

SECTION OF STATISTICS

DEPARTMENT OF MATHEMATICS

KATHOLIEKE UNIVERSITEIT LEUVEN



TECHNICAL REPORT

TR-06-10

Censored Depth Quantiles

Debruyne, M., Hubert, M., Portnoy, S. and
Vanden Branden, K.

<http://wis.kuleuven.be/stat/>

Censored Depth Quantiles

M. Debruyne^a, M. Hubert^{a,*}, S. Portnoy^b,
K. Vanden Branden^a

^a*Department of Mathematics - University Center for Statistics, K.U.Leuven,
Celestijnenlaan 200B, B-3001 Leuven, Belgium*

^b*Statistics Department, University of Illinois at Urbana-Champaign, Champaign,
IL 61820, USA*

Abstract

Quantile regression is a wide spread regression technique which allows to model the entire conditional distribution of the response variable. A natural extension to the case of censored observations has been introduced using a reweighting scheme based on the Kaplan-Meier estimator. The same ideas can be applied to depth quantiles. This leads to regression quantiles for censored data which are robust to both outliers in the predictor and the response variable. For their computation, a fast algorithm over a grid of quantile values is proposed. The robustness of the method is shown in a simulation study and on two real data examples.

Key words: Regression depth, quantile regression, censoring, robustness.

1 Introduction

Since its introduction by [Koenker and Bassett \(1978\)](#), quantile regression has become more and more popular. The possibility to estimate the entire conditional distribution, instead of only the conditional mean as in e.g. ordinary least squares regression, has proven to be advantageous in many applications. In recent years, quantile regression has been extended to many possible settings, such as non-linear and non-parametric regression, time series, etc. ([Koenker, 2005](#)). In this paper we focus on linear quantile regression with right censored observations. These are observations for which the true value of the response variable is not measured, but only a lower limit is given. This

* Corresponding author. Tel: +32-(0)16-322023. Fax: +32-(0)16-322831
Email address: mia.hubert@wis.kuleuven.be

kind of data is frequently encountered in many domains. In medicine for example, when time until healing is measured, patients might not yet be healed when finishing the study. In that case, the exact healing time is not observed, but we do know it is at least the time the patient spent in the study. Also in economics, censoring can be an issue. The first example we will give in Section 6 considers sales prices in auction type sales. When a bid is running on an object, but the deadline is not reached yet, this is a right censored observation. The true sales price is not known, but we do know it will be at least the current bid.

Let us first describe the model under consideration. We want to estimate the conditional quantiles of a real random variable Y given $\mathbf{x} \in \mathbb{R}^{p+1}$ where we take $x_{i1} = 1$. However, also models through the origin can be considered. We suppose throughout that these conditional quantiles denoted by $Q_\tau(Y|\mathbf{x})$ are linear in \mathbf{x} . So for $\tau \in (0, 1)$

$$Q_\tau(Y|\mathbf{x}) = \inf\{y : P(Y \leq y|X = \mathbf{x}) = \tau\} = \mathbf{x}'\boldsymbol{\beta}(\tau), \quad (1)$$

for $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \dots, \beta_{p+1}(\tau))'$ the τ th regression quantile. These assumptions correspond to the problem of estimating the regression quantiles in a linear, possibly heterogeneous, regression setting with response variable Y and covariates \mathbf{x} . Especially in the case of heterogeneous data these quantiles offer an overall view on the data as they catch much more the variability present in the sample when τ varies over the interval $(0, 1)$.

In [Koenker and Bassett \(1978\)](#), a consistent estimator $\hat{\boldsymbol{\beta}}(\tau)$ for $\boldsymbol{\beta}(\tau)$ has been defined. We will however assume that the observations can be right censored. This implies that instead of observing the true response y_i , we observe $\tilde{y}_i = \min(y_i, c_i)$ for a set of covariates $\mathbf{x}_i \in \mathbb{R}^{p+1}$ where c_i is the censoring time. A censoring indicator $\Delta_i = I(y_i \leq c_i)$, with I the indicator function, denotes whether observation i is censored ($\Delta_i = 0$) or observed ($\Delta_i = 1$). We assume independence between the response variable and the censoring times, conditionally on the covariates \mathbf{x} . The censoring times are however allowed to depend on the covariates, contrary to most other censored regression methods, e.g. [Honoré et al. \(2002\)](#), assuming censoring at random.

In [Portnoy \(2003\)](#) a reweighting scheme based on the Kaplan-Meier estimator has been developed for adapting Koenker and Bassett's L_1 -quantiles to the censored case. In Section 2 we review this L_1 -methodology and its extension towards censoring. A serious drawback of these estimators is the lack of robustness. Although they are resistant to vertical outliers, i.e. observations that are outlying in y given \mathbf{x} , L_1 -quantiles can be heavily influenced by leverage points, i.e. observations outlying in \mathbf{x} -space.

A more robust quantile estimator has been proposed in [Rousseeuw and Hubert \(1999\)](#), based on the concept of regression depth. The main goal of this

paper is to extend these depth quantiles to the framework of censored observations, using the same reweighting scheme as [Portnoy \(2003\)](#). This is done in Section 3. The major difficulties of this extension can be found in the computations. In contrast to [Portnoy \(2003\)](#) we can not rely on linear programming techniques. Instead we consider a direct approach in which we include the objective function of the depth quantile estimator. In order to speed up the computations an updating step is included. A detailed description of the algorithm can be found in Section 4. We illustrate our method with a simulation study in Section 5 and with two real data examples in Section 6.

2 L_1 -quantiles

A consistent estimator of the vector $\boldsymbol{\beta}(\tau)$ was proposed in [Koenker and Bassett \(1978\)](#) as the solution of

$$\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}'_i \boldsymbol{\beta}), \quad (2)$$

for a sample $(\mathbf{x}'_i, y_i)' \in \mathbb{R}^{p+2}$ where $i = 1, \dots, n$ and $\rho_{\tau}(u) = u(\tau - I(u < 0))$. When τ equals $\frac{1}{2}$, the function $\rho_{\frac{1}{2}}(u)$ reduces to the absolute value. Thus for the special case of the median, this estimator corresponds to the L_1 -estimator. Therefore the solution for general τ will be denoted by L_1 -quantiles further on. The asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau))$ has also been derived in [Koenker and Bassett \(1978\)](#). In the same article it was shown that L_1 -quantiles only depend on the sign of the residuals and not on the exact value of the response variable.

This is a very important observation when it comes to censored data. First of all, it allows an easy start for the lowest quantiles. Remember, the exact value y_i of the response variable is unknown for a censored observation, but we do know a lower limit c_i . Thus, as long as c_i lies above the τ th regression quantile, y_i certainly will. Hence the residual $y_i - \mathbf{x}'_i \boldsymbol{\beta}(\tau)$ will be positive no matter the true value of y_i . Therefore we can just use ordinary quantile regression for the smallest quantiles.

This changes of course as τ increases. Sooner or later a censored observation will have a negative residual $c_i - \mathbf{x}'_i \boldsymbol{\beta}(\tau)$. Then the true residual $y_i - \mathbf{x}'_i \boldsymbol{\beta}(\tau)$ might be either negative or positive, and there is no way of knowing the sign for sure. We will call such observations *crossed* from now on. The quantile at which the i th censored observation is crossed, will be denoted $\hat{\tau}_i$, thus

$$c_i - \mathbf{x}'_i \boldsymbol{\beta}(\hat{\tau}_i) \geq 0 \quad \text{and} \quad c_i - \mathbf{x}'_i \boldsymbol{\beta}(\tau) \leq 0 \quad \text{for all } \tau > \hat{\tau}_i.$$

The crucial idea explained in [Portnoy \(2003\)](#) is to estimate the probabilities of

crossed censored observations having a positive respectively negative residual, using these estimates as weights further on. More precisely, such a crossed censored observation is split into two new pseudo-observations, one at (\mathbf{x}_i, c_i) with weight $w_i(\tau) \approx P(y_i - \mathbf{x}'_i\boldsymbol{\beta}(\tau) \leq 0)$ and one at (\mathbf{x}_i, ∞) with weight $1 - w_i(\tau)$. Finally it is noted that the weights $w_i(\tau)$ can easily be found in quantile regression, since the number $1 - \hat{\tau}_i$ is an estimate of the censoring probability $P(y_i > c_i)$. Thus we can define

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} \quad \tau > \hat{\tau}_i. \quad (3)$$

This leads to the following method to deal with censored observations in linear L_1 -quantile regression:

- As long as no censored observations are crossed, use ordinary quantile regression as in (2).
- When the i th censored observation is crossed at the τ th regression quantile, store this value as $\hat{\tau}_i = \tau$.
- When estimating the τ th regression quantile and censored observations have been crossed, optimize a weighted version of (2):

$$\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \left\{ \sum_{i \in K_\tau^c} \rho_\tau(\tilde{y}_i - \mathbf{x}'_i\boldsymbol{\beta}) + \sum_{i \in K_\tau} [w_i(\tau)\rho_\tau(\tilde{y}_i - \mathbf{x}'_i\boldsymbol{\beta}) + (1 - w_i(\tau))\rho_\tau(y^* - \mathbf{x}'_i\boldsymbol{\beta})] \right\}, \quad (4)$$

where the set K_τ represents the crossed and censored observations at τ and K_τ^c is the complement of K_τ . The weights $w_i(\tau)$ are as defined in (3). The number y^* is any value sufficiently large to exceed all $\{\mathbf{x}'_i\boldsymbol{\beta}\}$.

To compute the regression quantile function in practice, a sequence of breakpoints $\{\tau_1^*, \dots, \tau_{L^*}^*\}$ is defined such that $\hat{\boldsymbol{\beta}}(\tau)$ is piecewise constant between these breakpoints. By simplex pivoting we can move from one breakpoint to another, using the subgradients of (4). Luckily, the resulting gradient conditions are linear in τ , making this linear programming approach possible. A detailed description of the algorithm can be found in [Portnoy \(2003\)](#), together with some consistency results (see also [Neocleous et al. \(2006\)](#)).

3 Depth quantiles

As already mentioned, L_1 -quantiles only depend on the sign of the residuals, not on the exact values of the response variable. Therefore one can immediately see that observations with outlying y -value will not have a large impact on

the estimates. In contrast to for example linear least squares regression, L_1 -quantiles can resist vertical outliers. However, L_1 -quantiles are sensitive to data points outlying in \mathbf{x} -space. This is reflected by its breakdown point, which equals 0. This means the slightest amount of contamination can have a disastrous effect on the resulting estimates. A more robust method was proposed in [Rousseeuw and Hubert \(1999\)](#), based on the concept of regression depth.

3.1 Definition

The regression depth of a hyperplane $\boldsymbol{\beta}$ with respect to a sample $Z_n = \{(\mathbf{x}'_i, y'_i)' \in \mathbb{R}^{p+2}\}$ is defined as

$$rdepth(\boldsymbol{\beta}, Z_n) = \min_{\boldsymbol{\lambda} \in \mathbb{R}^{p+1}} \left(\#\{\mathbf{x}_i : \text{sign}(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \neq \text{sign}(\mathbf{x}'_i \boldsymbol{\lambda})\} \right),$$

with $\text{sign}(u) = -1$ if $u < 0$, $\text{sign}(u) = 0$ if $u = 0$ and $\text{sign}(u) = 1$ if $u > 0$. It has a nice geometrical interpretation as it represents the smallest number of observations one has to pass in order to turn the hyperplane $\boldsymbol{\beta}$ into vertical position. As such, regression depth gives an indication of how well the data surrounds the hyperplane.

The maximal depth (or deepest regression) estimator $\hat{\boldsymbol{\beta}}(\frac{1}{2})$ is defined as

$$\hat{\boldsymbol{\beta}}(\frac{1}{2}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{argmax}} \{rdepth(\boldsymbol{\beta}, Z_n)\},$$

or equivalently ([Bai and He, 2000](#)):

$$\hat{\boldsymbol{\beta}}(\frac{1}{2}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{argmax}} \inf_{\boldsymbol{\gamma} \in S^p} \sum_{i=1}^n \text{sign}(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \text{sign}(\mathbf{x}'_i \boldsymbol{\gamma}), \quad (5)$$

with $S^p = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\| = 1\}$. Properties of this estimator have been studied in [Rousseeuw and Hubert \(1999\)](#), [Van Aelst et al. \(2002\)](#), [Van Aelst and Rousseeuw \(2000\)](#) and [Bai and He \(2000\)](#). In the latter paper it is proven that deepest regression is a consistent estimator of the median regression quantile $\beta(\frac{1}{2})$. The first papers show that the breakdown value of the deepest regression is around 33%. This means that theoretically smaller percentages of outliers cannot completely destroy the fit. Practical results show that the method can indeed easily resist at least up to 20% of outliers (vertical outliers as well as leverage points). Note that this is a big difference compared to the L_1 -estimator, which has breakdown value 0%.

The deepest regression estimator can be extended to the regression quantile setting by introducing the idea behind the function ρ_τ in definition (2)

(Rousseeuw and Hubert, 1999). This ρ_τ function can be seen as a weight function where a weight τ is given to positive residuals, and a weight $1 - \tau$ to negative residuals. Let $\Psi_\tau(u) = \tau - I(u < 0)$. Then the τ th regression depth quantile $\hat{\beta}(\tau)$ is defined as that value $\beta \in \mathbb{R}^{p+1}$ for which

$$\inf_{\gamma \in S^p} \sum_{i=1}^n \Psi_\tau(y_i - \mathbf{x}'_i \beta) \text{sign}(\mathbf{x}'_i \gamma)$$

is maximized. The case $\tau = 0.5$ coincides with the conditional median as defined in (5). In Bai and He (2000) the \sqrt{n} -consistency of $\hat{\beta}(\tau)$ has been shown and the limiting distribution of $\sqrt{n}(\hat{\beta}(\tau) - \beta)$ has been characterized.

An extension of regression depth to the case of censored data is proposed in Park and Hwang (2003). Their approach however only covers the special case $\tau = 0.5$. We introduce a general extension to censored data for all quantiles in the next section.

3.2 Censored depth quantiles

Similarly to the method defined above for the L_1 -regression quantiles, we introduce the reweighting scheme for the depth quantiles. Since depth quantiles also depend on the sign of the residuals only, the same idea can be used as in the previous section. The only thing changing is the objective function (4). We replace this expression by the corresponding depth quantile objective function, defined as

$$\begin{aligned} \operatorname{argmax}_{\beta \in \mathbb{R}^{p+1}} \inf_{\gamma \in S^p} & \left\{ \sum_{i \in K_\tau^c} \Psi_\tau(\tilde{y}_i - \mathbf{x}'_i \beta) \text{sign}(\mathbf{x}'_i \gamma) \right. \\ & \left. + \sum_{i \in K_\tau} [w_i(\tau) \Psi_\tau(\tilde{y}_i - \mathbf{x}'_i \beta) \text{sign}(\mathbf{x}'_i \gamma) + \tau(1 - w_i(\tau)) \text{sign}(\mathbf{x}'_i \gamma)] \right\}, \end{aligned} \quad (6)$$

where the set K_τ and the weights $w_i(\tau)$ are as defined in (3) and (4).

4 Computation

Although the idea for censored depth quantiles is the same as for L_1 -quantiles, the computation has to be done differently. Working with breakpoints τ_j^* is impossible, since gradient conditions cannot be obtained for depth quantiles. As such, the linear programming algorithm from Section 2 cannot be used. We now introduce another algorithm, using a grid $\{t_j : 0 < t_1 < t_2 < \dots < t_M < 1\}$ where M is the total number of grid points. Each censored observation receives a weight $w_i(\tau) = 1$ as long as it is not crossed by $\beta(\tau)$, i.e. $c_i -$

$\mathbf{x}'_i \boldsymbol{\beta}(\tau) > 0$. Once the censored observation c_i is crossed, say at grid point t_j , a weight $w_i(\tau)$ that varies along the grid is assigned to that observation c_i and a weight $1 - w_i(\tau)$ is placed at infinity. This weight $w_i(\tau)$ is defined as

$$w_i(\tau) = \frac{\tau - t_j}{1 - t_j},$$

for grid points $\tau > t_j$, corresponding to (3).

4.1 Algorithm

We will now in detail list all the steps of the algorithm for obtaining the censored depth quantiles. The `Matlab` routine (`cdq.m`) is part of LIBRA, Matlab Library for Robust Analysis (Verboven and Hubert, 2005), freely available at <http://wis.kuleuven.be/stat/robust.html>.

STEP 1 Choose a set of grid points $\{0 < t_1 < \dots < t_M < 1\}$. Estimate the t_1 th regression quantile using the regression depth quantile for uncensored data. Crossed censored observations can be ignored since they almost do not contain any information, if t_1 is small enough.

STEP 2 Suppose we have estimated the t_l th regression quantile $\hat{\boldsymbol{\beta}}(t_l)$. Then we also know the set of crossed censored observations $K_{t_l} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}(t_l) \leq 0\}$. For each of these crossed censored observations a number $\hat{\tau}_i$ has been given following equation (8) that will be explained in step 3 of the algorithm. The according weight is

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i}.$$

STEP 3 Estimate the t_{l+1} th regression quantile using expression (6), which in practice can be implemented as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}(t_{l+1}) = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} \inf_{\boldsymbol{\lambda} \in \mathbb{R}^p} & \left(t_{l+1} \#\{\tilde{y}_i \notin K_{t_l} : (r_i(\boldsymbol{\beta}) > 0, \mathbf{x}'_i \boldsymbol{\lambda} < 0)\} \right. \\ & + (1 - t_{l+1}) \#\{\tilde{y}_i \notin K_{t_l} : (r_i(\boldsymbol{\beta}) < 0, \mathbf{x}'_i \boldsymbol{\lambda} > 0)\} \\ & + t_{l+1} w_i(t_{l+1}) \#\{\tilde{y}_i \in K_{t_l} : (r_i(\boldsymbol{\beta}) > 0, \mathbf{x}'_i \boldsymbol{\lambda} < 0)\} \\ & + (1 - t_{l+1}) w_i(t_{l+1}) \#\{\tilde{y}_i \in K_{t_l} : (r_i(\boldsymbol{\beta}) < 0, \mathbf{x}'_i \boldsymbol{\lambda} > 0)\} \\ & \left. + t_{l+1} (1 - w_i(t_{l+1})) \#\{\tilde{y}_i \in K_{t_l} : (\mathbf{x}'_i \boldsymbol{\lambda} < 0)\} \right). \quad (7) \end{aligned}$$

The maximization is performed on a random grid of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ vectors and will be further explained in Section 4.2.

Consider the set $K_{\tau_{l+1}} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}(t_{l+1}) \leq 0\}$.

IF $K_{t_{l+1}} = K_{t_l}$,

the current estimate $\hat{\boldsymbol{\beta}}(t_{l+1})$ was found using the correct weights and is therefore a correct solution.

IF $K_{t_{l+1}} \neq K_{t_l}$,

then the weights should be changed. Observations in $K_{t_l} \setminus K_{t_{l+1}}$ are censored observations that were crossed but are not anymore. These receive weight 1 again. Observations from $K_{t_{l+1}} \setminus K_{t_l}$ are censored observations that are crossed just now, during the transition from t_l to t_{l+1} . We define the number

$$\hat{\tau}_i = t_l, \tag{8}$$

for each of these observations. Their weight is then

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} = \frac{\tau - t_l}{1 - t_l}.$$

The remaining weight $1 - w_i(\tau)$ is assigned to a pseudo-observation arbitrarily far away. Thus we find a new set of crossed censored observations. The regression quantile $\hat{\boldsymbol{\beta}}_{\tau_{l+1}}$ is then recomputed with this new set of weights.

We repeat this step until we find an estimate for which the weights remain the same, or until a predefined number of iterations is exceeded.

STEP 4 The algorithm stops when we have dealt with the last grid point τ_M , or when all observations with a positive residual are censored.

□

Note that at any grid point t_{l+1} , step 3 of the algorithm is repeated until a stable solution is found with $K_{t_l} = K_{t_{l+1}}$. The existence of such a stable solution is not guaranteed. However, this situation rarely occurred in our examples and simulations. Moreover, when this happened it was always for a very restricted area of τ values. Thus, by changing the problematic t_{l+1} value by a few thousandths, a stable solution can usually be found.

4.2 Optimization

One part of the algorithm still needs some explanation, i.e. the optimization of (7). We propose two methods.

A straightforward approach is to define a set of N hyperplanes at the start of the algorithm. We take $B = \{\beta_j, j = 1, \dots, N\}$ with each β_j a hyperplane through p randomly chosen data points. This way affine equivariance of the depth quantiles is retained. We can then maximize the objective function in (7) over this finite set B :

$$\hat{\beta}(t_{l+1}) = \operatorname{argmax}_{\beta_j \in B} \inf_{\beta_k \in B} \left(\right).$$

This approach was also used in [Adrover et al. \(2004\)](#) to compute regression quantiles in the uncensored case. Note that it is usually not necessary to scan all possibilities. Take for example $\beta_j \in B$, and suppose we kept track of the current maximum over $\beta_i \in B$, $i = 1, \dots, j - 1$. Then we do not really need to compute the infimum over all $\beta_k \in B$. As soon as we find a β_k such that the objective function is smaller than the current maximum, the infimum will certainly be smaller. Thus we can immediately discard β_j and proceed with β_{j+1} . As noted in [Adrover et al. \(2004\)](#), this leads to roughly $O(N \log(N))$ calculations to find $\hat{\beta}(t_l)$. Since we have M grid points, we roughly need $O(MN \log(N))$ calculations.

We propose a faster approach, explicitly making use of the iterative character of our algorithm. Suppose we want to compute the regression quantile at a grid point t_{l+1} . Then we already have an estimate $\hat{\beta}(t_l)$ of the regression quantile at t_l . Since t_l and t_{l+1} will not differ a lot, we can expect $\hat{\beta}(t_{l+1})$ to be close to $\hat{\beta}(t_l)$. We therefore suggest to perform the maximization in (7) over a set $B(\hat{\beta}(t_l))$ of N^* hyperplanes close to $\hat{\beta}(t_l)$:

$$\hat{\beta}(t_{l+1}) = \operatorname{argmax}_{\beta_j \in B(\hat{\beta}(t_l))} \inf_{\beta_k \in B} \left(\right).$$

The complexity of this algorithm is very roughly $O(MN^* \log(N))$. The gain in speed comes from the fact that N^* can usually be chosen much smaller than N . The set $B(\hat{\beta}(t_l))$ of hyperplanes close to $\hat{\beta}(t_l)$ can be obtained in several ways. We take hyperplanes that have $p - 1$ observations in common with $\hat{\beta}(t_l)$. Such a set can be constructed very fast using updating techniques.

5 Simulation study

We compare three algorithms: the L_1 -estimator using the package *crq* in R ([Portnoy, 2003](#)), the depth estimator using the basic algorithm and the depth estimator using the faster updating algorithm. The setting for our simulations is as follows: let ϵ be the percentage contamination and n the sample

size with $m = \text{round}(n\epsilon)$, then we generated $n - m$ datapoints $(\mathbf{x}_i, \tilde{y}_i)$ as follows:

- $\mathbf{x}_i = (1, x_{i2}, \dots, x_{ip})'$ with each $x_{ij} \sim N(0, 1)$.
- $y_i = x_{i2} + e_i$ with $e_i \sim N(0, 1)$.
- $c_i = 0.8x_{i2} + b + f_i$ with $f_i \sim N(0, 1)$. We considered different values for b , controlling the amount of censored observations in the data. All values reported are for $b = 1$, corresponding to roughly 20% of censored data points. Other percentages yielded similar results, at least up to 40%.
- $\tilde{y}_i = \min(c_i, y_i)$.

Note that the censoring c_i depends on the covariates and on the response variable. In other methods this is often not allowed, but the regression quantile approach only assumes conditional independence of c_i and y_i , given \mathbf{x}_i .

- We considered 2 cases: we took the m outliers all coinciding in (\mathbf{x}_0, y_0) (point contamination), but we also distributed the m outliers around (\mathbf{x}_0, y_0) . Since there was no big difference in results, we only report the case of point contamination. The outlier location was taken equal to $((1, -5, 0, \dots, 0)', 10)'$. This is motivated by [Adrover et al. \(2002\)](#), where this setup appeared as the worst case scenario in a very similar simulation study for robust (but uncensored) quantile regression.

Each simulation consists of 50 replications. We report the median of the squared errors ($\|\hat{\beta}_\tau - \beta_\tau\|^2$) in the grid points 0.1, 0.2, \dots , 0.8. We considered sample sizes $n = 50, 100$, dimensions $p = 2, 5, 10$ and percentages of contamination $\epsilon = 0, 0.05, 0.1, 0.2$. We took $M = 20$ equally spaced grid points, although even $M = \sqrt{n}$ is probably sufficient, as already proposed in [Portnoy \(2003\)](#). In the basic algorithm, we took $N = 500$. In the updating algorithm we chose $N = 500$ and $N^* = 100$.

Results for $n = 100$ are summarized in Figure 1. At the left side of the figure, we compare L_1 -quantiles (thick lines) to depth quantiles (thin lines), for $\tau = 0.1, \dots, 0.8$. Solid lines correspond to the case $\epsilon = 0$, dotted lines to $\epsilon = 0.05$ and dashed lines to $\epsilon = 0.1$. Plot A_1 shows results in 2 dimensions, B_1 in 5 and C_1 in 10 dimensions. It is clear that L_1 -quantiles are superior when there is no contamination. The medians of squared errors are uniformly smaller. Especially for lower and higher quantiles the difference can be quite significant.

When contamination is added, the situation changes. With 5% of contamination, depth quantiles are more efficient from the 0.4-quantile on. At the 0.6-quantile, L_1 even breaks down, as it is completely attracted towards the outliers. In case of 10% of outliers, breakdown occurs already at the 0.4 quantile. Depth quantiles on the other hand, suffer very little from outliers. Their efficiencies remain about the same, no matter the value of $\epsilon \leq 0.1$, showing the robustness of our method.

L_1 vs. depth quantiles
(updating algorithm U)

Updating (U) algorithm vs.
basic (B) algorithm

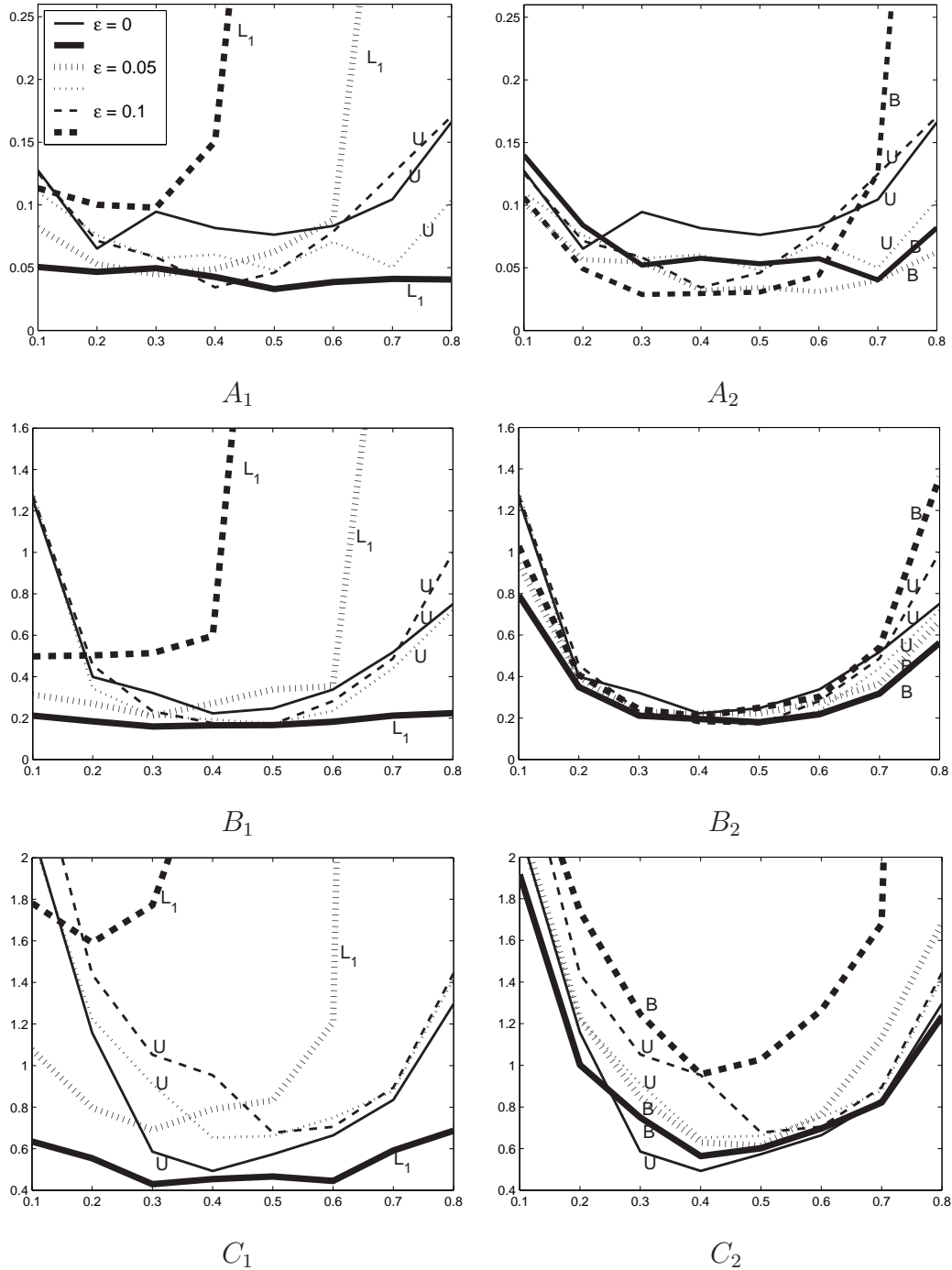


Fig. 1. Summary of simulation results. Left side plots: L_1 (thick) versus depth (thin lines). Right side: updating versus basic algorithm. Upper: $p = 2$, middle: $p = 5$, lower: $p = 10$. Solid lines: $\epsilon = 0$, dotted: $\epsilon = 0.05$, dashed: $\epsilon = 0.1$.

At the right side of Figure 1, we compare the basic algorithm (thick lines) with the updating algorithm (thin lines). Otherwise the setting is the same as previously, with solid/dotted/dashed lines corresponding to $\epsilon = 0/0.05/0.1$ and A_2, B_2, C_2 plots for $p = 2, p = 5$ and $p = 10$ respectively. The difference between both algorithms is not too big. In lower dimensions, the naive approach is slightly better. In higher dimensions, the updating algorithm sometimes even improves on the basic algorithm. In any event, the updating algorithm is certainly not much worse than the basic approach. Note however that it only took about half as much time. In the updating algorithm we constructed sets of 100 hyperplanes having $p - 1$ point in common. We also tried this with hyperplanes having $p - 2$ points in common. The results were however almost the same, whereas the computation time slightly increased. Therefore we propose to stick to the algorithm replacing only one point at a time.

The effective computation time of the algorithm of course highly depends on the choice of parameters. Setting the parameters as in our simulations and examples, i.e. $N = 500, N^* = 100$ and 20 grid points, the Matlab routine takes about 15 seconds on average, running Matlab 6.1 on a 2.4Ghz pc. For moderate-sized data sets decent results can thus be obtained in feasible time. However, the L_1 -quantiles are obviously much faster: a similar analysis took us 0.02 seconds with the *crq* implementation for R 2.4.1.

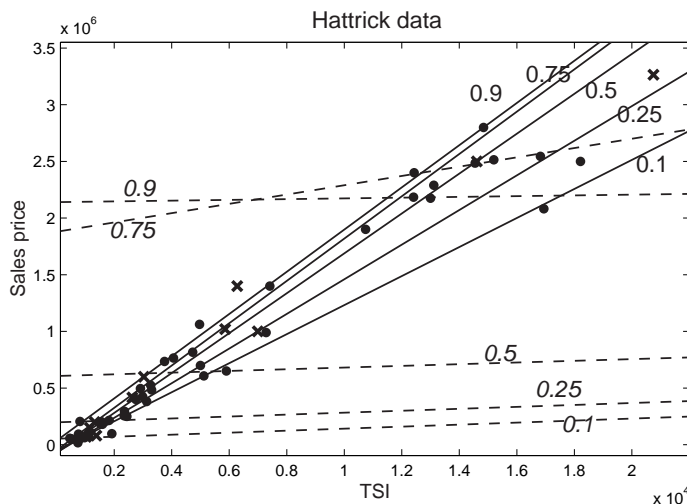


Fig. 2. Hatrick data. Censored observations are shown as stars, uncensored ones as dots. One data point at $(37.5 \times 10^4, 3,38 \times 10^6)$ (not visible on plot) destroys L_1 -quantiles (dashed lines), but is resisted by the depth quantiles (solid lines).

6 Examples

6.1 Hattrick data

Hattrick is a free online soccer game at www.hattrick.org, played by over half a million of people worldwide. Each participant owns one team, consisting of virtual soccer players, fans, money etc. Just like in real soccer, all teams are put into divisions and play weekly competition games. The winner can promote to a higher division. The ultimate goal is to become one of the top teams of your country.

An important and interesting aspect of Hattrick is its economy. Players can be sold and bought on the transfer market. The transfer system works as follows. A team can put a player on the transfer list with a starting price and a certain deadline. Other teams can make a bid; the highest bid at the deadline buys the player. All bids are publicly available to all users.

In our study, we followed 90 players that were on the transfer list on May 15, 2005. On May 17, 2005, we stopped our study. At that point, 55 players were sold, so we know their true sales price. For 14 players, at least one bid was already made, but the deadline was not yet reached. These are censored observations: we do not know the true sales price, but we do know it will be at least the current highest bid. The remaining 21 players did not receive a bid higher than their starting price, and thus were of no use in our study. This way, we obtained a data set of 69 observations, of which 14 are censored. As a covariate, we took the Total Skill Index (TSI) of each player, an in-game statistic measuring the quality of a player.

Our data contains one outlier: a player with a TSI of 374 600. Note that its sales price is relatively low: 3 381 000 euros. This is explained by the players' wage, which is closely related to their TSI. Moderate players with TSI around 5000 earn 4000 euro a week, better players with TSI around 20000 make about 15000. Since a team in the game typically has a budget of a few million euros, this difference is negligible. Our outlying player on the other hand has a wage of 299 496 euro a week. Although this player makes your team perform much better on the (virtual) pitch, his high wage becomes disadvantageous from an economical viewpoint. Therefore, the linear structure between TSI and market value is violated for these extremely good players.

The data is plotted in Figure 2. The outlier is not visible for aesthetic purposes, but was taken into account in our analysis. The dashed lines are the 0.1, 0.25, 0.5, 0.75 and 0.9 regression quantiles, estimated by *crq*, the method from [Portnoy \(2003\)](#) using the L_1 -quantiles. As one can see, the outlier has a huge effect. The estimates clearly make no sense.

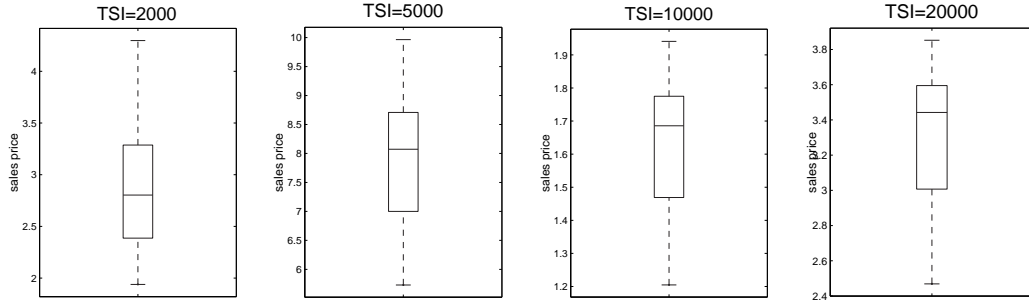


Fig. 3. Boxplots of conditional distribution of sales price given TSI= 2000, 5000, 10000 and 20000.

The depth quantiles are plotted as solid lines. They provide quite a nice view of the linear and heteroscedastic nature of the data. The effect of the outlier is minimal, showing the robustness of our approach. Also note that the outlier is not outlying in y direction. Furthermore, its L_1 -residual is not outlying compared to the other residuals, since the L_1 -quantiles are completely tilted towards the outlier. Thus our extremely good player can only be detected in \mathbf{x} -space. In simple regression as in this example, this is of course very easy. However, in a higher dimensional \mathbf{x} -space, outlier detection might be far from trivial. In that case, depth quantiles provide a way out.

An interesting feature of quantile regression is that one can visualize the entire conditional distribution at a certain point in x -space. This is shown in Figure 3 by means of boxplots at four different values of TSI: 2000, 5000, 10000 and 20000. Note the evolution from right skewed to left skewed as the value of TSI increases. This might reflect that people are more careful when dealing with more money. When a team with a budget of a few million euros wants to buy a player worth around 30 000, he might easily pay 50 000. When buying one worth around 3 000 000, he will probably make more effort to search for a relative good buy.

As a final note we should mention that the proposed model is probably not very optimal. Due to the rightskewness of both variables, a log transform would be a logical option. In that case the effect of the outlier is reduced, making the difference between L_1 -quantiles and depth quantiles rather small.

6.2 Granule data

Our second example concerns a process in pharmacy called fluidized bed granulation. The data were taken from [Rambali et al. \(2003\)](#). In an experimental design, they studied the granulation process on a fluidized bed in a semi-full scale (30 kg batch). There are 30 observations of which 8 are censored, because the process conditions were too bad to determine the granule size correctly.

In an empirical model they consider four variables: airflow rate, inlet air temperature, scaled spray rate and inlet air humidity. Their final proposal yields a 9 dimensional model including these four standardized variables and some interaction terms as independent variables. The response variable of interest is the observed granule size. [Rambali et al. \(2003\)](#) estimated the conditional median in two steps. First they obtained estimates for the value of the response variable for the censored observations. Then they used ordinary deepest regression on this completed data set.

We will now compare these results to the ones obtained by our algorithm. The right column of Table 1 shows the original results from [Rambali et al. \(2003\)](#). The middle column shows the results from our censored depth quantiles. Note that the results are pretty similar. We also performed ordinary deepest regression on the data set without the censored observations (see the first column in Table 1). Some estimates are completely different, eg. the coefficients of A^2 , S and AS . This shows that deleting censored observations can lead to severely biased estimates. It is absolutely necessary to use specific methods dealing with censored observations.

	Censoring ignored	Censored Depth	Rambali et al.
intercept	537.0	520.8	536.2
airflow rate (A)	-221.4	-309.3	-326.1
inlet air temperature (T)	-231.3	-166.5	-184.6
spray rate (S)	134.3	215.3	226.5
inlet air humidity (H)	35.9	28.4	30.6
A^2	52.6	197.6	164.4
T^2	172.3	123.8	145.4
AT	135.6	118.5	123.3
AS	4.5	-118.4	-110.7

Table 1
Granule data: parameter estimates of the conditional median.

7 Conclusion

We extended the idea of regression depth quantiles to data sets where censored observations are present. A grid algorithm was used and we introduced a relatively fast way of optimizing the objective function over this grid. A simulation study showed that this updating algorithm is particularly useful

in higher dimensions, when similar efficiency as the naive approach is reached twice as fast. The simulation study revealed a loss in efficiency compared to the L_1 -quantiles for normally distributed data, especially at lower and higher values of τ . When contamination was added on the other hand, depth quantiles performed better. They appear to have excellent robustness properties whereas L_1 -quantiles break down.

References

- Adrover, J., Maronna, R., Yohai, V., 2002. Relationships between maximum depth and projection regression estimates. *J. Stat. Plan. Infer.* 105, 363–375.
- 2004. Robust regression quantiles. *J. Stat. Plan. Infer.* 122, 187–202.
- Bai, Z., He, X., 2000. Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Ann. Stat.* 27, 1616–1637.
- Honoré, B., Khan, S., Powell, J., 2002. Quantile regression under random censoring. *J. Econometrics*, 109, 67–105.
- Koenker, R., 2005. *Quantile Regression*. Cambridge University Press.
- Koenker, R., Bassett, G., 1978. Regression quantiles. *Econometrica*, 46, 33–50.
- Neocleous, T., Vanden Branden, K., Portnoy, S., 2006. Correction to censored regression quantiles. *J. Am. Stat. Assoc.* 101, 860–861.
- Park, J., Hwang, J., 2003. Regression depth with censored and truncated data. *Commun. Stat - Theory M.* 32, 997–1008.
- Portnoy, S., 2003. Censored regression quantiles. *J. Am. Stat. Assoc.* 98, 1001–1012.
- Rambali, B., Van Aelst, S., Baert, L., Massart, D., 2003. Using deepest regression method for optimization of fluidized bed granulation on semi-full scale. *Int. J. Pharm.* 258, 85–94.
- Rousseeuw, P.J., Hubert, M., 1999. Regression depth. *J. Am. Stat. Assoc.* 94, 388–402.
- Van Aelst, S., Rousseeuw, P.J., 2000. Robustness of deepest regression. *J. Multivariate Anal.* 73, 82–106.
- Van Aelst, S., Rousseeuw, P.J., Hubert, M., Struyf, A., 2002. The deepest regression method. *J. Multivariate Anal.* 81, 138–166.
- Verboven, S., Hubert, M., 2005. LIBRA: a Matlab Library for Robust Analysis. *Chemometr. Intel. Lab.* 75, 127–136.