

Since we are interested on variable selection, which is a common practice in major studies, the main objective is to choose a set of relevant variables that explain the response variable y. In some situations we need to select more than n variables, the fact that is not possible for Lasso or adaptive Lasso when $p > n$. Moreover, in high dimension $p \geq n$, some variables can be correlated and form

groups as can be seen in our examples. So, we need to take into account these information in our variable selection criterion by a judicious choice of a penalty.

4.3 QR with Elastic net penalty

We first consider this version of the quantile regression elastic net problem :

$$\hat{\beta}_\tau = \text{argmin}_{\beta_\tau \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \parallel \beta_2 \parallel_2^2 \right\}. \tag{4.1}$$

for $\lambda_1 \geq 0$, $\lambda_2 > 0$ and which has the advantage to regularize l_1 QR solutions when we have strongly correlated variables. Its theoretical properties have not yet been studied in the literature, the same for its grouping effect. The technical difficulty for QR is that we cannot transform our problem into an artificial augmented data problem as it is the case for the classical Elastic net. As advocated by Zou and Hastie[160], the grouping effect is the ability for a method to select highly correlated group of variables which should have similar regression coefficients (in terms of absolute value or magnitude). Regarding the signs of regression coefficients, their relative sign will be depending on the sign of the correlation coefficient. For the classical case, it is well known that the Lasso has not the grouping effect property and tends to select a single predictor among a highly correlated set (see [Zou and Hastie[160]]). This facts also seem to be checked for L_1 and elastic net penalized quantile regression as it can be seen on figures (4.1), (4.2), (4.4) and figure (4.5) in illustration example section.

4.3.1 Grouping Effect

The QR elastic net solutions can be denoted by $\hat{\beta}_{0,\tau}$ and $\hat{\beta}_\tau(\lambda_1, \lambda_2)$, but for sake of simplicity we consider the notation $\hat{\beta}_0$ and $\hat{\beta}$. Then, we have the following

proposition.

Proposition 4.3.1 *(Grouping effect) For any pair* (j, l), $1 \leq j, l \leq p$ *we have*

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{\mid L_\tau \mid}{\lambda_2} \parallel \mathbf{x}_j - \mathbf{x}_l \parallel_1 = \frac{\mid L_\tau \mid}{\lambda_2} \sum_{i=1}^{n} |x_{ij} - x_{il}| \tag{4.2}$$

where $\mid L_\tau \mid = max(\tau, 1-\tau)$, *and* $0 < \tau < 1$.

Furthermore, if the predictors $\mathbf{x}_j$, $\mathbf{x}_l$ *are centered and normalized, then*

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{\sqrt{n} \mid L_\tau \mid}{\lambda_2} \sqrt{2(1-\rho)}, \tag{4.3}$$

where $\rho = cor(\mathbf{x}_j, \mathbf{x}_l)$ *is the sample correlation between* $\mathbf{x}_j$ *and* $\mathbf{x}_l$.

Remark 4.3.1 *This result also stands for Huber loss with elastic net penalty, since the Huber loss function is convex so Lipschitz continuous (see Wang et al.[143] for the non adaptive case). We can also see that grouping effect is obtained if we can take great correlation* ($\rho \to 1$) *for other fixed parameters or great* λ_2 *for other fixed parameters (in this case we can set* $\rho = 0$*). We also mention that this result holds for all* $\lambda_1 \geq 0$, *so that the grouping effect is from the quadratic term penalty.*

Proof of Proposition 4.3.1.

We firstly used Theorem 1 in Wang et al.[143] due to the fact that the quantile regression loss function is a positive convex combination of convex functions so, Lipschitz continuous with Lipschitz coefficient $\mid L_\tau \mid = max(\tau, 1-\tau)$.

Consider another set of coefficients $\hat{\beta}_0^* = \hat{\beta}_0$,

$$\hat{\beta}_{j'}^* = \begin{cases} \frac{1}{2}(\hat{\beta}_j + \hat{\beta}_l) & \text{if } j' = j \text{ or } j' = l \\ \hat{\beta}_{j'} & \text{otherwise.} \end{cases}$$

By definition of $\hat{\beta}_0$ and $\hat{\beta}$, we have

$$\sum_{i=1}^{n}\rho_\tau(y_i-\hat{\beta}_0^*-\mathbf{x}_i^T\hat{\beta}^*)+\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j^*|+\frac{\lambda_2}{2}\parallel\hat{\beta}^*\parallel_2^2-\sum_{i=1}^{n}\rho_\tau(y_i-\hat{\beta}_0-\mathbf{x}_i^T\hat{\beta})-\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j|-\frac{\lambda_2}{2}\parallel\hat{\beta}\parallel_2^2\geq 0. \tag{4.4}$$

We have that

$$\sum_{i=1}^{n}[\rho_\tau(y_i-\hat{\beta}_0^*-\mathbf{x}_i^T\hat{\beta}^*)-\rho_\tau(y_i-\hat{\beta}_0-\mathbf{x}_i^T\hat{\beta})]$$

$$\leq\sum_{i=1}^{n}\mid\rho_\tau(y_i-\hat{\beta}_0^*-\mathbf{x}_i^T\hat{\beta}^*)-\rho_\tau(y_i-\hat{\beta}_0-\mathbf{x}_i^T\hat{\beta})\mid$$

$$\leq\mid L_\tau\mid\sum_{i=1}^{n}\mid y_i-\hat{\beta}_0^*-\mathbf{x}_i^T\hat{\beta}^*-y_i-\hat{\beta}_0-\mathbf{x}_i^T\hat{\beta}\mid=\mid L_\tau\mid\sum_{i=1}^{n}\mid\mathbf{x}_i^T(\hat{\beta}^*-\hat{\beta})\mid$$

$$=\mid L_\tau\mid\sum_{i=1}^{n}\mid x_{ij}(\hat{\beta}_j^*-\hat{\beta}_j)+x_{il}(\hat{\beta}_l^*-\hat{\beta}_l)\mid$$

$$=\mid L_\tau\mid\sum_{i=1}^{n}\mid x_{ij}(1/2\hat{\beta}_l-1/2\hat{\beta}_j)+x_{il}(1/2\hat{\beta}_j-1/2\hat{\beta}_l)\mid$$

$$=\frac{\mid L_\tau\mid}{2}\mid\hat{\beta}_j-\hat{\beta}_l\mid\sum_{i=1}^{n}\mid x_{ij}-x_{il}\mid$$

$$=\frac{\mid L_\tau\mid}{2}\mid\hat{\beta}_j-\hat{\beta}_l\mid\parallel\mathbf{x}_j-\mathbf{x}_l\parallel_1. \tag{4.5}$$

We also have

$$\parallel\hat{\beta}^*\parallel_1-\parallel\hat{\beta}\parallel_1=\mid\hat{\beta}_j^*\mid+\mid\hat{\beta}_l^*\mid-\mid\hat{\beta}_j\mid-\mid\hat{\beta}_l\mid=\mid\hat{\beta}_j+\hat{\beta}_l\mid-\mid\hat{\beta}_j\mid-\mid\hat{\beta}_l\mid\leq 0, \tag{4.6}$$

and

$$\parallel\hat{\beta}^*\parallel_2^2-\parallel\hat{\beta}\parallel_2^2=\mid\hat{\beta}_j^*\mid^2+\mid\hat{\beta}_l^*\mid^2-\mid\hat{\beta}_j\mid^2-\mid\hat{\beta}_l\mid^2=-1/2\mid\hat{\beta}_j-\hat{\beta}_l\mid^2\leq 0. \tag{4.7}$$

Combining (4.5),(4.6) and (4.7), (4.4) implies that

$$\frac{\mid L_\tau\mid}{2}\mid\hat{\beta}_j-\hat{\beta}_l\mid\parallel\mathbf{x}_j-\mathbf{x}_l\parallel_1-\frac{\lambda_2}{2}|\hat{\beta}_j-\hat{\beta}_l|^2\geq 0. \tag{4.8}$$

Hence (4.2) is obtained. For (4.3), as in Wang et al.[143] we simply use the inequality (due to equivalency of norms in finite dimension) :

$$\| \mathbf{x}_j - \mathbf{x}_l \|_1 \leq \sqrt{n}\sqrt{\| \mathbf{x}_j - \mathbf{x}_l \|_2^2} = \sqrt{n}\sqrt{2(1-\rho)}.$$

■

4.3.2 Illustration example

We propose to illustrate the elastic net quantile regression grouping effect on both simulated and real data sets.

Example 4.3.1 *We consider the following model* $y_i = \boldsymbol{x}_i^T\beta + \epsilon_i$, *where* $\boldsymbol{x}_i = (x_{i1}, ..., x_{i50})$ *is a multinormal vector with zero mean and covariance matrix* C *including three highly correlated groups* G_1, G_2 *and* G_3.

For each group we have a pairwise correlation $\rho_{ij} = 0.9$ *between* $\boldsymbol{x}_i$ *and* $\boldsymbol{x}_j$ *and each variable is independent with other variables not in the same group.*

Errors $\epsilon_i, 1 \leq i \leq n$ *are obtained as follows : for each observation* i, *we randomly choose (sampling with replacement) a value* $a_i \in \{-1, 1\}$ *and set* $\epsilon_i = a_i u_i$, *where* u_i *follows a standard exponential distribution* ($u_i \sim exp(\lambda = 1)$). *The sample size is fixed to* $n = 101$ *(odd sample size for which* $n\tau$ *is noninteger ensures uniqueness of the initial value of the intercept* β_0 *when computing the solution of the penalized problem[103]). The vector* β *has ten nonzero components and the remaining components are set to be equal to zero, so*

$$\beta = (\underbrace{4,4,4,4,4}_{G_1}, 0,0,0,0,0, \underbrace{7,7,7}_{G_2}, 0,0, \underbrace{-10,-10}_{G_3}, \underbrace{0,...,0}_{33}).$$

Plots for this example are given in figure 4.1 and figure 4.2 where we compare Lasso and Elastic net (for $\lambda_2 = 15$*) quantile regression regularization paths for* $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

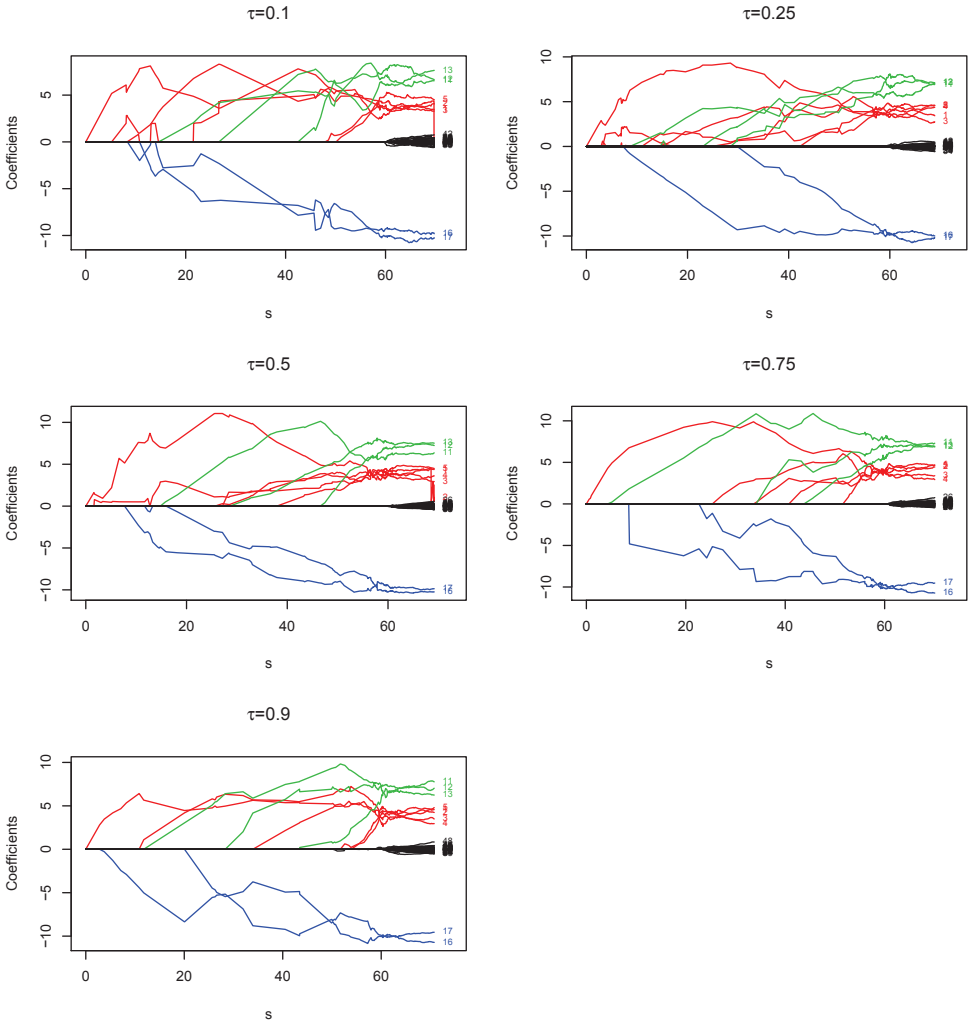

FIGURE 4.1 – Lasso QR regularization paths for simulated example.

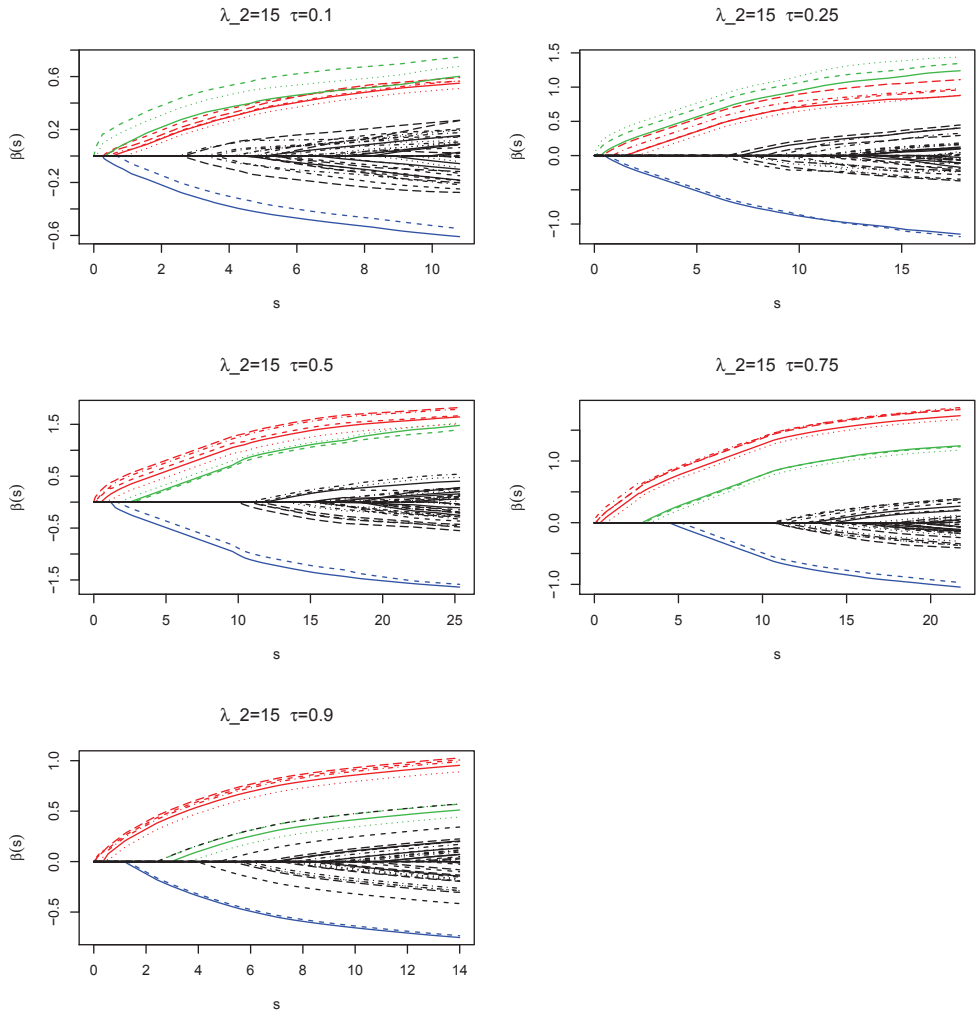

FIGURE 4.2 – Enet ($\lambda_2 = 15$) QR regularization paths for simulated example.

As excepted, only the Elastic net has the grouping effect property, since highly correlated group of variables which should have similar regression coefficients (in terms of absolute value or magnitude) are selected simultaneously or not. This property is clearly visible for $\tau = 0.75$*. Another interesting fact for Enet is that noisy variables seems to enter the selected model for* $\tau = 0.9$ *and* $\tau = 0.1$*, by taking for example the solution corresponding to* $s = 4$*. We recall that the parameter* s *on this figure plays the same role as* λ_1*. Finally, estimated coefficients by Enet are underestimated in absolute value for this choice of* λ_2*. We expect that lower value of* λ_2 *will give best estimates as for the* L_1 *QR but the grouping effect is not guaranteed according to our previous experiments for* $\lambda_2 = 5$ *and* $\lambda_2 = 10$*. This example confirms a known fact that* L_2 *part of the penalty stabilizes the regularization path.*

Example 4.3.2 *For this example we propose to illustrate the grouping effect property of the QR Enet on merlan data set. According to previous chapters we have seen that some volatile compounds are highly correlated, so it is natural to check if the QR Enet has the ability to select groups of highly correlated variables or not. We first set the correlation heat map of all the predictors and a reduced set of highly correlated predictors in figure (4.3).*

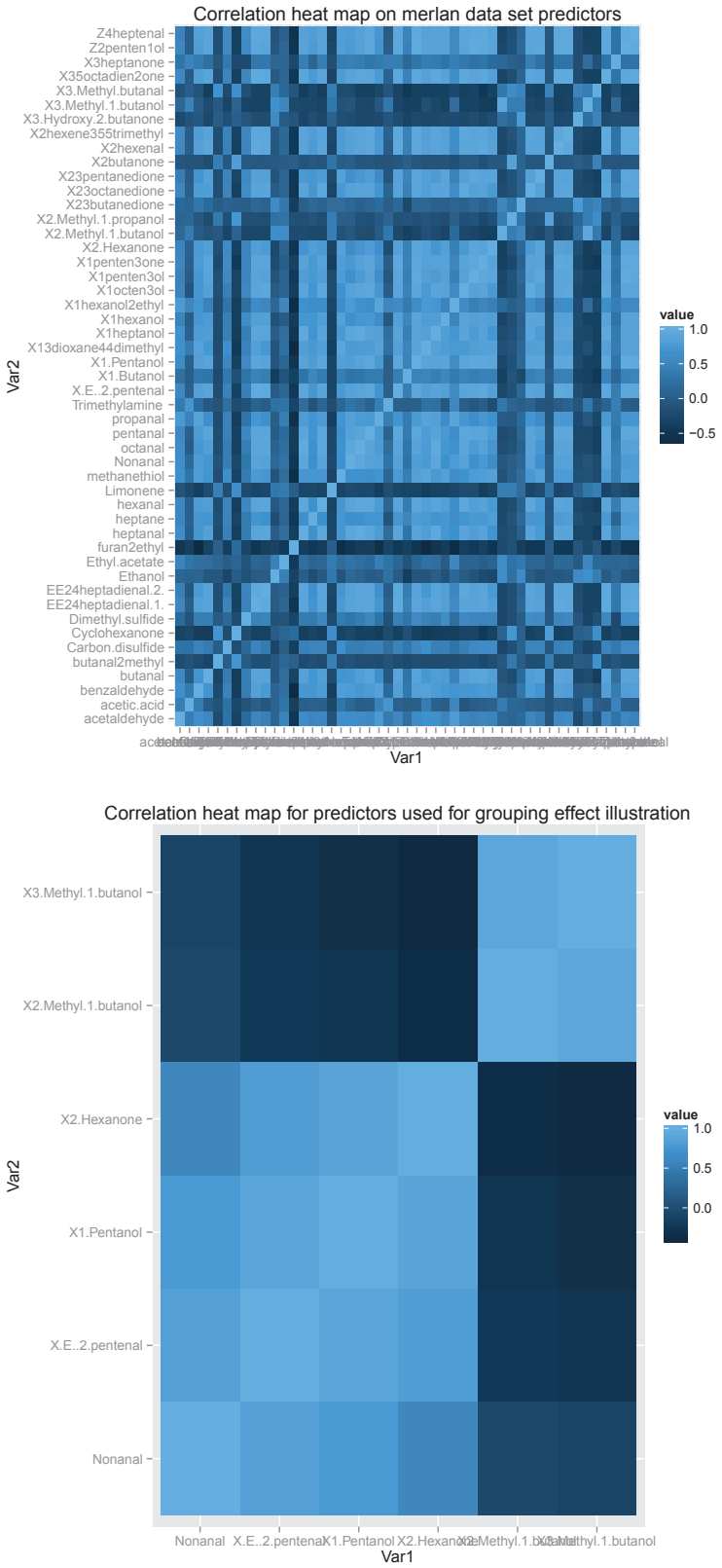

FIGURE 4.3 – Correlation heat map on merlan data set.

As it can be seen, we have many predictors that are correlated, but our previous findings suggest us to focus on the compounds 3 Methyl-1-butanol (ag)*,* 2 Methyl-1-butanol (ah)*,* (E) :2-pentenal (aj)*,* 1-Pentanol(ak)*,* 2-Hexanone (am) *and* nonanal (bh)*. According to the heat map, we clearly see that those six compounds can be classified in two groups :*

- *first group : (ag, ah) with correlation* $\rho = 0.942$
- *second group : (aj, ak, am, bh) with minimum and maximum pairwise correlations* $\rho_{min} = 0.797$ *and* $\rho_{max} = 0.937$

where as before, ρ *is the linear correlation coefficient.*
We consider different colors according to freshness or quality index response variable. In the case of freshness, figure (4.4) gives regularization paths for Lasso and Enet ($\lambda_2 \in \{5, 10, 15\}$) *QR for* $\tau \in \{0.1, 0.25\}$*. As expected only the Enet QR has the grouping effect, with first group compounds are in red color, second group ones in green and other variables are in black. The first group is selected at the same time, but this is not the case for the second group.*

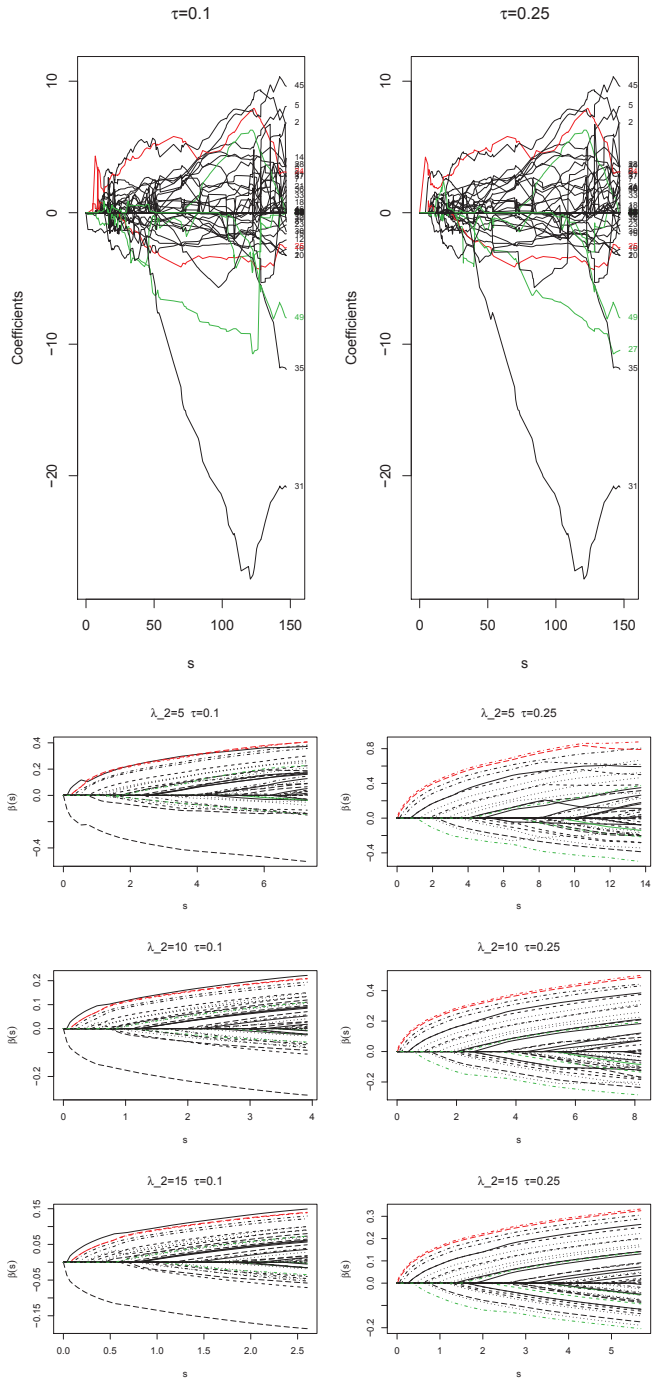

FIGURE 4.4 – Lasso and Enet QR on merlan data for freshness index.

The same remarks hold for the quality index regularization paths in figure (4.5), where the first group is colored in blue and the second one in blue light. For this example, first group enters the model first for $\tau \in \{0.1, 0.25\}$*, but it seems not the case for freshness when* $\tau = 0.1$*, where some variables in black first enter the model.*

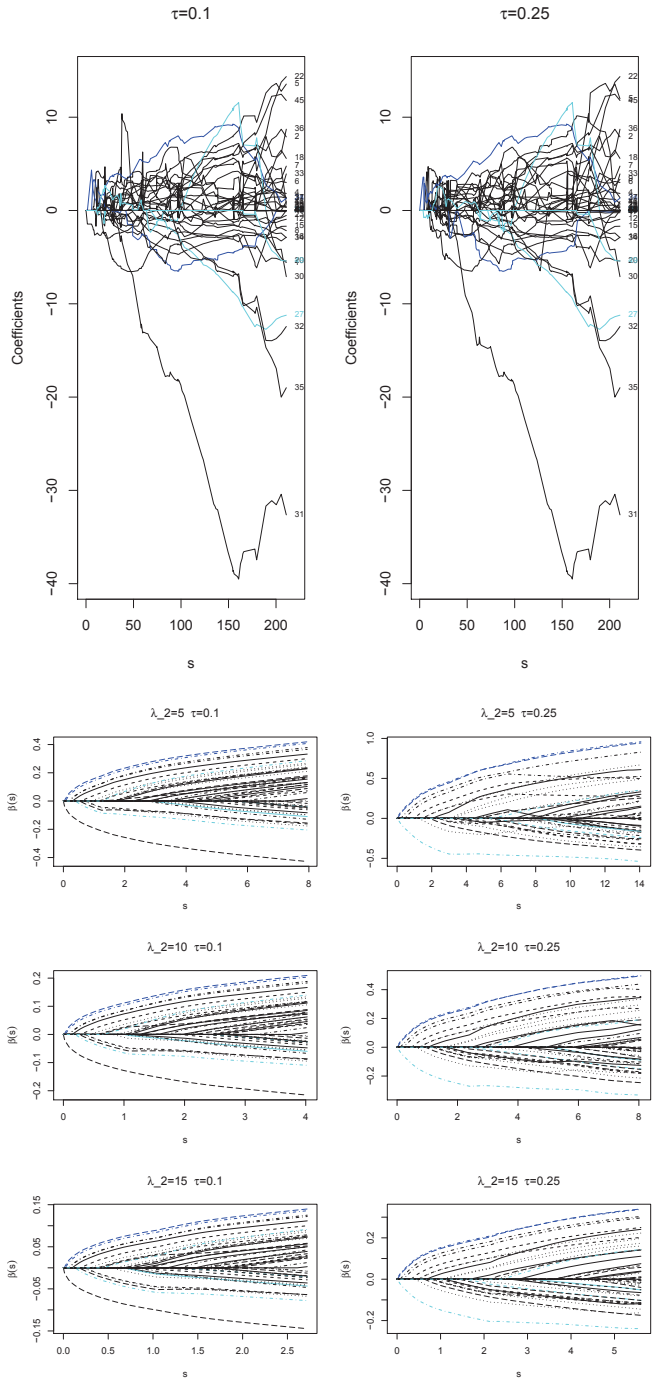

FIGURE 4.5 – Lasso and Enet QR on merlan data for quality index.

Finally, the stabilization property of Enet is also clearly visible on this real data set example. We also give classical Lasso and Enet regularization paths (glmnet version) for penalized linear regression. Surprisingly, the grouping effect is not guaranteed on this data set for both methods (see figure(4.6)).

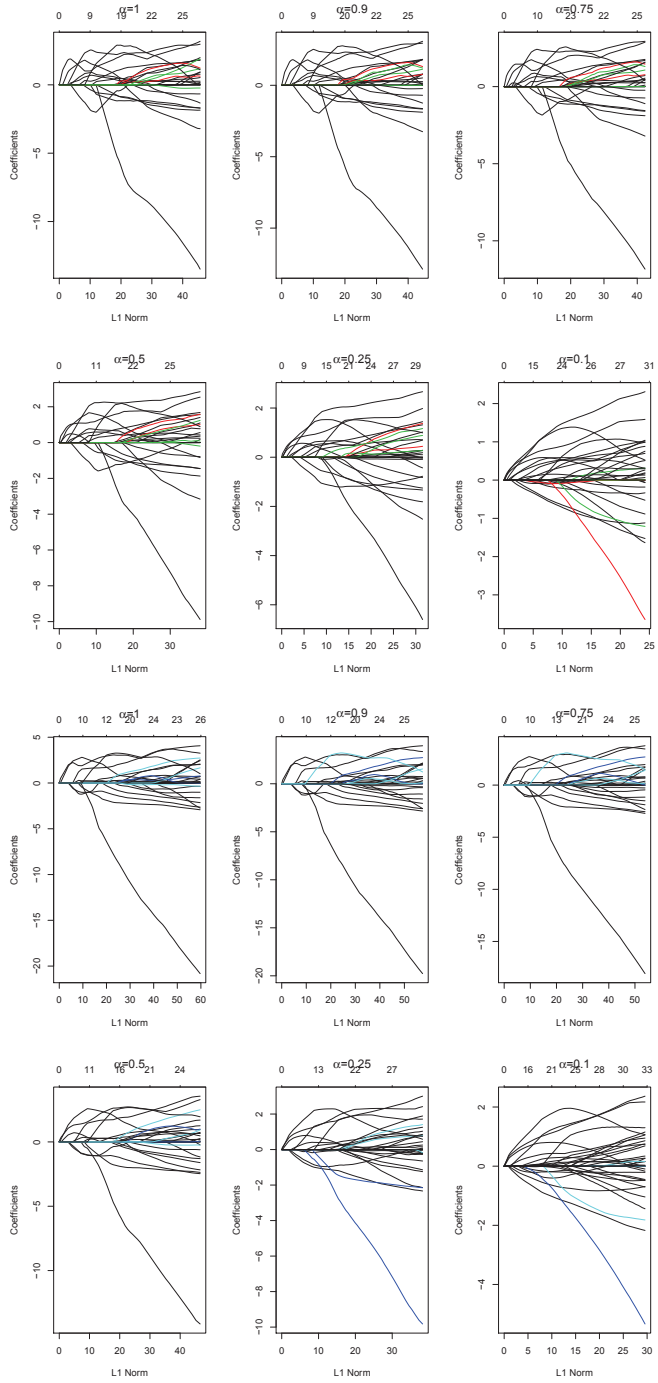

FIGURE 4.6 – L_1 ($\alpha = 1$) and Enet($\alpha \in \{0.9, 0.75, 0.5, 0.25, 0.1\}$) penalized least squares on merlan data for freshness(six first top plots) and quality index(six bottom plots).

This real data example also confirms the fact that Lasso regularization paths are unstable when predictors are highly correlated.

4.4 QR with Adaptive Enet penalty

For this point we consider the following problem :

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + \lambda_1 \sum_{j=1}^{p} \hat{w}_j |\beta_j| + \frac{\lambda_2}{2} \parallel \beta \parallel_2^2 \right\}. \tag{4.9}$$

which is a generalization of the quantile regression Elastic net ; as mentioned in Ghosh[58], Zou and Zhang[165], El Anbari and Mkhadri[43]. The non adaptive Elastic net has not oracle properties for penalized least squares problem. For quantile regression the oracle properties had not yet been studied to our knowledge for both adaptive and non adaptive cases. Theoretically, a first attempt had been done by Wu and Liu[148] for the adaptive lasso quantile regression which has the oracle properties. Oracle properties are difficult to establish for the adaptive elastic net unless for the cases where asymptotically this problem is equivalent to adaptive lasso quantile regression problem.

4.4.1 Grouping Effect

We also adopt previous notations for the quantile regression elastic net solution and the following proposition holds :

Proposition 4.4.1 *(Grouping effect) For any pair* (j, l), $1 \leq j, l \leq p$, *verifying* $\hat{\beta}_j \hat{\beta}_l > 0$, *we have*

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{1}{\lambda_2} \left[\sqrt{n} \mid L_\tau \mid\mid\mid \mathbf{x}_j - \mathbf{x}_l \parallel_2 + \lambda_1 \mid \hat{w}_j - \hat{w}_l \mid \right]. \tag{4.10}$$

Furthermore, if we assume $\hat{w}_k = \mid \hat{\beta}_k^o \mid^{-\gamma}$, $\hat{\beta}_k^o, 1 \leq k \leq p$ *is a consistent estimator*

of β_k, $\gamma > 0$, the input variables $\mathbf{x}_j$, $\mathbf{x}_l$ are centered and normalized, then

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{1}{\lambda_2}\left[\sqrt{n} \mid L_\tau \mid \sqrt{2(1-\rho)} + \lambda_1\gamma \mid \hat{\beta}_j^o - \hat{\beta}_l^o \mid\right]. \tag{4.11}$$

Remark 4.4.1 *As a remark from previous result, setting $\gamma = 0$ or taking $\hat{w}_j = \hat{w}_l$ gives us the previous non adaptive case result. For the present situation, grouping effect take into account both adaptive linear and quadratic part of the penalty.*

Proof of Proposition 4.4.1

We recall that the quantile regression loss function can be written as $\rho_\tau(u) = u(\tau - 1_{u<0})$; since $\hat{\beta}_j\hat{\beta}_l > 0$, $\hat{\beta}_j \neq 0$ and $\hat{\beta}_l \neq 0$. Moreover we have $sign(\hat{\beta}_j) = sign(\hat{\beta}_l)$.

Using KKT conditions by differentiating the objective function with respect to β_j and β_l we have :

$$-\mathbf{x}_j^T\left(\tau 1_n - 1_{\{y-X\hat{\beta}<0\}}\right) + \lambda_1\hat{w}_j sign(\hat{\beta}_j) + \lambda_2\hat{\beta}_j = 0 \tag{4.12}$$

$$-\mathbf{x}_l^T\left(\tau 1_n - 1_{\{y-X\hat{\beta}<0\}}\right) + \lambda_1\hat{w}_l sign(\hat{\beta}_l) + \lambda_2\hat{\beta}_l = 0 \tag{4.13}$$

Subtracting (4.12) from (4.13) gives the equality :

$$\lambda_2(\hat{\beta}_j - \hat{\beta}_l) = \left(\mathbf{x}_j^T - \mathbf{x}_l^T\right)\left(\tau 1_n - 1_{\{y-X\hat{\beta}<0\}}\right) - \lambda_1(\hat{w}_j - \hat{w}_l)sign(\hat{\beta}_j) \tag{4.14}$$

which leads to :

$$\mid \hat{\beta}_j - \hat{\beta}_l \mid \leq \frac{1}{\lambda_2}\left[\mid \left(\mathbf{x}_j^T - \mathbf{x}_l^T\right)\left(\tau 1_n - 1_{\{y-X\hat{\beta}<0\}}\right) \mid + \lambda_1 \mid \hat{w}_j - \hat{w}_l \mid\right] \tag{4.15}$$

due to $\mid a \pm b \mid \leq \mid a \mid + \mid b \mid, \forall a, b \in \mathbb{R}$.

Using Cauchy-Schwarz inequality and the fact that $\parallel \tau 1_n - 1_{\{y-X\hat{\beta}<0\}} \parallel_2 \leq \sqrt{n} \mid L_\tau \mid$ (where $\mid L_\tau \mid = max(\tau, 1-\tau)$) implies that

$$\mid \hat{\beta}_j - \hat{\beta}_l \mid \leq \frac{1}{\lambda_2}\left[\sqrt{n} \mid L_\tau \mid \parallel \mathbf{x}_j - \mathbf{x}_l \parallel_2 + \lambda_1 \mid \hat{w}_j - \hat{w}_l \mid\right]. \tag{4.16}$$

For the particular case when we assume that $\hat{w}_k = \mid \hat{\beta}_k^o \mid^{-\gamma}$ for some consistent estimator $\hat{\beta}^o$ and $\gamma > 0$. As in Ghosh[58], the fact that the function $f(x) = x^{-\gamma}$ is Lipschitz continuous, for $x > 0$ and $\gamma > 0$ leads to $\mid f(x) - f(y) \mid \leq \gamma \mid x - y \mid$.
In this case we have :

$$\mid \hat{w}_j - \hat{w}_l \mid = \mid\mid \hat{\beta}_j^o \mid^{-\gamma} - \mid \hat{\beta}_l^o \mid^{-\gamma} \mid \leq \gamma \mid\mid \hat{\beta}_j^o \mid - \mid \hat{\beta}_l^o \mid \mid \leq \gamma \mid \hat{\beta}_j^o - \hat{\beta}_l^o \mid . \quad (4.17)$$

So that, when the predictors X are standardized, inequality (4.16) gives :

$$\mid \hat{\beta}_j - \hat{\beta}_l \mid \leq \frac{1}{\lambda_2} \left[\sqrt{2n(1-\rho)} \mid L_\tau \mid + \gamma\lambda_1 \mid \hat{\beta}_j^o - \hat{\beta}_l^o \mid \right] . \blacksquare \quad (4.18)$$

4.4.2 Asymptotic properties

We consider that the data $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ consists of n observations from the linear model

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i = \mathbf{x}_{i\mathcal{A}}^T \beta_{\mathcal{A}} + \mathbf{x}_{i\mathcal{A}^c}^T \beta_{\mathcal{A}^c} + \epsilon_i, i = 1, ..., n, \quad (4.19)$$

with $P(\epsilon_i < 0) = \tau$ as in condition (i) defined below. We define $\mathcal{A} = \{1 \leq j \leq p, \beta_j \neq 0\}$, $s = \mid \mathcal{A} \mid$ and here $\mathbf{x}_i = (\mathbf{x}_{i\mathcal{A}}^T, \mathbf{x}_{i\mathcal{A}^c}^T)^T, \beta = (\beta_{\mathcal{A}}^T, \beta_{\mathcal{A}^c}^T)^T, \mathbf{x}_{i\mathcal{A}} \in \mathbb{R}^s, \mathbf{x}_{i\mathcal{A}^c} \in \mathbb{R}^{p-s}$, and the true regression coefficients are $\beta_{\mathcal{A}} = \beta_{\mathcal{A}0}$ with components being nonzero, and $\beta_{\mathcal{A}^c} = \beta_{\mathcal{A}^c 0} = \mathbf{0}$ so that $\beta_0 = (\beta_{\mathcal{A}0}^T, \beta_{\mathcal{A}^c 0}^T)^T$.
We also assume the following conditions :

(i) The regression errors $\{\epsilon_i\}$ are independent and identically distributed, with τth quantile zero and a continuous positive density $f(.)$ in a neighborhood of zero.

(ii) The design $\mathbf{x}_i$, $i = 1, ..., n$, is a deterministic sequence for which there exists a positive definite matrix C such that $lim_{n\to\infty}(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)/n = C$. Denote the top-left $s \times s$ submatrix of C by C_{11} and the right-bottom $(p-s) \times (p-s)$ submatrix of C by C_{22}.

Lambert-Lacroix and Zwald[100] have used other regularity conditions for the Huber loss and specify that for the penalized least squares estimator, we generally assume that ϵ_i are independent identically distributed random variables with mean 0 and has a finite variance.
We first prove the oracle property of adaptive Elastic net ; so we consider this formulation of our problem :

$$Q_1(\beta) = \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\beta) + n\lambda_{1n} \sum_{j=1}^{p} \tilde{w}_j|\beta_j| + n\frac{\lambda_{2n}}{2} \parallel \beta \parallel_2^2 \right\}. \tag{4.20}$$

Let $\hat{\beta}^{Aenet} = argmin_\beta Q_1(\beta)$ for $\tilde{w}_j = \mid \tilde{\beta}_j \mid^{-\gamma}$.
It means that we consider a non penalized quantile regression estimator $\tilde{\beta}$ as initial estimator for weights (case $p < n$). In general, when $p >> n$, we can use the L_2 penalized quantile regression estimator as initial estimator ; but this approach leads to a supplementary tuning parameter selection for this ridge type problem. In what follows we consider the case $p < n$.

Theorem 4.4.1 *Assume that for a sample* $\{(x_i, y_i), i = 1, ..., n\}$ *conditions* (i) *and* (ii) *are satisfied. If* $\sqrt{n}\lambda_{1n} \to 0$, $n^{(\gamma+1)/2}\lambda_{1n} \to \infty$, $\lambda_{2n} = \frac{k}{n^\nu}$ *and* $\sqrt{n}\lambda_{2n} \to 0$ *for some constant* $k > 0$ *and* $\nu > 1/2$, *then we have*
1) Sparsity : $\hat{\beta}_2^{Aenet} = 0$.
2) Asymptotic normality : $\sqrt{n}(\hat{\beta}_1^{Aenet} - \beta_{10}) \to_{\mathcal{L}} N(0, \tau(1-\tau)C_{11}^{-1}/f^2(0))$.

Remark 4.4.2 *According to previous theorem, the value* $\lambda_{2n} = \frac{k}{n^\nu}$ *ensures that* $\sqrt{n}\lambda_{2n} \to 0$. *This is the main difference between our result and result in Wu and Liu[148]. This fact ensures that the quadratic term of the Elastic net penalty goes to zero asymptotically, so the adaptive Elastic net quantile regression is reduced to adaptive Lasso quantile regression problem. Surprisingly with this condition we have asymptotic results similar to Wu and Liu's results with the adaptive quantile regression Lasso when* $p < n$. *Similar fact had been previously advocated by Zou and Zhang[165], Zou and Hastie[160] in the context of the least squares where if we have an orthogonal design, the adaptive Elastic net reduces to the adaptive Lasso, independently of the value of* λ_2. *Zou[159] also shows that the adaptive Lasso achieves the optimal minimax risk bound, but for quantile regression this fact had*

not been advocated. Finally, we recall that the quadratic part of Elastic net penalty is to "further regularize the adaptive lasso fit whenever the collinearity may cause serious trouble" *[Zou and Zhang[165]].*

Proof of Theorem 4.4.1

Considering the difference

$$Q_1(\beta_0 + \mathbf{u}/\sqrt{n}) - Q_1(\beta_0),$$

for any fixed $\mathbf{u} \in \mathbb{R}^p$. First, for $j = 1, 2, 3..., s$ we have $\beta_{j0} \neq 0$; as a result, $\tilde{w}_j \rightarrow_P \mid \beta_{j0} \mid^{-\gamma}$, hence

$$n\lambda_{1n}[\tilde{w}_j \mid \beta_{j0} + u_j/\sqrt{n} \mid -\tilde{w}_j \mid \beta_{j0} \mid] \rightarrow_P 0$$

due to

$$\sqrt{n}(\mid \beta_{j0} + u_j/\sqrt{n} \mid - \mid \beta_{j0} \mid) \rightarrow u_j sign(\beta_{j0})$$

and

$$\sqrt{n}\lambda_{1n} \rightarrow 0.$$

Moreover, we have

$$\sqrt{n}((\beta_{j0} + u_j/\sqrt{n})^2 - \beta_{j0}^2) \rightarrow 2u_j\beta_{j0}$$

and $\sqrt{n}\lambda_{2n} = \frac{k}{n^{\nu-1/2}} \rightarrow 0$ for fixed $k > 0$ and $\nu > 1/2$. This fact implies that

$$\frac{n\lambda_{2n}}{2}((\beta_{j0} + u_j/\sqrt{n})^2 - \beta_{j0}^2) \rightarrow 0.$$

On the other hand, for $j = s + 1, ..., p$, the true coefficient $\beta_{j0} = 0$; so

$$\sqrt{n}\lambda_{1n}\tilde{w}_j = n^{(1+\gamma)/2}\lambda_{1n}(\sqrt{n} \mid \tilde{\beta}_j \mid)^{-\gamma}$$

with $\sqrt{n}\tilde{\beta}_j = O_p(1)$; so it follows that

$$n\lambda_{1n}[\tilde{w}_j(\mid \beta_{j0} + u_j/\sqrt{n} \mid - \mid \beta_{j0} \mid)] \rightarrow_P \infty$$

when $u_j \neq 0$ and $= 0$ otherwise due to $\sqrt{n} \mid u_j/\sqrt{n} \mid = \mid u_j \mid$. We also have that

$$n\frac{\lambda_{2n}}{2}[(\beta_{j0} + u_j/\sqrt{n})^2 - \beta_{j0}^2] = \frac{\lambda_{2n}u_j^2}{2} = \frac{\sqrt{n}\lambda_{2n}}{2}.\frac{u_j^2}{\sqrt{n}} \to 0,$$

due to the fact that when $u_j \neq 0$, $\frac{u_j^2}{\sqrt{n}} \to 0, \sqrt{n}\lambda_2 n \to 0$ for large n and 0 otherwise. These facts and Lemma 3 in Wu and Liu[148] imply that

$$Q_1(\beta_0+\mathbf{u}/\sqrt{n})-Q_1(\beta_0) \to_{\mathcal{L}} V(\mathbf{u}) = \begin{cases} \frac{f(0)}{2}\mathbf{u}_1^T C_{11}\mathbf{u}_1 + W_{n,11}^T\mathbf{u}_1 & \text{if} \quad u_j = 0, j \geq s+1; \\ \infty & \text{otherwise.} \end{cases}$$

where $\mathbf{u}_1 = (u_1, u_2, ..., u_s)^T$. Noticing that $Q_1(\beta_0 + \mathbf{u}/\sqrt{n}) - Q_1(\beta_0)$ is convex in $\mathbf{u}$ and V has a unique minimizer, the epi-convergence results (see Geyer[57]) imply that

$$argminQ_1(\beta_0 + \mathbf{u}/\sqrt{n}) = \sqrt{n}(\hat{\beta}^{(Aenet)} - \beta_0) \to_{\mathcal{L}} argminV(u),$$

which establishes the asymptotic normality part.

For the consistency in variable selection, we have for any given β satisfying $\beta_1 - \beta_{10} = O_p(n^{-1/2})$ and $0 < \| \beta_2 \|_2 \leq Cn^{-1/2}$ for any constant C,

$$\begin{aligned} & Q_1((\beta_1^T, 0^T)^T) - Q_1((\beta_1^T, \beta_2^T)^T) \\ = & [Q_1((\beta_1^T, 0^T)^T) - Q_1((\beta_{10}^T, 0^T)^T)] - [Q_1((\beta_1^T, \beta_2^T)^T) - Q_1((\beta_{10}^T, 0^T)^T)] \\ = & G_n(\sqrt{n}((\beta_1 - \beta_{10})^T, 0^T)^T) - G_n(\sqrt{n}((\beta_1 - \beta_{10})^T, \beta_2^T)^T) - n\lambda_{1n}\sum_{j=s+1}^{p} \tilde{w}_j \mid \beta_j \mid \\ & -\frac{n\lambda_{2n}}{2}\sum_{j=s+1}^{p} \beta_j^2. \end{aligned}$$

The first two terms can be bounded (see Wu and Liu[148]). However the third term goes to $-\infty$ as $n \to \infty$ due to the following fact

$$n\lambda_{1n}\sum_{j=s+1}^{p} \tilde{w}_j \mid \beta_j \mid = (n^{(1+\gamma)/2}\lambda_{1n})\sqrt{n}\sum_{j=s+1}^{p} \mid (\sqrt{n} \mid \tilde{\beta}_j \mid)^{-\gamma} \mid\mid \beta_j \mid \to \infty.$$

$$\frac{n\lambda_{2n}}{2}\sum_{j=s+1}^{p}\beta_j^2 \leq \frac{n\lambda_{2n}}{2} \parallel \beta_2 \parallel_2^2 \leq \frac{n\lambda_{2n}}{2}C^2 n^{-1} = \frac{\lambda_{2n}C^2}{2} \to 0$$

due to

$$\lambda_{2n} = \frac{k}{n^{\nu}} \to 0,$$

for large n, $\nu > 1/2$.
Finally, the condition $n^{(1+\gamma)/2}\lambda_{1n} \to \infty$ implies that $n\lambda_{1n}\sum_{j=s+1}^{p}\tilde{w}_j \mid \beta_j \mid$ is of higher order than any other term and dominates. This in turn implies that

$$Q_1((\beta_1^T, 0^T)^T) - Q_1((\beta_1^T, \beta_2^T)^T) < 0$$

for large n. This proves the consistency of model selection of the adaptive elastic net penalized quantile regression. ■

4.5 QR with Berhu penalty

4.5.1 The procedure

It will be interesting to have a penalty which uses a single tuning parameter (contrarily to the Elastic net) and combines L_1 for relatively small coefficients and L_2 for large ones penalties properties (hybrid penalty). So, when performing variable selection the following problem is considered :

$$\hat{\beta}_\tau = \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_i^t\beta\right) + \lambda P(\beta)\right\}. \tag{4.21}$$

where $P(.) = \mathfrak{B}_M(.)$ is the Berhu function defined by

$$\mathfrak{B}_M(z) = M\mathfrak{B}_M\left(z/M\right) = \begin{cases} |z|, & |z| \leq M \\ \frac{z^2+M^2}{2M}, & |z| > M \end{cases}$$

where $M > 0$.

We consider the scaled version of this function (as it is the case for Huber criterion to ensure hybrid property) defined by

$$\sum_{j=1}^{p} \mathfrak{B}_M \left(\frac{\beta_j}{\theta} \right),$$

with θ is a scale parameter to be determined. As suggested by Lambert-Lacroix and Zwald[100] we can replace the penalty term (see [Owen[114]]) by

$$P(\beta) = min_{\theta>0} \left(p\theta + \theta \sum_{j=1}^{p} \mathfrak{B}_M \left(\frac{\beta_j}{\theta} \right) \right),$$

due to convexity arguments that ensure that $P(\beta)$ is convex by considering the fact that function $p\theta + \theta \sum_{j=1}^{p} \mathfrak{B}_M \left(\frac{\beta_j}{\theta} \right)$ is jointly convex as a function of $(\beta, \theta) \in R^p \times (0, \infty)$. This form of the penalty is called Berhu penalty with concomitant scale. One important behavior of this penalty is implicitly to create one group with the largest coefficients. As advocated by Lambert-Lacroix and Zwald[100], the created group is penalized in L_2 way like the grouping Lasso (Yuan and Lin[153]) to avoid to remove anyone of these largest coefficients.

In what follows we consider, as in Lambert-Lacroix and Zwald[100] the penalty $min_{\theta \in \mathbb{R}} P^{adb}(\beta, \theta)$ in order to make the Berhu penalty adaptive with

$$P^{adb}(\beta, \theta) = \begin{cases} \theta \left(\sum_{j=1}^{p} \frac{1}{\hat{w}_j^{adb}} + \sum_{j=1}^{p} \hat{w}_j^{adb} \mathfrak{B}_M \left(\frac{\beta_j}{\theta} \right) \right) & \text{if } \theta > 0, \\ 0 & \text{if } \beta = 0, \theta = 0, \\ +\infty & \text{if } \beta \neq 0, \theta = 0. \end{cases}$$

where $\hat{w}^{adb} = (\hat{w}_1^{adb}, ..., \hat{w}_p^{adb})$ is a known weights vector. We recall that Owen[114] used the Berhu penalty in its non adaptive form. Therefore, as in Lambert-Lacroix and Zwald[100], we need an adaptive form of Berhu penalty for asymptotic feature. The quantile regression loss function does not need to be scaled since scaling for L_1 loss will lead to degeneracy (Owen[114]). In many studies, the weights vector is given by $\hat{w}_j^{adb} = \mid \hat{\beta}_j^{unpen} \mid^{-\gamma}$, $j = 1, ..., p$, where $\gamma > 0$ and $\hat{\beta}^{unpen}$ denotes the unpenalized estimator which must only be root-n consistent estimator of β.

The adaptive Berhu penalty also behaves like the Lasso on the smallest coefficients and does not delete the largest ones, whatever the correlation structure. This interpretation is based on the following fact : in the non adaptive case, when β is fixed with $k(\beta)$ is the number of its nonzero components, we can sort the absolute values of its components, say

$$\mid \beta_{(p)} \mid \leq ... \leq \mid \beta_{(1)} \mid .$$

Then the minimum defined in $P(\beta)$ is achieved at (Lambert-Lacroix and Zwald[100])

$$\hat{\theta}(\beta) = \sqrt{\frac{1}{2Mp + M^2(q(\beta)-1)} \sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2},$$

if $\beta \neq 0$ and where $q(\beta)$ is the unique integer between 2 and $k(\beta)+1$ such that

$$\mid \beta_{(q(\beta))} \mid /M \leq \hat{\theta}(\beta) \leq \mid \beta_{(q(\beta)-1)} \mid /M.$$

Consequently,

$$P(\beta) = \sqrt{\frac{2p}{M} + q(\beta) - 1} \sqrt{\sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2 + \sum_{j=q(\beta)}^{k(\beta)} \mid \beta_{(j)} \mid} . \tag{4.22}$$

Previous form of $P(\beta)$ in (4.22) is closely related to Elastic Corr-Net procedure proposed by El Anbari and Mkhadri[44]. We can also see from (4.22) that the Behru penalty with concomitant term implicitly creates one group with the largest coefficients. This group is penalized in a L_2 way as for the case of grouping Lasso penalty (Yuan and Lin[153]). To avoid to remove anyone of these largest coefficients and the smallest ones are penalized individually in L_1 way, whatever the structure of the correlation matrix.

4.5.2 Grouping effect

As with Enet, we study in this section the property of grouping effect of Berhu QR. We have the following theorem for grouping effect of adaptive Berhu penalty.

Proposition 4.5.1 *(Grouping effect) Let $\gamma > 0$ and*

$$(\hat{\beta}^{adb}, \hat{\theta}^{adb}) = argmin_{(\beta,\theta)\in\mathbb{R}^p\times\mathbb{R}_+^*} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\beta) + \lambda_n P^{adb}(\beta, \theta).$$

We suppose that $\lambda_n > 0$, $\hat{\beta}_i^{adb} \neq 0$ and $\hat{\beta}_j^{adb} \neq 0$ for $i \neq j$. Then, the following bound holds

$$|\hat{\beta}_i^{adb}\hat{w}_i^{adb} - \hat{\beta}_j^{adb}\hat{w}_j^{adb}| \leq \frac{M\hat{\tau}^{adb}}{\lambda_n}\sqrt{\| \mathbf{x}_i \|_2^2 + \| \mathbf{x}_j \|_2^2 - 2K_{i,j}\mathbf{x}_i^T\mathbf{x}_j}, \tag{4.23}$$

where $K_{i,j} = min\left(1, \frac{|\hat{\beta}_i^{adb}|}{M\hat{\theta}^{adb}}, \frac{|\hat{\beta}_j^{adb}|}{M\hat{\theta}^{adb}}, \frac{|\hat{\beta}_i^{adb}\hat{\beta}_j^{adb}|}{(M\hat{\theta}^{adb})^2}\right)$.
When the variables are standardized in L_2-norm, this leads to

$$|\hat{\beta}_i^{adb}\hat{w}_i^{adb} - \hat{\beta}_j^{adb}\hat{w}_j^{adb}| \leq \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{2(1 - K_{i,j}\mathbf{x}_i^T\mathbf{x}_j)}. \tag{4.24}$$

Remark 4.5.1 *When considering $\gamma = 0$, we exactly get the grouping effect property in the non-adaptive case. Let now $\gamma \in \mathbb{R}_+^*$, the upper bound of the previous equation is a decreasing function of the correlation $\rho_{ij} = \mathbf{x}_i^T\mathbf{x}_j$ between variables i and j (since $K_{i,j} > 0$). Similar bounds have already been given by Zou and Hastie[160], Ghosh[58] for penalized least squares and by Lambert-Lacroix and Zwald[100] for the Huber loss function. However, as advocated by Lambert-Lacroix and Zwald[100], we do not have to assume that the coefficients $\hat{\beta}_i^{adb}$ and $\hat{\beta}_j^{adb}$ have the same sign. Ghosh[58] also assume that for the adaptive elastic net, the initial estimator satisfies the grouping effect property. We also recall that for the adaptive elastic net, Zou and Zhang[165] recommend to choose a non-adaptive elastic net estimator as an initial estimator for the weights vector.*
In this chapter we will not present details of Berhu computations for quantile regression and plots considered for Elastic net are only for the non adaptive case. Another very interesting discussion can be found in Lambert-Lacroix and Zwald[100] around the choice of the initial estimator to use for weights.

Proof of Proposition 4.5.1

The proof is very close to those used by Lambert-Lacroix and Zwald[100]. We recall that $\rho_\tau(u) = u(\tau - 1_{u<0})$; since $\hat{\beta}_i^{adb} \neq 0$, we have $\hat{\beta}^{adb} \neq 0$ and $\hat{\theta}^{adb} > 0$. Using KKT conditions by differentiating with respect to β_i, β_j and θ :

$$- \mathbf{x}_i^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right) + \lambda_n \hat{w}_i^{adb} \mathfrak{B}'_M \left(\frac{\hat{\beta}_i^{adb}}{\hat{\theta}^{adb}}\right) = 0 \tag{4.25}$$

$$- \mathbf{x}_j^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right) + \lambda_n \hat{w}_j^{adb} \mathfrak{B}'_M \left(\frac{\hat{\beta}_j^{adb}}{\hat{\theta}^{adb}}\right) = 0 \tag{4.26}$$

$$\sum_{j:|\hat{\beta}_j^{adb}|>M\hat{\theta}^{adb}} \hat{w}_j^{adb} \left(\frac{1}{2M}\left(\frac{\hat{\beta}_j^{adb}}{\hat{\theta}^{adb}}\right)^2 - \frac{M}{2}\right) = \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{adb}} \tag{4.27}$$

The last score equation implies that the set $G = \{j \in 1, ..., p, \mid \hat{\beta}_j^{adb} \mid > M\hat{\theta}^{adb}\}$ is non empty.
If i and $j \in G$, we have

$$- \mathbf{x}_i^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right) + \lambda_n \hat{w}_i^{adb} \frac{\hat{\beta}_i^{adb}}{M\hat{\theta}^{adb}} = 0 \tag{4.28}$$

and

$$- \mathbf{x}_j^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right) + \lambda_n \hat{w}_j^{adb} \frac{\hat{\beta}_j^{adb}}{M\hat{\theta}^{adb}} = 0. \tag{4.29}$$

Subtracting the second one to the first one, we get :

$$- \mathbf{x}_i^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right) + \mathbf{x}_j^T \left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb}<0\}}\right)$$

$$+\lambda_n \left(\hat{w}_i^{adb} \frac{\hat{\beta}_i^{adb}}{M\hat{\theta}^{adb}} - \hat{w}_j^{adb} \frac{\hat{\beta}_j^{adb}}{M\hat{\theta}^{adb}}\right) = 0. \tag{4.30}$$

Previous equation implies that

$$\mid \hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} \mid \leq \frac{M\hat{\theta}^{adb}}{\lambda_n} \mid \left(\mathbf{x}_i^T - \mathbf{x}_j^T\right)\left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb} < 0\}}\right) \mid . \quad (4.31)$$

Using Cauchy-Schwarz inequality, we obtain that

$$\mid \hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} \mid \leq \frac{M\hat{\theta}^{adb}}{\lambda_n} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel_2 \parallel \tau 1_n - 1_{\{y - X\hat{\beta}^{adb} < 0\}} \parallel_2 . \quad (4.32)$$

Moreover, we have

$$\parallel \tau 1_n - 1_{\{y - X\hat{\beta}^{adb} < 0\}} \parallel_2 \leq \sqrt{n} \mid L_\tau \mid,$$

where $\mid L_\tau \mid = max(\tau, 1 - \tau)$.
Then from equation (4.32), we have

$$\mid \hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} \mid \leq \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{n} \mid L_\tau \mid \parallel \mathbf{x}_i - \mathbf{x}_j \parallel_2,$$

which means that

$$\mid \hat{w}_i^{adb}\hat{\beta}_i(\theta) - \hat{w}_j^{adb}\hat{\beta}_j(\theta) \mid \leq \sqrt{n} \mid L_\tau \mid \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{\parallel \mathbf{x}_i \parallel_2^2 + \parallel \mathbf{x}_j \parallel_2^2 - 2\mathbf{x}_i^T\mathbf{x}_j}.$$

If only one index among i, j belongs to G, suppose that $i \in G$ and $j \notin G$, then in this case (4.25) and (4.26) become (4.28) and (4.29) ; equation (4.27) becomes

$$-x_j^T\left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb} < 0\}}\right) + \lambda_n \hat{w}_j^{adb} sign(\hat{\beta}_i^{adb}) = 0.$$

Previous equality leads to

$$\hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} = \frac{M\hat{\theta}^{adb}}{\lambda_n}\left(\mathbf{x}_i - \frac{\mid \hat{\beta}_j^{adb} \mid}{M\hat{\theta}^{adb}}\mathbf{x}_j\right)^T\left(\tau 1_n - 1_{\{y - X\hat{\beta}^{adb} < 0\}}\right), \quad (4.33)$$

i.e.

$$\begin{aligned} \mid \hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} \mid &\leq \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{n} \mid L_\tau \mid \parallel \frac{\mathbf{x}_j \mid \hat{\beta}_j^{adb} \mid}{M\hat{\theta}^{adb}} - \mathbf{x}_i \parallel_2 \\ &\leq \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{n} \mid L_\tau \mid \sqrt{\parallel \mathbf{x}_i \parallel_2^2 + \parallel \mathbf{x}_j \parallel_2^2 - 2\frac{\mid \hat{\beta}_j^{adb} \mid}{M\hat{\theta}^{adb}}\mathbf{x}_j^T\mathbf{x}_i}. \end{aligned} \quad (4.34)$$

Finally, when indices i and j do not belong to G, using similar arguments we obtain

$$\mid \hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} \mid \leq \frac{M\hat{\theta}^{adb}}{\lambda_n}\sqrt{n} \mid L_\tau \mid \sqrt{\parallel \mathbf{x}_i \parallel_2^2 + \parallel \mathbf{x}_j \parallel_2^2 - 2\frac{\mid \hat{\beta}_i^{adb}\hat{\beta}_j^{adb} \mid}{(M\hat{\theta}^{adb})^2}\mathbf{x}_i^T\mathbf{x}_j}.\blacksquare \tag{4.35}$$

4.5.3 Equivariance

As it is the case for nonpenalized quantile regression, we also have scale invariant properties for penalized quantile regression. As discussed by Lambert-Lacroix and Zwald[99], it is important to consider estimators having this property so that if y is replaced by $cy, c > 0$, the selected variables are the same and the prediction is affected similarly. Then, LAD criterion (as a special case of quantile regression) with Lasso penalty is always scale invariant contrarily to the classical Lasso procedure (OLS with Lasso penalty) which is not scale invariant. They also provide issue to ensure this property for the adaptive penalties cases.

4.6 Computations and selection of tuning parameters

4.6.1 Elastic net and Adaptive Elastic net QR

Slawski[134] have considered the following penalty

$$\Omega(\beta) = \alpha \parallel \beta \parallel_1 + (1-\alpha)\beta^T\Lambda\beta, \alpha \in (0,1),$$

where Λ is a $p \times p$ symmetric and positive definite matrix. He advocated that the doubly regularization choosing can be difficult due to the fact that we have to use two-dimensional grid search. The author also claim that for the quantile regression or SVM loss function, we have a simplification due to the fact that for

one tuning parameter kept fixed, the whole range of solutions can be obtained by tracking a piecewise linear solution path (voir Rosset and Zhu[125]). So we can use one-dimensional grid search.

In the case of k-fold cross-validation, one specifies a grid of values to be explored for one of the two tuning parameters. In an outer loop, one runs through the different folds, fixing a training and a test sample (the sample hold out) each time. In an inner loop, one runs through the grid points. Having one of the two parameters fixed within each inner iteration, a solution path of the second parameter is computed. For the regularization path, Wang et al.[143] have proved that in the case of the SVM that, as long as the amount of L_1 or L_2 regularization is kept fixed, the construction of piecewise linear solution paths is possible.
For the solution path in $\| \beta \|_1$, Slawski[134] considered the following problem

$$(\hat{\beta}_0(s), \hat{\beta}(s)) = argmin_{\beta_0,\beta} L\psi(y_i, f(\mathbf{x}_i; \beta_0, \beta)) + \frac{\lambda_2}{2}\beta^T \Lambda \beta, \lambda_2 > 0 \qquad (4.36)$$

subject to $\| \beta \|_1 \leq s$.
The objective function is well-defined and piecewise linear if $L\psi$ is the QR or SVM loss function and the related regularity Assumption 1 holds. He also recalls that above formulation is equivalent to a structured elastic net problem in the sense that for each s in (4.36) there exist a corresponding Lagrangian multiplier $\lambda_1(s)$, which will be shown to depend in a piecewise constant manner on s, such that

$$\lambda(s) = \lambda_1(s) + \frac{\lambda_2}{2}$$

and

$$\alpha(s) = \frac{\lambda_1(s)}{\lambda_1(s) + \frac{\lambda_2}{2}}.$$

On the other hand, in Wang et al.[143] author's advocate the fact that in practice, people can pre-specify a finite grid of values for λ_1 and λ_2 that covers a wide range, then use either a separate validation data set or cross-validation to do a grid search to find values for the (λ_1, λ_2) pair that give the best performance across the given grid. This is also the preconized approach by Lambert-Lacroix and Zwald[100] for Elastic net type procedures (adaptive or not).

For the adaptive Elastic net, we can take weights $\hat{w}_j = \mid \hat{\beta}_j^{unpen} \mid^{-\gamma}, j = 1, ..., p$, $\gamma > 0$ and $\hat{\beta}^{unpen}$ denotes the unpenalized estimator. For the adaptive Elastic net, Zou and Zhang[165] considered $\hat{w}_j = \mid \hat{\beta}_j(Enet) \mid^{-\gamma}, j = 1, ..., p$, but since the Elastic net naturally adopts a sparse representation, the weights $\hat{w}_j = (\mid \hat{\beta}_j(Enet) \mid +1/n)^{-\gamma}$ can be used to avoid dividing by zero. The value $\hat{w}_j = \infty$ can also be used when $\hat{\beta}_j(Enet) = 0$.
In this chapter, we only illustrate that the solution path for a fixed value of λ_2, denoted as $\beta_2(\lambda_1)$, is piece-wise linear as a function of λ_1 (in the $\mathbb{R}^p$ space) ; nevertheless for a fixed value of λ_1, the solution path, denoted as $\beta_1(\lambda_2)$, is also piece-wise linear as a function of $1/\lambda_2$. More details about two dimensional grid search can be found in Slawski[134] and Wang et al.[143]. For situations were we have to choose an initial estimator based on penalized ridge estimator as it is generally the case for $p > n$, the regularization parameter (for the ridge) can be chosen by 5-fold cross validation on training data.
We propose the following linear program to compute the adaptive Elastic net QR estimator :

$$minimize \left[\tau 1_n^T \mathbf{r}^+ + (1-\tau) 1_n^T \mathbf{r}^- + \lambda_2 \beta^T \beta\right]$$

subject to :

$\parallel \hat{w}^T \beta \parallel_1 \leq s$
$y_i = \mathbf{x}_i^T \beta + r_i^+ - r_i^-, 1 \leq i \leq n$
$u_j, v_j \geq 0, 1 \leq j \leq p$
$(\beta, \mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}$.

4.6.2 Computations with Berhu QR

Computations for Behru penalized quantile regression, require to determine the weights vector, regularization parameters λ_n and parameter M for Berhu's penalty. We can fix the Berhu penalty parameter $M = 1.345$ as it is the case for Huber function (Lambert-Lacroix and Zwald[100]), in general, interesting cases are for $M > 1$ (Owen[114]). The parameter M can also been chosen from data (for

example by cross validation simultaneously with the tuning parameter). Nevertheless, Lambert-Lacroix and Zwald[100] claimed that in practice, they do not observe some improvement to make it data adaptive. Optimal tuning parameter for Berhu penalty can be obtained by considering BIC-type criterions. More precisely, we have to select λ_n minimizing

$$log\left(\sum_{i=1}^{n}\rho_\tau\left(y_i-\hat{\beta}_{0\lambda_n}(\tau)-\mathbf{x}_i^T\hat{\beta}_{\lambda_n}\right)\right)+k_{\lambda_n}\frac{log(n)}{2n},$$

over λ_n, where k_{λ_n} denotes the model dimension. Practically k_{λ_n} can be determined by the non zero coefficient of the estimator $\hat{\beta}_{\lambda_n}$. This BIC type criterion can also been applied for Lasso or adaptive Lasso QR type problems. In the underlying model, all residuals are supposed to have a double exponential distribution (Lambert-Lacroix and Zwald[99]). For the computations, we can use R software or Matlab. For example numerical experiments by Lambert-Lacroix and Zwald ([99],[100]), are based on Matlab CVX package which uses the methodology of disciplined convex programming that imposes a limited set of convention or rules (DCP ruleset). In order to simplify the computations, they use the fact that the Huber function is the Moreau-Yosida regularization of the absolute value function (Rockafellar[124]), so that the Berhu penalty satisfies

$$\mathcal{B}_M(z)=min_{w\geq M\vee|z|}\left(\frac{w^2}{2M}-w+\mid z\mid+\frac{M}{2}\right),$$

where $a\vee b=max(a,b)$ with $a,b\in\mathbb{R}$.
We have not yet established a R code for quantile regression with Berhu penalty and the eventual using of Matlab software is considered. The major difficulty will certainly be in performing the regularization paths. Actually, an alternative for computing penalized quantile regression with Berhu penalty is to consider the quadratic programming formulation of our problem :

$$\text{argmin}_\beta\left\{\sum_{i=1}^{n}\rho_\tau\left(y_i-\mathbf{x}_i^T\beta\right)+\lambda_n\sum_{j=1}^{p}(\theta+\mathfrak{B}_M(\beta_j/\theta)\theta)\right\}.$$

which is equivalent to :

$$minimize \left[\tau 1_n^T \mathbf{r}^+ + (1-\tau) 1_n^T \mathbf{r}^- + \lambda_n \left(p\theta + 1_p^T \mathbf{u} + 1_p^T \frac{\mathbf{v}^2}{2M\theta} + 1_p^T \mathbf{v} \right) \right]$$

subject to :

$y_i = \mathbf{x}_i^T \beta + r_i^+ - r_i^-, 1 \leq i \leq n$
$\mid \beta_j \mid \leq u_j + v_j$
$u_j \leq M\theta$
$\theta \geq 0$
$u_j, v_j \geq 0, 1 \leq j \leq p$
$(\beta, \mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}$.

The computation for adaptive Behru penalty can be directly obtained by using CVX as in Lambert-Lacroix and Zwald[100], it suffices to replace the loss function by quantile regression loss.

One interesting discussion is around the fact that predictors have to be centered or not ? Should we include the intercept or not ? How can we estimate the intercept for penalized quantile regression ? In the case of penalized OLS when predictors are centered ($\Sigma_{i=1}^n \mathbf{x}_i = 0$), $\hat{\alpha} = \bar{y}$ and $\hat{\beta}$ value is not affected due to the fact that intercept is not penalized and the squared loss is used. For penalized quantile regression, the intercept term is equal to the τth quantile of the residual vector $y - X\hat{\beta}$. As claimed by Lambert-Lacroix and Zwald[99] the centering affects $\hat{\beta}$ without sufficient information on how it works, due to the fact that the loss is no more the squared loss.

4.7 Conclusion

Variables selection methods having grouping effects and oracle ones are important for both theoretical and practical purposes. Using Matlab software seems relatively easy due to CVX package but computation of regularization paths can be

more difficult. On the other hand, actually the majority of regularization methods can be adapted under R software which makes easy computation of regularization paths, prediction performances, cross validation errors and risk functions.

Cinquième partie

Penalized Nonconcave Likelihood

Chapitre 5

A mixture of local and quadratic approximation variable selection algorithm in nonconcave penalized regression

We consider the problem of variable selection via penalized likelihood using nonconvex penalty functions. To maximize the non-differentiable and nonconcave objective function, an algorithm based on local linear approximation and which adopts a naturally sparse representation was recently proposed. However, although it has promising theoretical properties, it inherits some drawbacks of Lasso in high dimensional setting. To overcome these drawbacks, we propose an algorithm (MLLQA) for maximizing the penalized likelihood for a large class of nonconvex penalty functions. The convergence property of MLLQA and oracle property of one-step MLLQA estimator are established. Some simulations and application to a real data set are also presented.

5.1 Introduction

Variable selection plays an important role in statistical modeling. In genomics and preteomics studies, functional MRI, tumor classification and signal processing[Zou and Hastie[160]], it is very common that a large number p of candidate predictors are included in the model. However, when p is large, selection of a small number of predictors that contribute to the response leads often to a parsimonious model. It amounts to assuming that the true model has a sparse representation, i. e. some components of the parameter vector β of regression coefficients are exactly zero. In this setting, variable selection can improve on both estimation accuracy and interpretation. Our objective is to find the set $\mathcal{A}$ of the nonzero components of β and to estimate the true corresponding coefficients.

Recently, variable selection for high dimensional data has received a lot of attention. In the last decade interest has focused on penalized regression methods which implement both variable selection and coefficient shrinkage in a single procedure. The most well known of these procedures are Lasso[Tibshirani[138] ; Chen et al.[27]] and SCAD[Fan and Li[48]], which have good computational and statistical properties. The Lasso sparse estimates minimize the penalized least squares with L_1 penalty. While SCAD is presented as a unified approach, via nonconcave penalized likelihood which simultaneously performs variable selection and coefficient estimation. By a judicious choice of nonconvex penalty function, SCAD keeps many merits of the best subset selection and ridge regression. A similar nonconvex penalty MCP[Zhang[155]] has been proposed to overcome the Lasso bias [Radchenko and James[123]]. SCAD and MCP enjoy the oracle property, that is, the SCAD and MCP estimators can perform as well as the oracle if the penalization parameter is appropriately chosen.

The SCAD (and also MCP) penalty is nonconvex, and consequently it is hard to compute the solution of the optimization problem. To facilitate the use of Newton-Raphson algorithm, Fan and Li[48] proposed to approximate the nonconvex penalty by the local quadratic approximation (LQA). However, the drawback of this approximation is that the estimate of the regression coefficient has to end up being 0 once it reached 0 at any step of the LQA algorithm. So, the LQA algorithm inhe-

rits the drawback of backward stepwise variable selection : if a covariate is eliminated at any step in the LQA algorithm, it will necessarily be deleted from the final selected model. To alleviate this problem, Hunter and Li[74] proposed a minorize-maximize (MM) algorithm to compute the nonconcave penalized estimator. In this algorithm, the LQA approximation is improved with a small perturbation $\epsilon > 0$ to overcome the non-differentiability at zero.

On the other hand, Zou and Li[161] proposed a local linear approximation (LLA) algorithm that recasts the computation of nonconcave penalized likelihood problems into a sequence of penalized L_1-likelihood problems. The LLA algorithm enjoys some significant advantages over LQA and the perturbed LQA and produces a sparse estimates via continuous penalization. Moreover, the efficient LARS algorithm[Efron et al.[42]] for solving Lasso is used to compute the one-step LLA estimator. Consequently, the LLA algorithm will inherit similar limitations of Lasso in high dimensional setting : for $p > n$, it selects at most n variables before it puts all coefficients to zero and a second limitation is that group of variables can not enter in the same time with Lasso.

In this chapter, we propose an efficient one-step sparse estimation procedure in nonconcave penalized likelihood models, which is based on the mixture of local linear and quadratic approximation penalties (MLLQA). The new iterative MLLQA enjoys the advantages of both LLA and the perturbed LQA algorithms. As with LLA, MLLQA does not delete any small coefficient and it produces a sparse estimates via continuous penalization. Its convergence property is shown and the oracle property of one-step MLLQA estimator is established. Computationally, we take advantage of the efficient coordinate descent algorithm for Lasso penalized regression to compute the one-step MLLQA estimator in high dimension.

In Section 2, we present the local linear and quadratic approximation algorithms for SCAD penalty. In Section 3, we present our mixture of the local linear and quadratic approximation algorithm for SCAD penalty and study its various properties. In particular, we show that the MLLQA algorithm is an instance of MM algorithms[Hunter and Lange[72]] which converges to a stationary point of the likelihood solutions. In Section 4, we study the statistical properties of the one-step

MLLQA estimator. In particular, we show that the one-step MLLQA estimator enjoys the oracle property : consistence of selection and asymptotical normality. Numerical study is presented in Section 5 and we end with a brief discussion in Section 6.

5.2 Linear and quadratic approximation algorithms

In this Section, we consider the problem of variable selection in generalized linear model based on penalized likelihood approach. Two useful nonconvex penalties (SCAD and MCP) and various local linear and quadratic approximation algorithms for computing the maximum penalized likelihood are briefly presented.

5.2.1 Penalized likelihood with concave penalty

Let $(\mathbf{x}^i, y_i), i = 1, \ldots, n$ be n i. i. d. predictive-response observation pairs that are assumed to be a random sample where $\mathbf{x}^i \in I\!R^p$, and $y_i \in I\!R$. We assume that the observation y_i depend on $\mathbf{x}^i$ through a linear combination of $(\mathbf{x}^i)^t\beta, \beta \in I\!R^p$ and t stands for the transpose. That is, we assume that given $\mathbf{x}^i$, y_i has the density $f_i(g((\mathbf{x}^i)^t\beta), y_i)$ where g is a known link function. The conditional log-likelihood given $\mathbf{x}^i$ can be written as

$$\ell_i(\beta) = \ell_i(\beta, \phi) = \ell_i((\mathbf{x}^i)^t\beta, y_i, \phi) \tag{5.1}$$

where ϕ is a dispersion parameter which is assumed to be known. Our objective is the estimation of the parameter vector β and the identification of the subset model.

We consider the estimating of parameter vector β by maximizing the penalized log-likelihood

$$P\ell(\beta) = \sum_{i=1}^{n} \ell_i(\beta) - n\sum_{j=1}^{p} J_\lambda(|\beta_j|), \tag{5.2}$$

for a penalty function $J_\lambda(.)$. In the linear model case, the penalty $J_\lambda(|\beta_j|) = \lambda|\beta_j|^\gamma, \gamma \geq 0$ leads the bridge estimator[Frank and Friedman[56]]. In the same setting, when $\gamma = 1$ the penalty yields the Lasso estimator[Tibshirani[138]]. Fan and Li[48] proposed the SCAD penalty which is a continuously differentiable concave function defined by : $J_\lambda(0) = 0$ and for $|\beta_j| > 0$

$$J_\lambda^{'}(|\beta_j|) = \lambda \mathbf{I}(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{a-1}\mathbf{I}(|\beta_j| > \lambda), \tag{5.3}$$

where $(z)_+ = \max(z, 0)$, $a = 3.7, J_\lambda^{'}(r) \geq 0$ for $r > 0$ and $\mathbf{I}(|x| \leq \lambda) = 1$ if $|x| \leq \lambda$ and 0 otherwise. So, with the penalty (5.3) the penalized likelihood function (5.2) is a nonconcave function. The hard thresholding estimator corresponds to the penalty $J_\lambda(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2\mathbf{I}(|\beta_j| < \lambda)$. Moreover, a new nonconvex penalty MCP[Zhang[155]] derived from SCAD penalty can be easily understood by considering its derivative $J_{\lambda_1,a}^{'}(z) = \lambda_1(1 - |z|/(a\lambda_1))_+\text{sgn}(z)$ where $\text{sgn}(z) = -1, 0$ or 1 if $z < 0, = 0$ or > 0, respectively. It begins by applying the same rate of penalization as Lasso, but continuously relaxes that penalization, until $|z| > a\lambda_1$, until the rate of penalization drops to zero. In the literature, the penalty J_λ produces estimates with basic properties such that : unbiasedness, sparsity and continuity [Zhang[155]].

5.2.2 Local approximation algorithms

The function (5.2) is non-differentiable at the origin and nonconcave with respect to β. Suppose given an initial value $\beta^{(0)}$ that is close to the true value of β. To run easily the Newton-Raphson algorithm, Fan and Li[48] propose the following quadratic approximation

$$[J_\lambda(|\beta_j|)]^{'} = J_\lambda^{'}(|\beta_j|)\text{sign}(\beta_j) \approx \left\{ J_\lambda^{'}(|\beta_j^{(0)}|)/|\beta_j^{(0)}| \right\} \beta_j. \tag{5.4}$$

It leads to $J_\lambda(|\beta_j|) \approx J_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\left\{J'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\right\}(\beta_j^2 - {\beta_j^{(0)}}^2)$ for $\beta_j \approx \beta_j^{(0)}$. Then the iterative procedure LQA (Local Quadratic Approximation) solves

$$\beta^{(k+1)} = \text{argmax}_\beta \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{i=1}^{n} \frac{J'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|} \beta_j^2 \right\}. \tag{5.5}$$

When $\beta_j^{(k)}$ is close to zero, i. e. $|\beta_j^{(k)}| < \epsilon_0$ (pre-specified value), then $\hat{\beta}_j = 0$ and delete the jth component of $\mathbf{x}^i$ from the iteration. However, LQA has two drawbacks : the choice of ϵ_0 and similarity with the backward stepwise variable selection. Hunter and Li[74] studied the convergence property of the LQA algorithm. They described a minorize-maximize (MM) algorithm[Hunter and Lange[72]] to compute the penalized nonconcave likelihood estimator. In this algorithm, the latter approximation (5.4) is improved with a small perturbation τ_0 to handle the nondifferentiability at 0. This prevents the estimation from being trapped at 0. Then, the new iterative perturbed LQA algorithm solves

$$\beta^{(k+1)} = \text{argmax}_\beta \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} \frac{J'_\lambda(|\beta_j^{(k)}|)}{2(|\beta_j^{(k)}| + \tau_0)} \beta_j^2 \right\}, \tag{5.6}$$

for a fixed size perturbation τ_0. Hunter and Li[74] noted that a suitable choice of the size of τ_0 is essential for the good degree of sparsity of the solution as well as the speed of convergence. To overcome the limitations of LQA algorithms, Zou and Li[161] described a new algorithm based on local linear approximation (LLA) to the penalty function :

$$J_\lambda(|\beta_j|) \approx J_\lambda(|\beta_j^{(0)}|) + J'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|) \quad \text{for} \quad \beta_j \approx \beta_j^{(0)}. \tag{5.7}$$

Then, the iterative LLA procedure becomes

$$\beta^{(k+1)} = \text{argmax}_\beta \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} J'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right\}. \tag{5.8}$$

As with Lasso, the ℓ_1 penalty in the LLA algorithm naturally leads to a sparse representation of the estimates of the vector parameter β. So the LLA algorithm shares the good properties of Lasso in terms of computational efficiency, and therefore the efficient least angle regression shrinkage (LARS) algorithm can be used

to solve the equation (5.8). Zou and Li[161] confirm that the LLA algorithm is numerically stable, and so, the limitations of backward variable selection can be avoided in the LLA algorithm. However, the LLA algorithm inherits the drawbacks of Lasso in high dimensional setting in the presence of strong correlated variables. Moreover, when $p > n$ Lasso will selects at least n variables[Zou and Hastie[160]].

In the same setting, other new algorithms have been recently proposed to find a minimizer of the SCAD (or MCP) penalized likelihood function[Kim et al.[80] ; Kwon et al.[98] ; Schifano et al.[128]]. However, even if these new procedures provide simple and efficient computational algorithms, but they inherit the drawbacks of LARS in high dimension with correlated predictors.

5.3 Mixture of Local Linear and Quadratic Approximations

In this Section, we propose our MLLQA procedure and establish its convergence property.

5.3.1 MLLQA procedure

To overcome the drawbacks of the LLA algorithm in high dimensional setting with correlated variables, we propose a new unified algorithm which is a mixture of local linear and quadratic approximations. Indeed, from the approximation (5.4), we obtain that $J''_\lambda(|\beta_j|) \approx J'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|$ for $\beta_j \approx \beta_j^{(0)}$. Then, we consider the following local quadratic approximation of the penalty function

$$J_\lambda(|\beta_j|) \approx J_\lambda(|\beta_j^{(0)}|) + J'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|) + \frac{J'_\lambda(|\beta_j^{(0)}|)}{2|\beta_j^{(0)}|}(|\beta_j|^2 - |\beta_j^{(0)}|^2)$$

for $\beta_j \approx \beta_j^{(0)}$. Finally, the iterative mixture of local linear and quadratic approximations (MLLQA) procedure is defined by

$$\beta^{(k+1)} = \text{argmax}_\beta \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} J'_\lambda(|\beta_j^{(k)}|)|\beta_j| - \frac{n}{2} \sum_{j=1}^{p} \frac{J'_\lambda(|\beta_j^{(k)}|) + \tau_0}{|\beta_j^{(k)}| + \tau_0} |\beta_j|^2 \right\}. \tag{5.9}$$

The small perturbation τ_0 is introduced, in the numerator and denominator of the 3th term of (5.9), to handle the non-differentiability at 0 and in order to ensure convergence of our algorithm as we will see further, respectively. Consequently, the penalty in (5.9) is a combination of the weighted ℓ_1 and ℓ_2 norms. So, MLLQA is similar to the Elastic Net[Zou and Hastie[160]] which is more adapted to strong correlated variables in high dimensional linear regression setting. Thus, MLLQA inherits the good properties of LLA algorithm and corrects its difficulties in high dimension.

5.3.2 Convergence property of MLLQA algorithm

Following Schifano et al.[128], we assume that $J'_\lambda(0+) \in [C_\lambda^{-1}, C_\lambda]$ for some finite $C_\lambda > 0$. So, $J_\lambda(.)$ satisfies the condition $(P1)$ in Schifano et al.[128], which implies that $J'_\lambda(r) > 0$ for $r \in (0, K_\lambda)$, where $K_\lambda > 0$ may be finite or infinite. The positivity of the right derivative at zero ensures that $\sum_{j=1}^{p} J_\lambda(|\beta_j|)$ is not identically zero for $|\beta_j| > 0$. Denote

$$H(\beta|\beta^{(k)}) = \left\{ \sum_{i=1}^{n} \ell_i(\beta) - n \sum_{j=1}^{p} \Upsilon_{\tau_0}(\beta_j|\beta_j^{(k)}) \right\} \tag{5.10}$$

where

$$\Upsilon_{\tau_0}(\beta_j|\beta_j^{(k)}) = J_\lambda(|\beta_j^{(k)}|) + J'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) + \frac{J'_\lambda(|\beta_j^{(k)}|) + \tau_0}{2(|\beta_j^{(k)}| + \tau_0)}(|\beta_j|^2 - |\beta_j^{(k)}|^2).$$

The following theorem states that MLLQA algorithm is an instance of *MM* algorithms and has the ascent property.

Theorem 5.3.1 *For a differentiable concave penalty function $J_\lambda(.)$ on $[0,\infty)$, we have*

$$P\ell(\beta) \geq H(\beta|\beta^{(k)}) \quad \textit{and} \quad P\ell(\beta^{(k)}) = H(\beta^{(k)}|\beta^{(k)}). \tag{5.11}$$

Furthermore, the MLLQA has the ascent property, i.e, for all k=0,1,2,...

$$P\ell(\beta^{(k+1)}) \geq P\ell(\beta^{(k)}). \tag{5.12}$$

Proof of Theorem 5.3.1

We recall that

$$P\ell(\beta) - H(\beta|\beta^{(k)}) = n\{\sum_{i=1}^{n}(J_\lambda(|\beta_j^{(k)}|) + J'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) + \frac{J'_\lambda(|\beta_j^{(k)}|) + \tau_0}{2(|\beta_j^{(k)}| + \tau_0)}(|\beta_j|^2 - |\beta_j^{(k)}|^2) - J_\lambda(|\beta_j|))\}.$$

By the concavity of $J_\lambda(.)$, we have

$$J_\lambda(|\beta_j^{(k)}|) + J'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) - J_\lambda(|\beta_j|) \geq 0 \quad \text{for} \quad j = 1, ..., p.$$

When $\beta_j^{(k)} = 0$, we use the right derivative. The quadratic term

$$\frac{J'_\lambda(|\beta_j^{(k)}|) + \tau_0}{2(|\beta_j^{(k)}| + \tau_0)}(|\beta_j|^2 - |\beta_j^{(k)}|^2)$$

is always positive due to the approximation $(|\beta_j| - |\beta_j^{(k)}|)^2 \approx |\beta_j|^2 - |\beta_j^{(k)}|^2$ for $\beta_j \approx \beta_j^{(k)}$. We hence have $P\ell(\beta) \geq H(\beta|\beta^{(k)})$ and it's easy to verify that $P\ell(\beta^{(k)}) = H(\beta^{(k)}|\beta^{(k)})$.
The second inequality holds by the fact that $\beta^{(k+1)} = \text{argmax}_\beta H(\beta|\beta^{(k)})$, which leads to $P\ell(\beta^{(k+1)}) \geq H(\beta^{(k+1)}|\beta^{(k)}) \geq H(\beta^{(k)}|\beta^{(k)}) = P\ell(\beta^{(k)})$. ■

Let $M(\beta^{(k)})$ denote the map defined by the MLLQA algorithm from $\beta^{(k)}$ to $\beta^{(k+1)}$. Note that the penalty function has continuous first derivative and solving $\beta^{(k+1)}$ is a convex optimization problem, so M is a continuous map. We assume

that the set $\mathfrak{S}$ of stationary points for $\xi(\beta) = -P\ell(\beta)$ is both non empty and finite.

Theorem 5.3.2 *Let $\beta^{(k+1)} = argmin_{\beta} - H(\beta|\beta^{(k)})$. Then, using the condition (iii) of Theorem* 2.1 *in Schifano et al.[128], $\beta^{(k+1)}$ converges to a stationary point of $\xi(\beta) = -P\ell(\beta)$.*

Proof of Theorem 5.3.2.

We recall that

$$-P\ell(\beta) = -\ell(\beta) + n\sum_{j=1}^{p} J_{\lambda}(|\beta_j|).$$

Let $\xi(\beta) = -P\ell(\beta) = g(\beta) + nS_{\lambda}(\beta)$. The negative log-likelihood function $g(.)$ is strictly convex and $S_{\lambda}(\beta) = \sum_{j=1}^{p} J_{\lambda}(|\beta_j|)$ satisfies the assumptions needed for Theorem 2.1 of Schifano et al.[128]. On the other hand, let
$U_{\lambda}(\beta|\beta^{(k)}) = \sum_{j=1}^{p} \tilde{u}_{\lambda}(|\beta_j|, |\beta_j^{(k)}|)$, where

$$\tilde{u}_{\lambda}(r,s) = J_{\lambda}(s) + J_{\lambda}^{'}(s)(r-s) + \frac{J_{\lambda}^{'}(s)+\tau_0}{2(s+\tau_0)}(r-s)^2$$

for $r \approx s$, with r and s are taken in a compact set of $(0,\infty)$. Then, we have $U_{\lambda}(\beta|\beta^{(k)}) - S_{\lambda}(\beta^{(k)}) > 0$. This strict inequality is obtained by the concavity of $J_{\lambda}(.)$ on $(0,\infty)$ which leads to $J_{\lambda}(r) \leq J_{\lambda}(s) + J_{\lambda}^{'}(s)(r-s)$ for each $r, s > 0$ and the fact that $(J_{\lambda}^{'}(s)+\tau_0)(r-s)^2/2(s+\tau_0) > 0$ for each $r \approx s$. Hence, $-H(\beta|\beta^{(k)})$ strictly locally majorizes $\xi(\beta)$ in an open neighborhood containing $\beta^{(k+1)}$. We mention here that as far as strictly local majorization holds at each iteration, we don't need to use the function $h(\boldsymbol{\beta}, \alpha)$ used in Theorem 2.1 of Schifano et al.[128] to majorize $g(\beta)$. In fact, one can consider that, $\sum_{j=1}^{p} \frac{J_{\lambda}^{'}(|\beta_j^{(k)}|)+\tau_0}{2(|\beta_j^{(k)}|+\tau_0)}(|\beta_j| - |\beta_j^{(k)}|)^2 > 0$ instead of $h(\beta, \beta^{(k)})$. This is the reason of introducing the perturbation to the numerator, as previously mentionned. Finally, strict convexity of $-H(\beta|\beta^{(k)})$ in β leads to unique minimum $\beta^{(k+1)}$. With locally strict majorization, we conclude that the MM algorithm derived from $-H(\beta|\beta^{(k)})$ converges to a stationary point of $\xi(\beta)$.
■

5.4 Statistical study of one-step MLLQA estimator

In this Section, we establish the oracle property of the one-step MLLQA estimator in the case of linear regression models based on the penalized least squares and in the most general penalized likelihood setting.

5.4.1 Linear regression case

In the case of linear models, the one step MLLQA estimator $\hat{\beta}$ verifies

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \frac{1}{2} \parallel y - X\beta \parallel^2 + n \sum_{j=1}^{p} J'_{\lambda}(|\beta_j^{(0)}|)|\beta_j| + \frac{n}{2} \sum_{j=1}^{p} \frac{J'_{\lambda}(|\beta_j^{(0)}|) + \tau_0}{|\beta_j^{(0)}| + \tau_0} |\beta_j|^2 \right\}. \tag{5.13}$$

We remark that solving (5.13) is similar to elastic net problem. In fact unlike the elastic net based on the choice of two regularization parameters, here we deal with a single parameter λ due to the behavior of the SCAD (and also MCP) penalty. It is easy to see that solving problem (5.13) is equivalent to find

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \frac{1}{2} \parallel y^* - X^*\beta \parallel^2 + n \sum_{j=1}^{p} J'_{\lambda}(|\beta_j^{(0)}|)|\beta_j| \right\}, \tag{5.14}$$

where

$$y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}, X^* = \begin{pmatrix} X \\ S \end{pmatrix}$$

and S is a diagonal matrix with $S_{jj} = \sqrt{n \frac{J'_{\lambda}(|\beta_j^{(0)}|) + \tau_0}{|\beta_j^{(0)}| + \tau_0}}$, $j = 1, ...p$.

Thus, maximizing $P\ell(\beta)$, as defined in (5.2) via the one-step MLLQA algorithm, is equivalent to use one-step LLA on an augmented data.
It's now interesting to see if $\hat{\beta}$ enjoys oracle properties. So, we assume the two following regularity conditions (A.1) and (A.2) used by Zou and Li[161].

(A.1). $y_i = \mathbf{x}_i^t \beta_0 + \epsilon_i$, where $\epsilon_1, ..., \epsilon_n$ are independent and identically distributed

random variables with mean 0 and variance σ^2,

(A.2).

$$\frac{1}{n}X^tX \to C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where C is a positive definite matrix, $\beta_0 = (\beta_{01}, ..., \beta_{0p})^t = (\beta_{10}^t, \beta_{20}^t)^t$ and $\beta_{20} = 0$.

Theorem 5.4.1 *Assume that the previous assumptions (A.1) and (A.2) are satisfied and that if* $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $\sqrt{n}\tau_0 \to 0$ *as* $n \to \infty$, *then with probability tending to one we obtain that :*

(a) Sparsity : $\hat{\beta}_2 = 0$.

(b) Asymptotic normality : $\sqrt{n}(\hat{\beta}_1 - \beta_{10}) \to \mathrm{N}(0, \sigma^2 C_{11}^{-1})$.

We omit the proof of Theorem 4.1 since it is similar to that of Theorem 4.2 defined in Section 4.2.

5.4.2 Generalized linear model case

In the penalized likelihood setting, we assume that the log-likelihood function $\ell(\beta) = \sum_{i=1}^n \ell_i(\beta)$ is twice differentiable according to β. For a given initial value $\beta^{(0)}$, we can use the following local approximation :

$$\ell(\beta) \approx \ell(\beta^{(0)}) + \nabla\ell(\beta^{(0)})^t(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})^t\nabla^2\ell(\beta^{(0)})(\beta - \beta^{(0)}). \quad (5.15)$$

Starting from $\beta^{(0)} = \hat{\beta}(\text{mle})$ the maximum likelihood estimator, with $\nabla\ell(\beta^{(0)}) = 0$, then $\hat{\beta}$ verifies

$$\begin{aligned}\hat{\beta} = \text{argmin}_\beta\{&\frac{1}{2}(\beta-\beta^{(0)})^t[-\nabla^2\ell(\beta^{(0)})](\beta-\beta^{(0)})\\ &+n\sum_{j=1}^{p}J'_\lambda(|\beta_j^{(0)}|)|\beta_j| + \frac{n}{2}\sum_{j=1}^{p}\frac{J'_\lambda(|\beta_j^{(0)}|)+\tau_0}{|\beta_j^{(0)}|+\tau_0}|\beta_j|^2\},\end{aligned} \tag{5.16}$$

which can be written as

$$\begin{aligned}\hat{\beta} = \text{argmin}_\beta\Bigg\{&\frac{1}{2}(\beta-\beta^{(0)})^t[-\nabla^2\ell(\beta^{(0)})](\beta-\beta^{(0)})\\ &+\frac{n}{2}\beta^t\mathbf{Q}_{\boldsymbol{\tau_0}}\beta + n\sum_{j=1}^{p}J'_\lambda(|\beta_j^{(0)}|)|\beta_j|\Bigg\}\end{aligned} \tag{5.17}$$

where $\mathbf{Q}_{\boldsymbol{\tau_0}} = \text{diag}(\mathbf{Q}_{\boldsymbol{\tau_0}11},\ldots,\mathbf{Q}_{\boldsymbol{\tau_0}pp})$ and $\mathbf{Q}_{\boldsymbol{\tau_0}jj} = \frac{J'_\lambda(|\beta_j^{(0)}|)+\tau_0}{|\beta_j^{(0)}|+\tau_0}$.

We denote $I(\beta_0)$ the $p\times p$ Fisher information matrix and the submatrix $I_1(\beta_{10}) = I(\beta_{10},0)$, the Fisher information knowing $\beta_{20} = 0$. As advocated by Lehmann and Casella[102], under some regularity conditions, we have $n^{-1}\nabla^2\ell(\hat{\beta}(mle)) \to_P -I(\beta_0)$, and $\sqrt{n}(\beta_0 - \hat{\beta}(mle)) \to_D W = N(0, I^{-1}(\beta_0))$. The following theorem assesses the oracle property of the one step MLLQA estimator for penalized likelihood.

Theorem 5.4.2 *Under the previous assumptions, if* $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $\sqrt{n}\tau_0 \to 0$ *as* $n \to \infty$, *then with probability tending to one,* β *satisfies :*

(a) Sparsity : $\hat{\beta}_2 = 0$.

(b) Asymptotic normality : $\sqrt{n}(\hat{\beta}_1 - \beta_{10}) \to \text{N}(0, I_1^{-1}(\beta_{10}))$.

According to the two previous theorems, we see that oracle properties require for the penalty function to be twice differentiable, λ_n is chosen as in Theorem 2 of Fan and Li[48] and we have used a supplementary condition $\sqrt{n}\tau_0 \to 0$ as $n \to \infty$.

This is justified by the fact that our algorithm is based on linear and quadratic approximation of the SCAD penalty function, which uses the second order derivatives. We recall that results for the one step LLA require less regularity conditions than results given in Fan and Li[48].

Remark 5.4.1 *In their earlier work, Fan and Li[48] showed that continuity for the nonconcave penalized likelihood estimates is guaranteed by the condition that the minimum of the function* $|\theta| + J'_\lambda(|\theta|)$ *must be attained at 0. Since our MLLQA one step estimator is based on a mixture of linear and quadratic approximation of the penalty function, continuity of* $\hat{\beta}$ *only requires that* $J'_\lambda(|\theta|)$ *is continuous for* $|\theta| > 0$*, as with the one step LLA estimator [Zou and Li[161]]. Since the computation of the sparse one-step MLLQA estimator is based on the LLA algorithm which uses an* L_1 *penalized criterion. The quadratic term of the approximation only contributes in the non penalized part of the objective function.*

Proof of Theorem 5.4.2.

We only demonstrate oracle properties for the penalized likelihood estimates. The proof for linear regression model is similar. The following proof is based on a slightly modified version of Theorem 5 in [Zou and Li[161]]. Let us define

$$\begin{aligned} K_n(u) &= \frac{1}{2}(\frac{u}{\sqrt{n}} + \beta_0 - \beta^{(0)})^t[-\nabla^2\ell(\beta^{(0)})](\frac{u}{\sqrt{n}} + \beta_0 - \beta^{(0)}) + \\ & \quad n\sum_{j=1}^{p} J'_{\lambda_n}(|\beta_j^{(0)}|)|\beta_{0j} + \frac{u_j}{\sqrt{n}}| + \frac{n}{2}\sum_{j=1}^{p}\frac{J'_{\lambda_n}(|\beta_j^{(0)}|) + \tau_0}{|\beta_j^{(0)}| + \tau_0}(\beta_{0j} + \frac{u_j}{\sqrt{n}})^2. \end{aligned}$$

Then, we have

$$\begin{aligned} K_n(u) - K_n(0) &= \frac{1}{2}(\frac{u^t}{\sqrt{n}}[-\nabla^2\ell(\beta^{(0)})]\frac{u}{\sqrt{n}} + (\beta_0 - \beta^{(0)})^t[-\nabla^2\ell(\beta^{(0)})]\frac{u}{\sqrt{n}} \\ &+ n\sum_{j=1}^{p} J'_{\lambda_n}(|\beta_j^{(0)}|)(|\beta_{0j} + \frac{u_j}{\sqrt{n}}| - |\beta_{0j}|) \\ &+ \frac{n}{2}\sum_{j=1}^{p}\frac{J'_{\lambda_n}(|\beta_j^{(0)}|) + \tau_0}{|\beta_j^{(0)}| + \tau_0}((\beta_{0j} + \frac{u_j}{\sqrt{n}})^2 - \beta_{0j}^2) \\ &\equiv T_1 + T_2 + T_3 + T_4. \end{aligned}$$

Moreover, it's easy to see that

$$\hat{u}(n) = \text{argmin}_u[K_n(u) - K_n(0)]$$

leads to $\hat{\beta} = \beta_0 + \frac{\hat{u}(n)}{\sqrt{n}}$ with $\hat{\beta}$ the one step MLLQA estimator.

By Slutsky's theorem and using the same argument as in the proof of Theorem 5 in Zou and Li[161], it follows that

$$T_1 = \frac{1}{2}(\frac{u^t}{\sqrt{n}}[-\nabla^2\ell(\beta^{(0)})]\frac{u}{\sqrt{n}} \to_P \frac{1}{2}u^t I(\beta_0)u,$$

$$T_2 = (\beta_0 - \beta^{(0)})^t[-\nabla^2\ell(\beta^{(0)})]\frac{u}{\sqrt{n}} = \sqrt{n}(\beta_0 - \beta^{(0)})^t[\frac{-\nabla^2\ell(\beta^{(0)})}{n}]u \to_D -W^t I(\beta_0)u$$

and

$$T_3 \to_P \begin{cases} 0 & \text{if } u_{20} = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{5.18}$$

The last term can be written as

$$T_4 = \frac{1}{2}\sum_{j=1}^{p}\sqrt{n}\frac{J'_{\lambda_n}(|\beta_j^{(0)}|) + \tau_0}{|\beta_j^{(0)}| + \tau_0}(\frac{(\beta_{0j} + \frac{u_j}{\sqrt{n}})^2 - \beta_{0j}^2}{\frac{1}{\sqrt{n}}}) = \frac{1}{2}\sum_{j=1}^{p} T_{4j}.$$

First, it can be seen that

$$\frac{(\beta_{0j} + \frac{u_j}{\sqrt{n}})^2 - \beta_{0j}^2}{\frac{1}{\sqrt{n}}} \to 2u_j\beta_{0j}I(\beta_{0j} \neq 0) + \frac{u_j^2}{\sqrt{n}}I(\beta_{0j} = 0).$$

We now examine the behavior of $\sqrt{n}(J'_{\lambda_n}(|\beta_j^{(0)}|)+\tau_0)/(|\beta_j^{(0)}|+\tau_0)$.

When $\beta_{0j} \neq 0$, we have $|\beta_j^{(0)}| \to_P |\beta_{0j}| > 0$ and $|\beta_j^{(0)}|+\tau_0$ remains bounded away from zero. Moreover, using the fact that $J'_{\lambda_n}(\theta) = 0$ if $\theta > a\lambda_n$ and if $\sqrt{n}\tau_0 \to 0$ and $\lambda_n \to 0$ as $n \to \infty$ we conclude that $\sqrt{n}\frac{J'_{\lambda_n}(|\beta_j^{(0)}|)+\tau_0}{|\beta_j^{(0)}|+\tau_0}u_j\beta_{0j} \to_P 0$.

When $\beta_{0j} = 0$, $T_{4j} = 0$ if $u_j = 0$ else we have $|\beta_j^{(0)}| = O_P(1/\sqrt{n})$. On the other hand, $J'_{\lambda_n}(\theta) = \lambda_n$ for all $0 < \theta < \lambda_n$ and $\sqrt{n}\lambda_n \to \infty$ and $\sqrt{n}\tau_0 \to 0$ as $n \to \infty$, then

$$T_{4j} = (\sqrt{n}\lambda_n + \sqrt{n}\tau_0)\frac{u_j^2}{\sqrt{n}(|\beta_j^{(0)}|+\tau_0)} \to_P \infty.$$

So

$$T_4 \to_P \begin{cases} 0 & \text{if } u_{20} = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{5.19}$$

By taking $W = (W_{10}^t, W_{20}^t)$, from T_1, T_2, T_3 and T_4 convergence results we conclude that for each fixed u,

$$K_n(u) - K_n(0) \to_d K(u) \equiv \begin{cases} \frac{1}{2}u_{10}^t I_1(\beta_{10})u_{10} - W_{10}^t u_{10} & \text{if} \quad u_{20} = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{5.20}$$

The unique minimum of $K(u)$ is $u = (u_{10} = I_1^{-1}(\beta_{10})W_{10}, u_{20} = 0)$. Since $K_n(u) - K_n(0)$ is a convex function of u, we conclude by epiconvergence as in Zou and Li[161], that

$$\hat{u}(n)_{10} \to_d I_1^{-1}(\beta_{10})W_{10} \tag{5.21}$$

$$\hat{u}(n)_{20} \to_d 0. \tag{5.22}$$

Considering $W_{10} = N(0, I_1(\beta_{10}))$, (5.21) is equivalent to $\sqrt{n}(\hat{\beta}_1 - \beta_{10}) \to N(0, I_1^{-1}(\beta_{10}))$ and (5.22) implies that $\sqrt{n}\hat{\beta}_2 \to_P 0$.

Now we have to show that $P(\hat{\beta}_2 = 0) \to 1$, which is stronger statement than (5.22). We just have to show that if $\beta_{0j} = 0$, then $P(\hat{\beta}_j \neq 0) \to 0$. Assume $\hat{\beta}j \neq 0$, by (KKT) conditions of (5.16), we must have

$$\frac{1}{\sqrt{n}}([-\nabla^2\ell(\beta^{(0)})](\hat{\beta} - \beta^{(0)}))_j = \sqrt{n}\lambda(J'_{\lambda_n}(|\beta_j^{(0)}|) + \frac{J'_{\lambda_n}(|\beta_j^{(0)}|)+\tau_0}{|\beta_j^{(0)}|+\tau_0}|\hat{\beta}_j|). \tag{5.23}$$

Since $\frac{J'_{\lambda_n}(|\beta_j^{(0)}|)+\tau_0}{|\beta_j^{(0)}|+\tau_0}|\hat{\beta}_j| \geq 0$, (5.23) leads to

$$\frac{1}{\sqrt{n}}([-\nabla^2\ell(\beta^{(0)})](\hat{\beta}-\beta^{(0)}))_j \geq \sqrt{n}\lambda J'_{\lambda_n}(|\beta_j^{(0)}|).$$

When $\beta_{0j}=0$, $\lambda\sqrt{n}J'_{\lambda_n}(|\beta_j^{(0)}|)$ goes to ∞ in probability. Moreover, the left hand side of (5.23) can be written as $([\frac{-\nabla^2\ell(\beta^{(0)})}{n}]\sqrt{n}(\hat{\beta}-\beta_0))_j-([\frac{-\nabla^2\ell(\beta^{(0)})}{n}]\sqrt{n}(\beta^{(0)}-\beta_0))_j$. From (5.21) and (5.22), the first term converges in law to some normal, and so does the second term. Thus

$$P(\hat{\beta}_j \neq 0) \leq P(\text{KKT condition} \quad (5.23) \quad \text{holds}) \to 0. \qquad \blacksquare$$

5.5 Numerical experiments

In this section, we study the performances of our one step MLLQA and its competitors on simulated and real data sets. As competitors we consider one step LLA, LQA, Perturbed LQA (PLQA), LASSO and ENET. For the choice of the initial parameter vector $\boldsymbol{\beta}^{(0)}$, we use the Maximum Likelihood Estimator (MLE) in the classical case when $n > p$, otherwise we consider the L_2-Penalized MLE. In all of the experiments, computations are conducted using R software.

5.5.1 Simulation study

In all of the examples the correlation matrix (Σ) is defined by $\Sigma_{ij} = \rho^{|i-j|}$, $1 \leq i, j \leq n$ and $\rho \in \{0, 0.5, 0.75, 0.9\}$. For illustration, we consider the setting of linear regression and logistic regression models. In all of the simulated examples, the perturbation $\tau_0 = 10^{-6}$ for one step MLLQA, and the same value is considered for the PLQA with ϵ chosen according to equation (3.12) in Hunter and Li[74]. The tuning parameters are selected by ten-fold cross validation. The statistics considered here are the prediction error (MSE_y), the false positive (FP), which is the number of true zero coefficients incorrectly estimated as nonzero, and false nega-

Method	MSE_y	Number of Zeros FP	FN
		$\rho = 0$	
LASSO	1.144(0.018)	2.50(0.17)	0(-)
LLA	1.139(0.018)	1.91(0.16)	0(-)
LQA	1.156(0.017)	4.26(0.09)	0(-)
PLQA	1.152(0.019)	4.24(0.09)	0(-)
MLLQA	**1.132**(0.018)	**1.77**(0.16)	0(-)
ENET	1.142(0.019)	2.49(0.15)	0(-)
		$\rho = 0.5$	
LASSO	1.119(0.015)	2.36(0.15)	0(-)
LLA	1.112(0.016)	1.67(0.15)	0(-)
LQA	1.118(0.015)	3.90(0.10)	0(-)
PLQA	1.114(0.016)	3.84(0.10)	0(-)
MLLQA	**1.108**(0.016)	**1.58**(0.15)	0(-)
ENET	1.119(0.015)	2.38(0.14)	0(-)

TABLE 5.1 – Simulation results for Example 1.

tive (FN), which is the number of true nonzero coefficients incorrectly estimated to zero value. All simulations are performed 100 times and standard error related to an estimation, presented in brackets, is obtained by 500 bootstrap resampling.

Linear regression case

Example 1 : $n > p$.

We consider the following model $y_i = \boldsymbol{x}_i^t\beta + \epsilon_i$ where $\boldsymbol{x}_i = (x_{i1}, ..., x_{i8})$ is a multinormal vector with correlation matrix (Σ), $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\epsilon_i \sim N(0, 1)$ $1 \leq i \leq n$. The sample size n is set to be $n = 60$ for training sample and the test sample has size $ntest = 2\times n$. The initial vector $\beta^{(0)}$ corresponds to the Ordinary Least Square estimate. Table 5.1 summarizes the results for $\rho \in \{0, 0.5\}$.

As presented in Table 5.1, one step MLLQA performs slightly better in terms of MSE_y and FP followed by the one step LLA, where the difference in terms of FP rate with the four other methods is clear. While the methods are all comparable in terms of FN. In addition, the performance of LASSO and ENET are relatively

similar but slightly better than the performance of LQA and PLQA.

Example 2 : $p > n$.

We consider the high dimensional model $y = \boldsymbol{x}^t\beta + \epsilon$ where ϵ follows a standard gaussian distribution and the predictors are generated as in the previous example. The number of predictors is fixed to $p = 120$ and the vector β has nine nonzero components of different signs and the remaining components set to be equal to zero, so $\beta = (\underbrace{3, 3, -1/3, ..., 3, 3, -1/3}_{9}, \underbrace{0, ..., 0}_{111})$. Sample sizes considered for both training and test sets are $n = p/3 = 40$ and $n = p/2 = 60$. The results are summarized in Table 5.2 and Table 5.3 for $\rho \in \{0.5, 0.75, 0.9\}$. For this example, we use a ridge regression estimate for $\beta^{(0)}$ as initial value instead of the classical OLS estimate.

As can be seen from Table 5.2 for $\rho = 0.5$, ENET and LASSO performs slightly better than MLLQA in terms of MSE_y, respectively. Moreover, MLLQA does better in terms of FP and FN rates with large difference in terms of FP. However, MLLQA performs better than all other methods in terms of MSE_y and FP for $\rho \in \{0.75, 0.9\}$. It is followed by LLA in terms of FP and ENET in terms of MSE_y. Unlike Example 1, the difference between MLLQA and LLA appears to be slightly greater in terms of errors and it is relatively low in terms of FP. In terms of FN, PLQA and LQA perform surprisingly better than all other methods, but their performance is bad in all other three measures.

Table 5.3 displays the results for the case where the sample size $n = 60$ and $p = 120$. In all terms and for all ρ values, it can be seen that the one step MLLQA is better than its competitors, except in terms of FN rate where PLQA and LQA are slightly better than MLLQA. Furthermore, we note that when the sample size increases the gap between LLA and MLLQA is increased in terms of prediction and estimation errors particularly for $\rho = 0.9$. Globally, it can be seen that the FN rates of all methods decrease when the sample size increases.

Example 3 : $p > n$.

We consider the high dimensional model $y = \boldsymbol{x}^t\beta + \sigma \times \epsilon$ where ϵ follows a standard gaussian distribution and the predictors are generated as in the previous examples with correlation matrix (Σ) defined by $\Sigma_{ij} = 0.9^{|i-j|}$, $1 \leq i, j \leq n$. The number of predictors is fixed to $p = 120$ with the nine first components of β with nonzero values and the remaining components set to be equal to zero, so that $\beta = (\underbrace{10, ..., 10}_{9}, \underbrace{0, ..., 0}_{111})$. Sample size considered is $n = p/2 = 60$ for both training and test sets and $\sigma \in \{3, 5\}$.

Results in Table 5.4 show that one step MLLQA is also the winner in terms of FP for $\sigma \in \{3, 5\}$ followed by one step LLA. Considering MSE_y, ENET performs better for $\sigma = 3$ followed respectively by one step MLLQA and LASSO ; for $\sigma = 5$ our one step MLLQA is the winner followed by ENET and LASSO. Globally for this example, one step LLA, LQA and PLQA seem not having good performances except for FN where PLQA and LQA perform better.

Method	MSE_y	Number of Zeros FP	FN
		$\rho = 0.5$	
LASSO	2.35(0.10)	16.26(0.56)	2.62(0.06)
LLA	3.22(0.70)	7.18(0.60)	2.41(0.12)
LQA	22.02(2.92)	18.72(1.66)	1.68(0.18)
PLQA	20.96(2.95)	20.60(1.75)	**1.56**(0.16)
MLLQA	2.40(0.17)	**4.60**(0.57)	2.30(0.09)
ENET	**2.33**(0.08)	12.95(0.52)	2.54(0.06)
		$\rho = 0.75$	
LASSO	1.91(0.06)	13.78(0.47)	2.45(0.07)
LLA	2.51(0.48)	2.99(0.45)	2.19(0.12)
LQA	14.44(2.54)	19.94(1.52)	1.47(0.13)
PLQA	12.55(2.13)	21.03(1.87)	**1.26**(0.10)
MLLQA	**1.74**(0.06)	**1.33**(0.23)	1.85(0.11)
ENET	1.82(0.06)	9.20(0.57)	2.20(0.08)
		$\rho = 0.9$	
LASSO	1.71(0.05)	14.87(0.51)	2.15(0.07)
LLA	4.00(1.01)	1.38(0.24)	2.05(0.17)
LQA	5.68(0.59)	26.26(1.73)	1.29(0.09)
PLQA	5.05(0.57)	28.88(1.61)	**1.08**(0.09)
MLLQA	**1.47**(0.05)	**0.44**(0.10)	1.37(0.09)
ENET	1.62(0.05)	6.97(0.48)	1.46(0.08)

TABLE 5.2 – Simulation results for Example 2 : $n = 40$, $p = 3 * n$.

Method	MSE_y	Number of Zeros FP	FN
		$\rho = 0.5$	
LASSO	1.87(0.05)	23.17(0.81)	2.35(0.07)
LLA	3.19(0.80)	5.98(0.77)	2.32(0.14)
LQA	13.81(2.13)	21.07(1.89)	**1.17**(0.15)
PLQA	14.57(2.36)	18.12(1.62)	1.34(0.13)
MLLQA	**1.57**(0.05)	**3.76**(0.58)	2.08(0.11)
ENET	1.78(0.05)	14.40(0.70)	2.60(0.06)
		$\rho = 0.75$	
LASSO	1.72(0.04)	21.36(0.62)	2.47(0.06)
LLA	2.94(0.69)	3.21(0.54)	1.97(0.14)
LQA	5.87(0.74)	18.44(1.74)	**1.20**(0.10)
PLQA	6.12(0.93)	21.38(2.10)	1.25(0.10)
MLLQA	**1.34**(0.03)	**1.18**(0.24)	1.73(0.10)
ENET	1.57(0.04)	11.45(0.64)	2.31(0.07)
		$\rho = 0.9$	
LASSO	1.54(0.03)	21.22(0.53)	2.28(0.07)
LLA	2.49(0.45)	1.38(0.28)	1.61(0.14)
LQA	3.19(0.27)	26.93(2.12)	1.46(0.09)
PLQA	3.18(0.26)	27.69(2.06)	**1.44**(0.09)
MLLQA	**1.31**(0.04)	**0.46**(0.12)	1.52(0.10)
ENET	1.40(0.03)	12.71(0.37)	1.69(0.08)

TABLE 5.3 – Simulation results for Example 2 : $n = 60$, $p = 2 * n$.

Logistic regression

The response variable is generated from a binomial distribution with parameter

$$\Pi = P(y = 1|X = \mathbf{x}) = e^{\mathbf{x}^t\beta}/(1 + e^{\mathbf{x}^t\beta}).$$

The predictors $\mathbf{x}$ are generated following the similar model as in examples before. On the other hand, the sample size considered is $n = 200$ for training and test sets. The parameter vector considered is $\beta = (3, 1.5, 0, 0, 2, 0, 0, -1)$. The misclassification error rate, FP and FN rates are the three measures used for comparing the performance of different methods on test data set. In this example, best results are obtained for weighted methods (LLA, LQA, PLQA, MLLQA) by

Method	MSE_y	Number of Zeros FP	FN
		$\sigma = 3$	
LASSO	14.75(0.36)	25.90(0.48)	2.36(0.07)
LLA	18.19(4.02)	4.15(0.13)	3.05(0.06)
LQA	33.28(7.96)	47.39(1.61)	1.39(0.12)
PLQA	32.69(8.67)	47.43(1.80)	**1.39**(0.11)
MLLQA	14.15(3.17)	**3.57**(0.06)	3.04(0.04)
ENET	**12.14**(0.32)	12.03(0.73)	2.70(0.05)
		$\sigma = 5$	
LASSO	40.47(0.93)	24.07(0.45)	2.42(0.07)
LLA	49.33(5.70)	4.49(0.29)	3.14(0.09)
LQA	100.23(8.16)	33.98(1.86)	1.21(0.11)
PLQA	92.82(8.75)	35.50(1.77)	**1.18**(0.14)
MLLQA	**32.81**(0.67)	**3.59**(0.10)	2.98(0.02)
ENET	34.80(0.76)	11.02(0.93)	2.73(0.05)

TABLE 5.4 – Simulation results for Example 3.

Method	Misclassification error	FP	FN
LASSO	24.14(0.53)	2.48(0.10)	0(-)
LLA	24.17(0.55)	**2.47**(0.11)	0.01(0.01)
LQA	27.12(0.65)	2.95(0.07)	0.02(0.01)
PLQA	27.12(0.68)	2.64(0.08)	0.02(0.01)
MLLQA	**23.90**(0.55)	2.75(0.10)	0(-)
ENET	24.26(0.54)	3.19(0.08)	0(-)

TABLE 5.5 – Simulation results for logistic regression model.

using $\exp(\hat{\lambda})$ rather than $\hat{\lambda}$ when evaluating weights related to each coefficient, where $\hat{\lambda}$ is the optimal tuning parameter selected by tenfold cross validation.

For this example, the results in Table 5.5 show that the one step MLLQA is the winner in terms of misclassification error rate followed by the LASSO, while one step LLA is the winner in terms of false positive. Moreover, LASSO, the one step MLLQA and ENET have a zero false negative rate, which is not the case for one step LLA, LQA and PLQA.

5.5.2 Real data experiments : p»n

We propose to test the performance of our method on ARCENE dataset which is one of the five NIPS 2003 feature selection challenge data sets. The data set is obtained from UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. For this two-class classification problem with continuous input variables, the aim is to distinguish cancer versus normal patterns from mass-spectrometric data. The samples include patients with cancer (ovarian or prostate cancer), and healthy or control patients. ARCENE was obtained by merging three mass-spectrometry data sets to obtain enough training and test data for a benchmark. The original features indicate the abundance of proteins in human sera having a given mass value. Based on those features one must separate cancer patients from healthy patients. The sample size considered here is $n = 100$ for training and test sets with $p = 10000$ features. Before performing variables selection, we firstly select $p = 500, 1000, 1500$ and 2000 variables with the smallest p-value and compare our methods on the selected subset of variables. This pre-selection step is a common approach used in many papers with ultra high dimensional data sets.

According to results in Table 5.6, the one step MLLQA is very competitive in terms of misclassification error rate and has a tendency to select small number of variables in all settings (followed by LLA or LASSO). While, ENET is competitive in terms of misclassification error, but it has a tendency to introduce a lot of variables. The LQA and PLQA perform badly in terms of misclassification rate and select more variables than all other methods (except in $p = 500$).

About the computation speed, we take advantage of the diagonal structure of the augmented data when using MLLQA ; but the computation speed would not remain the same as with the LLA. This can be partly explain by the fact that for example in the case of the elastic net, the diagonal elements are the same which is not the case for the MLLQA in all situations. Table 5.7 gives computational times of methods on arcene data set.

Method	Misclassification error	♯ Selected variables
	$p = 500$	
LASSO	36	18
LLA	36	18
LQA	40	65
PLQA	38	123
MLLQA	**33**	**11**
ENET	37	68
	$p = 1000$	
LASSO	29	22
LLA	29	**20**
LQA	44	119
PLQA	44	119
MLLQA	**27**	21
ENET	28	108
	$p = 1500$	
LASSO	30	43
LLA	31	41
LQA	44	174
PLQA	44	174
MLLQA	29	**36**
ENET	**27**	114
	$p = 2000$	
LASSO	**30**	47
LLA	32	50
LQA	45	150
PLQA	44	244
MLLQA	31	**46**
ENET	32	122

TABLE 5.6 – Results for ARCENE dataset.

Methods	Computation time (in seconds)	
	$p = 500$	$p = 1000$
LASSO	0.62	0.85
LLA	0.57	1.01
LQA	13.61	16.55
PLQA	14.02	18.44
MLLQA	1.08	2.19
ENET	0.83	0.98

TABLE 5.7 – Computation times on arcene data set.

5.6 Conclusion

We have proposed an efficient one-step sparse estimation procedure in nonconcave penalized likelihood models, which is based on the mixture of local linear and quadratic approximation penalties (MLLQA). The new iterative MLLQA enjoys the advantages of both LLA and the perturbed LQA algorithms. Its convergence property is shown. As with LLA, MLLQA does not delete any small coefficient and it produces a sparse estimates via continuous penalization. Computationally, we take advantage of the efficient coordinate descent algorithm for LASSO penalized regression to compute the one-step MLLQA estimator. Moreover, the oracle property of one-step MLLQA estimator is established. Empirically, the proposed method provides smaller models with better prediction accuracy in comparison with its principal competitors.

Conclusion générale et perspectives

Au terme de ce travail, les constats suivants peuvent être faits. Les méthodes basées sur les poids adaptatifs ou aléatoires semblent diminuer le nombre de variables bruitées. D'autre part, le bootstrap contribue à la sélection des variables pertinentes tout en contrôlant le nombre de ces variables bruitées. Un type de données dites éparses peuvent conduire à un gain de temps énorme notamment en grande dimension. Cependant, ceci pose un problème si l'on doit effectuer l'échantillonnage ou subdiviser l'échantillon en plusieurs sous échantillons.

Une facilité de programmation peut être obtenue en utilisant l'approche d'optimisation par composantes (*coordinate descent*) dans le cadre de la vraisemblance pénalisée non concave. Toutefois, l'utilisation des résultats obtenus suite à l'application de ces méthodes doit se faire en prenant en compte l'aspect corrélation si nous nous intéressons aux modèles explicatifs. A ceci, il faut ajouter les points aberrants qui peuvent dans certains cas influencer les résultats, notamment dans le cas des moindres carrés.

La régression sur les quantiles a l'avantage d'être robuste même en présence de ces points aberrants. Une fois un groupe de variables sélectionnées, la régression sur les quantiles permet d'avoir une vue globale sur les différentes variations des coefficients de régression ainsi que leur seuil de significativité. Cette information est généralement masquée lorsqu'on utilise la régression linéaire classique. Le choix du paramètre de régularisation peut s'avérer très difficile dans certains cas. Généralement on utilise la validation croisée ou la validation croisée généralisée en se basant sur un critère à optimiser. Récemment, Koenker[87] a introduit une

approche basée sur le quantile d'une certaine variable aléatoire afin de sélectionner le paramètre de régularisation optimal. Les approches classiques basées sur le critère BIC sont également utilisés dans ce cadre. S'agissant du chapitre 2, la régression sur les quantiles fournit plus d'information que la régression classique via les moindres carrés. Le modèle de régression établi pour l'indice de fraîcheur semble plus adapté que celui de l'indice de qualité quant aux hypothèses requises pour la validité du modèle linéaire en considérant le graphique 2.5 de la régression. Cependant, il aurait été intéressant d'évaluer les performances de ces modèles en terme d'erreur de prédiction. Ce qui nécessite l'acquisition de nouvelles données.

Le chapitre 3 illustre l'avantage des approches de sélection de variables basées sur le bootstrap afin de limiter le nombre de variables indésirables. En particulier, les méthodes de sélection de stabilité randomisées ou non semblent être plus performantes. Avec l'approche du bootstrap, la majorité des composés volatiles intervenant dans les modèles du chapitre 2 ont été retrouvés via une autre approche plus rigoureuse, ce qui nous a permis d'identifier les composés les plus stables.

Le chapitre 4 regroupe l'essentiel des résultats sur l'effet groupement ainsi que les propriétés asymptotiques dans le cas de la régression quantile pénalisée. La continuité de type Lipschitz de la fonction de perte de régression quantile a permis d'affaiblir les conditions permettant d'avoir l'effet groupement pour l'Elastic net QR. La pénalité du Berhu a également certaines caractéristiques similaires à l'Elastic net mais elle a l'avantage de n'utiliser qu'un seul paramètre de régularisation. De plus, cette pénalité crée implicitement un groupe de variables ayant les plus grands coefficients sur lesquels agira une pénalité de groupe. Alors que les autres variables à faibles coefficients sont traitées individuellement par une pénalité de type Lasso.

L'implémentation notamment pour le cas de la pénalité du Berhu adaptative ou non qui représente l'un des points les plus difficiles de cette partie peut être réalisée directement en utilisant le package CVX sous Matlab. Cependant, cette approche ne permet pas à priori d'avoir les chemins de régularisation avec la même facilité que les méthodes classiques implémentées sous le logiciel R. A cela il faut ajouter une lenteur de calculs en utilisant CVX.

Enfin, le chapitre 5 traite l'aspect théorique, incluant la convergence de l'algorithme one step MLLQA, ainsi que les propriétés d'oracle de l'estimateur correspondant. L'algorithme one step LLA initialement proposé par Zou and Li[161] ne peut être utilisé dans le cas où $p > n$. Dans ce cas, une des alternatives est l'approche via l'optimisation par composantes. D'un point de vue empirique l'approche one step MLLQA tend à produire des modèles plus réduits avec une erreur de prédiction compétitive.

Du point de vue pratique, les outils de programmation linéaire ont beaucoup facilité la mise en oeuvre des méthodes proposées dans le cas de la régression sur les quantiles pénalisée ou non.

Contrairement aux moindres carrés pénalisés, les résultats sur les bornes d'estimation et de prédiction dans le cadre de la régression quantile pénalisée ne sont pas nombreux. Il serait alors intéressant de proposer de nouvelles bornes d'estimation et de prédiction en considérant d'autres types de pénalités autres que L_1 comme étudié récemment par Belloni and Chernozhukov[16], Kato[79].

D'autre part l'étude des propriétés d'oracle et éventuellement l'effet groupement des pénalités SCAD et MCP combinées avec la pénalité L_2 serait intéressante. Un travail précurseur a été initié par Becker et al.[14] dans le cadre de la classification via les SVM en grande dimension. La notion de point de changement (change point) a également été exploitée par Chenxi et al.[28] dans le cas où la régression quantile linéaire classique n'est pas adaptée ; par exemple dans le cas où la variable réponse est linéaire par morceaux mais dépend des prédicteurs de manière continue. Il serait intéressant de considérer les méthodes de régression quantile pour lesquelles nous avons l'existence d'un point de changement.

L'approche bayésienne pourrait aussi être exploitée dans le cadre de la régression quantile vu le nombre d'écrits récents dans ce sujet. Introduite par Yu and Moyeed[151] dans le cas non pénalisé, cette approche a fait l'objet d'extension au cas pénalisé grâce aux travaux d'Alhamzawi et al.[4], Qing et al.[121] entre autres.

Enfin, il serait également intéressant de considérer la technique du boosting

utilisée par Burgette et al.[24] afin de calculer les solutions du problème de régression quantile avec la pénalité du Berhu. Les détails de la technique du boosting peuvent être consultés dans les références Songfeng[135], Zhao and Yu[157], et Bühlmann[20]. De même nous envisageons une extension de l'algorithme MLLQA (Chapitre 5) au cadre de la régression quantile.

Annexe

5.7 Fonctions de perte presque quadratiques avec pénalité de type L_1

En considérant la pénalité de type L_1 définie par

$$J(\beta) = \| \beta \|_1 = \sum_j \mid \beta_j \mid \tag{5.24}$$

et la fonction de perte différentiable et quadratique par morceaux définie par

$$L(y, X\beta) = \sum_i l(y_i, \mathbf{x}_i^T \beta), \tag{5.25}$$

avec $l(y, \mathbf{x}^T\beta) = a(r)r^2 + b(r)r + c(r)$ et $r = (y - \mathbf{x}^T\beta)$ est le résidu pour la régression et $r = (y\mathbf{x}^T\beta)$ est la zone de séparation pour la classification. La fonction $l(r)$ est une spline quadratique, c'est à dire que $a(.), b(.), c(.)$ sont des fonctions constantes par morceaux définies de telle sorte que l soit différentiable. A titre d'exemple, nous avons :

- L'erreur quadratique $l(y, \mathbf{x}^T\beta) = (y - \mathbf{x}^T\beta)^2$ $(a \equiv 1, b \equiv 0, c \equiv 0)$.
- La fonction de perte de Huber de paramètre t fixé,

$$l(y, \mathbf{x}^T\beta) = \begin{cases} (y - \mathbf{x}^T\beta)^2, & \text{si } \mid y - \mathbf{x}^T\beta \mid \leq t, \\ 2t \mid y - \mathbf{x}^T\beta \mid -t^2, & \text{sinon.} \end{cases}$$

- La fonction de perte quadratique en classification $l(y, \mathbf{x}^T\beta) = (1 - y\mathbf{x}^T\beta)_+^2$.

Les fonctions de perte dans le cas de la régression quantile et des SVM n'appartiennent pas à cette famille du fait de la non différentiabilité en $y - \mathbf{x}^T\beta = 0$ et $y\mathbf{x}^T\beta = 1$ respectivement. Tous les problèmes d'optimisation de la classe des fonctions presque quadratiques utilisant la pénalité de type L_1 ont des chemins de régularisation optimaux $\hat{\beta}(\lambda)$ linéaires par morceaux lorsque λ varie. La fonction de Huber comme cas particulier de cette famille a l'avantage dans le cas de la regression d'être robuste du fait de la linéarité pour les grandes valeurs des résidus. En se basant sur le théorème 2 de Rosset and Zhu[126] ainsi que les arguments utilisés dans la démonstration, l'algorithme suivant a été fourni afin de générer le chemin de régularisation pour la classe des fonctions presque quadratiques utilisant la pénalité L_1. La particularité par rapport à l'algorithme LARS-Lasso (Efron et al.[42]) (qui est relativement plus simple) est la notion de "knot crossing" qui est absente pour le Lasso classique du fait de la double différentiabilité de la fonction de perte dans ce cas. La notion de "knot" est liée aux points de non double différentiabilité par exemple la valeur t pour la fonction de Huber correspondant au point de transition entre la partie linéaire et quadratique. Le principe de l'algorithme repose sur les points suivants :

- Démarrer d'une solution initiale nulle ($\lambda = \infty$)
- Identification des évènements
- Calcul des directions.

A titre indicatif, l'algorithme ne calcule pas λ de manière explicite du fait que l'évaluation des directions ainsi que les distances associées (step-length) ne font pas appel aux valeurs de λ.

La démonstration de ce théorème est nécessaire afin de comprendre l'algorithme.

Démonstration(Th 2 Rosset and Zhu[125])

En adoptant l'écriture

$min_{\beta^+,\beta^-} \sum_i l(y_i, x_i^T(\beta^+ - \beta^-)) + \lambda \sum_j (\beta_j^+ + \beta_j^-)$

s.t $\beta^+ \geq 0$, $\beta^- \geq 0$, $\forall j$

Ainsi la fonction duale de Lagrange de notre problème de minimisation est :

$min_{\beta^+,\beta^-} \sum_i l(y_i, x_i^T(\beta^+ - \beta^-)) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+ \beta_j^+ - \sum_j \lambda_j^- \beta_j^-$

Les conditions d'optimalité de KKT impliquent :

$$(\nabla L(\beta))_j + \lambda - \lambda_j^+ = 0$$

$$-(\nabla L(\beta))_j + \lambda - \lambda_j^- = 0$$

$$\lambda_j^+ \beta_j^+ = 0$$

$$\lambda_j^- \beta_j^- = 0$$

De ce fait, en une solution optimale pour λ fixé, les scénarios suivants sont susceptibles de se réaliser.

$$\lambda = 0 \Rightarrow (\nabla L(\beta))_j = 0 \forall j$$

(solution sans contrainte)

$$\beta_j^+ > 0,$$

$$\lambda > 0 \Rightarrow \lambda_j^+ = 0 \Rightarrow (\nabla L(\beta))_j = -\lambda < 0 \Rightarrow$$

$$\Rightarrow \lambda_j^- > 0 \Rightarrow \beta_j^- = 0$$

$$\beta_j^- > 0,$$

$$\lambda > 0 \Rightarrow \lambda_j^- = 0 \Rightarrow (\nabla L(\beta))_j = \lambda > 0 \Rightarrow$$

$$\Rightarrow \lambda_j^+ > 0 \Rightarrow \beta_j^+ = 0$$

$$\mid (\nabla L(\beta))_j \mid > \lambda \Rightarrow contradiction.$$

En se basant sur ces éventuels scénarios il est aisé de voir que :

– Les variables peuvent avoir des coefficients non nuls en $\hat{\beta}(\lambda)$ si seulement leur correlation absolue généralisée $\mid (\nabla L(\hat{\beta}(\lambda))_j \mid$ est égale à λ. Ainsi, pour toute valeur de λ nous avons un ensemble de variables actives $\mathcal{A} = \{j : \hat{\beta}_j(\lambda) \neq 0\}$ de sorte que :
$j \in \mathcal{A} \Rightarrow \mid \nabla L(\hat{\beta}(\lambda))_j \mid = \lambda$, $sign(\nabla L(\hat{\beta}(\lambda))_j) = -sign(\hat{\beta}(\lambda)_j)$
$j \notin \mathcal{A} \Rightarrow \mid \nabla L(\hat{\beta}(\lambda))_j \mid \leq \lambda$

– La direction dans laquelle $\hat{\beta}(\lambda)$ se déplace $\frac{\partial \hat{\beta}(\lambda)}{\partial \lambda}$ quand λ varie doit maintenir les conditions $\mid \nabla L(\hat{\beta}(\lambda))_{\mathcal{A}} \mid = \lambda$, $\mid \nabla L(\hat{\beta}(\lambda))_{\mathcal{A}^c} \mid \leq \lambda$.

Le terme $\frac{\partial \hat{\beta}(\lambda)}{\partial \lambda}$ à la forme suivante sur l'ensemble actif $\mathcal{A}$:

$$\frac{\partial \hat{\beta}(\lambda)_{\mathcal{A}}}{\partial \lambda} = -(\nabla^2 L(\hat{\beta}(\lambda))_{\mathcal{A}})^{-1} sign(\hat{\beta}(\lambda)_{\mathcal{A}}).$$

Par exemple dans le cas des fonctions presque quadratiques

$$\nabla^2 L(\hat{\beta}(\lambda))_{\mathcal{A}} = \sum_i 2a(r(y_i, x_{\mathcal{A}i}^T \hat{\beta}(\lambda)_{\mathcal{A}}))x_{\mathcal{A}i} x_{\mathcal{A}i}^T.$$

Le fait que la fonction $a(.)$ soit constante par morceaux, les termes $\nabla^2 L(\hat{\beta}(\lambda))_{\mathcal{A}}$ et $\frac{\partial \hat{\beta}(\lambda)_{\mathcal{A}}}{\partial \lambda}$ sont également constants par morceaux. Ainsi le chemin de solution $\hat{\beta}(\lambda)$ est linéaire par morceaux.

La double différentiabilité n'est plus assurée quand l'un des cas suivants se produit (l'expression précédente de $\frac{\partial \hat{\beta}(\lambda)_{\mathcal{A}}}{\partial \lambda}$ changera dans ce cas) :

– Ajout d'une variable : une nouvelle variable doit s'ajouter à $\mathcal{A}$, i.e atteinte d'un point où l'inégalité $\mid \nabla L(\hat{\beta}(\lambda))_{\mathcal{A}^c} \mid \leq \lambda$ ne sera plus vérifiée si $\hat{\beta}(\lambda)$ varie dans la même direction.
– Suppression d'une variable (annulation d'un élément de $\mathcal{A}$) : un point de non différentiabilité est atteint au niveau de la pénalité, ainsi la condition $sign(\nabla L(\hat{\beta}(\lambda))_{\mathcal{A}}) = -sign(\hat{\beta}(\lambda)_{\mathcal{A}})$ ne sera plus vérifiée si l'on évolue dans la même direction. Le coefficient s'annulant doit être exclu de $\mathcal{A}$.

– Traverser un noeud : un résidu généralisé $r(y_i, x_i^T\hat{\beta}(\lambda))$ atteint un point de non double différentiabilité (noeud) dans la fonction de perte L. A titre d'exemple le point t pour Huber ou la valeur 0 dans le cas de la régression quantile.

Algorithme 1 : Cas des fonctions de perte presque quadratiques combinant la pénalité L_1

1. Initialisation : $\beta = 0$, $\mathcal{A} = argmax_j \mid \nabla L(\beta) \mid_j$,$\gamma_{\mathcal{A}} = -sign(\nabla L(\beta))_{\mathcal{A}}$, $\gamma_{\mathcal{A}^c} = 0$.

2. Si $max\{\nabla L(\beta)\} > 0$

(a) $d_1 = min\{d > 0 :\mid \nabla L(\beta + d\gamma)_j \mid=\mid \nabla L(\beta + d\gamma)_{\mathcal{A}} \mid, j \notin \mathcal{A}\}$

(b) $d_2 = min\{d > 0 : (\beta + d\gamma)_j = 0, j \in \mathcal{A}\}$ (hit 0)

(c) $d_3 = min\{d > 0 : r(y_i, x_i^T(\beta + d\gamma))$, hits a "knot", $i = 1, ..., n\}$

(d) poser $d = min(d_1, d_2, d_3)$ (longueur du pas)

(e) $\beta \leftarrow \beta + d\gamma$ (choix d'étape)

(f) Si $d = d_1$ alors ajouter la variable réalisant l'égalité $\mathcal{A}$ en d.

(g) Si $d = d_2$ alors retirer de $\mathcal{A}$ la variable atteignant 0 en d.

(h) Si $d = d_3$ pour l'observation i^*, décider alors d'une valeur appropriée pour $a(r(y_{i^*}, x_{i^*}^T\beta))$.

(i) $C = \sum_i a(r(y_i, x_i^T\beta))x_{\mathcal{A},i}x_{\mathcal{A},i}^T$ (calcul d'une nouvelle direction (i) à (k))

(j) $\gamma_{\mathcal{A}} = C^{-1}(-sign(\beta_{\mathcal{A}}))$

(k) $\gamma_{\mathcal{A}^c} = 0$

5.7.1 Coûts numériques

La complexité du LARS-Lasso est approximativement équivalente à celle d'une évaluation par moindres carrés, c'est à dire $O(p^3+np^2) = O(np^2)$ quand $n > p$. Un point important à souligner est que les évènements liés à l'exclusion d'une variable dans l'algorithme précédent (cas où $d = d_2$ à l'étape 2(b)) ne sont pas comptés et il n'y a pas une borne théorique simple relative à la fréquence de leur réalisation. Par conséquent ils sont considérés comme rares en moyenne ($O(1)$, voir discussion dans Efron et al.[42], Rosset et Zhu[126]) bien que cela ne puisse être garanti. S'inspirant de ce fait, Rosset et Zhu[126] affirment qu'en moyenne la complexité de l'algorithme précédent est $O(n^2p)$ quand $n > p$. La complexité supplémentaire liée à l'algorithme en question double :

- La détermination de la longueur de pas nécessite la considération de tous les $O(n)$ évènements possibles liés au "knot crossing" (étape 2(c)) ; ce qui ne se produit pas dans le cas du Lasso.
- Les évènements relatifs au "knot crossing" augmentent le nombre d'étapes de l'algorithme. La supposition de rareté des évènements d'exclusion implique que le nombre d'étapes du Lasso est O(p). Pour le présent algorithme il est également supposé que les "knots crossing" apparaissent uniquement $O(n)$ fois. Comme spécifié par Rosset et Zhu[126], en moyenne, cette deuxième hypothèse est très raisonnable puisque les résidus tendent à varier de façon monotone autour de 0 quand le paramètre de régularisation diminue.

Globalement, les points précédemment évoqués supposent que nous avons $O(n+p)$ étapes, chacune nécessitant :

- $O(np + p) = O(np)$ évaluations afin d'avoir la longueur du pas(étapes 2(a)-2(d)).
- $O(\mid \mathcal{A} \mid^2) = O(p^2)$ évaluations (inverse updating and downdating) pour calculer la nouvelle direction (étape 2(j)) en utilisant le lemme de Sherman-

Morrison-Woodbury et aussi le fait que nous ajoutons ou excluons une seule variable ou une seule observation à la fois.

Ce qui nous donne une complexité globale de $O((n+p)np + (n+p)p^2) = O(n^2p)$ quand $n > p$. Dans le cas où $n > p$ l'algorithme nécessite $O(p)$ étapes pour ajouter toutes les variables et $O(n)$ étapes pour le franchissement de noeud. Si $n < p$, du fait qu'au plus n variables peuvent intégrer le modèle estimé, l'algorithme requiert $O(n)$ étapes pour l'ajout de variables et le franchissement des noeuds combinés. Les évènements de suppression sont habituellement rares ($O(1)$). Comme la valeur maximale de $\mid \mathcal{A} \mid$ est $min(n,p)$, l'ensemble des coûts de calcul est $O(n^2p)$. Enfin d'une manière générale s'il n'y a pas de franchissement de noeuds (exemple du Lasso), le nombre total de steps est $O(min(n,p))$ et l'ensemble des coûts de calcul est ainsi de $O(npmin(n,p))$

Dans le cas $p \gg n$ la solution des moindres carrés peut être obtenue avec un coût de calcul de seulement $O(n^3)$ pour tout choix de n prédicteurs linéairement indépendants. En utilisant les mêmes hypothèses que précédemment (le nombre d'éléments d'exclusion est $O(1)$ et le nombre de steps est $O(n)$), la complexité est alors de $O(pn^2)$ tant pour le Lasso que pour l'algorithme précédent.

5.7.2 Fonctions de perte linéaires avec pénalité de type L_1

Nous nous intéresserons aux fonctions de perte linéaires par morceaux et non différentiables rencontrées aussi bien en régression (exemple de la régression sur les quantiles) qu'en classification (cas du SVM). Par exemple dans le cas de la régression sur les quantiles la fonction de perte a la forme :

$$l(y, x^T\beta) = \begin{cases} \tau. \mid y - x^T\beta \mid & \text{si } y - x^T\beta \geq 0 \\ (1-\tau). \mid y - x^T\beta \mid & \text{si } y - x^T\beta < 0 \end{cases}$$

et dans le cas des SVM la fonction de perte considérée (hinge loss) à l'expression $l(y, x^T\beta) = (1 - yx^T\beta)_+$. Les deux fonctions précédentes peuvent être généralisées dans l'écriture suivante :

$$l(r) = \begin{cases} b_1. \mid a + r \mid & \text{si } a + r \geq 0 \\ b_2. \mid a + r \mid & \text{si } a + r < 0 \end{cases}$$

avec le résidu r ayant la forme générale suivante :

$$r = \begin{cases} y - x^T\beta & \text{cas de la régression} \\ y.x^T\beta & \text{cas de la classification} \end{cases}$$

L'algorithme suivant nous donne le chemin de régularisation pour tous les problèmes de cette catégorie utilisant la pénalité L_1. Dans l'algorithme précédent, la direction du chemin de régularisation est uniquement déterminée par les points des parties quadratiques des fonctions de perte. Dans le cas présent, le sens de la trajectoire est uniquement déterminée par les points se trouvant sur le "coude" (elbow) $\mathcal{E}$ représentant le point de non-différentiabilité de la fonction de perte.

L'algorithme du chemin de régularisation utilise les ensembles ci dessous définis :

$\mathcal{A} = \{j : \beta_j \neq 0\}$ l'ensemble des variables actives. $\mathcal{E} = \{i : a+r_i = 0\}$ l'ensemble coude des observations. $\mathcal{L} = \{i : a + r_i < 0\}$ gauche du coude. $\mathcal{R} = \{i : a + r_i > 0\}$ droite du coude.

$$\Delta r(y, \Delta f) = \begin{cases} -\Delta f & \text{cas de la régression} \\ y.\Delta f & \text{cas de la classification} \end{cases}$$

$\Delta L(\gamma) = b_1 \sum_{\mathcal{R}} \Delta r(y_i, x_i^T\gamma) - b_2 \sum_{\mathcal{L}} \Delta r(y_i, x_i^T\gamma)$. L'idée de l'algorithme est de commencer avec une variable qui est susceptible de diminuer la fonction de perte avec le plus de rapidité par unité de variation du coefficient de cette variable. La solution se déplace le long de cette direction et s'arrête lorsque l'un des deux événements suivants se produit :

– une observation atteint le point de non différentiabilité de la perte, i.e $a+r \neq 0$ devient $a + r = 0$ pour un certain i. Ce point est ainsi ajouté à l'ensemble

$\mathcal{E}$.

- Un coefficient estimé atteint le point de non différentiabilité de la pénalité, i.e β_j change d'une valeur non nulle à une valeur nulle pour un certain j. Cette variable est alors exclue de l'ensemble $\mathcal{A}$.

Il y'a donc deux types d'actions à envisager :

- Ajouter une variable dans $\mathcal{A}$.
- Exclure un point de l'ensemble $\mathcal{E}$.

Une fois de plus, l'action choisie est déterminée par la variable susceptible de diminuer la perte le plus rapidement par unité d'augmentation de la norme L_1 dans le vecteur de coefficients, et la solution changerait le long d'une nouvelle direction de telle sorte que les points de $\mathcal{E}$ restent dans $\mathcal{E}$.

Nous présentons ci dessous l'algorithme détaillé, tout en précisant qu'il est fondamentalement différent de l'algorithme LARS-Lasso et de l'algorithme 1 précédent du fait que la fonction de perte soit non différentiable.

Algorithme 2 : Cas des fonctions de perte L_1 combinant la pénalité L_1

1. Initialisation

$\beta = 0$

$\mathcal{A} = argmax_j \mid \Delta L(e_j) \mid$

$\gamma_j = -sign(\Delta L(e_j)), j \in \mathcal{A}$

$\gamma_j = 0, j \in \mathcal{A}^c$

$\Delta L^* = max_j \mid \Delta L(e_j) \mid$

où e_j est le vecteur de composantes nulles sauf la j ème qui est égale à 1.

2. Tant que $\Delta L^* \neq 0$

(a) $d_1 = min\{d > 0 : (\beta + d\gamma)_j = 0, j \in \mathcal{A}\}$

(b) $d_2 = min\{d > 0 : a + r(y_i, x_i^T(\beta + d\gamma)) = 0,\ i = 1, ..., n\}$

(c) Prendre $d = min(d_1, d_2)$

(d) $\beta \leftarrow \beta + d\gamma$

(e) Si $d = d_1$, exclure alors de $\mathcal{A}$ la variable s'annulant en d ; si $d = d_2$, ajouter à $\mathcal{E}$ l'observation atteignant le coude.

(f) Pour chaque $j^* \in \mathcal{A}^c$, résoudre pour γ

$$\begin{cases} \Sigma_{j\in\mathcal{A}}\gamma_j x_{ij} + \gamma_{j^*} x_{ij^*} & = 0 \quad \text{for } i \in \mathcal{E} \\ \Sigma_{j\in\mathcal{A}} sign(\beta_j)\gamma_j + \mid \gamma_{j^*} \mid & = 1 \\ \gamma_j & = 0 \quad \text{for } j \notin \mathcal{A} \cup \{j^*\} \end{cases}$$

Calculer

$\Delta L(\gamma) = b_1 \sum_{\mathcal{R}} \Delta r(y_i, x_i^T\gamma) - b_2 \sum_{\mathcal{L}} \Delta r(y_i, x_i^T\gamma)$.

(g) Pour chaque $i^* \in \mathcal{E}$, résoudre pour γ

$$\begin{cases} \Sigma_{j\in\mathcal{A}}\gamma_j x_{ij} & = 0 \quad \text{for } i \in \mathcal{E} \setminus \{i^*\} \\ \Sigma_{j\in\mathcal{A}} sign(\beta_j)\gamma_j & = 1 \\ \gamma_j & = 0 \quad \text{for } j \in \mathcal{A}^c \end{cases}$$

Calculer

$\Delta L(\gamma) = b_1 \sum_{\mathcal{R}} \Delta r(y_i, x_i^T\gamma) - b_2 \sum_{\mathcal{L}} \Delta r(y_i, x_i^T\gamma)$.

(h) Choisir la plus petite valeur de $\Delta L(\gamma)$ à partir des étapes 2(f) et 2(g). Notons $\Delta L^* = min\Delta L(\gamma)$.

- Si ΔL^* correspond à un indice j^* dans l'étape 2(f), actualiser γ et $\mathcal{A} \leftarrow \mathcal{A} \cup \{j^*\}$
- Si ΔL^* correspond à un indice i^* dans l'étape 2(g), actualiser γ et $\mathcal{E} \leftarrow \mathcal{E} \setminus \{i^*\}$, $\mathcal{L} \leftarrow \mathcal{L} \cup \{i^*\}$ or $\mathcal{R} \leftarrow \mathcal{R} \cup \{i^*\}$.
- Si ΔL^* est positif, poser $\Delta L^* = 0$.

(i) Retourner à l'étape 2.

A titre de remarque, pour la fonction de perte L_∞ qui est également linéaire par morceaux et non différentiable, le chemin de régularisation est également linéaire par morceaux et un algorithme similaire pourrait être adopté comme évoqué dans Rosset and Zhu[126]. L'algorithme s'arrête quand les taux de variation dans la fonction de perte L, $\Delta L/\Delta\lambda$ sont tous positifs ou quand L ne peut plus décroître. Plus de détails algorithmiques peuvent être trouvés pour le cas de SVM dans Zhu et al.[158] et pour la régression sur les quantiles dans Li and Zhu[103] où le cas de l'intercept, l'unicité ou non de la solution initiale ainsi que certains aspects théoriques ont été traités en détails. Enfin, l'unicité de la solution dans l'algorithme précédent suppose que $\mid \mathcal{A} \mid = \mid \mathcal{E} \mid$ comme évoqué par Yao and Lee[150]. D'autres algorithmes ont été également proposés par Kato[78] ; Osborne and Turlach[113] via l'algorithme d'homotopie dans le cadre de la régression quantile.

5.8 Sélection de variables, méthodes de régularisation et programmation linéaire

La sélection de variables peut être également vue sous l'angle de la programmation linéaire comme détaillée dans Yao and Lee[150]. Pour des raisons de simplicité, les notions de solution de base réalisable, solution de base réalisable non dégénérée,

solution de base réalisable dégénérée et solution de base optimale ne seront pas développées. Dans le cas de la régression quantile pénalisée avec pénalité de type L_1, nous considérons la formulation suivante :

$$\begin{cases} min_{z\in\mathbb{R}^n} & (\mathbf{c}+\lambda\mathbf{a})^T z \qquad (*) \\ s.t. & \mathbf{A}z = \mathbf{b} \\ & z \geq \mathbf{0} \end{cases}$$

En général, les méthodes classiques utilisées pour la résolution des problèmes de programmation linéaire incluent la méthode du simplex, la méthode dual simplex, la méthode tableau et les méthodes du point intérieur. D'un point de vue géométrique, le problème de programmation linéaire standard cherche le minimum d'une fonction linéaire sur un polyhèdre dont les sommets sont définis par des hyperplans. Ainsi, si notre problème admet une solution, au moins un des points d'intersection avec les hyperplans (solutions de base) doit réaliser le minimum. Du fait que l'ensemble des indices de base ne dépende pas du paramètre λ, un ensemble optimal d'indices de base $\mathcal{B}^*$ pour une certaine valeur λ fixée restera optimal sur tout un intervalle $[\underline{\lambda}, \overline{\lambda}]$ appelé intervalle d'optimalité de $\mathcal{B}^*$ pour le problème de programmation linéaire paramétrique. Pour un hyperplan optimal nous avons $\beta_0 + \mathbf{x}^T\beta = 0$. Nous avons ci dessous une autre formulation du problème précédent :

$$\begin{cases} \quad min & \mathbf{c}^T\mathbf{z} \qquad (**) \\ \mathbf{z}\in\mathbb{R}^n, \delta\in\mathbb{R} & \\ \quad s.t. & \mathbf{A}\mathbf{z} = \mathbf{b} \\ & \mathbf{a}^T\mathbf{z} + \delta = s \\ & \mathbf{z} \geq \mathbf{0}, \delta \geq 0. \end{cases}$$

qui peut être transformée sous la formulation standard avec second membre (right-hand-side LP problem) :

$$\begin{cases} min_{\mathbb{Z}\in\mathbb{R}^{n+1}} & \mathbb{C}^T\mathbb{Z} \\ s.t. & \mathbb{A}\mathbb{Z} = \mathbb{B} + \omega\mathbb{B}^* \\ & \mathbb{Z} \geq \mathbf{0} \end{cases}$$

en posant $\omega = s$, $\mathbb{Z} = \begin{bmatrix} \mathbf{z} \\ \delta \end{bmatrix}$, $\mathbb{C} = \begin{bmatrix} \mathbf{c} \\ 0 \end{bmatrix}$, $\mathbb{B} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$, $\mathbb{B}^* = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$, et $\mathbb{A} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{a}^T & 1 \end{bmatrix}$. Quand $\mathbf{A}$ est de plein rang, $\mathbb{A}$ l'est aussi. Soit $\mathcal{B}^*$ un ensemble de base optimal du problème précédent en $\omega = \omega^*$, Yao and Lee[150] ont affirmé qu'on peut montrer que $\mathcal{B}^*$ est optimal pour tout ω satisfaisant $\mathbb{Z}_{\mathcal{B}^*} = \mathbb{A}_{\mathcal{B}^*}^{-1}(\mathbb{B} + \omega\mathbb{B}^*) \geq \mathbf{0}$, et il existe $\underline{\omega}$ et $\overline{\omega}$ de sorte que $\mathcal{B}^*$ soit optimal pour $\omega \in [\underline{\omega}, \overline{\omega}]$. Ce qui implique qu'une version du chemin de solution de ce programme linéaire est une fonction linéaire par morceaux. Lors de la résolution, on utilise l'algorithme simplex ou l'algorithme tableau-simplex. Par exemple quand l'algorithme tableau-simplex converge en J itérations, la complexité algorithmique du L_1 QR est $O((p+1)nJ)$. Les chemins de régularisation ont fait l'objet des théorèmes 6 et 7 dans Yao and Lee[150]. Dans le cas du problème $(*)$ le chemin de solution est :

$$\begin{cases} \mathbf{z}^0 & pour \lambda > \lambda_0 \\ \mathbf{z}^l & pour \lambda_l < \lambda < \lambda_{l-1}, l = 1, ..., J \\ \tau\mathbf{z}^l + (1-\tau)\mathbf{z}^{l+1} & pour \lambda = \lambda_l, \text{ et } \tau \in [0,1],\ l = 0, ..., J-1 \end{cases}$$

Il est possible d'obtenir le chemin de solution comme fonction de s en s'intéressant au problème $(**)$. Du fait de l'équivalence des deux formulations, le l^{eme} point de jonction de la solution linéaire par morceaux est donné par $s_l = \mathbf{a}^T\mathbf{z}^l$, et la solution entre les points de jonction est une combinaison linéaire de $\mathbf{z}^l$ et $\mathbf{z}^{l+1}$ comme spécifié dans le théorème 7 évoqué précédemment et spécifiant que pour $s \geq 0$, le chemin de solution peut prendre la forme suivante

$$\begin{cases} \frac{s_{l+1}-s}{s_{l+1}-s_l}\mathbf{z}^l + \frac{s-s_l}{s_{l+1}-s_l}\mathbf{z}^{l+1} & sis_l \leq s \leq s_{l+1} \text{ et } l = 0, ..., J-1 \\ \mathbf{z}^J & sis \geq s_J \end{cases}$$

Dans le cas des quantiles en considérant le problème d'optimisation sous contrainte :

$$\begin{cases} min & \sum_{i=1}^n \rho_\tau(y_i - \beta_0 - \mathbf{x}_i^T\beta) \\ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p & \\ s.t. & \| \beta \|_1 \geq s, \end{cases}$$

où $s > 0$ est un paramètre de régularisation. De manière équivalente, en considérant $\lambda_Q = \tau/(1-\tau)$ et λ un autre paramètre de régularisation. Nous pouvons avoir la

formulation suivante

$$\begin{cases} min & \sum_{i=1}^{n}\{(\zeta_i)_+ + \lambda_Q(\zeta_i)_-\} + \lambda \parallel \beta \parallel_1 \\ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \zeta \in \mathbb{R}^n & \\ s.t. & \beta_0 + \mathbf{x}_i^T\beta + \zeta_i = y_i \text{ pour } i = 1, ..., n, \end{cases}$$

permettant d'avoir la forme standard du programme linéaire pour la régression quantile combinée à la norme L_1 avec les paramètres suivants :

$$\mathbf{z} := \begin{pmatrix} \beta_0^+ & \beta_0^- & (\beta^+)^T & (\beta^-)^T & (\zeta^+)^T & (\zeta^-)^T \end{pmatrix}^T$$

$$\mathbf{c} := \begin{pmatrix} 0 & 0 & \mathbf{0}^T & \mathbf{0}^T & \mathbf{1}^T & \lambda_Q\mathbf{1}^T \end{pmatrix}^T$$

$$\mathbf{a} := \begin{pmatrix} 0 & 0 & \mathbf{1}^T & \mathbf{1}^T & \mathbf{0}^T & \mathbf{0}^T \end{pmatrix}^T$$

$$\mathbf{A} := \begin{pmatrix} 1 & -1 & \mathbf{X} & -\mathbf{X} & \mathbf{I} & -\mathbf{I} \end{pmatrix}$$

$$\mathbf{b} := \mathbf{Y}$$

composés de $N = 2(1 + p + n)$ variables et $M = n$ contraintes d'égalité.

5.9 Algorithme One step LLA (Zou and Li[161])

Grandes lignes de l'algorithme one-step LLA :
- $U = \{j : J'_\lambda(|\beta_j^{(0)}|) = 0\}$
- $V = \{j : J'_\lambda(|\beta_j^{(0)}|) > 0\}$
- $X^* = [X_U^*, X_V^*]$
- $\boldsymbol{\beta}^{(1)} = (\beta_U^{(1)\mathbf{T}}, \beta_V^{(1)\mathbf{T}})^T$
- Evaluer $\hat{\beta}_V^*$ via le LARS
- $\hat{\beta}_U^* = (X_U^{*T}X_U^*)^{-1}X_U^{*T}(Y^* - X_V^*\hat{\beta}_V^*)$
- $\beta_U^{(1)} = \hat{\beta}_U^*$
- $\beta_j^{(1)} = \hat{\beta}_j^* \frac{\lambda}{J'_\lambda(|\beta_j^0|)}$, $j \in V$

Bibliographie

[1] B. Abdous and B. Rémillard. Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, 47(2) :371–384, 1995.

[2] J. Adrover, R.A. Maronna, and V.J. Yohai. Robust regression quantiles. *Journal of Statistical Planning and Inference*, 122 :187–202, 2004.

[3] C. Alasalvar, T. Aishima, and P.C. Quantick. Dynamic headspace analysis of volatile aroma products in fresh and deteriorated mackerel (scomber scombrus). *Food.Sci.Technol.Int*, 1 :125–127, 1995.

[4] R. Alhamzawi, K. Yu, and D.F. Benoit. Bayesian adaptive lasso quantile regression. *Statistical Modelling*, 12(3) :279–297, March 2012.

[5] O.G. Alma. Comparison of robust regression methods in linear regression. *Int.J. Contemp.Math.Sciences*, 6(9) :409–421, 2011.

[6] L.T.H. An and P.D. Tao. Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. *Journal of Global Optimization*, 11 :253–285, 1997.

[7] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455) :437–447, September 2001.

[8] T. Aro, R. Tahovonen, L. Koskinen, and H. Kallio. Volatile compounds of baltic herring analysed by dynamic headspace sampling-gas chromatography-mass spectrometry. *Eur.Food.Res.Technol*, 216 :483–488, 2003.

[9] F.R. Bach. Bolasso : Model consistent lasso estimation through the bootstrap. *Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland*, 2008.

[10] B. Bailar. Salary survey of u.s. colleges and universities offering degrees in statistics. *Amstat News*, 182 :3–10, 1991.

[11] I. Barrodale and F.D.K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5) :839848, 1973. doi :10.1137/0710069.

[12] G.W. Bassett and R. Koenker. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77 :407–415, 1982.

[13] P. Bühlmann and S. Van de Geer. *Statistics for High-Dimensional Data*. Methods, Theory and Applications Series : Springer Series in Statistics, 2011.

[14] N. Becker, G. Toedt, P. Lichter, and A. Benner. Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC Bioinformatics*, 2011.

[15] A. Beinrucker, U. Dogan, and G. Blanchard. A simple extension of stability feature selection. *Pattern Recognition Lecture Notes in Computer Science*, 7476 :256–265, 2012.

[16] A. Belloni and V. Chernozhukov. l_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39(1) :82–130, 2011.

[17] A. Bene, A. Fornage, J. Luisier, P. Pichler, and J. Villettaz. A new method for the rapid determination of volatile substances : the spme-direct method. part i : Apparatus and working conditions. *Sensors Actuators B*, 72 :184–187, 2001.

[18] A. Bene, A. Hayman, E. Reynard, J. Luisier, and J. Villettaz. A new method for the rapid determination of volatile substances : the spme-direct method. part ii : Determination of the freshness of fish. *Sensors Actuators B*, 72 :204–207, 2001.

[19] A.K. Bera, A.F. Galvao, G. Montes-Rojas, and S.Y. Park. Which quantile is the most informative ? maximum likelihood, maximum entropy, and quantile regression. Technical report, Working Papers from Department of Economics, N°10/08. City University London, 2010.

[20] P. Bühlmann. Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2) :559–583, 2006.

[21] L. Breiman. Regularization of wavelet approximations. *Ann. Statist.*, 24(6) :2350–2383, 1996.

[22] L. Breiman and J. Friedman. Predicting multiple responses in multiple linear regression(with discussion). *Journal of the Royal Statistical Society : Series B*, 59 :3–54, 1997.

[23] M. Buchinsky. Changes in the u.s. wage structure 19631987 : Application of quantile regression. *Econometrica*, 62 :405458, 1994.

[24] L.F. Burgette and M.L. Miranda. Exploratory quantile regression with many covariates : an application to adverse birth outcomes. *Epidemiology*, 22(6) :859–866, November 2011.

[25] C. Chen and Y. Wei. Computational issues for quantile regression. *The Indian Journal of Statistics*, 67 :399–417, 2005.

[26] L.A. Chen, L.T. Tran, and L.C. Lin. Symmetric regression quantile and its application to robust estimation for the nonlinear regression model. *Journal of Statistical Planning and Inference*, 126 :423–440, 2004.

[27] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. on Sci. Comp*, 20(1) :33–61, 1998.

[28] L. Chenxi, W. Ying, C. Rick, and H. Xuming. Bent line quantile regression with application to an allometric study of land mammals' speed and mass. *Biometrics*, 67(1) :242249, March 2011.

[29] V. Chernozhukov. Extremal quantile regression. *The Annals of Statistics*, 33(2) :806–839, 2005.

[30] V. Chernozhukov, I. Fernández-Val, and A. Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3) :1093–1125, 2010.

[31] V. Chernozhukov and H. Hong. An mcmc approach to classical estimation. Technical report, Massachusetts Institute of Technology, Department of Economics,Working Paper Series, 2003.

[32] D.I. Clark and M.R. Osborne. Finite algorithms for hubers m-estimator. *SIAM J.Sci.Statist.Comput.*, 6 :7285, 1986.

[33] T.J. Cole. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society, Ser. A*, 151 :385–418, 1988.

[34] B. Crépon and N. Jacquemet. *Econométrie : Méthode et Applications*. De Boeck, 2010.

[35] X. D'haultfœuille and P. Givord. La régression quantile en pratique. Technical report, 2011/2012. CREST,INSEE, Direction de la méthodologie.

[36] G. Duflos, V.M. Coin, F. Moine, and P. Malle. Determination of volatile compounds in whiting using spme gc-ms. *J.Chromatogr.Sci*, 43 :304–312, 2005.

[37] G. Duflos, M. Cornu, V.M. Coin, J.F. Antinelli, and P. Malle. Determination of volatile compounds to characterize fish spoilage using hs/ms and spme-gc/ms analysis. *J.Agric.Food.Chem*, 86 :600–611, 2006.

[38] G. Duflos, F. Leduc, A. N'Guessan, F. Krezewinski, O. Kol, and P. Malle. Freshness characterisation of whiting (merlangius merlangus) using an spme/gc/pm method and a statistical multivariate approach. *J.Sci.Food.Agric*, 90 :2568–2575, 2010.

[39] G. Duflos, F. Leduc, A. N'Guessan, F. Krezewinski, K. Ossarath, and P. Malle. Freshness characterisation of whiting (merlangius merlangus) using an spme/gc/pm method and a statistical multivariate approach. *J Sci Food Agric*, 90 :2568–2575, 2010.

[40] R.K.B. Edirisinghe, A.J. Graffham, and S.J. Taylor. Characterisation of the volatiles of yellow fin tuna (thunnus albacares) during storage by solid phase microextraction and gc-ms and their relationship to fish quality parameters. *Int.J.Food.Sci.Technol*, 42 :1139–1147, 2007.

[41] B. Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1 :93–125, 1991.

[42] B. Efron, T. Hastie, I.M. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), 2004.

[43] M. El Anbari and A. Mkhadri. The adaptive gril estimator with a diverging number of parameters. arXiv preprint arXiv :1302.6390., 2013.

[44] M. El Anbari and A. Mkhadri. Penalized regression combining the l_1 norm and a correlation based penalty. *Sankhya B*, 2013. (to appear).

[45] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *Technical Report, Electrical Engineering and Compu-*

ter Sciences Department, University of California at Berkeley, Berkeley, UC/EECS, 126, 2010.

[46] J. Fan. Comments on "wavelets in statistics : A review", by a.antoniadis. *Journal of Italian Statistical Society*, 6 :131–138, 1997.

[47] J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. In press*, 2011.

[48] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 :1348–1360, 2001.

[49] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society B*, 70 :849–911, 2008.

[50] J. Fan and J. Lv. Properties of non-concave penalized likelihood with np-dimensionality. arXiv :0910.1119v1 [math.ST], 2009.

[51] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, 20 :101–148, 2010.

[52] J. Fan and J. Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57 :5467–5484, 2011.

[53] J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics (invited review article). *Annual Review of Economics*, 3 :291–317, 2011.

[54] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32 :928–961, 2004.

[55] J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional variable selection : beyond the linear model. *Journal of Machine Learning Research*, 10 :1829–1853, 2009.

[56] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35 :109–148, 1993.

[57] C.J. Geyer. On the asymptotics of constrained m-estimation. *Amer. Statist*, 22 :1993–2010, 1994.

[58] S. Ghosh. Adaptive elastic net : An improvement of elastic net to achieve oracle properties. *Tech. rep., Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis*, 2007.

[59] A. Giloni, S. Jeffrey Simonoff, and S. Bhaskar. Robust weighted lad regression. *Computational Statistics and Data Analysis*, 50(11) :3124–3140, 2006.

[60] M. Grant and S. Boyd. The cvx users' guide. November 03, 2012.

[61] C. Gutenbrunner and J. Jureckova. Regression rank scores and regression quantiles. *The Annals of Statistics*, 20(1) :305–330, 1992.

[62] Jurecková J. Koenker R. Gutenbrunner, C. and S. Portnoy. Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2 :307–333, 1993.

[63] J. HÁjek and Z. Šidák. Theory of rank tests. *Academia, Prague*, 1967.

[64] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path of the support vector machine. *Journal of Machine Learning Research*, 5 :1391–1415, 2004.

[65] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical Learning : data mining, inference and prediction.* Springer, 2009.

[66] X. He. Quantile curves without crossing. *The American Statistician*, 52(2) :186–192, 1997.

[67] X. He and Q.-M. Shao. On parameters of increasing dimensions. *J. Multivariate Anal.*, 73 :120135, 2000.

[68] M. Hebiri. *Quelques Questions De Sélection De Variables Autour De LEstimateur LASSO.* PhD thesis, Paris Diderot-Paris 7, 2009.

[69] L. Hongbo. A study of robust hybrids of lasso and ridge regression and applications. Master's thesis, 2010.

[70] P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1) :73–101, 1964.

[71] P. Huber. Robust regression : Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1 :799–821, 1988.

[72] D. Hunter and K. Lange. A tutorial on mm algorithms. *American. Statistician*, 58 :30–37, 2004.

[73] D.R. Hunter and K. Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1) :60–77, 2000.

[74] D.R. Hunter and R. Li. Variable selection using mm algorithms. *Ann.Statist*, 33(4) :1617–1642, 2005.

[75] R. Jonsdottir, M. Bragadottir, and G. Olafsdottir. The role of volatile compounds in odor development during hemoglobin-mediated oxidation of cod muscle membrane lipids. *J.Aquat.Food.Prod.Technol*, 16 :67–86, 2007.

[76] J. Jurecková. Regression quantiles and trimmed least squares estimator under a general design. *Kybernetika*, pages 345–357, 1984.

[77] N. Karmarker. A new polynomial time algorithm for linear programming. *Combinatorica*, 4 :373–395, 1984.

[78] K. Kato. Solving l_1 regularization problems with piecewise linear losses. *J. Comp. Graph. Statist*, 19(4) :1024–1040, 2010.

[79] K. Kato. Group lasso for high dimensional sparse quantile regression models. arXiv :1103.1458v2[stat.ME], 2011.

[80] Y. Kim, H. Choi, and H. Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103 :1656–1673, 2008.

[81] M. Kocherginsky, X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14 :41–55, 2005.

[82] R. Koenker. A note on l-estimates for linear models. *Statist. Probab. Lett*, 2 :323–325, 1984.

[83] R. Koenker. A note on l-estimators for linear models. *Statistics and Probability Letters*, 2 :323–325, 1984.

[84] R. Koenker. Confidence intervals for regression quantiles. *in Asymptotic Statistics, P. Mandl and M. Huskova, eds., Springer-Verlag, New York.*, 3 :349–359, 1994.

[85] R. Koenker. quantreg : Quantile regression. *R package version 4.94 (http ://www.econ.uiuc.edu/ roger/research/rq/rq.html)*, 2004.

[86] R. Koenker. *Quantile Regression.* Cambridge University Press, 2005.

[87] R. Koenker. Additive models for quantile regression : Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3) :239–262, 2011.

[88] R. Koenker and G.S. Bassett. Regression quantiles. *Econometrica*, 46 :33–50, 1978.

[89] R. Koenker and G.W. Bassett. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50 :43–61, 1982.

[90] R. Koenker and V. D'Orey. Computing regression quantiles. *Applied Statistics*, 36 :383–393, 1987.

[91] R. Koenker and V. d'Orey. A remark on algorithm as 229 : Computing the dual regression quantiles and regression rank scores. *Appl. Statist.*, 43 :410–414, 1994.

[92] R. Koenker and K. Hallock. Quantile regression : An introduction. *Journal of Economic Perspectives*, 15 :143–156, 2001.

[93] R. Koenker and A.F. Machado. Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.*, 94 :1296–1310, 1999.

[94] R. Koenker, P. Ng, and S. Portnoy. Quantiles smoothing splines. *Biometrika*, 81(4) :673–680, 1994.

[95] R. Koenker and B. Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71 :265–283, 1996.

[96] R. Koenker and S. Portnoy. Quantile regression. *Unpublished Manuscript, University of Illinois*, 1999.

[97] I. Komunjer. Quasi-maximum likelihood estimation for conditional quantiles. *Journal of Econometrics, Elsevier*, 128(1) :137–164, 2005.

[98] S. Kwon, H. Choi, and Y. Kim. Quadratic approximation on scad penalized estimation. *Computational Statistics and Data Analysis*, 55(1) :421–428, 2011.

[99] S. Lambert-Lacroix and L. Zwald. Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics Vol*, 5 :1015–1053, 2011.

[100] S. Lambert-Lacroix and L. Zwald. The berhu penalty and the grouped effect. ArXiv :1207.6868v1 [math.ST], 2012.

[101] D. Lamparter. Stability selection for error control in high-dimensional regression. Master's thesis, 2011. Seminar For Statistics.

[102] E. Lehmann and G. Casella. *Theory of point estimation*. Second edition, Springer, 1998.

[103] Y. Li and J. Zhu. l_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17 :163–185, 2008.

[104] Y. Liu and Y. Wu. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computation and Graphical Statistics*, 16 :782–798, 2007.

[105] K. Madsen and H.B. Nielsen. A finite smoothing algorithm for linear l_1 estimation. *SIAM J. Optimization*, 3 :223235, 1993.

[106] M.A. Mansur, A. Bhadra, H. Takamura, and T. Matoba. Volatile flavor compounds of some sea fish and prawn species. *Fish.Sci*, 69 :864–866, 2003.

[107] T. Maozai. A quantile regression analysis of family background factor effects on mathematical achievement. *Journal of Data Science*, 4 :461–478, 2006.

[108] N. Meinshausen and P. Buhlmann. Stability selection. *Journal of the Royal statistical Society, B*, 72(4) :417–473, 2010.

[109] T. Neocleous and S. Portnoy. On monotonicity of regression quantile functions. *Statistics and Probability Letters*, 78(10) :1226–1229, 2008.

[110] W.K. Newey and J.L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4) :819–847, 1987.

[111] N.M. Neykov, P. Cizek, P. Filzmoser, and P.N. Neytchev. The least trimmed quantile regression. *Computational Statistics and Data Analysis*, 56 :1757–1770, 2012.

[112] G. Olafsdottir, R. Jonsdottir, H.L. Lauzon, J. Luten, and K. Kristbergsson. Characterization of volatile compounds in chilled cod (gadus morhua) fillets by gas chromatography and detection of quality indicators by an electronic nose. *J.Agric.Food.Chem.*, 53 :10140–10147, 2005.

[113] M.R Osborne and B.A. Turlach. A homotopy algorithm for the quantile regression lasso and related piecewise linear problems. *Journal of Computational and Graphical Statistics*, 20(4) :972–987, 2011.

[114] A.B. Owen. A robust hybrid of lasso and ridge regression. Technical report, 2006.

[115] E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365) :105–121, 1979.

[116] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(02) :186–199, June 1991.

[117] S. Portnoy. Nearly root-n approximation for regression quantile processes. *The Annals of Statistics*, 40(3) :1714–1736, 2012.

[118] S. Portnoy and R. Koenker. The gaussian hare and the laplacian tortoise : Computability of squared-error versus absolute-error estimators, with discusssion. *Stat. Science*, 12 :279–300, 1997.

[119] B. Procházka. Regression quantiles and trimmed least squares estimator in the nonlinear regression model. *Computational Statistics and Data Analysis*, 6(4) :385–391, 1988.

[120] C. Prost, A. Hallier, M. Cardinal, T. Serot, and P. Courcoux. Effect of storage time on raw sardine (sardina pilchardus) flavor and aroma quality. *J.Food.Sci*, 69 :198–204, 2004.

[121] L. Qing, L. Nan, and Xi. Ruibin. Bayesian regularized quantile regression. *Bayesian Anal.*, 5(3) :533–556, 2010.

[122] Development Core Team. R. R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2012.

[123] P. Radchenko and G.M. James. Variable inclusion and shrinkage algorithms. *Journal of the americain statistical association*, 103(483) :1304–1315, 2008.

[124] R. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton, NJ : Princeton University Press, 1970.

[125] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Ann. Statist*, 35 :1012–1030, 2007.

[126] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3) :1012–1030, Jun 2007.

[127] D. Ruppert and R.J. Carroll. Trimmed least squares estimation in the linear model. *J.Amer.Statist.Assoc*, 75 :828–838, 1980.

[128] E.D. Schifano, R.L. Strawderma, and M.T. Wells. Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics*, 4 :1258–1299, 2010.

[129] G. Schwarz. Estimating the dimension of a model. *The annals of Statistics*, 6 :461–464, 1978.

[130] R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[131] R. Shah and R. Samworth. Variable selection with error control : Another look at stability selection. *J. Roy. Statist. Soc, Ser. B*, pages 55–80, 2012.

[132] I. Sidi Zakari, A. Mkhadri, and A. N'Guessan. A mixture of local and quadratic approximation variable selection algorithm in nonconcave penalized regression. *ARIMA journal*, 16 :29–46, January 2013.

[133] I. Sidi Zakari, A. Mkhadri, and A. N'Guessan. Stability selection and randomization in l_1 quantile regression. In *to appear in Proceedings of the 15th Applied Stochastic Models and Data Analysis International Conference, June, Spain*, 2013.

[134] M. Slawski. The structured elastic net for quantile regression and support vector classification. *Statistics and Computing*, 22, 153-168 2012. DOI 10.1007/s11222-010-9214-z.

[135] Z. Songfeng. Boosting based conditional quantile estimation for regression and binary classification. *Lecture Notes in Computer Science*, 6438 :67–79, 2010.

[136] J.F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. package, 1999.

[137] I. Takeuchi, Q.Y. Le, T. Sears, and A.J. Smola. Nonparametric quantile regression. *Journal of Machine Learning Research*, 7 :1001–1032, 2005.

[138] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal statistical Society, B*, 58 :267–288, 1996.

[139] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R.J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society*, 74(Part 2) :245–266, 2012.

[140] B. Turlach, W. Venables, and S. Wright. Simultaneous variable selection. *Technometrics*, 47 :349–363, 2005.

[141] H. Wang and C. Leng. Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, 102 :1039–1048, 2007.

[142] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics*, 25(3) :347–355, 2007.

[143] L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16 :589–615, 2006.

[144] L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3) :412–419, 2008.

[145] S. Wang, B. Nan, S. Rosset, and J. Zhu. Random lasso. *Annals of Applied Statistics*, 5(1) :468–485, 2011.

[146] Y. Wei, A. Pere, R. Koenker, and X. He. Quantile regression methods for reference growth charts. *Stat Med.*, 25(8) :1369–82, April 2006.

[147] R.L. Wierda, G. Fletcher, L. Xu, and J.P. Dufour. Analysis of volatile compounds as spoilage indicators in fresh king salmon (oncorhynchus tshawytscha) during storage using spme-gc-ms. *J.Agric.Food.Chem*, 54 :8480–8490, 2006.

[148] Y. Wu and Y. Liu. Variable selection in quantile regression. *Statistica Sinica*, 19 :801–817, 2009.

[149] Q. Yao and H. Tong. Asymmetric least squares regression estimation : a nonparametric approach. *Journal of nonparametric statistics*, 6(2-3) :273–292, 1996.

[150] Y. Yao and Y. Lee. Another look at linear programming for feature selection via methods of regularization. Technical report, Technical Report 800, Department of Statistics, The Ohio State University, 2007.

[151] K. Yu and R.A. Moyeed. Bayesian quantile regression. *Statist. Probab. Letters*, 54 :437–447, 2001.

[152] M. Yuan. Gacv for quantile smoothing splines. *Computational Statistics and Data Analysis*, 5 :813–829, 2006.

[153] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68 :49–67, 2006.

[154] Y. Yuan and G. Yin. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66 :105–114, 2010.

[155] C.H. Zhang. Nearly unbiased variable selection minimax concave penalty. *Annals of Statistics*, 38 :894–942, 2010.

[156] H. Zhang, Y. Liu, Y. Wu, and J. Zhu. Variable selection for multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2 :149–167, 2008.

[157] P. Zhao and B. Yu. Boosted lasso. Technical report, Journal of Machine Learning Research, 2004.

[158] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *In Advances in Neural Information Processing Systems*, 16, 2003.

[159] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 :1418–1429, 2006.

[160] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67 :301–320, 2005.

[161] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist*, 36(4) :1509–1533, 2008.

[162] H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36 :1108–1126, 2008.

[163] H. Zou and M. Yuan. The f_∞-norm support vector machine. *Statistica Sinica*, 18 :379–398, 2008.

[164] H. Zou and M. Yuan. Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics and Data Analysis*, 52 :5296–5304, 2008.

[165] H. Zou and H.H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4) :1733–1751, 2009.