Conditional quantile estimation using optimal quantization: a numerical study

Isabelle Charlier, Davy Paindaveine, Jérôme Saracco

Conditional quantile estimation using optimal quantization: a numerical study

Isabelle Charlier, *Université Libre de Bruxelles and Université de Bordeaux*, ischarli@ulb.ac.be
Davy Paindaveine, *Université Libre de Bruxelles*, dpaindav@ulb.ac.be
Jérôme Saracco, *Université de Bordeaux*, jerome.saracco@math.u-bordeaux1.fr

Abstract. We construct a nonparametric estimator of conditional quantiles of Y given $X = x$ using optimal quantization. Conditional quantiles are particularly of interest when the conditional mean is not representative of the impact of the covariable X on the dependent variable Y. L_p-norm optimal quantization is a discretizing method used since the 1950's in engineering. It allows to construct the best approximation of a continuous law with a discrete law with support of size N. The aim of this work is then to use optimal quantization to construct conditional quantile estimators. We study the convergence of the approximation ($N \to \infty$) and the consistency of the resulting estimator for this fixed-N approximation. This estimator was implemented in R in order to evaluate the numerical behavior and to compare it with existing methods.

Keywords. Nonparametric estimation, Conditional quantile, Optimal quantization.

1 Introduction

Quantile regression allows to assess the impact of a covariable X on a (scalar) response variable Y and is an alternative to standard regression. It is particularly of interest when the mean does not provide an enough satisfactory picture of the distribution. We then get a more complete picture of the conditional distribution if we consider the conditional quantile functions

$$x \mapsto q_\alpha(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \alpha\}, \tag{1}$$

for various $\alpha \in (0, 1)$, where $F(\cdot|x)$ stands for the conditional distribution of Y given $X = x$. They are equivalently defined by solving the following optimization problem:

$$q_\alpha(x) = \arg\min_{a \in \mathbb{R}} \mathrm{E}[\rho_\alpha(Y - a)|X = x], \tag{2}$$

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

An important application of conditional quantiles is that they provide reference hypersurfaces (curves when $d = 1$) if we consider the quantile functions $x \mapsto q_\alpha(x)$ when x varies, and confident intervals of the form $I_\alpha = [q_\alpha(x), q_{1-\alpha}(x)]$ when x is fixed, which are widely used in many domains, as medicine, economics or lifetime analysis.

There exist many approaches to define conditional quantile estimators since the literature on quantile regression became really large in recent years. For example, [1] focuses on nearest-neighbor estimators of a conditional quantile while local linear estimator is investigated in [6].

We define in [2] a new estimator of conditional quantiles based on optimal quantization and we perform a numerical study of this estimator in [3]. Optimal quantization is a tool allowing to discretize any continuous distribution of a random vector X. It then provides an approximation of X by a discrete random vector with support of size N. This approximation is obtained by projecting X on a set of N points, called a grid. This grid is chosen in such a way that the L_p-norm difference between X and its discretized version is minimal. The reader can refer to [4, 5] for more details on optimal quantization.

We will first briefly recall in Section 2 the general idea of our method and explain the different steps in the construction of our estimator. Then, in Section 3, we provide a numerical comparison of our estimator with three alternative quantiles estimators.

2 Conditional quantile estimation through optimal quantization

In this section, we first explain the general idea of the construction of our estimator introduced in [2]. We then detail point by point this construction that is implemented in a R package called `QuantifQuantile` (available on the CRAN).

In the sequel, we denote by Y a real random variable and X a d-dimensional random vector. We define an estimator of conditional quantiles thanks to L_p-norm quantization. The idea is to replace X in (2) by a discrete version, obtained by projecting X on an optimal quantization grid. We then take an empirical version of this approximation. Let us specify this construction.

Assume that X belongs to L_p, *i.e.* $\|X\|_p := \mathrm{E}[|X|^p]^{1/p} < \infty$. Let $\gamma^N \in (\mathbb{R}^d)^N$ a set of N points of $\mathbb{R}^d$, called a *grid*. We approximate X by the projection of X onto this grid, that we denote $\widetilde{X}^{\gamma^N} := \mathrm{Proj}_{\gamma^N}(X)$. Obviously, the quality of this approximation depends hugely on the choice of the grid. We then choose γ^N as the grid minimizing the *quantization error* $\|X - \widetilde{X}^{\gamma^N}\|_p$. Classic result in quantization ensures the existence (but not the unicity) of such grid under the assumption that the law of X does not charge any hyperplanes. We will denote in the sequel $\widetilde{X}^N$ the projection of X onto an optimal grid. In practice, an optimal grid is constructed using a *stochastic gradient algorithm*. This algorithm is detailed further. The reader may refer to [4] for more details on the concept of quantization. We then define

$$\widetilde{q}_\alpha^N(x) := \arg\min_{a\in\mathbb{R}} \mathrm{E}[\rho_\alpha(Y-a)|\widetilde{X}^N = \tilde{x}], \tag{3}$$

where $\tilde{x}$ is the projection of x onto γ^N.

Let us now assume that we have n independent copies $(X_1', Y_1)', \ldots, (X_n', Y_n)'$. We define an estimator of conditional quantiles by taking an empirical version of this approximation, denoted $\widehat{q}_\alpha^{N,n}(x)$. Its construction is provided in the sequel.

We derived the following theorems for the convergence of $\widetilde{q}_\alpha^N(x)$ when $N \to \infty$ and of $\widehat{q}_\alpha^{N,n}(x)$ when $n \to \infty$ and N fixed. We need the following assumptions.

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 hyperplanes.

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 $\mathbb{R}$).

To obtain rates of convergence, we will need the following reinforcement of Assumption (A).

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

ASSUMPTION (C) P_X is continuous and has a compact support.

Under these assumptions, the underlying curve m is quite smooth, which avoids possible peaks or jumps.

Theorem 2.1. *Fix $\alpha \in (0, 1)$. Then (i) under Assumptions (A)-(B),*

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

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),

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

Theorem 2.2. *Fix $\alpha \in (0, 1)$. Then, under Assumptions (A)-(B),*

$$\sup_{x \in S_X} \big|\widetilde{q}_\alpha^N(x) - q_\alpha(x)\big| \to 0, \quad \text{as } N \to \infty.$$

Theorem 2.3. *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 $p = 2$.

More details on these theorems and their proofs can be found in [2].

We will now explain step by step the construction of $\widehat{q}_\alpha^{N,n}(x)$. We will then complete this section with an illustration on some dataset.

Determining an optimal N-grid

Since the starting idea of our method consists in replacing X with a discrete version with support of size N, the first step is naturally dedicated to the choice of an optimal N-grid for X, with N fixed. Since no theoretical quantization result provides such a grid, the only way at our disposal

to get it is to use a stochastic gradient algorithm. Starting from an initial grid denoted $\hat{\gamma}^{N,0}$, we update the grid at step $t-1$ thanks to the observation X_t, playing the role of stimuli.We then obtain the grid at step t, for $t = 1, \ldots, n$. After n steps, we thus get a grid $\hat{\gamma}^{N,n}$ considered as optimal. Let us make it more precise.

Let $(\delta_t), t \in \mathbb{N}_0$, be a deterministic sequence in $(0,1)$ such that

$$\sum_t \delta_t = \infty \quad \text{and} \quad \sum_t \delta_t^2 < \infty.$$

For N fixed, the algorithm works as follows.

Algorithm 2.1.
For $t = 1 \ldots, n$,

Step 0 *The initial grid $\hat{\gamma}^{N,0}$ in $(\mathbb{R}^d)^N$ is chosen by sampling randomly among the X_i's without replacement.*

Step t *The grid at step t is defined recursively as*

$$\hat{\gamma}_i^{N,t} = \begin{cases} \hat{\gamma}_i^{N,t-1} - \delta_t |\hat{\gamma}_i^{N,t-1} - X_t|^{p-1} \frac{\hat{\gamma}_i^{N,t-1} - X_t}{|\hat{\gamma}_i^{N,t-1} - X_t|} & \text{if } Proj_{\hat{\gamma}^{N,t-1}}(X_t) = \hat{\gamma}_i^{N,t-1} \\ \hat{\gamma}_i^{N,t-1} & \text{otherwise} \end{cases},$$

where $\hat{\gamma}_i^{N,t} \in \mathbb{R}^d$ denotes the ith component of $\hat{\gamma}^{N,t}$, $i = 1, \ldots, N$.

We observe that only one point of the grid at step $t-1$ moves at each step t: the one on which the stimuli X_t is projected.

The resulting grid $\hat{\gamma}^{N,n}$ allows thus to quantize X: we define $\widehat{X}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}} X$. This is important to point out that this quantization step provides a grid that is chosen independently of Y. Thus, the link function m does not play any role in this step.

Estimating conditional quantiles

As above-mentioned, an approximation of conditional quantiles is defined by replacing X by its projection on the optimal N-grid in the definition. An estimator is then constructed by taking an empirical version of this approximation, as follows :

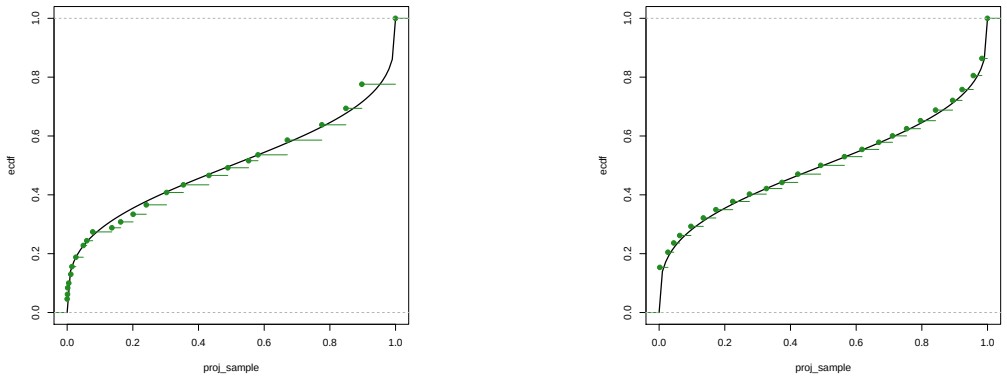
Algorithm 2.2.
Let $(X_1', Y_1)', \ldots, (X_n', Y_n)'$ be n independent copies of (X, Y).

Step 1 *We project each X_i on the grid $\hat{\gamma}^{N,n}$ and we write $\widehat{X}_i^N = Proj_{\hat{\gamma}^{N,n}}(X_i)$. We then work with the projected sample $\{(\widehat{X}_i^{N\prime}, Y_i)'\}_{i=1,\ldots,n}$.*

Step 2 *The conditional quantiles are then estimated by*

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

where $\hat{x}^N = Proj_{\hat{\gamma}^{N,n}}(x)$. In practice, $\widehat{q}_\alpha^{N,n}(x)$ is simply evaluated as the sample α-quantile of the Y_i's whose corresponding X_i admits $\hat{x}^N$ as projection onto $\hat{\gamma}^{N,n}$.



(a) (b)

Figure 1: Comparison of population (black) and grid-projected sample (green) cdf for samples of size 500 (left) and 5000 (right) generated from a beta distribution (for the green one, the sample was projected onto an optimal quantization grid of size $N = 25$).

The first step of this algorithm is illustrated in Figure 1. We generate observations with law Beta(0.3,0.3) and we consider $N = 25$. Using Algorithm 2.1, an optimal grid is constructed and we project the sample onto this grid. The left graph represents the grid-projected sample cumulative distribution function (cdf) in green and the population one in black for a sample size $n = 500$. The right one is similar with $n = 5000$. We observe that the grid-projected sample versions fit very well the population ones (better and better when n increases).

Nevertheless, the grid provided by the stochastic gradient algorithm may be a poor approximation of the optimal one when the sample size is small (when n is equal to 300 or less). Indeed, $\hat{\gamma}^{N,n}$ is constructed after n iterations. As the choice of the grid is the basis in the construction of our estimator, it has a major impact on the resulting reference curves that are not smooth. For this reason, we use bootstrap to introduce a more appropriate conditional quantile estimator.

For some integer B, we generate B samples of size n from our original sample $\{(X_i,' Y_i)'\}_{i=1,\ldots,n}$ with replacement. Each bootstrap sample is then used as stimuli to construct a grid by performing the stochastic gradient algorithm. Thanks to these B grids, we get B estimations of $q_\alpha(x)$ thanks to Algorithm 2.2, that we denote $\hat{q}_\alpha^{(1)}(x), \ldots, \hat{q}_\alpha^{(B)}(x)$. The bootstrap version of our estimator is then defined as:

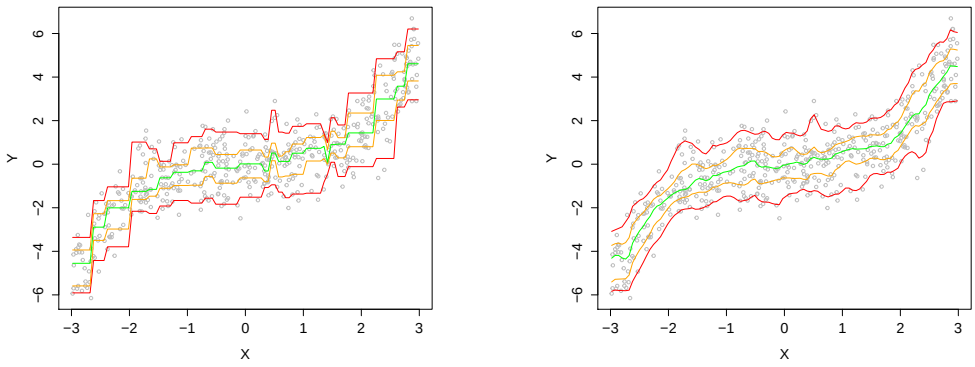
$$\bar{q}_{\alpha,B}^{N,n}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{q}_\alpha^{(b)}(x). \tag{4}$$

We usually take $B = 50$ when X is univariate.

Figure 2 represents the curves of estimated conditional quantiles for a sample of size $n = 500$. The left panel of Figure 2 is obtained without bootstrap and the right one with bootstrap. This bootstrap version provides clearly smoother curves.

Selecting the number N of quantizers

The choice of the number N of quantizers is crucial: for too small N, the curves show a large bias and for too large N, the variability is important but the bias smaller.



(a) Without bootstrap (b) $B = 50$

Figure 2: For $n = 500$, $X \sim U(-3, 3)$, $Y = X^3/5 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$ independent of X, the curves of estimated conditional quantiles, without and with bootstrap respectively. They are obtained with $\alpha = 0.05, 0.25, 0.5, 0.75$ and 0.95 respectively (upwards).

We propose a data driven selection criterion for N. As explained in [3], we first investigate the MSE (Mean Squared Error) as a function of N (by taking some suitable family of possible values for N). These curves are actually convex and we choose an optimal value for N as the "arg min" of $\mathrm{MSE}(N)$. Of course, the MSE is calculated using the true conditional quantiles. We then propose a bootstrap estimate of the MSE that only uses the observations. We observe that the corresponding curves are convex and minimized for a N close the optimal one for the true MSE (see [3] for more details). Let us specify this criterion.

Let $\{x_1, \ldots, x_{\mathcal{N}_x}\}$ be a set of $\mathcal{N}_x$ deterministic points for which we want to estimate $q_\alpha(x)$ (generally equispaced on the support of X). We actually generate $B + \widetilde{B}$ bootstrap samples of size n from the initial sample. The first B bootstrap samples allows to construct $\bar{q}^{N,n}_{\alpha,B}(x_j)$ as above explained. The last $\widetilde{B}$ are used to calculate $\widetilde{B}$ estimations of $q_\alpha(x_j)$, that we denote $\hat{q}^{(\tilde{b})}_\alpha(x_j)$, for $\tilde{b} = 1 \ldots, \tilde{B}$. The true conditional quantiles are replaced by $\hat{q}^{(\tilde{b})}_\alpha(x_j)$ in the expression of the MSE, and we take the mean of these $\tilde{B}$ versions. More precisely, we define

$$\widehat{\mathrm{MSE}}(N) = \frac{1}{\mathcal{N}_x} \sum_{j=1}^{\mathcal{N}_x} \left(\frac{1}{\widetilde{B}} \sum_{\tilde{b}=1}^{\widetilde{B}} \left(\hat{q}^{(\tilde{b})}_\alpha(x_j) - \bar{q}^{N,n}_{\alpha,B}(x_j) \right)^2 \right). \tag{5}$$

We then select the optimal number N of quantizers as

$$\widehat{N}^* = \arg\min_{N \in \mathfrak{N}} \widehat{\mathrm{MSE}}(N), \tag{6}$$

where $\mathfrak{N}$ denotes a grid of values for N chosen according to the sample size of the considered dataset.

3 Comparison with alternative conditional quantile estimators

We explained in the previous section the construction of our estimator and we proposed a selection criterion for the tuning parameter N. We now recall three well-known conditional quantile

estimators and the selection criteria for their own tuning parameters. We then summarize the boxplot comparison realized in [3] and specify which estimators seem preferable in each situation.

The k nearest-neighbor is introduced in [1]. This estimator of $q_\alpha(x)$ is defined as follows. Let $X_i^* = |X_i - x|$ for $i = 1, \ldots, n$ and let $X_{n1}^* < \cdots < X_{nn}^*$ denote the order statistics of $X_1^*, \ldots, X_n^*$ and $Y_{n1}, \ldots, Y_{nn}$ the induced order statistics of $(X_1^*, Y_1), \ldots, (X_n^*, Y_n)$, i.e. $Y_{ni} = Y_j$ if $X_{ni}^* = X_j^*$. For any positive integer $k \leq n$, the k nearest-neighbor estimator $\widehat{q}_\alpha^k(x) = \widehat{q}_\alpha^{k,n}(x)$ is the $[k\alpha]$th order statistics of $Y_{n1}, \ldots, Y_{nn}$. The idea is to select the k points of the data such that their X's are the nearest of x, whence the name, and to calculate the quantile of order α of their Y's. Of course, k plays the role of tuning parameter and must be specified. Since we did not find in the literature an efficient method to select k only based on the data, we choose k by taking it minimizing the mean squared error among an set of values for k, that we will denote k^*.

The kernel weighted local linear estimator introduced by [6] is the second competitor. This estimator is defined as $\widehat{q}_\alpha^{\mathrm{YJ}}(x) = \hat{a}$, with

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

where K is a kernel function, choosen as the standard normal density, and where h is the bandwidth. We choose h according to α as

$$h_\alpha = h_{\mathrm{mean}} \left(\frac{\alpha(1-\alpha)}{\varphi(\Phi^{-1}(\alpha))^2}\right),$$

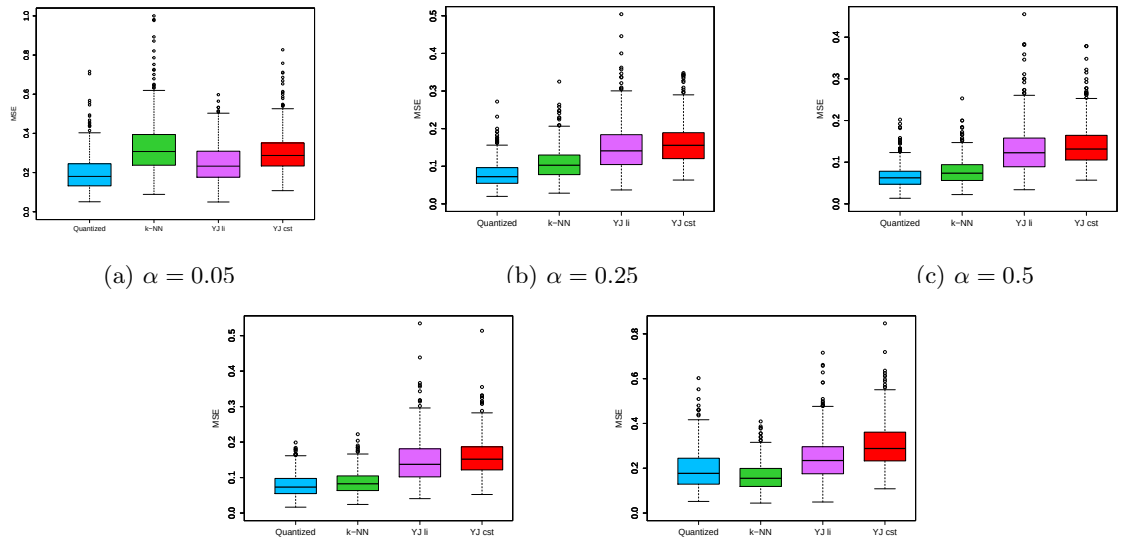
where φ and Φ are respectively the standard normal density and distribution functions, and where h_{mean} is the optimal choice of h for regression mean estimation, selected thanks to a cross-validation criteria. We also consider the local constant version of this estimator. More precisely, it is defined as $\widehat{q}_\alpha^{\mathrm{YJc}}(x) = \hat{a}$, with

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

and where the kernel function and the bandwidth are chosen as in the local linear case.

Notice that an important point in conditional quantile estimation is the choice of the observations X_i that will be taken into account when estimating $q_\alpha(x)$. We see that $\widehat{q}_\alpha^{\mathrm{YJ}}(x)$ and $\widehat{q}_\alpha^{\mathrm{YJc}}(x)$ choose it thanks to some bandwidth while $\widehat{q}_\alpha^k(x)$ is constructed using the k observations whose X-part is the closest to x. Our method is based on the observations whose X-part is projected on the same point of the grid as x (we call the set of such points a quantization cell). The main advantage of our method is then that the number of observations used to estimate $q_\alpha(x)$ is adaptive with x. The choice of a bandwidth is felt to be interesting when the observations X are quite uniformly distributed on the support of X but it may be disadvantageous when the density of the points is smaller in some regions of the support.

We then compare our estimator with these competitors. We consider different models (homoscedastic and heteroscedastic) and sample sizes ($n = 300$, 500 and 1000). For each of them, we generate 500 samples and we calculate the estimates $\bar{q}_{\alpha,B}^{N,n}(x)$, $\widehat{q}_\alpha^{\mathrm{YJ}}(x)$, $\widehat{q}_\alpha^{\mathrm{YJc}}(x)$ and $\widehat{q}_\alpha^k(x)$. We then realize the boxplots of the mean squared error (MSE) according to each estimator. We generally observe that $\bar{q}_{\alpha,B}^{N,n}(x)$ generally outperforms its competitors when the covariate is



(a) $\alpha = 0.05$ (b) $\alpha = 0.25$ (c) $\alpha = 0.5$

(d) $\alpha = 0.75$ (e) $\alpha = 0.95$

Figure 3: For 500 replications of sample of size $n = 300$ from model $Y = \frac{1}{5}X^3 + \varepsilon$, the boxplots of the MSE in the estimation of the conditional quantile curves: in blue, with $\bar{q}^{N,n}_{\alpha,B}(x)$, in green, with $\widehat{q}^{k}_{\alpha}(x)$, in purple, with $\widehat{q}^{\mathrm{YJ}}_{\alpha}(x)$ and in red, with $\widehat{q}^{\mathrm{YJc}}_{\alpha}(x)$.

not uniformly distributed. In case of uniformly distributed X, $\widehat{q}^{\mathrm{YJ}}_{\alpha}(x)$ is often better. We illustrate it in Figure 3 where we generate 500 samples of size $n = 300$ with $X = 6Z - 3$,with $Z \sim \mathrm{Beta}(0.3, 0.3)$ and $Y = X^3/5 + \varepsilon$, where ε is a normal error term independent of X. We observe that $\bar{q}^{N,n}_{\alpha,B}(x)$ provides the smallest MSE, followed by $\widehat{q}^{k}_{\alpha}(x)$. More details on this comparison study can be found in [3].

Bibliography

[1] Bhattacharya, P.K. and Gangopadhyay, A.K. (1990) *Kernel and nearest-neighbor estimation of a conditional quantile.* Annals of Statistics, **7**(3), 1400–1414.

[2] Charlier, I., Paindaveine, D. and Saracco, J. (2014) *Conditional quantile estimation through optimal quantization.* Submitted.

[3] Charlier, I., Paindaveine, D. and Saracco, J. (2014) *Numerical study od a conditional quantile estimator based on optimal quantization.* Manuscript in preparation.

[4] Pagès, G. (1998) *A space quantization method for numerical integration.* Journal of Computational and Applied Mathematics, **89**(1), 1–38.

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

[6] Yu, K. and Jones, M.C. (1998) *Local linear quantile regression.* Journal of the American Statistical Association, **93**(441), 228–237.