Conditional quantile estimation through optimal quantization

Isabelle Charlier

Thèse en cotutelle présentée à

l'Université de Bordeaux

et à l'Université Libre de Bruxelles

pour obtenir le grade de

Docteur

Spécialité : Mathématiques appliquées

École doctorale de Mathématiques et Informatique de Bordeaux
École doctorale de Statistiques et Sciences Actuarielles de l'ULB

Conditional quantile estimation through optimal quantization

par Isabelle Charlier

sous la direction de Davy Paindaveine / Jérôme Saracco

Soutenue le 17 décembre 2015

Membres du jury :

M. Rodolphe Thiebaut	*INSERM Bordeaux*	Président du Jury
M. Davy Paindaveine	*Université Libre de Bruxelles*	Directeur de thèse
M. Jérôme Saracco	*Bordeaux INP*	Directeur de thèse
Mme Benoîte de Saporta	*Université Montpellier 2*	Rapportrice
M. Keming Yu	*Brunel University London*	Rapporteur
M.Cédric Heuchenne	*Université de Liège*	Examinateur
M. Yvik Swan	*Université de Liège*	Examinateur
M. Thomas Verdebout	*Université Libre de Bruxelles*	Examinateur

Estimation de quantiles conditionnels basée sur la quantification optimale

Résumé : Les applications les plus courantes des méthodes non paramétriques concernent l'estimation d'une fonction de régression (i.e. de l'espérance conditionnelle). Cependant, il est souvent intéressant de modéliser les quantiles conditionnels, en particulier lorsque la moyenne conditionnelle ne permet pas de représenter convenablement l'impact des covariables sur la variable dépendante. De plus, ils permettent d'obtenir des graphiques plus compréhensibles de la distribution conditionnelle de la variable dépendante que ceux obtenus avec la moyenne conditionnelle.

À l'origine, la « quantification » était utilisée en ingénierie du signal et de l'information. Elle permet de discrétiser un signal continu en un nombre fini de quantifieurs. En mathématique, le problème de la quantification optimale consiste à trouver la meilleure approximation d'une distribution continue d'une variable aléatoire par une loi discrète avec un nombre fixé de quantifieurs. Initialement utilisée pour des signaux univariés, la méthode a été étendue au cadre multivarié et est devenue un outil pour résoudre certains problèmes en probabilités numériques.

Le but de cette thèse est d'appliquer la quantification optimale en norme L_p à l'estimation des quantiles conditionnels. Différents cas sont abordés : covariable uni- ou multidimensionnelle, variable dépendante uni- ou multivariée. La convergence des estimateurs proposés est étudiée d'un point de vue théorique. Ces estimateurs ont été implémentés et un package R, nommé **QuantifQuantile**, a été développé. Leur comportement numérique est évalué sur des simulations et des données réelles.

Mots-clés : Régression quantile, Quantification optimale, Estimation non paramétrique.

Conditional quantile estimation through optimal quantization

Abstract : One of the most common applications of nonparametric techniques has been the estimation of a regression function (i.e. a conditional mean). However it is often of interest to model conditional quantiles, particularly when it is felt that the conditional mean is not representative of the impact of the covariates on the dependent variable. Moreover, the quantile regression function provides a much more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function.

Originally, the "quantization" was used in signal and information theories since the fifties. Quantization was devoted to the discretization of a continuous signal by a finite number of "quantizers". In mathematics, the problem of optimal quantization is to find the best approximation of the continuous distribution of a random variable by a discrete law with a fixed number of charged points. Firstly used for a one-dimensional signal, the method has then been developed in the multi-dimensional case and extensively used as a tool to solve problems arising in numerical probability.

The goal of this thesis is to study how to apply optimal quantization in L_p-norm to conditional quantile estimation. Various cases are studied: one-dimensional or multidimensional covariate, univariate or multivariate dependent variable. The convergence of the proposed estimators is studied from a theoretical point of view. The proposed estimators were implemented and a R package, called **QuantifQuantile**, was developed. Numerical behavior of the estimators is evaluated through simulation studies and real data applications.

Keywords : Quantile regression, Optimal quantization, Nonparametric estimation.

Centre de recherche INRIA Bordeaux Sud-Ouest
200, Avenue Vieille Tour
33 405 Talence
France

Institut de Mathématiques de Bordeaux
Université de Bordeaux
351, Cours de la Libération
33 405 Talence
France

Université Libre de Bruxelles
Département de Mathématique
Boulevard du Triomphe
Campus Plaine, CP 210
1050 Brussels
Belgium

Université Libre de Bruxelles
ECARES
Avenue F.D. Roosevelt, 50, CP114/04
1050 Brussels
Belgium

Ne confonds pas ton chemin avec ta destination. Ce n'est pas parce que c'est orageux aujourd'hui que cela signifie que tu ne te diriges pas vers le soleil.

Anthony Fernando

Remerciements

Cette thèse représente l'aboutissement de plusieurs années de travail, marquées par un apprentissage et des expériences plus que variés. C'est pourquoi, avant de rentrer dans le vif du sujet, je tiens à remercier toutes les personnes qui m'ont entourée et qui ont contribué à l'accomplissement de ce doctorat.

Ce travail a été réalisé sous la direction de Messieurs les Professeurs Davy Paindaveine et Jérôme Saracco. Tout d'abord, je remercie Davy d'avoir accepté d'encadrer ma thèse, d'avoir en premier abordé la possibilité d'une cotutelle et d'avoir pris contact avec Jérôme à ce sujet. Je remercie ensuite Jérôme d'avoir accepté de m'encadrer également, sans jamais m'avoir rencontrée auparavant. Merci à tous les deux de m'avoir proposé ce sujet passionnant. J'ai particulièrement apprécié la grande variété qu'il offrait, demandant des résultats très mathématiques tout en ayant de nombreuses applications concrètes. Pouvoir passer de la théorie à la pratique et à l'implémentation (relativement) à ma guise était très plaisant. De plus, malgré les quelques problèmes administratifs qu'elle comporte, la cotutelle a, à mon sens, apporté un réel plus à ma thèse, stimulée par les différentes périodes passées dans l'une et l'autre universités. Merci à tous les deux de m'avoir toujours accordé du temps malgré vos emplois du temps souvent chargés, merci d'avoir été en permanence à l'écoute et de m'avoir apporté votre aide aussi souvent que nécessaire. Merci aussi pour vos relectures attentives de mes différents manuscrits. Cette thèse doit beaucoup à votre pédagogie, à la qualité de votre encadrement, à votre gentillesse et à vos encouragements. Je pense avoir beaucoup appris et évolué pendant ces trois ans et demi, et pas uniquement d'un point de vue scientifique, et c'est essentiellement à vous que je le dois. Encore merci pour tout.

Je voudrais exprimer ma gratitude à tous les membres du jury pour l'intérêt porté à ce travail. Je remercie en particulier Monsieur le Professeur Rodolphe Thiebaut d'en avoir accepté la présidence, et Madame la Professeure Benoîte de Saporta et Monsieur le Professeur Keming Yu d'avoir accepté de rapporter ma thèse. Je leur sais gré du temps qu'ils ont passé à considérer ce document.

Mes remerciements vont aussi au Fonds National pour la Recherche Scientique – F.N.R.S. (Belgique) pour le soutien financier apporté via une bourse FRIA durant ces années de thèse.

Je me tourne désormais vers mes différents collègues, à la fois de Bordeaux (et en particulier

de l'open-space de l'Inria) et de l'ULB, pour les fréquentes discussions autour d'un tableau ou d'une feuille de papier sur différents problèmes, et tout simplement parfois pour leur oreille attentive, puisqu'il suffit souvent d'énoncer et d'expliquer ce qui nous bloque à un tiers pour le voir sous un autre jour et entrevoir une solution. Mais surtout, merci aussi pour les discussions non-mathématiques et les pauses plus que nécessaires, qui ont agrémenté ce doctorat de très bons moments. Plus spécialement, merci à Romain Azaïs de m'avoir guidée dans l'apprentissage de la quantification et de m'avoir permis, par nos nombreuses discussions, de cerner au mieux ce sujet, et merci à Robin Genuer, Amaury Labenne et Adrien Todeschini de m'avoir fait profiter de leur expérience en R, plus précisément pour l'élaboration de mon package et pour la parallélisation. J'adresse aussi un merci particulier aux doctorants de l'Inria présents lors de ma première année de thèse, qui terminaient tous la leur, et qui m'ont offert, par leur accueil, leur expérience et leur bonne humeur, la meilleure entame possible pour la mienne. Enfin, merci spécialement à Audrey, Keno et Manon, pour leur amitié et le partage d'expériences depuis les bancs du bachelier à l'ULB jusqu'aux couloirs du NO.

Je tiens évidemment à remercier ma famille et mes amis de m'avoir soutenue tout au long de cette thèse. Même s'il n'était pas facile pour eux de comprendre les problèmes que je traversais, leur présence et leurs encouragements, inconditionnels et constants, étaient très importants. Merci particulièrement à mes parents, pour tout, et à ma nièce Joline, petit rayon de soleil d'un an.

Jonathan, tu es la personne qui me connaît le mieux et ton soutien le plus total et permanent me donne de la force au quotidien. Je n'en serais pas là aujourd'hui sans toi à mes côtés. Merci d'être toujours là, pour tout ce que tu m'apportes et merci de m'encourager à chaque baisse de moral. Enfin, merci à Elia dont seuls la présence et le sourire illuminent mon quotidien depuis quelques mois.

Résumé substantiel

La première partie de cette thèse a pour but de définir et d'étudier (d'un point de vue théorique dans le Chapitre 2 et numérique dans le Chapitre 3) un nouvel estimateur des quantiles conditionnels en utilisant un outil de discrétisation appelé *quantification optimale en norme* L_p.

La notion de quantile conditionnel, ou régression quantile, a tout d'abord été introduite par Koenker and Bassett (1978). Une motivation majeure était le besoin d'alternatives robustes à la moyenne conditionnelle. Les quantiles conditionnels permettent en effet de mieux représenter l'impact des variables explicatives X sur la variable dépendante Y. Depuis le travail de Koenker and Bassett (1978), qui les ont étudiés dans un contexte paramétrique, la littérature sur ce sujet s'est fortement développée ces dernières années et des estimateurs non paramétriques ont également été étudiés.

Dans la suite, nous appellerons X le vecteur de variables explicatives (de dimension d) et Y la variable dépendante, définie comme une fonction de X et d'une erreur. Dans des situations pratiques, nous ne connaissons pas la distribution conditionnelle de Y sachant $X = x$ et nous voulons l'étudier à partir d'un échantillon de taille n en estimant des fonctions de quantiles conditionnels. Un intérêt majeur de ceux-ci est qu'ils permettent de construire des courbes de référence et des intervalles de confiance. Ces graphes de référence sont largement utilisés dans de nombreux domaines comme en médecine, économie, écologie, analyse de durée de vie, etc.

Les quantiles conditionnels peuvent être définis de deux façons équivalentes. Nous avons fait le choix dans notre étude de nous baser sur la définition de ceux-ci comme solution du problème d'optimisation suivant :

$$q_\alpha(x) = \arg\min_{a\in\mathbb{R}} \mathrm{E}[\rho_\alpha(Y-a)|X=x],$$

où la fonction ρ_α est une fonction de perte qui généralise la valeur absolue.

Lorsque nous voulons estimer le quantile conditionnel sachant $X = x$ à partir d'un échantillon, nous devons sélectionner les observations dont la partie en X est proche de ce point x. Le choix de cette sélection est la clé de nombreuses méthodes d'estimation non paramétrique des quantiles. Certaines méthodes existantes utilisent par exemple un noyau ou sélectionnent les plus proches voisins. Notre méthode utilise le concept de quantification. Plus de détails sur cette notion se trouvent dans la Section 1.1. Avant d'expliquer l'avantage de notre méthode de sélection sur d'autres méthodes, commençons par définir précisément ce concept.

Dès les années 1950, la quantification optimale était utilisée en ingénierie et permettait de discrétiser un signal continu grâce à un nombre fini de quantifieurs. Ces quantifieurs doivent être placés de manière à transmettre le signal aussi efficacement que possible, d'où le côté optimisation de la quantification. En mathématique, cela consiste à trouver la meilleure approximation de la distribution d'un vecteur aléatoire (continue, le plus souvent) par une loi discrète. Le support de cette loi discrète porte le nom de grille, et est de taille N. Nous projetons alors le vecteur sur cette grille pour le discrétiser. Un résultat classique en quantification assure l'existence d'une grille optimale (c'est-à-dire, une grille minimisant l'erreur de projection commise) à condition que la loi de ce vecteur ne charge pas les hyperplans. Lorsque nous considérons une grille de quantification fixée, nous pouvons définir des cellules de quantification, caractérisées par un centre. À chaque point de la grille correspond une cellule dont le centre est ce point, et cette cellule est composée de l'ensemble des points de l'espace dont la projection sur la grille est le centre de cette cellule. Plus de détails sur ce concept peuvent être trouvés dans la Section 1.2.

Le point de départ du travail a consisté à remplacer dans la définition des quantiles conditionnels le vecteur de variables explicatives X par une version discrétisée obtenue par quantification optimale (cf. Section 2.2). Ceci nous a permis de définir une approximation des quantiles conditionnels, et la convergence de cette approximation a été étudiée lorsque N tend vers l'infini. Deux résultats de convergence ont été prouvés. Tout d'abord, un résultat de consistance à x fixé a été obtenu (cf. Théorème 2.2.2). Plus précisément, nous avons montré que notre approximation converge uniformément en x vers les quantiles conditionnels théoriques. Nous avons également étudié la convergence de cette approximation en X, et dans ce cas nous avons obtenu une vitesse de convergence (cf. Théorème 2.2.1).

Dans un second temps, nous avons défini dans la Section 2.3 un estimateur des quantiles conditionnels en prenant une version empirique de cette approximation. Partant d'un échantillon (X_i, Y_i) de taille n, le quantile conditionnel de Y sachant $X = x$ est alors estimé en pratique simplement en prenant le quantile empirique des Y_i pour lesquels les X_i correspondants sont projetés sur le même point de la grille optimale que le point x (en d'autres mots, ceux pour lesquels les X_i appartiennent à la même cellule que le x considéré). La convergence en probabilité de cet estimateur vers notre approximation a également été obtenue lorsque la taille de l'échantillon n tend vers l'infini et pour N fixé (cf. Théorème 2.3.1). Cependant, si ces résultats asymptotiques étaient très intéressants, il est apparu en pratique que, pour une petite taille d'échantillon, les courbes correspondantes n'étaient pas suffisamment lisses. Ceci était principalement dû au fait que l'algorithme utilisé pour construire notre grille de quantification optimale ne tournait pas suffisamment longtemps que pour que la grille fournie soit effectivement optimale. Pour cette raison, nous avons défini une version bootstrap de notre estimateur, ce qui en a amélioré significativement les performances. Une illustration de cette amélioration apparaît dans la Figure 2.2. Les courbes de quantiles conditionnels des graphes de gauche ont été estimées en utilisant la méthode initiale, et celles de droite en utilisant la version bootstrap. Au vu de cette amélioration, nous n'utiliserons plus dans le suite que la version bootstrap. Les preuves des théorèmes ont été regroupées dans les Sections 2.4 et 2.5. Un premier article reprenant ces différents résultats et définitions a été publié dans le Journal of Statistical Planning and Inference (Charlier et al., 2015a).

Après cette étude théorique, il était important de se pencher sur le comportement numérique de notre estimateur. Ceci fait l'objet du Chapitre 3 et d'un deuxième papier qui a été publié dans Computational Statistics and Data Analysis (Charlier et al., 2015b). Nous avons d'abord mis au point une méthode de sélection du paramètre N intervenant dans la construction de celui-ci, méthode n'utilisant que les observations (cf. Section 3.2). Cette méthode a démontré son efficacité sur 500 réplications de différents modèles et pour différentes tailles d'échantillon. Nous avons donc été en mesure de comparer les estimations fournies par notre estimateur bootstrap avec celles d'autres estimateurs classiques des quantiles conditionnels (cf. Section 3.3). À nouveau, nous avons généré 500 réplications de divers modèles et observé de très bons résultats. Plus précisément, nous avons noté de très bonnes performances de notre estimateur face aux compétiteurs classiques, et parfois une domination sur son principal concurrent, l'estimateur de type spline, plus particulièrement quand la fonction de lien du modèle est complexe. Les figures de la Section 3.3.3 illustrent ces différentes remarques et représentent les boxplots des ISE (Integrated Squared Error), une mesure de l'erreur commise. Notre estimateur bootstrap est en bleu, un estimateur de type spline en orange, un estimateur de type plus proches voisins est en vert, et deux estimateurs de type noyau sont en mauve et en rouge. Ces bons résultats s'expliquent notamment par le fonctionnement même de notre estimateur, et plus précisément par le choix des observations prises en compte lors de la sélection. En effet, notre méthode sélectionne les observations pour lesquelles la composante X_i est dans la même cellule de quantification que le point x considéré. Cela présente l'avantage d'être adaptatif avec x. D'une part, le nombre de points sélectionnés dépend de la taille de la cellule et varie donc avec x, au contraire d'une sélection de type plus proches voisins où le nombre de voisins est fixé. D'autre part, le « rayon » de la cellule varie également avec x (c'est-à-dire la distance entre le centre et le point de la cellule qui lui est le plus éloigné), au contraire des méthodes de type noyau où la taille de fenêtre est le plus souvent choisie de manière identique pour tout x.

L'étude numérique et les simulations de ces sections se concentrent sur le cas $d = 1$, d étant la dimension de X. Nous étudions de plus brièvement dans la Section 3.4 le cas $d = 2$ puisque tous les définitions et résultats théoriques sont valides en dimension quelconque.

Un package, nommé **QuantifQuantile**, reprenant notre méthode d'estimation a été développé sous R et est désormais en libre accès sur le CRAN, ce qui assure la disponibilité de cet outil à tout utilisateur de R. Ce package est composé de plusieurs fonctions, à choisir en fonction de la dimension du vecteur X. Ces fonctions fournissent des estimations des quantiles conditionnels après avoir sélectionné automatiquement la valeur optimale pour N grâce à notre méthode. De plus, une représentation graphique des courbes (dans le cas $d = 1$) ou surfaces (dans le cas $d = 2$) de quantiles conditionnels peut être obtenue via la fonction `plot`. Tous les détails sur ce package et ses différentes fonctionnalités se trouvent dans la Section 3.5. Un troisième papier expliquant le fonctionnement de ce package et fournissant des exemples et illustrations a été accepté pour publication au R journal (Charlier et al., 2015c).

L'approche numérique de la Section 3.3 a été complétée par une application sur des échantillons de données réelles dans la Section 3.6.

Après cette première partie permettant de caractériser le lien entre une variable dépendante unidimensionnelle et un vecteur de covariables de dimension quelconque, il était naturel de s'in-

téresser au cas d'une variable dépendante de dimension m, avec $m > 1$. Ceci a fait l'objet de la Partie II.

Cependant, si la notion de quantile est standard pour une variable réelle, il n'en est pas de même en dimension supérieure. En effet, cette notion repose sur l'ordre naturel dans les réels et la généralisation à un contexte multivarié n'est donc pas directe. Une littérature assez vaste a été consacrée à l'extension de ce concept au cadre multivarié. Nous présentons et comparons dans la Section 4.1 différentes notions de quantiles multivariés existantes afin de choisir celle qui sera utilisée dans cette partie. Notre choix s'est porté sur celle définie par Hallin, Paindaveine, and Šiman (2010a), qui se démarque des autres en satisfaisant toutes les propriétés généralement attendues de la part d'un concept de quantiles, comme des propriétés d'affine-équivariance, tout en mettant en exergue un lien avec l'important concept de profondeur statistique.

La définition des quantiles multivariés ayant été fixée (et leur version conditionnelle s'obtenant directement), nous avons généralisé les résultats du Chapitre 2 dans ce contexte à réponse multiple. Plus précisément, nous avons remplacé dans la définition des quantiles conditionnels à réponse multiple le vecteur de variables explicatives X par sa version discrétisée obtenue par quantification optimale (cf. Section 5.2). Ceci a à nouveau permis de définir une approximation des quantiles conditionnels qui a été étudiée lorsque N tend vers l'infini. Nous avons prouvé la consistance à x fixé de cette approximation (cf Théorème 5.2.1).

Dans un second temps, nous en avons déduit un estimateur des quantiles conditionnels à réponse multiple dans la Section 5.3 en prenant une version empirique de cette approximation. L'estimation du quantile conditionnel de Y sachant $X = x$ est en pratique obtenue comme dans le cas d'une réponse scalaire : en prenant le quantile multivarié empirique des Y_i pour lesquels les X_i correspondants sont projetés sur le même point de la grille optimale que le x considéré. Nous avons obtenu la convergence en probabilité de cet estimateur vers notre approximation à N fixé et pour $n \to \infty$ (cf. Théorème 5.3.1). Vu l'amélioration obtenue dans la première partie en définissant une version bootstrap de l'estimateur, nous avons fait de même dans ce contexte. Les preuves des Sections 5.2 et 5.3 ont été regroupées respectivement dans les Sections 5.4 et 5.5.

Nous nous sommes alors penchés sur le comportement numérique de cet estimateur dans le Chapitre 6. Suivant le même plan que dans le Chapitre 3, il a d'abord fallu développer une méthode de sélection du paramètre N. Celle-ci est simplement une extension de celle utilisée dans le cas d'une réponse scalaire. La différence majeure consiste en l'ajout d'une moyenne car l'ordre des quantiles est maintenant également composé d'une direction $u \in \mathcal{S}^{d-1}$, l'hypersphère unité dans $\mathbb{R}^d$. L'estimateur que nous proposons a été implémenté dans R. Avec cette méthode, nous avons donc été en mesure de comparer notre estimateur aux alternatives existantes (cf. Section 6.3). Étant donné que la notion de quantile multivarié n'est pas standard, la littérature en régression quantile multivariée est bien moins importante que précédemment. Nous nous sommes donc uniquement comparés aux estimateurs de type noyau introduits dans Hallin, Lu, Paindaveine, and Šiman (2015). Cette étude comparative a été réalisée en deux temps : d'abord d'un point de vue graphique, où nous avons représenté pour chaque méthode les contours de quantiles estimés, et ensuite du point de vue de l'erreur commise, en calculant des ISE empiriques. Nous avons à nouveau noté de très bonnes performances de notre estimateur qui domine souvent ses compétiteurs de type noyau. Cette observation généralise donc les résultats obtenus pour une

réponse scalaire.

Une application sur un jeu de données réelles a terminé cette étude numérique dans la Section 6.4. Le jeu de données considéré avait également été étudié par nos compétiteurs, ce qui a permis de compléter l'étude comparative réalisée précédemment et de confirmer les bons résultats de notre estimateur. Un article reprenant les résultats des Chapitres 5 et 6 est en cours de rédaction.

Enfin, le Chapitre 7 fournit une conclusion et discute des perspectives de ce travail.

Contents

1 Introduction

Contents

In this work, we develop a new nonparametric estimator of conditional quantiles, based on optimal quantization. The first part is dedicated to single-output conditional quantile regression, while the second part aims to extend the results in a multiple-output framework. In each part, this work will be twofold. First, a theoretical investigation of the estimator is realized, and uses classical results in optimal quantization. Then, numerical simulations are performed, since it is crucial to evaluate the finite-sample performances of our estimator for practical purposes.

This first chapter aims to define the key concepts of this thesis. We first recall the notion of quantiles (ordinary and conditional) and optimal quantization.

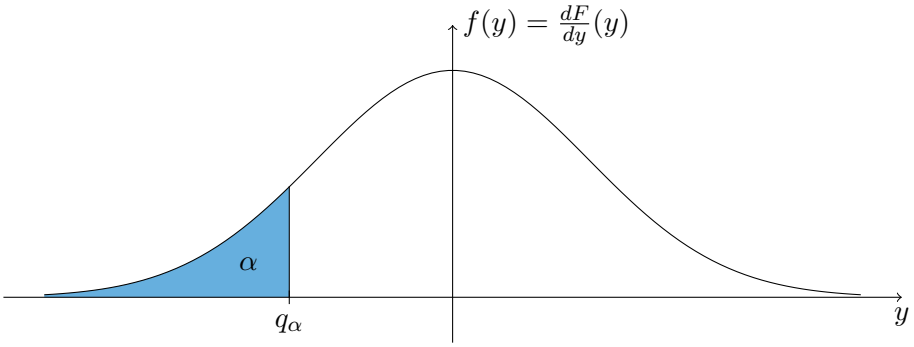


Figure 1.1 – The 0.3-quantile for the Normal distribution $\mathcal{N}(0,1)$, where f is the probability density function (pdf) of this distribution.

1.1 The notion of conditional quantiles

The notion of *conditional quantiles*, or *regression quantiles*, was first introduced by Koenker and Bassett (1978) in the linear model. These conditional quantiles are the natural generalization of ordinary quantiles when the interest variable is a dependent variable. A major motivation was the need for robust alternatives to the conditional mean and the least square estimator. As one will explain in more details at the end of this section, conditional quantiles are particularly interesting to model when it is felt that the conditional mean is not representative of the impact of the covariates on the dependent variable that one considers (see Figure 1.3). Indeed, conditional quantiles enable us to explore more thoroughly heterogeneous covariates effects. The literature on quantile regression became really large in recent years. The work of Koenker and Bassett (1978) introduced it in a parametric framework and since then it was investigated in other frameworks (see, e.g., Koenker (2005) for a review).

We first define ordinary quantiles (see Section 1.1.1) and we then explain how to generalize this notion to a framework with a dependent variable in Section 1.1.2. We then conclude this section on conditional quantiles with the importance of estimating them and some examples of applications (see Section 1.1.3).

1.1.1 Definition of ordinary quantiles

Let Y be a real random variable with cumulative distribution function (cdf) F.

Definition 1.1.1 (Indirect characterization). *For $\alpha \in (0,1)$, the* quantile of order α *of Y is defined as the real number q_α such that $q_\alpha = \inf\{y \in \mathbb{R} : F(y) \geq \alpha\}$.*

When the cdf of Y, i.e. F, is invertible, it can be rewritten in a simpler way as $q_\alpha = F^{-1}(\alpha)$. This definition is illustrated in Figure 1.1.

There exists another equivalent definition of the α-quantiles, where they are seen as the solutions of an optimization problem. Let us motivate this definition.

Let us consider the expectation $\mathrm{E}[Y]$. It is well known as the solution of the optimization problem $\mathrm{E}[Y] = \arg\min_{a\in\mathbb{R}} \mathrm{E}[(Y-a)^2]$. Similarly, the median is the solution of $\mathrm{med}(Y) = \arg\min_{a\in\mathbb{R}} \mathrm{E}[|Y-a|]$ and these two results are easily obtained by derivation. This property can be

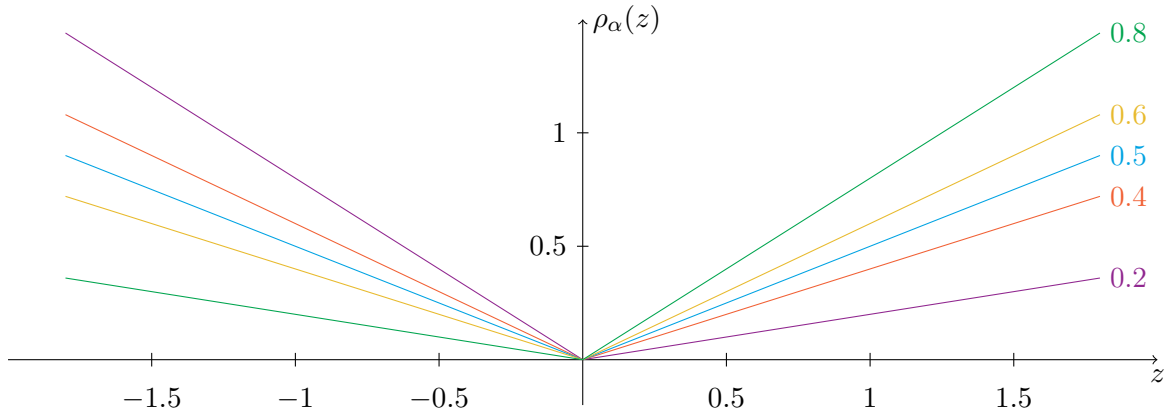


Figure 1.2 – The convex loss function $\rho_\alpha(z)$ for different values of α: in purple, $\alpha = 0.2$; in orange, $\alpha = 0.4$; in blue, $\alpha = 0.5$ (the absolute value function with a factor of $1/2$); in yellow, $\alpha = 0.6$ and in green, $\alpha = 0.8$.

generalized for the quantiles: since the median is nothing else that the $1/2$-quantile, one can show that the α-quantiles are the solutions of the optimization problem $q_\alpha = \arg\min_{a\in\mathbb{R}} \mathrm{E}[\rho_\alpha(Y-a)]$ where $\rho_\alpha(z)$ is a convex loss function that generalizes the absolute value appearing in the median case. Let us denote by $\mathbb{I}_A(x) = \mathbb{I}_{[x\in A]}$ the indicator function of x belonging to A.

Definition 1.1.2 (Direct characterization)**.** *The* quantile of order α *of* Y *can be equivalently defined as the solution of the following optimization problem:*

$$q_\alpha = \arg\min_{a\in\mathbb{R}} \mathrm{E}[\rho_\alpha(Y-a)],$$

where $\rho_\alpha(z) = -(1-\alpha)z\mathbb{I}_{[z<0]} + \alpha z\mathbb{I}_{[z\geq 0]} = z\big(\alpha - \mathbb{I}_{[z<0]}\big)$ *is called the* check function.

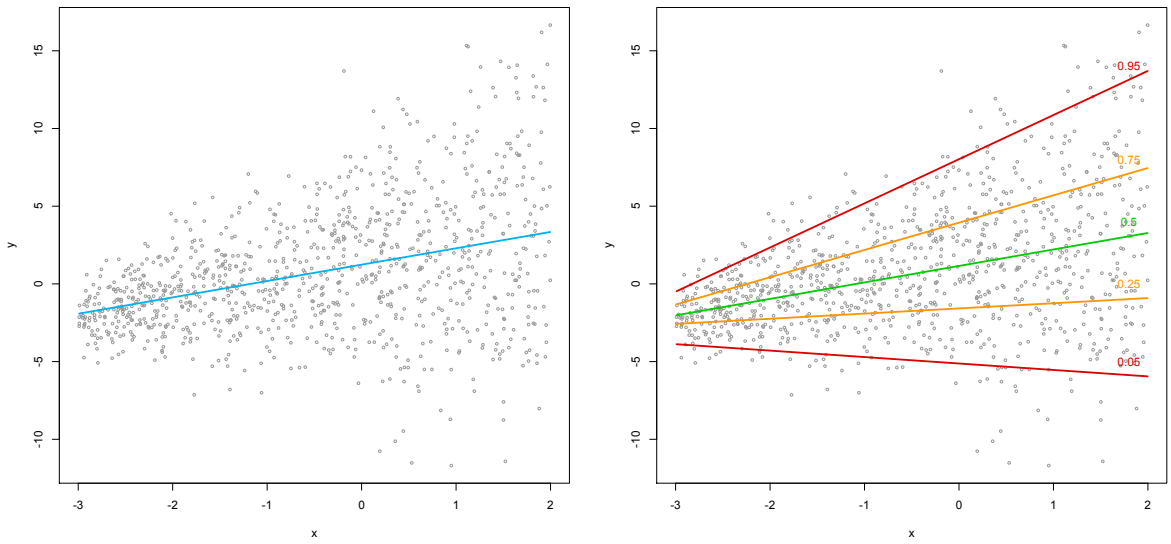
This check function also rewrites as $\rho_\alpha(z) = \big(|z| + (2\alpha-1)z\big)/2$ and this rewriting points out the link with the absolute value for the median case. Indeed, taking $\alpha = 0.5$ gives $\rho_{0.5}(z) = |z|/2$. The function ρ_α is thus a generalization of the absolute value when α is between 0 and 1 (up to a factor of $1/2$). This convex loss function is represented in Figure 1.2 for different values of α.

1.1.2 Definition of conditional quantiles

However, it happens regularly that there exist some variables providing information on the variable Y that one considers. These variables are called *covariates* and have to be used when one investigates the behavior of Y. It is then classical to consider the conditional mean and variance functions

$$x \mapsto \mathrm{E}[Y|X=x] \quad \text{and} \quad x \mapsto \mathrm{Var}[Y|X=x].$$

However, a more thorough picture can be obtained by considering some extension of the notion of quantile functions in this setting, for various $\alpha \in (0,1)$. This motivates us to define *conditional quantiles* of Y given $X = x$, where X is a covariate taking values in $\mathbb{R}^d$. The conditional quantiles are defined analogously to the standard case: one replaces the cdf F and the expectation by their conditional versions in Definitions 1.1.1 and 1.1.2 respectively. Denote by $F(y|x)$ the conditional distribution of Y given $X = x$. The conditional quantiles are defined as follows.



(a) (b)

Figure 1.3 – (a) The conditional mean curve of Y given X is represented in blue. (b) Five different conditional quantile curves: in green, $\alpha = 0.5$, in orange, $\alpha = 0.25$ for the lower one and α =0.75 for the upper one, in red, $\alpha = 0.05$ for the lower one and α =0.95 for the upper one.

Definition 1.1.3 (Indirect characterization)**.** *For $\alpha \in (0,1)$, the* conditional quantile of order α *of Y given $X = x$ is defined as the number $q_\alpha(x)$ such that $q_\alpha(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \alpha\}$.*

Here again, it can be more simply rewritten as $q_\alpha(x) = F^{-1}(\alpha|x)$ when $F(\cdot|x)$ is invertible. The following definition gives us the conditional quantiles as solution of an optimization problem and is then analogous to Definition 1.1.2.

Definition 1.1.4 (Direct characterization)**.** *The* conditional quantile of order α *of Y given $X = x$ can be equivalently defined as the solution of the following optimization problem:*

$$q_\alpha(x) = \arg\min_{a \in \mathbb{R}} \mathrm{E}[\rho_\alpha(Y - a)|X = x].$$

When one considers these conditional quantiles with fixed x, they provide a confidence interval for Y given $X = x$. If one considers them when x varies, they provide reference curves (if $d = 1$, hypersurfaces otherwise). More precisely, when x is fixed, the interval $I_\alpha(x) = [q_{1-\alpha}(x), q_\alpha(x)]$ contains $100(2\alpha - 1)\%$ if $\alpha > 1/2$ and the reference curves are defined by the sets of points $\{(x, q_{1-\alpha}(x))\}$ and $\{(x, q_\alpha(x))\}$.

1.1.3 Estimation of conditional quantiles

In practice, one does not know the conditional distribution $F(\cdot|x)$ and one wants to investigate it from a sample $\{(X_i, Y_i)\}$ of size n. The aim of quantile regression is thus to quantify the relationship between Y and X thanks to conditional quantiles. A major motivation is that a collection of conditional quantiles gives us an impression of the entire conditional distribution

when graphed. Therefore, it is a more complete and informative method than standard regression which fits only the average relationship between them. It can then happen that some changes in shape will not be displayed using standard regression. In particular, standard regression does not allow to detect heteroscedasticity. This fact is illustrated in Figure 1.3. One has represented the same data points in both subfigures. In the left one, the blue curve represents the conditional mean of Y given X. In the right one, the five curves represent conditional quantile curves for different values of α. It is obvious that the conditional quantile curves allow us to quantify in a better way the shape, the location and the propagation of the data. Another important interest of conditional quantiles is that they allow us to construct reference curves and conditional prediction intervals within which one expects the majority of points to lie, as explains after Definition 1.1.4.

These reference charts are widely used in many different areas, as in medicine, economics, ecology, lifetime analysis, etc. Let us illustrate it on two examples. Figure 1.4 represents the reference curves for the weight given the age of boys from birth to 5 years. This graph comes from the World Health Organization[1] and are incorporated in national health cards in several countries. These child growth standards were developed using data collected by the WHO. It allows the parents to control the growth of their child (their weight here, but similar charts exist for height or head circumference).

Figure 1.5 provides another application of conditional quantiles for the annual compensation in millions of Chief Executive Officers (CEO) given the firm market value in billions. Instead of looking for conditional quantile curves, this example provides boxplots of the compensation given the firm market value. To do so, the covariate (i.e. the firm market value) was clustered into ten groups, and then, looking at each group separately, the boxplot of the corresponding compensations was calculated. For each boxplot, the horizontal lines delimiting the dotted line correspond to the 0.1-quantile (lower one) and 0.9-quantile (upper one), while the lower and upper limits of the box represent the quantiles of order 0.25 and 0.75 respectively. The horizontal line in each box is the median (i.e. the 0.5-quantile). This figure comes from Koenker and Hallock (2001) and allows to put the light on some features, as a tendency for compensation to rise with firm size.

There exist several methods to estimate conditional quantiles. As above mentioned, quantile regression was introduced in the seminal paper of Koenker and Bassett (1978) where the focus was on linear regression. Since then, the literature became very large on this subject and more general regression frameworks were considered, as the nonparametric one. Kernel and nearest-neighbor estimators of conditional quantiles were investigated in Bhattacharya and Gangopadhyay (1990), while Yu and Jones (1998) focused on local linear quantile regression and double-kernel approaches. Koenker et al. (1994) and He et al. (1998) developed spline-based methods. Many other estimators were also considered; we refer to Fan et al. (1994), Hallin et al. (2009), Gannoun et al. (2002), Heagerty and Pepe (1999), or Yu et al. (2003), and Koenker (2005) for a review. In this work, we introduce a new nonparametric quantile regression method, based on *optimal quantization*. This concept is defined in the following subsection.

[1]The WHO child growth standards: `http://www.who.int/childgrowth/standards/en/`

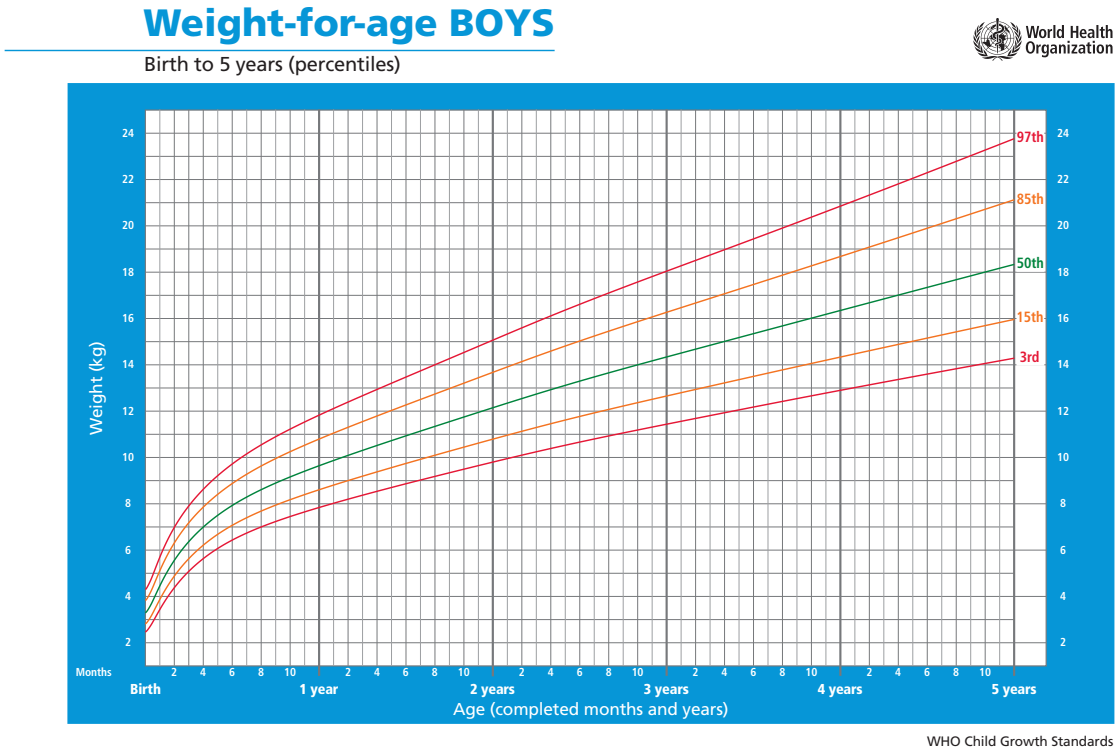


Figure 1.4 – The World Health Organization child growth standards for weight given the age of boys from birth to five years. The conditional quantile curves considered here are for $\alpha = 0.03$, 0.15, 0.5, 0.85 and 0.97 (from bottom to top).

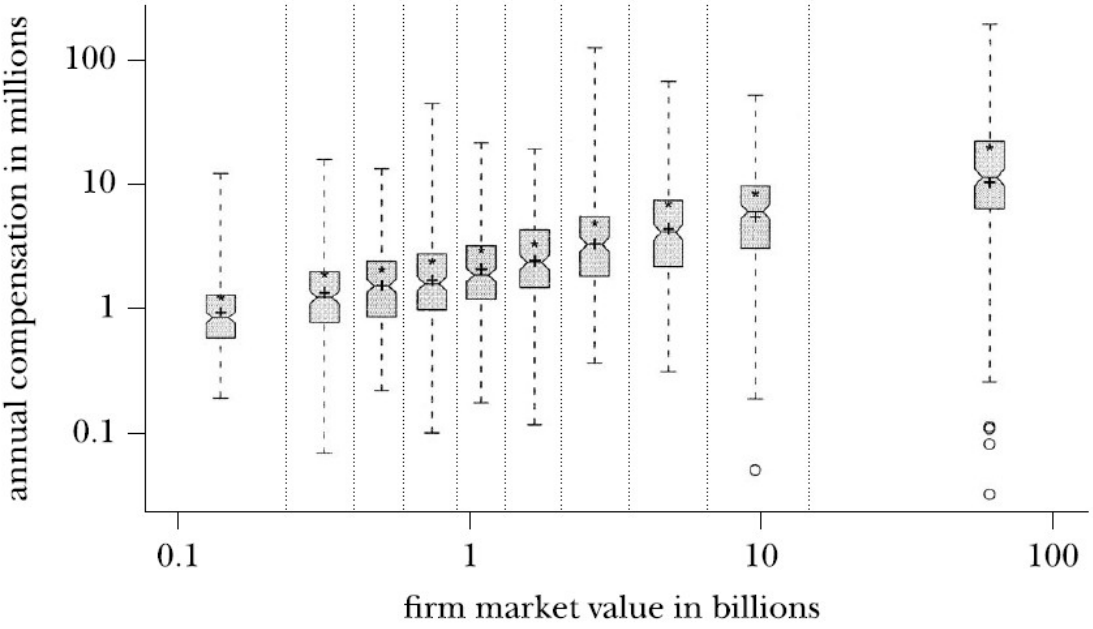


Figure 1.5 – Boxplot of annual compensation of CEO given 10 clusterings of firms ranked by market value. This graph comes from Koenker and Hallock (2001).

1.2 L_p-norm optimal quantization

We now give an overview on the concept of L_p-norm optimal quantization. After a brief description of the origin of quantization, we define it precisely and give the main property and results. As we will see, the choice of the quantization grid is of major importance and there exists no theoretical result characterizing an *optimal* grid. We then discuss about an algorithm allowing to construct such a grid and about the convergence of this algorithm.

1.2.1 Origin of quantization

The term *quantization* originates in the fifties where it was mainly used in signal and information theory. In this context, the word quantization meant discretising a continuous signal using a finite number of points, called the *quantizers*. The aim was obviously to have an efficient and parsimonious transmission of the signal. Consequently, the number of quantizers (and more precisely their geometric location) has to be optimized for this purpose.

As above mentioned, this tool was first used by engineers in signal compression, but it was later used in cluster analysis, pattern and speech recognition. The quantization occurred more recently in mathematics. In this case, the problem of optimal quantization consists in finding the best approximation of the continuous distribution P_X of a d-dimensional random vector X by a discrete distribution with a fixed number N of charged points. In other words, it concerns the best approximation of X by a random vector $\widetilde{X}^N$ which has at most N values in its image. The vector $\widetilde{X}^N$ is called the *quantized version* of X and the quantization is thus a spatial discretization method. Quantization was then considered in probability theory; see, e.g., Zador (1964) or Pagès (1998) and was extensively investigated in numerical probability, finance, stochastic processes, and numerical integration ; see, e.g., Pagès et al. (2004a), Pagès et al. (2004b), or Bally et al. (2005).

Quantization, however, was seldom used in statistics. To the best of our knowledge, its applications in statistics are restricted to Sliced Inverse Regression (Azaïs et al., 2012) and clustering (Fischer, 2010, 2014). As announced above, we use in this work quantization in a nonparametric quantile regression framework.

Let us now give a more precise definition of this concept and the main properties and results. The following sections are based on Pagès (1998), Pagès and Printems (2003), Pagès et al. (2004b), Bally et al. (2005) and Graf and Luschgy (2000).

1.2.2 Definitions

Let X be a random vector on a probability space $(\Omega, \mathcal{F}, P)$ taking its values in $\mathbb{R}^d$ and denote by P_X the distribution of X on $\mathbb{R}^d$. Fix a real number $p \geq 1$ such that $\mathrm{E}[|X|^p] < \infty$ (throughout, $|\cdot|$ denotes the Euclidean norm). Optimal L_p-mean quantization consists in studying the best L_p-norm approximation of X by a random vector $X' = \pi(X)$ taking at most N values in $\mathbb{R}^d$. Our goal is then to find such a vector so that the L_p-mean quantization error $\|X - \pi(X)\|_p$ will be minimized, with $\|Z\|_p := \left(\mathrm{E}[|Z^p|]\right)^{1/p}$. This optimization problem is decomposed in two phases.

The first one consists in considering a fixed number N of values and a fixed subset of N points in $\mathbb{R}^d$, $x = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N$, that we call a *grid*. This N-tuple being set, one wants

to find a quantizer $\pi_x : \mathbb{R}^d \to \{x_1, \dots, x_N\}$ such that

$$\|X - \pi_x(X)\|_p = \inf \left\{ \|X - \pi(X)\|_p \, , \, \pi : \mathbb{R}^d \to \{x_1, \dots, x_N\} \text{ is a Borel function} \right\}.$$

The solution of this problem is geometric and consists in the projection on the closest neighbor, induced by the *Voronoi tessellation* of x.

Definition 1.2.1. *Let $x = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N$. A Borel partition $\big(C_i(x)\big)_i$, $i = 1, \dots, N$, of $\mathbb{R}^d$ is a* Voronoi tessellation *of x if, for every $i \in \{1, \dots, N\}$, $C_i(x)$ satisfies*

$$C_i(x) \subset \left\{ y \in \mathbb{R}^d : |x_i - y| = \min_{1 \le j \le N} |x_j - y| \right\}.$$

The set $C_i(x)$ is called the Voronoi cell of center x_i. This Voronoi tessellation allows us to define the closest neighbor projection.

Definition 1.2.2. *The* closest neighbor projection *or* Voronoi quantizer function π_x *induced by the Voronoi tessellation $\big(C_i(x)\big)_{i=1,\dots,N}$, is defined for every $\xi \in \mathbb{R}^d$ by*

$$\pi_x(\xi) = \sum_{1 \le i \le N} x_i \mathbb{I}_{C_i(x)}(\xi).$$

Thanks to those definitions, one is now able to define the quantized version of the random vector X.

Definition 1.2.3. *The random vector $\widetilde{X}^x = \pi_x(X) = \sum_{1 \le i \le N} x_i \mathbb{I}_{C_i(x)}(X)$ is called a* Voronoi quantization *of X. The N-tuple x is often called a N-quantizer.*

It is important to note that, whatever the choice of the Voronoi tessellation, the closure and the boundary of the ith cell $C_i(x)$ are the same. Moreover, this boundary is included into at most $N - 1$ hyperplanes. If the distribution of X, P_X, does not charge hyperplanes, i.e. if $P_X(H) = 0$ for each hyperplane H of $\mathbb{R}^d$, then all the Voronoi tessellations are P_X-equal and all the Voronoi quantizations $\widetilde{X}^x$ have the same distribution.

The second optimization problem consists in minimizing the function $x \to \|X - \widetilde{X}^x\|_p$, that is minimizing the L_p-mean quantization error. One can notice that this error (at power p) can be rewritten as follows.

$$\begin{aligned} \|X - \widetilde{X}^x\|_p^p &= \left\| X - \sum_{i=1}^N x_i \mathbb{I}_{C_i(x)}(X) \right\|_p^p = \mathrm{E}\left[\left| X - \sum_{i=1}^N x_i \mathbb{I}_{C_i(x)}(X) \right|^p \right] \\ &= \mathrm{E}\left[\left| \sum_{i=1}^N (X - x_i) \mathbb{I}_{C_i(x)}(X) \right|^p \right], \end{aligned}$$

and, given that every mixed terms are equal to 0 since the $C_i(x)$'s are a partition of $\mathbb{R}^d$, this becomes

$$\|X - \widetilde{X}^x\|_p^p = \sum_{i=1}^N \mathrm{E}\left[\left| (X - x_i) \mathbb{I}_{C_i(x)}(X) \right|^p \right] = \sum_{i=1}^N \mathrm{E}\left[|X - x_i|^p \, \mathbb{I}_{C_i(x)}(X) \right].$$

Notice that the only term in this sum that is different of 0 is the term corresponding to the index j such that $X \in C_j(x)$. However, by definition of $C_j(x)$, it means that $|X-x_j| = \min_{i=1\dots,N} |X-x_i|$, which allows to rewrite $\|X - \widetilde{X}^x\|_p^p$ as

$$\|X - \widetilde{X}^x\|_p^p = \mathrm{E}\left[\min_{1\leq i\leq N} |X - x_i|^p\right] = \int_{\mathbb{R}^d} \min_{1\leq i\leq N} |x_i - \xi|^p P_X(d\xi).$$

These definitions lead to two natural questions: does this function actually reach a minimum? How does this minimum behave as N goes to infinity? One will answer these questions in the following section.

1.2.3 Main properties and results

Thanks to the previous rewriting of the L_p-mean quantization error, one notices that this one depends on X through its distribution P_X. The first result sets that the function $x \to \|X - \widetilde{X}^x\|_p^p$ reaches a minimum, called the L_p*-optimal* N*-quantizer* and denoted γ^N. Let $\mathcal{D}_N = \{x \in (\mathbb{R}^d)^N : x_i \neq x_j \text{ iff } i \neq j\}$ and, for the sake of simplicity of notation, $D_N^p = D_N^{p,P_X}(x) = \|X - \widetilde{X}^x\|_p^p$ the pth power of the quantization error. This letter D refers to the word "distortion" used in Information Theory. In addition, we write $C := \mathrm{Conv}(S_X)$, where $S_X := \mathrm{supp}(X)$.

Proposition 1.2.4 (Pagès, 1998, Prop. 5)**.** *Assume that S_X is infinite and that P_X admits a p-moment. Then, (i) the function $x \to D_N^p(x)$ is continuous on $(\mathbb{R}^d)^N$ and reaches its minimum on C. Moreover, $\arg\min_{x\in(\mathbb{R}^d)^N} D_N^p(x) \subset \mathcal{D}_N$; (ii) the sequence $N \to \min_{C^N} D_N^p(x)$ decreases to 0.*

A proof is given in Pagès (1998). This proposition implies that the function $x \to \|X - \widetilde{X}^x\|_p^p$ reaches an absolute minimum on $(\mathbb{R}^d)^N$ at some point $\gamma^N \in (\mathbb{R}^d)^N$, and that this minimum has pairwise distinct components, i.e. $|\widetilde{X}^{\gamma^N}(\Omega)| = N$. Moreover, the quantization error tends to zero as N goes to infinity.

Recall that the L_p-mean quantization error is an error bound since one approximates the vector X by $\widetilde{X}^{\gamma^N}$. Thanks to the Proposition 1.2.4, one already knows that $\min_{C^N} \|X - \widetilde{X}^x\|_p^p$ goes to 0 as N goes to infinity. However, the question of the *rate* of convergence is a much more challenging problem. Its solution is often referred to as Zador's theorem (Zador, 1964) and provides the rate of convergence of the quantization error; see, e.g., Graf and Luschgy (2000) for a proof.

Theorem 1.2.5 (Graf and Luschgy, 2000, Th. II.6.2)**.** *Assume that $\|X\|_{p+\delta} < \infty$ for some $\delta > 0$. Let $P_X(du) = f(u)\lambda_d(du) + \nu(du)$ be the Lebesgue decomposition of P_X, where λ_d is the Lebesgue measure on $\mathbb{R}^d$ and $\nu \perp \lambda_d$. Then*

$$\lim_{N\to\infty} \left(N^{\frac{p}{d}} \min_{x\in(\mathbb{R}^d)^N} \|\widetilde{X}^x - X\|_p^p \right) = J_{p,d} \left(\int_{\mathbb{R}^d} (f(u))^{\frac{d}{p+d}}\, du \right)^{1+\frac{p}{d}},$$

with $J_{p,d} = \min_N \left(N^{p/d} \min_{x\in(\mathbb{R}^d)^N} D_N^{p,U}(x)\right)$, where $D_N^{p,U}(x)$ denotes the (pth power of the) quantization error, obtained for the uniform distribution over $[0,1]^d$, when considering the grid $x \in (\mathbb{R}^d)^N$.

In dimension $d = 1$, one has $J_{p,d} = \frac{1}{2^p(p+1)}$. For $d > 1$, little is known about $J_{p,d}$, but it can be shown that $J_{p,d} \sim \left(\frac{d}{2\pi e}\right)^{p/2}$ as $d \to \infty$ (Graf and Luschgy, 2000).

This theorem says that $\min_{x\in(\mathbb{R}^d)^N} \|X - \widetilde{X}^x\|_p = C_{X,p,d}N^{-1/d} + o(N^{-1/d})$. Consequently, one obtained that, for any distribution P_X, optimal quantization produces for every N the best-matching N-grid for P_X and that, asymptotically, a sequence of optimal quantizers yields the lowest possible constant $C_{X,p,d}$, which is of high numerical interest. The following proposition is proved in Graf and Luschgy (2000) and is a generalization of a result of Pierce (1970).

Proposition 1.2.6 (Graf and Luschgy, 2000, Cor. II.6.7). *Assume that $\|X\|_{p+\delta} < \infty$ for some $\delta > 0$. Then, for some $C, D \in \mathbb{R}$ and $N_0 \in \mathbb{N}$, we have that*

$$\|\widetilde{X}^{\gamma^N} - X\|_p^p \leq \frac{1}{N^{p/d}}\big(C\|X\|_{p+\delta}^{p+\delta} + D\big),$$

for all $N \geq N_0$.

In the sequel, one denotes by $\widetilde{X}^N$ the vector $\widetilde{X}^{\gamma^N}$ when one wants to emphasize the size of the optimal grid instead of the optimal grid itself.

At this stage, one knows that there exists some L_p-optimal N-quantizer γ^N, also mentioned as optimal N-grid, that reaches the minimum of the function $x \to \|X - \widetilde{X}^x\|_p$, but one has not obtained the uniqueness of such a minimizer. Up to now, no result of unicity for the optimal grid could be proved. The question of how to get an optimal N-grid comes then naturally in mind. The aim of the following section is to solve this optimization problem thanks to some algorithms.

1.2.4 How to determine an optimal N-grid?

Except in some very exceptional cases (such as the uniform over a compact interval of the real line), optimal N-grids have no closed form. That is, there exists no result that describes the geometric structure of such grids. However, one can attempt to obtain (approximations of) optimal N-grids through some algorithms. Let $d \geq 1$, X be a $\mathbb{R}^d$-valued random vector having an absolutely continuous distribution $\mu = P_X$. Let $N \geq 1$ be an integer and recall that $D_N^p(x)$ denotes the L_p-mean quantization error, that is, for every $x = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N$,

$$D_N^p(x) = \|X - \widetilde{X}^x\|_p = \left(\mathrm{E}\left[\min_{1\leq i\leq N} |X - x_i|^p\right]\right)^{1/p} = \left(\int_{\mathbb{R}^d} d_N^p(x,\xi)P_X(d\xi)\right)^{1/p},$$

where $d_N^p(x,\xi) := \min_{1\leq i\leq N} |x_i - \xi|^p$. The function $d_N^p(x,\xi)$ is often called *local* L_p*-distortion*. In this section, we will deal with the optimization problem

$$\begin{cases} \text{Find a } N\text{-tuple } \gamma^N = (\gamma_1^N, \ldots, \gamma_N^N) \text{ s.t.} \\ D_N^p(\gamma_1^N, \ldots, \gamma_N^N) \leq D_N^p(x_1, \ldots, x_N), \quad \forall x = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N. \end{cases}$$

The Lloyd's method I

Historically, the first attempt to solve this optimization problem was in the particular case of $p = 2$ and $d = 1$ and was called the *Lloyd's method I*. This method, also called the *fixed point*

approach, is an iterative procedure. It defines by induction the grid Γ^{s+1} of size N given a grid Γ^0 of size N as follows:

$$\Gamma^{s+1} = \mathrm{E}[X|\mathrm{Proj}_{\Gamma^s}(X)](\Omega) = (\mathrm{E}[X|X \in C_i(\Gamma^s)])_{1\leq i\leq N}, \quad s \in \mathbb{N}.$$

It can be shown that the sequence $\{\|X - \mathrm{Proj}_{\Gamma^s}(X)\|_2, s \in \mathbb{N}\}$ is non-increasing and converges toward some random vector $\widetilde{X}$ taking N values as s goes to infinity, under some appropriate assumptions. Moreover, this $\widetilde{X}$ satisfies the stationary quantizer property and is a solution of the original optimization problem $\arg\min\{\|X - Y\|_2, |Y(\Omega)| \leq N\}$.

Some problems appear if the dimension d becomes greater than 1. Indeed, the convergence may fail and, if some convergence holds, one keeps the stationarity property of the limit $\widetilde{X}$ but it is not guaranteed that it minimizes the quadratic quantization error. This procedure has actually two main drawbacks. The first one is that it is a local procedure since it does not explore the whole space. The second one, and not the least, is that it becomes numerically intractable since it requires the computation of d-dimensional integrals. However, when X is simulatable, one can randomize the Lloyd's method by using a Monte Carlo simulation to compute the above integrals. This version is sometimes used as a final step of the optimization procedure to refine locally the results obtained by other methods. More details on this method can be found in Pagès and Printems (2003).

The stochastic gradient algorithm

This procedure is based on the smoothness of the L_p-quantization error function. In this procedure, one uses an important property of this error function. Indeed, $D_N^p(x)$ is smoother than Lipschitz continuous and is continuously differentiable at every $x \in (\mathbb{R}^d)^N$ satisfying some admissibility conditions, for $p > 1$ (see Pagès (1998) for a proof).

Proposition 1.2.7 (Pagès, 1998, Prop. 9)**.** *If $p > 1$, D_N^p is continuously differentiable at every $x = (x_1, \ldots, x_N) \in \mathcal{D}_N$ such that $P_X(\bigcup_{i=1}^N \partial C_i(x)) = 0$. Furthermore,*

$$\nabla D_N^p(x) = p\left(\int_{C_i(x)} |x_i - \xi|^{p-1} \frac{x_i - \xi}{|x_i - \xi|} P_X(d\xi)\right)_{i=1,\ldots,N},$$

with the convention $\frac{0}{|0|} = 0$. This result holds for $p = 1$ with the additional condition P_X is continuous.

Thanks to this result, one obtained the gradient of D_N^p and one observes that ∇D_N^p has an integral representation with respect to P_X. If the distribution P_X is simulatable, this suggests to implement a stochastic gradient descent derived from this representation to approximate some local minimum of D_N^p (more details on stochastic gradient descent in this context can be found in Pagès and Printems (2003)). However, this would require the computation of many multidimensional integrals and then becomes unrealistic as soon as $d \geq 2$. For this reason, the stochastic gradient algorithm is defined as the corresponding stochastic procedure, obtained by replacing $\nabla D_N^p(x)$ by its empirical version in the gradient descent approach.

Let $(X^t)_{t\in\mathbb{N}^*}$, $\mathbb{N}^* = \mathbb{N}\setminus\{0\}$, be a sequence of i.i.d. P_X-distributed random variables and let $(\delta_t)_{t\in\mathbb{N}^*}$ be a deterministic (0,1)-valued sequence satisfying

$$\sum_{t=1}^{\infty}\delta_t = +\infty \quad \text{and} \quad \sum_{t=1}^{\infty}\delta_t^2 < +\infty.$$

These X^t, $t\in\mathbb{N}^*$, are used as stimuli in this algorithm. One starts with a deterministic initial N-tuple $\hat{\gamma}^{N,0} = x^0$ with N pairwise distinct components, since it is the admissibility condition for D_N^p to be continuously differentiable at $x\in(\mathbb{R}^d)^N$. In other words, x^0 is our initial N-grid that one wants to transform in an optimal way. One then defines recursively for every $t\geq 0$

$$\hat{\gamma}^{N,t+1} = \hat{\gamma}^{N,t} - \frac{\delta_{t+1}}{p}\nabla_x d_N^p(\hat{\gamma}^{N,t}, X^{t+1}), \tag{1.2.1}$$

where $\nabla_x d_N^p(x,\xi)$ stands for the gradient with respect to the x-argument of the local quantization error, with $x=(x_1,\dots,x_N)\in(\mathbb{R}^d)^N$ and $\xi\in\mathbb{R}^d$. Note that, for any ξ, the ith entry of this gradient is

$$\big(\nabla_x d_N^p(x,\xi)\big)_i = p|x_i-\xi|^{p-1}\frac{x_i-\xi}{|x_i-\xi|}\mathbb{I}_{C(x_i)}(\xi).$$

Therefore, this gradient is a vector with exactly one non-zero component, the one corresponding to the point of the grid x that is the closest to ξ. This implies that, at each step of the algorithm, only one point of the grid $\hat{\gamma}^{N,t}$ is updated when defining the grid $\hat{\gamma}^{N,t+1}$. This point is the point from $\hat{\gamma}^{N,t}$ that is the closest to X^{t+1}. Moreover, it is really important to note that this recursive formula provides at every step a new grid with pairwise distinct points. This is crucial since one wants to take advantage of the differentiability of D_N^p.

The problem of this procedure is that the assumptions usually made that ensure the a.s. convergence of such a procedure are not fulfilled by D_N^p. It comes from the fact that $D_N^p(x_1,\dots,x_N)$ does not go to infinity as $\max_{1\leq i\leq N}|x_i|$ goes to infinity and ∇D_N^p is not a Lipschitz function. However, some weaker a.s. convergence results have been obtained under some assumptions on P_X in the case $p=2$, as we will detail in Section 1.2.5. The quadratic case $p=2$ is then the most commonly implemented for applications, since one also checks on simulation that it does behave better, and it is known as the *Competitive Learning Vector Quantization* (CLVQ) algorithm.

Let us now explain in more details the recursive formula (1.2.1). Set $\hat{\gamma}^{N,t} := (\hat{\gamma}_1^{N,t},\dots,\hat{\gamma}_N^{N,t})\in(\mathbb{R}^d)^N$ the grid at step t.

Step 1 (Competitive Phase) Select the index $i(t+1)\in\arg\min_i|\hat{\gamma}_i^{N,t}-X^{t+1}|$, i.e. the index corresponding to the closest neighbor of X^{t+1} among the points of the grid determined by $\hat{\gamma}^{N,t}$, called the "winning index". This index depends both on the grid $\hat{\gamma}^{N,t}$ and on X^{t+1}.

Step 2 (Learning Phase) Update the grid $\hat{\gamma}^{N,t}$ as follows:

$$\begin{cases} \hat{\gamma}_{i(t+1)}^{N,t+1} &:= \hat{\gamma}_{i(t+1)}^{N,t} - \delta_{t+1}\frac{\hat{\gamma}_{i(t+1)}^{N,t}-X^{t+1}}{|\hat{\gamma}_{i(t+1)}^{N,t}-X^{t+1}|}|\hat{\gamma}_{i(t+1)}^{N,t}-X^{t+1}|^{p-1} \\ \hat{\gamma}_i^{N,t+1} &:= \hat{\gamma}_i^{N,t} \qquad \text{if } i\neq i(t+1) \end{cases}.$$

The fact that only one point moves at each step appears here clearly.

Step 3 (Companion Procedure) One can define some "companion" procedures that are costless since they use some by-products of the two previous phases, and they yield parameters of numerical interest. Assume that $X \in \mathrm{L}_{p(1+\eta)}$ for some $\eta \in (0,1]$ and let $(\widetilde{\delta}_t)_{t\geq 1}$ be a deterministic (0,1)-valued sequence satisfying

$$\sum_{t=1}^{\infty} \widetilde{\delta}_t = +\infty \quad \text{and} \quad \sum_{t=1}^{\infty} \widetilde{\delta}_t^{1+\eta} < +\infty.$$

Then, one defines recursively the following sequences

$$\begin{array}{rlcl} & p_i^0 & := & 0, \\ \forall t \geq 0, & p_i^{t+1} & := & p_i^t(1-\widetilde{\delta}_{t+1}) + \widetilde{\delta}_{t+1}\mathbb{I}_{i=i(t+1)}, \\ & D_N^{r,0} & := & 0, \\ \forall t \geq 0, & D_N^{r,t+1} & := & D_N^{r,t}(1-\widetilde{\delta}_{t+1}) + \widetilde{\delta}_{t+1}|\hat{\gamma}_{i(t+1)}^{N,t} - X^{t+1}|^r, \end{array}$$

for all $i = 1, \ldots, N$ and with $r \in [1, p]$. One can prove that, under the event $\{\hat{\gamma}^{N,t} \xrightarrow[t\to\infty]{\text{a.s.}} \gamma^N\}$, one has

$$\begin{array}{lll} \forall i \in \{1,\ldots,N\}, & p_i^t \xrightarrow{\text{a.s.}} P_X\big(C_i(\gamma^N)\big), & \text{as } t \to \infty, \\ \forall r \in [1,p], & D_N^{r,t} \xrightarrow{\text{a.s.}} D_N^r(\gamma^N), & \text{as } t \to \infty. \end{array}$$

In the particular case where $\widetilde{\delta}_t = 1/t$, one can rewrite the recursive formula for p_i^t and $D_N^{r,t}$ as simple synthetic expressions, namely

$$p_i^t = \frac{1}{t}\big|\{s \in \{1,\ldots,t\} | X^s \in C_i(\hat{\gamma}^{N,s-1})\}\big| \quad \text{and} \quad D_N^{r,t} = \frac{1}{t}\sum_{s=1}^{t} |\hat{\gamma}_{i(s)}^{N,s-1} - X^s|^r. \quad (1.2.2)$$

These companion procedures yield (approximations of) the P_X-weight of the Voronoi cells $C_i(\gamma^N)$ and the L_p-mean quantization error $\|X - \widetilde{X}^{\gamma^N}\|_p$ needed for a numerical use of the quantizer γ^N. An important fact is that they work under the assumption $\{\hat{\gamma}^{N,t} \xrightarrow{\text{a.s.}} \gamma^N\}$ whatever the limiting N-tuple γ^N is, i.e. whatever the grid provided by the algorithm. This fact shows their consistency.

Concerning the practical implementation of this algorithm, it has to be noticed that, in the particular quadratic case $p = 2$, i.e. the CLVQ algorithm, the N-tuple $\hat{\gamma}^{N,t+1}$ remains at each step in the convex hull of $\hat{\gamma}^{N,t}$ and X^{t+1}. This induces a stabilizing effect on the procedure observed in simulations which explains why the CLVQ algorithm is more often implemented than its non-quadratic counterparts.

One now turns to the discussion about a problem that arises in practice: the choice of the initialization grid and of the step. Let start by the initialization of the N-quantizer. In the case of a Normal distribution, Pagès and Printems (2003) suggests to consider two methods according to the value of N. Indeed, if $N \leq 10$, i.e. if N is small, it proposes to adopt a *random initialization*, so that $\hat{\gamma}^{N,0} \sim (\mathcal{N}(0, I_d))^{\otimes N}$. However, when N gets larger, the so-called *splitting initializing method* should be preferred. This method consists in adding one further point to the optimal N-quantizer in order to obtain a starting quantizer of this procedure with $N+1$ components. This further point is usually the optimal 1-quantizer, i.e. $0_{\mathbb{R}^d}$ for the Normal distribution (the

expectation of the law in a general setting). The problem of the splitting method is that it could provide only sub-optimal stationary quantizers. However, it turns out to be a good compromise between stability and efficiency. In this work, one will use quantization for estimating conditional quantile. Therefore, this method cannot be used since we only have at our disposal a sample and we do not know the underlying distribution. The initialization in our setting will be detailed in the corresponding chapter.

One now considers the choice of the step parameter. A common choice is to take $\delta_t > \frac{2N^{3/d}}{\pi^2 t}$ which is the critical step for the uniform distribution to get a Central Limit Theorem for large enough t. Consequently, the parameter sequence $(\delta_t)_{t\geq 0}$ will be set equal to $\delta_t = \delta_0 \frac{a}{a+\delta_0 bt}$, with $a = 4N^{1/d}$ and $b = \pi^2 N^{-2/d}$. It remains to choose the initial step δ_0. It is chosen either equal to the square root of the distortion computed at last step, or to 1 if this distortion is greater that 1. This choice is motivated by the fact that $\min_{i\neq j} |x_i - x_j| \leq 2(\mathrm{E}|X - x_i|^2)^{1/2}$.

One applied this algorithm to the bivariate standard Normal distribution. One chose the step parameter as explained above and $N = 500$. This algorithm is performed with 10^6 iterations. The obtained optimal grid is represented in Figure 1.6. One observes that the grid is more dense around the center of the distribution, $0_{\mathbb{R}^2}$, and that the grid is symmetric, which is not surprising since the Normal distribution is itself symmetric.

Finally, let us illustrate[2] the efficiency of this algorithm to provide an optimal grid on two examples (one with $d = 1$ and the other one with $d = 2$). In Figure 1.7, we used this algorithm to construct an optimal quantization grid of size $N = 15$ for the uniform law $U(-2, 2)$. The left panel is composed of three figures. The top one represents the initial grid $\hat{\gamma}^{N,0}$ in red and the optimal grid $\hat{\gamma}^{N,500}$ in green, provided by the algorithm after 500 replications (the grey points are the stimuli X^t). The middle and the bottom figures plot the empirical cumulative distribution functions of the projections of the stimuli X^t onto the initial grid and the optimal grid respectively. The right panel is obtained analogously, but with $n = 5,000$. The initial grids are of course quite far from the expected optimal ones, but the grid provided by the algorithm are really close to the optimal grid corresponding to the uniform law (i.e. an equispaced grid on the support). Of course, the larger n is, the more the green grid is close to the optimal one. Therefore, the right green grid is more desirable than the left one.

The second example is similar for $d = 2$, except that we do not represent the empirical cumulative distribution functions. The law of X is here a uniform law on the square $(-2, 2)^2$. The top figures of Figure 1.8 plot initial grids in red, and the bottom ones optimal grids in green, provided by the algorithm. The left ones are obtained after 2,000 replications of the algorithm, and the right ones 20,000.

[2]Figures 1.7 and 1.8 were obtained thanks to the `choice.grid` function of the R package **QuantifQuantile**, see Section 3.5 for more details.

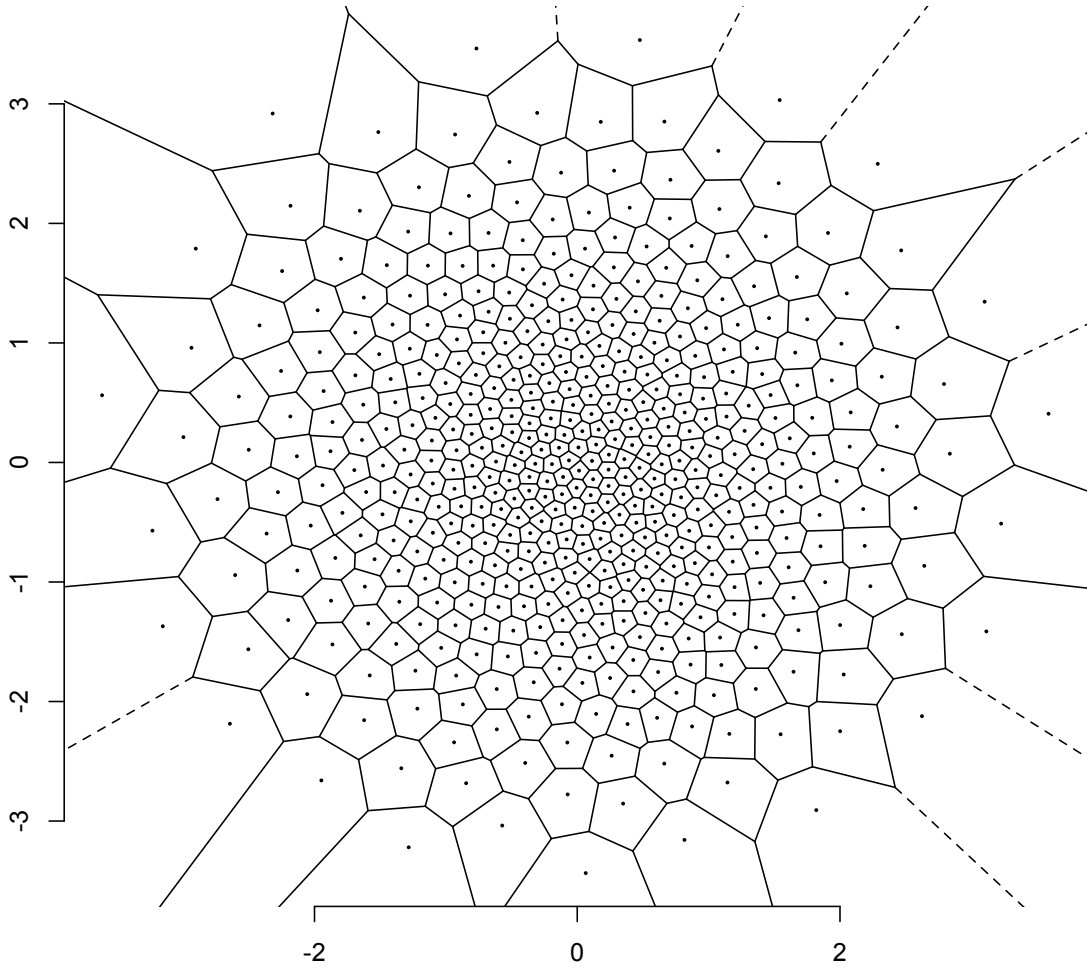


Figure 1.6 – Optimal quantization grid of size $N = 500$ for the bivariate standard Normal distribution (with Voronoi tessellation).

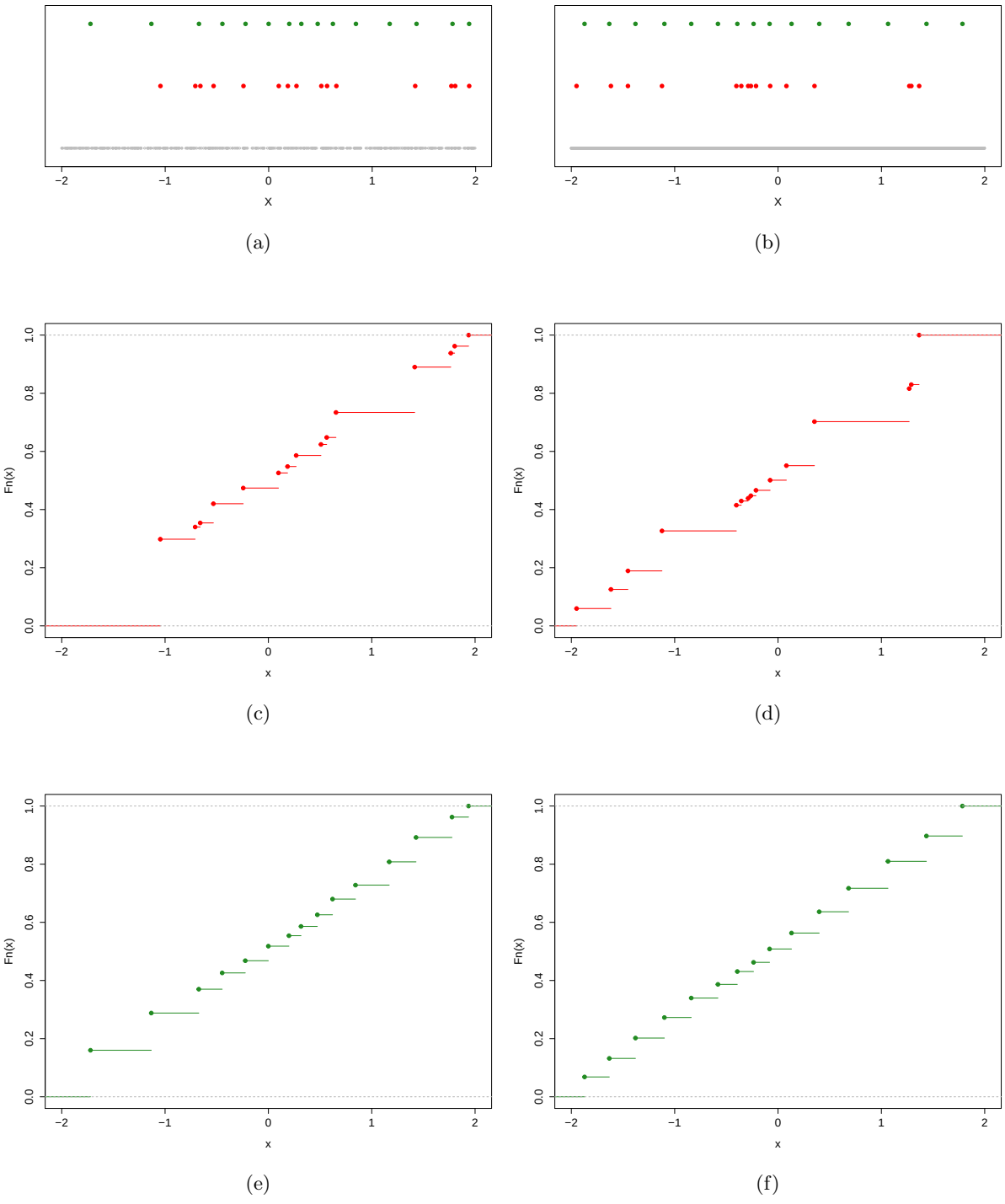


Figure 1.7 – The left panel corresponds to a number of stimuli $n = 500$ while the right one to $n = 5,000$. The top figures represent the stimuli X^t (grey), the initial grid (red) and the optimal grid (green). The middle and bottom figures plot the ecdf of the projections of the observations onto the initial grid and the optimal grid respectively.

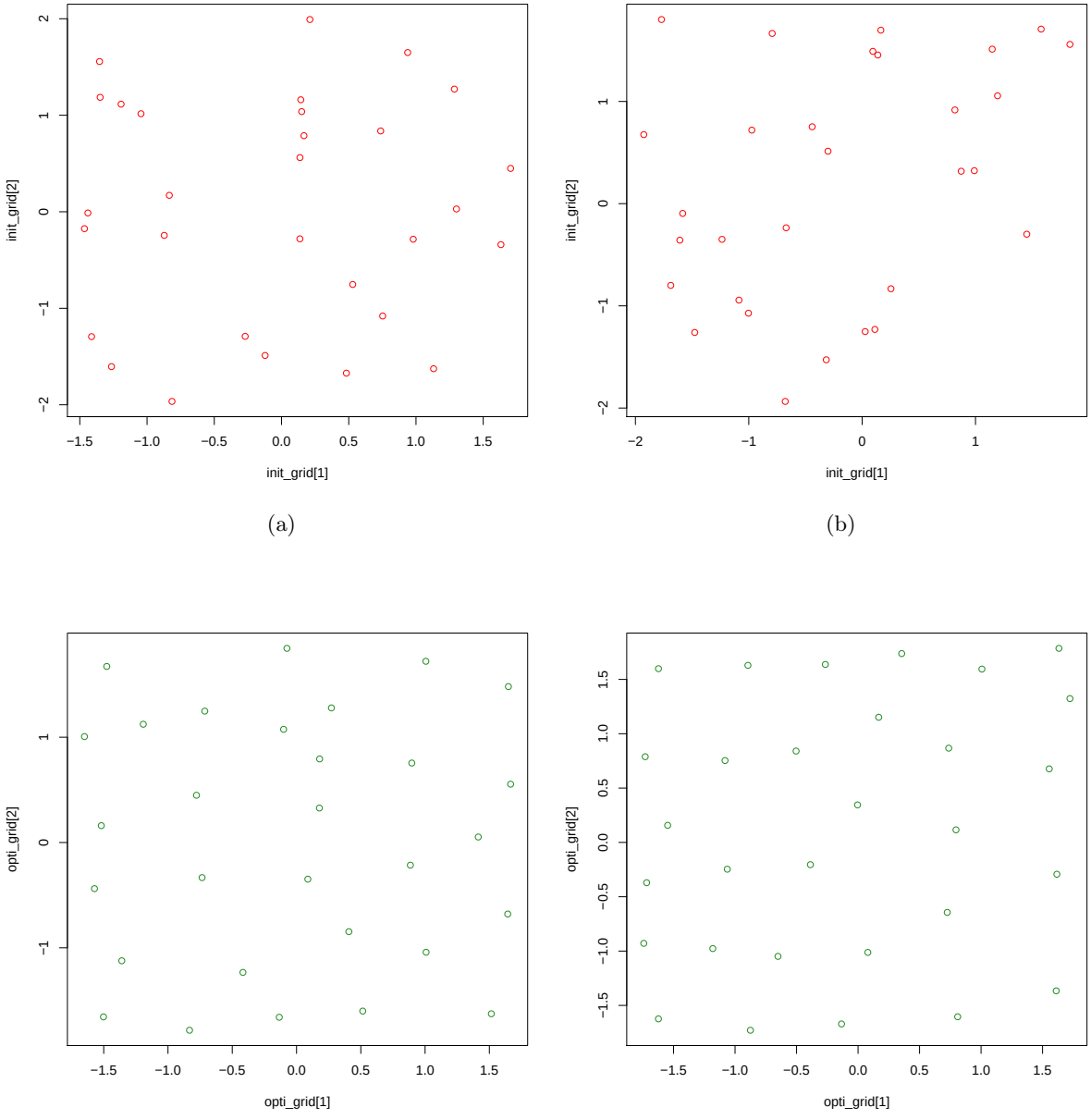


Figure 1.8 – For n stimuli uniformly distributed on $(-2, 2)^2$, with $n = 2,000$ (left) and $n = 20,000$ (right), the initial and optimal grids (top and bottom respectively).

1.2.5 Convergence of the CLVQ algorithm

Here we state several results showing that the grids provided by the CLVQ algorithm converge to optimal grids as the number of iterations t goes to infinity.

We start with the univariate case ($d = 1$). Assume that the support of P_X is compact and let its convex hull C be $[a, b]$. Write $F_N^+ := \{x = (x_1, \ldots, x_N) : a < x_1 < \cdots < x_N < b\}$ for the set of N-grids on C involving pairwise distinct points stored in ascending order, and let $\bar{F}_N^+$ be its closure; see Pagès (1998).

Theorem 1.2.8 (Pagès, 1998, Th. 27). *In the univariate setup above, we have the following.*

(i) Assume that P_X is absolutely continuous with a density $f : [a, b] \to \mathbb{R}^+$ that is positive on (a, b), and assume either that f is strictly log-concave or that it is log-concave with $f(a+) - f(b-) > 0$. Then $x \mapsto D_N^{2,P_X}(x)$ has a unique minimizer γ^N in $\bar{F}_N^+$, which coincides with the unique solution of $\nabla D_N^{2,P_X}(x) = 0$ in $\bar{F}_N^+$ (when P_X is the uniform over $[0, 1]$, the optimal grid is $\gamma^N = \big(a + \frac{2k-1}{2N}(b-a)\big)_{1\le k\le N}$).

(ii) Irrespective of the initial grid $\hat{\gamma}^{N,0} \in F_N^+$, every trajectory $(\hat{\gamma}^{N,0}, \hat{\gamma}^{N,1}, \hat{\gamma}^{N,2}, \ldots)$ of the CLVQ algorithm is a.s. such that $\hat{\gamma}^{N,t} \in F_N^+$ for all t. If P_X is absolutely continuous and if there are finitely many grids x $(\in \bar{F}_N^+)$ such that $\nabla D_N^{2,P_X}(x) = 0$, then $\hat{\gamma}^{N,t} \xrightarrow{a.s.} \gamma^N$ as $t \to \infty$, with $\nabla D_N^{2,P_X}(\gamma^N) = 0$.

Part (i) of the result provides a particular family of distributions for which the optimal grid is unique (recall that existence always holds). Beyond stating that trajectories of the CLVQ algorithm live in F_N^+ (with grids that therefore stay of size N), Part (ii) of the result provides mild conditions under which the algorithm almost surely provides a limiting grid that is a critical point of the quantization error, hence, under the assumptions of Part (i), is optimal.

Unfortunately, the picture is less clear for the multivariate case ($d > 1$). While it is still so that the grid $\hat{\gamma}^{N,t}$ will have pairwise distinct components for any t, some of the components of the limiting grid γ^N, if any, may coincide.

(a) If, parallel to the univariate case, this does not happen, then the a.s. convergence of $\hat{\gamma}^{N,t}$ to a critical point of the quantization error $D_N^{2,P_X}(\cdot)$ can be established under the assumption that P_X has a bounded density with a compact and convex support.

(b) Otherwise, no convergence results are available; the only optimality results that can then be obtained relate to approximations involving grids of size $k < N$, where k is the number of distinct components in the limiting grid γ^N, which is quite different from the original N-quantization problem considered initially. Indeed, one cannot prove that the points of the grid stay at strictly positive distance from each other when the number of iterations tends to infinity. Therefore, it could appear that the limiting grid has strictly less than N points.

These features are detailed in Pagès (1998). For practical purposes, though, one should not worry too much, as all numerical exercices we conducted were compatible with case (a) (with

	$d = 1$ and $N = 15$			$d = 2$ and $N = 30$		
	$n = 100$	$n = 1,000$	$n = 10,000$	$n = 200$	$n = 5,000$	$n = 20,000$
CPU	0.012	0.116	1.145	0.556	0.883	3.468
AVE$[e]$	0.132	0.0993	0.0843	0.364	0.322	0.315
SD$[e]$	0.0196	0.0115	0.00657	0.0315	0.0130	0.00835

Table 1.1 – Averaged computation times (CPU, in seconds) dedicated to the CLVQ algorithm with grid size N, along with the averages (AVE$[e]$) and standard deviations (SD$[e]$) of the sample quantization errors in (1.2.3), obtained from 50 independent samples of size n; see Section 1.2.6 for details.

increasing t, the smallest distance between two components of $\hat{\gamma}^{N,t}$ always seemed to stabilize rather than decrease to zero).

In the sequel, we will write throughout that the grids provided by the CLVQ algorithm are optimal. Strictly speaking, though, the grid $\hat{\gamma}^{N,n}$ resulting from this algorithm, for any fixed n, does not minimize the quantization error, hence is not optimal. Yet, the previous results show that $\hat{\gamma}^{N,n}$ converges to an optimal grid as the number of iterations n tends to infinity (with N fixed), under mild assumptions. Clearly, only moderate-to-large values of n are expected to provide good approximations of genuine optimal N-grids.

1.2.6 Numerical performances of CLVQ

We end this section by quickly evaluating the performances of the CLVQ algorithm, both for $d = 1$ and $d = 2$. To do so, we generated, for the various combinations of n and N considered in Table 1.1, 50 mutually independent random samples of size n from the uniform distribution over $[-2, 2]^d$ and performed in each case the CLVQ algorithm with the corresponding value of N. For each (n, N), this leads to 50 CPU times to completion and 50 sample quantization errors

$$e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathrm{Proj}_{\hat{\gamma}^{N,n}}(X_i)\right)^2}. \tag{1.2.3}$$

Table 1.1 reports the corresponding averaged CPU times, along with the averages and standard deviations of the various collections of 50 quantization errors. Since the number of iterations of the algorithm is equal to n, it is no surprise that the error decreases with n while the computation time increases with n.

1.3 Objectives and structure of the thesis

This thesis consists in two parts, each of them composed of several chapters. We motivated in Section 1.1 the importance of conditional quantiles and gave an overview of the numerous applications of this concept. Efficient estimators of conditional quantiles are then highly desirable and the goal of the first part of this thesis consists in constructing new estimators of conditional quantiles, using L_p-mean optimal quantization.

Chapter 2 aims to define this quantization-based estimator. The starting point consists in replacing X in the definition of conditional quantiles by its quantized version. The resulting approximation of conditional quantiles is studied from a theoretical point of view in Section 2.2 and convergence results are derived when N tends to infinity. From this approximation, an estimator of conditional quantiles is defined in Section 2.3. The convergence in probability of this estimator toward the approximation is obtained when n tends to infinity and for N fixed. However, even if these asymptotic results are really interesting, we observe in practice that the resulting conditional quantile curves are not sufficiently smooth for small-to-moderate sample sizes n. For this reason, we define a bootstrap version of our estimator that considerably improves its performances. Sections 2.4 and 2.5 gather the proofs of Sections 2.2 and 2.3 respectively.

This theoretical investigation naturally requires to address the numerical behavior of our estimator. This is the object of Chapter 3. We first develop in Section 3.2 a data-driven selection method for the parameter N (the size of the quantization grid) required in the construction of our estimator. This method shows its efficiency on several replications. With this selection method, our estimation procedure is entirely functional and we are then able to compare our performances with the ones of alternatives estimators in Section 3.3. The comparison study of this section only deals with univariate covariates. Therefore, Section 3.4 briefly compares our estimator with its main competitor in the case of a bidimensional covariate. We also developed an R package, called **QuantifQuantile**, allowing anyone to perform our method in a quite automatic way. Section 3.5 focuses on the different functions of this package and illustrates them. The numerical investigation is completed in Section 3.6 with real data applications.

These first chapters tackle the regression setup of a univariate response variable and a uni- or multivariate covariate. It is then natural to wonder whether it is possible to extend the results of the first part to a multivariate response variable. This requires to generalize the concept of quantile in a multidimensional setting, which is not direct due to the lack of natural ordering in the Euclidean space for dimension greater than one. This is then the subject of Part II of this thesis where we first discuss in Chapter 4 the choice of a concept of multivariate quantiles and compare the different existing notions.

Chapter 5 adresses the theoretical investigation of some quantization-based estimator of multiple-output conditional quantiles and is the natural generalization of Chapter 2. It follows the same structure since 1) an approximation is defined and studied in Section 5.2, and 2) a new estimator is introduced in Section 5.3 by taking an empirical version of this approximation. A bootstrap version of the estimator is also defined. Sections 5.4 and 5.5 gather the proofs of Sections 5.2 and 5.3 respectively.

Finally, Chapter 6 focuses on the numerical behavior of our estimator. It generalizes the first sections of Chapter 3. Indeed, Section 6.2 extends the data-driven selection method for the parameter N (the size of the quantization grid) while Section 6.3 compares our performances with the ones of alternative estimators. The numerical investigation is completed in Section 6.4 with a real data application.

We then conclude this thesis in Chapter 7 and we point out some perspectives.


FUNCTION
CONDITIONAL
quantization
OPTIMAL
QUANTILES
VECTOR
regression
DEFINED
used
obtained
PROBLEM
DEFINITION
VALUES
considers
called
error
distribution
ESTIMATOR
Voronoi
PROCEDURE
FIXED
GENERALIZATION
EXISTS
FRAMEWORK
MEAN
NONPARAMETRIC
ABSOLUTE
GRADIENT
APPROXIMATION
REFERENCE
SECTION
RANDOM
variable
CONTINUOUS
CORRESPONDING
DEPENDENT
MEDIAN
NOTION
INTEGRALS
FIGURE
PROPERTY
THANKS
CASE
LINE
REWRITTING
CONSISTS
best
reaches
DEVELOPED
LLOYD
humerical
CLUSTERED
COVARIATES
MINIMIZING
ORIGIN
DIFFERENT
GIVEN
LINEAR
IMPORTANT
INVERTIBLE
TAKING
PERFORMED
introduced
representative
LITERATURE
EXTENSIVELY
TERM
MAIN
FINITE
GOES
DESCENT
EXPLORE
neighbor
signal
CURVES
CONCEPT
WORD
CHARGED
CHARTS
boxplat
STANDARDS
QUANTILESTHE
QUANTIZATIONTHE
derived
rewritten
PROPOSITION
THEOREM
appearing
questions
ILLUSTRATED
reaches
HEALTH
INFORMATION
ORGANIZATION
PARTITION
MULTIDIMENSIONAL
HISTORICALLY
ADMISSIBILITY
computation
output
CLOSEST
LOCAL
growth
CONSEQUENTLY
ordinary
HORIZONTAL
examples
GEOMETRIC
SIMULATIONS
context
REQUIRES
boundary
EXCEPTIONAL
KNOWS
approach
HETEROGENEOUS
INSTEAD
RECENTLY
characterizing
COMPENSATION
HYPERPLANES
PURPOSES
DETAILS
choice
infinity
ANALYSIS
STATISTICS
CONVEX
PROVED
EXPLAIN
alternatives
DISCRETIZATION
ASYMPTOTICALLY
LOOKING
MENTIONED
TRANSMISSION
theoretical
CLASSICAL
sample
FIRM
PROJECTION
weight
REAL
PROBABILITY
COMES
LIMITS
ATTEMPT
theory
HYPERSURFACES
REPRESENTATION
RESPECTIVELY
APPLICATIONS
humber
PRACTICE
ANALOGOUSLY
market
sets
incorporated
ACTUALLY
distortion
harm
stationarity
RECALL
safeguarding
REPLACING
LINE
kernel
simulatable
DISCRETISING
INDUCED
DIFFERENTIABLE
expectation
INTERESTING
notice
CONTINUOUS
SEQUENCE
groups
CHARGED
EXTENSIVELY
FINITE
EXPLORE
DESCENT
CORRESPONDING
discretising
INDUCED

Glossary

Notation	Meaning
q_α	Quantile of order α of the random variable Y
$q_\alpha(x)$	Conditional quantile of order α of the random variable Y given $X = x$
$\mathbb{I}_A$	Indicator function of the set A
$\lvert z\rvert$, $\lVert z\rVert_p$	Euclidean norm, L_p-norm of z
P_X, S_X	Distribution of the random vector X and its support
γ^N	Optimal quantization grid of size N
$\widetilde{X}^N$	Projection of X onto the quantization grid γ^N
C_x	Quantization cell containing all points projected on the same grid point as x
$R(C_x)$	Radius of C_x, i.e. largest distance between the center and a point of the cell
$\mathcal{B}(z, r)$	Ball of center z and radius r
$\hat{\gamma}^{N,n}$	Approximation of γ^N provided by CLVQ after n replications
$\widehat{X}^{N,n}$	Projection of X onto the approximated quantization grid $\hat{\gamma}^{N,n}$
$\widetilde{q}_\alpha^N(x)$	Approximation of $q_\alpha(x)$ constructed using $\widetilde{X}^N$
$\widehat{q}_\alpha^{N,n}(x)$	Estimator of $q_\alpha(x)$ constructed from a sample of size n, using $\hat{\gamma}^{N,n}$
$\bar{q}_{\alpha,B}^{N,n}(x)$	Bootstrap-based estimator of $q_\alpha(x)$
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution of mean μ and variance σ^2
$U(a, b)$	Uniform distribution on the interval (a, b)
$\mathrm{Beta}(a, b)$	Beta distribution with shape parameters a and b
χ^2_k	Chi-square distribution with k degrees of freedom
$\mathrm{ISE}^{\hat{}}_\alpha(N)$	Infeasible criterion for selecting N based on $\widehat{q}_\alpha^{N,n}(x)$
$\mathrm{ISE}^{\bar{}}_{\alpha,B}(N)$	Infeasible criterion for selecting N based on $\bar{q}_{\alpha,B}^{N,n}(x)$
$\widehat{\mathrm{ISE}}^{\hat{}}_\alpha(N)$	Data-driven criterion for selecting N based on $\widehat{q}_\alpha^{N,n}(x)$
$\widehat{\mathrm{ISE}}^{\bar{}}_{\alpha,B}(N)$	Data-driven criterion for selecting N based on $\bar{q}_{\alpha,B}^{N,n}(x)$

$\widehat{q}_{\alpha}^{\mathrm{sp}}(x)$	Spline method-based estimator of $q_{\alpha}(x)$
$\widehat{q}_{\alpha;k\mathrm{NN}}(x)$	k-nearest neighbor estimator of $q_{\alpha}(x)$
$\widehat{q}_{\alpha}^{\mathrm{YJ}}(x)$	Local linear estimator of $q_{\alpha}(x)$
$\widehat{q}_{\alpha}^{\mathrm{YJc}}(x)$	Local constant estimator of $q_{\alpha}(x)$

2

Conditional quantile estimation through optimal quantization

Contents

2.1 Introduction

In this chapter, we define a new estimator of conditional quantile. As above mentioned, this estimator is constructed using optimal quantization. In Section 2.2, we explain the first step of the construction. It consists in replacing the covariate X in Definition 1.1.4 of conditional quantile to get an approximation of them. An estimator is then defined in a second step by taking an empirical version of this approximation, as detailed in Section 2.3. This estimator and the asymptotic results given below are the subject of the paper *Conditional quantile estimation through optimal quantization* (Charlier et al., 2015a) with Davy Paindaveine and Jérôme Saracco, published in *Journal of Statistical Planning and Inference*.

Throughout the chapter, X is a d-dimensional random vector of covariates while Y is a scalar response variable, unless otherwise stated.

2.2 Approximation of conditional quantile

Fix $p \geq 1$ such that $\|X\|_p < \infty$. The idea of this approximation is to replace X by a discrete random vector in the definition of conditional quantile. This discrete random vector is its quantized

version $\widetilde{X}^N$. As explained in Section 1.2, $\widetilde{X}^N$ is the projection of X onto an L_p-norm optimal quantization N-grid γ^N. Therefore, $\widetilde{X}^N$ can only take N values (the N points of γ^N) and the x for which we consider $q_\alpha(x)$ has also to be replaced by its projection on the grid in the conditioning of the expectation since $P(\widetilde{X}^N = x) = 0$ for all $x \in \mathbb{R}^d \setminus \gamma^N$. Then, for any positive integer N, one may consider the approximation

$$\widetilde{q}_\alpha^N(x) = \arg\min_{a\in\mathbb{R}} \mathrm{E}\big[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}\big], \tag{2.2.1}$$

where $\tilde{x}$ is the projection of x onto γ^N. Since $\widetilde{X}^N - X$ goes to zero as $N \to \infty$ by Proposition 1.2.4, one may expect that $\widetilde{q}_\alpha^N(x)$ provides a better and better approximation of $q_\alpha(x)$ as N increases. The main goal of this section is to quantify the quality of this approximation.

Let us denote $\mathbb{R}_0^+ := \{z \in \mathbb{R} : z > 0\}$. We will need several assumptions and we will motivate each of them.

Assumption (A) (i) The random vector (X, Y) is generated through $Y = m(X, \varepsilon)$, where the d-dimensional covariate vector X and the error ε are mutually independent; (ii) the link function $(x, z) \mapsto m(x, z)$ is of the form $m_1(x) + m_2(x)z$, where the functions $m_1(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and $m_2(\cdot) : \mathbb{R}^d \to \mathbb{R}_0^+$ are Lipschitz functions; (iii) $\|X\|_p < \infty$ and $\|\varepsilon\|_p < \infty$; (iv) the distribution of X does not charge any hyperplane.

Point (i) of this assumption simply sets that Y depends on the covariate X and also on an error term independent of X. Point (ii) characterizes the link function m. The assumptions on m are quite mild and we observe that choosing non-constant $m_2(\cdot)$ provides heteroscedastic models. They are particularly of interest since detecting heteroscedasticity is a major advantage of quantile regression with respect to standard regression. Combining Point (ii) with second part of Point (iii) implies that the link function $m(\cdot, \varepsilon)$ of the model is a Lipschitz function, i.e. that there exists $C > 0$ such that $m(\cdot, \varepsilon)$ satisfies

$$\forall u, v \in \mathbb{R}^d,\ \ \|m(u, \varepsilon) - m(v, \varepsilon)\|_p \le C|u - v|. \tag{2.2.2}$$

The resulting Lipschitz constant — that is, the smallest real number C for which (2.2.2) holds — is $[m]_{\mathrm{Lip}} \le [m_1]_{\mathrm{Lip}} + [m_2]_{\mathrm{Lip}}\|\varepsilon\|_p$, where $[m_1]_{\mathrm{Lip}}$ and $[m_2]_{\mathrm{Lip}}$ are the corresponding Lipschitz constants of m_1 and m_2, respectively. Indeed, for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned}
\big\|m(u,\varepsilon) - m(v,\varepsilon)\big\|_p &= \big\|\big(m_1(u) - m_1(v)\big) + \big(m_2(u) - m_2(v)\big)\varepsilon\big\|_p \\
&\le \big\|m_1(u) - m_1(v)\big\|_p + \big\|\big(m_2(u) - m_2(v)\big)\varepsilon\big\|_p \\
&\le \big(\mathrm{E}\big[|m_1(u) - m_1(v)|^p\big]\big)^{1/p} + \big(\mathrm{E}\big[|m_2(u) - m_2(v)|^p|\varepsilon|^p\big]\big)^{1/p} \\
&\le [m_1]_{\mathrm{Lip}}|u - v| + \big(\mathrm{E}\big[[m_2]_{\mathrm{Lip}}^p|u - v|^p|\varepsilon|^p\big]\big)^{1/p} \\
&\le [m_1]_{\mathrm{Lip}}|u - v| + [m_2]_{\mathrm{Lip}}|u - v|\big(\mathrm{E}[|\varepsilon|^p]\big)^{1/p} \\
&= \big([m_1]_{\mathrm{Lip}} + [m_2]_{\mathrm{Lip}}\|\varepsilon\|_p\big)|u - v|.
\end{aligned}$$

Finally, first part of Point (iii) is mandatory to perform quantization and Point (iv) is necessary to ensure the existence of an L_p-optimal quantization N-grid γ^N, as explained in Section 1.2.

The following reinforcement of Assumption (A) is necessary if we want to obtain convergence rates. This assumption actually comes from Theorem 1.2.5 allowing to get convergence rate for the quantization error.

ASSUMPTION (A′) Same as Assumption (A), but with (iii) replaced by (iii)′ there exists $\delta > 0$ such that $\|X\|_{p+\delta} < \infty$, and $\|\varepsilon\|_p < \infty$.

The last assumption of this section is a technical assumption used to prove both theorems.

ASSUMPTION (B) (i) The support S_X of P_X is compact; (ii) ε admits a continuous density $f^{\varepsilon} : \mathbb{R} \to \mathbb{R}_0^+$ (with respect to the Lebesgue measure on $\mathbb{R}$).

We can then prove the following result (see Section 2.4 for the proof). It gives an upper bound for the L_p-norm difference between $q_\alpha(X)$ and its approximation. In this bound appears the quantization error that we control as N goes to infinity thanks to Proposition 1.2.6.

Theorem 2.2.1. *Fix $\alpha \in (0,1)$. Then (i) under Assumptions (A)-(B),*

$$\|\widetilde{q}_\alpha^N(X) - q_\alpha(X)\|_p \leq 2 \max\left(\frac{\sqrt{\alpha}}{1-\alpha}, \frac{\sqrt{1-\alpha}}{\alpha}\right) [m]_{\mathrm{Lip}}^{1/2} \left\|L^N(X)\right\|_p^{1/2} \|X - \widetilde{X}^N\|_p^{1/2},$$

for N sufficiently large, where $(L^N(X))$ is a sequence of X-measurable random variables that is bounded in L_p; (ii) under Assumptions (A′)-(B),

$$\|\widetilde{q}_\alpha^N(X) - q_\alpha(X)\|_p = O(N^{-1/2d}),$$

as $N \to \infty$.

Of course, fixed-x consistency results are also quite appealing in quantile regression. Such a result is provided in the following theorem (see Section 2.4 for the proof).

Theorem 2.2.2. *Fix $\alpha \in (0,1)$. Then, under Assumptions (A)-(B),*

$$\sup_{x \in S_X} \left|\widetilde{q}_\alpha^N(x) - q_\alpha(x)\right| \to 0,$$

as $N \to \infty$.

Unlike in Theorem 2.2.1, Theorem 2.2.2 does not provide any rate of convergence. This is a consequence of the fact that, while the convergence of $\widetilde{X}^N$ towards X can be shown to imply the convergence of $\tilde{x}$ towards x for each fixed x, it does not seem possible to show that the rate of convergence in the fixed-x convergence is inherited from the convergence involving X. In other words, one does not have fixed-x analogous result to Proposition 1.2.6 that is the major result that provides the rate of convergence.

We then constructed an approximation of conditional quantile using optimal quantization, that converges to conditional quantile as the tuning parameter N tends to infinity. It was a first step to construct what mostly interests us: a conditional quantile estimator. This will be done in the next section by taking an empirical version of this approximation.

2.3 Estimation of conditional quantile

2.3.1 The proposed estimators and their consistency

Consider now the problem of estimating the conditional quantile $q_\alpha(x)$ on the basis of independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). For any $N(< n)$, the approximation $\widetilde{q}_\alpha^N(x)$ in (2.2.1) leads to an estimator $\widehat{q}_\alpha^{N,n}(x)$ of the conditional quantile $q_\alpha(x)$, through the following two steps.

Step 1: Selection of the grid

As mentioned in Section 1.2, there does not exist any result allowing to calculate an optimal quantization grid and we have to use the stochastic gradient algorithm to construct it. Moreover, in our setting, we only have at our disposal n copies of the X and we do not know its distribution. Therefore, the number of replications for the algorithm is fixed by the sample size: each X_i, $i = 1, \ldots, n$, will play the role of stimuli to construct an "optimal" grid. Optimal is now in quotes since, with the number n of iterations fixed, the grid provided by the algorithm is only an approximation of an optimal one. It is thus expected that only moderate-to-large n will provide reasonable approximations of optimal N-grids.

First of all, we have to choose an initial grid to perform the stochastic gradient algorithm. Since we do not know the distribution of X, the methods described in 1.2.4 and usually applied do not seem suitable here. Therefore, a reasonable choice appears to be the following: the initial grid $\hat{\gamma}^{N,0}$ is obtained by sampling randomly among the X_i's without replacement, and with the constraint[1] that the same x-values cannot be picked more than once. This choice ensures us to start with an initial grid with pairwise distinct components.

Afterwards, the algorithm is performed with n iterations, based on $X^t = X_t$, $t = 1, \ldots, n$. We write $\hat{\gamma}^{N,n} = (\hat{x}_1^{N,n}, \ldots, \hat{x}_N^{N,n})$ for the resulting grid and $\widehat{X}^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(X)$ for the corresponding (empirical) quantization of X; to make the notation less heavy, we will stress the dependence on n in these quantities only when necessary.

Let us point out the notation. Throughout this chapter, we will use both optimal grid in the sense *minimizing the quantization error*, and optimal grid in the sense *provided by the algorithm*. We chose the notation in order to make clear what kind of optimality we are considering. The grid that minimizes the quantization error is always denoted γ^N and the projections on this grid wear the tilde symbol (as $\widetilde{X}^N$ and $\tilde{x}$ that are the projections of X and x on γ^N respectively). The grid provided by the algorithm is always denoted with a hat symbol and the projections on this grid also wear a hat symbol instead of a tilde symbol (as $\widehat{X}^{N,n}$ and $\hat{x}^{N,n}$ that are the projections of X and x on $\hat{\gamma}^{N,n}$ respectively).

Step 2: Estimation

After having constructed an optimal grid $\hat{\gamma}^{N,n}$, we project the X-part of the sample onto this grid. Let $\widehat{X}_i^N = \widehat{X}_i^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(X_i)$. We thus consider now the sample $\{(\widehat{X}_i^N, Y_i)\}_{i=1,\ldots,n}$. The

[1] If the X_i's are i.i.d. with a common density f, sampling without replacement among the X_i's of course implies that this constraint will be met with probability one. One often needs to impose it, however, in real-data examples (due to the possible presence of ties) or when performing bootstrap (see later).

approximation $\widetilde{q}_\alpha^N(x) = \arg\min_a \mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}]$ is then estimated by taking an empirical version, i.e.

$$\widehat{q}_\alpha^{N,n}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^n \rho_\alpha(Y_i - a)\, \frac{\mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}}{\#\left\{j\in\{1,\dots,n\}:\widehat{X}_j^N = \hat{x}^N\right\}},$$

where $\hat{x}^N = \hat{x}^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(x)$. Equivalently, we rather define

$$\widehat{q}_\alpha^{N,n}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^n \rho_\alpha(Y_i - a)\, \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}. \tag{2.3.1}$$

Of course, $\widehat{q}_\alpha^{N,n}(x)$, in practice, is simply evaluated as the sample α-quantile of the Y_i's whose corresponding $\widehat{X}_i^N$ is equal to $\hat{x}^N$.

This projection of X_i's thus proceeds according to their belongings to Voronoi cells that are partitioning the covariate space $\mathbb{R}^d$ (see Section 1.2). This estimator can therefore be viewed as a nonparametric partitioning regression estimator. Partitioning ideas were already used in the construction of least square estimators; see, e.g., Györfi et al. (2002) for a complete overview and Kohler et al. (2006) or Cattaneo and Farrell (2013) for more recent results.

For fixed N (and x), the convergence in probability of $\widehat{q}_\alpha^{N,n}(x)$ to $\widetilde{q}_\alpha^N(x)$ as $n\to\infty$ can be obtained by making use of the convergence results for the stochastic gradient algorithm discussed in Section 1.2.5. In order to do so, we need to restrict to $p=2$ (that is, to the CLVQ algorithm) since this is the only case for which convergence results were established. We also need to adopt the following assumption.

ASSUMPTION (C) P_X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

We then have the following result, whose proof is given in Section 2.5.

Theorem 2.3.1. *Fix $\alpha\in(0,1)$, $x\in S_X$ and $N\in\mathbb{N}_0$. Then, under Assumptions (A), (B)(i), and (C), we have that, as $n\to\infty$,*

$$|\widehat{q}_\alpha^{N,n}(x) - \widetilde{q}_\alpha^N(x)| \to 0,$$

in probability, provided that quantization is based on $p=2$.

In the previous section, we showed that, as $N\to\infty$, $\widetilde{q}_\alpha^N(x) - q_\alpha(x)$ goes to zero almost surely, hence in probability. Theorem 2.3.1 then suggests that $\widehat{q}_\alpha^{N,n}(x) - q_\alpha(x)$ might go to zero in probability as both n and N go to infinity in some appropriate way. Obtaining such a double asymptotic result, however, is extremely delicate, since, to the best of our knowledge, all convergence results for the stochastic gradient algorithm in the literature are as $n\to\infty$ with N fixed. In any case, this is of academic interest only; in practice, indeed, the sample size n is of course fixed and the user needs to choose appropriately the size N of the quantization grid. We propose in Chapter 3 a method that allows to choose this smoothing parameter N by achieving a bias-variance trade-off (see Figure 2.2 below).

2.3.2 Numerical example and bootstrap modification

For the sake of illustration, we evaluated the estimator $\widehat{q}_\alpha^{N,n}(x)$ for $N = 10$, 25 and 40, in a sample of $n = 300$ mutually independent observations (X_i, Y_i) obtained from the model

$$Y = \frac{1}{5}X^3 + \varepsilon, \tag{2.3.2}$$

where $X = 6Z - 3$, with $Z \sim \text{Beta}(0.3, 0.3)$, and $\varepsilon \sim \mathcal{N}(0, 1)$ are independent. The left panels in Figure 2.2 plot the corresponding quantile curves $x \mapsto \widehat{q}_\alpha^{N,n}(x)$ for $\alpha = 0.05, 0.25, 0.5, 0.75$, and 0.95 (actually, these curves were evaluated only at 300 equispaced points in $(-3, 3)$).

It is seen that the number N of quantizers used has an important impact on the curves. For small N, the curves are not smooth and show a large bias. For large N, bias is smaller but the variability is large. This may seem contradictory at first sight, since one stated in Section 1.2 that the quantization error decreases with N. However, this result only holds for an optimal grid γ^N and not for the grid provided by the algorithm. Indeed, the number of iterations of this algorithm is fixed by the sample size, and only one point moves at each iteration. It is then obvious that the larger N is, the more iterations one needs to have a good approximation of the optimal grid γ^N. One should then keep in mind that, for large N, the grid provided by the CLVQ algorithm poorly approximates the corresponding optimal N-grid, since a fixed number n of iterations is used in the algorithm (that should not be too small compared to N).

Smoother quantile curves can be obtained from the bootstrap. The starting idea of this procedure comes from the observation that performing the CLVQ algorithm several times provides a different optimal grid every time. This is due to the fact that the initial grid is chosen differently since one selects it randomly among the observations, as explained in Section 2.3.1. We also notice that, even if we start from the same initial grid, the resulting grid differs if we use the observations in a different order when updating the grid. However, there is no reason to prefer one of these grids rather than another one. The bootstrap method is then based on these remarks and generates B optimal grids, then B corresponding estimations of $q_\alpha(x)$. These estimations are all acceptable, and then we take the mean. Let us detail this procedure, that is illustrated step by step in Figure 2.1.

Generating B grids

For some integer B, we first generate B samples of size n with replacement from the initial sample $X_1, \ldots, X_n$, that we denote $(\xi_b^t)_t$, $t = 1, \ldots, n$ and $b = 1, \ldots, B$. We also generate initial grids $\hat{\gamma}_b^{N,0}$ as above, by sampling randomly among the corresponding $(\xi_b^t)_t$ under the constraints that the N values are pairwise distinct. We then perform B times CLVQ with iterations based on ξ_b^t, $t = 1, \ldots, n$ and with initial grid $\hat{\gamma}_b^{N,0}$. This provides B optimal grids $\hat{\gamma}_b^{N,n}$, $b = 1, \ldots, B$ (each of size N).

Bootstrap based estimator

Each of these grids is now used to estimate conditional quantiles. Working again with the original sample (X_i, Y_i), $i = 1, \ldots, n$, we project the X-part onto the grids $\hat{\gamma}_b^{N,n}$, $b = 1, \ldots, B$. Therefore,

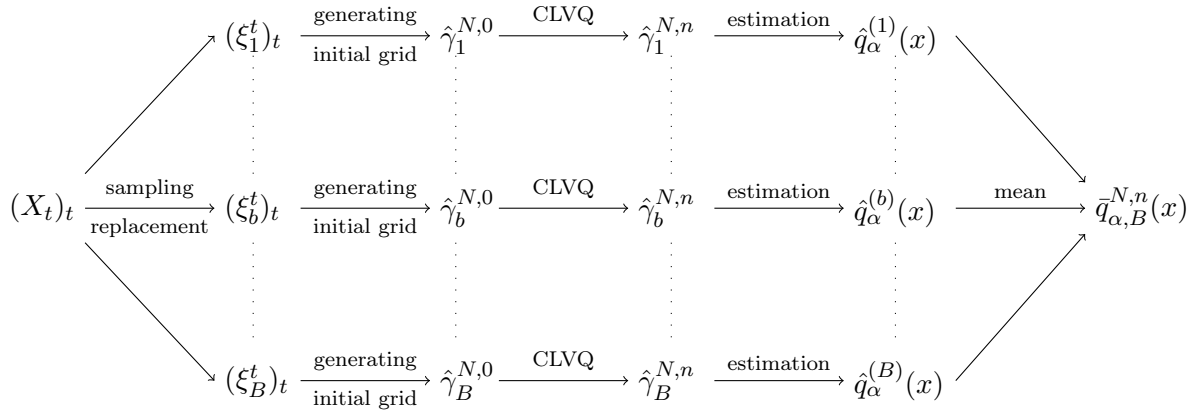


Figure 2.1 – Step by step construction of the bootstrap based estimator $\bar{q}_{\alpha,B}^{N,n}(x)$.

(2.3.1) provides B estimations, denoted $\hat{q}_\alpha^{(b)}(x) = \hat{q}_\alpha^{(b),N,n}(x)$ of $q_\alpha(x)$. This allows to consider the bootstrap estimator

$$\bar{q}_{\alpha,B}^{N,n}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{q}_\alpha^{(b)}(x), \tag{2.3.3}$$

obtained by taking the mean of these B estimations. Bootstrapping, thus, focuses on the construction of the grids and we come back to the original sample in the estimation step.

The right panels of Figure 2.2 plot the resulting bootstrapped quantile curves $x \mapsto \bar{q}_{\alpha,B}^{N,n}(x)$ for the same values of α and N as in the original (non-bootstrapped) versions. Bootstrapping clearly smooths all curves, and moreover significantly reduces boundary effects for small N. These advantages require to take B large enough. But of course, very large values of B should be avoided in order to keep the computational burden under control.

Figure 2.2 also shows that the number N of quantizers has an important impact on the resulting quantile curves. For small N ($N = 10$), the curves show a large bias and a small variability. For larger N ($N = 50$), the bias is smaller but the variability is larger (with or without bootstrap). This bias-variance trade-off is standard in the selection of a smoothing parameter. The intermediate value $N = 25$, that achieves a nice balance between those two extreme cases, was actually obtained from an original N-selection method we describe in Section 3.2 below.

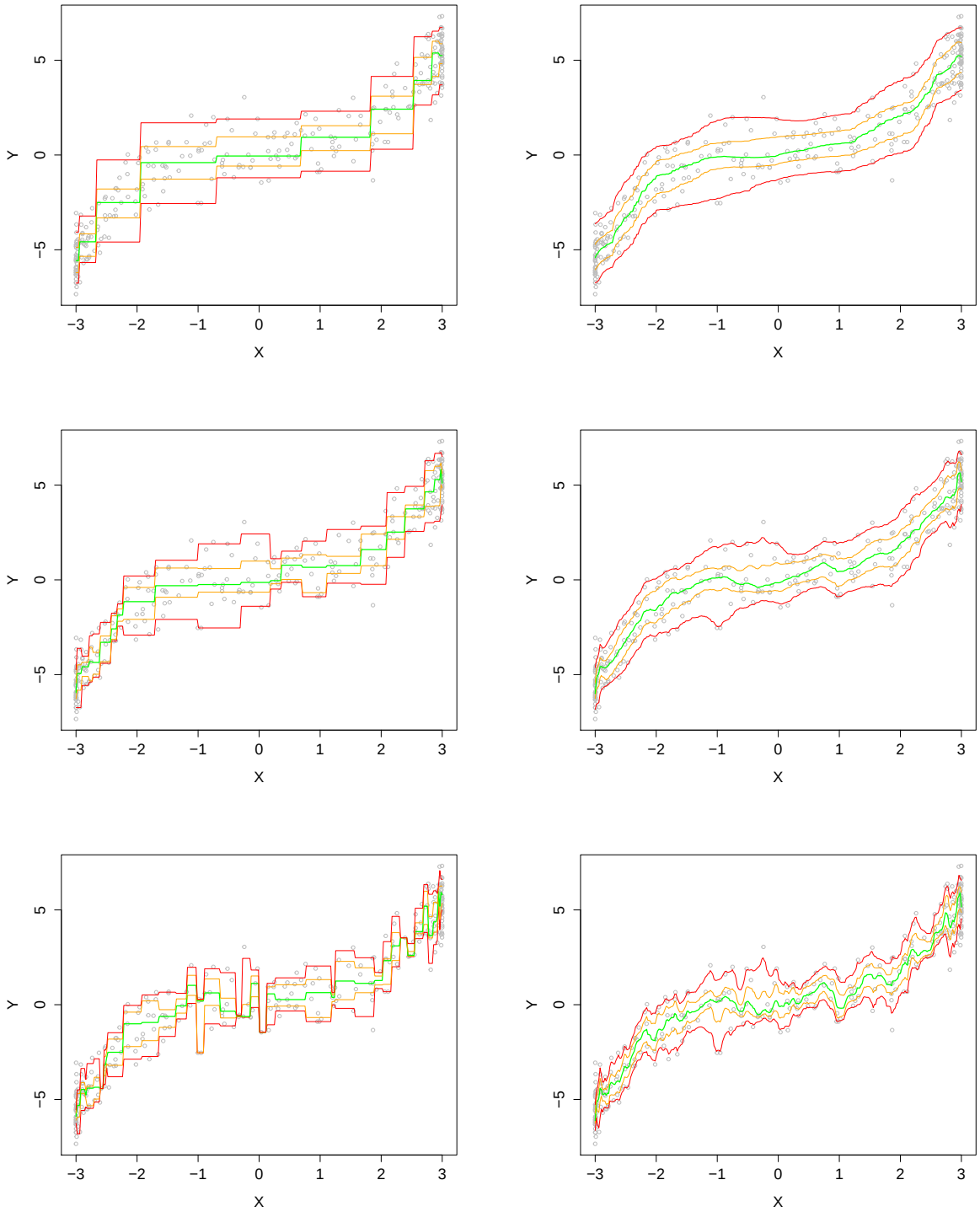
2.4 Proofs of Section 2.2

In this section, we prove Theorems 2.2.1 and 2.2.2, which requires to establish several lemmas. We first introduce some notation. Let

$$G_a(x) = \mathrm{E}[\rho_\alpha(Y - a)|X = x],$$

$$\widetilde{G}_a(\tilde{x}) = \mathrm{E}[\rho_\alpha(Y - a)|\widetilde{X}^N = \tilde{x}].$$

Since the proof of Theorem 2.2.1 requires Theorem 2.2.2 (i.e. its corresponding with x fixed), one will first prove Theorem 2.2.2.



(e) (f)

Figure 2.2 – Estimated conditional quantile curves $x \mapsto \widehat{q}_\alpha^{N,n}(x)$ (left) and their bootstrapped counterparts $x \mapsto \bar{q}_{\alpha,B}^{N,n}(x)$ for $B = 50$ (right), based on $N = 10$ (top), $N = 25$ (middle), and $N = 50$ (bottom). The sample size is $n = 300$, and the quantiles levels considered are α=0.05, 0.25, 0.5, 0.75, and 0.95. See (2.3.2) for the data generating model.

2.4.1 Proof of Theorem 2.2.2

Since $\widetilde{q} = \widetilde{q}_\alpha^N(x) = \arg\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x})$ and $q = q_\alpha(x) = \arg\min_{a\in\mathbb{R}} G_a(x)$, it is natural to try to control the distance between $\widetilde{G}_a(\tilde{x})$ and $G_a(x)$. This is achieved in Lemma 2.4.2, the proof of which requires the following preliminary result. This first lemma proves the Lipschitz property of the functions $\rho_\alpha(\cdot)$ and $G_\alpha(\cdot)$ and gives their corresponding Lipschitz constant. Recall that $m(\cdot,\varepsilon)$ is the link function of our model.

Lemma 2.4.1. *Fix $\alpha \in (0,1)$ and $a \in \mathbb{R}$. Then, under Assumption (A), (i) $\rho_\alpha : \mathbb{R} \to \mathbb{R}$ is Lipschitz, with Lipschitz constant $[\rho_\alpha]_{\text{Lip}} = \max(\alpha, 1-\alpha)$, and (ii) $G_a : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz, with Lipschitz constant equal at most to $[G_a]_{\text{Lip}} = \max(\alpha, 1-\alpha)[m]_{\text{Lip}}$.*

Proof of Lemma 2.4.1. (i) Fix $\alpha \in (0,1)$. We consider different cases.

- If z_1 and z_2 are both positive, then $|\rho_\alpha(z_1) - \rho_\alpha(z_2)| = \alpha|z_1 - z_2|$.
- If they are both negative, then $|\rho_\alpha(z_1) - \rho_\alpha(z_2)| = (1-\alpha)|z_1 - z_2|$.
- Finally, if $z_1 z_2 \leq 0$, then (without loss of generality, we may assume that $z_1 > 0$ and $z_2 \leq 0$), $|\rho_\alpha(z_1) - \rho_\alpha(z_2)| = |\alpha z_1 - (1-\alpha)z_2| \leq \alpha|z_1| + (1-\alpha)|z_2| \leq \max(\alpha, 1-\alpha)(|z_1| + |z_2|) = \max(\alpha, 1-\alpha)|z_1 - z_2|$.

Combining these three cases provides the result.
(ii) Note that, for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned} |G_a(u) - G_a(v)| &= \big|\mathrm{E}[\rho_\alpha(Y-a)|X=u] - \mathrm{E}[\rho_\alpha(Y-a)|X=v]\big| \\ &= \big|\mathrm{E}[\rho_\alpha(m(X,\varepsilon)-a)|X=u] - \mathrm{E}[\rho_\alpha(m(X,\varepsilon)-a)|X=v]\big|, \end{aligned}$$

so that the independence of X and ε entails

$$|G_a(u) - G_a(v)| \leq \mathrm{E}[|\rho_\alpha(m(u,\varepsilon)-a) - \rho_\alpha(m(v,\varepsilon)-a)|] \leq [\rho_\alpha]_{\text{Lip}}[m]_{\text{Lip}}|u-v|,$$

where the last inequality follows from Point (i) of the result and equation (2.2.2). □

We still need the following lemma to prove Theorem 2.2.2. The first point sets that the supremum of the Euclidean distance of any point x and its projection onto an optimal grid tends to zero as N goes to infinity. Point (ii) shows that the supremum of the radius of a Voronoi cell tends also to zero as N goes to infinity. Remember that a Voronoi cell is formed of all points in $\mathbb{R}^d$ that are projected onto the same point called the center of the cell (see Section 1.2). These results are natural: if N becomes greater, it is expected that the distance between any point and its projection will decrease. Similarly, if the number of cells increases, then their radius decreases since we assume that X has a compact support. These two first points are used to set the convergence to zero of the distance between $G_a(x)$ and its quantized version $\widetilde{G}_a(\tilde{x})$, hence of the distance between their respective minimum.

Lemma 2.4.2. *Fix $\alpha \in (0,1)$ and $x \in S_X$. For any integer N, let $\tilde{x} = \tilde{x}^N = \mathrm{Proj}_{\gamma^N}(x)$ and $C_x = C_x^N = \{z \in S_X : \mathrm{Proj}_{\gamma^N}(z) = \tilde{x}\}$. Then, under Assumptions (A)-(B),*
(i) $\sup_{x\in S_X} |x - \tilde{x}| \to 0$ *as* $N \to \infty$;
(ii) $\sup_{x\in S_X} R(C_x) \to 0$ *as* $N \to \infty$, *where we let* $R(C_x) := \sup_{z\in C_x} |z - \tilde{x}|$;
(iii) $\sup_{x\in S_X} \sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)| \to 0$ *as* $N \to \infty$;
(iv) $\sup_{x\in S_X} |\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)| \to 0$ *as* $N \to \infty$.

Proof of Lemma 2.4.2. (i) Assume by contradiction that there exists $\varepsilon > 0$ such that, for infinitely many N (for $N \in \mathcal{N}(\varepsilon)$, say), we have $\sup_{x\in S_X} |\tilde{x}^N - x| > \varepsilon$. For any such value of N, one can pick $x \in S_X$ (that may depend on N) with $|\tilde{x}^N - x| > \varepsilon$. Since $\tilde{x}^N$ is by definition the closest point of the grid to x, no point of the optimal grid γ^N belongs to the ball $\mathcal{B}(x,\varepsilon) = \{z \in \mathbb{R}^d : |z - x| < \varepsilon\}$. This implies that, for all $z \in \mathcal{B}(x, \varepsilon/2)$, $|\tilde{z}^N - z| > \varepsilon/2$, where $\tilde{z}$ is the projection of z onto γ^N. Therefore, for $N \in \mathcal{N}(\varepsilon)$,

$$\begin{aligned} \|\widetilde{X}^{\gamma^N} - X\|_p^p &= \int_{S_X} |\tilde{z}^N - z|^p \, dP_X(z) \geq \int_{\mathcal{B}(x,\varepsilon/2)} |\tilde{z}^N - z|^p \, dP_X(z) \\ &> \left(\frac{\varepsilon}{2}\right)^p \inf_{y\in S_X} P_X\big[\mathcal{B}(y, \varepsilon/2)\big] =: \delta_\varepsilon > 0, \end{aligned} \tag{2.4.1}$$

where the last inequality follows from the fact that $y \mapsto P_X\big[\mathcal{B}(y,\varepsilon/2)\big]$ is a continuous function taking only strictly positive values on the compact set S_X. Since the cardinality of $\mathcal{N}(\varepsilon)$ is infinite, (2.4.1) prevents $\|\widetilde{X}^{\gamma^N} - X\|_p$ to go to zero as $N \to \infty$, a contradiction.
(ii) The result directly follows from Point (i) since $C_x \subset S_X$.
(iii) Fix $a \in \mathbb{R}$. First note that it comes directly from the definitions that the event $[\widetilde{X}^N = \tilde{x}]$ is equivalent to $[X \in C_x]$. Then, one has

$$|\mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}] - \mathrm{E}[\rho_\alpha(Y-a)|X = \tilde{x}]| \leq \sup_{z\in C_x} |\mathrm{E}[\rho_\alpha(Y-a)|X = z] - \mathrm{E}[\rho_\alpha(Y-a)|X = \tilde{x}]|.$$

Therefore,

$$\begin{aligned} &|\widetilde{G}_a(\tilde{x}) - G_a(x)| = |\mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}] - \mathrm{E}[\rho_\alpha(Y-a)|X = x]| \\ &\leq |\mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}] - \mathrm{E}[\rho_\alpha(Y-a)|X = \tilde{x}]| + |\mathrm{E}[\rho_\alpha(Y-a)|X = \tilde{x}] - \mathrm{E}[\rho_\alpha(Y-a)|X = x]| \\ &\leq 2 \sup_{z\in C_x} |\mathrm{E}[\rho_\alpha(Y-a)|X = z] - \mathrm{E}[\rho_\alpha(Y-a)|X = \tilde{x}]|, \end{aligned}$$

where the last inequality comes from the fact that x is a particular point of C_x. Using the independence between X and ε and the Lipschitz properties of ρ_α and m then yields

$$\begin{aligned} |\widetilde{G}_a(\tilde{x}) - G_a(x)| &\leq 2 \sup_{z\in C_x} |\mathrm{E}[\rho_\alpha(m(z,\varepsilon) - a) - \mathrm{E}[\rho_\alpha(m(\tilde{x},\varepsilon) - a)]| \\ &\leq 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}} \sup_{z\in C_x} |z - \tilde{x}| \\ &= 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}} R(C_x). \end{aligned}$$

Since the last right hand side does not depend on a, it implies that

$$\sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)| \leq 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}} R(C_x).$$

Hence,

$$\sup_{x\in S_X}\sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)| \leq 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}} \sup_{x\in S_X} R(C_x).$$

The result then follows from Point (ii).

(iv) Letting $\mathbb{I}_+ = \mathbb{I}_{[\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) \geq \min_{a\in\mathbb{R}} G_a(x)]}$, we have

$$\begin{aligned}
|\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)|\mathbb{I}_+ &= \big(\widetilde{G}_{\widetilde{q}_\alpha^N(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)\big)\mathbb{I}_+ \\
&\leq \big(\widetilde{G}_{q_\alpha(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)\big)\big)\mathbb{I}_+ \\
&\leq \sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)|\mathbb{I}_+.
\end{aligned}$$

Proceeding similarly with $\mathbb{I}_- = \mathbb{I}_{[\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) < \min_{a\in\mathbb{R}} G_a(x)]}$, this yields

$$\begin{aligned}
|\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)|\mathbb{I}_- &= \big(G_{q_\alpha(x)}(x) - \widetilde{G}_{\widetilde{q}_\alpha^N(x)}(\tilde{x})\big)\mathbb{I}_- \\
&\leq \big(G_{\widetilde{q}_\alpha^N(x)}(x) - \widetilde{G}_{\widetilde{q}_\alpha^N(x)}(\tilde{x})\big)\mathbb{I}_- \\
&\leq \sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)|\mathbb{I}_-,
\end{aligned}$$

so that $|\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)| \leq \sup_{a\in\mathbb{R}} |\widetilde{G}_a(\tilde{x}) - G_a(x)|$. The result therefore follows from Point (iii). □

We can now prove Theorem 2.2.2. The idea of the proof is to derive the convergence to zero of the difference of the argmin from Points (iii) and (iv) of Lemma 2.4.2.

Proof of Theorem 2.2.2. First note that, for any $x \in S_X$,

$$\begin{aligned}
|G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| &\leq |G_{\widetilde{q}_\alpha^N(x)}(x) - \widetilde{G}_{\widetilde{q}_\alpha^N(x)}(\tilde{x})| + |\widetilde{G}_{\widetilde{q}_\alpha^N(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)| \\
&\leq \sup_{a\in\mathbb{R}} |G_a(x) - \widetilde{G}_a(\tilde{x})| + |\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)| \\
&\leq \sup_{x\in S_X}\sup_{a\in\mathbb{R}} |G_a(x) - \widetilde{G}_a(\tilde{x})| + \sup_{x\in S_X} |\min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x}) - \min_{a\in\mathbb{R}} G_a(x)|.
\end{aligned}$$

Therefore, Lemma 2.4.2(iii)-(iv) readily implies that, as $N \to \infty$,

$$\sup_{x\in S_X} |G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \to 0. \tag{2.4.2}$$

Now, let $\tilde{N}$ be such that, for any $N \geq \tilde{N}$, we have $|G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \leq 1$ for all $x \in S_X$. As we will show later in this proof, this implies that there exists M such that

$$|\widetilde{q}_\alpha^N(x) - q_\alpha(x)| \leq M, \tag{2.4.3}$$

for all $x \in S_X$ and $N \geq \tilde{N}$.

One can easily check that, for any $x \in S_X$, $a \mapsto G_a(x)$ is twice continuously differentiable, with derivatives

$$\frac{dG_a(x)}{da} = F^{\varepsilon}\left(\frac{a - m_1(x)}{m_2(x)}\right) - \alpha \quad \text{and} \quad \frac{d^2G_a(x)}{da^2} = \frac{1}{m_2(x)} f^{\varepsilon}\left(\frac{a - m_1(x)}{m_2(x)}\right),$$

where $F^{\varepsilon}(\cdot)$ denotes the cdf of ε. We then perform a second-order expansion about $a = q_\alpha(x)$. It provides

$$\begin{aligned} G_{\widetilde{q}_\alpha^N(x)}(x) &= G_{q_\alpha(x)}(x) + \left.\frac{dG_a(x)}{da}\right|_{a=q_\alpha(x)} (\widetilde{q}_\alpha^N(x) - q_\alpha(x)) + \left.\frac{d^2G_a(x)}{da^2}\right|_{a=q_{\alpha*}^N(x)} \frac{(\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2}{2} \\ &= G_{q_\alpha(x)}(x) + \frac{1}{2m_2(x)} f^{\varepsilon}\left(\frac{q_{\alpha*}^N(x) - m_1(x)}{m_2(x)}\right)(\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2, \end{aligned}$$

for some $q_{\alpha*}^N(x)$ between $\widetilde{q}_\alpha^N(x)$ and $q_\alpha(x)$, where the last equality comes from the fact that $q_\alpha(x)$ is a minimum in a of the function $G_a(x)$. The first derivative is then equal to zero in this point. Hence,

$$G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x) = \frac{1}{2m_2(x)} f^{\varepsilon}\left(\frac{q_{\alpha*}^N(x) - m_1(x)}{m_2(x)}\right)(\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2.$$

Isolating the term of interest $(\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2$ and taking the supremum over $x \in S_X$ yield

$$\sup_{x \in S_X} (\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2 \leq \frac{2 \sup_{x \in S_X} m_2(x)}{\inf_{x \in S_X} f^{\varepsilon}\left(\frac{q_{\alpha*}^N(x) - m_1(x)}{m_2(x)}\right)} \sup_{x \in S_X} |G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)|. \tag{2.4.4}$$

Since $m_2(\cdot)$ is a continuous function defined over the compact set S_X, it is bounded and we have

$$\sup_{x \in S_X} m_2(x) \leq C_a \tag{2.4.5}$$

for some positive constant C_a. Using (2.4.3), we have

$$\begin{aligned} |q_{\alpha*}^N(x) - m_1(x)| &\leq |q_\alpha(x)| + |q_{\alpha*}^N(x) - q_\alpha(x)| + |m_1(x)| \\ &\leq |q_\alpha(x)| + M + |m_1(x)|. \end{aligned}$$

Moreover, since $q_\alpha(\cdot)$, $m_1(\cdot)$, and $m_2(\cdot)$ are continuous functions also defined over this compact set (with $m_2(\cdot)$ taking strictly positive values), we get that, for $N \geq \tilde{N}$,

$$\sup_{x \in S_X} \frac{|q_{\alpha*}^N(x) - m_1(x)|}{m_2(x)} \leq \frac{\sup_{x \in S_X} \left(|q_\alpha(x)| + |m_1(x)| + M\right)}{\inf_{x \in S_X} m_2(x)} \leq C_b,$$

for some constant C_b. Jointly with the continuity of the (strictly) positive function $f^{\varepsilon}(\cdot)$, this implies that the infimum in (2.4.4) admits a strictly positive lower bound that is independent of N (for $N \geq \tilde{N}$). Using this, (2.4.2) and (2.4.5), we conclude from (2.4.4) that

$$\sup_{x \in S_X} (\widetilde{q}_\alpha^N(x) - q_\alpha(x))^2 \to 0,$$

as $N \to \infty$, which was to be proved.

It remains to show that the claim in (2.4.3) indeed holds true. By contradiction, assume that for all M, there exists $x = x_M$ (and $N \geq \tilde{N}$) such that $|\widetilde{q}_\alpha^N(x_M) - q_\alpha(x_M)| > M$. The convexity of $a \mapsto G_a(x)$ implies that, for all a between $q_\alpha(x_M)$ and $\widetilde{q}_\alpha^N(x_M)$, the image of a by $G_a(x_M)$ is less than its image by the straight line between $q_\alpha(x_M)$ and $\widetilde{q}_\alpha^N(x_M)$. This line admits as equation

$$y = G_{q_\alpha(x_M)}(x_M) + \frac{|G_{\widetilde{q}_\alpha^N(x_M)}(x_M) - G_{q_\alpha(x_M)}(x_M)|}{|\widetilde{q}_\alpha^N(x_M) - q_\alpha(x_M)|}|a - q_\alpha(x_M)|,$$

which implies

$$G_a(x_M) \leq G_{q_\alpha(x_M)}(x_M) + \frac{|G_{\widetilde{q}_\alpha^N(x_M)}(x_M) - G_{q_\alpha(x_M)}(x_M)|}{|\widetilde{q}_\alpha^N(x_M) - q_\alpha(x_M)|}\,|a - q_\alpha(x_M)|.$$

Given the fact that $|G_{\widetilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \leq 1$ for all $x \in S_X$ (for $N \geq \tilde{N}$) and that $|\widetilde{q}_\alpha^N(x_M) - q_\alpha(x_M)| > M$, we have, for any a with $|a - q_\alpha(x_M)| \leq M$,

$$G_a(x_M) \leq G_{q_\alpha(x_M)}(x_M) + \frac{1}{M}\,|a - q_\alpha(x_M)|. \tag{2.4.6}$$

This property is illustrated in Figure 2.3 for the case $q_\alpha(x_M) \leq a \leq \widetilde{q}_\alpha^N(x_M)$. In particular,

$$G_{q_\alpha(x_M)+1}(x_M) \leq G_{q_\alpha(x_M)}(x_M) + \frac{1}{M} \tag{2.4.7}$$

for any integer M. The smoothness of $a \mapsto G_a(x)$ ensures that, for any M, there exists q_M between $q_\alpha(x_M)$ and $q_\alpha(x_M) + 1$ such that

$$G_{q_\alpha(x_M)+1}(x_M) = G_{q_\alpha(x_M)}(x_M) + \frac{1}{2}\,\frac{d^2 G_a(x_M)}{da^2}\bigg|_{a=q_M}.$$

In view of (2.4.7), the sequences (x_M) and (q_M) are such that

$$\frac{1}{m_2(x_M)} f^\varepsilon\left(\frac{q_M - m_1(x_M)}{m_2(x_M)}\right) = \frac{d^2 G_a(x_M)}{da^2}\bigg|_{a=q_M} \to 0$$

as $M \to \infty$. Using again the fact that $m_2(\cdot)$ is a continuous function defined over the compact set S_X, we conclude that

$$f^\varepsilon\left(\frac{q_M - m_1(x_M)}{m_2(x_M)}\right) \to 0 \tag{2.4.8}$$

as $M \to \infty$. By proceeding as above, we can however show that the argument of f^ε in (2.4.8) remains bounded as $M \to \infty$, so that (2.4.8) contradicts the continuity of $f^\varepsilon : \mathbb{R} \to \mathbb{R}_0^+$. Therefore, the claim (2.4.3) is indeed true and this concludes the proof of Theorem 2.2.2. □

2.4.2 Proof of Theorem 2.2.1

The proof of Theorem 2.2.1 still requires the following three lemmas. Lemma 2.4.3 and 2.4.4 are the analogous to Lemma 2.4.2 (iii) and (iv) for global X and not fixed x. The third lemma that we will prove is a technical lemma.

In this section, q and $\tilde{q}$ will stand for $q_\alpha(X)$ and $\widetilde{q}_\alpha^N(X)$, respectively, for simplicity of notation.

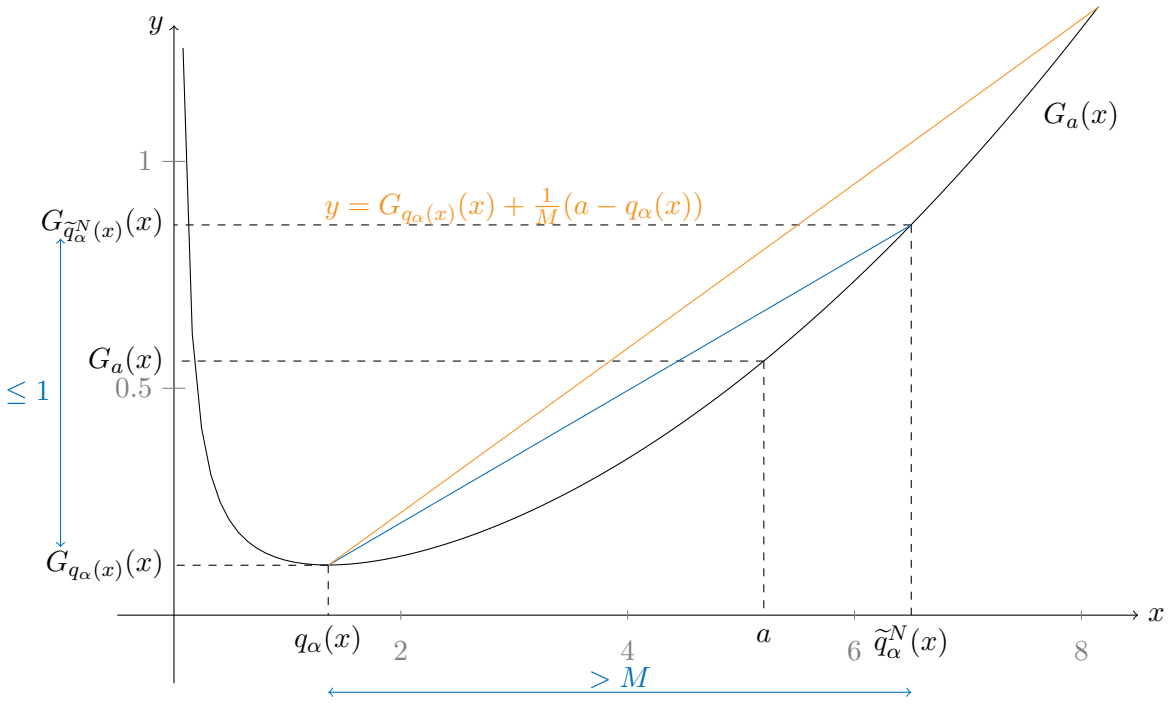


Figure 2.3 – Illustration of the convex function $G_a(x)$ and the resulting property (2.4.6).

Lemma 2.4.3. *Fix $\alpha \in (0,1)$. Then (i) under Assumptions (A)-(B), we have*

$$\big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N) - G_a(X)\big|\big\|_p \leq 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}}\big\|\widetilde{X}^N - X\big\|_p\,; \tag{2.4.9}$$

(ii) under Assumptions (A′)-(B), we have that

$$\big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N) - G_a(X)\big|\big\|_p = O\big(N^{-1/d}\big),$$

as $N \to \infty$.

Lemma 2.4.4. *Fix $\alpha \in (0,1)$. Then (i) under Assumptions (A)-(B), we have $\big\|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N) - G_q(X)\big\|_p \leq 2\max(\alpha, 1-\alpha)[m]_{\mathrm{Lip}}\big\|\widetilde{X}^N - X\big\|_p$; (ii) under Assumptions (A′)-(B), $\big\|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N) - G_q(X)\big\|_p = O(N^{-1/d})$ as $N \to \infty$.*

Lemma 2.4.5. *Let Assumption (B) hold and fix $\alpha, \beta \in (0,1)$. For any $x \in S_X$, let $L(x) = 1/f^{Y|X=x}(q_\alpha(x))$ and $L_\beta^N(x) = 1/f^{Y|X=x}(c_\beta^N(x))$, where $c_\beta^N(x)$ is the infimum of all c's between $q_\alpha(x)$ and $t_\beta(x) = t_\beta(\widetilde{q}_\alpha^N(x), q_\alpha(x)) = \beta\widetilde{q}_\alpha^N(x) + (1-\beta)q_\alpha(x)$ for which*

$$\int_{\min(q_\alpha(x),t_\beta(x))}^{\max(q_\alpha(x),t_\beta(x))} f^{Y|X=x}(y)\,dy = f^{Y|X=x}(c)\,|t_\beta(x) - q_\alpha(x)| \tag{2.4.10}$$

(existence follows from the mean value theorem). Then $\|L_\beta^N(X)\|_p \to \|L(X)\|_p$ as $N \to \infty$.

Notice that, under Assumption (A), the conditional density $f^{Y|X=x}(y)$ rewrites as $f^{Y|X=x}(y) = \frac{1}{m_2(x)} f^{\varepsilon}(\frac{y-m_1(x)}{m_2(x)})$.

Proof of Lemma 2.4.3. Point (ii) of the result readily follows from Point (i) and Proposition 1.2.6, so that we may focus on the proof of Point (i). Note that $G_a(\widetilde{X}^N)$ stands for the conditional expectation of $\rho_\alpha(Y-a)$ given that $X=\widetilde{X}^N$, which is different from $\mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N]$. For any a,

$$\begin{aligned}\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big| &\le \big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big|+\big|G_a(\widetilde{X}^N)-G_a(X)\big| \\ &\le \sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big|+\sup_a\big|G_a(\widetilde{X}^N)-G_a(X)\big|\end{aligned}$$

almost surely, so that

$$\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big| \le \sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big|+\sup_a\big|G_a(\widetilde{X}^N)-G_a(X)\big|$$

almost surely. The triangular inequality then yields

$$\big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big|\big\|_p \le \big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big|\big\|_p+\big\|\sup_a\big|G_a(\widetilde{X}^N)-G_a(X)\big|\big\|_p. \tag{2.4.11}$$

Since $\widetilde{X}^N$ is X-measurable, we have that

$$\widetilde{G}_a(\widetilde{X}^N)=\mathrm{E}\big[\rho_\alpha(Y-a)|\widetilde{X}^N\big]=\mathrm{E}\big[\mathrm{E}[\rho_\alpha(Y-a)|X]|\widetilde{X}^N\big]=\mathrm{E}\big[G_a(X)|\widetilde{X}^N\big],$$

which gives

$$\begin{aligned}\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big| &= \sup_a\big|\mathrm{E}[G_a(X)-G_a(\widetilde{X}^N)|\widetilde{X}^N]\big| \\ &\le \mathrm{E}\Big[\sup_a|G_a(X)-G_a(\widetilde{X}^N)|\,\Big|\widetilde{X}^N\Big],\end{aligned}$$

almost surely. From Jensen's inequality, we then obtain

$$\big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(\widetilde{X}^N)\big|\big\|_p \le \big\|\sup_a\big|G_a(X)-G_a(\widetilde{X}^N)\big|\big\|_p.$$

Substituting in (2.4.11) and using Lemma 2.4.1(ii) yields

$$\begin{aligned}\big\|\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big|\big\|_p &\le 2\big\|\sup_a\big|G_a(X)-G_a(\widetilde{X}^N)\big|\big\|_p \\ &\le 2\max(\alpha,1-\alpha)[m]_{\mathrm{Lip}}\big\|\widetilde{X}^N-X\big\|_p,\end{aligned}$$

which establishes the result. □

Proof of Lemma 2.4.4. The idea of this proof is really similar to the one of Lemma 2.4.2 (iv). Letting $\mathbb{I}_+=\mathbb{I}_{[\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)\ge G_q(X)]}$, note that

$$\begin{aligned}|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)-G_q(X)|\,\mathbb{I}_+ &\le \big(\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)-G_q(X)\big)\mathbb{I}_+ \le \big(\widetilde{G}_q(\widetilde{X}^N)-G_q(X)\big)\mathbb{I}_+ \\ &\le \Big(\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big|\Big)\mathbb{I}_+,\end{aligned}$$

almost surely. Similarly, with $\mathbb{I}_-=\mathbb{I}_{[\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)<G_q(X)]}$, we have

$$\begin{aligned}|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)-G_q(X)|\,\mathbb{I}_- &\le \big(G_q(X)-\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)\big)\mathbb{I}_- \le \big(G_{\tilde{q}}(X)-\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)\big)\mathbb{I}_- \\ &\le \Big(\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big|\Big)\mathbb{I}_-,\end{aligned}$$

almost surely, so that $|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N)-G_q(X)|\le\sup_a\big|\widetilde{G}_a(\widetilde{X}^N)-G_a(X)\big|$, almost surely. The result then directly follows from Lemma 2.4.3. □

Proof of Lemma 2.4.5. We have to prove that, as $N \to \infty$,

$$\int_{S_X} \frac{1}{\left(f^{Y|X=x}(c_\beta^N(x))\right)^p}\, dP_X(x) \to \int_{S_X} \frac{1}{\left(f^{Y|X=x}(q_\alpha(x))\right)^p}\, dP_X(x). \tag{2.4.12}$$

First note that Assumption (B) ensures that, for any $x \in S_X$,

$$y \mapsto f^{Y|X=x}(y) = \frac{1}{m_2(x)}\, f^\varepsilon\left(\frac{y - m_1(x)}{m_2(x)}\right)$$

is continuous. Therefore, Theorem 2.2.2, which clearly entails that $c_\beta^N(x) \to q_\alpha(x)$ for any x (and even uniformly in x) as $N \to \infty$, implies that

$$\frac{1}{\left(f^{Y|X=x}(c_\beta^N(x))\right)^p} \to \frac{1}{\left(f^{Y|X=x}(q_\alpha(x))\right)^p},$$

still for any x as $N \to \infty$. To establish (2.4.12), it is then sufficient — in view of Lebesgue's dominated convergence theorem — to prove that, for any x and any (sufficiently large) N,

$$\frac{1}{f^{Y|X=x}(c_\beta^N(x))} = \frac{m_2(x)}{f^\varepsilon\left(\frac{c_\beta^N(x) - m_1(x)}{m_2(x)}\right)} \leq C \tag{2.4.13}$$

for some constant C that does not depend on N.

To show (2.4.13), note that Theorem 2.2.2 and the continuity of $m_1(\cdot)$ and $m_2(\cdot)$ (with $m_2(\cdot)$ taking strictly positive values) over the compact set S_X entail that, for N sufficiently large,

$$\begin{aligned}
\left|\frac{c_\beta^N(x) - m_1(x)}{m_2(x)}\right| &= \left|\frac{(q_\alpha(x) - m_1(x)) + (c_\beta^N(x) - q_\alpha(x))}{m_2(x)}\right| \\
&= \left|\varepsilon_\alpha + \frac{c_\beta^N(x) - q_\alpha(x)}{m_2(x)}\right| \\
&\leq |\varepsilon_\alpha| + \frac{|c_\beta^N(x) - q_\alpha(x)|}{D_1} \leq |\varepsilon_\alpha| + \frac{1}{D_1} = D_2,
\end{aligned}$$

for some constants D_1, D_2 (that do not depend on N), where ε_α denotes the α-quantile of ε (it is direct from the expression of $m(\cdot, \varepsilon)$ that $\varepsilon_\alpha = (q_\alpha(x) - m_1(x))/m_2(x)$). Consequently, (2.4.13) directly follows from the continuity of $m_2(\cdot)$ and $f^\varepsilon(\cdot) : \mathbb{R} \to \mathbb{R}_0^+$. □

Finally, we prove Theorem 2.2.1.

Proof of Theorem 2.2.1. Throughout the proof, we write $q(x)$ and $\tilde{q}(x)$ for $q_\alpha(x)$ and $\widetilde{q}_\alpha^N(x)$, respectively. Remember that q and $\tilde{q}$ stand for $q_\alpha(X)$ and $\widetilde{q}_\alpha^N(X)$ respectively. For any $r, s \in \mathbb{R}$ and $\beta \in (0,1)$, we also let $t_\beta(r,s) := \beta r + (1-\beta)s$.

(i) Let first $r, s \in \mathbb{R}$ with $r \leq s$. It is then easy to show that, for all $y \in \mathbb{R}$, one has

$$\rho_\alpha(y - r) - \rho_\alpha(y - s) \geq -(1-\alpha)(s-r)\mathbb{I}_{[y \leq t_\alpha(r,s)]} + \alpha(s-r)\mathbb{I}_{[y > s]}. \tag{2.4.14}$$

Indeed,

- if $y > t_\alpha(r,s)$, then $\rho_\alpha(y-r) - \rho_\alpha(y-s)$ is positive. Multiplying by it the trivial inequality $1 \geq \mathbb{I}_{[y \leq t_\alpha(r,s)]} + \mathbb{I}_{[y>s]}$ provides directly (2.4.14);

- if $r < y \leq t_\alpha(r,s)$, then $\rho_\alpha(y-r) - \rho_\alpha(y-s) = \alpha(s-r) + (y-s)$. (2.4.14) follows since $y > r$;
- if $y \leq r$, then $\rho_\alpha(y-r) - \rho_\alpha(y-s) = -(1-\alpha)(s-r)$, hence the result.

Therefore, using (2.4.14),

$$\big\{\rho_\alpha(Y-\tilde{q}) - \rho_\alpha(Y-q)\big\}\mathbb{I}_{[\tilde{q}\leq q]} \geq \big\{-(1-\alpha)(q-\tilde{q})\mathbb{I}_{[Y\leq t_\alpha(\tilde{q},q)]} + \alpha(q-\tilde{q})\mathbb{I}_{[Y>q]}\big\}\mathbb{I}_{[\tilde{q}\leq q]}.$$

For simplicity of notations, we will write in the sequel $t_\alpha = t_\alpha(\tilde{q},q)$. Taking conditional expectation on X then yields

$$\begin{aligned}
|G_{\tilde{q}}(X) - G_q(X)|\mathbb{I}_{[\tilde{q}\leq q]} &= \big(G_{\tilde{q}}(X) - G_q(X)\big)\mathbb{I}_{[\tilde{q}\leq q]} \\
&\geq \big\{-(1-\alpha)(q-\tilde{q})P[Y\leq t_\alpha|X] + \alpha(q-\tilde{q})P[Y>q|X]\big\}\mathbb{I}_{[\tilde{q}\leq q]} \\
&= (1-\alpha)(q-\tilde{q})\big(\alpha - P[Y\leq t_\alpha|X]\big)\mathbb{I}_{[\tilde{q}\leq q]} \\
&\geq \min(\alpha,1-\alpha)|\tilde{q}-q|\big(P[Y\leq q|X] - P[Y\leq t_\alpha|X]\big)\mathbb{I}_{[\tilde{q}\leq q]} \\
&= \min(\alpha,1-\alpha)|\tilde{q}-q|P[t_\alpha < Y \leq q|X]\mathbb{I}_{[\tilde{q}\leq q]}, \qquad (2.4.15)
\end{aligned}$$

almost surely.

Now, for $r,s\in\mathbb{R}$ with $r>s$, we prove similarly that

$$\rho_\alpha(y-r) - \rho_\alpha(y-s) \geq -(1-\alpha)(s-r)\mathbb{I}_{[y\leq s]} + \alpha(s-r)\mathbb{I}_{[y>t_{1-\alpha}(r,s)]},$$

for all $y\in\mathbb{R}$. Hence,

$$\big\{\rho_\alpha(Y-\tilde{q}) - \rho_\alpha(Y-q)\big\}\mathbb{I}_{[\tilde{q}>q]} \geq \big\{-(1-\alpha)(q-\tilde{q})\mathbb{I}_{[Y\leq q]} + \alpha(q-\tilde{q})\mathbb{I}_{[Y>t_{1-\alpha}(\tilde{q},q)]}\big\}\mathbb{I}_{[\tilde{q}>q]}.$$

Taking expectation conditional on X, this gives (throughout with $t_{1-\alpha} = t_{1-\alpha}(\tilde{q},q)$, where we stress that $\tilde{q}$ and q still stand for $\widetilde{q}_\alpha^N(X)$ and $q_\alpha(X)$, respectively)

$$\begin{aligned}
|G_{\tilde{q}}(X) - G_q(X)|\mathbb{I}_{[\tilde{q}>q]} &= \big(G_{\tilde{q}}(X) - G_q(X)\big)\mathbb{I}_{[\tilde{q}>q]} \\
&\geq \big\{-(1-\alpha)(q-\tilde{q})P[Y\leq q|X] + \alpha(q-\tilde{q})P[Y>t_{1-\alpha}|X]\big\}\mathbb{I}_{[\tilde{q}>q]} \\
&= \alpha(q-\tilde{q})\big(P[Y>t_{1-\alpha}|X] - (1-\alpha)\big)\mathbb{I}_{[\tilde{q}>q]} \\
&\geq \min(\alpha,1-\alpha)|\tilde{q}-q|\big(P[Y\leq t_{1-\alpha}|X] - P[Y\leq q|X]\big)\mathbb{I}_{[\tilde{q}>q]} \\
&= \min(\alpha,1-\alpha)|\tilde{q}-q|P[q < Y \leq t_{1-\alpha}|X]\mathbb{I}_{[\tilde{q}>q]}, \qquad (2.4.16)
\end{aligned}$$

almost surely.

Our strategy consists in combining (2.4.15)-(2.4.16) to obtain an almost sure lower bound for $|G_{\tilde{q}}(X) - G_q(X)|$. We will need to consider the cases $\alpha\leq 1/2$ and $\alpha > 1/2$ separately.

(a) In the case $\alpha\leq 1/2$, we have, under $\tilde{q}>q$, that $t_\alpha \leq t_{1-\alpha}$, so that (2.4.16) entails

$$|G_{\tilde{q}}(X) - G_q(X)|\mathbb{I}_{[\tilde{q}>q]} \geq \min(\alpha,1-\alpha)|\tilde{q}-q|P[q<Y\leq t_\alpha|X]\mathbb{I}_{[\tilde{q}>q]}. \qquad (2.4.17)$$

Summing (2.4.15) and (2.4.17) then provides

$$\left|G_{\tilde{q}}(X) - G_q(X)\right| \geq \min(\alpha, 1-\alpha)|\tilde{q} - q| P[\min(t_\alpha, q) < Y \leq \max(t_\alpha, q)|X]. \tag{2.4.18}$$

Now, for any $x \in S_X$,

$$\begin{aligned} P[\min(t_\alpha, q) < Y \leq \max(t_\alpha, q)|X = x] &= \int_{\min(q(x), t_\alpha(x))}^{\max(q(x), t_\alpha(x))} f^{Y|X=x}(y)\, dy \\ &= f^{Y|X=x}(c_\alpha^N(x))\, |t_\alpha(x) - q(x)| = \frac{|t_\alpha(x) - q(x)|}{L_\alpha^N(x)}, \end{aligned}$$

where $t_\alpha(x) = t_\alpha(\tilde{q}(x), q(x))$, $c_\alpha^N(x)$ and $L_\alpha^N(x)$ were defined in Lemma 2.4.5. It follows that

$$P[\min(t_\alpha, q) \leq Y < \max(t_\alpha, q)|X] = \frac{|t_\alpha - q|}{L_\alpha^N(X)}$$

almost surely. Since $|t_\alpha - q| = \alpha|\tilde{q} - q| \geq \min(\alpha, 1-\alpha)|\tilde{q} - q|$, we have

$$P[\min(t_\alpha, q) \leq Y < \max(t_\alpha, q)|X] \geq \min(\alpha, 1-\alpha) \frac{|\tilde{q} - q|}{L_\alpha^N(X)}$$

almost surely.

Plugging into (2.4.18) yields that it almost surely holds that

$$|\tilde{q} - q|^2 \leq \frac{1}{(\min(\alpha, 1-\alpha))^2} L_\alpha^N(X) |G_{\tilde{q}}(X) - G_q(X)|$$

or equivalently, that

$$|\tilde{q} - q|^p \leq \frac{1}{(\min(\alpha, 1-\alpha))^p} (L_\alpha^N(X))^{p/2} |G_{\tilde{q}}(X) - G_q(X)|^{p/2}.$$

Taking expectations, applying the Cauchy-Schwarz inequality, then computing pth roots, provide

$$\|\tilde{q} - q\|_p \leq \frac{1}{\min(\alpha, 1-\alpha)} \left\|L_\alpha^N(X)\right\|_p^{1/2} \|G_{\tilde{q}}(X) - G_q(X)\|_p^{1/2}. \tag{2.4.19}$$

From Lemmas 2.4.3-2.4.4, we obtain

$$\begin{aligned} \left\|G_{\tilde{q}}(X) - G_q(X)\right\|_p &\leq \left\|G_{\tilde{q}}(X) - \widetilde{G}_{\tilde{q}}(\widetilde{X}^N)\right\|_p + \left\|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N) - G_q(X)\right\|_p \\ &\leq \left\|\sup_a |G_a(X) - \widetilde{G}_a(\widetilde{X}^N)|\right\|_p + \left\|\widetilde{G}_{\tilde{q}}(\widetilde{X}^N) - G_q(X)\right\|_p \\ &\leq 4 \max(\alpha, 1-\alpha) [m]_{\mathrm{Lip}} \left\|\widetilde{X}^N - X\right\|_p. \end{aligned}$$

The result (in the case $\alpha \leq 1/2$) then follows by plugging this into (2.4.19) (the boundedness of $L^N(X) = L_\alpha^N(X)$ in L_p for all $\beta \in (0,1)$ is a direct corollary of Lemma 2.4.5).

(b) We now turn to the case $\alpha > 1/2$. Here, we have that $t_\alpha \leq t_{1-\alpha}$ under $\tilde{q} \leq q$, so that (2.4.15) yields

$$|G_{\tilde{q}}(X) - G_q(X)| \mathbb{I}_{[\tilde{q} \leq q]} \geq \min(\alpha, 1-\alpha)|\tilde{q} - q| P[t_{1-\alpha} < Y \leq q|X] \mathbb{I}_{[\tilde{q} \leq q]}. \tag{2.4.20}$$

Summing (2.4.16) and (2.4.20) then provides

$$\left|G_{\tilde{q}}(X) - G_q(X)\right| \geq \min(\alpha, 1-\alpha)|\tilde{q} - q| P[\min(t_{1-\alpha}, q) < Y \leq \max(t_{1-\alpha}, q)|X]. \tag{2.4.21}$$

The rest of the proof is entirely similar to the case $\alpha \leq 1/2$. Indeed, for any $x \in S_X$,

$$P[\min(t_{1-\alpha}, q) < Y \leq \max(t_{1-\alpha}, q) | X = x] = \int_{\min(q(x), t_{1-\alpha}(x))}^{\max(q(x), t_{1-\alpha}(x))} f^{Y|X=x}(y)\, dy$$

$$= f^{Y|X=x}(c_{1-\alpha}^N(x))\, |t_{1-\alpha}(x) - q(x)| = \frac{|t_{1-\alpha}(x) - q(x)|}{L_{1-\alpha}^N(x)},$$

where $t_{1-\alpha}(x) = t_{1-\alpha}(\tilde{q}(x), q(x))$, $c_{1-\alpha}^N(x)$ and $L_{1-\alpha}^N(x)$ were also defined in Lemma 2.4.5. It follows that

$$P[\min(t_{1-\alpha}, q) \leq Y < \max(t_{1-\alpha}, q) | X] = \frac{|t_{1-\alpha} - q|}{L_{1-\alpha}^N(X)}$$

almost surely. Since $|t_{1-\alpha} - q| = (1-\alpha)|\tilde{q} - q| \geq \min(\alpha, 1-\alpha)|\tilde{q} - q|$, we have

$$P[\min(t_{1-\alpha}, q) \leq Y < \max(t_{1-\alpha}, q) | X] \geq \min(\alpha, 1-\alpha) \frac{|\tilde{q} - q|}{L_{1-\alpha}^N(X)}$$

almost surely. Then, we obtain similarly

$$\|\tilde{q} - q\|_p \leq \frac{1}{\min(\alpha, 1-\alpha)} \left\| L_{1-\alpha}^N(X) \right\|_p^{1/2} \|G_{\tilde{q}}(X) - G_q(X)\|_p^{1/2}, \tag{2.4.22}$$

and the result follows as in part (a).

(ii) The result directly follows from Point (i) and Corollary 1.2.6. □

2.5 Proofs of Section 2.3

Let $\gamma^N = \gamma^N(X) = \{\tilde{x}_1, \ldots, \tilde{x}_N\}$ be an optimal grid and denote by $\hat{\gamma}^{N,n} = \hat{\gamma}^{N,n}(X_1, \ldots, X_n) = (\hat{x}_1^{N,n}, \ldots, \hat{x}_N^{N,n})$ the grid provided by the CLVQ algorithm. Throughout this section, we assume the almost sure convergence of the empirical quantization of X to the population one, that is

$$\widehat{X}^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(X) \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{Proj}_{\gamma^N}(X) = \widetilde{X}^N, \tag{2.5.1}$$

which is justified by the discussion in Section 1.2.5.

The proof of Theorem 2.3.1 then requires Lemmas 2.5.1-2.5.2 below.

Lemma 2.5.1. *Let Assumptions (B)(i) and (C) hold. Fix $N \in \mathbb{N}_0$ and $x \in S_X$, and write $\tilde{x} = \mathrm{Proj}_{\gamma^N}(x)$ and $\hat{x}^N = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(x)$. Then, with $\widehat{X}_i^N = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(X_i)$, $i = 1, \ldots, n$, we have*

(i) $\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]} \xrightarrow[n\to\infty]{a.s.} P[\widetilde{X}^N = \tilde{x}]$;

(ii) after possibly reordering the $\tilde{x}_i$'s, $\hat{x}_i^{N,n} \xrightarrow[n\to\infty]{a.s.} \tilde{x}_i$, $i = 1, \ldots, N$ (hence, $\hat{\gamma}^{N,n} \xrightarrow[n\to\infty]{a.s.} \gamma^N$).

Proof. Under (2.5.1), Point (i) was shown in Bally et al. (2005) (see also Pagès (1998)) and Point (ii) only states the a.s. convergence of the supports $\hat{\gamma}^{N,n}$ to γ^N, which is a necessary condition for the corresponding convergence of random vectors in (2.5.1). □

Lemma 2.5.2. *Fix $\alpha \in (0,1)$, $x \in S_X$ and $N \in \mathbb{N}_0$, let $K(\subset \mathbb{R})$ be compact, and define*

$$\widehat{G}_a^{N,n}(\hat{x}^N) := \frac{\frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_i - a)\,\mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}}{\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}}.$$

Then, under Assumptions (A), (B)(i) and (C), (i) $\sup_{a\in K} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})| = o_{\mathrm{P}}(1)$ as $n \to \infty$; (ii) $|\min_{a\in\mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x})| = o_{\mathrm{P}}(1)$ as $n \to \infty$; (iii) $|\widetilde{G}_{\widehat{q}_\alpha^{N,n}(x)} - \widetilde{G}_{\widetilde{q}_\alpha^N(x)}| = o_{\mathrm{P}}(1)$ as $n \to \infty$.

Proof. (i) Since

$$\widetilde{G}_a(\tilde{x}) = \mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}] = \frac{\mathrm{E}[\rho_\alpha(Y-a)\mathbb{I}_{[\widetilde{X}^N=\tilde{x}]}]}{P[\widetilde{X}^N = \tilde{x}]},$$

it is sufficient — in view of Lemma 2.5.1(i) — to prove that, as $n \to \infty$,

$$\sup_{a\in K}\left|\frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_i - a)\,\mathbb{I}_{[\widehat{X}_i^N=\hat{x}^N]} - \mathrm{E}[\rho_\alpha(Y-a)\mathbb{I}_{[\widetilde{X}^N=\tilde{x}]}]\right| = o_P(1).$$

Notice that $\widehat{G}_a^{N,n}(\hat{x}^N)$ can be seen as an empirical version of $\widetilde{G}_a(\tilde{x})$, but taking the grid provided by the algorithm instead of γ^N since the latter is not available in practice. It is natural to consider the following decomposition, where we add and subtract the term corresponding to $\widehat{G}_a^{N,n}(\hat{x}^N)$ with $\hat{\gamma}^{N,n}$ replaced by γ^N:

$$\sup_{a\in K}\left|\frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_i - a)\,\mathbb{I}_{[\widehat{X}_i^N=\hat{x}^N]} - \mathrm{E}[\rho_\alpha(Y-a)\mathbb{I}_{[\widetilde{X}^N=\tilde{x}]}]\right| \le \sup_{a\in K}|T_{a1}| + \sup_{a\in K}|T_{a2}|,$$

with

$$T_{a1} = \frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_i - a)\left(\mathbb{I}_{[\widehat{X}_i^N=\hat{x}^N]} - \mathbb{I}_{[\widetilde{X}_i^N=\tilde{x}]}\right)$$

and

$$T_{a2} = \frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_i - a)\,\mathbb{I}_{[\widetilde{X}_i^N=\tilde{x}]} - \mathrm{E}[\rho_\alpha(Y-a)\mathbb{I}_{[\widetilde{X}^N=\tilde{x}]}],$$

where $\widetilde{X}_i^N = \mathrm{Proj}_{\gamma^N}(X_i)$ for $i = 1,\dots,n$. Using the fact that $m_1(\cdot)$ and $m_2(\cdot)$ are continuous functions defined over the compact set S_X, we obtain that, for any $a \in K$, there exist positive constants C_1 and C_2 such that

$$\rho_\alpha(Y-a) \le \max(\alpha, 1-\alpha)|Y-a| \le \max(\alpha,1-\alpha)(|m_1(X)| + |m_2(X)|\,|\varepsilon| + |a|) \le C_1 + C_2|\varepsilon|,$$

that is in L_1 (recall that ε is assumed to be in L_p, $p=2$), the uniform law of large numbers (see, e.g., Theorem 16(a) in Ferguson, 1996) shows that $\sup_{a\in K}|T_{a2}| = o_P(1)$ as $n \to \infty$.

Turning to T_{a1}, consider the set $\mathcal{I}_n = \{i = 1,\dots,n : \mathbb{I}_{[\widehat{X}_i^N=\hat{x}^N]} \neq \mathbb{I}_{[\widetilde{X}_i^N=\tilde{x}]}\}$ collecting the indices of observations that are projected on the same point as x for γ^N but not for $\hat{\gamma}^{N,n}$ (or

vice versa on the same point as x for $\hat{\gamma}^{N,n}$ but not for γ^N). Theses ones correspond to the terms of the sum in T_{a1} that are not zero. For any $a \in K$, we then have

$$|T_{a1}| \leq \frac{1}{n} \sum_{i \in \mathcal{I}_n} \big|\rho_\alpha(Y_i - a)\big| \leq \frac{\max(\alpha, 1-\alpha)}{n} \sum_{i \in \mathcal{I}_n} (|m_1(X_i)| + |m_2(X_i)|\,|\varepsilon_i| + |a|)$$

$$\leq \frac{\#\mathcal{I}_n}{n} \times \frac{1}{\#\mathcal{I}_n} \sum_{i \in \mathcal{I}_n} (C_1 + C_2\,|\varepsilon_i|) =: S_1 \times S_2.$$

Clearly, Lemma 2.5.1(ii) implies that $\#\mathcal{I}_n/n = o_P(1)$ as $n \to \infty$, while the independence between $\mathcal{I}_n$ (which is measurable with respect to the X_i's) and the ε_i's entails that $\mathrm{E}[S_2] = O(1)$ as $n \to \infty$, so that S_2 is bounded in probability. Consequently, $\sup_{a \in K} |T_{a1}|$ goes to zero in probability as $n \to \infty$. Point (i) of the result follows.

(ii) Fix $\delta > 0$ and $\eta > 0$. Writing $\hat{q} = \widehat{q}_\alpha^{N,n}(x)$ and, as in the previous section, $\tilde{q} = \widetilde{q}_\alpha^N(x)$, first choose n_1 and M large enough to have $|\tilde{q}| \leq M$ and $P[|\hat{q}| > M] < \eta/2$ for any $n \geq n_1$ (Lemma 2.5.1(i) implies that $\hat{q}$ is the sample quantile of a number of Y_i's that increases to infinity, so that $|\hat{q}|$, with arbitrarily large probability for n large, cannot exceed $2 \sup_{x \in S_X} |q_\alpha(x)|$). Then, with $\mathbb{I}_+ = \mathbb{I}_{[\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) \geq \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})]}$, we have

$$\begin{aligned} |\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})|\mathbb{I}_+ &= \big(\widehat{G}_{\hat{q}}^{N,n}(\hat{x}^N) - \widetilde{G}_{\tilde{q}}(\tilde{x})\big)\mathbb{I}_+ \\ &\leq \big(\widehat{G}_{\tilde{q}}^{N,n}(\hat{x}^N) - \widetilde{G}_{\tilde{q}}(\tilde{x})\big)\mathbb{I}_+ \\ &\leq \sup_{a \in [-M,M]} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})|\mathbb{I}_+, \end{aligned} \tag{2.5.2}$$

almost surely. Now, with $\mathbb{I}_- = \mathbb{I}_{[\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) < \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})]}$, we have that, under $|\hat{q}| \leq M$,

$$\begin{aligned} |\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})|\mathbb{I}_- &= \big(\widetilde{G}_{\tilde{q}}(\tilde{x}) - \widehat{G}_{\hat{q}}^{N,n}(\hat{x}^N)\big)\mathbb{I}_- \\ &\leq \big(\widetilde{G}_{\hat{q}}(\tilde{x}) - \widehat{G}_{\hat{q}}^{N,n}(\hat{x}^N)\big)\mathbb{I}_- \\ &\leq \sup_{a \in [-M,M]} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})|\mathbb{I}_-. \end{aligned} \tag{2.5.3}$$

By combining (2.5.2) and (2.5.3), we obtain that, under $|\hat{q}| \leq M$,

$$|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| \leq \sup_{a \in [-M,M]} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})|.$$

Consequently, for any $n \geq n_1$, we obtain

$$\begin{aligned} &P\Big[|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| > \delta\Big] \\ &= P\Big[|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| > \delta, |\hat{q}| \leq M\Big] + P\Big[|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| > \delta, |\hat{q}| > M\Big] \\ &\leq P\Big[\sup_{a \in [-M,M]} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})| > \delta\Big] + \frac{\eta}{2}. \end{aligned}$$

From Point (i) of the lemma, the first term is smaller than $\eta/2$ for any $n \geq n_2$. We conclude that, for any $n \geq n_0 := \max(n_1, n_2)$, we have

$$P\Big[|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| > \delta\Big] < \eta,$$

which shows Point (ii) of the result.

(iii) The proof proceeds in the same way as in (ii) above. First we pick n_1 and M large enough to have $P[|\hat{q}| > M] < \eta/2$ for any $n \geq n_1$, which yields

$$P\Big[|\widetilde{G}_{\tilde{q}}(\tilde{x}) - \widetilde{G}_{\hat{q}}(\tilde{x})| > \delta\Big] \leq P\Big[|\widetilde{G}_{\tilde{q}}(\tilde{x}) - \widetilde{G}_{\hat{q}}(\tilde{x})| > \delta, |\hat{q}| \leq M\Big] + \frac{\eta}{2}. \tag{2.5.4}$$

Now, from the triangular inequality, we obtain

$$\begin{aligned}
&P\Big[|\widetilde{G}_{\tilde{q}}(\tilde{x}) - \widetilde{G}_{\hat{q}}(\tilde{x})| > \delta, |\hat{q}| \leq M\Big] \\
&\quad \leq P\Big[|\widetilde{G}_{\tilde{q}}(\tilde{x}) - \widehat{G}^{N,n}_{\hat{q}}(\hat{x}^N)| > \delta/2, |\hat{q}| \leq M\Big] + P\Big[|\widehat{G}^{N,n}_{\hat{q}}(\hat{x}^N) - \widetilde{G}_{\hat{q}}(\tilde{x})| > \delta/2, |\hat{q}| \leq M\Big] \\
&\quad \leq P\Big[|\min_{a\in\mathbb{R}} \widehat{G}^{N,n}_a(\hat{x}^N) - \min_{a\in\mathbb{R}} \widetilde{G}_a(\tilde{x})| > \delta/2\Big] + P\Big[\sup_{a\in[-M,M]} |\widehat{G}^{N,n}_a(\hat{x}^N) - \widetilde{G}_a(\tilde{x})| > \delta/2\Big],
\end{aligned}$$

which, from Point (i) and Point (ii) of the lemma, can be made arbitrarily small for n large enough. Jointly with (2.5.4), this establishes the result. □

We can now conclude with the proof of Theorem 2.3.1.

Proof of Theorem 2.3.1. Since the function $\rho_\alpha(\cdot)$ is strictly convex, $\widetilde{G}_a(\tilde{x})$ is also strictly convex in a. Its minimum in a (for any fixed x) is therefore unique, and the convergence in probability of $\widetilde{G}_{\hat{q}}(\tilde{x})$ towards $\widetilde{G}_{\tilde{q}}(\tilde{x})$ implies the convergence in probability of the corresponding arguments. □

2.6 Final comments

The emphasis in this chapter was mainly on theoretical aspects. We derived and proved some convergence results. The first step consisted in replacing the covariate X by some quantized version $\widetilde{X}^N$ and to quantify how well the resulting quantization-based conditional quantile $\widetilde{q}^N_\alpha(x)$ approximates the original conditional quantile as N grows. This result had two parts: one with fixed x, one with global X. A second step aimed to define a new estimator of conditional quantile thanks to this approximation. Taking then an empirical version of the approximation allowed to construct a sample quantization-based conditional quantile $\widehat{q}^{N,n}_\alpha(x)$ and the convergence in probability of this estimator to its population version (i.e. $\widetilde{q}^N_\alpha(x)$) was proved. It is important to point out that all the theoretical results were obtained for any dimension d of the covariate X. The required assumptions were quite mild and were essentially assumptions on the link function m between the response variable Y and its covariate X, and assumptions necessary to perform quantization. Moreover, the convergence result for the estimator is only valid if the quantization is based on $p = 2$.

Practical implementation, however, was barely touched in this chapter, and finite-sample performances were not investigated. We only illustrated our estimator on a simulated sample, and improved it thanks to the bootstrap-based estimator $\bar{q}^{N,n}_{\alpha,B}(x)$. The practical investigation will then be the goal of the next chapter where our method will be widely applied on many simulated samples (for different models and sample sizes). As we will explain, we will also compare our

estimator $\bar{q}_{\alpha,B}^{N,n}(x)$ with some well-known competitors: a k- nearest neighbor estimator, two kernel estimators and a spline-based estimator. In particular, the key philosophy of each concept will be confronted.

Another point has not been treated yet: the choice of the tuning parameter N. As above explained, this parameter corresponds to the size of the quantization grid. We derived convergence results with N fixed when n tends to infinity for the estimator (and for N going to infinity for the approximation), but in practice we deal with finite sample with fixed size n. It then seems crucial to choose N with respect to n since we observe in Figure 2.2 that the choice of N has a significant impact on the conditional quantile curves. Indeed, we noticed larger bias when N is too small and larger variability when N is too large. Therefore, it appears of major interest to be able to select an optimal value of N for which a bias-variance compromise will be reached. Moreover, in practical situations, we only have at our disposal the sample and we do not know the theoretical conditional quantiles. For this reason, this N-selection method should be data-driven in order to be feasible in practice. This will be treated in the next chapter.

3

Numerical study of the estimator

Contents

3.1 Introduction

This chapter completes the theoretical study conducted in Chapter 2. As above mentioned, the first aim of this chapter is to provide a criterion for selecting the size N of the quantization grid. The selection method is constructed in two steps. The starting point consists in evaluating the square difference between our estimations and the true conditional quantiles on some grid of points of the support of X. The evaluation is realized for different values of N. As we will see, we observed on the basis of numerous simulations that there exists an optimal value of N that minimizes this error. This N should then be chosen. Section 3.2.1 details this first step.

Unfortunately, the theoretical conditional quantiles are unknown in practice – hence the need

of an estimator, and this selection method is unfeasible on real data sets. For this reason, we explain in Section 3.2.2 how replacing the theoretical conditional quantiles appearing in the first step by some bootstrap version depending only on the data. The optimal values for N selected by both methods are then compared and we will see that the data-driven selection method is really satisfying.

Once an efficient selection method for the tuning parameter is established, our estimation procedure is completely operational. We are then able to compare it with some famous competitors, which is the aim of Section 3.3. We first recall in Section 3.3.1 the competitors that we consider, i.e. the k-nearest neighbor estimator of Bhattacharya and Gangopadhyay (1990), the local linear and local constant estimators of Yu and Jones (1998) and quantile regression splines of Koenker et al. (1994). We then compare these competitors with our bootstrap based estimator $\bar{q}_{\alpha,B}^{N,n}(x)$. This comparison is based on visual inspection of estimated conditional quantile curves in Section 3.3.2 and on empirical integrated square errors in Section 3.3.3, for different models and sample sizes.

The numerical exercises and simulations of these sections focus on the case $d = 1$. We then briefly investigate the case $d = 2$ in Section 3.4 since the definitions and results of the last chapter are valid for any dimension d. These first sections (and a real data application) are the object of the paper Charlier et al. (2015b), published in *Computational Statistics and Data Analysis*.

We then constructed an R package allowing to perform quantization-based quantile regression, called **QuantifQuantile**. The different functions of the package are described in Section 3.5 and some illustrations are provided. This section is the subject of the paper Charlier et al. (2015c), accepted for publication in the R-journal.

After that, our quantization-based estimator is illustrated on some real data examples in Section 3.6. Section 3.6.1 is devoted to a sample corresponding to the concentration of immunoglobulin-G in children aged from 6 months to 6 years (investigated in Charlier et al. (2015b)). Section 3.6.2 investigates a list of several variables collecting in different towns of Gironde, France.

We conclude the chapter with some final comments.

3.2 Selection method of the tuning parameter N

In this section, we first define a natural criterion to select an optimal value for N by considering the square difference between our estimator and the corresponding theoretical conditional quantiles (Section 3.2.1). More precisely, we look at an *integrated square error (ISE)* quantity that is essentially convex in N, which allows to identify an optimal value N_{opt} for N. As we will see, ISEs involve the true unknown conditional quantiles, so that N_{opt} cannot be obtained from the data. This criterion is then infeasible in practice. Therefore, we propose a data-driven selection method for N obtained by replacing the true ISEs with bootstrap versions (Section 3.2.2). We then investigate how close the resulting $\hat{N}_{\text{opt}}$ is to N_{opt}.

3.2.1 Starting idea: infeasible selection of N

For any fixed N, we can consider the square difference between $\widehat{q}_\alpha^{N,n}(x)$ (resp. $\bar{q}_{\alpha,B}^{N,n}(x)$) and $q_\alpha(x)$ in any point x of the support S_X. In other words, the ISEs are given by

$$\int_{S_X} \left(\widehat{q}_\alpha^{N,n}(x) - q_\alpha(x)\right)^2 dx \quad \text{and} \quad \int_{S_X} \left(\bar{q}_{\alpha,B}^{N,n}(x) - q_\alpha(x)\right)^2 dx,$$

associated with the non-bootstrap and bootstrap estimators, respectively. Of course, we only investigate in practice these square differences for a finite set of points of the support S_X, that is these ISEs are approximated by discrete integrals of the form

$$\widehat{\mathrm{ISE}}_{\alpha,J}(N) = \frac{1}{J}\sum_{j=1}^{J} \left(\widehat{q}_\alpha^{N,n}(x_j) - q_\alpha(x_j)\right)^2 \quad \text{and} \quad \bar{\mathrm{ISE}}_{\alpha,B,J}(N) = \frac{1}{J}\sum_{j=1}^{J} \left(\bar{q}_{\alpha,B}^{N,n}(x_j) - q_\alpha(x_j)\right)^2,$$

where $x_1, \ldots, x_J$ are a finite set of J points of interest in S_X. In the sequel, we consider X univariate and we choose them equispaced between the minimum and the maximum values of $X_1, \ldots, X_n$. It is of course natural to consider optimal a value of N that minimizes these ISEs, which leads to considering

$$\widehat{N}_{\alpha,J;\mathrm{opt}} = \arg\min_{N\in\mathbb{N}_0} \widehat{\mathrm{ISE}}_{\alpha,J}(N) \quad \text{and} \quad \bar{N}_{\alpha,B,J;\mathrm{opt}} = \arg\min_{N\in\mathbb{N}_0} \bar{\mathrm{ISE}}_{\alpha,B,J}(N).$$

These optimal N-values may depend on α, which explains the notation. The dependence on J of these ISEs and of optimal N-values is way less crucial than their dependence on α and B; accordingly, we simply write $\widehat{\mathrm{ISE}}_\alpha(N)$, $\bar{\mathrm{ISE}}_{\alpha,B}(N)$, $\widehat{N}_{\alpha;\mathrm{opt}}$ and $\bar{N}_{\alpha,B;\mathrm{opt}}$ in the sequel.

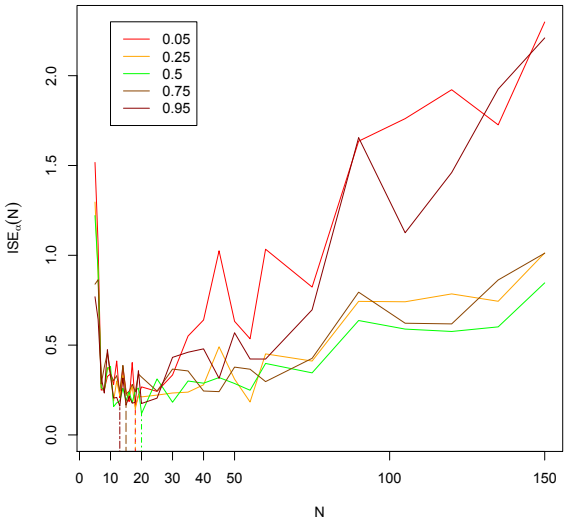
Looking at each $N \in \mathbb{N}_0$ is computationally heavy since our estimation procedure has to run separately for each value of N. In practice, we then look only at a finite subset of values for N chosen according to the sample size n.

To illustrate these definitions, we simulated random samples of size $n = 300$ according to the models

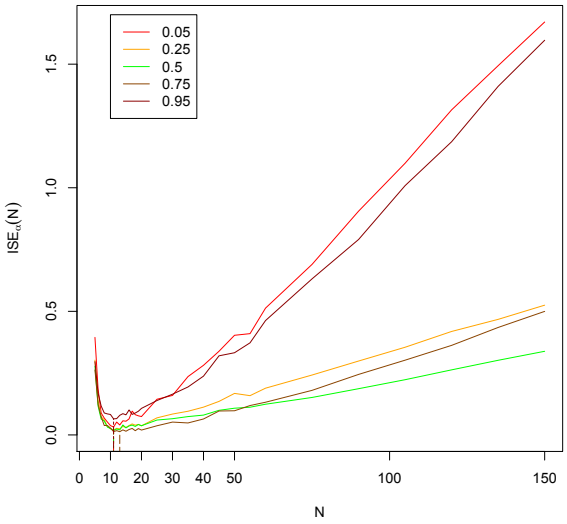
$(\mathcal{M}1)$ $Y = \frac{1}{5}X_1^3 + \varepsilon$,
$(\mathcal{M}2)$ $Y = f(X_2) + \varepsilon'$,

where $X_1 = 6Z_1 - 3$ (with $Z_1 \sim \mathrm{Beta}(0.3, 0.3)$), $X_2 = 3Z_2 - 1.5$ (with $Z_2 \sim \mathrm{Beta}(2,2)$), $\varepsilon \sim \mathcal{N}(0,1)$, and $\varepsilon' \sim \chi^2_2$ are mutually independent. Denoting the standard normal density as φ, the link function f is defined as $f(x) = \frac{1}{2}\varphi(x) + 10\sum_{\ell=1}^{3}(0.2)^\ell \varphi(10(x - \frac{\ell}{2} + 1))$, a choice that is inspired by Marron and Wand (1992). Obviously, $q_\alpha(x) = \frac{x^3}{5} + \Phi^{-1}(\alpha)$ for $(\mathcal{M}1)$ and $q_\alpha(x) = f(x) + \Psi_2^{-1}(\alpha)$ for $(\mathcal{M}2)$, where Φ and Ψ_2 denote the cumulative distribution functions of the $\mathcal{N}(0,1)$ distribution and of the χ^2_2 distribution, respectively. We evaluated the ISEs above with $J = 300$ points.

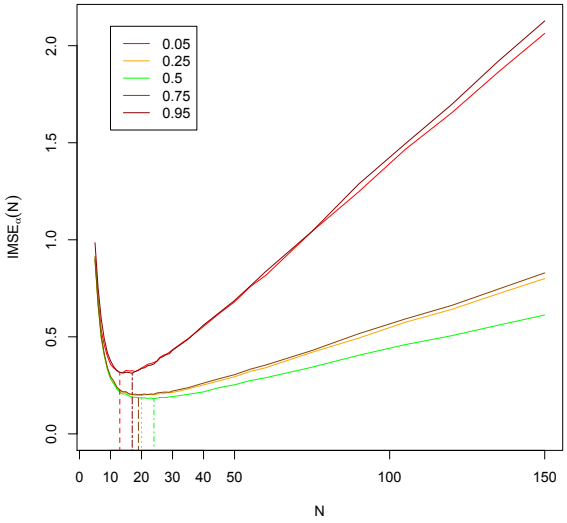
Figure 3.1 focuses on Model $(\mathcal{M}1)$. Its top panels plot the graphs of $N \mapsto \widehat{\mathrm{ISE}}_\alpha(N)$ and $N \mapsto \bar{\mathrm{ISE}}_{\alpha,B}(N)$ (with $B = 50$) for a single sample, while the bottom panels plot the corresponding graphs averaged (pointwise in N) over $m = 500$ independent samples. For the sake of brevity, Figure 3.2, that is related to Model $(\mathcal{M}2)$, only reports the corresponding plots for bootstrapped estimators.



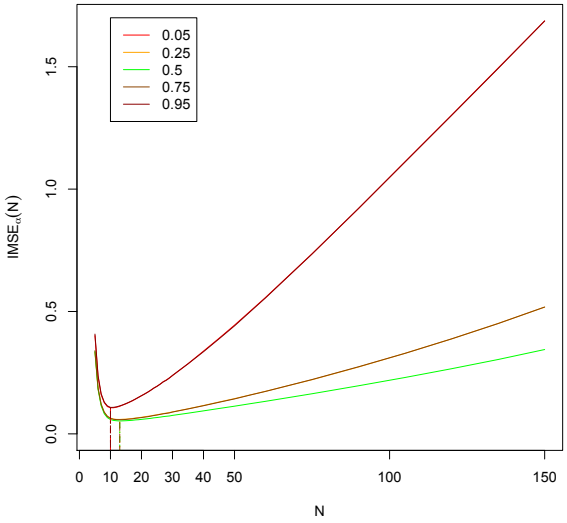
(a) $\widehat{\mathrm{ISE}}_\alpha(N)$, $m = 1$



(b) $\bar{\mathrm{ISE}}_{\alpha,B}(N)$, $m = 1$

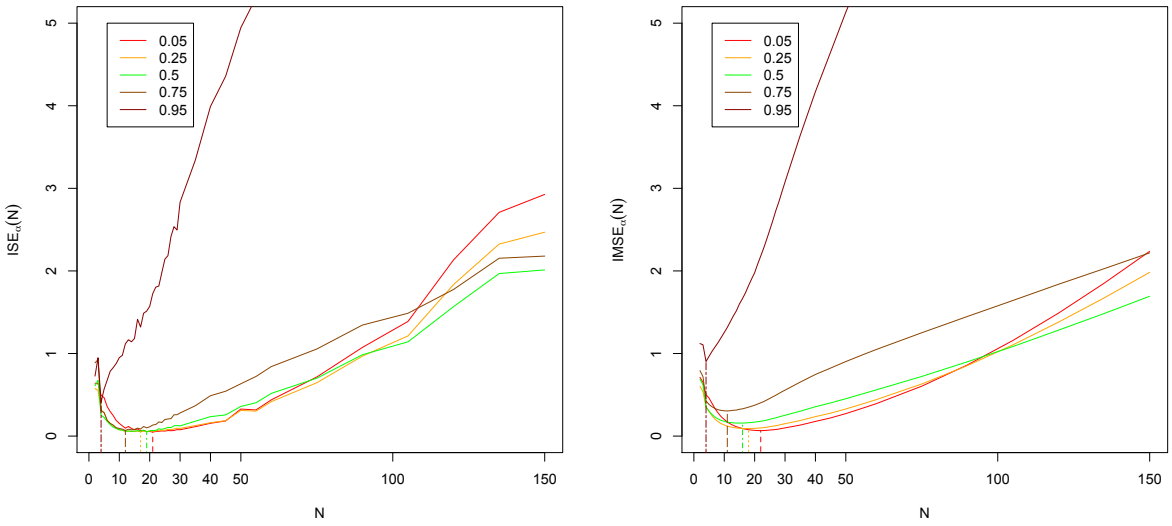


(c) $\widehat{\mathrm{ISE}}_\alpha(N)$, $m = 500$



(d) $\bar{\mathrm{ISE}}_{\alpha,B}(N)$, $m = 500$

Figure 3.1 – Plots of the mappings $N \mapsto \widehat{\mathrm{ISE}}_\alpha(N)$ (top left) and $N \mapsto \bar{\mathrm{ISE}}_{\alpha,B}(N)$ with $B = 50$ (top right) for a random sample of size $n = 300$ from Model ($\mathcal{M}1$). The bottom panels report the corresponding plots obtained by averaging these mappings over $m = 500$ mutually independent replications (all mappings are actually only evaluated at $N = 5, 6, 7, \ldots, 29, 30, 35, 40, \ldots, 150$).



(a) $\overset{-}{\text{ISE}}_{\alpha,B}(N)$, $m = 1$ (b) $\overset{-}{\text{ISE}}_{\alpha,B}(N)$, $m = 500$

Figure 3.2 – The same plots as in the right panels of Figure 3.1, but for Model ($\mathcal{M}2$).

Figures 3.1–3.2 show that ISE curves are indeed essentially convex in N: while random variability prevents perfect convexity for a single sample, the smoothest case based on 500 replications with $\bar{q}^{N,n}_{\alpha,B}$ leads to strictly convex IMSE curves. Besides, these curves are quite flat in a neighborhood of $\hat{N}_{\alpha;\text{opt}}$ or $\overset{-}{N}_{\alpha,B;\text{opt}}$ (except for $\alpha = 0.95$), so that picking a value of N that is close to (but not exactly equal to) the optimal N-value will not have an important impact on the resulting ISE values. It seems that, as expected, $\hat{N}_{1-\alpha;\text{opt}} = \hat{N}_{\alpha;\text{opt}}$ and $\overset{-}{N}_{1-\alpha,B;\text{opt}} = \overset{-}{N}_{\alpha,B;\text{opt}}$ when the error ε has a symmetric distribution (observed when looking at the smoothest case $\hat{\text{ISE}}_\alpha(N)$ with $m = 500$). In contrast, asymmetric errors (Figure 3.2) lead to optimal values of N that are not symmetric in α and depend more strongly of α. We notice in particular that the optimal value for N increases as α decreases. This is due to the dispersion of the data in a model with a chi-squared error term.

To conclude this section, we stress that the existence of such an optimal value of N does not contradict the theoretical results from Chapter 2. One might indeed have guessed from Theorem 2.2.1 that the ISE functions above would be monotone decreasing (rather than convex) in N. The result in Theorem 2.2.1, however, (i) involves the population quantile approximation $\widetilde{q}^N_\alpha(x)$ and not its sample counterpart $\widehat{q}^{N,n}_\alpha(x)$, and (ii) requires that the projection, for any N, is performed on an optimal quantization grid. As pointed out earlier, the CLVQ algorithm provides a good approximation of an optimal grid only if the number of iterations, that is equal to the sample size n, is large compared to N. Consequently, in the present setup where the sample size n is fixed, increasingly large values of N will result into CLVQ grids that are less and less optimal, which explains the increasing ISE values for such values of N.

3.2.2 Data-driven selection of N

In practice, observations $(X_1, Y_1), \dots, (X_n, Y_n)$ are available, but the population conditional quantile function $q_\alpha(\cdot)$ is unknown. This is the reason why we aim to estimate them and the criterion of Section 3.2.1 does not make sense since it uses the unknown quantity we want to estimate. It is therefore impossible to obtain optimal N-values by minimizing $\widehat{\mathrm{ISE}}_\alpha(N)$ and $\widehat{\mathrm{ISE}}_{\alpha,B}(N)$ as above. We then propose the following approach.

Consider again the grid $\{x_1 = X_{(1)}, x_2, \dots, x_{J-1}, x_J = X_{(n)}\}$ of equispaced points between the minimal and maximal observed X-values. We replace the theoretical conditional quantiles by some estimates using bootstrap, in the spirit of the construction of $\bar{q}^{N,n}_{\alpha,B}(x)$. This procedure works as follows.

Generating $\widetilde{B}$ grids

For some integer $\widetilde{B}$, we first generate $\widetilde{B}$ samples of size n with replacement from the initial sample $X_1, \dots, X_n$. We write here the newly generated samples $(\xi^t_{B+\tilde{b}})_t$, for $\tilde{b} = 1, \dots, \widetilde{B}$, to distinguish them from the corresponding samples obtained in the construction of $\bar{q}^{N,n}_{\alpha,B}(x)$. We also generate initial grids $\hat{\gamma}^{N,0}_{B+\tilde{b}}$, $\tilde{b} = 1, \dots, \widetilde{B}$, as before, by sampling randomly among the corresponding $(\xi^t_{B+\tilde{b}})_t$ under the constraint that the N values are pairwise distinct. We then perform $\widetilde{B}$ times CLVQ with iterations based on $\xi^t_{B+\tilde{b}}$, $t = 1, \dots, n$ and with initial grid $\hat{\gamma}^{N,0}_{B+\tilde{b}}$. This provides $\widetilde{B}$ optimal grids $\hat{\gamma}^{N,n}_{B+\tilde{b}}$, $\tilde{b} = 1, \dots, \widetilde{B}$.

Bootstrap based selection method for N

Each of these grids is now used to estimate conditional quantiles and to replace them in the infeasible criterion of Section 3.2.1. Working again with the original sample (X_i, Y_i), $i = 1, \dots, n$, we project the X-part onto the grids $\hat{\gamma}^{N,n}_{B+\tilde{b}}$, $\tilde{b} = 1, \dots, \widetilde{B}$. Therefore, for all $j = 1, \dots, J$, (2.3.1) provides $\widetilde{B}$ estimations, denoted $\hat{q}^{(B+\tilde{b})}_\alpha(x_j) = \hat{q}^{(B+\tilde{b}),N,n}_\alpha(x_j)$. This allows to consider the square difference between $\bar{q}^{N,n}_{\alpha,B}(x_j)$ and $\hat{q}^{(B+\tilde{b})}_\alpha(x_j)$, $\tilde{b} = 1, \dots, \widetilde{B}$ instead of the one between $\bar{q}^{N,n}_{\alpha,B}(x_j)$ and $q_\alpha(x_j)$. We then take the mean of these $\widetilde{B}$ differences and we define

$$\widehat{\mathrm{ISE}}_{\alpha,\widetilde{B},J}(N) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\widetilde{B}}\sum_{\tilde{b}=1}^{\widetilde{B}}\left(\widehat{q}^{N,n}_\alpha(x_j) - \hat{q}^{(B+\tilde{b})}_\alpha(x_j)\right)^2\right) \tag{3.2.1}$$

and

$$\widehat{\mathrm{ISE}}_{\alpha,B,\widetilde{B},J}(N) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\widetilde{B}}\sum_{\tilde{b}=1}^{\widetilde{B}}\left(\bar{q}^{N,n}_\alpha(x_j) - \hat{q}^{(B+\tilde{b})}_\alpha(x_j)\right)^2\right). \tag{3.2.2}$$

As in their infeasible counterparts, we will not stress the dependence on J in these samples ISEs, nor the dependence on $\widetilde{B}$, that we choose equal to 30 throughout.

It is important to notice that evaluating $\widehat{\mathrm{ISE}}_{\alpha,B}(N)$ thus requires generating $B+\widetilde{B}$ bootstrap samples of size n: B for the construction of $\bar{q}^{N,n}_\alpha(x_j)$, and $\widetilde{B}$ to obtain $\hat{q}^{(B+\tilde{b})}_\alpha(x_j)$, $\tilde{b} = 1, \dots, \widetilde{B}$. The construction of $\widehat{\mathrm{ISE}}_{\alpha,B}(N)$ (jointly with the one of $\bar{q}^{N,n}_{\alpha,B}(x)$) is illustrated in Figure 3.3. These

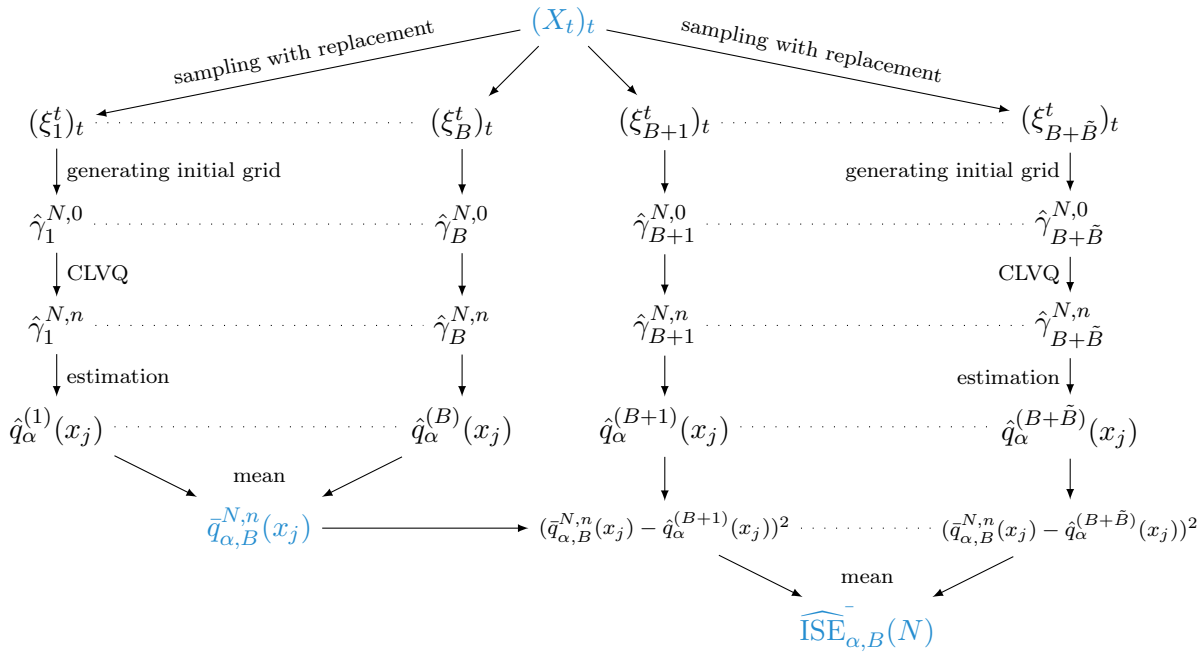


Figure 3.3 – Construction step by step of the sample ISE, $\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$.

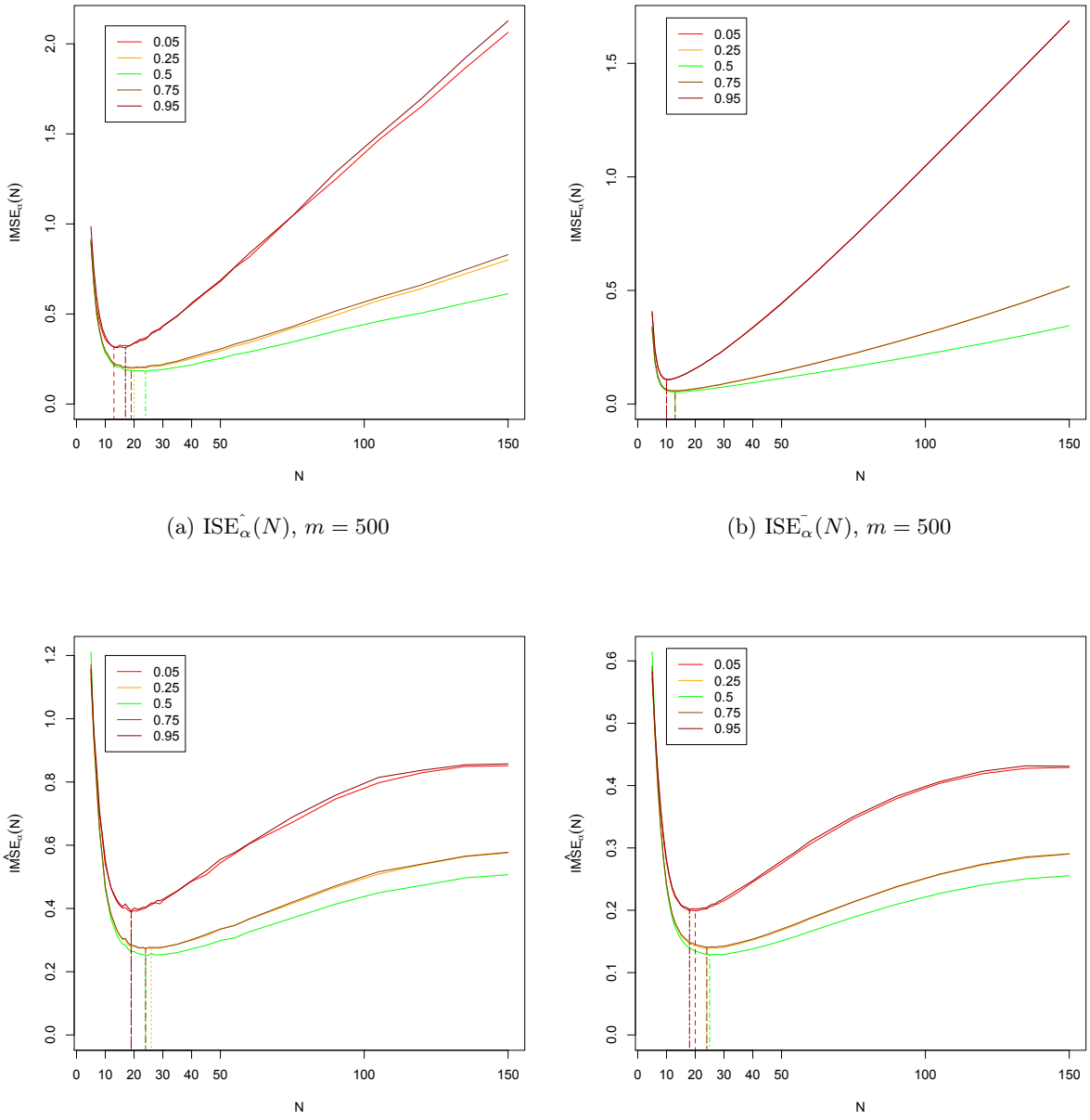
sample ISEs are to be minimized in N. Since not all values of N can be considered in practice, we rather consider

$$\hat{N}^{\hat{}}_{\alpha;\mathrm{opt}} = \arg\min_{N\in\mathcal{N}} \widehat{\mathrm{ISE}}^{\hat{}}_{\alpha}(N) \quad \text{and} \quad \hat{N}^{-}_{\alpha,B;\mathrm{opt}} = \arg\min_{N\in\mathcal{N}} \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N), \tag{3.2.3}$$

where the cardinality of $\mathcal{N}(\subset \mathbb{N}_0)$ is finite (and may be chosen as a function of n).

We now use these sample ISE to construct some graphs analogous to Figure 3.1. Our aim is to check if the curves of $\widehat{\mathrm{ISE}}^{\hat{}}_{\alpha}(N)$ and $\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ are convex as their theoretical versions and if they are minimized for the same value of N. Figure 3.4 plots the mappings $N \mapsto \widehat{\mathrm{ISE}}^{\hat{}}_{\alpha}(N)$ and $N \mapsto \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ and, for the sake of comparison, the (infeasible) mappings $N \mapsto \mathrm{ISE}^{\hat{}}_{\alpha}(N)$ and $N \mapsto \mathrm{ISE}^{-}_{\alpha,B}(N)$, in the setup of Model $(\mathcal{M}1)$ with sample size $n = 300$ (more precisely, the average of the corresponding plots, over 500 mutually independent replications, are plotted there). It is seen that, for large N, the (averaged) sample ISE functions quite poorly estimate their theoretical versions. Our primary interest, however, rather is in the agreement of the corresponding argmins, which are associated with the feasible or infeasible optimal N-values. In that respect, the results show that the values of N minimizing the sample ISE functions tend to over-estimate the optimal N-values, with a bias of about 10. Note that this bias is not alarming since Figure 3.1 shows that virtually the same performances in terms of ISE are obtained with or without bias (particularly so for intermediate values of α). Finally, we point out that, when considering a single sample (instead of 500 replications), the estimated optimal N-value stays in the same interval (the plots are not provided here).

To sum up, the sample ISE functions introduced above allow us to obtain a data-driven value of N that, for all practical purposes, is as satisfactory as the infeasible optimal N-value based on the original ISE functions.



(a) $\widehat{\mathrm{ISE}}_{\alpha}(N)$, $m = 500$

(b) $\overline{\mathrm{ISE}}_{\alpha}(N)$, $m = 500$

(c) $\widehat{\widehat{\mathrm{ISE}}}_{\alpha}(N)$, $m = 500$

(d) $\widehat{\overline{\mathrm{ISE}}}_{\alpha,B}(N)$, $m = 500$

Figure 3.4 – Plots of the mappings $N \mapsto \widehat{\mathrm{ISE}}_{\alpha}(N)$ (top left), $N \mapsto \overline{\mathrm{ISE}}_{\alpha,B}(N)$ with $B = 50$ (top right), and of the sample mappings $N \mapsto \widehat{\widehat{\mathrm{ISE}}}_{\alpha}(N)$ with $\widetilde{B} = 30$ (bottom left), $N \mapsto \widehat{\overline{\mathrm{ISE}}}_{\alpha,B}(N)$ with $\widetilde{B} = 30$ and $B = 50$ (bottom right), averaged over 500 mutually independent replications of Model ($\mathcal{M}1$) with sample size $n = 300$.

3.3 Comparison with some famous competitors

Thanks to Chapter 2 and Section 3.2, we have now developed a conditional quantile estimation method that 1) possesses interesting convergence results and 2) provides an efficient data-driven selection method for the tuning parameter. Our method is then entirely functional. It is then natural to wonder now if our estimator is competitive with respect to some well-known other conditional quantile estimators. Let us first define the competitors considered.

3.3.1 The competitors considered

The first competing estimator we consider is associated with quantile regression *spline methods* (Koenker et al., 1994). More precisely, this estimator, $\widehat{q}_\alpha^{\text{sp}}$, say, is defined as

$$\widehat{q}_\alpha^{\text{sp}} = \widehat{q}_{\alpha,\lambda}^{\text{sp}} = \arg\min_{g\in\mathcal{G}} \left\{ \sum_{i=1}^{n} \rho_\alpha(Y_i - g(X_i)) + \lambda P(g) \right\},$$

where λ is a nonnegative real number, $\mathcal{G}$ is an appropriately chosen space of functions, and the penalty $P(g) = \int_0^1 |g''(x)|\,dx$ is the total variation of the first derivative of g. It can be showed that $\widehat{q}_\alpha^{\text{sp}}$ is always piecewise polynomial. As usual in penalization methods, λ governs the trade-off between fidelity to the sample and smoothness of the resulting estimator. Several methods to select an appropriate value of λ are available in the literature. Below, we will use the AIC criterion, which selects the value $\lambda = \hat{\lambda}_{\alpha;\text{opt}}$ minimizing

$$\text{AIC}(\lambda) = \log\left[\frac{1}{n}\sum_i \rho_\alpha\big(Y_i - \widehat{q}_{\alpha,\lambda}^{\text{sp}}(X_i)\big)\right] + \frac{p_\lambda}{n},$$

where p_λ is the number of knots defining the spline. The reader can refer to Koenker et al. (1994) and Koenker and Mizera (2004) for more details on how to choose λ (or $\mathcal{G}$). When implementing this method, we used the `rqss` function from the R package `quantreg` (Koenker, 2015) to compute $\widehat{q}_\alpha^{\text{sp}}$ and performed the λ-selection via the R function `AIC` (the package `quantreg` does not propose an automatic λ-selection procedure). Notice that the current implementation of `quantreg` provides piecewise linear estimated curves.

Another competitor is the k-nearest neighbor estimator, mentioned in the sequel as the *kNN estimator*, which is defined as follows (Bhattacharya and Gangopadhyay, 1990). Let k be any integer between 1 and n. The idea is to select the k observations for which the X_i is the nearest of the considered x, and to take the α-quantile of the associated Y_i. More precisely, we permute the observations $\{(X_i, Y_i)\}_{i=1,\dots,n}$ in such a way that $|X_i - x|$ is increasing in i. After this permutation, we define the kNN estimator, denoted $\hat{q}_{\alpha;k\text{NN}}^k(x)$, as the α-quantile of $Y_1, \dots, Y_k$.

In many conditional quantile estimation methods, the crucial point is to select, for x fixed, the observations (X_i, Y_i) that will be taken into account in the calculation of the quantile. Of course, it is felt that the X_i should be quite close to the x, but the definition of this closeness is not trivial. Actually, any method uses a different choice. For the kNN estimator, the choice is to use the k observations for which X_i is the closest to x. Therefore, if many observations are close to x, it corresponds to a small bandwidth while if there are few observations around x, this is a large bandwidth. The bandwidth is then in a way adaptive with x.

As the size N of our quantization grid, the number k of neighbors is the tuning parameter of this method and it has to be selected. There does not seem to exist an efficient and well-accepted data-driven method to choose k in the literature. For this reason, we select an optimal value for k, denoted $k_{\alpha;\mathrm{opt}}$ as follows:

$$k_{\alpha;\mathrm{opt}} = \arg\min_{k\in\mathcal{K}} \frac{1}{J}\sum_{i=1}^{J}\big(\hat{q}^{k}_{\alpha;k\mathrm{NN}}(x_j) - q_\alpha(x_j)\big)^2,$$

where $\{x_1,\ldots,x_J\}$ is chosen as in Section 3.2 and $\mathcal{K}$ denotes some set of possible values for k. In other words, this selection method consists in evaluating the square difference between the estimation and the true value for a grid of values for x, and to choose the k that minimizes this difference. Clearly, k cannot be chosen this way in practical situations since the theoretical conditional quantiles $q_\alpha(x)$ are typically unknown. In the sequel, we throughout considered

$$\hat{q}_{\alpha;k\mathrm{NN}}(x) = \hat{q}^{k_{\alpha;\mathrm{opt}}}_{\alpha;k\mathrm{NN}}(x),$$

the kNN estimator for the choice $k = k_{\alpha;\mathrm{opt}}$. To the best of our knowledge, no R package allows to compute $\widehat{q}_{\alpha;k\mathrm{NN}}(x)$ and we therefore wrote our own implementation to conduct the simulations below.

The last estimators we consider are the *kernel (local linear or local constant) estimators* introduced in Yu and Jones (1998). The local linear estimator is of the form $\widehat{q}^{\mathrm{YJ}}_\alpha(x) = \hat{a}$, with

$$(\hat{a},\hat{b}) = \arg\min_{(a,b)\in\mathbb{R}\times\mathbb{R}} \sum_{i=1}^{n}\rho_\alpha\big(Y_i - a - b(X_i - x)\big)K\left(\frac{X_i - x}{h}\right),$$

where K is a kernel function and h is the bandwidth. In the sequel, K will be the standard normal density, and we choose

$$h = \hat{h}_{\alpha;\mathrm{opt}} = \frac{\alpha(1-\alpha)}{\big(\varphi(\Phi^{-1}(\alpha))\big)^2}\,h_{\mathrm{mean}},$$

where φ and Φ are respectively the standard normal density and distribution functions, and where h_{mean} is the optimal choice of h for mean regression, selected through cross-validation; see Yu and Jones (1998). The local constant version of this estimator is defined as $\widehat{q}^{\mathrm{YJc}}_\alpha(x) = \hat{a}$, with

$$\hat{a} = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n}\rho_\alpha\big(Y_i - a\big)K\left(\frac{X_i - x}{h}\right),$$

where K and h will throughout be chosen as for the local linear estimator. Here, the X_i's taken into account are chosen thanks to a bandwidth that appears in the weight function K. As for $\widehat{q}_{\alpha;k\mathrm{NN}}(x)$, the simulations below are based on our own R implementation of $\widehat{q}^{\mathrm{YJ}}_\alpha(x)$ and $\widehat{q}^{\mathrm{YJc}}_\alpha(x)$.

3.3.2 Comparison of estimated quantile curves

In this section, we now compare our proposed quantization-based estimators with their competitors described above. Since we saw in Section 2.3 (and particularly in Figure 2.2) that the

bootstrapped estimators $\bar{q}^{N,n}_{\alpha,B}(x)$ are to be favored over their original versions $\widehat{q}^{N,n}_{\alpha}(x)$, we restrict to $\bar{q}^{N,n}_{\alpha,B}(x)$ below, with the corresponding data-driven value of N, namely $\hat{N}^{-}_{\alpha,B;\mathrm{opt}}$, that was proposed in Section 3.2. In this section, we also do not consider the local constant estimator $\widehat{q}^{\mathrm{YJc}}_{\alpha}(x)$ since the results in Section 3.3.3 below show that it is usually outperformed by its local linear version in terms of ISEs. We start the comparison by investigating estimated quantile curves computed from n independent observations generated according to the models

$(\mathcal{M}1)$ $Y = \frac{1}{5}X_1^3 + \varepsilon$,
$(\mathcal{M}2)$ $Y = f(X_2) + \varepsilon'$,
$(\mathcal{M}3)$ $Y = \sin(X_3) + \big(0.5 + 1.5\sin^2(\frac{\pi}{2}X_3)\big)\varepsilon$,

where $X_1 = 6Z_1 - 3$ (with $Z_1 \sim \mathrm{Beta}(0.3, 0.3)$), $X_2 = 3Z_2 - 1.5$ (with $Z_2 \sim \mathrm{Beta}(2,2)$), $X_3 = 6Z_2 - 3$, $\varepsilon \sim \mathcal{N}(0,1)$, and $\varepsilon' \sim \chi^2_2$ are mutually independent. The link function f is the same as in Section 3.2. For $n = 300$, the resulting estimated quantile curves are plotted in Figures 3.5-3.7 for Models $(\mathcal{M}1)$-$(\mathcal{M}3)$, respectively. For the sake of comparison, Figure 3.8 provides the results for Model $(\mathcal{M}1)$ and $n = 1,000$.

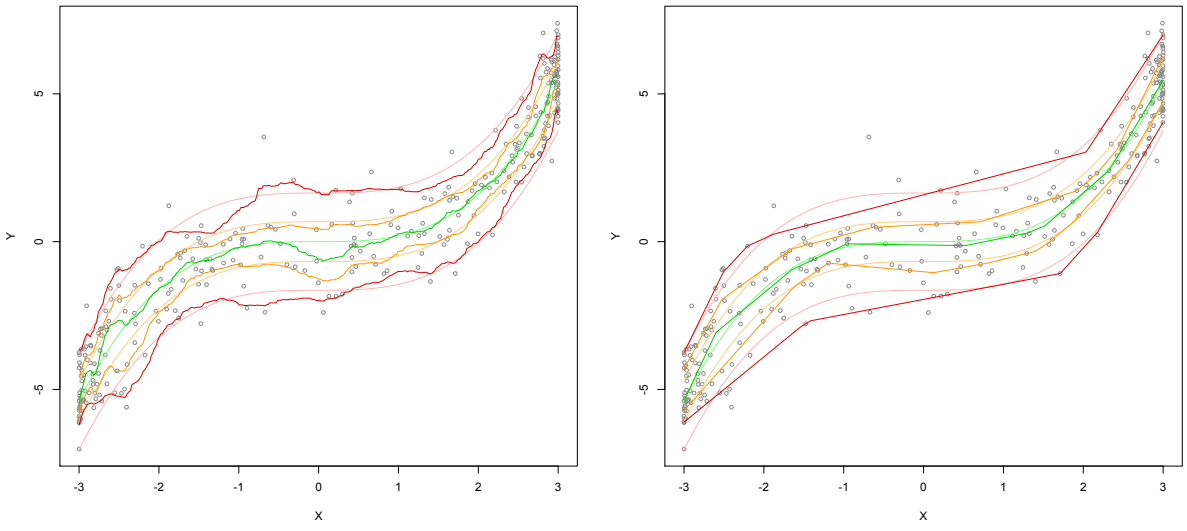
The quantization-based quantile curves are smooth and adapt well to the polynomial or more complex nature of the link function at hand. In contrast, while the piecewise linear curves obtained from the spline estimator are beneficial for polynomial link functions (see Figure 3.5), they hurt for more complex link functions (Figure 3.6, e.g. shows that the spline-based curves miss some of the underlying bumps). The curves resulting from kNN estimation are relatively close to the theoretical ones in each model, but show some peaks or constant pieces, hence are less pleasant from a visual point of view. Eventually, the curves associated to $\widehat{q}^{\mathrm{YJ}}_{\alpha}(x)$ lack smoothness for medium values of the covariate in Figure 3.5 (further numerical experiments, not reported here, reveal that this follows from the non-uniform covariate distribution). Also, the local linear estimator $\widehat{q}^{\mathrm{YJ}}_{\alpha}(x)$ does not catch correctly the link function in Figure 3.7.

The empirical quantile curves obtained through quantization in Figures 3.5-3.7 do not present crossings, which is in line with what occurs at the population level. In principle, there may be such crossings, though, since, in order to achieve highest possible efficiency for any given α, we allowed the optimal data-driven N-value to depend on α. An easy way to guarantee that no quantile crossings occur is to select (possibly at the expense of efficiency) an optimal N-value that does not depend on α, as follows:

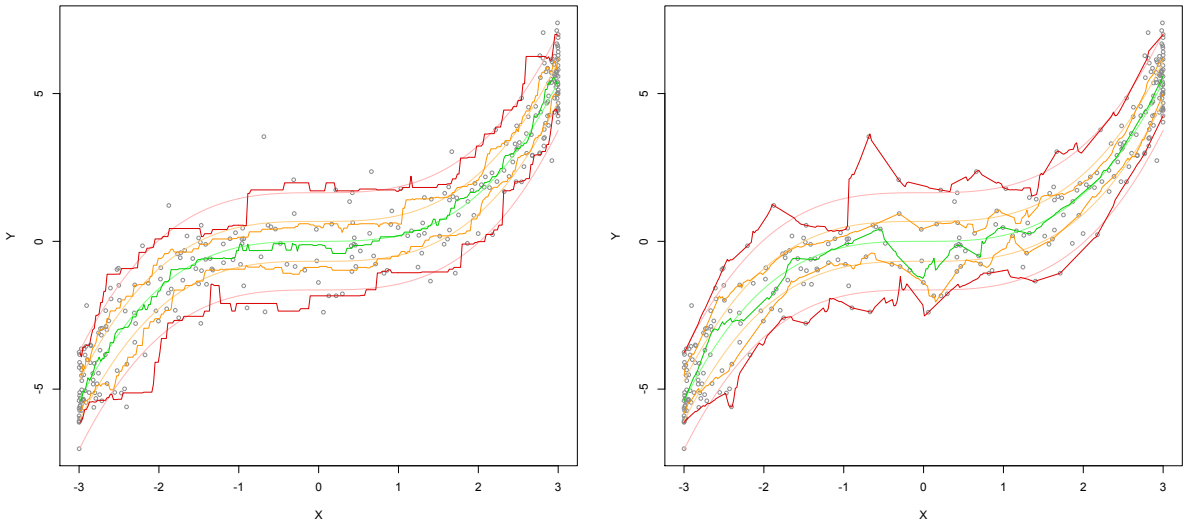
$$\hat{N}^{-}_{\mathrm{opt}} = \arg\min_{N\in\mathcal{N}} \widehat{\mathrm{ISE}}^{-}(N), \quad \text{with} \quad \widehat{\mathrm{ISE}}^{-}(N) = \mathrm{AVE}_{\alpha}\big[\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)\big], \tag{3.3.1}$$

where AVE_{α} denotes the average over all the alpha values considered.

Let us conclude this section with some observations on the tuning parameters gathered in Table 3.1. If it provides more details on the values used to construct the different figures, it also allows to notice some coherence between the different methods. Indeed, some of them have to select the observations that will be taken into account when estimating $q_\alpha(x)$ with respect to x. It would then be natural that if one method selects a large number of points, other methods

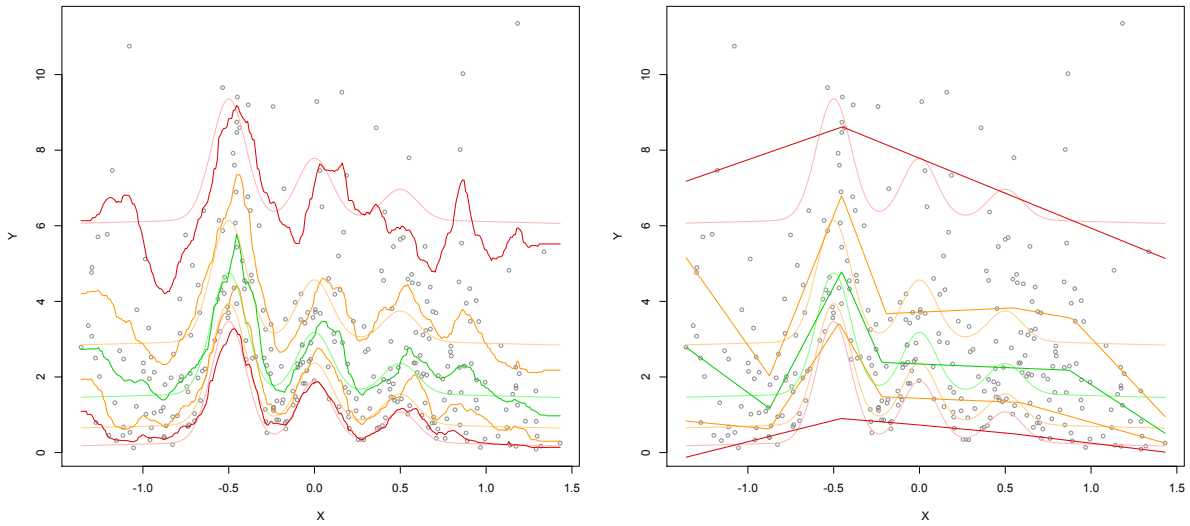


(a) $\widehat{q}^{N,n}_{\alpha,B}$ (b) $\widehat{q}^{\mathrm{sp}}_{\alpha}$

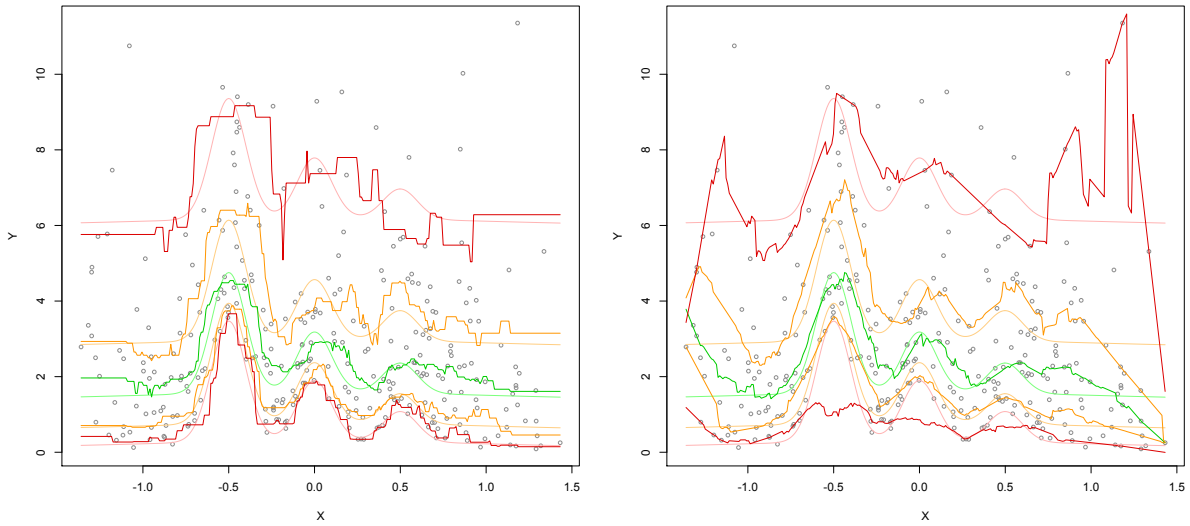


(c) $\widehat{q}_{\alpha;k\mathrm{NN}}$ (d) $\widehat{q}^{\mathrm{YJ}}_{\alpha}$

Figure 3.5 – For a random sample of size $n = 300$ generated according to Model ($\mathcal{M}1$), conditional quantile curves obtained from quantization-based estimation (upper left), spline methods (upper right), nearest-neighbor estimation (lower left), and local linear kernel methods (lower right). In all cases, the quantile levels considered are $\alpha = 0.05$ (red), 0.25 (orange), 0.5 (green), 0.75 (orange), and 0.95 (red). The lighter curves correspond to population conditional quantiles.

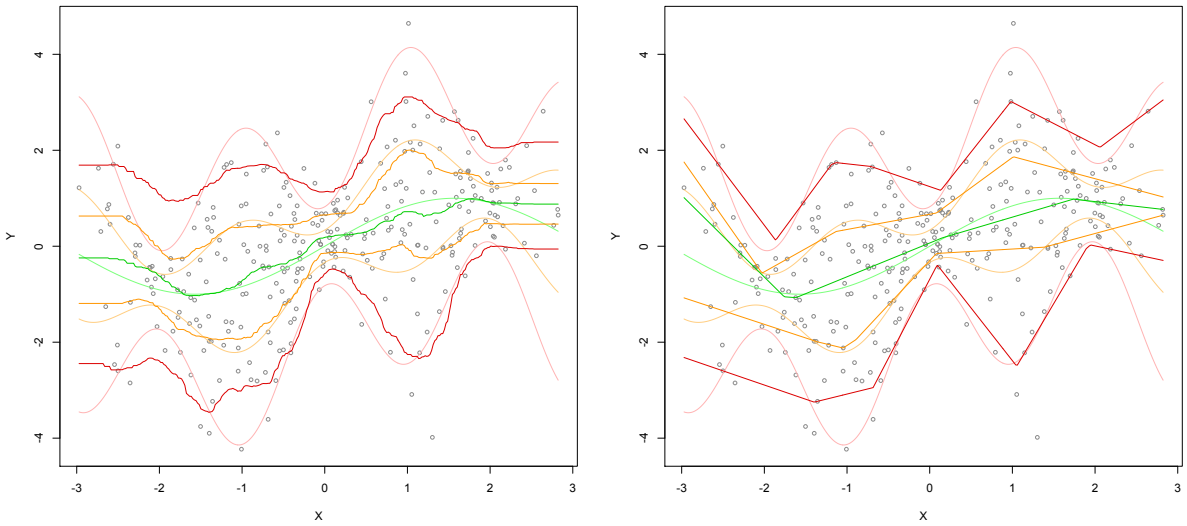


(a) $\bar{q}^{N,n}_{\alpha,B}$ (b) $\widehat{q}^{\mathrm{sp}}_{\alpha}$

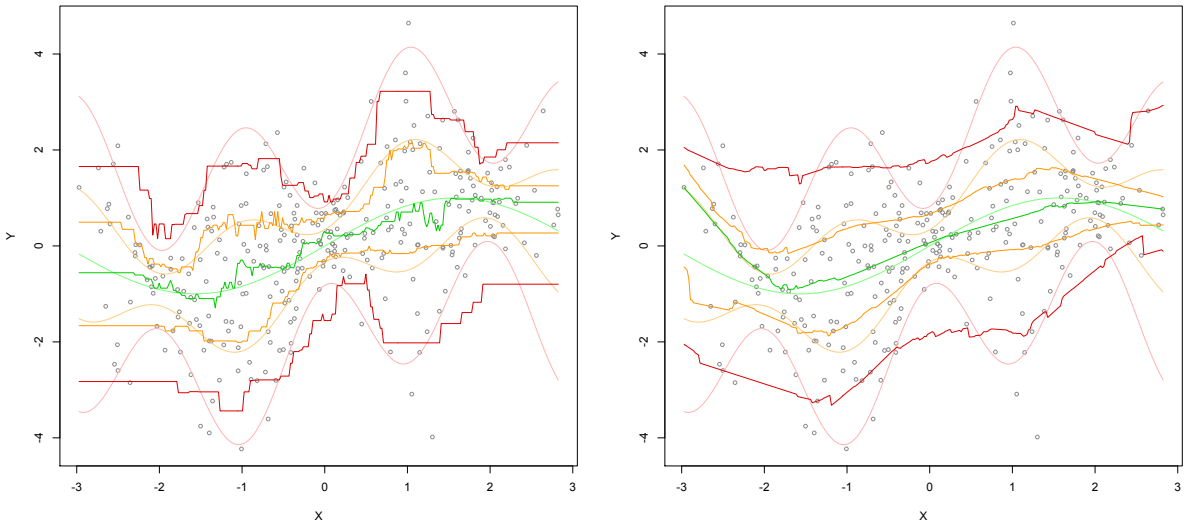


(c) $\widehat{q}_{\alpha;k\mathrm{NN}}$ (d) $\widehat{q}^{\mathrm{YJ}}_{\alpha}$

Figure 3.6 – The same plots as in Figure 3.5, but for Model ($\mathcal{M}2$) and $n = 300$.

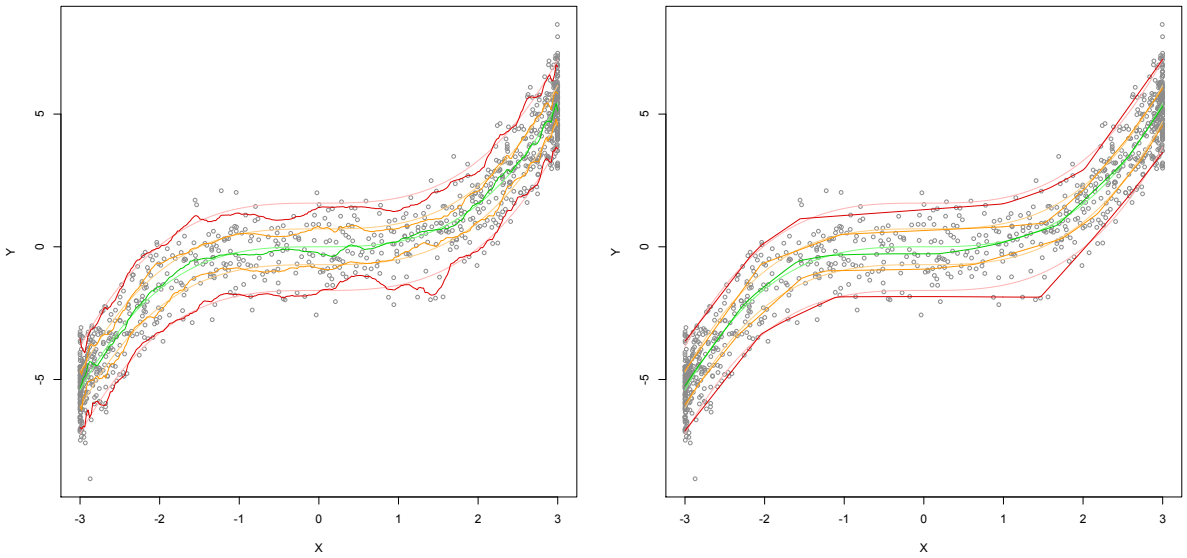


(a) $\bar{q}^{N,n}_{\alpha,B}$ (b) $\widehat{q}^{\mathrm{sp}}_{\alpha}$

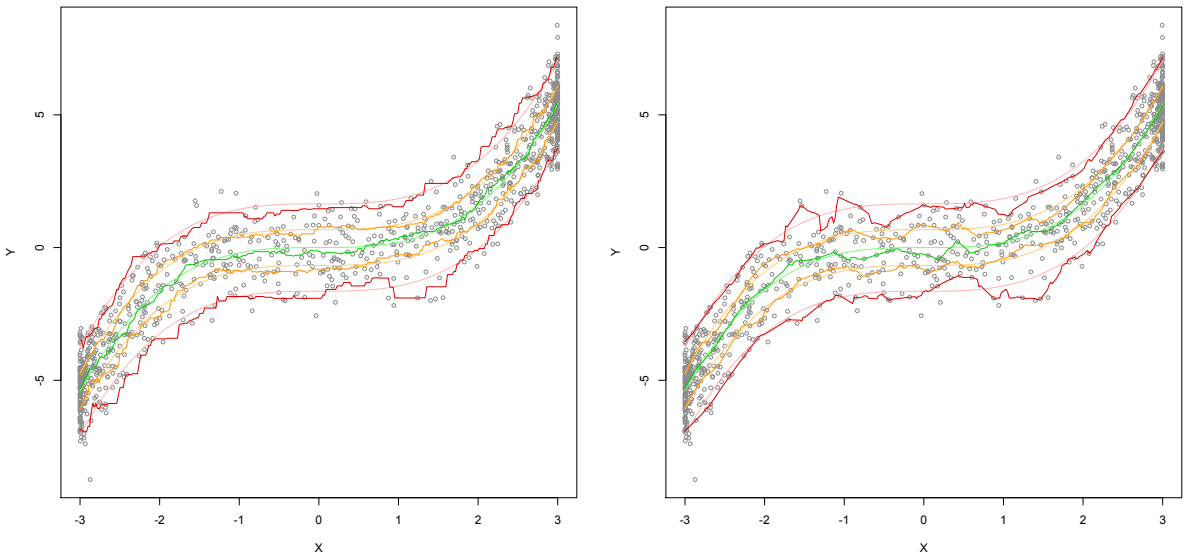


(c) $\widehat{q}_{\alpha;k\mathrm{NN}}$ (d) $\widehat{q}^{\mathrm{YJ}}_{\alpha}$

Figure 3.7 – The same plots as in Figures 3.5-3.6, but for Model ($\mathcal{M}3$) and $n = 300$.



(a) $\bar{q}^{N,n}_{\alpha,B}$ (b) $\widehat{q}^{\mathrm{sp}}_{\alpha}$



(c) $\widehat{q}_{\alpha;k\mathrm{NN}}$ (d) $\widehat{q}^{\mathrm{YJ}}_{\alpha}$

Figure 3.8 – The same plots as in Figures 3.5-3.7, but for Model ($\mathcal{M}1$) and $n = 1,000$.

		$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 0.95$
	$\hat{N}^{-}_{\alpha,B;\text{opt}}$	20	23	23	18	18
$(\mathcal{M}1)$	$\hat{\lambda}_{\alpha;\text{opt}}$	0.595	0.971	1.075	0.638	1.033
$n = 300$	$k_{\alpha;\text{opt}}$	25	32	30	28	27
	$\hat{h}_{\alpha;\text{opt}}$	0.202	0.170	0.164	0.170	0.202
	$\hat{N}^{-}_{\alpha,B;\text{opt}}$	23	23	19	16	15
$(\mathcal{M}2)$	$\hat{\lambda}_{\alpha;\text{opt}}$	0.556	0.672	0.632	0.504	0.716
$n = 300$	$k_{\alpha;\text{opt}}$	16	38	42	43	55
	$\hat{h}_{\alpha;\text{opt}}$	0.135	0.113	0.109	0.113	0.135
	$\hat{N}^{-}_{\alpha,B;\text{opt}}$	13	10	9	10	9
$(\mathcal{M}3)$	$\hat{\lambda}_{\alpha;\text{opt}}$	0.518	1.127	1.490	0.840	0.516
$n = 300$	$k_{\alpha;\text{opt}}$	70	61	38	28	34
	$\hat{h}_{\alpha;\text{opt}}$	0594	0.498	0.482	0.498	0.594
	$\hat{N}^{-}_{\alpha,B;\text{opt}}$	32	43	43	41	33
$(\mathcal{M}1)$	$\hat{\lambda}_{\alpha;\text{opt}}$	1.122	1.117	0.927	1.120	1.208
$n = 1,000$	$k_{\alpha;\text{opt}}$	72	54	84	74	94
	$\hat{h}_{\alpha;\text{opt}}$	0.148	0.124	0.120	0.124	0.148

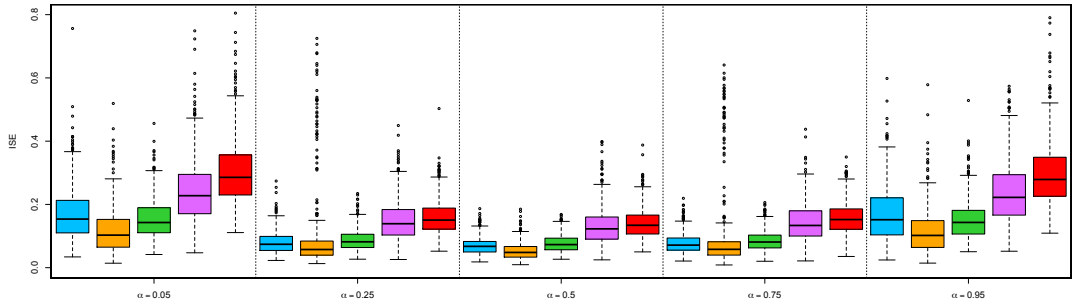
Table 3.1 – Values of the various tuning parameters involved in the conditional quantile estimators considered in Figures 3.5-3.8.

act similarly (and vice versa). For Model $(\mathcal{M}2)$, we see that the optimal value for N decreases with α, which is due to the chi-square error term, implying that estimation of conditional α-quantiles should be based on increasingly many data points as α grows. In accordance with this, the optimal number of neighbors $k_{\alpha;\text{opt}}$ in kNN estimation increases with α. Such a monotonic pattern is not observed for the bandwidth $\hat{h}_{\alpha;\text{opt}}$ used in kernel estimators since it is chosen such a way that $\hat{h}_{\alpha;\text{opt}} = \hat{h}_{1-\alpha;\text{opt}}$. Finally, the parameter $\hat{\lambda}_{\alpha;\text{opt}}$, that is a monotone decreasing function of the number of knots of the splines, is related to α in a quite unclear fashion.

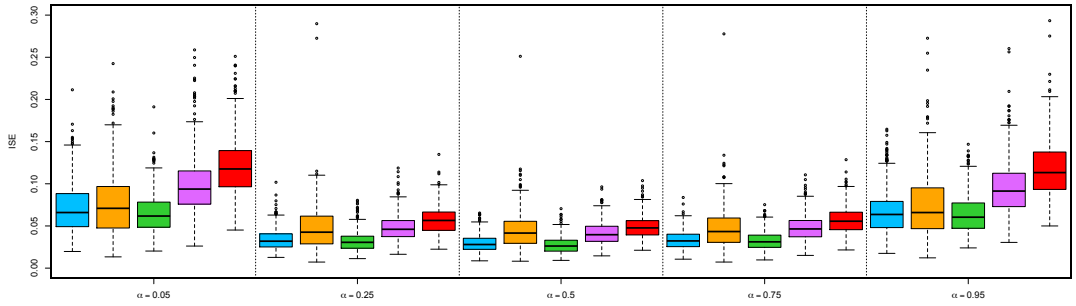
3.3.3 Comparison of the ISEs

Obtaining well-behaved curves is of course desirable, particularly so in applied statistics, but this should not be achieved at the expense of efficiency. That is why we now compare the various estimators in terms of ISEs. To do so, we generated 500 independent samples from Models $(\mathcal{M}1)$-$(\mathcal{M}3)$ with sample sizes $n = 300$ and $n = 1,000$. In each case, we evaluated the ISEs corresponding to the quantization-based estimators and to their four competitors defined in Section 3.3.1. For each model, sample size, and quantile order α considered, it provides a series of 500 observed ISEs for each estimator. Figures 3.9-3.11 draw the boxplots of those 500 ISE values for Models $(\mathcal{M}1)$-$(\mathcal{M}3)$, respectively. ISEs are, as expected, smaller for $n = 1,000$ than for $n = 300$, but otherwise are quite similar.

For most models and quantile orders considered, the quantization-based estimators, spline estimators, and kNN estimators provide smaller ISEs than the local constant and local linear



(a) $n = 300$



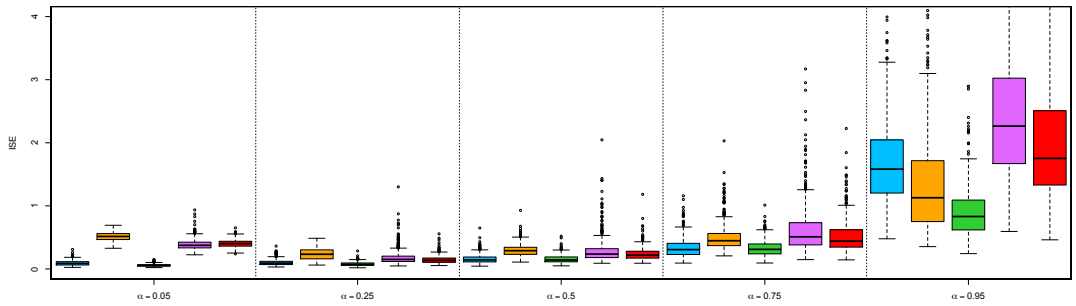
(b) $n = 1,000$

Figure 3.9 – Boxplots, for $\alpha = 0.05$, 0.25, 0.50, 0.75, and 0.95, of the ISEs of various conditional α-quantile estimators obtained from 500 independent random samples from Model ($\mathcal{M}1$), with size $n = 300$ (top) and $n = 1,000$ (bottom). The estimators considered are the quantization-based estimator $\bar{q}^{N,n}_{\alpha,B}$ (in blue), the spline estimator $\widehat{q}^{\mathrm{sp}}_{\alpha}$ (in orange), the kNN estimator estimator $\widehat{q}_{\alpha;k\mathrm{NN}}$ (in green), the local linear estimator $\widehat{q}^{\mathrm{YJ}}_{\alpha}$ (in purple) and the local constant estimator $\widehat{q}^{\mathrm{YJc}}_{\alpha}$ (in red).

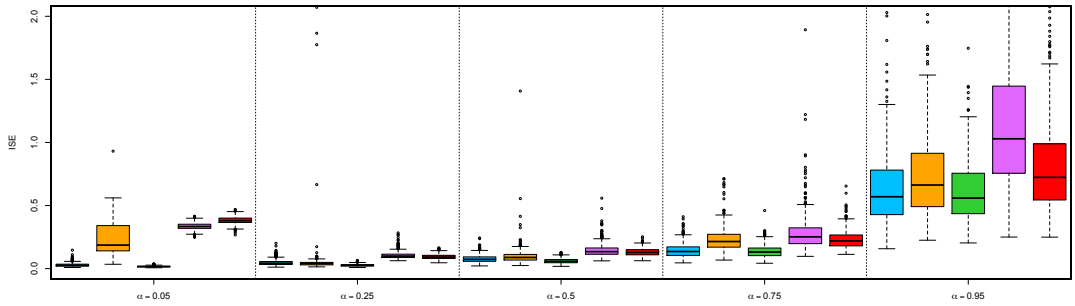
estimators. Since the kNN estimators are based on a selection of k that is infeasible, the main competitors to our quantization-based estimators are those based on splines. Results reveal that polynomial link functions (see Figure 3.9) usually are more favorable to spline-based methods (Figure 3.9b provides an exception, though, which shows that the sample size may also play a role). On the contrary, quantization-based estimators are better for more complex link function (see Figures 3.10-3.11), which is in line with the comparison of the quantization-based and spline-based estimated curves in Section 3.3.2 (note, however, that an exception also appears in Figure 3.11b).

Let us comment the good performances of $\bar{q}^{N,n}_{\alpha,B}(x)$ with respect to the ones of kernel estimators. The main advantage of quantization is that the radius of a Voronoi cell (i.e. the larger distance between a point of the grid and any point projected on it) is adaptive: when there are many points in a region, the number of quantizers is more important in this region than in other less dense regions. The radius of the cells can be seen as a data-adaptive bandwidth parameter. It is then not surprising that $\bar{q}^{N,n}_{\alpha,B}$ outperforms the kernel estimators $\widehat{q}^{\mathrm{YJc}}_{\alpha}$ and $\widehat{q}^{\mathrm{YJ}}_{\alpha}$ in situations where this feature appears, since the bandwidth h is the same for any x in $\widehat{q}^{\mathrm{YJ}}_{\alpha}(x)$.

Since the computational burden is also an important issue, we gather in Table 3.2 the compu-



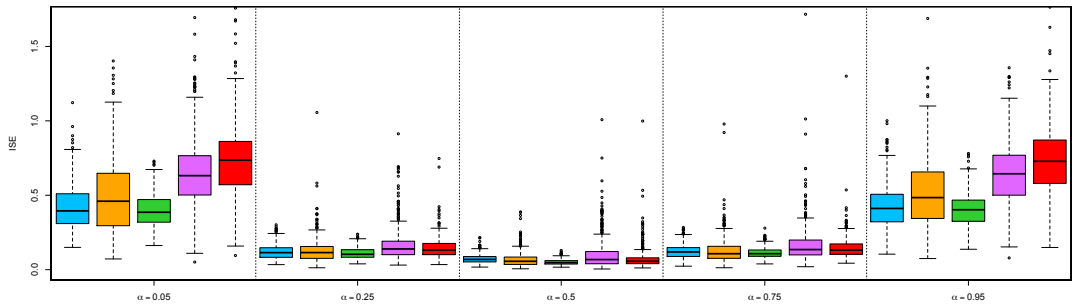
(a) $n = 300$



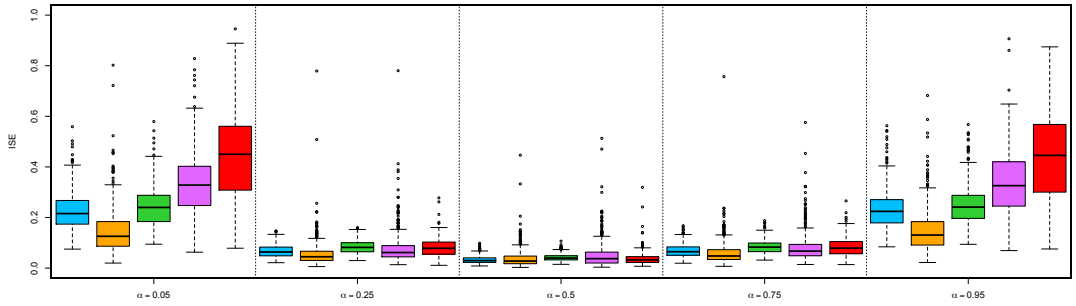
(b) $n = 1,000$

Figure 3.10 – The same boxplots as in Figure 3.9, but for Model ($\mathcal{M}2$).

tation times (in seconds) used by each estimator to produce Figure 3.9a. To study computation times in the light of efficiency, we further report there, for each estimator, a global measure of efficiency (Eff), defined as the sum, over the five α-values considered in that figure, of the medians of the 500 fixed-α observed ISEs. Since the choice of the grid $\mathcal{N}$ plays an important role in the selection of the optimal N-value (at least from a computational point of view), we considered two grids, namely $\mathcal{N}_1 = \{5, 6, \ldots, 29, 30\}$ and $\mathcal{N}_5 = \{5, 10, \ldots, 25, 30\}$ (the latter one, that may seem too coarse, is actually the one that led to the good ISE performances in Figures 3.9). Table 3.2 confirms that there is no free lunch, as it shows that the gain in terms of efficiency (for Model ($\mathcal{M}1$) and $n = 300$) has a price in terms of computation time. This price, however, is quite reasonable. Moreover, the Mac and Linux versions of our R package **QuantifQuantile** offer an option for parallel computing, which divides the computing times shown in Table 3.2 by a factor of 4, which clearly makes our estimators very competitive in this respect. In addition, the computation times for the kNN estimator should be considered with care, since the procedure here is not based on a (typically always quite computationally intensive) data-driven selection of smoothing parameters. This table also gathers the efficiency measures Eff associated with Figures 3.10a and 3.11a (we do not provide the corresponding CPU times, that barely depend on the model considered). The grids $\mathcal{N}_1$ and $\mathcal{N}_5$ actually differ across models but not their size.



(a) $n = 300$



(b) $n = 1,000$

Figure 3.11 – The same boxplots as in Figures 3.9-3.10, but for Model ($\mathcal{M}3$).

		$\bar{q}^{N,n}_{\alpha,B}$ ($\mathcal{N}_1$)	$\bar{q}^{N,n}_{\alpha,B}$ ($\mathcal{N}_5$)	$\widehat{q}^{\text{sp}}_{\alpha}$	$\widehat{q}_{\alpha;k\text{NN}}$	$\widehat{q}^{\text{YJ}}_{\alpha}$	$\widehat{q}^{\text{YJc}}_{\alpha}$
CPU		11,108.3	2,583.8	2,258.6	541.4	1,319.8	1,178.4
Eff	($\mathcal{M}1$)	0.503	0.518	0.368	0.522	0.845	1.001
	($\mathcal{M}2$)	1.938	2.211	2.614	1.412	3.535	2.950
	($\mathcal{M}3$)	1.108	1.109	1.223	1.046	1.618	1.784

Table 3.2 – (First line:) Computation times (CPU, in seconds) used by each estimator to obtain Figure 3.9a. (Subsequent lines:) ISE-based global efficiency measures (Eff) associated with Figures 3.9a, 3.10a, and 3.11a; see Section 3.3.3 for details.

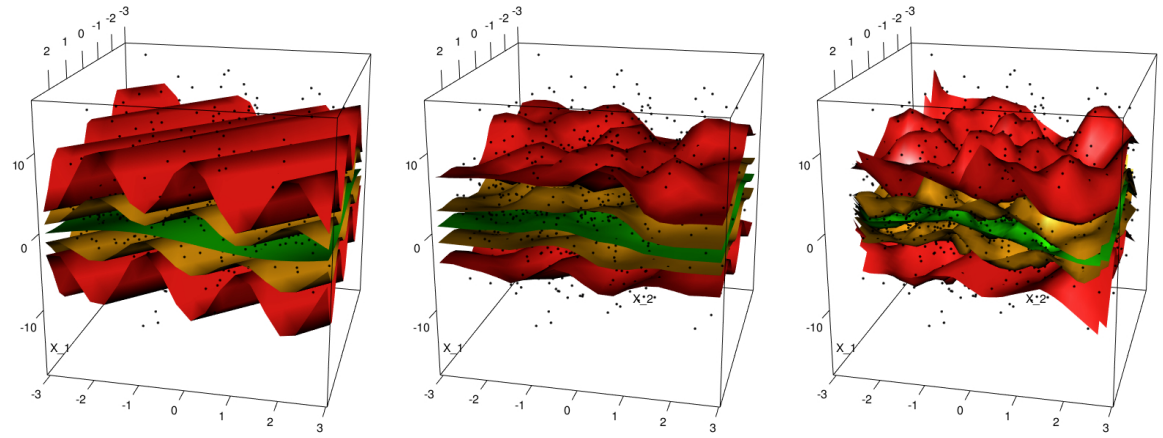
3.4 Extension to multivariate regressors ($d > 1$)

The numerical exercises and simulations in Sections 3.3 focused on the case of a single covariate ($d = 1$). All previous definitions and results, however, cover the general case ($d \geq 1$): this not only includes the theoretical results from Chapter 2, but also the CLVQ algorithm and the proposed data-driven method to select N. It is therefore natural to investigate how well quantization-based conditional quantile estimation performs for $d > 1$.

To do so, we focus on the bivariate case $d = 2$ to be able to plot the resulting conditional quantile hypersurfaces. We first generated a random sample of size $n = 3,000$ from the model

$$(\mathcal{M}4)\quad Y = \sin(X_1 + X_2) + \big(0.5 + 1.5\sin^2(\tfrac{\pi}{2}(X_1 + X_2))\big)\varepsilon,$$

with $X = (X_1, X_2)' = (6Z_1 - 3, 6Z_2 - 3)'$ ($Z_i \sim \text{Beta}(2, 2)$, $i = 1, 2$) and $\varepsilon \sim \mathcal{N}(0, 9)$, where Z_1, Z_2 and ε are mutually independent. Conditional quantile hypersurfaces of this model for α =0.05,



(a) (b) (c)

Figure 3.12 – Various conditional α-quantile surfaces for Model ($\mathcal{M}4$) and $\alpha = 0.05$, 0.25, 0.5, 0.75 and 0.95. The quantile surfaces are the population ones (left), the ones estimated (from 3,000 independent observations) through quantization (center), and through triogram-based splines (right); see Section 3.4 for details.

0.25, 0.5, 0.75, and 0.95 are plotted in Figure 3.12a. Figure 3.12b provides the corresponding sample quantile surfaces obtained from our estimator $\bar{q}^{N,n}_{\alpha,B}$, with $B = 50$ and the optimal N-values that are obtained through our data-driven N-selection procedure for the various values of α. The quantile surfaces were readily obtained from the R package **QuantifQuantile**, that can therefore also be used for $d > 1$. Clearly, $\bar{q}^{N,n}_{\alpha,B}$ provides very smooth surfaces that properly catch the link function, even though the amplitude, for extreme values of α, is a bit under-estimated.

The results of Section 3.3 suggest restricting the comparison to spline-based estimators. For this purpose, we consider the triogram-based bivariate splines from Koenker and Mizera (2004), that is implemented in the `rqss` function of the R package `quantreg`. The resulting estimated quantile surfaces, based on the same λ-selection method as for $d = 1$ in Section 3.3, are plotted in Figure 3.12c. Clearly, these quantile surfaces exhibit more variability than the quantization-based ones. Incidentally, note that, unlike the proposed estimator, this spline-based estimator does not extend easily to $d > 2$.

Parallel to our investigation of the case $d = 1$, we complement the results above with some efficiency results in terms of ISE. To do so, we generated 50 independent random samples from Model ($\mathcal{M}4$) above with sample size $n = 3,000$ and computed, for each sample and various values of α, the quantization-based and spline-based conditional quantile estimators. The corresponding 50 resulting ISEs for each estimator and every α considered, that are defined in a similar way as in Section 3.2, are represented in Figure 3.13. In line with the $d = 1$ results in Figure 3.11, this shows that, for the complex link function considered, $\bar{q}^{N,n}_{\alpha,B}$ dominates $\widehat{q}^{\,\mathrm{sp}}_{\alpha}$ in terms of efficiency. Note that this is particularly true for extreme values of α.

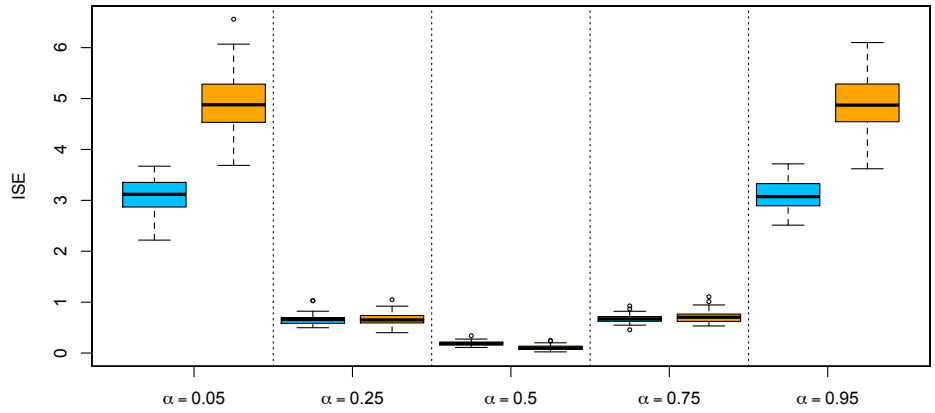


Figure 3.13 – Boxplots, for $\alpha = 0.05$, 0.25, 0.50, 0.75, and 0.95, of the ISEs of quantization-based (in blue) and triogram-based spline (in orange) conditional α-quantile estimators obtained from 500 independent random samples from Model ($\mathcal{M}4$), with size $n = 3,000$.

3.5 R-package QuantifQuantile

For the sake of ease of use, we developed a package in R called **QuantifQuantile**, that allows to perform quantization-based quantile regression (Charlier et al., 2015d). We will see in the sequel that this package aims to help the user compute the corresponding estimators in a quite straightforward way.

This section provides a description of the various functions offered by the R package **QuantifQuantile**. We first detail the three functions that allow to estimate conditional quantiles through quantization (Section 3.5.1). Then we describe a function computing optimal quantization grids (Section 3.5.2).

3.5.1 Conditional quantile estimation

QuantifQuantile is composed of three main functions that each provides estimated conditional quantiles in (2.3.1)-(2.3.3). These functions work in a similar way but address different values of d (recall that d is the dimension of the covariate vector X):

- The function `QuantifQuantile` is suitable for $d = 1$.
- The function `QuantifQuantile.d2` addresses the case $d = 2$.
- Finally, `QuantifQuantile.d` can deal with an arbitrary value of d.

Notice that, if `QuantifQuantile.d` can be used whatever the dimension d, it is highly recommended to use it only for $d \geq 3$ since the codes inside the first two functions are more adapted for $d = 1$ and $d = 2$, providing a gain in computing time in comparison with `QuantifQuantile.d`.

Combined with the `plot` function, the first two functions provide reference curves and reference surfaces, respectively. No graphical outputs can be obtained from the third function.

The three functions share the same arguments, but not necessarily the same default values. For each function, using `args()` displays the various arguments and corresponding default values:

```
R> args(QuantifQuantile)
```

```
function (X, Y, alpha = c(0.05, 0.25, 0.5, 0.75, 0.95), x =
        seq(min(X), max(X), length = 100), testN = c(35, 40, 45,
        50, 55), p = 2, B = 50, tildeB = 20, same_N = TRUE,
        ncores = 1)
```

```
R> args(QuantifQuantile.d2)
```

```
function (X, Y, alpha = c(0.05, 0.25, 0.5, 0.75, 0.95), x =
        matrix(c(rep(seq(min(X[1, ]), max(X[1, ]), length = 20),
        20), sort(rep(seq(min(X[2, ]), max(X[2, ]), length = 20),
        20))), nrow = 2, byrow = TRUE), testN = c(110, 120, 130,
        140, 150), p = 2, B = 50, tildeB = 20, same_N = TRUE,
        ncores = 1)
```

```
R> args(QuantifQuantile.d)
```

```
function (X, Y, x, alpha = c(0.05, 0.25, 0.5, 0.75, 0.95),
        testN = c(35, 40, 45, 50, 55), p = 2, B = 50, tildeB = 20,
        same_N = TRUE, ncores = 1)
```

We now give more details on these arguments.

- `X`: a $d \times n$ real array (a vector of length n for `QuantifQuantile`, required by all three functions). The columns of this matrix are the X_i's, $i = 1, \ldots, n$.
- `Y`: an $n \times 1$ real array (required by all three functions). This vector collects the Y_i's, $i = 1, \ldots, n$.
- `alpha`: an $r \times 1$ array with components in $(0, 1)$ (optional for all three functions, with r the number of α-values considered). This vector contains the orders for which $q_\alpha(x)$ should be estimated.
- `x`: a $d \times J$ real array (optional for `QuantifQuantile` and `QuantifQuantile.d2`, required by `QuantifQuantile.d`). The columns of this matrix are the x_j's at which the quantiles $q_\alpha(x_j)$ are to be estimated. If `x` is not provided when calling `QuantifQuantile`, then it is set to a vector of $J = 100$ equispaced values between the minimum and the maximum of the X_i's. If this argument is not provided when calling `QuantifQuantile.d2`, then the default for `x` is a matrix whose $J = 20^2 = 400$ column vectors are obtained as follows: 20 equispaced values are considered between the minimum and maximum values of the $(X_i)_1$'s and similarly for the second component. The 400 column vectors of the default `x` are obtained by considering all combinations of those 20 values for the first component with the 20 values for the second one[1].

[1]Since the number J of points in a default value of `x` obtained in this fashion would increase exponentially with the dimension d, we did not adopt the same approach for $d \geq 3$.

- `testN`: an $m \times 1$ vector of pairwise distinct positive integers (optional for all three functions). The entries of this vector are the elements of the set $\mathcal{N}$ in (3.2.3)-(3.3.1), hence are the N-values for which the $\widehat{\text{ISE}}^{-}_{\alpha}$ quantity considered will be evaluated. The default is $(35, 40, \ldots, 55)$ but it is strongly recommended to adapt it according to the sample size n at hand.
- `p`: a real number larger than or equal to one (optional for all three functions). It is the parameter p to be used when performing optimal quantization in L_p-norm.
- `B`: a positive integer (optional for all three functions). It is the number of bootstrap replications B to be used in (2.3.3).
- `tildeB`: a positive integer (optional for all three functions). It is the number of bootstrap replications $\widetilde{B}$ to be used when determining $\hat{N}^{-}_{\alpha;\text{opt}}$ or $\hat{N}^{-}_{\text{opt}}$.
- `same_N`: a boolean variable (optional for all three functions). If `same_N=TRUE`, then a common value of N (that is, $\hat{N}^{-}_{\text{opt}}$ in (3.3.1)) will be selected for all α's. If `same_N=FALSE`, then optimal values of N will be chosen independently for the various α (which will provide several $\hat{N}^{-}_{\alpha;\text{opt}}$, as in (3.2.3)).
- `ncores`: number of cores to use. These functions can use parallel computation to save time by increasing this parameter. Parallel computation relies on `mclapply` from **parallel** package (R Core Team, 2015), hence is not available on Windows.

All three functions return the following list of objects, which is of class `"QuantifQuantile"`:

- `hatq_opt`: an $r \times J$ real array. If `same_N=TRUE`, then the entry (i, j) of this matrix is $\bar{q}^{\hat{N}^{-}_{\text{opt}},n}_{\alpha_i,B}(x_j)$. If `same_N=FALSE`, then it is rather $\bar{q}^{\hat{N}^{-}_{\alpha_i;\text{opt}},n}_{\alpha_i,B}(x_j)$. This object can also be returned using the usual `fitted.values` function.
- `N_opt`: a positive integer (if `same_N=TRUE`) or an $r \times 1$ array of positive integers (if `same_N=FALSE`). Depending on `same_N`, this provides the value of $\hat{N}^{-}_{\text{opt}}$ or the vector $(\hat{N}^{-}_{\alpha_1;\text{opt}}, \ldots, \hat{N}^{-}_{\alpha_r;\text{opt}})$.
- `hatISE_N`: an $r \times m$ real array. The entry (i, j) of this matrix is $\widehat{\text{ISE}}^{-}_{\alpha_i}(N_j)$. Plotting this for fixed α or plotting its average over the various α, in both cases over `testN`, allows to assess the global convexity of these ISEs. Hence, it can be used to illustrate whether or not the choice of `testN` was adequate. This will be illustrated in the examples below.
- `hatq_N`: an $r \times J \times m$ real array. The entry (i, j, ℓ) of this matrix is $\bar{q}^{N_\ell,n}_{\alpha_i,B}(x_j)$, where N_ℓ is the ℓ^{th} entry of the argument `testN`. From this output, it is easy by fixing the third entry to get the matrix of the $\bar{q}^{N,n}_{\alpha_i,B}(x_j)$ values for any N in `testN`.
- The arguments `X`, `Y`, `x`, `alpha`, and `testN` are also reported in this response list.

Moreover, when the optimal value `N_opt` selected is on the boundary of `testN`, which means that `testN` most likely was not well chosen, a warning message is printed.

The "`QuantifQuantile`" class response can be used as argument of the functions `plot` (only for $d \leq 2$), `summary` and `print`. The `plot` function draws the observations and plots the estimated conditional quantile curves ($d = 1$) or surfaces ($d = 2$) — for $d = 2$, the **rgl** package is used (Adler

et al., 2015), which allows to change the perspective in a dynamic way. In order to illustrate the selection of N, the function `plot` also has an optional argument `ise`. Setting this argument to `TRUE` (the default is `FALSE`), this function, that can be used for any dimension d, provides the plot (against N) of the $\widehat{\mathrm{ISE}}_{\alpha}$ and $\widehat{\mathrm{ISE}}$ quantities in (3.2.3) or in (3.3.1), depending on the choice `same_N=FALSE` or `same_N=TRUE`, respectively; see the examples below for details. If $d \leq 2$, it also returns the fitted quantile curves or surfaces.

3.5.2 Computing optimal grids

Besides the functions that allow to estimate conditional quantiles and plot the corresponding reference curves/surfaces, **QuantifQuantile** provides a function that computes optimal quantization grids. This function, called `choice.grid`, admits the following arguments:

- `X`: a $d \times n$ real array (required). The columns of this matrix are the X_i's, $i = 1, \ldots, n$, for which the optimal quantization grid should be determined. Each point of `X` is used as a stimulus in the stochastic gradient algorithm to get an optimal grid.
- `N`: a positive integer (required). The size of the desired quantization grid.
- `ng`: a positive integer (optional). The number of desired quantization grids. The default is 1.
- `p`: a real number larger than or equal to one (optional). It is the parameter p used in the quantization error. The default is 2.

Let us detail the parameter `ng`. In some cases, it may be necessary to have several quantization grids. For example, in the use of this function inside `QuantifQuantile` and its multidimensional versions, we need `B + tildeB` grids. If `ng` > 1, the different grids are obtained using as stimuli a resampling version of `X` (the X_t's in Section 1.2.4).

The output is a list containing the following elements:

- `init_grid`: a d×`N`×`ng` real array. The entry (i, j, ℓ) of this matrix is the i^{th} component of the j^{th} point of the ℓ^{th} initial grid.
- `opti_grid`: a d×`N`×`ng` real array. The entry (i, j, ℓ) of this matrix is the i^{th} component of the j^{th} point of the ℓ^{th} optimal grid provided by the algorithm.

3.5.3 Illustrations

In this section, we illustrate on several examples the use of the functions described above. Examples 1 and 2 restrict to `QuantifQuantile`/`QuantifQuantile.d2` and provide graphical representations in each case. We conclude this section with an illustration of the function `choice.grid`.

Example 1: one-dimensional covariate

We generate a random sample of size $n = 300$, where X is uniformly distributed on the interval $(-2, 2)$ and where Y is obtained by adding to X^2 an independent standard normal error term:

```
R> set.seed(258164)
R> n <- 300
R> X <- runif(n, -2, 2)
R> Y <- X^2 + rnorm(n)
```

We test the number N of quantizers between 10 and 30 by steps of 5 and we do not change the default values of the other arguments. We then run the function `QuantifQuantile` with these arguments and we stock the response in `res`.

```
R> testN <- seq(10, 30, by = 5)
R> res <- QuantifQuantile(X, Y, testN = testN)
```

No warning message is printed, which means that this choice of `testN` was adequate. To assess this in a graphical way, we use the function `plot` with `ise` argument set to `TRUE` that plots `hatISEmean_N` (obtained by taking the mean of `hatISE_N` over `alpha`) against the various N-values in `testN`.

```
R> plot(res, ise = TRUE)
```

Figure 3.14a provides the resulting graph, which confirms that `testN` was well chosen since `hatISEmean_N` is larger for smaller and larger values of N than `N_opt`. We then plot the corresponding estimated conditional quantiles curves in Figure 3.14b. The default colors of the points and of the curves are changed by using the `col.plot` argument. This argument is a vector of size 1+`length(alpha)`, whose first component fixes the color of the data points and whose remaining components determine the colors of the various reference curves.

```
R> col.plot <- c("grey", "red", "orange", "green", "orange", "red")
R> plot(res, col.plot = col.plot, xlab = "X", ylab = "Y")
```

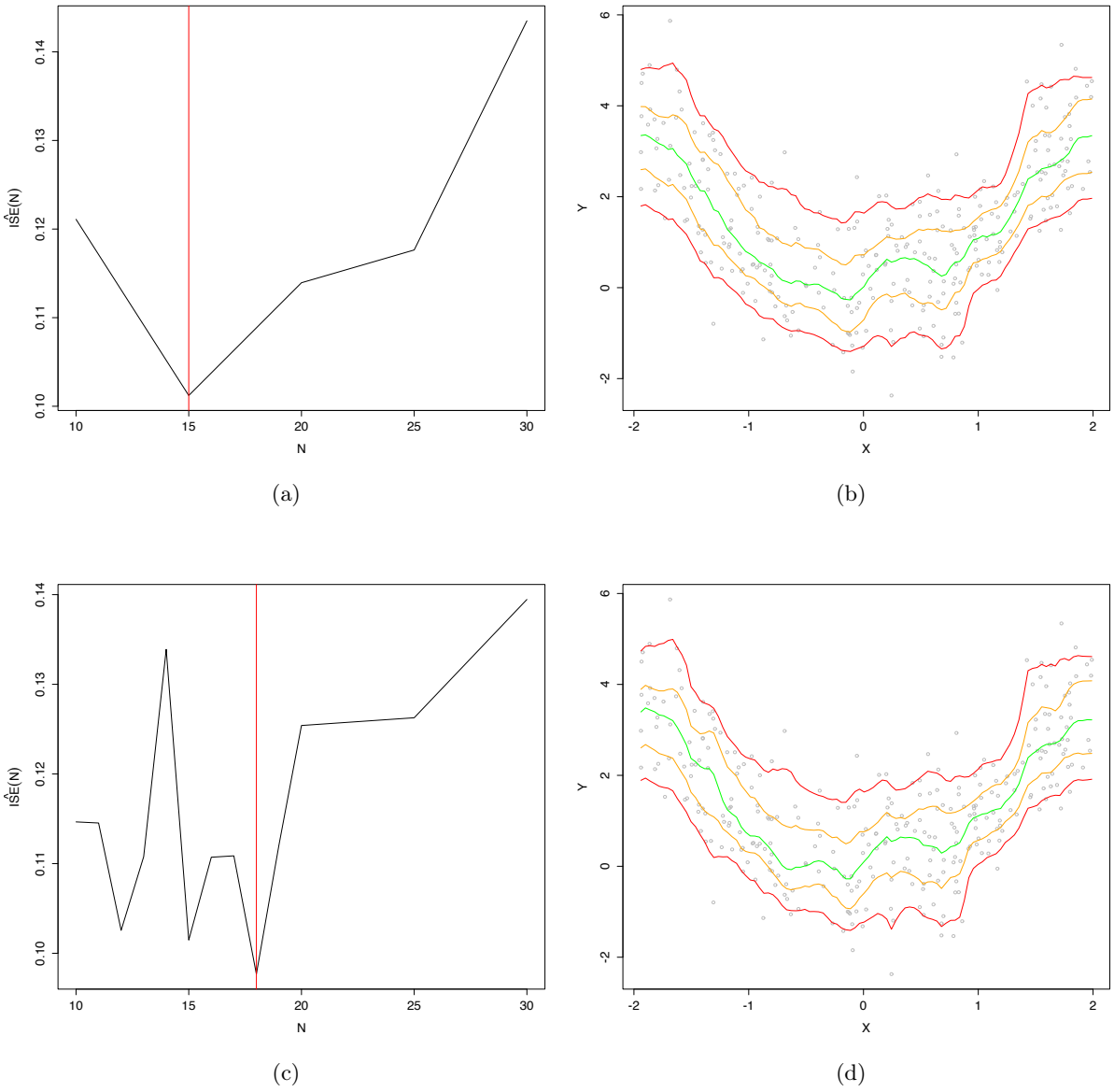
It is natural to make the grid `testN` finer. Of course, the more N-values we test, the longer it takes. It is why we adopted this two-stage approach, where the goal of first stage was to get a rough approximation of the optimal N-value. In the second stage, we can then refine the grid only in the vicinity of the value `N_opt` obtained in the first stage.

```
R> testN <- c(seq(10, 20, by = 1), seq(25, 30, by = 5))
R> res_step1 <- QuantifQuantile(X, Y, testN = testN)
R> plot(res_step1, ise = TRUE, col.plot = col.plot, xlab = "X",
        ylab = "Y")
```

The resulting graphs are plotted in Figure 3.14c and 3.14d respectively. We observe that the value of `N_opt` is made more precise, since we now get `N_opt` = 18 instead of 15. The resulting estimated conditional quantiles curves in Figure 3.14d are very similar to the ones in Figure 3.14b.

So far, we used the default value `same_N=TRUE`, which leads to selecting an N-value that is common to all α's. For the sake of comparison, we now explore the results for `same_N=FALSE`.

```
R> testN <- c(seq(10, 30, by = 5))
R> res2 <- QuantifQuantile(X ,Y ,testN = testN, same_N = FALSE)
R> plot(res2, ise = TRUE, col.plot = col.plot, xlab = "X",
```



(a) (b)

(c) (d)

Figure 3.14 – For the sample considered in Example 1, this figure provides (a) the plot of $N \mapsto \widehat{\overline{\mathrm{ISE}}}(N)$ with $N \in \{10, 15, 20, 25, 30\}$, and (b) the resulting reference curves. The panels (c)-(d) provide the corresponding plots when taking $N \in \{10, 11, 12, \ldots, 19, 20, 25, 30\}$.

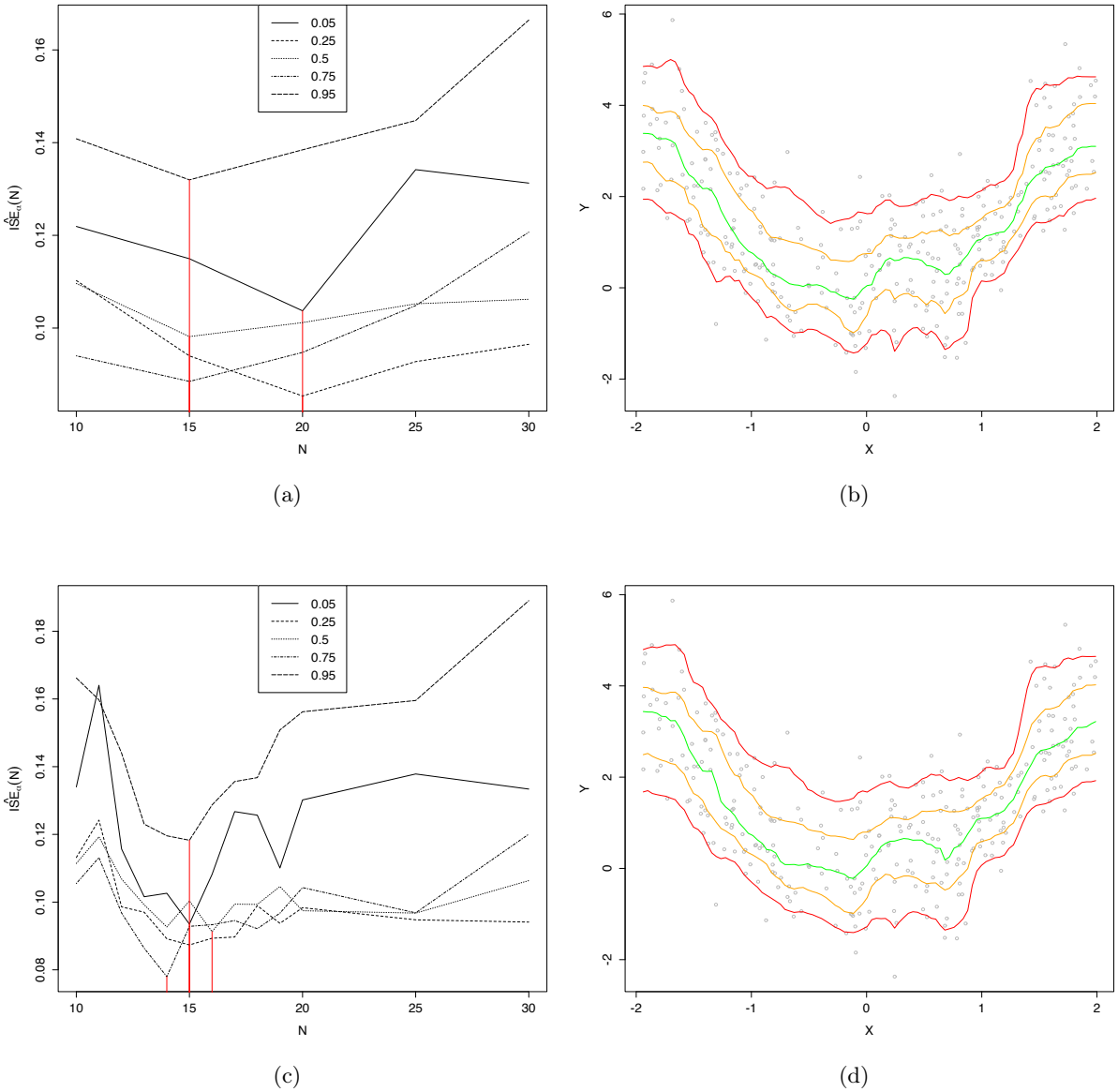


Figure 3.15 – The same results as in Figure 3.14, but when selecting optimal values of N separately for each α.

```
        ylab = "Y")
R> testN <- c(seq(10, 20, by = 1), seq(25, 30, by = 5))
R> res2_step1 <- QuantifQuantile(X,Y,testN = testN,same_N = FALSE)
R> plot(res2_step1, ise = TRUE, col.plot = col.plot, xlab = "X",
        ylab = "Y")
```

The results are provided in Figure 3.15. Comparing the left panels of Figures 3.14 and 3.15, we see that when choosing N by steps of five, we find `N_opt` = 15 with `same_N=TRUE` and `N_opt` = 15 or 20 (depending on `alpha`) for `same_N=FALSE`. When we refine the grid `testN`, we find analogously `N_opt` = 18 for `same_N=TRUE` and `N_opt` = 14, 15, or 16 for `same_N=FALSE`. In the present setup, thus, both methods provide relatively close optimal N-values, which explains why the corresponding estimated reference curves are so similar (see the right panels of Figures 3.14 and 3.15). Therefore, the grid of N-values tested in Figure 3.14, that may seem too coarse at first sight, actually provides fitted curves that are as satisfactory as those associated with the finer grid in Figure 3.15.

Example 2: two-dimensional covariate

The sample considered here is made of $n = 1,000$ independent realizations of a random vector $(X', Y)'$, where $X = (X_1, X_2)'$ is uniformly distributed on the square $(-2, 2)^2$ and where Y is obtained by adding to $X_1^2 + X_2^2$ an independent standard normal error term.

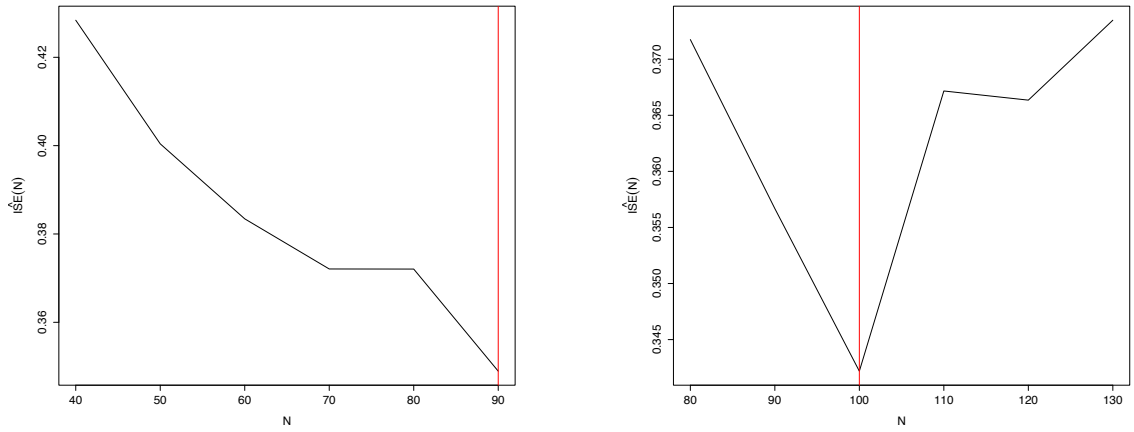
```
R> set.seed(642516)
R> n <- 1000
R> X <- matrix(runif(n*2, -2, 2), ncol = n)
R> Y <- apply(X^2, 2, sum) + rnorm(n)
```

We test N between 40 and 90 by steps of 10. We change the values of `B` and `tildeB` to reduce the computational burden, that is heavier when $d = 2$ than when $d = 1$. We keep the default values of all other arguments when running the function `QuantifQuantile.d2`. Here, a warning message is printed informing us that `testN` was not well-chosen. We confirm it with the function `plot` with `ise` argument set to `TRUE`.

```
R> testN <- seq(40, 90, by = 10)
R> B <- 20
R> tildeB <- 15
R> res <- QuantifQuantile.d2(X, Y, testN = testN, B = B,
                  tildeB = tildeB)
R> plot(res, ise = TRUE)
```

Figure 3.16a provides the resulting graph. The parameter `testN` was not well chosen since `hatISEmean_N` becomes smaller and smaller as `N_opt` increases. We then adapt the choice of `testN` accordingly and rerun the procedure, which identifies an optimal N-value equal to 100; see Figure 3.16b.

```
R> testN <- seq(80, 130, by = 10)
R> res <- QuantifQuantile.d2(X, Y, testN = testN, B = B,
                  tildeB = tildeB)
```



(a) (b)

Figure 3.16 – For the sample considered in Example 2, this figure plots $N \mapsto \widehat{\overline{\mathrm{ISE}}}(N)$ (a) for $N \in \{40, 50, 60, 70, 80\}$ and (b) for $N \in \{80, 90, \ldots, 120, 130\}$.

```
R> plot(res, ise = TRUE)
```

We then plot the corresponding estimated conditional quantile surfaces in Figure 3.17. This figure gathers two prints of the **rgl** plot, but this plot is actually interactive and the user can move it with the mouse in order to view it in any orientation.

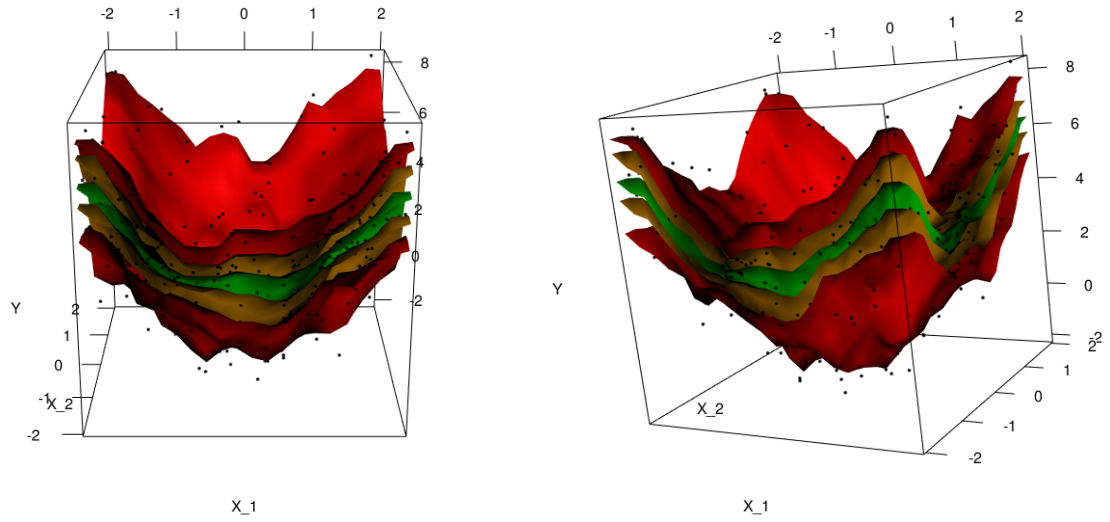
```
R> col.plot <- c("black", "red", "orange", "green", "orange", "red")
R> plot(res, col.plot = col.plot, xlab = "X_1", ylab = "X_2",
        zlab = "Y")
```

Illustration of choice.grid

We now focus on the function `choice.grid` in the univariate and bivariate settings. This function helps obtain optimal quantization grids and was actually used to construct the grids of Section 1.2.4. We now give and detail the code allowing to obtain these illustrations, but we refer the reader to Section 1.2.4 for the graphical outputs and observations.

We start with the univariate case and generate a random sample of size $n = 500$ from the uniform distribution over $(-2, 2)$. With `N` $= 15$ and `ng` $= 1$, this function provides a single initial grid (obtained by sampling without replacement among the uniform sample) and the corresponding optimal grid returned by the algorithm. Figure 1.7 of Chapter 1 (top left) represents the observations (in grey), the initial grid (in red), and the optimal grid (in green). The middle and bottom left figures plot the empirical cdf of the observations projected onto the initial grids and the optimal grids, respectively. The same exercise is repeated with sample size $n = 5,000$, and the results are also given in Figure 1.7 (right panel).

```
R> set.seed(643625)
R> n <- 500
R> X <- runif(n, -2, 2)
```



(a) (b)

Figure 3.17 – For the sample considered in Example 2, this figure plots (with two different views) the estimated conditional quantile surfaces obtained with the `plot` function for α =0.05, 0.25, 0.50, 0.75 and 0.95.

```
R> N <- 15
R> ng <- 1
R> res <- choice.grid(X, N, ng)
R> # Plots of the initial and optimal grids
R> plot(X, rep(0, n), col = "grey", cex = 0.5, ylim = c(-0.1,
         1.1), yaxt = "n", ylab = " ")
R> points(res$init_grid, rep(0.5, N), col = "red", pch = 16,
        cex = 1.2)
R> points(res$opti_grid, rep(1, N), col = "forestgreen", pch = 16,
        cex = 1.2)
R> # To plot the ecdf
R> projX_init <- array(0, dim=c(n, 1))
R> projX_opti <- array(0, dim=c(n, 1))
R> i_init <- array(0, dim = c(n, 1))
R> i_opti <- array(0, dim = c(n, 1))
R> for (i in 1:n) {
R>   RepX <- rep(X[i], N)
R>   diff_init <- sqrt((RepX - res$init_grid)^2)
R>   diff_opti <- sqrt((RepX - res$opti_grid)^2)
R>   i_init[i] <- which.min(diff_init)
R>   i_opti[i] <- which.min(diff_opti)
R>   projX_init[i] <- res$init_grid[i_init[i]]
R>   projX_opti[i] <- res$opti_grid[i_opti[i]]
R> }
R> plot(ecdf(projX_init), main = " ", col = "red", xlim = c(-2, 2))
R> plot(ecdf(projX_opti), main = " ", col = "forestgreen",
        xlim = c(-2, 2))
```

We conclude this section with an illustration of `choice.grid` in the bivariate case. We gen-

erate 2,000 points from the uniform distribution over the square $(-2,2)^2$, and we choose `N` $= 30$ and `ng` $= 1$. The resulting initial and optimal grids are plotted in the left panels of Figure 1.8 (top and bottom, respectively). The right panels are obtained similarly from 20,000 points instead of 2,000.

```
R> set.seed(345689)
R> n <- 2000
R> X <- matrix(runif(n*2, -2, 2), nc = n)
R> N <- 30
R> ng <- 1
R> res <- choice.grid(X, N, ng)
R> col <- c("red", "forestgreen")
R> l <- c("init_grid[1]", "init_grid[2]", "opti_grid[1]",
          "opti_grid[2]")
R> plot(res$init_grid[1, , 1], res$init_grid[2, , 1],
        col = col[1], xlab = l[1], ylab = l[2])
R> plot(res$opti_grid[1, , 1], res$opti_grid[2, , 1],
        col = col[2], xlab = l[3], ylab = l[4])
```

3.6 Real data examples

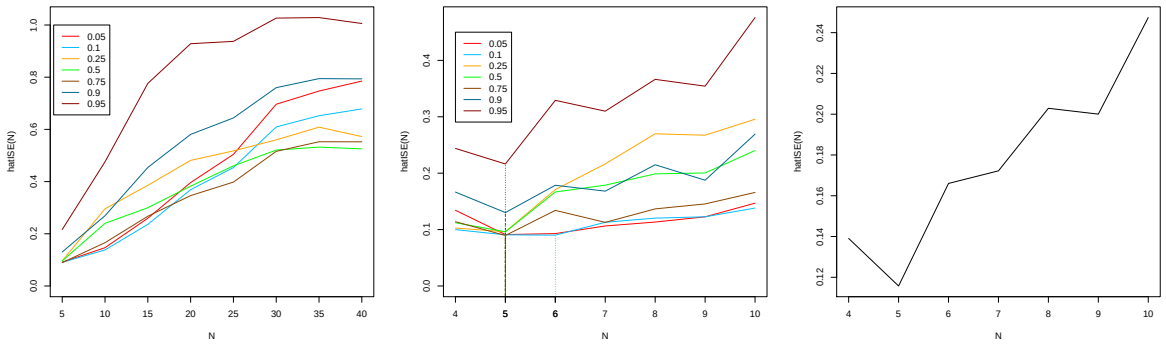
In this section, we illustrate the behavior of $\bar{q}_{\alpha,B}^{N,n}$ on several real data sets. Section 3.6.1 is devoted to a first real data set that was kindly sent to us by Dr. Keming Yu. He used this data set in a previous work with his own estimation method of conditional quantile, which allows us to compare our results with his (see Yu and Jones, 1998). We put the accent on the statistical methodology for this first real data example. In other words, we explain how works our estimation procedure by detailing each step separately. We also compare it with spline methods.

The investigation of Section 3.6.2 is based on a list of data sets collected in towns located in Gironde, France. This list will be detailed more precisely in the corresponding section. This second example aims to highlight the package instead of the methodology (which is always similar). We then focus on the implementation part in this second section.

3.6.1 Concentration of immunoglobulin-G given age for children

This data set, of size $n = 298$, corresponds to the serum concentration, in grams per liter, of immunoglobulin-G in children aged from 6 months to 6 years. It was already investigated in Yu and Jones (1998), where two estimators were considered, namely a kernel local linear estimator and a double kernel estimator. The reader can refer to this paper where (smoothed versions of) the corresponding quantile curves are plotted. Also, we do not plot the quantile curves resulting from the kNN estimator $\widehat{q}_{\alpha;k\mathrm{NN}}$ since the selection of the parameter k is not data-driven, hence cannot be achieved on a real data set. Consequently, we only compare the proposed quantization-based quantile curves with those obtained from their main competitor, namely the spline estimator $\widehat{q}_{\alpha}^{\mathrm{sp}}$.

We now describe how we obtained quantization-based quantile curves in this context, for each α= 0.05, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.95. We chose these quantile orders with respect to



(a) (b) (c)

Figure 3.18 – For the real-data example of Section 3.6.1, these figures provide (a) the plot of $N \mapsto \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ over $N \in \mathcal{N} = \{5, 10, \ldots, 35, 40\}$, (b) the plot of $N \mapsto \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ over $N \in \mathcal{N} = \{4, 5, \ldots, 9, 10\}$, and (c) the plot of $N \mapsto \widehat{\mathrm{ISE}}^{-}(N)$, still over $N \in \mathcal{N} = \{4, 5, \ldots, 9, 10\}$.

the choice of Yu and Jones (1998).

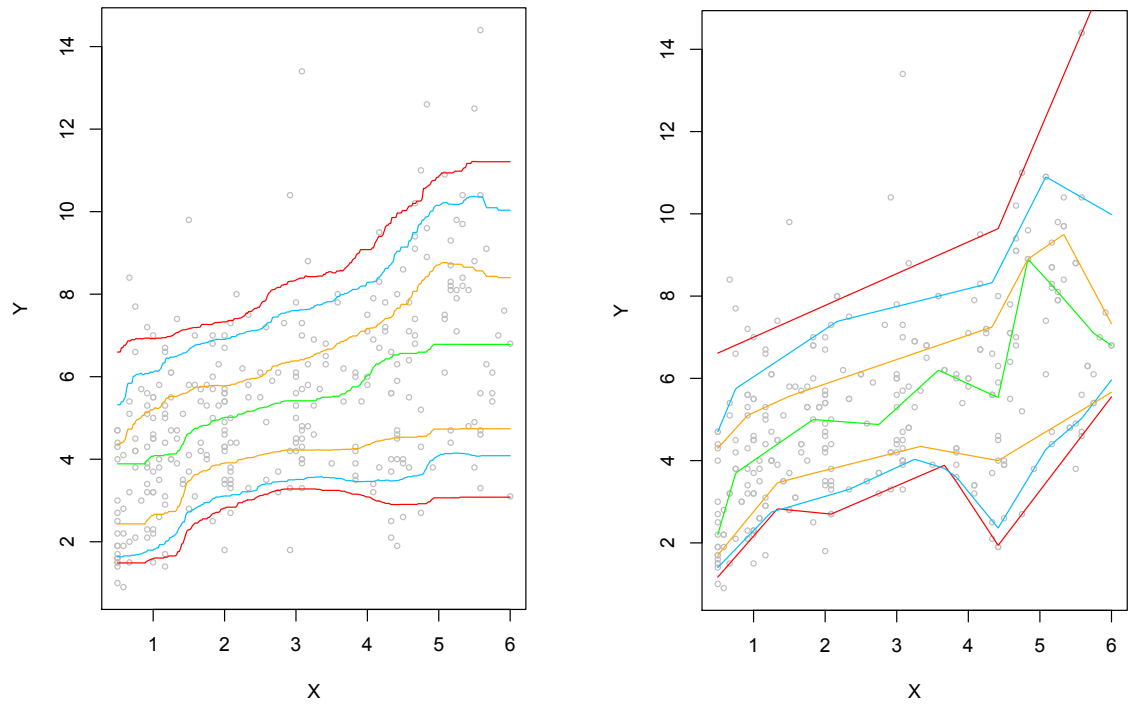
Selection of an optimal N

The first step consists in choosing the optimal number N of quantizers. To do so, we adopted the method proposed in Section 3.2.2, based on the minimization of $\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ (see (3.2.2)), for which we considered throughout $B = 50$, $\widetilde{B} = 30$, and a grid of $J = 300$ equispaced points between the minimum and maximum values of the X-part of the sample. We first evaluated $\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N)$ for all $N \in \mathcal{N} = \{5, 10, \ldots, 35, 40\}$. Figure 3.18a plots the resulting ISE curves for each α. Irrespective of α, these curves are monotone increasing, which means that the tested values of N are too large: we feel that the optimal value is among the smallest and also that a finer grid is desirable. We therefore did the same exercise for $\mathcal{N} = \{4, 5, \ldots, 9, 10\}$, which led to Figure 3.18b. For each α, this provides an optimal N (that is equal to 5 or 6).

As we explained in Section 3.3.2, identifying an optimal N-value separately for each α leads to estimated quantile curves that in principle may cross. If one wants to protect against this, a single optimal value of N should be chosen for all α's using 3.3.1. The resulting ISE curve is plotted in Figure 3.18c and leads to an optimal N-value equal to 5.

Estimation of the conditional quantiles

The second step of course consists in obtaining the estimated quantile curves themselves, based on the selected values of N above. These are the plots of the mappings $x \mapsto \bar{q}^{N,n}_{\alpha,B}(x)$ in (2.3.3), where we chose $B = 50$. The resulting quantile curves are plotted in Figure 3.19. As announced, we compare these curves with those associated with the spline estimator $\widehat{q}^{\mathrm{sp}}_{\alpha}$. For each α considered, the parameter λ was selected according to the AIC procedure described in Section 3.3.1, which led to $\hat{\lambda}_{\alpha;\mathrm{opt}} = 0.40, 0.42, 0.96, 0.63, 0.96, 0.52$, and 0.69. Figure 3.19b plots the resulting quantile curves. Clearly, these piecewise linear curves are more irregular than the quantization-based ones



(a) (b)

Figure 3.19 – Conditional α-quantile curves for the real-data example of Section 3.6.1 and $\alpha = 0.05$, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.95. The estimators considered are the quantization-based estimator $\bar{q}_{\alpha,B}^{N,n}$ with $B = 50$ (left) and the spline estimator $\widehat{q}_\alpha^{\mathrm{sp}}$ (right).

and show some important peaks and slight crossings.

3.6.2 Employment, housing and environment in Gironde, France

This section is based on a list of data sets in R called `gironde`. This list is composed of several variables, numerical or categorical. The variables of type employment, housing and services were collected by the INSEE[2] while the environmental ones by the IGN[3]. These data are freely available for research institutes and were conditioned into this list by the IRSTEA[4], that kindly sent it to us. These data were joined to the **QuantifQuantile** package.

We of course only use in our approach the 16 different numerical variables. To each variable corresponds a data set of size $n = 542$ (the number of towns or villages). Table 3.3 gathers the different variables with their names in the list, a short description and the group to which they belong.

Among these 16 variables, any of them could play the role of the response variable or the

[2]French National Institute of Statistics and Economic Studies
[3]French National Institute of Geographical and Forest Information
[4]French National Research Institute of Science and Technology for Environment and Agriculture

R Names	Description	Group
farmers	Percentage of farmers	employment
tradesmen	Percentage of tradesmen and handicraftsmen	employment
managers	Percentage of managers and executives	employment
workers	Percentage of workers and employees	employment
unemployed	Percentage of unemployed workers	employment
middleemp	Percentage of middle-range employees	employment
retired	Percentage of retired people	employment
employrate	Employment rate	employment
income	Average income	employment
density	Population density	housing
primaryres	Percentage of primary residences	housing
owners	Percentage of home owners living in their primary residence	housing
building	Percentage of buildings	environment
water	Percentage of water	environment
vegetation	Percentage of vegetation	environment
agricul	Percentage of agricultural land	environment

Table 3.3 – Numerical variables of the list `gironde`.

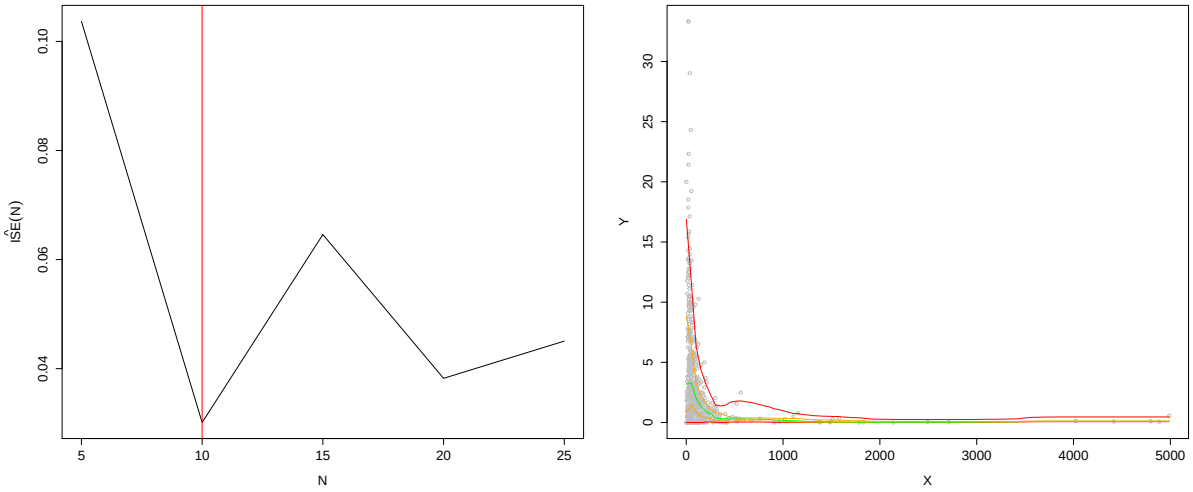
covariate. Therefore, we can consider any of them given another one, and observe if a possible link is highlighted. For example, we choose as Y the percentage of farmers and we consider it given X equal to the population density. We then perform quantization-based quantile regression using the `QuantifQuantile` function. We only change the `testN` parameter. As explained in Section 3.5, the `plot` function (with `ise` set to `TRUE`) allows to check the choice of `testN` (see Figure 3.20a) and to obtain directly conditional quantile curves (see Figure 3.20b).

```
R> set.seed(564346)
R> X <- gironde[[2]]$density
R> Y <- gironde[[1]]$farmers
R> testN <- seq(5, 25, by = 5)
R> res <- QuantifQuantile(X, Y, testN = testN)
R> col.plot <- c("grey", "red", "orange", "green", "orange", "red")
R> plot(res, ise = TRUE, col.plot = col.plot)
```

We observe that the percentage of farmers is inversely proportional with the population density, which was expected.

We can perform this method with any choice of X and Y among the list. The different functions of the package allow to massively generate reference curves for a variable given another one of interest. We only have to focus on the rightness of the parameter `testN` that might be adapted if a warning message is printed (and checked with `plot` with `ise=TRUE`).

Figure 3.21 gathers different types of reference curves that we obtained. We do not represent here the illustration of the selection criterion but only the resulting curves. Each subfigure was



(a) (b)

Figure 3.20 – Illustration of the plots provided by the `plot` function of the package (with `ise=TRUE`) for the particular choice X = farmers and Y = density.

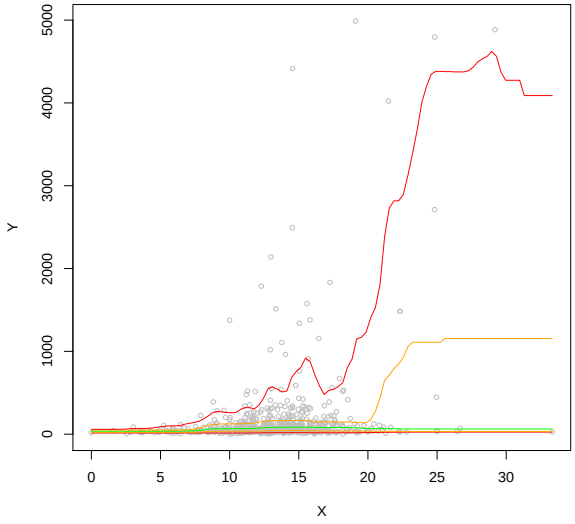
obtained by selecting N with respect to α between 5 and 45.

As mentioned in Section 3.5, the `QuantifQuantile.d` function allows to perform quantile regression with $d > 2$. To illustrate it, we consider the response $Y =$ `population density` and the three covariates $X_1 =$ `percentage of unemployed workers`, $X_2 =$ `percentage of farmers`, and $X_3 =$ `percentage of managers`. In this setup, no graphical output is available. We therefore restrict to a finite collection of x-values where conditional quantiles are to be estimated. Denoting by M_j and $\overline{X}_j$, $j = 1, 2, 3$, the maximal value and the average of X_{ij}, $i = 1, \ldots, n = 542$, respectively, we consider the following eight values of x:

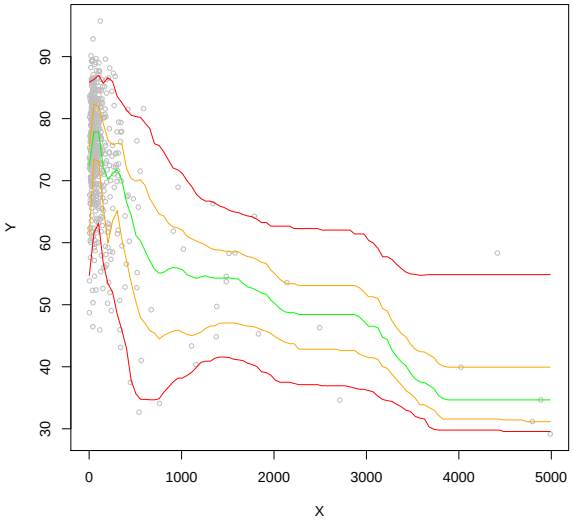
$$x_1 = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix}, \; x_2 = \begin{pmatrix} \frac{1}{2}(\overline{X}_1 + M_1) \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix}, \; x_3 = \begin{pmatrix} \overline{X}_1 \\ \frac{1}{2}(\overline{X}_2 + M_2) \\ \overline{X}_3 \end{pmatrix}, \; x_4 = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \frac{1}{2}(\overline{X}_3 + M_3) \end{pmatrix},$$

$$x_5 = \begin{pmatrix} \frac{1}{2}(\overline{X}_1 + M_1) \\ \frac{1}{2}(\overline{X}_2 + M_2) \\ \overline{X}_3 \end{pmatrix}, \; x_6 = \begin{pmatrix} \frac{1}{2}(\overline{X}_1 + M_1) \\ \overline{X}_2 \\ \frac{1}{2}(\overline{X}_3 + M_3) \end{pmatrix}, \; x_7 = \begin{pmatrix} \overline{X}_1 \\ \frac{1}{2}(\overline{X}_2 + M_2) \\ \frac{1}{2}(\overline{X}_3 + M_3) \end{pmatrix}, \; x_8 = \begin{pmatrix} \frac{1}{2}(\overline{X}_1 + M_1) \\ \frac{1}{2}(\overline{X}_2 + M_2) \\ \frac{1}{2}(\overline{X}_3 + M_3) \end{pmatrix}.$$

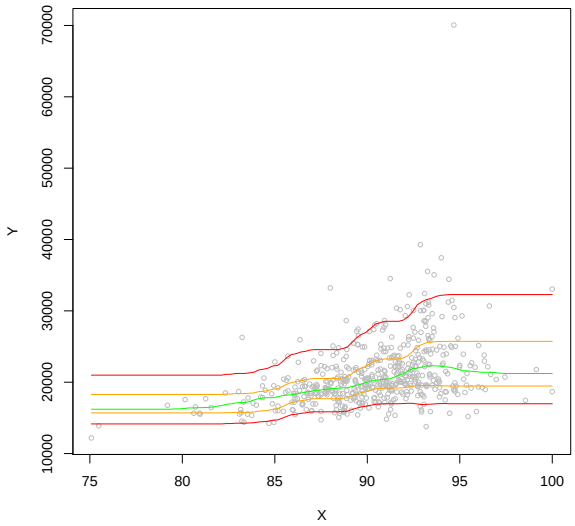
The function `QuantifQuantile.d` is then evaluated for the response and covariates indicated above, and with the arguments `alpha` $= (0.25, 0.5, 0.75)'$, `testN` $= (5, 6, 7, 8, 9, 10)'$, `x` being the 3×8 matrix whose columns are the vectors $x_1, x_2, \ldots, x_8$ just defined and `ncores` being the number of cores detected by R minus 1. This provided $\hat{N}^{-}_{\alpha;\text{opt}} = 8$, 8 and 7, for $\alpha = 0.25$, 0.50 and 0.75, respectively. The total computation time is 6.84 seconds. The `fitted.values` function then allowed to return the following matrix `hatq_opt` of estimated conditional quantiles:



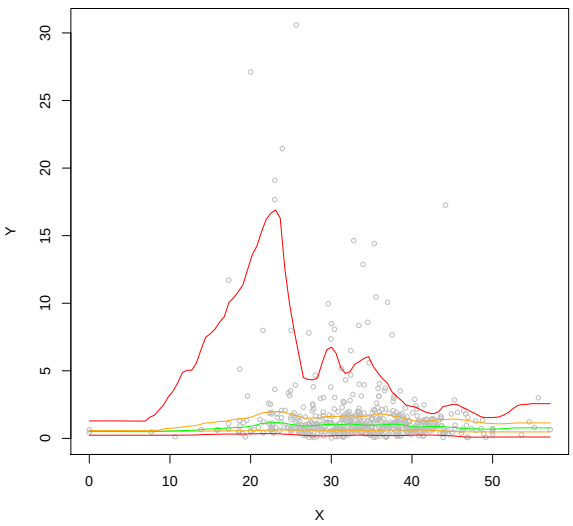
(a) Density given unemployed



(b) Building given workers



(c) Income given employrate



(d) Owners given density

Figure 3.21 – Conditional quantile curves $x \mapsto \bar{q}^{N,n}_{\alpha,B}(x)$ for different choices of X and Y among the list `gironde` and $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$.

```
       [,1]   [,2]    [,3]    [,4]   [,5]   [,6]    [,7]   [,8]
[1,]  43.39  22.22   39.19   71.88  25.19  22.33   66.14  24.45
[2,]  78.48  30.01   78.23  156.90  34.32  30.45  151.50  33.74
[3,] 142.30  45.66  306.33  321.66  53.26  46.17  450.17  53.00
```

This collection of (estimated) conditional quantiles allows to appreciate the impact of a marginal perturbation of the covariates on Y's conditional median (location) or interquartile range (scale). For instance, the results suggest that Y's conditional median decreases with X_1, is stable with X_2, and increases with X_3, whereas its conditional interquartile range decreases with X_1 but increases much with X_2 and with X_3. The eight x-values considered further allow to look at the joint impact of two or three covariates on Y's conditional location and scale. Of course, other shifts in the covariates (and other orders α) should further be considered to fully appreciate the dependence of Y on X.

3.7 Final comments

In this chapter, we investigated the empirical performances of the quantization-based estimators of conditional quantile that were recently introduced in Chapter 2. This led us to design an efficient data-driven method to select the number N of quantizers to be used, that plays the role of a smoothing parameter (Section 3.2). From extensive simulations, we concluded that quantization-based estimators compete well with alternative conditional quantile estimators and sometimes dominate its main competitor, spline estimators, particularly when the link function is complex (Section 3.3). We treated real data examples, in which we showed that the proposed methodology provides very satisfactory conditional quantile curves, and we saw that the good properties of quantization-based estimators extend to the bivariate covariate case (Sections 3.4 and 3.6). This should make conditional quantile estimation based on quantization of interest to practitioners; in this spirit, we wrote an R package, named **QuantifQuantile**, that allows to compute in a straightforward way the proposed estimators (the data-driven selection of N is included) and to plot the resulting quantile curves/surfaces. This package is already available on the CRAN and was described in Section 3.5.

We conclude this chapter with a last remark. Both in Chapter 2 and in this chapter, quantization was applied to the covariate only. One may wonder whether or not it may be of interest to move entirely to a discrete framework by applying quantization to both the covariate and the response. Such a double quantization might involve different numbers of quantizers for X and Y (N_X and N_Y, say) and would lead to the approximated conditional α-quantile

$$\widetilde{q}_\alpha^{Y,N_X,N_Y}(x) = \arg\min_{a\in\mathbb{R}} \mathrm{E}\big[\rho_\alpha(\widetilde{Y}^{N_Y} - a)|\widetilde{X}^{N_X} = \tilde{x}\big],$$

where $\widetilde{Y}^{N_Y} = \mathrm{Proj}_{\delta^{N_Y}}(Y)$ denotes the projection of Y onto $\delta^{N_Y} = (\tilde{y}_1, \ldots, \tilde{y}_{N_Y})$, an optimal N-grid for Y ($\tilde{x}$ still denotes the projection of x onto the optimal N_X-grid for X). If observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are available, then the resulting estimator of $\widetilde{q}_\alpha^{Y,N_X,N_Y}(x)$ is the sample α-quantile, $\widehat{q}_\alpha^{Y,N_X,N_Y,n}(x)$ say, of the $\widehat{Y}_i^{N_Y,n} = \mathrm{Proj}_{\hat{\delta}^{N_Y,n}}(Y_i)$'s corresponding to the indices i for which $\widehat{X}_i^{N,n} = \hat{x}^{N,n}$ (the optimal y-grid $\hat{\delta}^{N_Y,n} = (\hat{y}_1^{N_Y,n}, \ldots, \hat{y}_{N_Y}^{N_Y,n})$ can be obtained

from the CLVQ algorithm). Of course, we can also define a bootstrap version $\bar{q}_{\alpha,B}^{Y,N_X,N_Y,n}(x)$ of this estimator by proceeding as in Section 2.3.2.

To investigate the performance of double quantization, we generated a random sample of size $n = 500$ from the model

$$Y = \frac{1}{5}X^3 + \varepsilon, \tag{3.7.1}$$

where $X \sim U(-3, 3)$ and $\varepsilon \sim \mathcal{N}(0, 1)$ are independent. Figure 3.22 plots, for five values of α, the $\widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N_X, N_Y)$ quantities achieved by the double-quantization estimator $\bar{q}_{\alpha,B}^{Y,N_X,N_Y,n}(x)$ with $B = 50$; we only considered $N_X \in \{5, 10, \ldots, 75, 80\}$ and $N_Y \in \{5, 10, \ldots, 245, 250\}$. In order to get smoother surfaces, we actually averaged sample ISEs over 5 independent random samples. Denote $\hat{N}^{-}_{\alpha,B;\mathrm{opt}}(N_Y)$ the optimal value for N_X at the given N_Y (i.e. the one that minimizes the ISE function $N \mapsto \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(N, N_Y)$. The white curve in each panel is the curve associated with $N_Y \mapsto \widehat{\mathrm{ISE}}^{-}_{\alpha,B}(\hat{N}^{-}_{\alpha,B;\mathrm{opt}}(N_Y), N_Y)$. It is seen that, irrespective of N_Y and α, the optimal number of quantizers for X stays close to 25. However, there is no optimal value for N_Y: the larger N_Y, the smaller the averaged ISE. In conclusion, an optimal choice for (N_X, N_Y) appears to be N_X close to 25 and $N_Y \to \infty$. Since letting the size of the quantization grid go to infinity is equivalent to not quantizing at all, this leads to favoring the solution adopted in Chapters 2 and 3 over double quantization. Quantizing X only also makes perfect sense in the problem of estimating nonparametrically a given conditional quantile $q_\alpha(x)$. Most estimation methods indeed consist in selecting the observations whose X-part is the closest to x (where the meaning of "close" depends on the method at hand) and to take the sample quantile of the corresponding Y's. Quantizing Y thus does not seem natural, as it adds an unnecessary approximation error (see Figure 3.23).

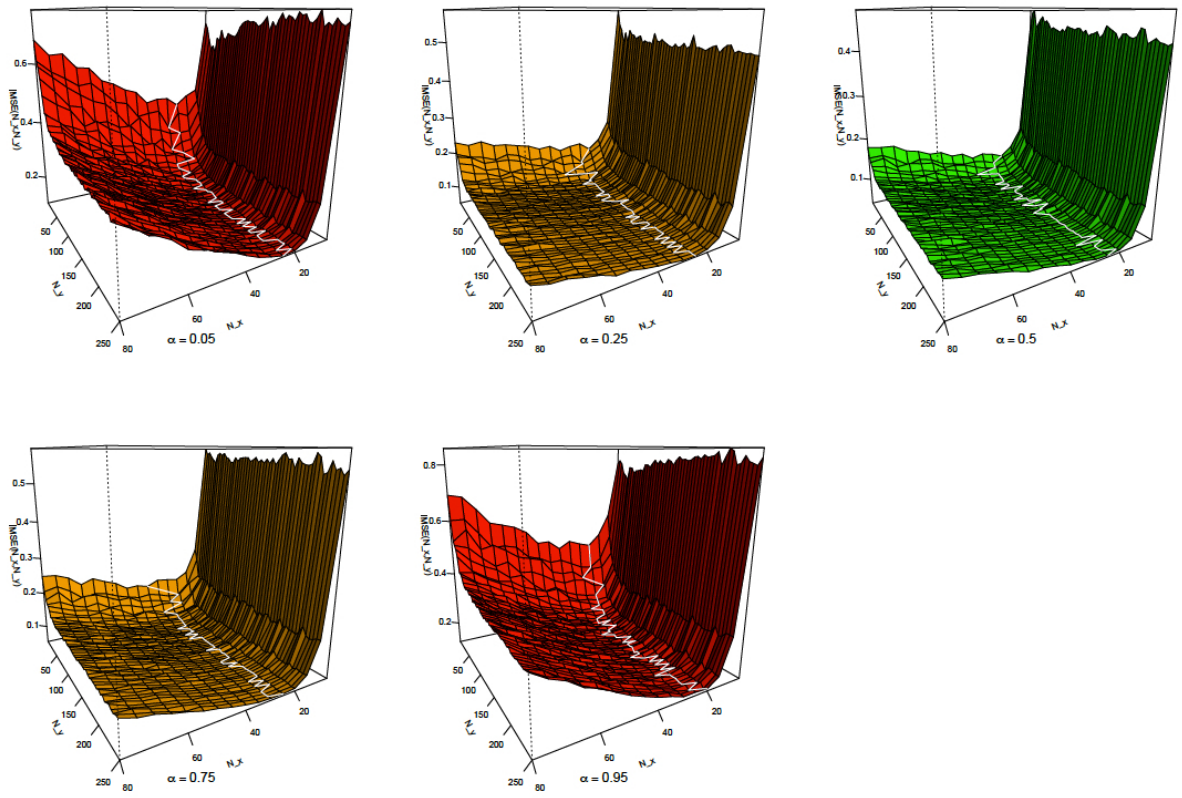


Figure 3.22 – Plots of the mapping $(N_X, N_Y) \mapsto \widehat{\mathrm{ISE}}_{\alpha,B}(N_X, N_Y)$ (for the estimator $\bar{q}_{\alpha,B}^{Y,N_X,N_Y,n}(x)$ with $B = 50$), for $\alpha = 0.05$, 0.25, 0.50, 0.75, and 0.95, averaged over 5 mutually independent random samples from (3.7.1) with sample size $n = 500$.

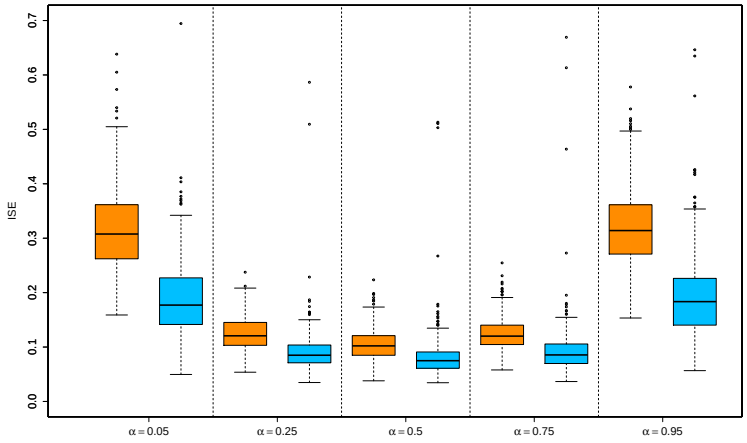


Figure 3.23 – For 500 replications of sample of size $n = 300$ from (3.7.1), the boxplots of the ISE in the estimation of conditional quantiles curves: in orange, $\bar{q}_{\alpha,B}^{Y,N_X,N_Y,n}$ and in blue, $\bar{q}_{\alpha,B}^{N,n}$.


DEFINE
quantile
MULTIVARIATE
MULTIPLE
RESULTS
DEFINITION
estimator
output
CONDITIONAL
empirical
function
DEPTH
selection
regression
SATISFYING
GENERATED
QUANTIZATION
SECTION
GRID
POINTS
REGIONS
LOCAL
HYPERPLANES
PROJECTION
optimal
based
corresponding
setting
method
vector
different
properties
APPROXIMATION
EQUIVARIANT
DATA
VALUE
HALFSPACE
NOTION
SEQUEL
BOOTSTRAP
allows
original
BOUNDARY
CONSTRUCTION
random
DIMENSION
single
CASE
constant
MEASURE
INTERESTING
NOTICE
DENOTE
FIXED
SIZE
NATURAL
FIGURE
approach
existing
ANSWER
assumption
OBSERVATIONS
INNER
POSITIVE
STEP
CONCEPT
COVARIATE
UNIVARIATE
MODEL
USED
DRIVEN
distribution
affine
REQUIRES
EXPECTED
CHOICE
LINK
PLOTS
LIPSCHITZ
FRAMEWORK
probability
consists
compare
competitors
CHARACTERIZED
CONNECTION
PERFORMANCES
proposed
MENTIONED
GRADIENT
EXPLAIN
EXTREMELY
PROCEDURE
CRITERION
EVALUATE
COLLECTION
INTRODUCED
DIMENSIONAL
work
interpretation
parameter
minimization
asymptotic
ORIENTED
SUPPORT
example
analogous
SOURCE
PARTICULARLY
LIBRARIE
CHOOSE
SPATIAL
QUESTIONS
SIMULATIONS
BILINEAR
REASON
INCREASES
independent
PREVIOUS
CONTINUOUS
CENTER
comparison
ALGORITHM
CONTEXT
ADMITS
DERIVED
EUCLIDEAN
ahes
extend
counterpart
ERRORS
RELATION
THEORETICAL
MOMENTS
variables
PRECISELY
REPRESENTATION
ILLUSTRATE
IMPORTANT
GREATER
showed
PROBABILISTIC
POSSIBLE
dependence
information
HAVING
PROBLEM
EFFICIENCY
ENTIRELY
differentiable
alternative
numerical
LINK
AUXILIARY
MOTIVATED
PROVED
ASSOCIATED
INVERTIBLE
CHAPTER
NATURAL
FIGURE
approach
existing
EFFICIENCY
ENTIRELY
differentiable
alternative
numerical
LINK
AUXILIARY
MOTIVATED
PROVED
ASSOCIATED
INVERTIBLE
CHAPTER
HETEROSCEDASTICITY
transformation
SECOND
Introduction
EXTREMELY
EXPLAIN
PROCEDURE
CRITERION
EVALUATE
COLLECTION
INTRODUCED
DIMENSIONAL
work
interpretation
parameter
minimization
asymptotic
ORIENTED
SUPPORT
example
analogous
SOURCE
PARTICULARLY
LIBRARIE
CHOOSE
SPATIAL
QUESTIONS
SIMULATIONS
BILINEAR
REASON
INCREASES
independent
PREVIOUS
CONTINUOUS
CENTER
comparison
ALGORITHM
CONTEXT
ADMITS
DERIVED
EUCLIDEAN
ahes
extend
counterpart
ERRORS
RELATION
THEORETICAL
MOMENTS
variables
PRECISELY
REPRESENTATION
ILLUSTRATE
IMPORTANT
GREATER
showed
PROBABILISTIC
POSSIBLE
dependence
information
HAVING
PROBLEM
EFFICIENCY
ENTIRELY
differentiable
alternative
numerical
LINK
AUXILIARY
MOTIVATED
PROVED
ASSOCIATED
INVERTIBLE
CHAPTER
NATURAL
FIGURE
approach
existing
EFFICIENCY
ENTIRELY
differentiable
alternative
numerical
LINK
AUXILIARY
MOTIVATED
PROVED
ASSOCIATED
INVERTIBLE
CHAPTER

Glossary

Notation	Meaning
$\boldsymbol{Y}$	Random vector of dimension m, $m > 1$.
$\mathcal{S}^{m-1}$, $\mathcal{B}^m$	Unit sphere, open unit ball (deprived of the origin) of $\mathbb{R}^m$
$\lvert\boldsymbol{z}\rvert$, $\lVert\boldsymbol{z}\rVert_p$	Euclidean norm, L_p-norm of $\boldsymbol{z}$
$\boldsymbol{\alpha} = \alpha\boldsymbol{u}$	Order of multivariate quantile, $\alpha \in (0,1)$, $\boldsymbol{u} \in \mathcal{S}^{m-1}$
$\boldsymbol{\Gamma_u}$	Arbitrary $m \times (m-1)$ matrix such that $(\boldsymbol{u}, \boldsymbol{\Gamma_u})$ is an orthonormal basis of $\mathbb{R}^m$
$(a_{\boldsymbol{\alpha}}, c'_{\boldsymbol{\alpha}})'$	Coefficient of the quantile of $\boldsymbol{Y}$ of order $\boldsymbol{\alpha}$ in the sense of HPŠ10
$\boldsymbol{X}$	Covariate of dimension d
$P_{\boldsymbol{X}}$, $S_{\boldsymbol{X}}$	Distribution of the random vector $\boldsymbol{X}$ and its support
$\begin{pmatrix} a_{\boldsymbol{\alpha}}(\boldsymbol{x}) \\ \boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}) \end{pmatrix}$	Coefficients of the conditional quantile of $\boldsymbol{Y}$ of order $\boldsymbol{\alpha}$ given $\boldsymbol{X} = \boldsymbol{x}$
$\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Equivalent notation for $(a_{\boldsymbol{\alpha}}(\boldsymbol{x}), \boldsymbol{c}'_{\boldsymbol{\alpha}}(\boldsymbol{x}))'$
$\boldsymbol{\gamma}^N$	Optimal quantization grid of size N
$\widetilde{\boldsymbol{X}}^N$	Projection of $\boldsymbol{X}$ onto the quantization grid $\boldsymbol{\gamma}^N$
$\boldsymbol{C_x}$	Quantization cell containing all points projected on the same grid point as $\boldsymbol{x}$
$R(\boldsymbol{C_x})$	Radius of $\boldsymbol{C_x}$, i.e. largest distance between the center and a point of the cell
$\begin{pmatrix} \widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) \\ \widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) \end{pmatrix}$	Approximation of the coefficients of the conditional quantile of $\boldsymbol{Y}\lvert\boldsymbol{X} = \boldsymbol{x}$
$\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Equivalent notation for $(\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}), (\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}))')'$
$\hat{\boldsymbol{\gamma}}^{N,n}$	Approximation of $\boldsymbol{\gamma}^N$ provided by CLVQ after n replications
$\widehat{\boldsymbol{X}}^{N,n}$	Projection of $\boldsymbol{X}$ onto the approximated quantization grid $\hat{\boldsymbol{\gamma}}^{N,n}$
$\begin{pmatrix} \widehat{a}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}) \\ \widehat{\boldsymbol{c}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}) \end{pmatrix}$	Estimation of the coefficients of the conditional quantile of $\boldsymbol{Y}\lvert\boldsymbol{X} = \boldsymbol{x}$
$\widehat{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Equivalent notation for $(\widehat{a}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}), (\widehat{\boldsymbol{c}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}))')'$
$\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Bootstrap-based estimator of $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$

$\widehat{\mathbf{ISE}}^{-}_{\alpha,B}(N)$	Data-driven criterion to select N based on $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x})$
$\widehat{\boldsymbol{q}}^{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Local constant estimator of $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$
$\widehat{\boldsymbol{q}}^{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{x})$	Local bilinear estimator of $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$

4 Introduction

Contents

The first part of this thesis was dedicated to the framework of a unidimensional response variable Y, while the covariate X could be multidimensional. It is then natural to be interested in a method allowing to characterize the link between X and Y when Y is of dimension m greater than 1. Since Section 1.1 and Figure 1.3 convince us of the advantages of quantile regression over standard mean regression, we then aim to generalize quantization-based quantile regression to multiple-output (i.e. to $m > 1$).

However, generalizing the concept of quantile in a multivariate setting is not direct due to the lack of natural ordering of Euclidean space for dimension greater than one. Section 4.1 is thus devoted to the definitions and comparison of the different existing notions, in order to choose one of them in our forthcoming investigation. With the choice of Section 4.1, Section 4.2 aims to define precisely the notion of conditional quantile in a multiple-output setting.

For the sake of clarity of notation, multivariate variables will be represented with bold font in this part.

4.1 Multivariate quantiles

Many approaches of multivariate quantiles have been developed and an attractive comparison of them has been proposed in Serfling (2002). Other interesting multivariate quantiles having a connection with the notion of statistical depth have been defined since then. The aim of this section is to complete the comparison study of Serfling (2002) with some recent concepts of multivariate quantiles and thus to provide a detailed overview of existing definitions of multivariate quantiles.

Let us start with recalling some important properties and features of quantiles (thus in a univariate context). We saw in Section 1.1 that the quantiles of order α of a random variable Y can be defined equivalently by inverting the cumulative distribution function F (Definition 1.1.1) or by solving an optimization problem (Definition 1.1.2).

The notion of *quantile function* is also of major importance. We define a quantile function associated to F as

$$\begin{array}{rccl} Q(\cdot, F): & (-1,1) & \to & \mathbb{R} \\ & u & \mapsto & Q(u,F) = F^{-1}\left(\frac{u+1}{2}\right). \end{array} \tag{4.1.1}$$

We notice that $Q(u, F)$ corresponds to the quantile of order $\alpha = (u+1)/2$. Therefore, a value of u near to zero provides a central quantile when a large value of u (i.e. near to 1 in absolute value) indicates an extreme quantile. Moreover, u gives two different information: a direction (positive or negative) and a magnitude (between 0 and 1).

We now state some crucial properties of univariate quantiles. First, the quantile function is affine equivariant, that is

$$Q(v, F_{aY+b}) = aQ(u, F_Y) + b,$$

where $-1 < u < 1$, $a, b \in \mathbb{R}$ and $v = v(u, a, b, F_Y)$ such that $|v| = |u|$. We can then take $v = \text{sgn}(a)u$ or $v = -\text{sgn}(a)u$.

In addition, the resulting contours satisfy some nice geometric properties as nestedness and convexity. But a main characteristic of quantiles is their relation to depth. A depth function is a function providing a center-outward ordering of points, with respect to the distribution F. In other words, it is a measure of centrality valid in any dimension. A large depth corresponds with a "central" point while a depth near to zero indicates a point distant from the center. For more information about depth, see for instance Rousseeuw and Ruts (1999); Zuo and Serfling (2000). In this univariate setting, we define the depth of a point x as $\min(F(x), 1 - F(x))$. Therefore, the only points with depth α are $x_\alpha := F^{-1}(\alpha)$ and $x_{1-\alpha} := F^{-1}(1-\alpha)$, i.e. the quantiles of order α and $1-\alpha$ respectively. Depth is an interesting concept for its geometrical considerations for instance while the quantile approach provides analytical and algorithmic tools. A connection between these two concepts allows then to transfer at each the advantages of the other. For this reason, such a connection for multivariate quantiles is highly desirable.

As we have seen in (1.1.1), the fundamental concept of quantiles relies on the natural ordering of $\mathbb{R}$. For this reason, it cannot be easily generalized in a multivariate setting since the Euclidean space for greater dimension suffers from the lack of a natural ordering.

A huge literature has been devoted to the problem of extending this one-dimensional concept to a multivariate setting. Different approaches were investigated. A possibility is starting from a

property of one-dimensional quantiles and attempting to generalize it. The following step consists then in evaluating how well the other properties expected from a quantile are satisfied.

For this purpose, we start in Section 4.1.1 with listing some questions to be asked when we regard a new definition of multivariate quantiles and proposed by Serfling (2002). We explain a mathematical tool introduced in this paper to evaluate and compare different notions of multivariate quantiles. Therefore, we will answer these questions in the sequel for each notion of multivariate quantiles.

A first possible approach consists in generalizing (1.1.1) by defining multivariate quantiles as inversion of mappings (see for instance Koltchinskii (1997)). We can also define them from a norm minimization as an extension of (1.1.2) (Abdous and Theodorescu, 1992; Chaudhuri, 1996). These constructions are investigated in Section 4.1.2. However, as we will see in the sequel, these multivariate quantiles do not satisfy crucial properties as affine equivariance and have no link with a depth function. We then define in Section 4.1.3 multivariate quantiles based on some depth function. Another step in this direction was realized by Kong and Mizera (2012) who defined multivariate quantiles by applying quantile to some univariate functions, the more natural choice being projections. We see in Section 4.1.4 that their multivariate quantiles have an interesting link with halfspace depth but unfortunately the resulting contours do not enjoy properties as nestedness or affine equivariance. A solution was proposed by Hallin, Paindaveine, and Šiman (2010a) (hereafter, HPŠ10) where they define multivariate quantiles as hyperplanes. This vision of multivariate quantiles solves the problems arisen with the concept of Kong and Mizera (2012) and the corresponding multivariate quantile satisfies all desirable properties of a quantile. Moreover, the inner region characterized by the collection of such hyperplanes coincides with halfspace depth regions, as we explain in Section 4.1.5.

4.1.1 Criterion for comparing different versions of multivariate quantiles

In this section, we state the questions that we will investigate for each definition of multivariate quantiles, and explain the criterion developed by Serfling (2002) to compare and evaluate the different existing definitions. When considering a multivariate quantile, we will ask the following questions.

Q1 Does it support probabilistic interpretation analogous to the univariate case ?

Q2 Does it satisfy suitable equivariance properties, such as affine equivariance ?

Q3 Does its empirical version support asymptotic results, as strong consistency, asymptotic normality, Bahadur-type representation ?

Q4 Does it provide a connection to the concept of depth ?

We focus here on these questions, but other features could also be considered (see Serfling (2002)), as the possible formulation of multivariate notions of nonparametric descriptive statistics for location or scale, as the median, trimmed means, IQR (InterQuartile Range), etc. The computational feasibility is also of considerable practical importance.

Serfling (2002) proposes to use a *median-oriented quantile function* to compare various definitions of quantiles. Let us first introduce it in the univariate setting and we will then explain how to extend it to a multivariate context.

A median-oriented quantile function is a function $\mathcal{Q}(u,\alpha)$ defined by

$$\begin{cases} \mathcal{Q}(u,0) = M, \text{ for } u = \pm 1, \\ \mathcal{Q}(-1,\alpha) = F^{-1}\left(\frac{1-\alpha}{2}\right), \\ \mathcal{Q}(1,\alpha) = F^{-1}\left(1-\frac{1-\alpha}{2}\right), \end{cases} \tag{4.1.2}$$

where $\alpha \in (0,1)$ and $M = F^{-1}(1/2)$ is the median. This definition is motivated by the fact that $F^{-1}\big((1-\alpha)/2\big)$ and $F^{-1}\big(1-(1-\alpha)/2\big)$ correspond to the points with lower and upper tail regions of probability $(1-\alpha)/2$. The closed interval $\big[F^{-1}\big((1-\alpha)/2\big), F^{-1}\big(1-(1-\alpha)/2\big)\big]$ has therefore a probability α and is called *median-oriented* αth quantile inner region having probability α. In the particular case $\alpha = 1/2$, this interval is the interquartile region and when $\alpha \to 0$, it reduces to the median. Actually, each point $x \in \mathbb{R}$ can be characterized by a value of α, determining the "outlyingness" of x. Therefore, these αth quantile inner regions entirely represent the quantile information about F, for $\alpha \in (0,1)$. This quantile information is then expressed in terms of this median-oriented quantile function.

The median-oriented quantile function $\mathcal{Q}(\cdot,\cdot)$ satisfies the following properties.

P1 For each $\alpha \in [0,1)$, the set $\big\{\mathcal{Q}(u,t) : 0 \le t \le \alpha, u \in \{-1,1\}\big\}$ comprises a αth quantile inner region with boundary points $\mathcal{Q}(u,\alpha)$ and with $\mathcal{Q}(u,0) \equiv M$.

P2 For each fixed direction u from M, $|\mathcal{Q}(u,\alpha) - M|$ increases with α, $0 \le \alpha < 1$.

P3 Considered as sets, the αth quantile inner regions $\big\{\mathcal{Q}(u,t) : 0 \le t \le \alpha, u \in \{-1,1\}\big\}$ have suitable structure and interpretations.

We notice that the name *median-oriented quantile function* comes from the fact that there is a correspondence with the quantile function $Q(\cdot,F)$, with the difference here that it is defined with respect to the median, whence median-oriented.

It is then desirable to define in $\mathbb{R}^m$ the notion of median-oriented quantile functions $\mathcal{Q}(u,\alpha)$ having probabilistic interpretations, satisfying directional monotonicity and having suitable set-theoretic interpretations, i.e. satisfying multivariate extensions of P1–P3 respectively. Now, $\boldsymbol{M}$ denotes any multivariate median and $\boldsymbol{u} \in \mathcal{S}^{m-1}(\boldsymbol{M})$, since we can see $\{-1,1\}$ as the sphere of null dimension. Let us introduce a family $\mathcal{A} = \{\boldsymbol{A}_\gamma, 0 < \gamma < \infty\}$ of regions nested around $\boldsymbol{M}$, with $\boldsymbol{A}_0 = \{\boldsymbol{M}\}$ and $\boldsymbol{A}_\gamma \subset \boldsymbol{A}_{\gamma'}$ for $0 < \gamma < \gamma'$. For any $\alpha \in (0,1)$, define $\gamma_\alpha = \inf\{\gamma : P(\boldsymbol{A}_\gamma) > \alpha\}$. Then, $\mathcal{Q}(\boldsymbol{u},\alpha)$ is given by the boundary point of $\boldsymbol{A}_{\gamma_\alpha}$ in the direction $\boldsymbol{u}$ from $\boldsymbol{M}$. This quantile function is then defined in order to satisfy P1 and P2 but P3 depends crucially on the choice of the family $\mathcal{A}$.

We now consider different multivariate quantiles and try to construct such a family satisfying P1 and P2, and then check if it satisfies P3. It helps answer Q1. We also investigate the three other questions on affine equivariance property, asymptotic results for the empirical version, and relation with any concept of depth.

In the sequel, $\boldsymbol{Y}$ denotes a random vector of $\mathbb{R}^m$ with F the corresponding probability measure. Let $\boldsymbol{Y}^{(n)} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a n-tuple of m-dimensional random vectors with $m < n$. The usual empirical measure is denoted $F^{(n)}$, corresponding to mass $1/n$ at each $\boldsymbol{Y}_i$. We will also write $\mathcal{S}^{m-1}(\boldsymbol{r})$ and $\mathcal{B}^m(\boldsymbol{r})$ for the unit sphere and open unit ball centered at $\boldsymbol{r} \in \mathbb{R}^m$ deprived of its center respectively. We will omit to precise the center when $\boldsymbol{r}$ is the origin.

4.1.2 Multivariate quantiles based on norm minimization

The aim of this section is to examine two possible extensions of (1.1.2) based on norm minimization.

A first approach is proposed in Abdous and Theodorescu (1992). For $1 \leq q \leq \infty$ and $0 < \alpha < 1$, they define the norm-like functions

$$\|\boldsymbol{y}\|_{q,\alpha} = \|(y_1, \ldots, y_d)\|_{q,\alpha} = \left\| \frac{|y_1| + (2\alpha - 1)y_1}{2}, \ldots, \frac{|y_d| + (2\alpha - 1)y_d}{2} \right\|_q,$$

where $\|\cdot\|_q$ denotes the usual L_q-norm on $\mathbb{R}^m$. Notice that this function is obtained by applying the check function ρ_α to each component, and then taking the L_q-norm.

Definition 4.1.1. *The multivariate αth quantile of $\boldsymbol{Y} \in \mathbb{R}^m$, $\alpha \in (0,1)$, is defined as*

$$\boldsymbol{Q}_{AT}(q, \alpha) \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^m} \mathrm{E}\big[\|\boldsymbol{Y} - \boldsymbol{\theta}\|_{q,\alpha} - \|\boldsymbol{Y}\|_{q,\alpha}\big]. \tag{4.1.3}$$

Notice that the term $\|\boldsymbol{Y}\|_{q,\alpha}$ is only here to circumvent the need of first-order finite moment for $\boldsymbol{Y}$. These quantiles are not rotationally equivariant and answer no to the third question of Section 4.1.1.

We aim now to associate to these quantiles a quantile function $\mathcal{Q}(\boldsymbol{u}, \alpha)$ in the sense of Section 4.1.1. Serfling (2002) indicated that we cannot find a suitable family of nested inner regions. This notion fails then to conform any satisfactory definition of median-oriented quantile function and it fails to answer positively to Q1. We will then not investigate any longer this definition of multivariate quantile.

Another approach is due to Chaudhuri (1996). Setting $u = 2\alpha - 1$ allows a re-indexing of the univariate quantiles by $u \in (-1, 1)$ as in (4.1.1). The m-dimensional quantiles are then obtained by extending the set $(-1, 1)$ to $\mathcal{B}^m \cup \{\boldsymbol{0}\}$.

Definition 4.1.2. *The $\boldsymbol{u}$th quantile $\boldsymbol{Q}_C(\boldsymbol{u})$ of $\boldsymbol{Y} \in \mathbb{R}^m$, for $\boldsymbol{u} \in \mathcal{B}^m$, is defined as*

$$\boldsymbol{Q}_C(\boldsymbol{u}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^m} \mathrm{E}\big[\Phi(\boldsymbol{u}, \boldsymbol{Y} - \boldsymbol{\theta}) - \Phi(\boldsymbol{u}, \boldsymbol{Y})\big], \tag{4.1.4}$$

where $\Phi(\boldsymbol{u}, \boldsymbol{t}) = |\boldsymbol{t}| + < \boldsymbol{u}, \boldsymbol{t} >$, with $|\cdot|$ denoting the usual Euclidean norm and $< \cdot, \cdot >$ the usual Euclidean inner product. They are called spatial quantiles.

This Φ function thus generalizes the check function $\rho_\alpha(z) = \big(|z| + (2\alpha - 1)z\big)/2$ (up to a factor 2). Indeed, the Euclidean norm and the Euclidean inner product are natural generalizations of the absolute value and of the product respectively, with $u = 2\alpha - 1$.

As for the Abdous and Theodorescu (1992) approach, the term $\Phi(\boldsymbol{u}, \boldsymbol{Y})$ allows not to assume finite first-order moment. When the expectation of $\boldsymbol{Y}$ is finite, we simply look for a minimizer of $\mathrm{E}\big[\Phi(\boldsymbol{u}, \boldsymbol{Y} - \boldsymbol{\theta})\big]$. Unicity of this minimizer was proven under the assumption that the distribution of $\boldsymbol{Y}$ is not supported on a straight line in $\mathbb{R}^m$ (see Chaudhuri (1996) for a proof).

We define in an analogous way the empirical version of these spatial quantiles.

Definition 4.1.3. *The empirical $\boldsymbol{u}$th quantile of $\boldsymbol{Y}^{(n)}$ is defined as*

$$\boldsymbol{Q}_C^{(n)}(\boldsymbol{u}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^m} \sum_{i=1}^{n} \Phi(\boldsymbol{u}, \boldsymbol{Y}_i - \boldsymbol{\theta}). \tag{4.1.5}$$

The following assumption is necessary to obtain asymptotic results for the empirical spatial quantiles.

Assumption (D_n). The observations $\boldsymbol{Y}_i$, $i = 1, \ldots, n$, are i.i.d and the common law is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$. The density f is bounded on every bounded subsets of $\mathbb{R}^m$.

Under Assumption (D_n), Chaudhuri (1996) proved consistency and asymptotic normality of $\boldsymbol{Q}_{\mathrm{C}}^{(n)}(\boldsymbol{u})$, and derived a Bahadur-type representation. These spatial quantiles then answer yes to Q3.

These quantiles are not affine equivariant in general but are rotationally and location equivariant. It is already an important improvement in comparison with the quantiles defined by Abdous and Theodorescu (1992). However, this can be fixed thanks to some "Transformation-Retransformation" tool. It thus answers only partially to Q2. Let us explain briefly how works this Transformation-Retransformation tool (TR). More details can be found in Chakraborty (2001).

Let $\beta = \{i_0, \ldots, i_m\}$ a subset of size $m+1$ of $\{1, \ldots, n\}$. The m-dimensional data points $\boldsymbol{Y}_{i_0}, \ldots, \boldsymbol{Y}_{i_m}$ form a *data-driven coordinate system*, where $\boldsymbol{Y}_{i_0}$ is the origin and $\boldsymbol{Y}_{i_1}, \ldots, \boldsymbol{Y}_{i_m}$ the coordinate axes. We define the transformation matrix as the $m \times m$ matrix $\boldsymbol{Y}(\beta)$ whose columns are the differences $\boldsymbol{Y}_{i_j} - \boldsymbol{Y}_{i_0}$, $j = 1, \ldots, m$. This transformation matrix is used to rewrite the remaining data points in terms of the new coordinate system as follows:

$$\boldsymbol{Z}_j^{(\beta)} := \big(\boldsymbol{Y}(\beta)\big)^{-1} \boldsymbol{Y}_j,$$

for all $j \in \{1, \ldots, n\} \setminus \beta$. Notice that, under Assumption (D_n), the matrix $\boldsymbol{Y}(\beta)$ is invertible with probability one. This is the Transformation step. Moreover, a transformation of the data requires a transformation of the indexing vector $\boldsymbol{u}$ as

$$\boldsymbol{v}(\beta) = \begin{cases} \dfrac{|\boldsymbol{u}|}{\left|\big(\boldsymbol{Y}(\beta)\big)^{-1}\boldsymbol{u}\right|} \big(\boldsymbol{Y}(\beta)\big)^{-1}\boldsymbol{u}, & \boldsymbol{u} \neq 0, \\ 0, & \boldsymbol{u} = 0, \end{cases}$$

Then, we write the empirical spatial quantile of order $\boldsymbol{v}(\beta)$ as $\boldsymbol{R}_\beta^{(n)}(\boldsymbol{v})$, based on $\boldsymbol{Z}_j^{(\beta)}$, $j \in \{1, \ldots, n\} \setminus \beta$. Finally, we come back in the original system with a Retransformation step, by defining the empirical TR spatial quantile of order $\boldsymbol{u}$ as

$$\boldsymbol{Q}_{C;\beta}^{(n)}(\boldsymbol{u}) = \boldsymbol{Y}(\beta)\boldsymbol{R}_\beta^{(n)}(\boldsymbol{v}).$$

It is proved in Chakraborty (2001) that the empirical TR spatial quantile $\boldsymbol{Q}_{C;\beta}^{(n)}(\boldsymbol{u})$ is affine equivariant. The choice of the index family β is of course an important issue, and Chakraborty (2001) suggests to take β such that $\big(\boldsymbol{Y}(\beta)\big)'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}(\beta)$ is the closest to a diagonal matrix of the form $\lambda \boldsymbol{I}_d$, where $\boldsymbol{\Sigma}$ is the variance covariance matrix if the underlying distribution admits second moments.

We now construct a quantile function as above. We start with $\mathcal{Q}(\boldsymbol{u}, 0) = \boldsymbol{Q}_{\mathrm{C}}(0) \equiv \boldsymbol{M}$, and naturally associated nested median-oriented inner regions are

$$\boldsymbol{B}_t = \{\boldsymbol{Q}_{\mathrm{C}}(\boldsymbol{u}') : |\boldsymbol{u}'| < t\},$$

for $0 \leq t < 1$. Let $t_\alpha = \inf\{t : P(\boldsymbol{B}_t) \geq \alpha\}$. We define a median-oriented function $\mathcal{Q}(\boldsymbol{u}, \alpha)$ by associating the point of the boundary of $\boldsymbol{B}_{t_\alpha}$ in the direction $\boldsymbol{u}$ from $\boldsymbol{M}$. Notice that the points of this boundary are those for which $|\boldsymbol{u}'| = t_\alpha$. By construction, $\mathcal{Q}(\boldsymbol{u}, \alpha)$ satisfies P1 and P2. However, the link between $\boldsymbol{u}'$ ($\in \mathcal{B}^m \cup \{\boldsymbol{0}\}$) defining $\boldsymbol{Q}_{\mathrm{C}}(\boldsymbol{u}')$ and $(\boldsymbol{u}, \alpha) \in \mathcal{S}^{m-1} \times (0, 1)$ for which $\boldsymbol{Q}_{\mathrm{C}}(\boldsymbol{u}') = \mathcal{Q}(\boldsymbol{u}, \alpha)$ is not clear and P3 is not well satisfied. The probabilistic interpretation is then not so obvious, and the answer to Q1 is no.

In addition, these multivariate quantiles provide no connection to the concept of depth, failing to satisfy Q4.

To sum up, this definition of multivariate quantiles constructed by generalizing the optimization problem (1.1.2) is a really interesting concept but unfortunately fails to satisfy several properties expected from a quantile.

We conclude this section with a link with quantiles as inversion of mapping. Indeed, Koltchinskii (1997) suggests to extend the notion of distribution function (of a probability measure) to a multivariate setting in such a way that the corresponding inverse function is viewed as a multivariate quantile. As an example, one can prove that $\boldsymbol{Q}_{\mathrm{C}}(\boldsymbol{u})$ is the unique solution $\boldsymbol{\theta}$ of the equation

$$\boldsymbol{u} = -\mathrm{E}\left[\frac{\boldsymbol{Y} - \boldsymbol{\theta}}{|\boldsymbol{Y} - \boldsymbol{\theta}|}\right].$$

Consequently, $\boldsymbol{Q}_{\mathrm{C}}(\boldsymbol{u})$ is obtained by inverting the mapping $\boldsymbol{\theta} \mapsto G_F(\boldsymbol{\theta}) := -\mathrm{E}\left[(\boldsymbol{Y} - \boldsymbol{\theta})/|\boldsymbol{Y} - \boldsymbol{\theta}|\right]$ from $\mathbb{R}^m$ to $\mathbb{R}^m$. Of course, different choices of G_F may lead to different notions of quantiles. We do not investigate it any further but we refer the reader to Koltchinskii (1997).

4.1.3 Multivariate quantiles functions based on depth functions

Let us denote by $D(\boldsymbol{y}, F)$ a nonnegative real-valued depth function on $\mathbb{R}^m$. This depth function provides a center-outward ordering based on F. This center is interpreted as a multivariate median. The reader can refer to Rousseeuw and Ruts (1999) or Zuo and Serfling (2000) for a precise definition and some examples of depth function. The definition of such depth-based multivariate quantiles is explained in Liu et al. (1999) and Serfling (2002). Notice that these quantiles answer then yes to Q4.

For any such depth function, we define the corresponding *β-depth inner region* by $\boldsymbol{I}(\beta, D, F) = \{\boldsymbol{y} : D(\boldsymbol{y}, F) \geq \beta\}$, with $\beta > 0$. Notice that $\boldsymbol{I}(0, D, F) = \mathbb{R}^m$.

For the particular case of the halfspace depth, we have

$$HD(\boldsymbol{y}, F) = \inf\{P(\boldsymbol{H}) : \boldsymbol{H} \text{ a closed halfspace}, \boldsymbol{y} \in \boldsymbol{H}\}, \boldsymbol{y} \in \mathbb{R}^m,$$

and the identity

$$\boldsymbol{I}(\beta, HD, F) = \bigcap\{\boldsymbol{H} : \boldsymbol{H} \text{ a closed halfspace}, P(\boldsymbol{H}) > 1 - \beta\}.$$

Denote by $\beta_\alpha = \sup\{\beta : P(\boldsymbol{I}(\beta, D, F)) \geq \alpha\}$ the largest boundary depth associated with inner region of probability at least α. Since the inner regions decrease as β increases, a smallest inner region having probability at least α exists and is $\boldsymbol{C}(\alpha, D, F) := \boldsymbol{I}(\beta_\alpha, D, F)$. We call it the *$\alpha$th central region*. The boundary $\partial\boldsymbol{I}(\beta, D, F)$ of the β-depth inner region is called *β-depth contour*.

Definition 4.1.4. *The αth quantile surface, with $0 < \alpha < 1$, is defined as*

$$\boldsymbol{Q}(\alpha, D, F) = \partial \boldsymbol{I}(\beta_\alpha, D, F) = \partial \boldsymbol{C}(\alpha, D, F), \tag{4.1.6}$$

where D is any depth function based on F.

Notice that this notion of quantile can be called center-outward, since these surfaces become larger as α increases, containing all the center.

We define analogously the empirical version of these quantile surfaces. It suffices to replace F and D by their empirical versions $F^{(n)}$ and $D^{(n)}$. We then get directly empirical versions of β-depth inner region, αth central region and αth depth contour, denoted $\boldsymbol{I}(\beta, D^{(n)}, F^{(n)})$, $\boldsymbol{C}(\alpha, D^{(n)}, F^{(n)})$ and $\partial\boldsymbol{C}(\alpha, D^{(n)}, F^{(n)})$.

Definition 4.1.5. *The empirical αth quantile surface, with $0 < \alpha < 1$, is defined as*

$$\boldsymbol{Q}^{(n)}(\alpha, D^{(n)}, F^{(n)}) = \partial \boldsymbol{I}(\beta_\alpha, D^{(n)}, F^{(n)}) = \partial \boldsymbol{C}(\alpha, D^{(n)}, F^{(n)}). \tag{4.1.7}$$

The β_α-depth contour is then used to construct a depth-based median-oriented quantile function. Now indexing each point of $\boldsymbol{Q}(\alpha, D, F)$ by α and its direction $\boldsymbol{u} \in \mathcal{S}^{m-1}(\boldsymbol{M})$ allows to define $\mathcal{Q}(\boldsymbol{u}, \alpha)$, $\boldsymbol{u} \in \mathcal{S}^{m-1}(\boldsymbol{M})$ satisfying P1 and P2 of Section 4.1.1. Moreover, the depth at $\boldsymbol{x}$ decreases monotonically in any direction from $\boldsymbol{M}$. Therefore, the αth quantile inner region can be interpreted as connected regions of higher depth, whose interior points having depth larger than β_α and boundary points having common depth β_α. In other words, P3 is well satisfied and Q1 is positively answered.

The usual results of depth function imply that the αth quantile inner regions are nested, connected, compact and affine equivariant. These multivariate quantiles are then affine equivariant and the answer to Q2 is yes.

However, characterizing the asymptotic behavior of sample versions is also of major interest but the asymptotic distribution theory remains unexplored, excepted for halfspace depth. The answer to Q3 is rather no.

Moreover, these multivariate quantiles are defined from depth regions and not from some analytical characterization as in (1.1.1) or (1.1.2). Therefore, they do not inherit the analytical and algorithmic tools of the quantile approach.

4.1.4 Multivariate quantiles as projection quantiles

Kong and Mizera (2012) introduced a new quantile function by using directional quantiles. As we will see, they obtained some link with halfspace depth. Let $\boldsymbol{Y}$ be a random vector of $\mathbb{R}^m$. Recall that we denote by $q(\alpha, Z)$ the univariate quantile of order α of the random variable Z.

Definition 4.1.6. *The directional quantile of Kong and Mizera of $\boldsymbol{Y} \in \mathbb{R}^m$ of order $\alpha \in (0, 1)$ and of direction $\boldsymbol{u} \in \mathcal{S}^{m-1}$ is*

$$\boldsymbol{Q}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F) = q(\alpha, \boldsymbol{u}'\boldsymbol{Y})\boldsymbol{u}. \tag{4.1.8}$$

The idea of this definition is to project the random vector in direction $\boldsymbol{u}$ in order to be able to calculate univariate quantile, and to come back in the m-dimensional setting by multiplying by the direction $\boldsymbol{u}$. We also define quantile hyperplanes.

Definition 4.1.7. *The quantile hyperplane of order* $\boldsymbol{\alpha} = \alpha\boldsymbol{u}$ *is the hyperplane* $\boldsymbol{\pi}_{\mathrm{KM};\boldsymbol{\alpha}}$ *of equation* $\boldsymbol{u}'\boldsymbol{y} = \boldsymbol{u}'\boldsymbol{Q}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F)$, $\boldsymbol{y} \in \mathbb{R}^m$.

They are often referred in the literature as *projection quantiles*, given their construction. We can define empirical versions of these conditional quantiles.

Definition 4.1.8. *The empirical directional quantile of* $\boldsymbol{Y}^{(n)}$ *of Kong and Mizera of order* $\alpha \in (0,1)$ *and of direction* $\boldsymbol{u} \in \mathcal{S}^{m-1}$ *is*

$$\boldsymbol{Q}^{(n)}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F^{(n)}) = q^{(n)}(\alpha; \boldsymbol{u}'\boldsymbol{Y}_1, \ldots, \boldsymbol{u}'\boldsymbol{Y}_n)\boldsymbol{u}, \tag{4.1.9}$$

where $q^{(n)}(\alpha; Z_1, \ldots, Z_n)$ *is the empirical univariate* α*th quantile of* $\boldsymbol{Z}^{(n)} = (Z_1, \ldots, Z_n) \in \mathbb{R}^n$.

The resulting contours of these multivariate quantiles do not satisfy some nice properties expected from quantile contours. Indeed, they depend on the choice of the origin and can intersect each other (the authors proved that they do not only under the assumption that the origin has a depth greater than α).

The affine equivariance of $\boldsymbol{Q}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F)$ is not entirely clear. Indeed, Kong and Mizera (2012) only proved affine equivariance for the directional quantiles envelopes generated by some estimator $\widehat{\boldsymbol{Q}}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F)$ of $\boldsymbol{Q}_{\mathrm{KM}}(\alpha, \boldsymbol{u}, F)$ under the assumption that these estimators are translation and scale equivariant for all $\boldsymbol{u} \in \mathcal{S}^{m-1}$. Therefore, the answer to Q2 is rather no.

Nevertheless, they provide a connection with halfspace depth via quantile envelopes.

Definition 4.1.9. *The quantile envelope of order* α *of* $\boldsymbol{Y}$ *is defined as the set*

$$\boldsymbol{R}_{KM}(\alpha) = \bigcap_{\boldsymbol{u} \in \mathcal{S}^{m-1}} \boldsymbol{H}(\boldsymbol{u}, Q(\alpha, \boldsymbol{u}'\boldsymbol{Y})),$$

where $\boldsymbol{H}(\boldsymbol{u}, q) = \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} \geq q\}$.

We then have the following proposition.

Proposition 4.1.10. *For all* $\alpha \in (0,1)$, $\boldsymbol{R}_{KM}(\alpha) = \boldsymbol{D}(\alpha) := \{\boldsymbol{y} \in \mathbb{R}^m : HD(\boldsymbol{y}, F) \geq \alpha\}$.

Denote by $\boldsymbol{D}^{(n)}(\alpha)$ the empirical version of $\boldsymbol{D}(\alpha)$. The previous theorem also holds for the empirical version.

Proposition 4.1.11. *Assume that the* n $(\geq m+1)$ *data points are in general position. Then, for any* $l \in \{1, \ldots, n-m\}$ *such that* $\boldsymbol{D}^{(n)}(l/n)$ *has nonempty interior, we have that* $\boldsymbol{R}^{(n)}_{KM}(\alpha) = \boldsymbol{D}^{(n)}(l/n)$ *for all positive* $\alpha \in \left[\frac{l-1}{n}, \frac{l}{n}\right)$.

This proposition provides an interpretation of depth level sets. Indeed, any face of the polyhedral corresponding to the upper level set of depth α coincides with the quantile hyperplane of order $\boldsymbol{\alpha} = \alpha\boldsymbol{u}$ where the direction $\boldsymbol{u}$ is orthogonal to this face. However, quantile envelopes are defined as an intersection on each direction of $\mathcal{S}^{m-1}$: their construction is then not easy and requires a discretization of the sphere in practice. The quality of the resulting approximation deteriorates extremely fast as m increases. We refer the reader to HPŠ10 or Paindaveine and Šiman (2011) for more details.

Concerning asymptotic results, Kong and Mizera (2012) mentioned consistency of the empirical versions of the quantile envelope but did not derive asymptotic normality or Bahadur-type representation. The answer to Q3 is then no.

To conclude, these projection quantiles are interesting since it is a first approach to bridge the gap between the two concepts of quantile and depth, but we have seen that they do not satisfy several expected properties.

4.1.5 Multivariate quantile based on quantile regression

These multivariate quantiles were introduced by HPŠ10. The major interest of their definition of quantiles is that they satisfy all properties usually expected from a quantile while providing a link with halfspace depth.

These quantiles are directional quantiles as those of Kong and Mizera (2012), but stay in a m-dimensional setting. Indeed, they are $(m-1)$-dimensional hyperplanes. Since the median-oriented quantile function used in previous sections was developed for multivariate quantile defined as points once the direction from the median is fixed, it is not well adapted for such multivariate directional quantiles defined as hyperplanes. We will then answer the questions of Section 4.1.1 in a different way.

We will denote by $\boldsymbol{\alpha} = \alpha\boldsymbol{u}$ a m-vector with $\alpha = |\boldsymbol{\alpha}| \in (0,1)$ and $\boldsymbol{u} \in \mathcal{S}^{m-1}$. Let $\boldsymbol{\Gamma}_{\boldsymbol{u}}$ be an arbitrary $m \times (m-1)$ matrix of unit vectors such that $(\boldsymbol{u}, \boldsymbol{\Gamma}_{\boldsymbol{u}})$ is an orthonormal basis of $\mathbb{R}^m$. The $\boldsymbol{\alpha}$-quantile of a m-variate random vector $\boldsymbol{Y}$ is defined as the regression α-quantile hyperplane obtained when regressing $Y_{\boldsymbol{u}} = \boldsymbol{u}'\boldsymbol{Y}$ on $\boldsymbol{Y}^{\perp}_{\boldsymbol{u}} = \boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y}$ and a constant term, whence its name. The vector $\boldsymbol{u}$ indicates now the vertical axis in the regression.

Definition 4.1.12. *The $\boldsymbol{\alpha}$-quantile of $\boldsymbol{Y} \in \mathbb{R}^m$, with $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$, is any element of the collection $\boldsymbol{\Pi}_{\mathrm{HPS};\boldsymbol{\alpha}}$ of hyperplanes $\boldsymbol{\pi}_{\mathrm{HPS};\boldsymbol{\alpha}} := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} = \boldsymbol{c}'_{\boldsymbol{\alpha}}\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{y} + a_{\boldsymbol{\alpha}}\}$ with*

$$(a_{\boldsymbol{\alpha}}, \boldsymbol{c}'_{\boldsymbol{\alpha}})' \in \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \Psi_{\boldsymbol{\alpha}}(a, \boldsymbol{c}), \tag{4.1.10}$$

where $\Psi_{\boldsymbol{\alpha}}(a, \boldsymbol{c}) := \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}^{\perp}_{\boldsymbol{u}} - a)]$.

Notice that this definition extends the traditional univariate quantiles. Indeed, when $m = 1$, hyperplanes of dimension $m-1$ are then points and $\mathcal{B}^m$ is $(-1,0) \cup (0,1)$. Therefore, $\boldsymbol{\pi}_{\boldsymbol{\alpha}}$ is then a classical quantile of order $1 - |\boldsymbol{\alpha}|$ (when $\boldsymbol{\alpha}$ is pointing to the left) or $|\boldsymbol{\alpha}|$ (when $\boldsymbol{\alpha}$ is pointing to the right).

Each quantile hyperplane (i.e. each element $(a_{\boldsymbol{\alpha}}, \boldsymbol{c}'_{\boldsymbol{\alpha}})'$ satisfying (4.1.10)) defines *lower (open) and upper (closed) quantile halfspaces* as

$$\boldsymbol{H}^{-}_{\boldsymbol{\alpha}} = \boldsymbol{H}^{-}_{\boldsymbol{\alpha}}(a_{\boldsymbol{\alpha}}, \boldsymbol{c}_{\boldsymbol{\alpha}}) := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} < \boldsymbol{c}'_{\boldsymbol{\alpha}}\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{y} + a_{\boldsymbol{\alpha}}\}, \tag{4.1.11}$$

$$\boldsymbol{H}^{+}_{\boldsymbol{\alpha}} = \boldsymbol{H}^{+}_{\boldsymbol{\alpha}}(a_{\boldsymbol{\alpha}}, \boldsymbol{c}_{\boldsymbol{\alpha}}) := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} \geq \boldsymbol{c}'_{\boldsymbol{\alpha}}\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{y} + a_{\boldsymbol{\alpha}}\}, \tag{4.1.12}$$

respectively.

Notice that, if $\boldsymbol{u}$ ranges over $\mathcal{S}^{m-1}$, we need finite first-order moment assumption for $\boldsymbol{Y}$. We then adopt the following assumption.

Assumption (E). The distribution of the random vector $\boldsymbol{Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, with a density f that has connected support, and admits finite first-order moments.

Proposition 4.1.13. *Let Assumption (E) hold. Then, for any $\boldsymbol{\alpha} \in \mathcal{B}^m$, the minimizer $(a_{\boldsymbol{\alpha}}, \boldsymbol{c}'_{\boldsymbol{\alpha}})'$ in (4.1.10), hence also the resulting quantile hyperplane $\boldsymbol{\pi}_{\mathrm{HPS};\boldsymbol{\alpha}}$, is unique.*

We will then work under Assumption (E) in the sequel.

When $\boldsymbol{u}$ is fixed, the subfamily $\boldsymbol{\Pi}_{\mathrm{HPS};\boldsymbol{u}} = \{\boldsymbol{\pi}_{\mathrm{HPS};\boldsymbol{\alpha}} : \boldsymbol{\alpha} = \alpha\boldsymbol{u}, \alpha \in (0,1)\}$ provides, for each $\boldsymbol{u}$, the usual interpretation of a collection of regression quantile hyperplanes. When α is fixed, the subfamily $\boldsymbol{\Pi}_{\mathrm{HPS};\alpha} = \{\boldsymbol{\pi}_{\mathrm{HPS};\boldsymbol{\alpha}} : \boldsymbol{\alpha} = \alpha\boldsymbol{u}, \boldsymbol{u} \in \mathcal{S}^{m-1}\}$ provides, for each α, a quantile contour.

These $\boldsymbol{\alpha}$-quantiles can be defined in an alternative but equivalent way with a constrained optimization formulation. The two definitions coincide and also the corresponding lower and upper quantile halfspaces, see HPŠ10 for more details.

We can define analogously empirical versions of these quantile hyperplanes. Let $\boldsymbol{Y}^{(n)} := (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ a n-tuple (with $n > m$) of m-dimensional random vectors, and let $Y_{i\boldsymbol{u}} := \boldsymbol{u}'\boldsymbol{Y}_i$ and $\boldsymbol{Y}^{\perp}_{i\boldsymbol{u}} := \boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y}_i$.

Definition 4.1.14. *The empirical $\boldsymbol{\alpha}$-quantile of $\boldsymbol{Y}^{(n)}$ is any element of the collection $\boldsymbol{\Pi}^{(n)}_{\mathrm{HPS};\boldsymbol{\alpha}}$ of hyperplanes $\boldsymbol{\pi}^{(n)}_{\mathrm{HPS};\boldsymbol{\alpha}} := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} = \boldsymbol{c}^{(n)\prime}_{\boldsymbol{\alpha}}\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{y} + a^{(n)}_{\boldsymbol{\alpha}}\}$ such that*

$$(a^{(n)}_{\boldsymbol{\alpha}}, \boldsymbol{c}^{(n)\prime}_{\boldsymbol{\alpha}})' \in \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \Psi^{(n)}_{\boldsymbol{\alpha}}(a, \boldsymbol{c}), \tag{4.1.13}$$

where $\Psi^{(n)}_{\boldsymbol{\alpha}}(a, \boldsymbol{c}) := \frac{1}{n}\sum_{i=1}^n \rho_\alpha(Y_{i\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}^{\perp}_{i\boldsymbol{u}} - a)$.

We then directly define empirical analogs $\boldsymbol{H}^{(n)-}_{\boldsymbol{\alpha}}$ and $\boldsymbol{H}^{(n)+}_{\boldsymbol{\alpha}}$ of the lower and upper halfspaces in (4.1.11) and (4.1.12). The following assumption is needed when considering empirical quantiles.

Assumption (E_n). The observations $\boldsymbol{Y}_i$, $i = 1, \ldots, n$ are i.i.d with a common distribution satisfying Assumption (E).

Let us now focus on usual properties of quantiles satisfied by these multivariate quantiles. Under Assumption (E), it is easy to check the affine equivariance property

$$\boldsymbol{\pi}_{\mathrm{HPS};\alpha\boldsymbol{M}\boldsymbol{u}/|\boldsymbol{M}\boldsymbol{u}|}(\boldsymbol{M}\boldsymbol{Y} + \boldsymbol{r}) = \boldsymbol{M}\boldsymbol{\pi}_{\mathrm{HPS};\alpha\boldsymbol{u}}(\boldsymbol{Y}) + \boldsymbol{r}, \tag{4.1.14}$$

where $\boldsymbol{M}$ is a $m \times m$ invertible matrix and $\boldsymbol{r}$ a m-vector in $\mathbb{R}^m$. A direct consequence of this equivariance property is that these multivariate quantiles are not localized at any point of $\mathbb{R}^m$. If the origin of $\mathbb{R}^m$ seems playing an important role in their definition, as center of the $(d-1)$-dimensional unit sphere, we now see that it is not the case. Therefore, it answers yes to Q2.

Under Assumption (E_n), HPŠ10 derived strong consistency results for sample $\boldsymbol{\alpha}$-quantiles and related quantities. The asymptotic normality and Bahadur-type representation results were also proven under a slightly stronger assumption. Then, Q3 is also positively answered.

These multivariate quantiles were defined as the solution of an optimization problem. Since the objective function $\Psi_{\boldsymbol{\alpha}}$ appearing in (4.1.10) is convex and continuously differentiable on $\mathbb{R}^m$,

the authors derived some subgradient conditions allowing to characterize the minimizer. Setting to zero the partial derivative according to a gives

$$P\big(\boldsymbol{Y} \in \boldsymbol{H}^{-}_{\boldsymbol{\alpha}}(a_{\boldsymbol{\alpha}}, \boldsymbol{c}_{\boldsymbol{\alpha}})\big) = \alpha, \tag{4.1.15}$$

which provides a natural probabilistic interpretation: the probability of the lower halfspace corresponding to the $\boldsymbol{\alpha}$-quantile, with $\boldsymbol{\alpha} = \alpha\boldsymbol{u}$, is equal to $\alpha = \|\boldsymbol{\alpha}\|$. This answers yes to Q1.

The last significant property of these quantiles is their connection to halfspace depth. We define, for any $\alpha \in (0,1)$, the *α-quantile region* $\boldsymbol{R}_{\mathrm{HPS}}(\alpha)$ as

$$\boldsymbol{R}_{\mathrm{HPS}}(\alpha) := \bigcap_{\boldsymbol{u}\in\mathcal{S}^{m-1}} \cap\{\boldsymbol{H}^{+}_{\alpha\boldsymbol{u}}\},$$

where $\cap\{\boldsymbol{H}^{+}_{\alpha\boldsymbol{u}}\}$ is the intersection of the collection $\{\boldsymbol{H}^{+}_{\alpha\boldsymbol{u}}\}$ of all closed upper $\alpha\boldsymbol{u}$-quantile halfspaces (4.1.12). For the particular case $\alpha = 0$, we let $\boldsymbol{R}_{\mathrm{HPS}}(0) = \mathbb{R}^m$. The empirical version of these regions are similarly defined as

$$\boldsymbol{R}^{(n)}_{\mathrm{HPS}}(\alpha) := \bigcap_{\boldsymbol{u}\in\mathcal{S}^{m-1}} \cap\{\boldsymbol{H}^{(n)+}_{\alpha\boldsymbol{u}}\},$$

for any $\alpha \in (0,1)$ and $\boldsymbol{R}^{(n)}_{\mathrm{HPS}}(0) := \mathbb{R}^m$.

Theorem 4.1.15. *Under Assumption (E), $\boldsymbol{R}_{\mathrm{HPS}}(\alpha) = \boldsymbol{D}(\alpha) := \{\boldsymbol{y} \in \mathbb{R}^m : HD(\boldsymbol{y}, F) \geq \alpha\}$ for all $\alpha \in (0,1)$.*

Still denoting by $\boldsymbol{D}^{(n)}(\alpha)$ the empirical version of $\boldsymbol{D}(\alpha)$, the previous theorem also holds for the empirical version.

Theorem 4.1.16. *Assume that the n ($\geq m+1$) data points are in general position. Then, for any $l \in \{1,\ldots,n-m\}$ such that $\boldsymbol{D}^{(n)}(l/n)$ has nonempty interior, we have that $\boldsymbol{R}^{(n)}_{\mathrm{HPS}}(\alpha) = \boldsymbol{D}^{(n)}(l/n)$ for all positive $\alpha \in \left[\frac{l-1}{n}, \frac{l}{n}\right)$.*

These theorems then show that the halfspace depth regions coincide with the upper envelope of directional quantile halfspaces. Moreover, the face of the polyhedral empirical depth contours are parts of empirical quantile hyperplanes. Therefore, Q4 is positively answered and it provides in addition nice geometrical characterizations.

We then showed that this last definition of multivariate quantiles is very attractive since it enjoys all the properties we expected in Section 4.1.1.

4.1.6 Choice of a notion of multivariate quantile

The previous sections gave an overview on different existing notions of multivariate quantiles, that we attempt to make as complete as possible. We first pointed out the properties expected from a quantile, and investigated one by one each definition to determine which properties were satisfied by this one. The original approach of HPŠ10 stands out by answering positively to all questions of Section 4.1.1, and thus providing a connection with halfspace depth while satisfying all nice properties expected as affine equivariance. Therefore, we choose to use the definition of multivariate quantiles of HPŠ10 in the sequel.

4.2 Multiple-output conditional quantiles

As for the single-output case, it happens regularly that there exist some variables providing information on the variable $\boldsymbol{Y}$ one considers. Our choice of multivariate quantiles allows to define *multiple-output conditional quantiles* that enjoy all nice properties of their single-output counterparts.

4.2.1 Definition of multiple-output conditional quantiles

In the sequel, $\boldsymbol{X}$ denotes a vector of covariates of dimension d, and $\boldsymbol{Y}$ a vector of response variables of dimension m. We thus consider the multivariate quantiles of HPŠ10 that are $(m-1)$-dimensional hyperplanes indexed by $\boldsymbol{\alpha}$, ranging over $\mathcal{B}^m$. The equivalent of the scalar order of quantiles in dimension one is then here the directional index $\boldsymbol{\alpha}$ that factorizes as $\boldsymbol{\alpha} = \alpha\boldsymbol{u}$, with $\alpha = |\boldsymbol{\alpha}| \in (0,1)$ and $\boldsymbol{u} \in \mathcal{S}^{m-1}$. We then define the conditional $\boldsymbol{\alpha}$-quantiles of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ as follows.

Definition 4.2.1. *For $\boldsymbol{\alpha} := \alpha\boldsymbol{u} \in \mathcal{B}^m$, the conditional quantile of order $\boldsymbol{\alpha}$ of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is any element of the collection $\boldsymbol{\Pi}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ of hyperplanes $\boldsymbol{\pi}_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} = (\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{y} + a_{\boldsymbol{\alpha}}(\boldsymbol{x})\}$ such that*

$$\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \begin{pmatrix} a_{\boldsymbol{\alpha}}(\boldsymbol{x}) \\ \boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}) \end{pmatrix} = \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{x}], \tag{4.2.1}$$

where $Y_{\boldsymbol{u}} = \boldsymbol{u}'\boldsymbol{Y}$ and $\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} = \boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}$.

Notice that such a quantile is entirely characterized by $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$. We adopt the following assumption in the sequel.

Assumption (F) The distribution of the random vector $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, with a density that has connected support, and admits finite first-order moments.

This assumption is analogous to Assumption (E) in this regression setup. Therefore, under Assumption (F), the minimization problem in (4.2.1) admits a unique solution. Moreover, the function $(a,\boldsymbol{c}) \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x}) = \mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a)|\boldsymbol{X} = \boldsymbol{x}]$ is convex and continuously differentiable on $\mathbb{R}^m$. Then, $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \big(a_{\boldsymbol{\alpha}}(\boldsymbol{x}), (\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))'\big)'$ satisfies the system of equations

$$\nabla_{(a,\boldsymbol{c}')'} G_{a,\boldsymbol{c}}(\boldsymbol{x}) = 0.$$

Thanks to Lemma 4.4.1 given in Appendix 4.4, the conditional quantiles are then characterized by the relations

$$0 = (\partial_a G_{a,\boldsymbol{c}}(\boldsymbol{x}))_{(a_{\boldsymbol{\alpha}}(\boldsymbol{x}),(\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))')'} = P[\boldsymbol{u}'\boldsymbol{Y} < a_{\boldsymbol{\alpha}}(\boldsymbol{x}) + (\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}] - \alpha \tag{4.2.2}$$

and

$$0 = (\nabla_{\boldsymbol{c}} G_{a,\boldsymbol{c}}(\boldsymbol{x}))_{(a_{\boldsymbol{\alpha}}(\boldsymbol{x}),(\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))')'} = \mathrm{E}\big[\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}\big(\alpha - \mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y} < a_{\boldsymbol{\alpha}}(\boldsymbol{x}) + (\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}]}\big)\big|\boldsymbol{X} = \boldsymbol{x}\big]. \tag{4.2.3}$$

As we will see in the sequel, twice differentiability of $(a, \boldsymbol{c}')' \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$, however, requires slightly stronger assumptions. Consider the following reinforcement of Assumption (F).

Assumption (F)' The distribution of the random vector $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, with a density $f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ that has connected support, admits finite *second*-order moments and, for some constants $C > 0$, $r > m - 1$ and $s > 0$, satisfies

$$\left|f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}_1) - f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}_2)\right| \leq C|\boldsymbol{y}_1 - \boldsymbol{y}_2|^s \left(1 + \left|\frac{\boldsymbol{y}_1 + \boldsymbol{y}_2}{2}\right|\right)^{-(3+r+s)/2}, \tag{4.2.4}$$

for all $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^m$.

More details on this Assumption (and particularly on (4.2.4)) can be found in HPŠ10. The lower limit for r was slightly changed for calculation purposes.

4.2.2 Estimation of multiple-output conditional quantiles

As we saw in the first part, quantile regression issue has motivated the definition of numerous estimators, as for instance nearest-neighbor, spline, kernel estimators, or of course quantization-based estimators. However, much less results are available in the multiple-output case. The main reason for this is that no definition of multivariate quantile is universally preferred, which does not encourage to consider multiple-output quantile regression. Moreover, it is often challenging to incorporate the covariate effect into the definition of multivariate quantiles.

We saw in Section 4.1 that the spatial quantiles defined in (4.1.4) are quite interesting since a Transformation-Retransformation tool can make them affine-equivariant. For this reason, much work has been devoted to the notion of spatial regression quantiles, as the ones of Chakraborty (2003) for linear regression and Cheng and De Gooijer (2007) for nonparametric regression. However, if the TR tool can be adapted in the linear regression framework, there does not exist such a tool working in the general nonparametric regression framework, to the best of our knowledge. Therefore, the resulting quantiles are not affine-equivariant.

Wei (2008) recently proposed a new concept of multiple-output regression contours. Even if they satisfy some interesting probability properties, they have no link with depth contours and they are not affine equivariant.

When defining multivariate quantiles, HPŠ10 also introduced multiple-output regression quantiles, extending the famous single-output concept of Koenker and Bassett (1978). This first approach for this definition of multivariate quantiles is quite simple and consists essentially in introducing the covariate into the equations that characterize the regions and contours (see HPŠ10 for a precise definition). Empirical versions of these, however, generally provide poor results and carry little information about the conditional distribution. For example, they could miss some important features as parabolic trend or heteroscedasticity.

A natural step was then to extend the single-output estimation procedure to the concept of multiple-output quantile regression of HPŠ10. Starting from the observation that local constant and local linear methods perform well in the single-output framework, Hallin et al. (2015) (hereafter, HPŠ15) developed local constant and local bilinear approaches in the multiple-output

context. Their estimators are then the main existing competitors. In this part, we introduce a new nonparametric multiple-output quantile regression method, that extends the quantization-based estimator of the first part in a multivariate setting.

4.3 Objective and structure of the part

This second part is dedicated to the extension in a multivariate setting of the quantization-based estimators defined in Part I. A motivation for such an extension of these estimators comes from our simulation results of Chapter 3. We have indeed shown that our quantization-based estimator $\bar{q}_{\alpha,B}^{N,n}(x)$ generally outperforms its local linear and local constant competitors. It is then natural to wonder if its dominance also appears when the dependent variable is of dimension greater than one. This part is composed of several chapters.

Chapter 5 is the natural generalization of Chapter 2 in a multiple-output framework. In the same spirit, we first replace $\boldsymbol{X}$ by its quantized version in Definition 4.2.1 of multiple-output conditional quantiles. The resulting approximation is then studied from a theoretical point of view in Section 5.2 where convergence results are obtained. In a second time, we define in Section 5.3 an estimator of multiple-output conditional quantiles by taking an empirical version of this approximation. It is then proven that this estimator is consistent for this approximation. However, even if these asymptotic results are really interesting, we again observe that the resulting conditional quantile contours can be improved by defining a bootstrap version of our estimator. Sections 5.4 and 5.5 gather the proofs of Sections 5.2 and 5.3 respectively.

This theoretical investigation naturally requires to address the numerical behavior of our estimator. This is the object of Chapter 6. We first extend in Section 6.2 the data-driven selection method for the parameter N (the size of the quantization grid) of Section 3.2, required in the construction of our estimator. This method showed its efficiency on several replications. With this selection method, our estimation procedure is entirely functional and we are then able to compare our performances with the ones of alternatives estimators in Section 6.3. The comparison study focuses on resulting conditional quantiles contours and on some integrated square error. The numerical investigation is completed in Section 6.4 with a real data application.

4.4 Proof of Section 4.2

This section aims to prove the relations (4.2.2) and (4.2.3) of Section 4.2. The following lemma gives the gradient and the Hessian matrix of the function $(a, \boldsymbol{c}')' \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$.

Lemma 4.4.1. *Let Assumption (F)' hold. For any $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $(a, \boldsymbol{c}')' \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$ is twice continuously differentiable with gradient vector*

$$\begin{aligned}\boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) &= \left(\frac{\partial G_{a,\boldsymbol{c}}(\boldsymbol{x})}{\partial a}, \left(\boldsymbol{\nabla}_{\boldsymbol{c}} G_{a,\boldsymbol{c}}(\boldsymbol{x})\right)'\right)', \\ &= \begin{pmatrix} P\left[\boldsymbol{u}'\boldsymbol{Y} < a + c\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}\right] - \alpha \\ \mathrm{E}\left[\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}\left(\mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y} < a + c\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}]} - \alpha\right) \middle| \boldsymbol{X} = \boldsymbol{x}\right] \end{pmatrix},\end{aligned}$$

and Hessian matrix

$$\boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big) = \int_{\mathbb{R}^{m-1}} \begin{pmatrix} 1 & \boldsymbol{t}' \\ \boldsymbol{t} & \boldsymbol{t}\boldsymbol{t}' \end{pmatrix} f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)d\boldsymbol{t}.$$

Proof. Let $\eta_{\boldsymbol{\alpha}}(a,\boldsymbol{c}) = -\big(\alpha - \mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a<0]}\big)\begin{pmatrix}1 & (\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y})'\end{pmatrix}'$. For all $(a,\boldsymbol{c}')'$ and $(a_0,\boldsymbol{c}_0')' \in \mathbb{R}^m$, we directly show that

$$\begin{aligned}
&\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a) - \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}_0'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a_0) - (a-a_0,\boldsymbol{c}'-\boldsymbol{c}_0')\eta_{\boldsymbol{\alpha}}(a_0,\boldsymbol{c}_0) \\
&\quad = (\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a)\{\mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}_0'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a_0<0]} - \mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a<0]}\} \\
&\quad \geq 0. \qquad (4.4.1)
\end{aligned}$$

Therefore, $\eta_{\boldsymbol{\alpha}}(a,\boldsymbol{c})$ is a subgradient for $(a,\boldsymbol{c}) \mapsto \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a)$. Hence, interchanging differentiation and expectation, which is justified in a standard way, provides

$$\boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) = \boldsymbol{\nabla}_{a,\boldsymbol{c}}\mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}-\boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}-a)|\boldsymbol{X}=\boldsymbol{x}] = \mathrm{E}[\eta_{\boldsymbol{\alpha}}(a,\boldsymbol{c})|\boldsymbol{X}=\boldsymbol{x}],$$

which entails

$$\boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) = \begin{pmatrix} P\big[\boldsymbol{u}'\boldsymbol{Y} < a+c\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}\big]-\alpha \\ \mathrm{E}\big[\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}\big(\mathbb{I}_{[\boldsymbol{u}'\boldsymbol{Y}<a+c\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}]}-\alpha\big)\big|\boldsymbol{X}=\boldsymbol{x}\big] \end{pmatrix}.$$

Let us now show that

$$|\boldsymbol{\nabla} G_{a+\Delta_a,\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}}}(\boldsymbol{x}) - \boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'| = o\big(|(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'|\big),$$

for $|(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'| \to 0$. Previous calculations allow to obtain

$$\begin{aligned}
&\boldsymbol{\nabla} G_{a+\Delta_a,\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}}}(\boldsymbol{x}) - \boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')' \\
&\quad = \mathrm{E}[\eta_{\boldsymbol{\alpha}}(a+\Delta_a,\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}}) - \eta_{\boldsymbol{\alpha}}(a,\boldsymbol{c})|\boldsymbol{X}=\boldsymbol{x}] - \int_{\mathbb{R}^{m-1}} \begin{pmatrix} 1 & \boldsymbol{t}' \\ \boldsymbol{t} & \boldsymbol{t}\boldsymbol{t}' \end{pmatrix} f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)d\boldsymbol{t} \\
&\qquad \times (\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')' \\
&\quad = \int_{\mathbb{R}^{m-1}}\int_{\mathbb{R}} \big(\mathbb{I}_{[z-(\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}})'\boldsymbol{t}-(a+\Delta_a)<0]} - \mathbb{I}_{[z-\boldsymbol{c}'\boldsymbol{t}-a<0]}\big)(1,\boldsymbol{t}')' f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(z\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t})dzd\boldsymbol{t} \\
&\qquad - \int_{\mathbb{R}^{m-1}} \begin{pmatrix} 1 & \boldsymbol{t}' \\ \boldsymbol{t} & \boldsymbol{t}\boldsymbol{t}' \end{pmatrix} f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)d\boldsymbol{t}(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'. \qquad (4.4.2)
\end{aligned}$$

Since one directly gets that

$$\int_{a+\boldsymbol{c}'\boldsymbol{t}}^{(a+\Delta_a)+(\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}})'\boldsymbol{t}} (1,\boldsymbol{t}')'dz = \begin{pmatrix} 1 & \boldsymbol{t}' \\ \boldsymbol{t} & \boldsymbol{t}\boldsymbol{t}' \end{pmatrix}(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')',$$

equation (4.4.2) becomes

$$\begin{aligned}
&\boldsymbol{\nabla} G_{a+\Delta_a,\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}}}(\boldsymbol{x}) - \boldsymbol{\nabla} G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')' \\
&\quad = \int_{\mathbb{R}^{m-1}}\int_{a+\boldsymbol{c}'\boldsymbol{t}}^{(a+\Delta_a)+(\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}})'\boldsymbol{t}} \big\{f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(z\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}) - f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)\big\}(1,\boldsymbol{t}')'dzd\boldsymbol{t}.
\end{aligned}$$

By Assumption (F)', one has

$$|f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(z\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}) - f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)| \leq C\frac{|z-a-\boldsymbol{c}'\boldsymbol{t}|^s}{\big(1+|\frac{1}{2}(z+a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}|^2\big)^{(3+r+s)/2}}$$
$$\leq C\frac{|\Delta_a+\boldsymbol{\Delta}_{\boldsymbol{c}}'\boldsymbol{t}|^s}{|(1,\boldsymbol{t}')'|^{(3+r+s)}},$$

for any z between $a+\boldsymbol{c}'\boldsymbol{t}$ and $(a+\Delta_a)+(\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}})'\boldsymbol{t}$. This entails

$$\begin{aligned}
&|\boldsymbol{\nabla}G_{a+\Delta_a,\boldsymbol{c}+\boldsymbol{\Delta}_{\boldsymbol{c}}}(\boldsymbol{x}) - \boldsymbol{\nabla}G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'| \\
&\quad\leq C\int_{\mathbb{R}^{m-1}}|\Delta_a+\boldsymbol{\Delta}_{\boldsymbol{c}}'\boldsymbol{t}|^{1+s}|(1,\boldsymbol{t}')'|^{-(2+r+s)}d\boldsymbol{t} \\
&\quad\leq C|(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'|^{1+s}\int_{\mathbb{R}^{m-1}}|(1,\boldsymbol{t}')'|^{-(1+r)}d\boldsymbol{t} \\
&\quad= o(|(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'|),
\end{aligned}$$

as $|(\Delta_a,\boldsymbol{\Delta}_{\boldsymbol{c}}')'|\to\infty$. Therefore, for any $\boldsymbol{x}\in S_{\boldsymbol{X}}$, $(a,\boldsymbol{c}')'\mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$ is twice continuously differentiable with Hessian matrix $\boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)$. □

5

Multiple-output quantile regression through optimal quantization

Contents

5.1 Introduction

In this chapter, we define a new estimator of multiple-output conditional quantile. As above mentioned, this estimator is constructed using optimal quantization, in the same spirit as the quantization-based estimator of Chapter 2. The structure of this chapter is thus analogous to the one of Chapter 2. First, Section 5.2 explains the first step of the construction: replacing the covariate $\boldsymbol{X}$ in Definition 4.2.1 of multiple-output conditional quantile to get an approximation of them. The convergence of this approximation as the size N of the quantization grid increases is also investigated. Further, an estimator is defined by taking an empirical version of this approximation in Section 5.3. Asymptotic results are derived, for N fixed and the sample size n going to infinity. A bootstrap version of this estimator is then defined, given that the numerical performances of the bootstrap estimator $\bar{q}_{\alpha,B}^{N,n}$ were significantly better than the ones of $\widehat{q}_{\alpha}^{N,n}$ in the first part. These two sections (and their proofs) constitute the first part of the working paper *Multiple-output quantile regression through optimal quantization* with Davy Paindaveine and Jérôme Saracco.

Throughout the chapter, $\boldsymbol{X}$ is a d-dimensional random vector of covariates while $\boldsymbol{Y}$ is a m-dimensional vector of response variables, unless otherwise stated.

5.2 Approximation of multiple-output conditional quantile

As we now explain, we construct an approximation of multiple-output conditional quantiles using optimal quantization. Similarly to the single-output case of Chapter 2, let $p \geq 1$ such that $\|\boldsymbol{X}\|_p < \infty$ and let $\boldsymbol{\gamma}^N$ be an optimal quantization grid of size N, with N a positive integer, that is $\boldsymbol{\gamma}^N \in (\mathbb{R}^d)^N$. Denote $\widetilde{\boldsymbol{X}}^N := \mathrm{Proj}_{\boldsymbol{\gamma}^N}(\boldsymbol{X})$ obtained by projecting $\boldsymbol{X}$ onto this quantization grid (see Section 1.2). Replacing $\boldsymbol{X}$ in (4.2.1) by this projection allows then to consider

$$\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \begin{pmatrix} \widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) \\ \widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) \end{pmatrix} = \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}^{\perp}_{\boldsymbol{u}} - a) | \widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}], \tag{5.2.1}$$

where $\tilde{\boldsymbol{x}}$ denotes the projection $\mathrm{Proj}_{\boldsymbol{\gamma}^N}(\boldsymbol{x})$. An approximation of multiple-output conditional quantiles of order $\boldsymbol{\alpha} = \alpha \boldsymbol{u}$ is then any element of the collection $\widetilde{\boldsymbol{\Pi}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$ of the following hyperplanes

$$\tilde{\boldsymbol{\pi}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} = (\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}))'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{y} + \widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})\}.$$

Notice that such an approximation is entirely characterized by $\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$. In the sequel, we will then investigate the quality of our approximation through $\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$. The classical convergence result of quantization allows us to expect that $\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ tends to zero. This will be the aim of this section.

The following assumptions are needed.

Assumption (G) (i) The random vector $(\boldsymbol{X}, \boldsymbol{Y})$ is generated through $\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{X}, \boldsymbol{\varepsilon})$, where the d-dimensional covariate vector $\boldsymbol{X}$ and the m-dimensional error $\boldsymbol{\varepsilon}$ are mutually independent; (ii) the link function $\boldsymbol{M} : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m : (\boldsymbol{x}, \boldsymbol{z}) \mapsto \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{z})$ is of the form $\boldsymbol{M}_1(\boldsymbol{x}) + \boldsymbol{M}_2(\boldsymbol{x})\boldsymbol{z}$, where the functions $\boldsymbol{M}_1 : \mathbb{R}^d \to \mathbb{R}^m$ and $\boldsymbol{M}_2 : \mathbb{R}^d \to M^+_m(\mathbb{R})$ admit Lipschitz property, with $M^+_m(\mathbb{R})$ denoting the set of square matrices of size $m \times m$ positive definite ; (iii) $\|\boldsymbol{X}\|_p < \infty$ and $\|\boldsymbol{\varepsilon}\|_p < \infty$; (iv) the distribution of $\boldsymbol{X}$ does not charge any hyperplane.

Notice that Assumptions (G)(iii) and (iv) are the same as Assumptions (A)(iii) and (iv), while Assumptions (G)(i) and (ii) are the generalization of Assumptions (A)(i) and (ii) in this multivariate setting. Let us comment Assumption (G)(ii). The Lipschitz property of $\boldsymbol{M}_2$ requires first to choose a matrix norm. We use here the operator norm on the space $M^+_m(\mathbb{R})$ defined, for any matrix $\boldsymbol{A} \in M^+_m(\mathbb{R})$, by

$$||\boldsymbol{A}|| = \sup_{\boldsymbol{x} \in \mathbb{R}^m, |\boldsymbol{x}|=1} |\boldsymbol{A}\boldsymbol{x}|.$$

Then, $[\boldsymbol{M}_2]_{\mathrm{Lip}}$ is the smallest real number C for which

$$\forall \boldsymbol{x}_1, \boldsymbol{x_2} \in \mathbb{R}^d, ||\boldsymbol{M}_2(\boldsymbol{x}_1) - \boldsymbol{M}_2(\boldsymbol{x}_2)|| \leq C|\boldsymbol{x}_1 - \boldsymbol{x}_2|.$$

Let $[\boldsymbol{M}_1]_{\mathrm{Lip}}$ denote the Lipschitz constant for $\boldsymbol{M}_1$. The Lipschitz property of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ directly implies that there exists $C > 0$ such that the link function $\boldsymbol{M}(\cdot, \boldsymbol{\varepsilon})$ is also Lipschitz. Indeed, we

have

$$\begin{aligned}\|\boldsymbol{M}(\boldsymbol{x}_1,\boldsymbol{\varepsilon})-\boldsymbol{M}(\boldsymbol{x}_2,\boldsymbol{\varepsilon})\|_p &= \|(\boldsymbol{M}_1(\boldsymbol{x}_1)-\boldsymbol{M}_1(\boldsymbol{x}_2))+(\boldsymbol{M}_2(\boldsymbol{x}_1)-\boldsymbol{M}_2(\boldsymbol{x}_2))\boldsymbol{\varepsilon}\|_p\\ &\leq |\boldsymbol{M}_1(\boldsymbol{x}_1)-\boldsymbol{M}_1(\boldsymbol{x}_2)|+\big\|\big(\boldsymbol{M}_2(\boldsymbol{x}_1)-\boldsymbol{M}_2(\boldsymbol{x}_2)\big)\boldsymbol{\varepsilon}\big\|_p\\ &\leq [\boldsymbol{M}_1]_{\mathrm{Lip}}|\boldsymbol{x}_1-\boldsymbol{x}_2|+||\boldsymbol{M}_2(\boldsymbol{x}_1)-\boldsymbol{M}_2(\boldsymbol{x}_2)||\,\|\boldsymbol{\varepsilon}\|_p\\ &\leq ([\boldsymbol{M}_1]_{\mathrm{Lip}}+[\boldsymbol{M}_2]_{\mathrm{Lip}}\|\boldsymbol{\varepsilon}\|_p)|\boldsymbol{x}_1-\boldsymbol{x}_2|,\end{aligned}$$

which implies that $\boldsymbol{M}(\cdot,\boldsymbol{\varepsilon})$ is a Lipschitz function with Lipschitz constant $[\boldsymbol{M}]_{\mathrm{Lip}} \leq [\boldsymbol{M}_1]_{\mathrm{Lip}} + [\boldsymbol{M}_2]_{\mathrm{Lip}}\|\boldsymbol{\varepsilon}\|_p$.

ASSUMPTION (H) (i) The support $S_{\boldsymbol{X}}$ of $P_{\boldsymbol{X}}$ is compact; (ii) $\boldsymbol{\varepsilon}$ admits a continuous density $f^{\boldsymbol{\varepsilon}} : \mathbb{R}^m \to \mathbb{R}_0^+$ (with respect to the Lebesgue measure on $\mathbb{R}^m$).

Notice that, under Assumption (G) and Assumption (H)(ii), the Hessian matrix of $(a,\boldsymbol{c}')' \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$ rewrites as

$$\boldsymbol{H}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big) = \frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))}\int_{\mathbb{R}^{m-1}} \begin{pmatrix}1 & \boldsymbol{t}'\\ \boldsymbol{t} & \boldsymbol{t}\boldsymbol{t}'\end{pmatrix} f^{\boldsymbol{\varepsilon}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-\boldsymbol{M}_1(\boldsymbol{x})\big)\big)d\boldsymbol{t}. \tag{5.2.2}$$

In addition, one directly has that the restriction of this Hessian matrix to the derivatives with respect to $\boldsymbol{c}$

$$\boldsymbol{H}^{\boldsymbol{c}}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big) = \frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))}\int_{\mathbb{R}^{m-1}} \boldsymbol{t}\boldsymbol{t}' f^{\boldsymbol{\varepsilon}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}\big((a+\boldsymbol{c}'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-\boldsymbol{M}_1(\boldsymbol{x})\big)\big)d\boldsymbol{t}$$

is positive definite under Assumption (G) and Assumption (H)(ii).

We would like to prove theorems analogous to Theorems 2.2.1 and 2.2.2 of Chapter 2. The following theorem focuses on fixed-$\boldsymbol{x}$ consistency (see Section 5.4 for the proof).

Theorem 5.2.1. *Fix $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$. Then, under Assumptions (F)', (G) and (H),*

$$\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} |\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})| \to 0,$$

as $N \to \infty$.

As for its single-output counterpart in Chapter 2, this last theorem does not provide any convergence rate, since the convergence rate for $\widetilde{\boldsymbol{X}}^N$ toward $\boldsymbol{X}$ does not imply any rate for the fixed-$\boldsymbol{x}$ convergence.

However, extending Theorem 2.2.1 in this multivariate setting appears extremely delicate. Indeed, the core of the proof relies on the following structure: 1) obtaining a rate of convergence for the difference $\big\|\sup_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} |G_{a,\boldsymbol{c}}(\boldsymbol{X}) - \widetilde{G}_{a,\boldsymbol{c}}(\widetilde{\boldsymbol{X}}^N)|\big\|_p$, hence for $\|G_{\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{X})}(\boldsymbol{X}) - G_{\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{X})}(\boldsymbol{X})\|_p$ and 2) deriving a rate of convergence for $\|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{X}) - \widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{X})\|_p$ from the one of $\|G_{\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{X})}(\boldsymbol{X}) - G_{\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{X})}(\boldsymbol{X})\|_p$. Going from $G_{\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{X})}(\boldsymbol{X})$ to $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{X})$ requires actually to go from $a_{\boldsymbol{\alpha}}(\boldsymbol{X}) + \boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{X})'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}$ to $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$, which is not possible since the matrix $(1 \;\; (\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y})')'(1 \;\; (\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y})')$ showing up is not of maximal rank. Nevertheless, if obtaining such a theorem would have been interesting for the convergence rate, the main result is Theorem 5.2.1 that provides fixed-$\boldsymbol{x}$ convergence.

5.3 Estimation of multiple-output conditional quantile

5.3.1 The proposed estimators and their consistency

Let $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, be n independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$. We consider now the problem of estimating multiple-output conditional quantiles using these independent copies. In this section, we fix N $(< n)$ and construct an estimator $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ of the conditional quantile $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ from the approximation $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})$. This construction is realized through the two following steps, in the same spirit as the single-output counterpart.

Step 1: Selection of the optimal grid

As in Chapter 2, we only have at our disposal these n independent copies of $\boldsymbol{X}$ whose distribution is unknown. Therefore, the number of replications of the stochastic gradient algorithm (used to construct approximation of optimal grids) is fixed by the sample size n and each $\boldsymbol{X}_i$ plays the role of stimuli.

First, an initial grid has to be chosen to perform the stochastic gradient algorithm, and again $\hat{\gamma}^{N,0}$ is selected by sampling randomly without replacement among the $\boldsymbol{X}_i$'s. Hence, this initial grid has pairwise distinct components and we start with a grid that is indeed composed of N points.

The stochastic gradient algorithm is then performed, based on $\boldsymbol{\xi}^t = \boldsymbol{X}_t$, for $t = 1, \ldots, n$. The resulting optimal grid is denoted $\hat{\gamma}^{N,n} = (\hat{\boldsymbol{x}}_1^{N,n}, \ldots, \hat{\boldsymbol{x}}_n^{N,n})$. In the sequel, we write $\widehat{\boldsymbol{X}}^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(\boldsymbol{X})$ the corresponding quantization of $\boldsymbol{X}$. To make the notation less heavy, we will only stress the dependence in n when necessary. However, notice that we use throughout this chapter notation coherent with the one of Chapter 2. Indeed, the tilde symbol always corresponds to optimal grids (or projections onto an optimal grid) while the hat symbol refers to approximations of an optimal grid obtained with the algorithm (or projections onto such a approximated optimal grid). Therefore, there should be no ambiguity concerning which kind of grid is considered, or which kind of optimality we consider.

Step 2: Estimation

We now project the $\boldsymbol{X}$-part of the sample onto this optimal grid $\hat{\gamma}^{N,n}$. Let $\widehat{\boldsymbol{X}}_i^N = \widehat{\boldsymbol{X}}_i^{N,n} = \mathrm{Proj}_{\hat{\gamma}^{N,n}}(\boldsymbol{X}_i)$. We thus consider now the sample $(\widehat{\boldsymbol{X}}_i^N, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$. The approximation $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x}) = \arg\min_{(a,\boldsymbol{c}')'} \mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a) | \widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]$ is then estimated by taking an empirical version, i.e.

$$\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) = \begin{pmatrix} \widehat{a}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) \\ \widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) \end{pmatrix}$$

$$= \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \sum_{i=1}^{n} \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}_i - a) \frac{\mathbb{I}_{[\hat{\boldsymbol{X}}_i^N = \hat{\boldsymbol{x}}]}}{\#\{j \in \{1, \ldots, n\} : \hat{\boldsymbol{X}}_j^N = \hat{\boldsymbol{x}}\}},$$

where $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(\boldsymbol{x})$. Equivalently, we rather define

$$\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) = \begin{pmatrix} \widehat{a}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) \\ \widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) \end{pmatrix} = \arg\min_{(a,\boldsymbol{c}')' \in \mathbb{R}^m} \sum_{i=1}^{n} \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}_i - a)\mathbb{I}_{[\hat{\boldsymbol{X}}_i^N = \hat{\boldsymbol{x}}]}. \tag{5.3.1}$$

An estimator of multiple-output conditional quantiles is then any element of the collection $\widehat{\boldsymbol{\Pi}}_{\boldsymbol{\alpha}}^N$ of the following hyperplanes

$$\hat{\boldsymbol{\pi}}_{\boldsymbol{\alpha}}^N := \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^m : \boldsymbol{u}'\boldsymbol{y} = (\widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}))'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{y} + \widehat{a}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})\}.$$

Notice that such an estimation is entirely characterized by $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$. In the sequel, we will then investigate the quality of our estimation through $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$.

For N and $\boldsymbol{x} \in S_{\boldsymbol{X}}$ fixed, the following theorem derives the convergence in probability of $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ toward $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})$ as $n \to \infty$. In this aim, we need to restrict to $p = 2$. Indeed, recall that this setting is the only one providing convergence results for the stochastic gradient algorithm (i.e, CLVQ) that are needed for our purposes (see Section 1.2). We also still make the following assumption (that is actually the same as Assumption (C)).

ASSUMPTION (I) $P_{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

We then have the following result (see Section 5.5 for the proof).

Theorem 5.3.1. *Fix $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$, $\boldsymbol{x} \in S_{\boldsymbol{X}}$ and $N \in \mathbb{N}_0$. Then, under Assumptions (F), (G), (H)(i) and (I), we have that, as $n \to \infty$,*

$$|\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) - \widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})| \to 0$$

in probability, provided that quantization is based on $p = 2$.

In the previous section, we showed that $|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})|$ goes to zero as N tends to infinity almost surely, hence in probability. Theorem 5.3.1 suggests that it could be combined with Theorem 5.2.1 to get an asymptotic result for $|\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})|$ in probability, as $n \to \infty$ and with $N = N_n$ going to infinity in an appropriate way. However, obtaining such a result is extremely delicate for the same reason as in Chapter 2: to the best of our knowledge, all convergence results for the CLVQ algorithm are as $n \to \infty$ with N fixed.

5.3.2 Numerical example and bootstrap modification

As for the single-output case, the stochastic gradient algorithm can provide a grid that is not optimal in the case of small sample sizes. Moreover, by construction, the conditional quantiles are constant on the quantization cell, inducing possible important jumps at the boundaries of the cells. The resulting conditional quantile contours may then suffer from these facts. A solution is again to perform bootstrap to obtain smoother quantile contours. The construction is the exact same one as in Section 2.3.2.

(S1) For some integer B, we first generate B samples of size n with replacement from the initial sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, that we denote $(\boldsymbol{\xi}_b^t)_t$, $t = 1, \ldots, n$ and $b = 1, \ldots, B$. We also generate initial grids $\hat{\boldsymbol{\gamma}}_b^{N,0}$ as above, by sampling randomly among the corresponding $(\boldsymbol{\xi}_b^t)_t$ under the constraints that the N values are pairwise distinct. We then perform B times CLVQ with iterations based on $\boldsymbol{\xi}_b^t$, $t = 1, \ldots, n$ and with initial grid $\hat{\boldsymbol{\gamma}}_b^{N,0}$. This provides B optimal grids $\hat{\boldsymbol{\gamma}}_b^{N,n}$, $b = 1, \ldots, B$ (each of size N).

(S2) Each of these grids is now used to estimate multiple-output conditional quantiles. Working again with the original sample $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, we project the $\boldsymbol{X}$-part onto the grids $\hat{\boldsymbol{\gamma}}_b^{N,n}$, $b = 1, \ldots, B$. Therefore, (5.3.1) provides B estimations, denoted $\hat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{(b)}(\boldsymbol{x}) = \hat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{(b),N,n}(\boldsymbol{x})$ of $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$.

This allows to consider the bootstrap estimator

$$\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{(b)}(\boldsymbol{x}), \tag{5.3.2}$$

obtained by taking the mean of these B estimations. Recall that bootstrapping focuses on the construction of the grids and that we come back to the original sample in the estimation step.

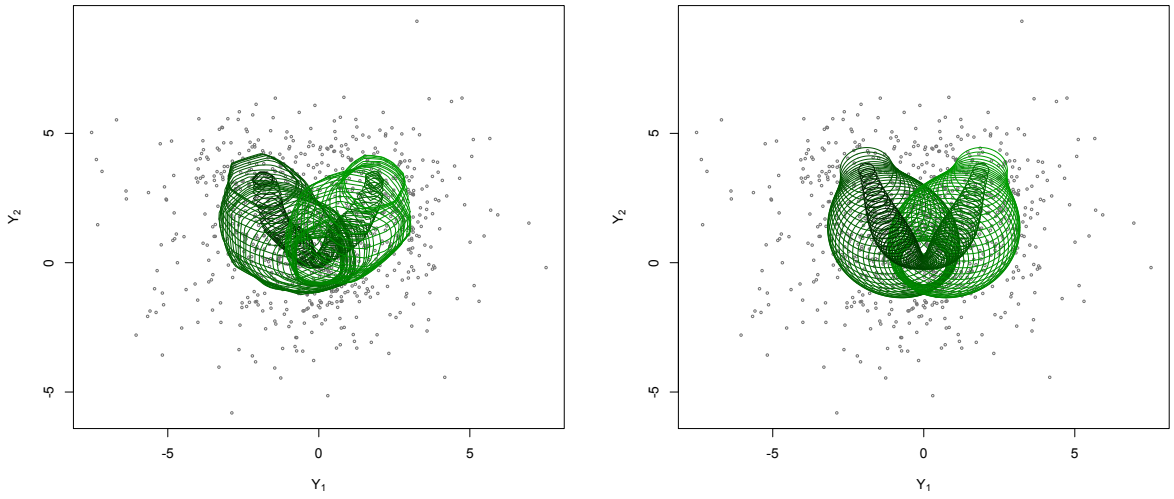
Remember that B should be chosen large enough to clearly smooth quantile contours, but not too large to keep the computational burden under control. In the sequel, we use $B = 100$ throughout. Similarly, the choice of the tuning parameter N has an important impact on the contours. We generalized the selection method of Section 3.2 in this multivariate setting (see Section 6.2 for a detailed description). Figure 5.1 illustrates this method. We generated $n = 999$ points from the model

$$(Y_1, Y_2) = (X, X^2) + \Big(1 + \frac{3}{2}\Big(\sin\Big(\frac{\pi}{2}X\Big)\Big)^2\Big)\boldsymbol{\varepsilon}, \tag{5.3.3}$$

where $X \sim U[-2, 2]$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)^2$ are independent. Figure 5.1a represents the regression quantile contours estimated with our method for $\alpha = 0.2$ et $\alpha = 0.4$. More precisely, there are the intersections of these contours with hyperplanes orthogonal to the $\boldsymbol{x}$-axis at fixed $\boldsymbol{x}$-values $-1.89, -1.87, \ldots, 1.87, 1.89$. We used 360 equispaced directions $\boldsymbol{u} \in \mathcal{S}^{m-1}$ to obtain results. Figure 5.1b illustrates the corresponding theoretical contours.

5.4 Proofs of Section 5.2

The proof of Theorem 5.2.1 requires several lemmas. First, recall that $G_{a,\boldsymbol{c}}(\boldsymbol{x}) = \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{x}]$ and consider the corresponding quantized quantity $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) = \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]$. Since $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}) = (a_{\boldsymbol{\alpha}}(\boldsymbol{x}), (\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x}))')'$ and $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x}) = (\widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x}), (\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x}))')'$ are defined as the vectors achieving the minimum of $G_{a,\boldsymbol{c}}(\boldsymbol{x})$ and $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ respectively, we naturally start with trying to control the distance between $G_{a,\boldsymbol{c}}(\boldsymbol{x})$ and $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$. This is the object of Lemma 5.4.2 but its proof first requires the following lemma.



(a) (b)

Figure 5.1 – Intersections, with hyperplanes orthogonal to the $\boldsymbol{x}$-axis at fixed $\boldsymbol{x}$-values $-1.89, -1.87, \ldots, 1.87, 1.89$, of (a) the regression quantile contours estimated with $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$, based on $B = 100$ and N selected using the data-driven method of Section 6.2, and (b) their theoretical counterparts. The sample size is $n = 999$, and the quantile levels considered are α=0.2 and 0.4. 360 equispaced directions $\boldsymbol{u} \in \mathcal{S}^{m-1}$ were used to obtain results. See (5.3.3) for the data generating model.

Lemma 5.4.1. *Fix $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$, $a \in \mathbb{R}$ and $\boldsymbol{c} \in \mathbb{R}^{m-1}$. Then, under Assumption (G), $G_{a,\boldsymbol{c}} : \mathbb{R}^d \to \mathbb{R}$ is Lispchitz, with Lipschitz constant equal at most to $[G_{a,\boldsymbol{c}}]_{\text{Lip}} = \max(\alpha, 1-\alpha)(1+|\boldsymbol{c}|)[\boldsymbol{M}]_{\text{Lip}}$.*

Proof of Lemma 5.4.1. Notice that ρ_α is a Lipschitz function with Lipschitz constant $[\rho_\alpha]_{\text{Lip}}$. For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, we have

$$
\begin{aligned}
&|G_{a,\boldsymbol{c}}(\boldsymbol{x}_1) - G_{a,\boldsymbol{c}}(\boldsymbol{x}_2)| \\
&\quad = \left|\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{x}_1] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{x}_2]\right| \\
&\quad = \left|\mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{M}(\boldsymbol{X},\boldsymbol{\varepsilon}) - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}(\boldsymbol{X},\boldsymbol{\varepsilon}) - a)|\boldsymbol{X} = \boldsymbol{x}_1] - \mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{M}(\boldsymbol{X},\boldsymbol{\varepsilon}) - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}(\boldsymbol{X},\boldsymbol{\varepsilon})\right. \\
&\qquad \left. - a)|\boldsymbol{X} = \boldsymbol{x}_2]\right| \\
&\quad \leq [\rho_\alpha]_{\text{Lip}}\left|\mathrm{E}[(\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}')(\boldsymbol{M}(\boldsymbol{x}_1,\boldsymbol{\varepsilon}) - \boldsymbol{M}(\boldsymbol{x}_2,\boldsymbol{\varepsilon}))]\right|,
\end{aligned}
$$

using independence of $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$. Applying Jensen inequality provides

$$
\begin{aligned}
|G_{a,\boldsymbol{c}}(\boldsymbol{x}_1) - G_{a,\boldsymbol{c}}(\boldsymbol{x}_2)| &\leq [\rho_\alpha]_{\text{Lip}}|\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'|\left|\mathrm{E}[\boldsymbol{M}(\boldsymbol{x}_1,\boldsymbol{\varepsilon}) - \boldsymbol{M}(\boldsymbol{x}_2,\boldsymbol{\varepsilon})]\right| \\
&\leq [\rho_\alpha]_{\text{Lip}}|\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'|[\boldsymbol{M}]_{\text{Lip}}|\boldsymbol{x}_1 - \boldsymbol{x}_2| \\
&\leq [\rho_\alpha]_{\text{Lip}}(1+|\boldsymbol{c}|)[\boldsymbol{M}]_{\text{Lip}}|\boldsymbol{x}_1 - \boldsymbol{x}_2|,
\end{aligned}
$$

where the second inequality follows from the discussion below Assumption (G). □

We still need the following lemma to prove Theorem 5.2.1.

Lemma 5.4.2. *Fix $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$ and $\boldsymbol{x} \in S_{\boldsymbol{X}}$. For $N \in \mathbb{N}_0$, let $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}^N = \mathrm{Proj}_{\gamma^N}(\boldsymbol{x})$ and $\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{C}_{\boldsymbol{x}}^N = \{\boldsymbol{z} \in S_{\boldsymbol{X}} : \mathrm{Proj}_{\gamma^N}(\boldsymbol{z}) = \tilde{\boldsymbol{x}}\}$. Let $\boldsymbol{K}(\subset \mathbb{R}^{m-1})$ be compact. Then, under Assumptions (F), (G) and (H),*
(i) $\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}} |\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})| \to 0$ *as* $N \to \infty$*;*
(ii) $\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} |\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})| \to 0$ *as* $N \to \infty$.

Proof of Lemma 5.4.2. (i) Fix $a \in \mathbb{R}$ and $\boldsymbol{c} \in \boldsymbol{K}$. Since $[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]$ is equivalent to $[\boldsymbol{X} \in \boldsymbol{C}_{\boldsymbol{x}}]$, one has

$$\begin{aligned}
&|\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}]| \\
&\quad \le \sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} |\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{z}] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}]|.
\end{aligned}$$

Hence, adding and subtracting $\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}]$ to the difference between $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ and $G_{a,\boldsymbol{c}}(\boldsymbol{x})$ provide

$$\begin{aligned}
&|\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})| \\
&\quad \le |\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}]| \\
&\qquad + |\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{x}]| \\
&\quad \le 2 \sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} |\mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \boldsymbol{z}] - \mathrm{E}[\rho_\alpha(Y_{\boldsymbol{u}} - \boldsymbol{c}'\boldsymbol{Y}_{\boldsymbol{u}}^{\perp} - a)|\boldsymbol{X} = \tilde{\boldsymbol{x}}]| \\
&\quad \le 2 \sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} \big|\mathrm{E}\big[\rho_\alpha\big((\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}')\boldsymbol{M}(\boldsymbol{z},\boldsymbol{\varepsilon}) - a\big)\big] - \mathrm{E}\big[\rho_\alpha\big((\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}')\boldsymbol{M}(\tilde{\boldsymbol{x}},\boldsymbol{\varepsilon}) - a\big)\big]\big|,
\end{aligned}$$

using independence of $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$. Lipschitz property of ρ_α and $\boldsymbol{M}$, and Jensen inequality entail

$$|\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})| \le 2[\rho_\alpha]_{\mathrm{Lip}}|\boldsymbol{u}' - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'|[\boldsymbol{M}]_{\mathrm{Lip}} \sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} |\boldsymbol{z} - \tilde{\boldsymbol{x}}|,$$

hence,

$$\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}} |\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})| \le 2[\rho_\alpha]_{\mathrm{Lip}} D\ [\boldsymbol{M}]_{\mathrm{Lip}} \sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} |\boldsymbol{z} - \tilde{\boldsymbol{x}}|,$$

where $D = \sup_{\boldsymbol{c}\in\boldsymbol{K}}(1+|\boldsymbol{c}|) < \infty$. Remember that $\sup_{\boldsymbol{z}\in\boldsymbol{C}_{\boldsymbol{x}}} |\boldsymbol{z}-\tilde{\boldsymbol{x}}|$ corresponds to the radius $R(\boldsymbol{C}_{\boldsymbol{x}})$ of $\boldsymbol{C}_{\boldsymbol{x}}$, defined by the largest distance between any point of the cell $\boldsymbol{C}_{\boldsymbol{x}}$ and the corresponding center $\tilde{\boldsymbol{x}}$. We proved in Lemma 2.4.2 that $\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} R(\boldsymbol{C}_{\boldsymbol{x}}) \to 0$ as $N \to \infty$. The result then follows.

(ii) For simplicity of notations, we write $\tilde{a} = \widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$ and $\tilde{\boldsymbol{c}} = \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$. First, notice that $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$ are the quantiles of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ and $\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}$ respectively: they are then bounded for any $\boldsymbol{x} \in S_{\boldsymbol{X}}$. We aim to prove that, for N_0 sufficiently large, there exists $\boldsymbol{K} = \boldsymbol{K}_{N_0}$ a compact set such that, for all $N \ge N_0$ and for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$ belong to $\boldsymbol{K}$.

Let us start with $\boldsymbol{x}$ fixed in $S_{\boldsymbol{X}}$. We know that $(a,\boldsymbol{c}')' \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$ and $(a,\boldsymbol{c}')' \mapsto \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ are strictly convex functions and we showed in Point (i) that $|\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})| \to 0$ as $N \to \infty$ uniformly in $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $a \in \mathbb{R}$ and $\boldsymbol{c}$ belonging to any compact set. We consider now a compact

set containing $\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$. Therefore, it implies that the minimizer of $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ must be in the same compact set for N sufficiently large, that we denote $\boldsymbol{K}(\boldsymbol{x})$.

We now want to prove that we can find a compact set working for any $\boldsymbol{x} \in S_{\boldsymbol{X}}$. The set $\cup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \boldsymbol{K}(\boldsymbol{x})$ is a natural candidate and we have to ensure that it is bounded. Since $\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is bounded for all $\boldsymbol{x}$, a sufficient condition in the case $m = 2$ is that

$$\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \frac{\partial^2 G_{a,c}(\boldsymbol{x})}{\partial c^2}\Big|_{a=a_{\boldsymbol{\alpha}}(\boldsymbol{x}),c=c_{\boldsymbol{\alpha}}(\boldsymbol{x})} > 0.$$

Remember that $\boldsymbol{H}^{\boldsymbol{c}}(G_{a,\boldsymbol{c}}(\boldsymbol{x}))$ is a positive definite matrix. Therefore, in the case $m = 2$, it implies that $\frac{\partial^2 G_{a,c}(\boldsymbol{x})}{\partial c^2}\Big|_{a=a_{\boldsymbol{\alpha}}(\boldsymbol{x}),c=c_{\boldsymbol{\alpha}}(\boldsymbol{x})} > 0$. Since $S_{\boldsymbol{X}}$ is compact, the sufficient condition above is met. This argument can be generalized in higher dimensional setting where it becomes

$$\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \boldsymbol{v}'\boldsymbol{H}^{\boldsymbol{c}}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)\Big|_{a=a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}=\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})} \boldsymbol{v} > 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^{m-1}, |\boldsymbol{v}| > 0.$$

Since $\boldsymbol{H}^{\boldsymbol{c}}\big(G_{a,\boldsymbol{c}}(\boldsymbol{x})\big)$ is positive definite for all $a \in \mathbb{R}$ and $\boldsymbol{c} \in \mathbb{R}^{m-1}$, and since $S_{\boldsymbol{X}}$ is compact, this condition is met.

Therefore, there exists a compact set $\boldsymbol{K}$ such that, for all $N \geq N_0$ and for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$ belong to $\boldsymbol{K}$. We denote this compact $\boldsymbol{K}_{\boldsymbol{\alpha}}$ in the sequel.

Then, with $\mathbb{I}_+ = \mathbb{I}_{[\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) \geq \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})]}$, we have

$$\begin{aligned}
&\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})\Big|\mathbb{I}_+ = (\widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}))\mathbb{I}_+ \\
&\quad \leq \big(\widetilde{G}_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\tilde{\boldsymbol{x}}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})\big)\mathbb{I}_+ \leq \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}_{\boldsymbol{\alpha}}} |\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})|\mathbb{I}_+. \qquad (5.4.1)
\end{aligned}$$

Now, with $\mathbb{I}_- = \mathbb{I}_{[\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) < \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})]}$, we have that,

$$\begin{aligned}
&\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})\Big|\mathbb{I}_- = \big(G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})\big)\mathbb{I}_- \\
&\quad \leq \big(G_{\tilde{a},\tilde{\boldsymbol{c}}}(\boldsymbol{x}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})\big)\mathbb{I}_- \leq \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}_{\boldsymbol{\alpha}}} |\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})|\mathbb{I}_-, \qquad (5.4.2)
\end{aligned}$$

so that

$$\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} G_{a,\boldsymbol{c}}(\boldsymbol{x})\Big| \leq \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}_{\boldsymbol{\alpha}}} |\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - G_{a,\boldsymbol{c}}(\boldsymbol{x})|.$$

The result then directly follows from Point (i). □

We can now prove Theorem 5.2.1.

Proof of Theorem 5.2.1. First note that, for any $\boldsymbol{x} \in S_{\boldsymbol{X}}$ and for $N \geq N_0$,

$$\begin{aligned}
&|G_{\tilde{a},\tilde{\boldsymbol{c}}}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})| \\
&\quad \leq |G_{\tilde{a},\tilde{\boldsymbol{c}}}(\boldsymbol{x}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})| + |\widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})| \\
&\quad \leq \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}_{\boldsymbol{\alpha}}} |G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| + |\min_{a,\boldsymbol{c}} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{a,\boldsymbol{c}} G_{a,\boldsymbol{c}}(\boldsymbol{x})| \\
&\quad \leq \sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \sup_{a\in\mathbb{R},\boldsymbol{c}\in\boldsymbol{K}_{\boldsymbol{\alpha}}} |G_{a,\boldsymbol{c}}(\boldsymbol{x}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| + \sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} |\min_{a,\boldsymbol{c}} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \min_{a,\boldsymbol{c}} G_{a,\boldsymbol{c}}(\boldsymbol{x})|
\end{aligned}$$

Therefore, Lemma 5.4.2(i) and (ii) implies that

$$\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} |G_{\tilde{a},\tilde{\boldsymbol{c}}}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})| \to 0, \tag{5.4.3}$$

as $N \to \infty$.

Now, let N_1 be such that, for any $N \geq \max(N_0, N_1)$, we have

$$|G_{\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})| \leq 1,$$

for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$. We will show later that it entails that there exists M such that

$$|\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})| \leq M, \tag{5.4.4}$$

for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$ and $N \geq \max(N_0, N_1)$.

Performing a second-order expansion about $(a, \boldsymbol{c}')' = (\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}))'$ provides

$$G_{\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) = \frac{1}{2}\big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big)' \boldsymbol{H}(G_{a,\boldsymbol{c}}(\boldsymbol{x}))\Big|_{a=a_*,\boldsymbol{c}=\boldsymbol{c}_*} \big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big),$$

where $\boldsymbol{q}^N_*(\boldsymbol{x}) = (a_*, \boldsymbol{c}'_*)' = (a_*(\boldsymbol{x}), \boldsymbol{c}'_*(x))'$ is such that $\boldsymbol{q}^N_*(\boldsymbol{x}) = \theta \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}) + (1-\theta)\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$, with $\theta \in (0,1)$. Since $(a, \boldsymbol{c}) \mapsto G_{a,\boldsymbol{c}}(\boldsymbol{x})$ is convex, this Hessian matrix, denoted $\boldsymbol{H}$ in the sequel, is positive definite and invertible. Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $\boldsymbol{H}$, sorted in descending order, and let $\boldsymbol{O}$ be an orthogonal matrix such that $\boldsymbol{H}$ can be decomposed as $\boldsymbol{H} = \boldsymbol{O}'\boldsymbol{\Lambda}\boldsymbol{O}$, with $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues. We then have

$$\begin{aligned}
G_{\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) &= \frac{1}{2}\big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big)' \boldsymbol{H} \big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big) \\
&= \frac{1}{2}\sum_{j=1}^{m} \lambda_j \Big(\big(\boldsymbol{O}'\big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big)\big)_j\Big)^2 \geq \frac{1}{2}\lambda_m \sum_{i=1}^{m} \Big(\big(\boldsymbol{O}'\big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big)\big)_j\Big)^2 \\
&= \frac{1}{2}\lambda_m \big|\boldsymbol{O}'\big(\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big)\big|^2 = \frac{1}{2}\lambda_m \big|\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big|^2,
\end{aligned}$$

hence,

$$\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \big|\widetilde{\boldsymbol{q}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big|^2 \leq \frac{2}{\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \lambda_m} \sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} |G_{\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - G_{a_{\boldsymbol{\alpha}}(\boldsymbol{x}),\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})|. \tag{5.4.5}$$

We then have to ensure that $\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \lambda_m > 0$. Since $\lambda_m = \lambda_m(\boldsymbol{x})$ has strictly positive values and the function that associates to a matrix its eigenvalues is continuous, it suffices to show that $\boldsymbol{x} \mapsto \boldsymbol{H}$ is continuous. It is enough to prove that each element of $\boldsymbol{H}$ is continuous in $\boldsymbol{x}$. Since $\boldsymbol{M}_1(\boldsymbol{x})$ and $\boldsymbol{M}_2(\boldsymbol{x})$ are Lipschitz functions, and since the density $f^{\boldsymbol{\varepsilon}}(\cdot)$ is also continuous by Assumption (H), the integrands of each element of $\boldsymbol{H}$ are continuous in $\boldsymbol{x}$ and it remains to ensure that the integrals with respect to $\boldsymbol{t}$ are still continuous in $\boldsymbol{x}$. This is achieved if the integrand is dominated by some integrable function that only depends on $\boldsymbol{t}$ (see Briane and Pagès (2012, Theorem 8.5)). Let us first consider the integrand of the (1,1)-entry of $\boldsymbol{H}$, denoted

$f^{1,1}(\boldsymbol{x},\boldsymbol{t})$. We have

$$\begin{aligned} f^{1,1}(\boldsymbol{x},\boldsymbol{t}) &= f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big) \\ &= \frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))} f^{\boldsymbol{\varepsilon}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-\boldsymbol{M}_1(\boldsymbol{x}))\big) \\ &= \frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))}\Big(f^{\boldsymbol{\varepsilon}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-\boldsymbol{M}_1(\boldsymbol{x}))\big)-f^{\boldsymbol{\varepsilon}}((1,\boldsymbol{t}')')\Big) \\ &\quad + \frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))} f^{\boldsymbol{\varepsilon}}((1,\boldsymbol{t}')'). \end{aligned}$$

Since $f^{\boldsymbol{\varepsilon}}(\cdot)$ is a density, one directly has that $f^{\boldsymbol{\varepsilon}}((1,\boldsymbol{t}')')$ is an integrable function and it remains to deal with the first part of this last equation that rewrites as

$$\begin{aligned} &\frac{1}{\det(\boldsymbol{M}_2(\boldsymbol{x}))}\Big(f^{\boldsymbol{\varepsilon}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-\boldsymbol{M}_1(\boldsymbol{x}))\big)-f^{\boldsymbol{\varepsilon}}((1,\boldsymbol{t}')')\Big) \\ &\quad = f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)-f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}(1,\boldsymbol{t}')'+\boldsymbol{M}_1(\boldsymbol{x})\big). \end{aligned}$$

Using Assumption (F)', the compactness of $S_{\boldsymbol{X}}$ and the Lipschitz property of $\boldsymbol{M}_1(\boldsymbol{x})$ and $\boldsymbol{M}_2(\boldsymbol{x})$ entails that there exist some constants $C>0$, $r>m-1$ and $s>0$ (independent of $\boldsymbol{x}$) such that

$$\begin{aligned} &\big|f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}\big)-f^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}\big((\boldsymbol{M}_2(\boldsymbol{x}))^{-1}(1,\boldsymbol{t}')'+\boldsymbol{M}_1(\boldsymbol{x})\big)\big| \\ &\quad \le C\frac{|(a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}-(\boldsymbol{M}_2(\boldsymbol{x}))^{-1}(1,\boldsymbol{t}')'-\boldsymbol{M}_1(\boldsymbol{x})|^s}{\Big(1+\Big|\frac{(a_*+\boldsymbol{c}_*'\boldsymbol{t})\boldsymbol{u}+\boldsymbol{\Gamma}_{\boldsymbol{u}}\boldsymbol{t}+(\boldsymbol{M}_2(\boldsymbol{x}))^{-1}(1,\boldsymbol{t}')'+\boldsymbol{M}_1(\boldsymbol{x})}{2}\Big|^2\Big)^{(3+r+s)/2}} \\ &\quad \le C\frac{|\boldsymbol{A}\boldsymbol{t}+\boldsymbol{b}|^s}{\big(1+|\boldsymbol{D}\boldsymbol{t}+\boldsymbol{e}|^2\big)^{(3+r+s)/2}} =: g(\boldsymbol{t}). \end{aligned}$$

The integrability of this function $g(\boldsymbol{t})$ requires that $3+r+s-s>m$, i.e. $r>m-3$, which results from Assumption (F)'. Therefore, the first entry of $\boldsymbol{x}\mapsto\boldsymbol{H}$ is continuous.

The other entries of this Hessian matrix can be treated in the exact same way to obtain a domination. The integrability condition however slightly changes, since the $\boldsymbol{t}$ or $\boldsymbol{t}\boldsymbol{t}'$ factor increases the power of the numerator. We obtain respectively that r should be greater than $m-2$ or $m-1$, which is ensured by Assumption (F)'. Thus, $\boldsymbol{x}\mapsto\boldsymbol{H}$ is continuous, hence $\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}}\lambda_m>0$. Using this and (5.4.3), we conclude from (5.4.5) that

$$\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}}\big|\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})-\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})\big|^2\to 0,$$

as $N\to\infty$, which has to be proved.

It remains to show the claim in (5.4.4). First notice that $|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})-\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})|\le|a_{\boldsymbol{\alpha}}(\boldsymbol{x})-\widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})|+|\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})-\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})|$ and we proved in Lemma 5.4.2(ii) that there exists some compact set $\boldsymbol{K}_{\boldsymbol{\alpha}}$ that contains $\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$ for $N\ge N_0$ and for all $\boldsymbol{x}\in S_{\boldsymbol{X}}$. Therefore, for $N\ge N_0$, there exists $M_1>0$ such that $|\boldsymbol{c}_{\boldsymbol{\alpha}}(\boldsymbol{x})-\widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})|\le M_1$ for all $\boldsymbol{x}\in S_{\boldsymbol{X}}$.

The same argument can be used analogously to deal with $|a_{\boldsymbol{\alpha}}(\boldsymbol{x})-\widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})|$. Indeed, using the strict convexity of $G_{a,\boldsymbol{c}}(\boldsymbol{x})$ and $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ and Lemma 5.4.2(i) entails, as in Lemma 5.4.2(ii), that there exists for all $\boldsymbol{x}\in S_{\boldsymbol{X}}$, a compact set $\boldsymbol{J}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ containing $a_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$ for N sufficiently

large. We now want to prove that $\boldsymbol{J}_{\boldsymbol{\alpha}} = \cup_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \boldsymbol{J}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is bounded. Given that $a \in \mathbb{R}$, a sufficient condition is that

$$\inf_{\boldsymbol{x}\in S_{\boldsymbol{X}}} \frac{\partial^2 G_{a,c}(\boldsymbol{x})}{\partial a^2}\Big|_{a=a_{\boldsymbol{\alpha}}(\boldsymbol{x}),c=c_{\boldsymbol{\alpha}}(\boldsymbol{x})} > 0,$$

which is direct from Lemma 4.4.1 since $S_{\boldsymbol{X}}$ is compact and $f^{\boldsymbol{\epsilon}}(\cdot) > 0$. Consequently, for all $N \geq N_1$, the compact set $\boldsymbol{J}_{\boldsymbol{\alpha}}$ contains $a_{\boldsymbol{\alpha}}(\boldsymbol{x})$ and $\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$, which implies that $|a_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})| \leq M_2$ for some $M_2 > 0$ and for all $\boldsymbol{x} \in S_{\boldsymbol{X}}$, hence (5.4.4). □

5.5 Proofs of Section 5.3

The proof of Theorem 5.3.1 requires two lemmas. The first one is actually Lemma 2.5.1. In the sequel, we write $\boldsymbol{\gamma}^N = (\tilde{\boldsymbol{x}}^N_1, \ldots, \tilde{\boldsymbol{x}}^N_N)$.

Lemma 5.5.1. *Let Assumption (I) hold. Fix $N \in \mathbb{N}_0$ and $\boldsymbol{x} \in S_{\boldsymbol{X}}$. We write $\tilde{\boldsymbol{x}}^N = \tilde{\boldsymbol{x}} = \mathrm{Proj}_{\boldsymbol{\gamma}^N}(\boldsymbol{x})$ and $\hat{\boldsymbol{x}}^{N,n} = \hat{\boldsymbol{x}} = \mathrm{Proj}_{\hat{\boldsymbol{\gamma}}^{N,n}}(\boldsymbol{x})$. Then, with $\widehat{\boldsymbol{X}}^{N,n}_i = \widehat{\boldsymbol{X}}^N_i = \mathrm{Proj}_{\hat{\boldsymbol{\gamma}}^{N,n}}(\boldsymbol{X}_i)$, $i = 1, \ldots, n$, we have*

(i) $\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[\hat{\boldsymbol{X}}^N_i = \hat{\boldsymbol{x}}^N]} \xrightarrow[n\to\infty]{a.s.} P[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]$;

(ii) after possibly reordering the $\tilde{\boldsymbol{x}}^N_i$'s, $\hat{\boldsymbol{x}}^{N,n}_i \xrightarrow[n\to\infty]{a.s.} \tilde{\boldsymbol{x}}^N_i$, $i = 1, \ldots, N$ (hence, $\hat{\boldsymbol{\gamma}}^{N,n} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\gamma}^N$).

Lemma 5.5.2. *Fix $\boldsymbol{\alpha} = \alpha\boldsymbol{u} \in \mathcal{B}^m$, $\boldsymbol{x} \in S_{\boldsymbol{X}}$ and $N \in \mathbb{N}_0$. Let $\boldsymbol{K}$ ($\subset \mathbb{R}^m$) be compact and define*

$$\widehat{G}^{N,n}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) = \widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) := \frac{\frac{1}{n}\sum_{i=1}^n \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y}_i - a)\,\mathbb{I}_{[\hat{\boldsymbol{X}}^N_i = \hat{\boldsymbol{x}}]}}{\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[\hat{\boldsymbol{X}}^N_i = \hat{\boldsymbol{x}}]}}.$$

Then, under Assumptions (F), (G) and (H),

(i) $\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}} |\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| = o_{\mathrm{P}}(1)$ *as* $n \to \infty$;

(ii) $|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m} \widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{a\in\mathbb{R}} \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| = o_{\mathrm{P}}(1)$ *as* $n \to \infty$;

(iii) $|\widetilde{G}_{\widetilde{a}^N_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^N_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\tilde{\boldsymbol{x}}) - \widetilde{G}_{\widehat{a}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widehat{\boldsymbol{c}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\tilde{\boldsymbol{x}})| = o_{\mathrm{P}}(1)$ *as* $n \to \infty$.

Proof. (i) Notice first that

$$\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) = \mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y} - a)|\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}] = \frac{\mathrm{E}[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y} - a)\mathbb{I}_{[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]}]}{P[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]}$$

by definition of the conditional expectation. Therefore, using Lemma 5.5.1(i), it is sufficient to show that

$$\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}} \Big|\frac{1}{n}\sum_{i=1}^n \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y}_i - a)\,\mathbb{I}_{[\hat{\boldsymbol{X}}^N_i = \hat{\boldsymbol{x}}]} - \mathrm{E}\big[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y} - a)\mathbb{I}_{[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]}\big]\Big| = o_{\mathrm{P}}(1),$$

as $n \to \infty$. As for the single-output case, it is natural to decompose it as

$$\begin{aligned}
&\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}} \Big|\frac{1}{n}\sum_{i=1}^n \rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y}_i - a)\,\mathbb{I}_{[\hat{\boldsymbol{X}}^N_i = \hat{\boldsymbol{x}}]} - \mathrm{E}\big[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}'_{\boldsymbol{u}}\boldsymbol{Y} - a)\mathbb{I}_{[\widetilde{\boldsymbol{X}}^N = \tilde{\boldsymbol{x}}]}\big]\Big| \\
&\qquad \leq \sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}} |T_{a\boldsymbol{c}1}| + \sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}} |T_{a\boldsymbol{c}2}|,
\end{aligned}$$

with

$$T_{a\boldsymbol{c}1} = \Big|\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}_i - a)\big(\mathbb{I}_{[\hat{\boldsymbol{X}}_i^N=\hat{\boldsymbol{x}}]} - \mathbb{I}_{[\tilde{\boldsymbol{X}}_i^N=\tilde{\boldsymbol{x}}]}\big)\Big|,$$

and

$$T_{a\boldsymbol{c}2} = \Big|\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}_i - a)\mathbb{I}_{[\tilde{\boldsymbol{X}}_i^N=\tilde{\boldsymbol{x}}]} - \mathrm{E}\big[\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a)\mathbb{I}_{[\tilde{\boldsymbol{X}}^N=\tilde{\boldsymbol{x}}]}\big]\Big|,$$

with $\widetilde{\boldsymbol{X}}_i^N = \mathrm{Proj}_{\boldsymbol{\gamma}^N}(\boldsymbol{X}_i)$, $i = 1, \ldots, n$.

For all $(a, \boldsymbol{c}')' \in \boldsymbol{K}$, one has

$$\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a) \le \max(\alpha, 1-\alpha)|\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a|$$

$$\le \max(\alpha, 1-\alpha)\big(|\boldsymbol{u}'\boldsymbol{M}_1(\boldsymbol{X})| + ||\boldsymbol{u}'\boldsymbol{M}_2(\boldsymbol{X})||\,|\boldsymbol{\varepsilon}| + |\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}_1(\boldsymbol{X})|\,|\boldsymbol{c}| + ||\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}_2(\boldsymbol{X})||\,|\boldsymbol{c}|\,|\boldsymbol{\varepsilon}| + |a|\big).$$

Since $\boldsymbol{u}$ (and then $\boldsymbol{\Gamma}_{\boldsymbol{u}}$) is fixed, and using Assumption (G) setting that $\boldsymbol{M}_1(\boldsymbol{X})$ and $\boldsymbol{M}_2(\boldsymbol{X})$ are continuous function defined over the compact set $S_{\boldsymbol{X}}$, there exist positive constants C_1 and C_2 such that

$$\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y} - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y} - a) \le C_1 + C_2|\boldsymbol{\varepsilon}|$$

for all $(a, \boldsymbol{c}') \in \boldsymbol{K}$, then is in L_1 since $\boldsymbol{\varepsilon}$ is assumed to be in L_p, $p = 2$. The uniform law of large numbers (see, e.g., Theorem 16(a) in Ferguson (1996)) then implies that $\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}}|T_{a\boldsymbol{c}2}| = o_{\mathrm{P}}(1)$ as $n \to \infty$.

It remains to treat $T_{a\boldsymbol{c}1}$. Denote by $\ell_n = \{i = 1, \ldots, n : \mathbb{I}_{[\hat{\boldsymbol{X}}_i^N=\hat{\boldsymbol{x}}]} \neq \mathbb{I}_{[\tilde{\boldsymbol{X}}_i^N=\tilde{\boldsymbol{x}}]}\}$ the set collecting the indices of observations that are projected on the same point as $\boldsymbol{x}$ for $\boldsymbol{\gamma}^N$ but not for $\hat{\boldsymbol{\gamma}}^{N,n}$, and inversely. Then, for any $(a, \boldsymbol{c}')' \in \boldsymbol{K}$,

$$|T_{a\boldsymbol{c}1}| \le \frac{1}{n}\sum_{i\in\ell_n}|\rho_\alpha(\boldsymbol{u}'\boldsymbol{Y}_i - \boldsymbol{c}'\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{Y}_i - a)|$$

$$\le \frac{\max(\alpha, 1-\alpha)}{n}\sum_{i\in\ell_n}\big(|\boldsymbol{u}'\boldsymbol{M}_1(\boldsymbol{X}_i)| + ||\boldsymbol{u}'\boldsymbol{M}_2(\boldsymbol{X}_i)||\,|\boldsymbol{\varepsilon}_i| + |\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}_1(\boldsymbol{X}_i)|\,|\boldsymbol{c}| + ||\boldsymbol{\Gamma}_{\boldsymbol{u}}'\boldsymbol{M}_2(\boldsymbol{X}_i)||\,|\boldsymbol{c}|\,|\boldsymbol{\varepsilon}_i|$$

$$+ |a|\big)$$

$$\le \frac{\#\ell_n}{n} \times \frac{1}{\#\ell_n}\sum_{i\in\ell_n}(C_1 + C_2|\boldsymbol{\varepsilon}_i|) =: S_1 \times S_2,$$

using similar argument as for $T_{a\boldsymbol{c}2}$. Lemma 5.5.1(ii) implies that $S_1 = \#\ell_n/n = o_{\mathrm{P}}(1)$ as $n \to \infty$. Concerning S_2, notice that ℓ_n is measurable with respect to the $\boldsymbol{X}_i$'s and then independent of $\boldsymbol{\varepsilon}_i$ for $i = 1, \ldots, n$. This entails that $\mathrm{E}[S_2] = O(1)$ as $n \to \infty$ and then S_2 is bounded in probability. Therefore, $\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}}|T_{a\boldsymbol{c}1}| = o_{\mathrm{P}}(1)$, the expected result.

(ii) Fix $\delta > 0$ and $\eta > 0$. For simplicity of notations, we write $\tilde{a} = \widetilde{a}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$, $\tilde{\boldsymbol{c}} = \widetilde{\boldsymbol{c}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})$, $\hat{a} = \widehat{a}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ and $\hat{\boldsymbol{c}} = \widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$. We first choose N_1 and M large enough to have $|\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^N(\boldsymbol{x})| \le M$ and $P[|\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})| > M] < \eta/2$ for any $N \ge N_1$. Indeed, $\widehat{\boldsymbol{q}} = \widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x}) = \big(\hat{a}, \hat{\boldsymbol{c}}'\big)'$ is the sample quantile of a number of $\boldsymbol{Y}_i$'s that increases to infinity so that, with arbitrary large probability

for n large, $|\widehat{\boldsymbol{q}}|$ cannot exceed $2\sup_{\boldsymbol{x}\in S_{\boldsymbol{X}}}|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})|$. Define $\boldsymbol{K}_M := \{\boldsymbol{z}\in\mathbb{R}^m : |\boldsymbol{z}|\le M\}$. Then, with $\mathbb{I}_+ = \mathbb{I}_{[\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}})\ge\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})]}$, we have

$$\begin{aligned}\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big|\mathbb{I}_+ &= (\widehat{G}_{\hat{a},\hat{\boldsymbol{c}}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}))\mathbb{I}_+\\ \le \big(\widehat{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})\big)\mathbb{I}_+ &\le \sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}_M}|\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})|\mathbb{I}_+. \end{aligned}\tag{5.5.1}$$

Now, with $\mathbb{I}_- = \mathbb{I}_{[\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}})<\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})]}$, we have that, under $|\widehat{\boldsymbol{q}}|\le M$,

$$\begin{aligned}\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big|\mathbb{I}_- &= \big(\widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widehat{G}_{\hat{a},\hat{\boldsymbol{c}}}(\hat{\boldsymbol{x}})\big)\mathbb{I}_-\\ \le \big(\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widehat{G}_{\hat{a},\hat{\boldsymbol{c}}}(\hat{\boldsymbol{x}})\big)\mathbb{I}_- &\le \sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}_M}|\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})|\mathbb{I}_-. \end{aligned}\tag{5.5.2}$$

By combining (5.5.1) and (5.5.2), we obtain that, under $|\widehat{\boldsymbol{q}}|\le M$,

$$\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big| \le \sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}_M}|\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})|.$$

Therefore, for any $N\ge N_1$, we get

$$\begin{aligned}&P\Big[\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big| > \delta\Big]\\ &= P\Big[\Big|\min_{a,\boldsymbol{c}}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{a,\boldsymbol{c}}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big| > \delta, |\widehat{\boldsymbol{q}}|\le M\Big] + P\Big[\Big|\min_{a,\boldsymbol{c}}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{a,\boldsymbol{c}}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big| > \delta, |\widehat{\boldsymbol{q}}| > M\Big]\\ &\le P\Big[\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}_M}|\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| > \delta\Big] + \frac{\eta}{2}.\end{aligned}$$

Point (i) of the lemma implies that the first term is smaller than $\eta/2$ for any $N\ge N_2$. Consequently, for any $N\ge N_0 = \max(N_1, N_2)$, we have

$$P\Big[\Big|\min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{(a,\boldsymbol{c}')'\in\mathbb{R}^m}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})\Big| > \delta\Big] < \eta.$$

(iii) This proof proceeds in the same way as (ii). We start with picking N_1 and M large enough so that $P[|\widehat{\boldsymbol{q}}| > M] < \eta/2$ for any $N\ge N_1$, with η fixed. This yields

$$P\Big[|\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})| > \delta\Big] \le P\Big[|\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})| > \delta, |\widehat{\boldsymbol{q}}|\le M\Big] + \frac{\eta}{2}.\tag{5.5.3}$$

Using now triangular inequality entails

$$\begin{aligned}&P\Big[|\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})| > \delta, |\widehat{\boldsymbol{q}}|\le M\Big]\\ &\le P\Big[|\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}}) - \widehat{G}_{\hat{a},\hat{\boldsymbol{c}}}(\hat{\boldsymbol{x}})| > \delta/2, |\widehat{\boldsymbol{q}}|\le M\Big] + P\Big[|\widehat{G}_{\hat{a},\hat{\boldsymbol{c}}}(\hat{\boldsymbol{x}}) - \widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})| > \delta/2, |\widehat{\boldsymbol{q}}|\le M\Big]\\ &\le P\Big[\sup_{(a,\boldsymbol{c}')'\in\boldsymbol{K}_M}|\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}}) - \widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}})| > \delta/2\Big] + P\Big[|\min_{a,\boldsymbol{c}}\widehat{G}_{a,\boldsymbol{c}}(\hat{\boldsymbol{x}}) - \min_{a,\boldsymbol{c}}\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})| > \delta/2\Big].\end{aligned}$$

Points (i) and (ii) of the lemma imply that both probabilities can be made arbitrarily small for N large enough. Combining this with (5.5.3) yields

$$|\widetilde{G}_{\widehat{a}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widehat{\boldsymbol{c}}^{N,n}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x}) - \widetilde{G}_{\widetilde{a}^{N}_{\boldsymbol{\alpha}}(\boldsymbol{x}),\widetilde{\boldsymbol{c}}^{N}_{\boldsymbol{\alpha}}(\boldsymbol{x})}(\boldsymbol{x})| \to 0,$$

as $N\to\infty$. □

Thanks to these lemmas, the proof of Theorem 5.3.1 is now direct and analogous to the single-output counterpart.

Proof of Theorem 5.3.1. The strict convexity of the loss function $\rho_\alpha(\cdot)$ entails the strict convexity of $\widetilde{G}_{a,\boldsymbol{c}}(\tilde{\boldsymbol{x}})$ in $(a, \boldsymbol{c}')'$. Then, its minimum in $(a, \boldsymbol{c}')'$ (for any fixed $\boldsymbol{x}$) is unique. Therefore, the convergence in probability of $\widetilde{G}_{\hat{a},\hat{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})$ toward $\widetilde{G}_{\tilde{a},\tilde{\boldsymbol{c}}}(\tilde{\boldsymbol{x}})$ implies the convergence in probability of the corresponding arguments. □

5.6 Final comments

As Chapter 2 for the single-output framework, this chapter mainly focused on the theoretical aspects of multiple-output conditional quantile estimation based on optimal quantization. We first replaced $\boldsymbol{X}$ by its quantized version in the Definition 4.2.1 of multiple-output conditional quantile (in the sense of HPŠ10) and we evaluated how well this quantity approximates the quantiles. Theorem 5.2.1 states the fixed-$\boldsymbol{x}$ convergence of $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})$. However, no convergence rate (and global-$\boldsymbol{X}$ result) could have been derived. In a second step, we introduced a new estimator $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ of multiple-output conditional quantile from this approximation and we showed the convergence in probability of this estimator $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ toward the approximation $\widetilde{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N}(\boldsymbol{x})$. These results were obtained for any dimension d of the covariate $\boldsymbol{X}$ and any dimension m (≥ 2) of the response vector $\boldsymbol{Y}$. The required assumptions were quite mild and generalized the ones necessary in the single-output part. Again, the convergence result for the estimator is only valid if the quantization is based on $p = 2$.

The following step consists in investigating the numerical performances of this new estimator (and particularly its bootstrap version $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$). This is the goal of the next chapter where our new method will be widely applied on several samples and for different sample sizes. We will first extend the data-driven selection method for N of Section 3.2 in this multiple-output setting. We will then compare our estimator with the main existing competitors, i.e. the local constant and local bilinear estimators of HPŠ15. This comparison will be based both on graphical inspection of estimated quantile contours (as the ones of Figure 5.3.3) and on empirical integrated square errors.

6 Numerical study of the estimator

Contents

6.1 Introduction

In Chapter 5, we introduced a new estimator of multiple-output conditional quantile based on optimal quantization, and we investigated it from a theoretical point of view. The aim of this chapter is to complete this investigation with a numerical angle.

We mentioned that the choice of the tuning parameter N is crucial. Since the data-driven method developed in Section 3.2 was really satisfactory in the single-output framework, Section 6.2 aims to extend it in this context.

With this efficient selection method of the tuning parameter, our estimating procedure is entirely functional. We then compare it with alternative estimators of multiple-output conditional quantiles in Section 6.3. We first define precisely the competitors considered in Section 6.3.1. We then compare them with our method both from a graphical point of view (Section 6.3.2) and by comparing the empirical integrated square errors (Section 6.3.3).

Finally, we illustrate our estimator on a real data example in Section 6.4 and we conclude the chapter with some final comments. This chapter constitutes the second part of the working paper *Multiple-output quantile regression through optimal quantization* with Davy Paindaveine and Jérôme Saracco.

6.2 Data-driven selection method of the tuning parameter N

This section aims to extend the N-selection criterion of Section 3.2 in this multiple-output context. This criterion was based on the minimization of an empirical integrated square error (ISE) quantity that is essentially convex in N, which allows to identify an optimal value N_{opt} of N. Since this empirical ISE provided very satisfying results, we expect that it will be similar here and we directly generalize the data-driven selection method (without comparing it with some infeasible criterion as in Section 3.2.1).

A major difference between the single and the multiple-output cases lies in the order of the quantile. We now consider $\boldsymbol{\alpha} = \alpha \boldsymbol{u}$, with $\alpha \in (0,1)$ and $\boldsymbol{u} \in \mathcal{S}^{m-1}$. We are actually interested in estimating $\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x})$ for many $\boldsymbol{u}$ ranging in $\mathcal{S}^{m-1}$ to obtain regression quantile contours (see Figure 5.1). Therefore, the criterion developed below contains an additional mean over the different directions $\boldsymbol{u}$ considered.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J$ be a finite set of J points of interest in $S_{\boldsymbol{X}}$ and $\mathcal{U}$ containing 40 equispaced directions $\boldsymbol{u} \in \mathcal{S}^{m-1}$. The procedure works again in two steps.

Generating $\widetilde{B}$ grids

For some integer $\widetilde{B}$, we first generate $\widetilde{B}$ samples of size n with replacement from the initial sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, denoted $(\boldsymbol{\xi}^t_{B+\tilde{b}})_t$, for $\tilde{b} = 1, \ldots, \widetilde{B}$. We also generate initial grids $\hat{\boldsymbol{\gamma}}^{N,0}_{B+\tilde{b}}$, $\tilde{b} = 1, \ldots, \widetilde{B}$, as before, by sampling randomly among the corresponding $(\boldsymbol{\xi}^t_{B+\tilde{b}})_t$ under the constraint that the N values are pairwise distinct. We then perform $\widetilde{B}$ times CLVQ with iterations based on $\boldsymbol{\xi}^t_{B+\tilde{b}}$, $t = 1, \ldots, n$, and with initial grid $\hat{\boldsymbol{\gamma}}^{N,0}_{B+\tilde{b}}$. This provides $\widetilde{B}$ optimal grids $\hat{\boldsymbol{\gamma}}^{N,n}_{B+\tilde{b}}$, $\tilde{b} = 1, \ldots, \widetilde{B}$.

Bootstrap based selection method for N

Each of these grids is now used to estimate multiple-output conditional quantiles. Working again with the original sample $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, we project the $\boldsymbol{X}$-part onto the grids $\hat{\boldsymbol{\gamma}}^{N,n}_{B+\tilde{b}}$, $\tilde{b} = 1, \ldots, \widetilde{B}$. Therefore, for all $j = 1, \ldots, J$, (5.3.1) provides $\widetilde{B}$ estimations, denoted $\hat{\boldsymbol{q}}^{(B+\tilde{b})}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j) = \hat{\boldsymbol{q}}^{(B+\tilde{b}),N,n}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j)$. It allows to consider the square difference between $\bar{\boldsymbol{q}}^{N,n}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j)$ and $\hat{\boldsymbol{q}}^{(B+\tilde{b})}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j)$, $\tilde{b} = 1, \ldots, \widetilde{B}$. We then take the mean of these $\widetilde{B}$ differences and we define

$$\widehat{\overline{\mathbf{ISE}}}_{\alpha,B,\widetilde{B},J,\mathcal{U}}(N) = \frac{1}{J} \sum_{j=1}^{J} \left(\frac{1}{|\mathcal{U}|} \sum_{\boldsymbol{u} \in \mathcal{U}} \left(\frac{1}{\widetilde{B}} \sum_{\tilde{b}=1}^{\widetilde{B}} \left| \bar{\boldsymbol{q}}^{N,n}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j) - \hat{\boldsymbol{q}}^{(B+\tilde{b})}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j) \right|^2 \right) \right).$$

As in its single-output counterpart, we do not stress in the sequel the dependence on J in these ISEs, nor the dependence on $\widetilde{B}$ and on $\mathcal{U}$, that we choose equal to 30 and 40 equispaced directions respectively throughout.

Evaluating $\widehat{\overline{\mathbf{ISE}}}_{\alpha,B}(N)$ thus requires generating $B + \widetilde{B}$ bootstrap samples of size n: B for the construction of $\bar{\boldsymbol{q}}^{N,n}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j)$, and $\widetilde{B}$ to obtain $\hat{\boldsymbol{q}}^{(B+\tilde{b})}_{\alpha \boldsymbol{u}}(\boldsymbol{x}_j)$, $\tilde{b} = 1, \ldots, \widetilde{B}$. These sample ISEs are to be minimized in N. Since not all values of N can be considered in practice, we rather consider

$$\hat{N}^{-}_{\alpha,B;\text{opt}} = \arg\min_{N \in \mathcal{N}} \widehat{\overline{\mathbf{ISE}}}_{\alpha,B}(N), \tag{6.2.1}$$

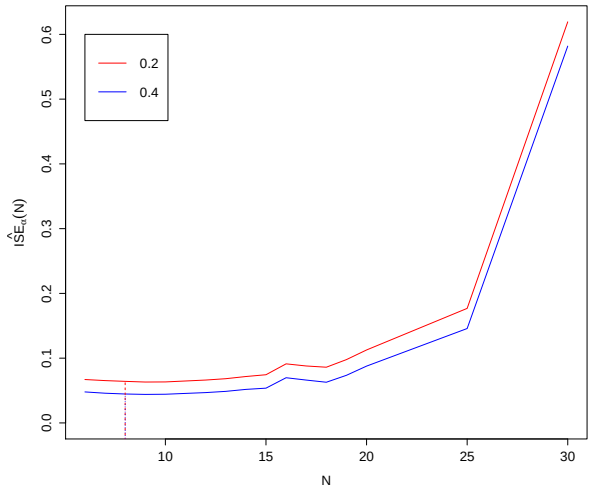


Figure 6.1 – Plot of the mappings $N \mapsto \widehat{\mathbf{ISE}}^{-}_{\alpha,B}(N)$ with $B = 50$, $\widetilde{B} = 30$ and $|\mathcal{U}| = 40$, averaged over 100 mutually independent replications of Model $(\mathcal{M}1)$ with sample size $n = 999$.

where the cardinality of $\mathcal{N}(\subset \mathbb{N}_0)$ is finite (and may be chosen as a function of n).

To illustrate these definitions, we simulated random samples of size $n = 999$ according to the model

$$(\mathcal{M}1)\ \ (Y_1, Y_2) = (X, X^2) + \left(1 + \tfrac{3}{2}\left(\sin\left(\tfrac{\pi}{2}X\right)\right)^2\right)\boldsymbol{\varepsilon},$$

where $X \sim U[-2, 2]$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)^2$ are independent. We evaluated the ISEs above for $x = -1.89, -1.83, -1.77, \ldots, 1.89$. Figure 6.1 plots the resulting graphs of $N \mapsto \widehat{\mathbf{ISE}}^{-}_{\alpha,B}(N)$, with $B = 50$ (more precisely, the average of the corresponding plots, over 100 mutually independent replications, are plotted there). It shows that ISE curves are indeed essentially convex in N and allows to select N equal to 8 for both values of α.

Let us motivate the choice $n = 999$ that could seem unnatural. We actually consider here models and sample sizes already investigated in HPŠ15 by our competitors (see their definition in the following section). They chose $n = 999$ (and $n = 499$) because there is a problem in the definition of multiple-output quantile regression if the order α is a multiple of $1/n$.

6.3 Comparison with some competitors

Chapter 5 and Section 6.2 described a new multiple-output conditional quantile estimator that 1) has interesting convergence results and 2) provides an efficient data-driven selection method for the tuning parameter. A natural step is then to wonder if our estimator is competitive with alternative estimators. We first define the competitors considered in the next section.

6.3.1 The competitors considered

As explained in Section 4.2.2, defining an estimator of multiple-output conditional quantiles first requires to choose a definition of multivariate quantiles. Since there is no universally accepted definition of them, the literature in multiple-output quantile regression is much less huge than in

its single-output counterpart, and our main competitors are the local constant and local *bilinear* estimators of HPŠ15, that extend the local constant and local linear estimators of Yu and Jones (1998) defined in Section 3.3.

Let K be a nonnegative kernel function over $\mathbb{R}^d$ and h be a bandwidth. Let

$$\mathcal{X}_{i\boldsymbol{u}}^c = \begin{pmatrix} 1 \\ \boldsymbol{Y}_{i\boldsymbol{u}}^{\perp} \end{pmatrix},$$

and

$$\mathcal{X}_{i\boldsymbol{u}}^l = \begin{pmatrix} 1 \\ \boldsymbol{Y}_{i\boldsymbol{u}}^{\perp} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ \boldsymbol{X}_i - \boldsymbol{x} \end{pmatrix},$$

where $\otimes$ denotes the tensor product and where the superscripts c and l stand respectively for the local constant and the local bilinear cases, and where $\boldsymbol{Y}_{i\boldsymbol{u}}^{\perp}$ is the projection of $\boldsymbol{Y}_i$ onto $\boldsymbol{\Gamma}_{\boldsymbol{u}}$. We denote analogously $\boldsymbol{Y}_{i\boldsymbol{u}}$ the projection of $\boldsymbol{Y}_i$ onto $\boldsymbol{u}$. The parameter of interest for the local constant case is $\boldsymbol{\theta}_{\boldsymbol{\alpha}}^c(\boldsymbol{x}) := \widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^c(\boldsymbol{x}) = \big(\widehat{a}_{\boldsymbol{\alpha}}^c(\boldsymbol{x}), (\widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^c(\boldsymbol{x}))'\big)'$ while, for the local bilinear case, it is

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x}) = \operatorname{vec} \begin{pmatrix} \widehat{a}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x}) & \widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x}) \\ \widehat{\dot{\boldsymbol{a}}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x}) & \widehat{\dot{\boldsymbol{c}}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x}) \end{pmatrix},$$

where $\widehat{\dot{\boldsymbol{a}}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$ stands for the gradient of $\boldsymbol{x} \mapsto \widehat{a}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$ and $\widehat{\dot{\boldsymbol{c}}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$ for the Jacobian of $\boldsymbol{x} \mapsto \widehat{\boldsymbol{c}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$. The local constant and local bilinear estimators are then defined as the minimizer $\theta_{\boldsymbol{\alpha}}^r(\boldsymbol{x})$ of

$$\sum_{i=1}^{n} K\Big(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h}\Big) \rho_{\alpha}\big(\boldsymbol{Y}_{i\boldsymbol{u}} - (\boldsymbol{\theta}_{\boldsymbol{\alpha}}^r(\boldsymbol{x}))' \mathcal{X}_{i\boldsymbol{u}}^r\big), \qquad r = c, \ell. \tag{6.3.1}$$

As explained in HPŠ15, the local bilinear approach is more informative than the local constant one and should be more reliable at boundary points, as in the single-output case. However, the price to pay is an increase of the covariate space dimension ($\mathcal{X}_{i\boldsymbol{u}}^c$ is of dimension m while $\mathcal{X}_{i\boldsymbol{u}}^{\ell}$ is of dimension md). More details on these approaches can be found in HPŠ15.

In the sequel, we consider $d = 1$. Hence, K will be the Gaussian kernel and we choose, as in some applications in HPŠ15,

$$h = 3\sigma_x n^{-1/5},$$

where σ_x stands for the empirical standard deviation of the regressor X.

To the best of our knowledge, no R package allows to compute $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^c(\boldsymbol{x})$ and $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$; we therefore wrote our own implementation to conduct the simulations below.

6.3.2 Comparison of estimated quantile contours

In this section, we compare our proposed quantization-based estimator with their competitors described above. We saw in the first part of this thesis that the bootstrap version should be favored to its original version in the single-output setting, and this observation is also valid in this multiple-output setting. Therefore, we do not investigate $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}$ in the sequel and restrict our comparison to $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}$, with the corresponding data-driven value of N proposed in Section 6.2. We start the comparison by investigating estimated regression quantile/depth contours computed from $n = 999$ independent observations generated according to the models

$(\mathcal{M}1)$ $(Y_1, Y_2) = (X, X^2) + \left(1 + \frac{3}{2}\left(\sin\left(\frac{\pi}{2}X\right)\right)^2\right)\boldsymbol{\varepsilon}_1$,
$(\mathcal{M}2)$ $(Y_1, Y_2) = (X, X^2) + \boldsymbol{\varepsilon}_2$,
$(\mathcal{M}3)$ $(Y_1, Y_2) = (X, X^2) + (1 + X^2)\boldsymbol{\varepsilon}_2$,

where $X \sim U[-2, 2]$, $\boldsymbol{\varepsilon}_1 \sim \mathcal{N}(0, 1)^2$, $\boldsymbol{\varepsilon}_2 \sim \mathcal{N}(0, 1/4)^2$ and X is independent of $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$. We actually estimate $\boldsymbol{q}_{\boldsymbol{\alpha}}(x) = \boldsymbol{q}_{\alpha\boldsymbol{u}}(x)$ for $x = -1.89, -1.83, -1.77, \ldots, 1.89$, $\alpha = 0.2$ and 0.4, and 360 equispaced directions $\boldsymbol{u} \in \mathcal{S}^1$. These models were actually already investigated in HPŠ15 where they compared the contours obtained with $\widehat{\boldsymbol{q}}^c_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$ with the ones from HPŠ10. Since the latter generally provide poorer results (as mentioned in Section 4.2.2, see also HPŠ15), we do not consider them in the sequel.

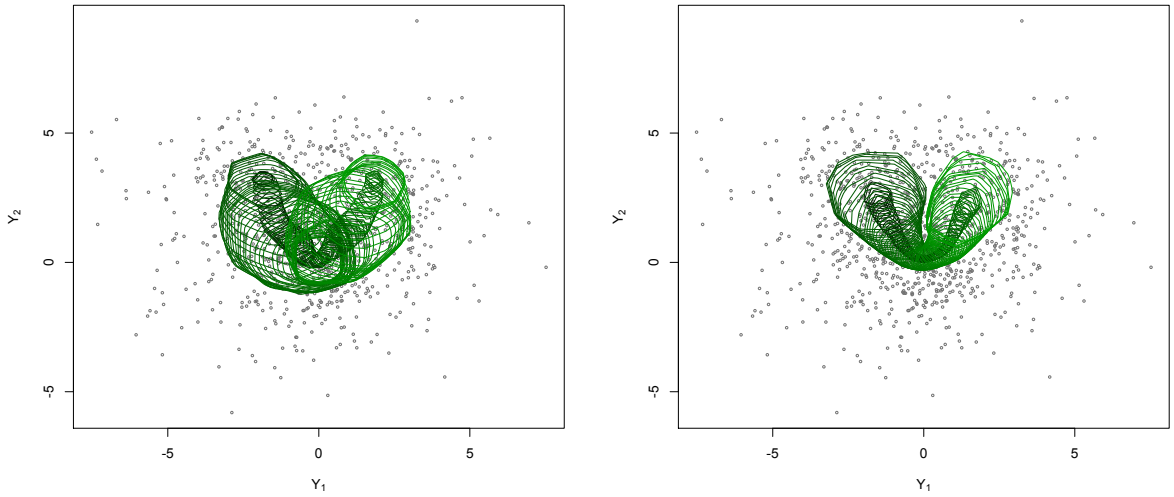
In Figure 6.2, we are plotting the regression quantile/depth contours for Model $(\mathcal{M}1)$, estimated with $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ (top left), with $\widehat{\boldsymbol{q}}^c_{\boldsymbol{\alpha}}$ (top right) and with $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$ (bottom left). The corresponding population halfspace depth regions are plotted in the bottom right graph. More precisely, these plots show the intersections, with hyperplanes orthogonal to the x-axis at fixed x-values $-1.89, -1.83, -1.77, \ldots, 1.89$, of regression quantile regions. Figures 6.3 and 6.4 are analogous for Models $(\mathcal{M}2)$ and $(\mathcal{M}3)$.

In each model, we did not select h following the data-driven procedure mentioned in Section 6.3.1, but chose it equal to 0.37, as proposed in HPŠ15, Figures 1–3. Concerning the choice of N, our method selected N equal to 10 in each model, among the grids $\mathcal{N}_1 = \{5, 10, 15, 20\}$ for Models $(\mathcal{M}1)$ and $(\mathcal{M}2)$, and $\mathcal{N}_2 = \{10, 15, 20\}$ for Models $(\mathcal{M}3)$. We notice that the selecting value is on the boundary of $\mathcal{N}_2$ for this last model. Indeed, it happens that the criterion selects a too small value in some cases, which induces not sufficiently smooth contours. Since it is reasonable given the sample size to select N greater than 10, we adapt the grid of values for N and we do not worry about the fact that the optimal N selected is on the boundary.

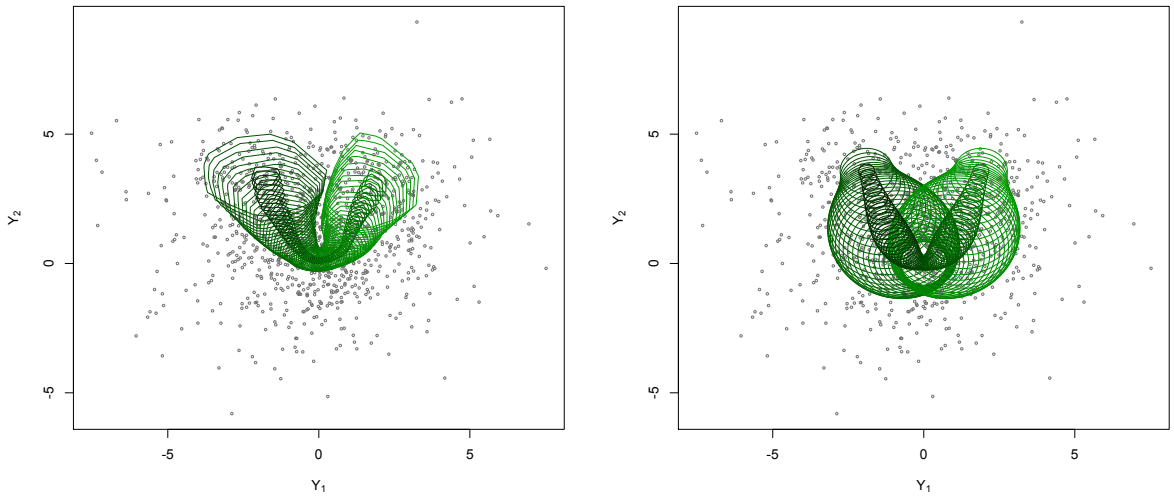
We observe that $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$ provide really nice contours close to the theoretical ones. They catch the heteroscedasticity in Models $(\mathcal{M}1)$ and $(\mathcal{M}3)$. Compared with the local constant ones, they do better, particularly in the boundary of the support of X, i.e. the most extreme contours.

It is also interesting to consider another sample size and to evaluate if the previous conclusions remain valid in this context. Figure 6.5 is then analogous to Figure 6.2 for $n = 499$ instead of $n = 999$. Here h is again chosen equal to 0.37 while N is equal to 6, selected by our criterion among the grid $\mathcal{N}_3 = \{6, 8, 10, 15, 20\}$ (as above, it is reasonable to select N greater than a certain value given the size $n = 499$). Here again, our estimator $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ provides really satisfactory contours.

We then conclude that $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$ are the most satisfactory from a graphical point of view. We also observe a slight advantage for $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$, particularly on the boundaries. This is not surprising since the bilinear estimator allows to improve estimation on the boundaries with respect to the local constant one. However, it has to be pointed out that N was selected through a data-driven procedure for $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ while it is not the case for h in $\widehat{\boldsymbol{q}}^\ell_{\boldsymbol{\alpha}}$. Therefore, we will compare the ISEs in the next section using the data-driven procedure for h for more equity.

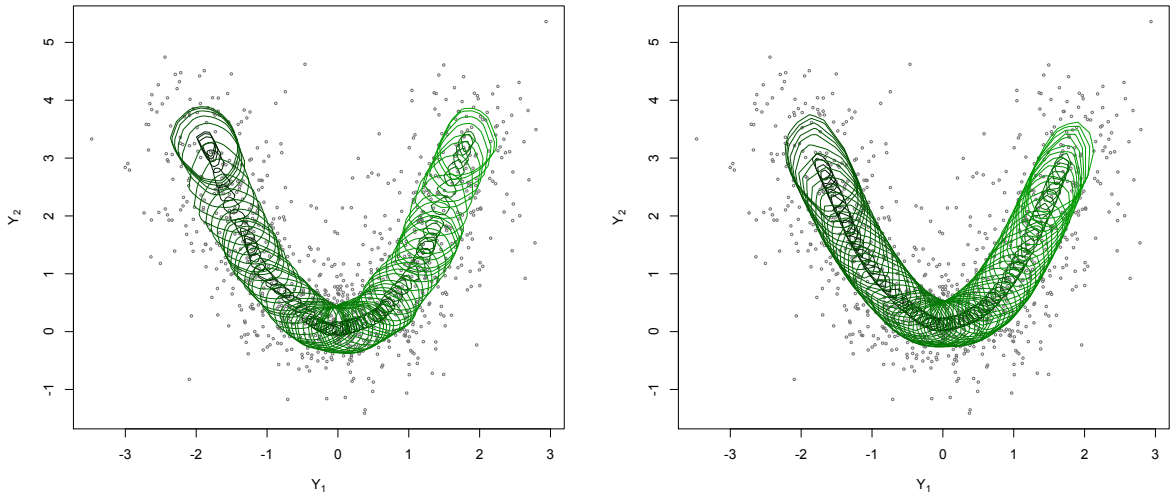


(a) (b)



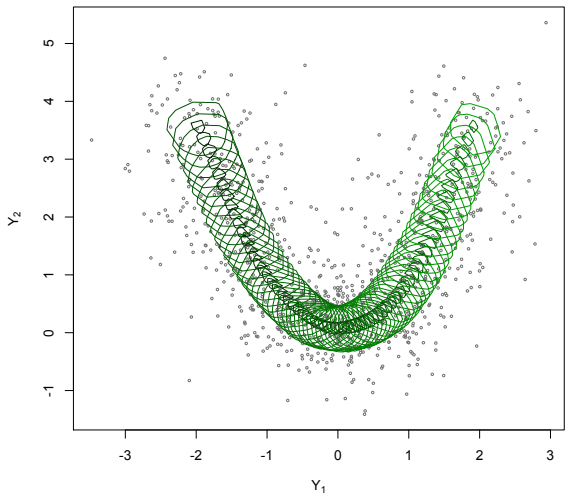
(c) (d)

Figure 6.2 – For $n = 999$ points following Model ($\mathcal{M}1$), the intersections, with hyperplanes orthogonal to the x-axis at fixed x-values $-1.89, -1.83, -1.77, \ldots, 1.89$, of regression quantile regions estimated with (a) $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}$, (b) $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{c}$ and (c) $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{\ell}$. For the sake of comparison, the corresponding population (conditional) halfspace depth regions are provided in (d).

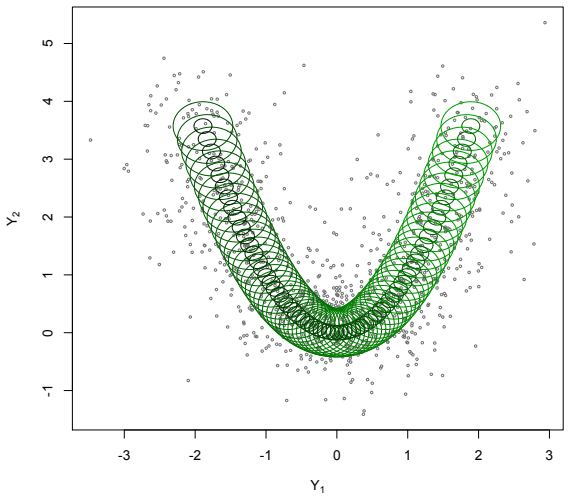


(a)

(b)

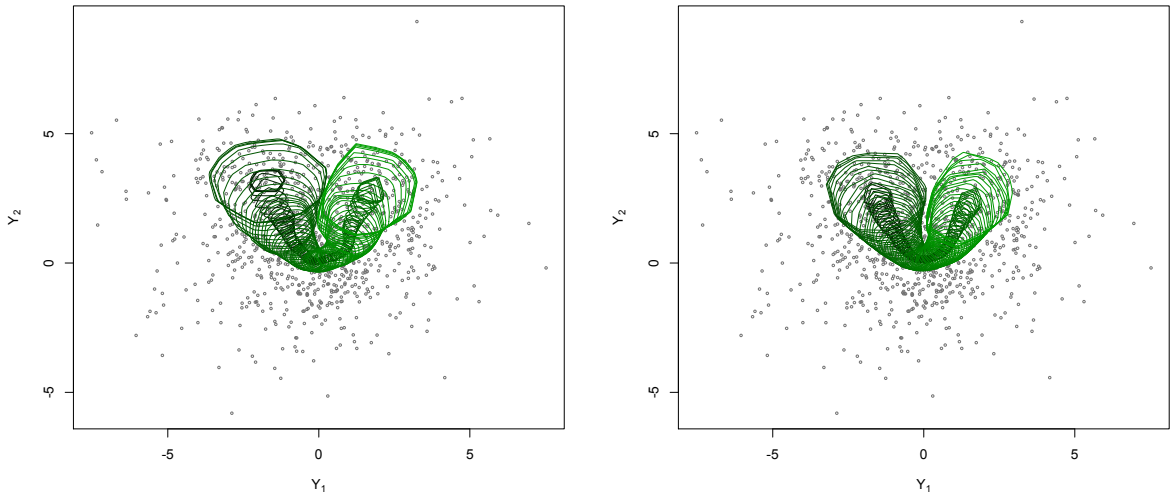


(c)



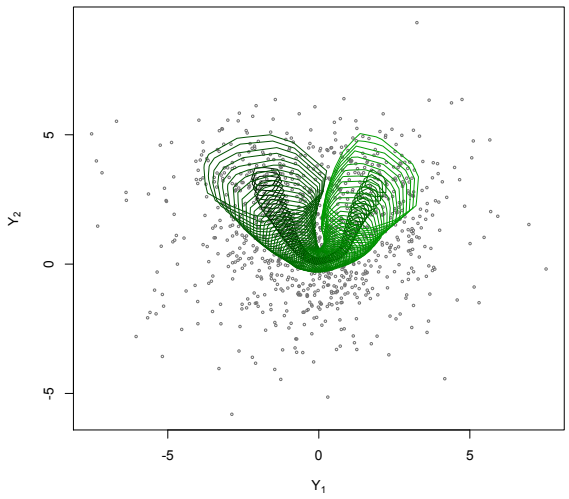
(d)

Figure 6.3 – For $n = 999$ points following Model ($\mathcal{M}2$), the same as Figure 6.2.

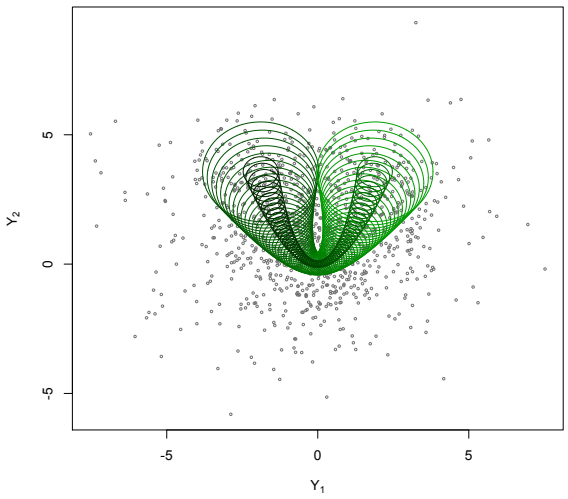


(a)

(b)

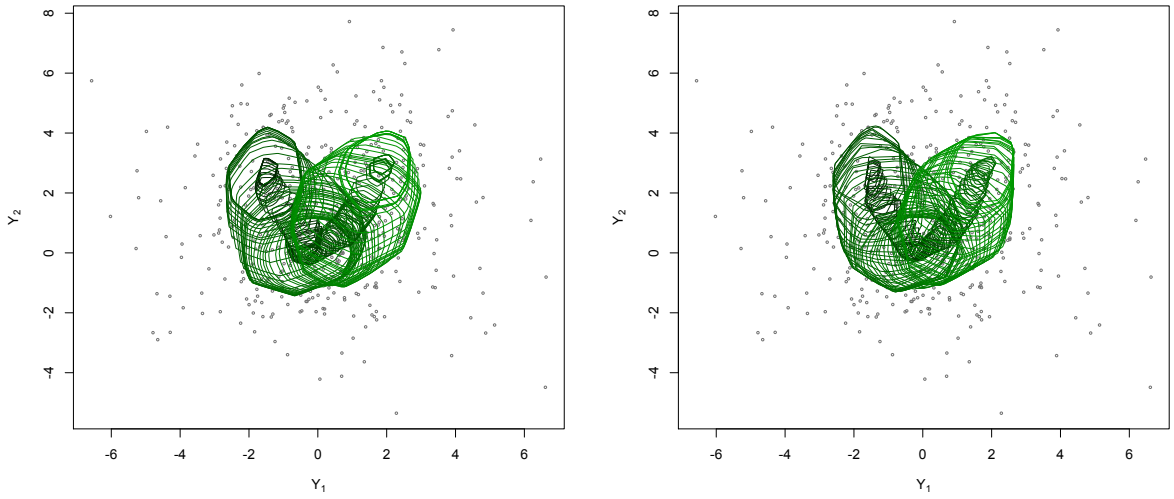


(c)



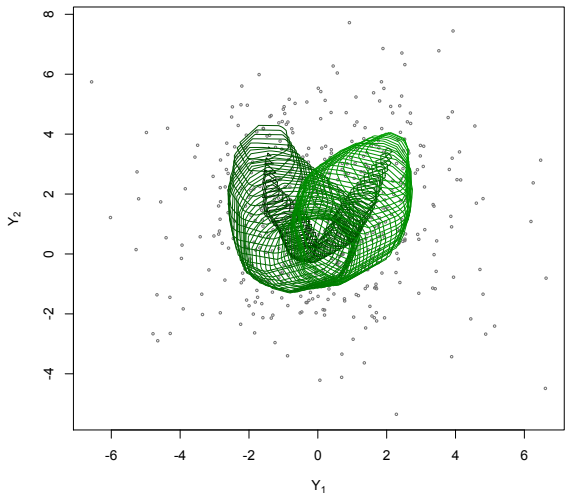
(d)

Figure 6.4 – For $n = 999$ points following Model ($\mathcal{M}3$), the same as Figure 6.2.

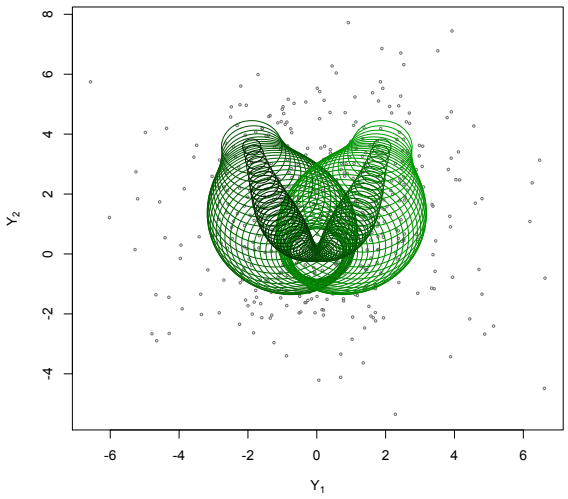


(a)

(b)



(c)



(d)

Figure 6.5 – For $n = 499$ points following Model ($\mathcal{M}1$), the same as Figure 6.2.

6.3.3 Comparison of the ISEs

Obtaining well-behaved regression quantile contours is of course desirable, but this should not be achieved at the expense of efficiency. That is why we now compare the various estimators in terms of ISEs. To do so, we generated 500 independent samples from Models $(\mathcal{M}1) - (\mathcal{M}3)$ with sample sizes $n = 999$, and from Model $(\mathcal{M}1)$ with $n = 499$. In each case, we evaluated the ISEs corresponding to the quantization-based estimators and to their two competitors defined in Section 6.3.1.

As above mentioned, the quantile order here is composed of both $\alpha \in (0, 1)$ and a direction $\boldsymbol{u} \in \mathcal{S}^1$. The aim of this section is to compare efficiency of estimators of *one* conditional quantile $\boldsymbol{q}_{\boldsymbol{\alpha}}(x)$, hence for a couple $(\alpha, \boldsymbol{u})$ fixed. Investigating the efficiency of the contours appears complicated since it will be sensitive to the number of directions $\boldsymbol{u}$ considered. Therefore, we only consider one direction $\boldsymbol{u}$, and we choose $\boldsymbol{u} = (0, 1) \in \mathcal{S}^1$ without loss of generality. We still work with $\alpha = 0.2$ and 0.4, and $x_1 = -1.89, x_2 = -1.83, \ldots, x_J = 1.89$.

In each model, we selected h following the data-driven procedure mentioned in Section 6.3.1, and N is chosen by our selection method among $\mathcal{N}_2$ for $n = 999$ and among $\mathcal{N}_3$ for $n = 499$.

For each model, sample size and quantile order $\boldsymbol{\alpha}$ considered, it provides a series of 500 observed ISEs for each estimator. More precisely, we evaluate

$$\mathrm{ISE}^a_{\boldsymbol{\alpha}} = \sum_{j=1}^{J} \big(a_{\boldsymbol{\alpha}}(x_j) - \widehat{a}_{\boldsymbol{\alpha}}(x_j)\big)^2,$$

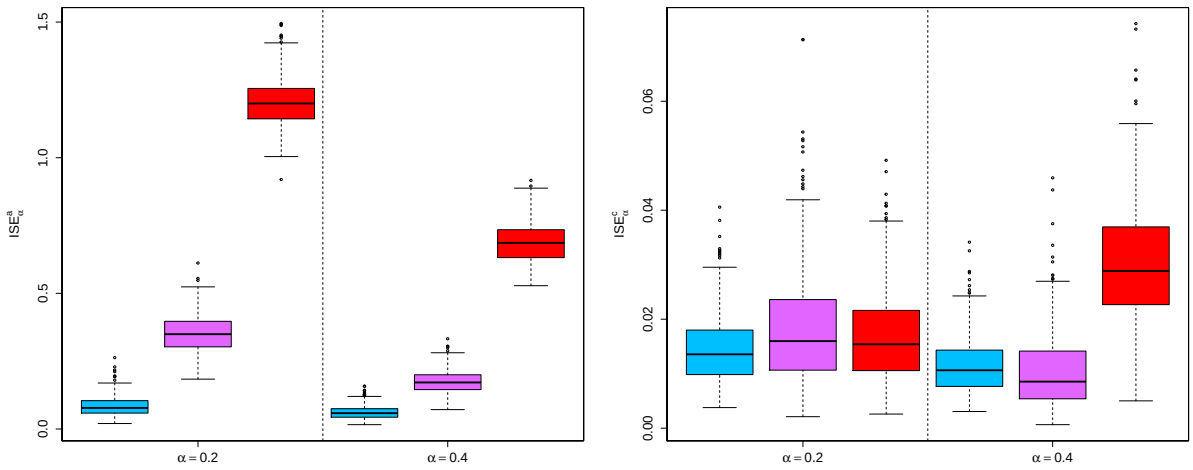
and

$$\mathrm{ISE}^c_{\boldsymbol{\alpha}} = \sum_{j=1}^{J} \big(c_{\boldsymbol{\alpha}}(x_j) - \widehat{c}_{\boldsymbol{\alpha}}(x_j)\big)^2,$$

where $\widehat{a}_{\boldsymbol{\alpha}}(x_j)$ stands for $\bar{a}^{N,n}_{\boldsymbol{\alpha}}(x_j)$, $\widehat{a}^{\ell}_{\boldsymbol{\alpha}}(x_j)$ or $\widehat{a}^{c}_{\boldsymbol{\alpha}}(x_j)$ and $\widehat{c}_{\boldsymbol{\alpha}}(x_j)$ for $\bar{c}^{N,n}_{\boldsymbol{\alpha}}(x_j)$, $\widehat{c}^{\ell}_{\boldsymbol{\alpha}}(x_j)$ or $\widehat{c}^{c}_{\boldsymbol{\alpha}}(x_j)$. Figures 6.6–6.8 draw the boxplots of those 500 ISE values for Models $(\mathcal{M}1) - (\mathcal{M}3)$, respectively. The left plots correspond to $\mathrm{ISE}^a_{\boldsymbol{\alpha}}$ (one for each α considered), while the right ones to $\mathrm{ISE}^c_{\boldsymbol{\alpha}}$. The boxplots of $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{q}}^{\ell}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{q}}^{c}_{\boldsymbol{\alpha}}$ are represented in blue, purple and red respectively.

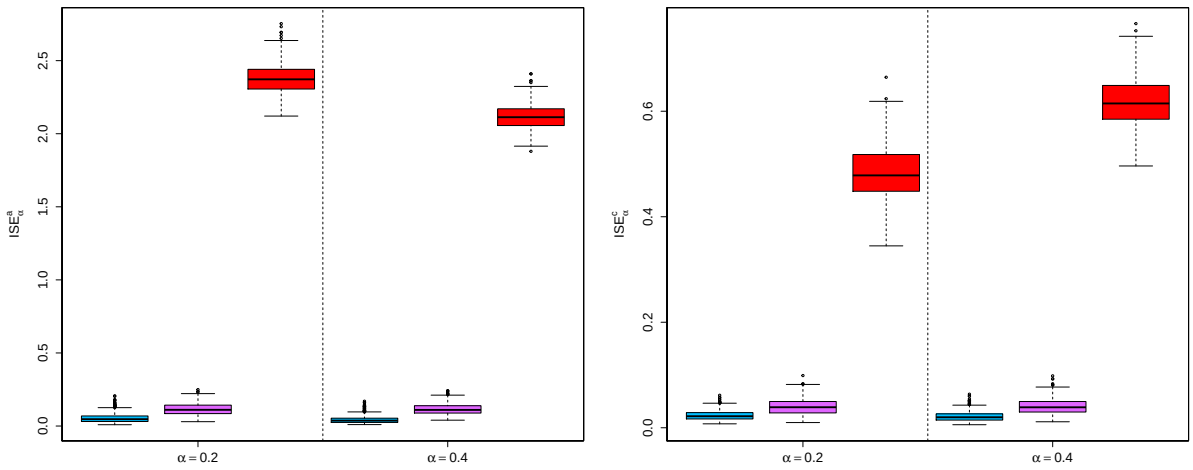
Results reveal that $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{q}}^{\ell}_{\boldsymbol{\alpha}}$ generally do clearly better than $\widehat{\boldsymbol{q}}^{c}_{\boldsymbol{\alpha}}$, particularly for the component $a_{\boldsymbol{\alpha}}(x)$ of $\boldsymbol{q}_{\boldsymbol{\alpha}}(x)$. Moreover, for most models, sample sizes and components of $\boldsymbol{q}_{\boldsymbol{\alpha}}(x)$, $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ also outperforms $\widehat{\boldsymbol{q}}^{\ell}_{\boldsymbol{\alpha}}$. The cases where this trend is inverted mostly concern the boxplots for the component $\boldsymbol{c}_{\boldsymbol{\alpha}}(x)$ and we notice there that the errors are really small (less than 0.05 in median). Hence, our estimator is still really efficient.

As in the first part, the good performances of $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}(x)$ with respect to the ones of those kernel estimators result from the principle of quantization: recall that the radius of the Voronoi cell is adaptive with x. It can then be seen as a data-adaptive bandwidth parameter while the bandwidth h for the kernel estimators is chosen in the same way for each x.



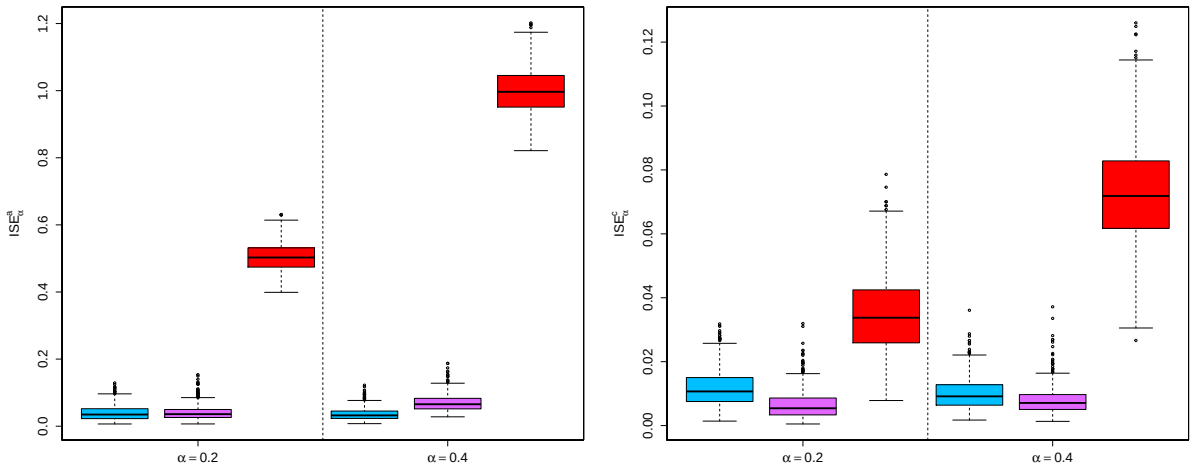
(a) (b)

Figure 6.6 – Boxplots, for $\alpha = 0.2$ and 0.4 and $\boldsymbol{u} = (0, 1)$, of $\mathrm{ISE}^a_{\boldsymbol{\alpha}}$ (left) and of $\mathrm{ISE}^c_{\boldsymbol{\alpha}}$ (right) for various conditional quantile estimators obtained for 500 independent random samples from Model ($\mathcal{M}1$), with size $n = 999$. The estimators considered are the quantization-based estimator $\bar{\boldsymbol{q}}^{N,n}_{\boldsymbol{\alpha}}$ (in blue), the local bilinear estimator $\widehat{\boldsymbol{q}}^{\ell}_{\boldsymbol{\alpha}}$ (in purple) and the local constant estimator $\widehat{\boldsymbol{q}}^{c}_{\boldsymbol{\alpha}}$ (in red).



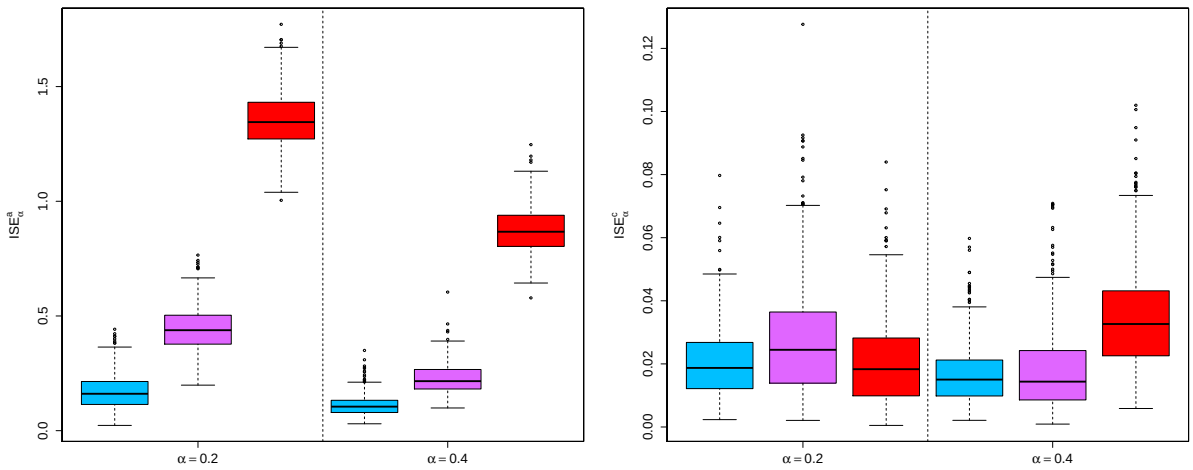
(a) (b)

Figure 6.7 – The same boxplots as in Figure 6.6, but for Model ($\mathcal{M}2$).



(a) (b)

Figure 6.8 – The same boxplots as in Figure 6.6, but for Model ($\mathcal{M}3$).



(a) (b)

Figure 6.9 – The same boxplots as in Figure 6.6, but for $n = 499$.

6.4 Real data applications

In this section, we illustrate the behavior of $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ on a real data set. We investigate the "body girth measurement" data set from Heinz et al. (2003), investigated also in HPŠ10 and HPŠ15. This data set consists of nine skeletal and twelve body girth dimensions, and of weight, height and age, in a group of 247 young men and 260 young women. As in HPŠ15, we get rid of the men observations. The response vector is composed of the calf maximum girth (Y_1) and the thigh maximum girth (Y_2) while we deal with a single random covariate (X = weight, age, BMI[1] or height). We consider for α the values 0.01, 0.03, 0.10, 0.25, 0.40 and for x different p-quantiles of the regressor X, i.e. $p = 0.1$ (black), 0.3 (blue), 0.5 (green), 0.7 (cyan) and 0.9 (yellow). We now describe how we obtained quantization-based quantile contours in this context.

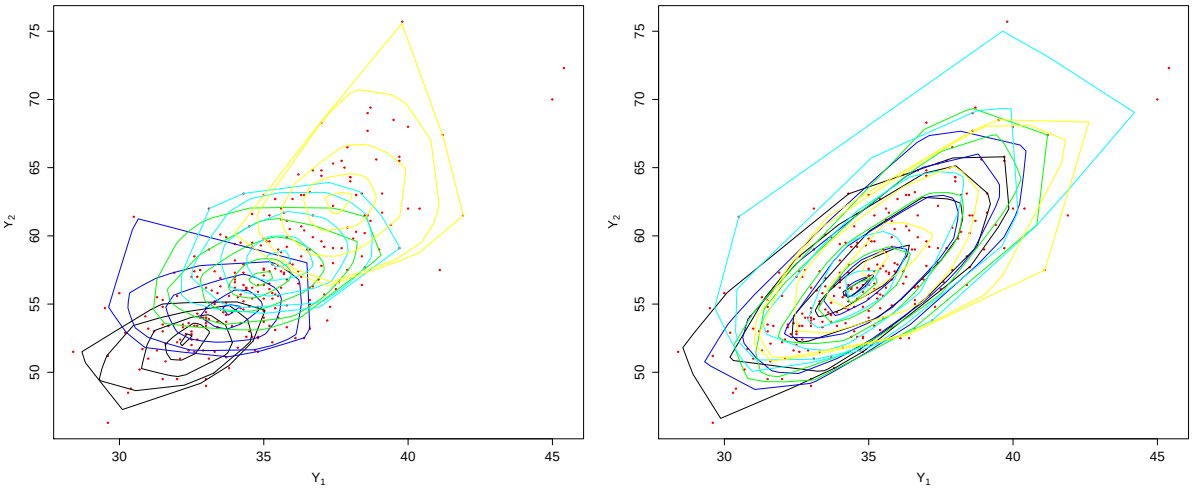
The first step consists in choosing, for each regression setup, the optimal number N of quantizers. We then adopt the method of Section 6.2, based on the minimization of $\widehat{\mathbf{ISE}}_{\boldsymbol{\alpha},B}(N)$ with the choice $B = 100$, $\widetilde{B} = 30$ and $|\mathcal{U}| = 360$. As above mentioned, we then take $J = 5$. We first evaluated $\widehat{\mathbf{ISE}}_{\boldsymbol{\alpha},B}(N)$ for all $N \in \mathcal{N}_1 = \{5, 10, \ldots, 30\}$ and the criterion selected $N = 5$ as optimal value. The tested values are then too large and it is felt that the optimal value is among the smallest ones. Moreover, as explained in Section 6.3.2, it is reasonable to select N greater than a certain value, otherwise it would induce not sufficiently smooth contours. Therefore, we did the same exercise for $N \in \mathcal{N}_2 = \{6, 7, \ldots, 10\}$, which led to $N_{\text{opt}} = 6$. Notice that we did not select N separately for each α. Indeed, we tested it as in the single-output case but we observed crossings of the contours, which makes no sense. We then protected against this by choosing the same N for each α.

In a second step, we obtain the different estimations $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(x)$ for each x and $\boldsymbol{\alpha}$ considered, based on the selection of N above. We plot in Figure 6.10 the resulting estimated quantile contours. Each graph corresponds to a different regressor (X = weight in (a), X = age in (b), X = BMI in (c) and X = height in (d)). We observe that the resulting contours are quite different with respect to the regressor considered. When X = weight (Figure 6.10a), we notice a positive trend in the location and an increase in dispersion. A rotation of the principal direction is also visible, going from horizontal to vertical as x increases. Similar remarks can be made for X = BMI in Figure 6.10c. It is totally different when X = age (Figure 6.10b) where the plot reveals no location trend and a constance for the principal direction and the dispersion. Finally, for X = height (Figure 6.10d), the conclusion are similar as the ones for X = age, excepted for the dispersion that seems to increase with x.

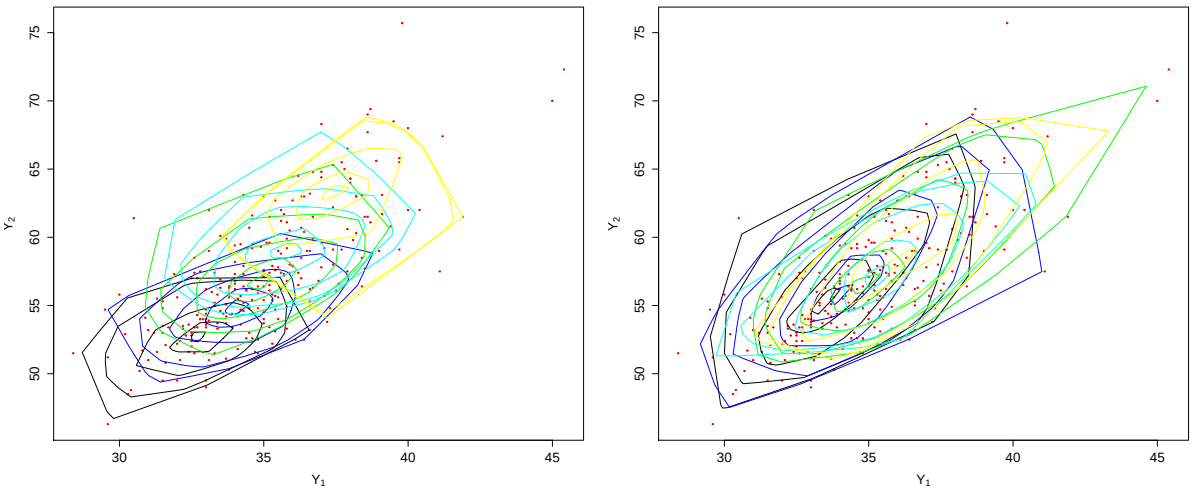
These different observations are in line with the ones of HPŠ15 for the local bilinear estimator (but not for the local constant one). In view of the good performances of $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}$ and $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{\ell}$ (and their dominance over $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{c}$) of the previous section, this encourages us to trust these results. The corresponding estimated contours with $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{\ell}$ can be found in HPŠ15.

We then conclude that our quantization-based estimator $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}$ provides really satisfying results since it reveals different features of the conditional distribution of $\boldsymbol{Y}$ given $X = x$.

[1] BMI=Body Mass Index, defined as weight/(height)^2



(a) X = weight (b) X = age



(c) X = BMI (d) X = height

Figure 6.10 – For $n = 260$ points from the body girth measurements data set (women subsample, see Heinz et al. (2003)), the intersections, with hyperplanes orthogonal to the x-axis at the empirical p-quantiles of the regressors, for $p = 0.10$ (black), 0.30 (blue), 0.50 (green), 0.70 (cyan) and 0.90 (yellow), of regression quantile regions estimated with $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(x)$,with $\alpha = 0.01$, 0.03, 0.10, 0.25 and 0.40. Throughout, the bivariate response (Y_1, Y_2) corresponds to the calf maximum girth (Y_1) and the thigh maximum girth (Y_2), while a single random regressor is used: (a) weight, (b) age, (c) BMI or (d) height.

6.5 Final comments

In this chapter, we investigated the empirical performances of the quantization-based estimator of multiple-output conditional quantiles that was introduced in Chapter 5. We first extended in Section 6.2 the data-driven selection method for the tuning parameter N of Section 3.2. Secondly, we realized extensive simulations and concluded that our estimator often dominates its competitors in terms of empirical integrated square errors. From a graphical point of view, we plotted the resulting estimated quantile contours that were really satisfying (e.g. we detected heteroscedasticity and were close to the theoretical ones). We then treated a real data example that was also investigated by the competitors considered, which allows to complete the comparison study based on simulations. We showed that the proposed methodology provides good results and allows to deduce some information about the variables considered, in concordance with the conclusions made by the competitors.

The major motivation for this last part was to determine if the dominance of our quantization-based estimator over the kernel estimators in the single-output framework was also observed in the multiple-output framework. We indeed obtained very satisfactory results and conclude that $\bar{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{N,n}(\boldsymbol{x})$ is a serious competitor for $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{c}(\boldsymbol{x})$ and $\widehat{\boldsymbol{q}}_{\boldsymbol{\alpha}}^{\ell}(\boldsymbol{x})$ in this context.

7

Conclusion and perspectives

Quantile regression allows to represent the impact of some covariate X on a response variable Y and is a robust alternative to standard (mean) regression. Hence, having an efficient estimator of them is highly desirable in order to quantify at best the conditional distribution of Y. The main goal of this thesis was to propose a new nonparametric estimator of conditional quantiles using L_p-norm optimal quantization, a discretizing method quite rarely used in statistics. The different preliminary notions were properly defined in Chapter 1, where this issue was also widely motivated with several examples.

The construction of this estimator is based on two steps. First, we defined an approximation of conditional quantiles by replacing X by its quantized version into the definition of conditional quantiles. Secondly, we took an empirical version of this approximation to get an estimation. Convergence of both this approximation and this estimation was studied and several theorems were obtained, with a rate of convergence for one of them. This was the aim of Chapter 2. A bootstrap version of our estimator was then defined, allowing to get smoother estimated quantile curves.

Since the empirical performances are also of major importance, Chapter 3 was devoted to numerical results. Starting with developing a data-driven method for selecting the tuning parameter of our procedure (i.e., the size of the quantization grid), we then compared our estimator with some well-known competitors, as spline, kernel and k-nearest neighbor estimators. This comparison study revealed very good performances (and sometimes dominance) of our bootstrap-based estimator. We treated some real data examples for which our methodology also provided very satisfactory estimated curves. Therefore, quantization-based estimation should be of interest for practitioners, which motivated the creation of an R package, named **QuantifQuantile**, whose functions make the application of our procedure quite easy. This package and its different functions were then introduced and illustrated in this chapter.

As above mentioned, quantile regression aims to quantify the relationship between some response *variable* Y and some vector of covariates X. A natural question was then to consider a multivariate response and this was the goal of the second part of this thesis. Since the notion of quantile cannot be directly generalized in higher dimension, the literature is much less huge than for a univariate response. Chapter 4 compared the different existing generalizations of the notion of quantiles in a multivariate setting and led to select the one from HPŠ10 that stands out by

satisfying all the properties expected from a notion of quantiles and by providing an interesting link to the concept of depth.

This definition of multivariate quantiles directly allowed to define multiple-output conditional quantile. Following the same construction as in Chapter 2, we defined in Chapter 5 successively an approximation and an estimation of these multiple-output conditional quantiles, and studied their convergence. In view of the improvement observed in the first part with bootstrap, a bootstrap version of this estimator was also introduced.

In the same spirit as in Chapter 3 for the single-output context, Chapter 6 aimed to 1) extend the data-driven procedure of the selection of the quantization grid size N for this new estimator and 2) compare it with the existing alternative estimators – mainly, the ones from HPŠ15. Also in this multiple-output context, the performances of our bootstrap-based estimator were really satisfactory since it generally dominates its competitors in terms of some empirical integrated square error. A real-data example confirmed these good results.

The main conclusion of this thesis is that the proposed quantization-based estimators are serious competitors for existing conditional quantile estimators, both in single and multiple-output frameworks. This should then be of high practical interest for practitioners that aim to extract from data sets several information on the dependence between variables and to obtain reference curves (or surfaces, contours).

Let us conclude this work with some perspectives. We mentioned in Section 2.3 the delicate problem of double asymptotic result for $\widehat{q}_\alpha^{N,n}(x)$. Recall that we proved 1) the convergence of $\widetilde{q}_\alpha^{N}(x)$ toward $q_\alpha(x)$ as N tends to infinity and 2) the convergence of $\widehat{q}_\alpha^{N,n}(x)$ toward $\widetilde{q}_\alpha^{N}(x)$ as n tends to infinity and for N fixed. Obtaining a result of convergence for $\widehat{q}_\alpha^{N,n}(x)$ as both N and n go to infinity was beyond the scope of this thesis. Indeed, this would require convergence results for the stochastic gradient algorithm for both N and n going to infinity. To the best of our knowledge, such results were not proved yet and would be the key to obtain such a convergence result for $\widehat{q}_\alpha^{N,n}(x)$.

Moreover, this work focused on numerical data that are fully observed. The good results encourage us to consider other types of data. For instance, it may happen that some variables are not fully observed due to the termination of the study or an early withdrawal of the subject. It could be the case, e.g., in medical studies if we are interested in the survival time. Hence, the data are only partially known (above a certain value in this example) and are called *censored*. Censoring can also occur if a measure instrument is not sufficiently accurate: if we do not detect some substance, it does not mean that the substance is absent but means that the quantity is too low to be detected. Quantile regression in the context of censored data is also of high interest since such data appear in many fields (medicine, economics or industrial context), see for instance Gannoun et al. (2002). However, we cannot apply directly our method to such data. An interesting perspective of work would then be to investigate how to adapt our estimator to such data sets. Likewise, remember that we throughout assumed that the data are independent and identically distributed. It would be interesting to extend this work for dependent data (coming from time series for example, see Cai (2002)).

Bibliography

Abdous, B. and R. Theodorescu (1992). Note on the spatial quantile of a random vector. *Statist. Probab. Lett. 13*(4), 333–336.

Adler, D., D. Murdoch, and others (2015). *rgl: 3D Visualization Using OpenGL*. R package version 0.95.1337.

Azaïs, R., A. Gégout-Petit, and J. Saracco (2012). Optimal quantization applied to sliced inverse regression. *J. Statist. Plann. Inference 142*(2), 481–492.

Bally, V., G. Pagès, and J. Printems (2005). A quantization tree method for pricing and hedging multidimensional American options. *Math. Finance 15*(1), 119–168.

Bhattacharya, P. K. and A. K. Gangopadhyay (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist. 18*(3), 1400–1415.

Bouton, C. and G. Pagès (1997). About the multidimensional competitive learning vector quantization algorithm with constant gain. *Ann. Appl. Probab. 7*(3), 679–710.

Breckling, J. and R. Chambers (1988). M-quantiles. *Biometrika 75*(4), 761–771.

Breckling, J., P. Kokic, and O. Lübke (2001). A note on multivariate M-quantiles. *Statist. Probab. Lett. 55*(1), 39–44.

Briane, M. and G. Pagès (2012). *Analyse - Théorie de l'intégration: Convolution et transformée de Fourier*. Vuibert.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory 18*(1), 169–192.

Cattaneo, M. D. and M. H. Farrell (2013). Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. *J. Econometrics 174*(2), 127–143.

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Ann. Inst. Statist. Math. 53*(2), 380–403.

Chakraborty, B. (2003). On multivariate quantile regression. *J. Statist. Plann. Inference 110*(1-2), 109–132.

Charlier, I., D. Paindaveine, and J. Saracco (2015a). Conditional quantile estimation through optimal quantization. *J. Statist. Plann. Inference 156*, 14–30.

Charlier, I., D. Paindaveine, and J. Saracco (2015b). Conditional quantile estimation based on

optimal quantization: from theory to practice. *Comput. Statist. Data Anal. 91*, 20–39.

Charlier, I., D. Paindaveine, and J. Saracco (2015c). QuantifQuantile: an R package for performing quantile regression through optimal quantization. *R journal, to appear*.

Charlier, I., D. Paindaveine, and J. Saracco (2015d). *QuantifQuantile: Estimation of Conditional Quantiles using Optimal Quantization*. R package version 2.2.

Chaudhuri, P. (1991a). Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Multivariate Anal. 39*(2), 246–269.

Chaudhuri, P. (1991b). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist. 19*(2), 760–777.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc. 91*(434), 862–872.

Cheng, Y. and J. G. De Gooijer (2007). On the uth geometric conditional quantile. *J. Statist. Plann. Inference 137*(6), 1914–1930.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossings. *Econometrica 78*, 1093–1125.

de Saporta, B., F. Dufour, and K. Gonzalez (2010). Numerical method for optimal stopping of piecewise deterministic Markov processes. *Ann. Appl. Probab. 20*(5), 1607–1637.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, Volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc. 87*(420), 998–1004.

Fan, J., T.-C. Hu, and Y. Truong (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics 21*(4), 433–446.

Fan, J., Q. Yao, and H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika 83*(1), 189–206.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall/CRC.

Fischer, A. (2010). Quantization and clustering with bregman divergences. *J. Multivariate Anal. 101*, 2207–2221.

Fischer, A. (2014). Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales. *J. Soc. Fr. Stat. 155*(2), 2–35.

Gannoun, A., S. Girard, C. Guinot, and J. Saracco (2002). Reference curves based on non-parametric quantile regression. *Statistics in Medicine 21*(4), 3119–3135.

Gannoun, A., S. Girard, C. Guinot, and J. Saracco (2004). Sliced inverse regression in reference curves estimation. *Comput. Statist. Data Anal. 46*(1), 103–122.

Gannoun, A., J. Saracco, A. Yuan, and G. E. Bonney (2005). Non-parametric quantile regression with censored data. *Scand. J. Statist. 32*(4), 527–550.

Graf, S. and H. Luschgy (2000). *Foundations of quantization for probability distributions*, Volume

1730 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York.

Hall, P., R. C. L. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc. 94*(445), 154–163.

Hallin, M., Z. Lu, D. Paindaveine, and M. Šiman (2015). Local constant and local bilinear multiple-output quantile regression. *Bernoulli 21*(3), 1435–1466.

Hallin, M., Z. Lu, and K. Yu (2009). Local linear spatial quantile regression. *Bernoulli 15*(3), 659–686.

Hallin, M., D. Paindaveine, and M. Šiman (2010a). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Ann. Statist. 38*(2), 635–669.

Hallin, M., D. Paindaveine, and M. Šiman (2010b). Rejoinder. *Ann. Statist. 38*(2), 694–703.

He, X. (1997). Quantile curves without crossing. *The American Statistician 51*(2), 186–192.

He, X., P. Ng, and S. Portnoy (1998). Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*(3), 537–550.

Heagerty, P. and M. Pepe (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *J. R. Stat. Soc. Ser. C Appl. Stat. 48*(4), 533–551.

Heinz, G., L. J. Peterson, R. W. Johnson, and C. J. Kerk (2003). Exploring relationships in body dimensions. *Journal of Statistics Education 11*(2).

Jones, M. C. and P. Hall (1990). Mean squared error properties of kernel estimates of regression quantiles. *Statist. Probab. Lett. 10*(4), 283–289.

Koenker, R. (2000). Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *J. Econometrics 95*(2), 347–374. Principles of econometrics (Madison, WI, 1998).

Koenker, R. (2005). *Quantile regression*, Volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Koenker, R. (2015). *quantreg: Quantile Regression*. R package version 5.19.

Koenker, R. and G. Bassett, Jr. (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and K. Hallock (2001). Quantiles regression. *Journal of Economic Perspective 15*(4), 143–156.

Koenker, R. and I. Mizera (2004). Penalized triograms: total variation regularization for bivariate smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(1), 145–163.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*(4), 673–680.

Kohler, M., A. Krzyżak, and H. Walk (2006). Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis 97*, 311–323.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *Ann. Statist. 25*(2), 435–477.

Kong, L. and I. Mizera (2012). Quantile tomography: using quantiles with multivariate data. *Statist. Sinica 22*(4), 1589–1610.

Lejeune, M. G. and P. Sarda (1988). Quantile regression: a nonparametric approach. *Comput. Statist. Data Anal. 6*(3), 229–239.

Li, Q. and J. S. Racine (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *J. Bus. Econom. Statist. 26*(4), 423–434.

Liu, R. Y., J. M. Parelius, and K. Singh (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist. 27*(3), 783–858. With discussion and a rejoinder by Liu and Singh.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *Ann. Statist. 20*(2), 712–736.

Pagès, G. (1998). A space quantization method for numerical integration. *J. Comput. Appl. Math. 89*(1), 1–38.

Pagès, G., H. Pham, and J. Printems (2004a). An optimal Markovian quantization algorithm for multi-dimensional stochastic control problems. *Stoch. Dyn. 4*(4), 501–545.

Pagès, G., H. Pham, and J. Printems (2004b). Optimal quantization methods and applications to numerical problems in finance. In *Handbook of computational and numerical methods in finance*, pp. 253–297. Boston, MA: Birkhäuser Boston.

Pagès, G. and J. Printems (2003). Optimal quadratic quantization for numerics: the Gaussian case. *Monte Carlo Methods Appl. 9*(2), 135–165.

Paindaveine, D. and M. Šiman (2011). On directional multiple-output quantile regression. *J. Multivariate Anal. 102*(2), 193–212.

Paindaveine, D. and M. Šiman (2012a). Computing multiple-output regression quantile regions. *Comput. Statist. Data Anal. 56*(4), 840–853.

Paindaveine, D. and M. Šiman (2012b). Computing multiple-output regression quantile regions from projection quantiles. *Comput. Statist. 27*(1), 29–49.

Pierce, J. N. (1970). Asymptotic quantizing error for unbounded random variables. *IEEE Trans. Information theory IT-16*, 81–83.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rousseeuw, P. J. and I. Ruts (1999). The depth function of a population distribution. *Metrika 49*(3), 213–244.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statist. Neerlandica 56*(2), 214–232. Special issue: Frontier research in theoretical statistics, 2000 (Eindhoven).

Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *J. Nonparametr. Stat. 22*(7), 915–936.

Wei, Y. (2008). An approach to multivariate covariate-dependent quantile contours with appli-

cation to bivariate conditional growth charts. *J. Amer. Statist. Assoc. 103*(481), 397–409.

Yu, K. (1999). Smoothing regression quantile by combining k-NN estimation with local linear kernel fitting. *Statist. Sinica 9*(3), 759–774.

Yu, K. (2002). Quantile regression using RJMCMC algorithm. *Comput. Statist. Data Anal. 40*(2), 303–315.

Yu, K. and M. C. Jones (1997). A comparison of local constant and local linear regression quantile estimators. *Comput. Statist. Data Anal. 25*, 159–166.

Yu, K. and M. C. Jones (1998). Local linear quantile regression. *J. Amer. Statist. Assoc. 93*(441), 228–237.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *J. R. Stat. Soc. Ser. D Statistician 52*(3), 331–350.

Zador, P. L. (1964). *Development and evaluation of procedures for quantizing multivariate distributions*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

Zuo, Y. and R. Serfling (2000). General notions of statistical depth function. *Ann. Statist. 28*(2), 461–482.