Estimation des quantiles conditionnels par quantification optimale : nouveaux résultats

Isabelle Charlier [1,2,3] & Davy Paindaveine [1,2] & Jérôme Saracco [3]

[1] *Université Libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus Plaine, CP210, B-1050, Bruxelles, Belgique.*
ischarli@ulb.ac.be, dpaindav@ulb.ac.be
[2] *ECARES, 50 Avenue F.D. Roosevelt, CP114/04, B-1050, Bruxelles, Belgique.*
[3] *Université de Bordeaux, Institut de Mathématiques de Bordeaux, UMR CNRS 5251 et INRIA Bordeaux Sud-Ouest, équipe CQFD, 351 Cours de la Libération, 33405 Talence.*
Jerome.Saracco@math.u-bordeaux1.fr

Résumé. Nous construisons un estimateur non-paramétrique des quantiles conditionnels de Y sachant X en utilisant la quantification optimale. Les quantiles conditionnels sont particulièrement intéressants lorsqu'il apparaît que la moyenne conditionnelle seule ne permet pas de représenter convenablement l'impact de la covariable X sur la variable dépendante Y. La quantification optimale en norme L^p est une méthode de discrétisation utilisée depuis les années 1950 en ingénierie. Elle permet d'obtenir la meilleure approximation d'une distribution continue par une distribution discrète de support de taille N.

Le but de ce travail est donc d'appliquer la quantification optimale à l'estimation de quantiles conditionnels. Nous étudions la convergence de l'approximation ainsi définie ($N \to \infty$) et de l'estimateur en découlant ($n \to \infty$). Celui-ci a été implémenté dans R afin d'en évaluer le comportement numérique et de réaliser une étude de simulations. Nous l'avons ensuite comparé aux méthodes existantes.

Mots-clés. Estimation non-paramétrique, Quantile conditionnel, Quantification optimale.

Abstract. We construct a nonparametric estimator of conditional quantiles of Y given X using optimal quantization. Conditional quantiles are particularly of interest when it is felt that conditonal mean is not representative of the impact of the covariable X on the dependent variable Y. Optimal quantization in L^p-norm is a discretizing method used since the fifties in engineering. We use it to find the best approximation of X by a discrete version with support of size N.

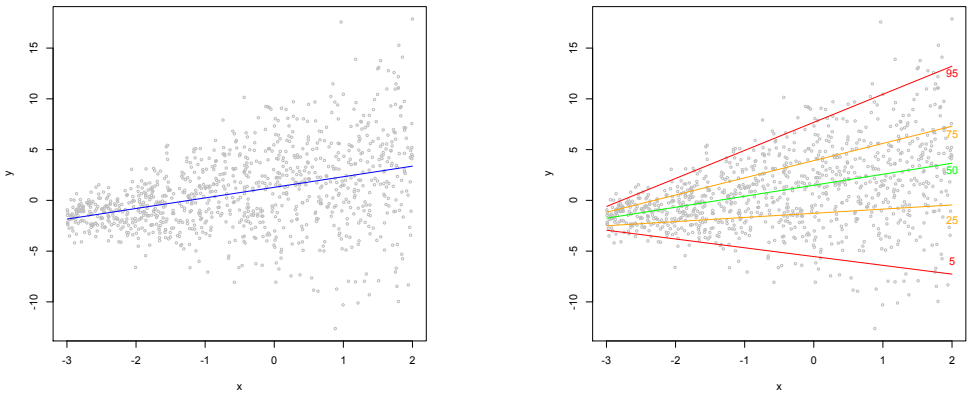
The aim of this work is to apply optimal quantization to conditional quantile estimation. We study the convergence of the approximation defined above ($N \to \infty$) and of the resulting estimator ($n \to \infty$). It was implemented in R in order to evaluate its numerical behavior and realize a simulation study. We then compare it with existing methods.

Keywords. Nonparametric estimation, Conditional quantile, Optimal quantization.

1 Les quantiles conditionnels

Introduite par Koenker et Bassett [3], la notion de quantiles conditionnels a pour principal intérêt de fournir une alternative à la moyenne conditionnelle en représentant de manière plus claire et complète l'impact des covariables sur la variable dépendante, notamment dans la cas de données hétéroscédastiques. Nous observons dans la Figure 1(b) une image beaucoup plus précise de la distribution conditionnelle de Y sachant X grâce aux courbes de quantiles que celle obtenue avec la moyenne conditionnelle à la Figure 1(a). La littérature sur ce sujet s'est depuis fortement développée et plusieurs estimateurs non-paramétriques ont été étudiés (voir notamment Bhattacharya et Gangopadhyay [2] et Yu et Jones [6]).

Les quantiles conditionnels sont définis de deux manières équivalentes. La définition en termes de solution d'un problème d'optimisation est celle que nous avons choisie dans notre étude.



(a) (b)

FIGURE 1 – (a) En bleu, la courbe de la moyenne conditionnelle de Y sachant X, (b) Cinq courbes de quantiles conditionnels : en vert, $\alpha = 0.5$; en orange, 0.25 pour la plus basse et $\alpha = 0.75$ pour la plus haute ; en rouge, $\alpha = 0.05$ pour la plus basse et 0.95 pour la plus haute.

Définition 1.1. *Le* quantile conditionnel d'ordre α *de* Y *sachant* $X = x$ *est défini par*

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

où $\rho_\alpha(z) = \alpha z \mathbb{I}_{[z \geq 0]} - (1 - \alpha) z \mathbb{I}_{[z > 0]}$ *est appelée la fonction de perte ou de coût.*

En pratique, la distribution conditionnelle de Y sachant $X = x$ est inconnue et nous voulons l'estimer à partir d'un échantillon de taille n, $\{(X_i, Y_i)\}_{i=1,\ldots,n}$, au moyen des fonctions de quantiles conditionnels. L'estimation de ces quantiles permet de construire

des courbes de référence à l'intérieur desquelles se trouvera une certaine proportion des observations. Ces courbes de référence ont de nombreuses applications et sont utilisées dans divers domaines comme en médecine (courbe de référence pour la croissance par exemple) mais aussi en économie, écologie, analyse de durée de vie, etc.

2 Quantification optimale en norme L^p

D'abord utilisée en ingénierie (traitement du signal et théorie de l'information), la quantification optimale consiste en la discrétisation d'un signal continu à l'aide d'un nombre fixé de points, les *quantifieurs*, dont la position doit être judicieusement choisie afin de rendre la transmission du signal la plus efficace possible. Elle a depuis été utilisée dans de nombreux autres domaines mais encore très rarement en statistique (voir par exemple Azaïs, Gégout-Petit et Saracco [1]).

Dans un contexte mathématique, la quantification optimale en norme L^p, $p \geq 1$, consiste à étudier la meilleure approximation d'un vecteur aléatoire X de dimension d par un vecteur $q(X)$ prenant au plus N valeurs dans $\mathbb{R}^d$. Nous cherchons donc un vecteur tel que l'erreur de quantification en norme L^p, $\|X - q(X)\|_p$, soit minimale, avec $\|Z\|_p = \mathrm{E}[|Z|^p]^{1/p}$, où $|\cdot|$ dénote la norme euclidienne dans $\mathbb{R}^d$.

Plus précisément, fixons N et considérons x une grille composée de N points $x_1, \ldots, x_N$ appartenant à $\mathbb{R}^d$. Nous cherchons la fonction $q_x : \mathbb{R}^d \to x = \{x_1, \ldots, x_N\}$ telle que

$$\|X - q_x(X)\|_p = \inf\{\|X - q(X)\|_p, q : \mathbb{R}^d \to x \text{ est une fonction de Borel}\}.$$

La solution de ce problème est la fonction q_x projetant sur le plus proche voisin. Plus précisément, soit $C_i(x)$, $i = 1, \ldots, N$ une partition de Borel de $\mathbb{R}^d$ satisfaisant, pour tout $i = 1, \ldots, N$, $C_i(x) \subset \{y \in \mathbb{R}^d : |x_i - y| = \min_{1 \leq j \leq N} |x_j - y|\}$. En clair, un point va appartenir à la cellule $C_i(x)$ si et seulement si le point x_i est le point de la grille qui lui est le plus proche. Une telle partition est appelée partition de Voronoi. La fonction q_x est donc définie comme

$$q_x(\xi) = \sum_{i=1}^{N} x_i \mathbb{I}_{C_i(x)}(\xi),$$

pour tout vecteur $\xi \in \mathbb{R}^d$. Nous discrétisons donc X de la façon suivante.

Définition 2.1. *Le vecteur aléatoire* $\widehat{X}^x = q_x(X) = \sum_{i=1}^{N} x_i \mathbb{I}_{C_i(x)}(X)$ *est appelé la quantification de Voronoi de* X*.*

Il reste à choisir la grille x de manière à minimiser l'erreur de quantification $\|X - \widehat{X}^x\|_p$. L'existence d'une grille atteignant le minimum a été obtenue sous la condition que la loi de X ne charge pas les hyperplans (voir Pagès [4]). Il n'existe cependant pas de résultat d'unicité ni de forme fermée nous permettant de déterminer une N-grille optimale. En pratique, nous utilisons l'algorithme du gradient stochastique permettant d'en obtenir une à N fixé (voir Pagès [4] et Pagès et Printems [5]).

3 Approximation des quantiles conditionnels

Soit X un vecteur aléatoire de dimension d, $X : (\Omega, \mathcal{F}, P) \to \mathbb{R}^d$ et soit $p \geq 1$ tel que $X \in \mathrm{L}^p(\Omega)$. Nous considérons le modèle non-paramétrique

$$Y = f(X, \epsilon),$$

où la variable réponse Y est liée à X par une fonction f inconnue et où ϵ est un terme d'erreur indépendant de X. Nous voulons approcher les quantiles conditionnels de Y sachant $X = x$. L'idée est de remplacer X dans la Définition 1.1 par une approximation discrète construite à l'aide de la quantification optimale. Plus précisément, nous nous plaçons sous l'hypothèse suivante.

Hypothèse (A) (i) Il existe $\delta > 0$ tel que $\|X\|_{p+\delta} < \infty$; (ii) la loi de X ne charge pas les hyperplans ; (iii) il existe une constante $C > 0$ telle que, pour tout $u, v \in \mathbb{R}^d$, $\|f(u, \epsilon) - f(v, \epsilon)\|_p \leq C|u - v|$.

Ces hypothèses nous permettent notamment d'utiliser les résultats classiques en quantification pour X. Par conséquent, pour N fixé, il existe $\gamma_N \in (\mathbb{R}^d)^N$ une grille optimale pour X et nous quantifions X en le projetant sur cette grille, ce qui nous donne le vecteur aléatoire discret $\widehat{X}^N$. Soient $x \in \mathbb{R}^d$ et $\hat{x}$ sa projection sur la grille γ_N, c'est-à-dire le point qui lui est le plus proche au sens de la norme euclidienne. Nous approchons alors $q_\alpha(x)$ par

$$\widehat{q}_\alpha^N(x) = \arg\min_{a \in \mathbb{R}} \mathrm{E}[\rho_\alpha(Y - a) | \widehat{X}^N = \hat{x}].$$

Le théorème suivant nous permet de contrôler l'écart entre $\widehat{q}_\alpha^N(x)$ et le vrai quantile conditionnel.

Théorème 3.1. *Sous l'hypothèse (A) et pour $\alpha \in (0, 1)$, nous avons*

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

pour $N \to \infty$.

4 Estimation des quantiles conditionnels

Supposons maintenant que $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ sont n copies indépendantes de (X, Y). Nous construisons une grille optimale γ_N à l'aide de l'algorithme du gradient stochastique [1]. Nous projetons les X_i sur cette grille, et nous définissons alors l'estimateur par

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

1. Nous ne détaillons pas plus ici l'initialisation et le fonctionnement de cet algorithme, voir Pagès [5] pour plus de détails.

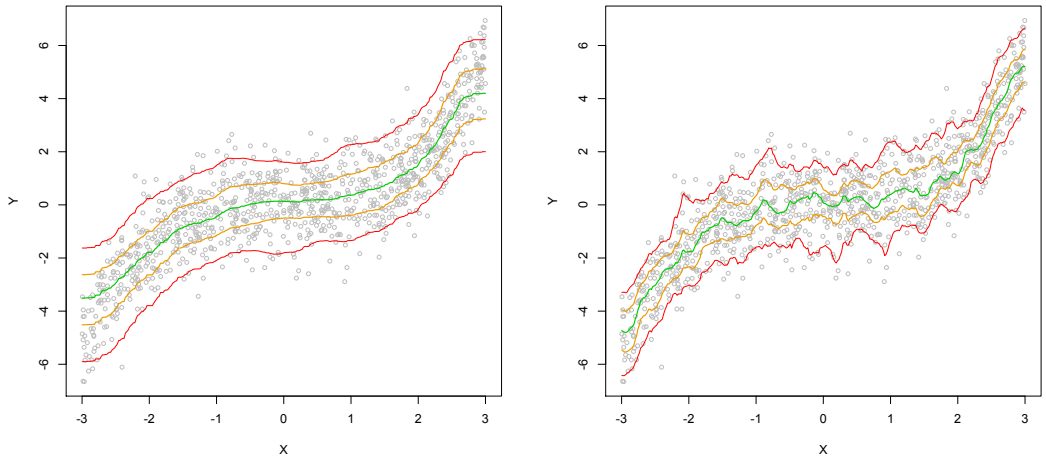
La convergence presque sûre de cet estimateur vers l'approximation définie plus haut a été obtenue pour $n \to \infty$ avec N fixé sous certaines hypothèses supplémentaires nous assurant que l'algorithme utilisé fournit (à $n \to \infty$) une grille optimale.

Dans le cas de petites tailles d'échantillons ($n < 2000$), l'algorithme du gradient stochastique peut fournir une grille qui n'est en réalité pas optimale, ce qui a un impact conséquent sur les courbes estimées. Pour corriger ce problème, nous avons eu recours à des estimations bootstrap. L'idée est de générer à partir de notre échantillon de départ B échantillons bootstrap de taille n avec remise. Nous procédons alors comme expliqué ci-dessus avec chacun de ces échantillons et nous obtenons B estimations, notées $\hat{q}^{(b)}_{\alpha,n}(x)$ pour $b = 1, \ldots, B$, que nous moyennons ensuite. Plus précisément, nous avons

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

En pratique, nous avons choisi $B = 50$ afin d'obtenir des courbes suffisamment lisses sans toutefois augmenter de manière conséquente le temps de calcul.

Un exemple de construction de courbes de référence fondées sur notre estimateur est donné à la Figure 2.



(a) (b)

FIGURE 2 – L'échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$ de taille $n = 1000$ est généré à partir du modèle $Y = \frac{1}{5}X^3 + \epsilon$, avec $X \sim U[-3, 3]$ et $\epsilon \sim \mathcal{N}(0, 1)$ indépendant de X. Les courbes estimées des quantiles conditionnels ont été obtenues avec $N = 10$ (à gauche) et $N = 45$ (à droite), et $\alpha = 0.05$ (en rouge, en bas), $\alpha = 0.25$ (en orange, en bas), $\alpha = 0.5$ (en vert), $\alpha = 0.75$ (en orange, en haut), $\alpha = 0.95$ (en rouge, en haut).

5 Étude de simulations et comparaison aux méthodes existantes

Nous avons considéré différents modèles et différentes tailles d'échantillons. Le nombre de quantifieurs N ayant un impact conséquent sur les courbes, nous avons proposé une méthode basée uniquement sur les données permettant de choisir N. Pour ce faire, nous avons étudiée les courbes du MSE (erreur quadratique moyenne) en fonction de N à n fixé, erreur commise en estimant le quantile conditionnel à l'aide de notre estimateur. Nous en avons observé la convexité : il existe donc une valeur de N minimisant l'erreur commise. Notre procédure de sélection de N est alors basée sur cette observation et utilise des réplications bootstrap pour remplacer les quantiles conditionnels théoriques apparaissant dans le MSE. Tout ceci a été implémenté en R et un package est en cours de développement.

Nous avons alors comparé notre méthode à deux estimateurs non-paramétriques des quantiles conditionnels : l'estimateur linéaire local de Yu et Jones [6] et l'estimateur des k plus proches voisins de Bhattacharya et Gangopadhyay [2]. Pour chaque modèle considéré, nous avons réalisé des graphiques de type « boxplot » pour différentes valeurs de n et α. Nous avons observé que nous obtenions de meilleurs résultats que l'estimateur des k plus proches voisins, et des résultats similaires voire meilleurs dans certains cas que l'estimateur linéaire local.

Bibliographie

[1] Romain Azaïs, Anne Gégout-Petit, and Jérôme Saracco. Optimal quantization applied to sliced inverse regression. *J. Statist. Plann. Inference*, 142(2) :481–492, 2012.

[2] P. K. Bhattacharya and Ashis K. Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist.*, 18(3) :1400–1415, 1990.

[3] Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1) :33–50, 1978.

[4] Gilles Pagès. A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89(1) :1–38, 1998.

[5] Gilles Pagès and Jacques Printems. Optimal quadratic quantization for numerics : the Gaussian case. *Monte Carlo Methods Appl.*, 9(2) :135–165, 2003.

[6] Keming Yu and M. C. Jones. Local linear quantile regression. *J. Amer. Statist. Assoc.*, 93(441) :228–237, 1998.