# Uniform Bahadur Representation for Nonparametric Censored Quantile Regression: A Redistribution-of-Mass Approach

Efang Kong
*University of Kent at Canterbury, UK*

Yingcun Xia
*National University of Singapore, Singapore*
and *University of Electronic Science and Technology, China*

The authors thank a Co-Editor, an Associate Editor and two referees for their thoughtful comments, and Dr Patrick Saart for his suggestions. Xia's research is partially supported by National Natural Science Foundation of China (71371095) and a research grant from the Ministry of Education, Singapore (MOE 2014-T2-1-072). Address correspondence to Efang Kong, School of Mathematics, Statistics and Actuarial Science, University of Kent at Canterbury, UK; e-mail: e.kong@kent.ac.uk.

**Abstract.** Censored quantile regressions have received a great deal of attention in the literature. In a linear setup, recent research has found that an estimator based on the idea of "redistribution-of-mass" (Efron, 1967) has better numerical performance than other available methods. In this paper, this idea is combined with the local polynomial kernel smoothing for nonparametric quantile regression of censored data. We derive the uniform Bahadur representation for the estimator and, more importantly, give theoretical justification for its improved efficiency over existing estimation methods. We include an example to illustrate the usefulness of such a uniform representation in the context of sufficient dimension reduction in regression analysis. Finally, simulations are used to investigate the finite sample performance of the new estimator.

*Key words and phrases.* Bahadur representation; empirical processes; generalized Kaplan-Meier estimator; local polynomial smoothing; quantile regression; random censoring; semiparametric models; stochastic equicontinuity; dimension reduction.

# 1 Introduction

Let $\mathbf{X}$ be a $p \times 1$ vector of covariates and $T$ be a dependent (scalar) variable of interest, such as a logarithmic or another monotonic transformation of survival time. In survival analysis, quite often the value of $T$ is not directly observable due to censoring. What is observed, instead, is the triple $\xi = (Y, \mathbf{X}, d)$ such that

$$Y = \min(T, C), \quad d = I(T \leq C), \tag{1}$$

where $I(.)$ is the indicator function and $C$ is the censoring variable, assumed to be conditionally independent of $T$ given $\mathbf{X}$. [1]

In investigating the functional relationship between $T$ and $\mathbf{X}$, many classical methods of analysis are rendered obsolete, as censoring can cause a breakdown in the normality assumptions or moment restrictions which underlie the construction of these methods. Alternatively, modelling could be based on the conditional hazard function of $T$ given $\mathbf{X}$, such as the semi-parametric Cox proportional hazard model which requires unconditional independence between $T$ and $C$, and the accelerated failure time models which are predominately fully parametric.

Since quantile regression (Koenker and Bassett, 1978; Koenker and Geling, 2001) is more robust to censoring, it has become a very popular choice in the analysis of censored data (Powell, 1984; Powell, 1986). In this case, the function of interest is the conditional $\tau$th quantile of $T$ given $\mathbf{X}$, defined as

$$Q_\tau(\mathbf{X}) = \min\{t : F_0(t|\mathbf{X}) \geq \tau\}, \tag{2}$$

where $F_0(.|.)$ denotes the conditional distribution function of $T$ given $\mathbf{X}$. Collectively, (1) and (2) specify the so-called censored quantile regression (CQR) model. Note that an equivalent but more commonly used form of (2) can be defined as follows

$$Q_\tau(\mathbf{X}) = \arg_a \min E[\rho_\tau(T - a)|\mathbf{X}], \tag{3}$$

where $\rho_\tau(u) = u[\tau - I(u < 0)]$.

A large body of research has been devoted to studying various aspects of the CQR; see, for example, Buchinsky and Hahn (1998), Chernozhukov and Hong (2002), Portnoy (2003), De Gooijer and Zerom (2003), Peng and Huang (2008), Zhou and Wu (2009), El Ghouch and Van Keilegom (2009) and Wang and Wang (2009). While most of these studies assume that the quantile function belongs to a fixed finite-dimensional space of functions, less work has been done for nonparametric censored regression models. Below, let us list a few well-known examples. Dabrowska (1987, 1992) and Van Keilegom and Veraverbeke (1998) proposed nonparametric censored regression estimators based on quantile methods. Lewbel and Linton (2002) considered the case where the censoring time is a degenerate random variable (i.e., it is constant), while Chen, Dahl and Khan (2005) allowed for heteroscedasticity. Heuchenne and Van Keilegom (2007, 2008) considered a nonparametric regression model, where the error term is independent of the covariates. Linton, Mammen, Nielsen and Van Keilegom (2011) studied univariate regression models with a variety of censoring schemes, while employing estimation methods based on hazard functions. The most noteworthy is the work by Kong, Linton and Xia (2013), who investigated the same estimation problem as will be discussed in this paper but with a critical shortfall in the sense that their estimation method suffered from poor efficiency.

The core issue in estimating the CQR model is how to handle the censored observations. While a naive approach (either these observations are simply thrown away or $Y$ is used as a substitute for $T$) is certainly inconsistent, a number of consistent methods have been proposed for linear CQR. Under the assumption that $C$ is independent of both $T$ and $\mathbf{X}$ (a relatively strong assumption), Honoré, Khan and Powell (2002) suggested replacing $\rho_\tau(Y - \boldsymbol{\beta}^\top \mathbf{X})$ with its conditional expectation given $\xi = (Y, \mathbf{X}, d)$. On the contrary, Lindgren (1997) and Bang and Tsiatis (2002) considered the use of $w(\xi) = d/[1 - G(Y|\mathbf{X})]$, where $G(.|\mathbf{X})$ is the conditional distribution function of $C$ given $\mathbf{X}$. Such a weighting scheme, which is referred to hereafter as the LBT weight, is derived from the fact that $E(d/[1 - G(Y|\mathbf{X})]) = 1$ and can lead to loss of information as only those uncensored observations with $d = 1$ are used in the estimation.

A heuristically more efficient weighting scheme introduced by Portnoy (2003), and Wang et al. (2009) was based on the idea of 'redistribution-of-mass' (RDW) of Efron (1967). This involves the redistribution to the right of the probability $1 - F_0(C|\mathbf{X}) = P(T > C|\mathbf{X})$ for each censored observation. Specifically, for quantile regression at level $\tau$, the RDW weight assigned to $\xi = (Y, \mathbf{X}, d)$ is

$$w(\tau|\xi) = \begin{cases} 1, & d = 1 \text{ or } F_0(C|\mathbf{X}) > \tau; \\ \frac{\tau - F_0(C|\mathbf{X})}{1 - F_0(C|\mathbf{X})}, & d = 0 \text{ and } F_0(C|\mathbf{X}) < \tau. \end{cases} \tag{4}$$

However, since $F_0(.|.)$ is usually unknown, a substitute, for example the generalized Kaplan-Meier (K-M) estimator, is often used. Wang et al. (2009) gave a comprehensive review of this weighting scheme for the estimation of linear CQR and demonstrated that in general the RDW weighting scheme should be recommended due to its superior numerical performance compared to its competitors.

This paper studies estimation of the nonparametric censored quantile regression (NCQR), where the function of interest is $Q_\tau(\mathbf{x})$ and (or) its partial derivatives, which are evaluated at any $\tau \in (0,1)$ and $\mathbf{x}$ in the support of $\mathbf{X}$. We propose an estimation procedure that combines local polynomial smoothing with the RDW weighting scheme. Furthermore, we derive a global Bahadur representation of the resulting estimator uniformly over both the value of the covariate $\mathbf{x}$ and the quantile level $\tau$.

The Bahadur (1966) representation is a useful tool to study the asymptotic properties of estimators when the loss function is not smooth as in M-estimation and quantile regression, for example. The Bahadur representation approximates an estimator by a sum of independent variables with a higher order remainder (He and Shao, 1996) so that the asymptotic properties of the estimator can be easily derived. In the literature, a number of different Bahadur representations have been obtained under various different settings. Carroll (1978) and Martinsek (1989) obtained strong representations for location and regression M-estimators with preliminary scale estimates, while Babu (1989) and Pollard (1991) obtained the Bahadur representation for the least absolute deviation regression. In addition, Portnoy (1997) obtained the Bahadur representation for quantile smoothing splines, while Portnoy (2003) did so for the censored quantile of the Cox model. Chaudhuri (1991) investigated the pointwise Bahadur representation of nonparametric kernel quantile regression, while Wu (2005) examined the representation for nonstationary time series data under both parametric and nonparametric settings; see also Zhou et al. (2009).

In nonparametric settings, 'global' (or equivalently 'uniform') asymptotic theory (Bickel and Rosenblatt, 1973; Mack and Silverman, 1982) is essential for conducting statistical inference. Hence, uniform Bahadur representations are often more useful than their pointwise counterparts. In the current paper, note that the term 'uniform' (or 'global') refers interchangeably to the order of the remainder term that is uniform over a collection of $\tau$ and $\mathbf{x}$. In such a sense, Kong, Linton and Xia (2010) and Guerre and Sabbah (2011) obtained the uniform Bahadur representation for the quantile local polynomial estimators. A more comprehensive survey of these kinds of Bahadur representations in the context of nonparametric quantile regression with complete data can be found in Kong et al. (2013); see also Kong et al. (2010) and the references therein.

In this paper, the RDW weighting scheme will be used in the derivation of the the global Bahadur representation instead of the LBT weighting scheme as in Kong et al. (2013). This seemingly small difference does have far-reaching implications as detailed below.

- In contrast to the LBT weight, which uses only uncensored data, the RDW weighting scheme uses all available observations. In addition, the weight assigned to each observation depends on the amount of information it contains. As a result, even a censored observation could be assigned a full weight of one provided that it is deemed to be sufficiently 'informative' at the given level $\tau$, i.e. $F_0(C_i|\mathbf{X}_i) > \tau$.

- The LBT weight of Kong et al. (2013) is susceptible to heavy influence from big values of $T$ for which $G(T|\mathbf{X})$ is close to one. This problem will be further exacerbated when $G(.|.)$ is unknown and thus replaced by an estimator, for example the well known K-M estimator, which in general suffers from weak performance towards the right tail of the distribution due to the small number of observations being available in that area.

- The weights for uncensored observations are always one for the RDW scheme, while they always need to be estimated with the LBT estimator.

- The LBT weight is smooth with respect to $G(.|.)$, which makes its asymptotic study relatively straightforward from a theoretical point of view. For example, to evaluate the error incurred in the weight by replacing $G(.|.)$ with a preliminary estimate $\hat{G}(.|.)$:

$$\frac{1}{1 - G(Y_i|\mathbf{X}_i)} - \frac{1}{1 - \hat{G}(Y_i|\mathbf{X}_i)} = \frac{G(Y_i|\mathbf{X}_i) - \hat{G}(Y_i|\mathbf{X}_i)}{[1 - G(Y_i|\mathbf{X}_i)][1 - \hat{G}(Y_i|\mathbf{X}_i)]}, \tag{5}$$

  all we need is the asymptotic expression for $G(Y_i|\mathbf{X}_i) - \hat{G}(Y_i|\mathbf{X}_i)$; evaluation of the congregated error, i.e. the summation of terms like (5) over $i = 1, \cdots, n$, then readily follows from an application of standard results of U-processes. However, the above technique is not applicable to the RDW weight because the weight itself is not smooth. We tackle this problem by employing key techniques of empirical process, such as the stochastic equi-continuity properties of non-smooth function of preliminary (nonparametric) estimators.

To summarize, the theoretical study of the RDW estimator is more challenging compared to that of Kong et al. (2013). This is clearly manifested in the proof of Proposition 5 in the supplement, which could be compared with that of Lemma A.2 in Kong et al. (2013). Nevertheless, the additional complexity is justified since it helps confirm our conjecture that the RDW weight does generally lead to better estimation efficiency; see Corollary 1 in Section 4 in particular. Another interesting theoretical discovery of this paper is that, in a nonparametric setting, the effect of substituting a generalized K-M estimator for the unknown $F_0(.|.)$

can take different forms depending on the convergence rate of the K-M estimator. A phenomenon that does not exist in a parametric setup, and in the case of a global polynomial model, however, our results do coincide with those in a parametric setup considered in Wang et al. (2009).

The remainder of this paper is organized as follows. Section 2 describes the proposed RDW-based estimation method. Section 3 contains a list of regularity conditions and also some notations. We provide in Section 4 the main theoretical results and demonstrate in Section 5 their applications in the research area of dimension reduction. A small scale numerical study is given in Section 6, in which the numerical performance of several competing methods is examined, including the one by Kong et al. (2013). Additional lemmas, propositions and all the proofs are provided online at Cambridge Journals Online (journals.cambridge.org/ect) in supplementary material to this article.

## 2 The estimation method

Suppose we have i.i.d. copies $\xi_i = (Y_i, \mathbf{X}_i, d_i)$, $i = 1, \cdots, n$, of $\xi = (Y, \mathbf{X}, d)$ generated according to (1) and (2). Our interest is in the the estimation of $Q_\tau(.)$ and its partial derivatives. For any given $\tau \in (0,1)$ and $\mathbf{x}$ in the compact support $\mathcal{D}$ of $\mathbf{X}$, the local polynomial smoothing estimator of $Q_\tau(\mathbf{x})$ is based on the assumption that there exists some positive integer $k$, such that in a neighbourhood of $\mathbf{x}$, $Q_\tau(.)$ is smooth enough to have partial derivatives up to order $k$. Consequently, for $\mathbf{X}$ close to $\mathbf{x}$, $Q_\tau(\mathbf{X})$ can be approximated by its $k-$th order Taylor series expansion:

$$Q_{n\tau}(\mathbf{X}) \overset{def}{=} Q_\tau(\mathbf{x}) + \sum_{1\le[\mathbf{r}]\le k} \frac{D^{\mathbf{r}} Q_\tau(\mathbf{x})}{\mathbf{r}!}(\mathbf{X}-\mathbf{x})^{\mathbf{r}}, \tag{6}$$

where $\mathbf{r} = (r_1, \cdots, r_p)$ denotes an arbitrary $p-$dimensional vector of nonnegative integers, $D^{\mathbf{r}}$ denotes the differential operator, $\partial^{[\mathbf{r}]}/\partial x_1^{r_1} \cdots \partial x_p^{r_p}$, $[\mathbf{r}] = \sum_{i=1}^p r_i$, $\mathbf{r}! = \prod_{i=1}^p r_i!$, $\mathbf{x}^{\mathbf{r}} = \prod_{i=1}^p x_i^{r_i}$ with the convention that $0^0 = 1$. Let $A = \{\mathbf{r} : [\mathbf{r}] \le k\}$ and $n(A) = \sharp(A)$, its cardinality. In the absence of censoring, the local polynomial estimates of the vector $[D^{\mathbf{r}} Q_\tau(\mathbf{x}),\ \mathbf{r} \in A]$ are obtained via the minimization of the function below with respect to $\boldsymbol{\beta} = (\beta_{\mathbf{r}})_{\mathbf{r}\in A} \in R^{n(A)}$:

$$\sum_{i=1}^n K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \rho_\tau(T_i - \boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})), \tag{7}$$

where $\mathbf{X}_{i\mathbf{x}} = \mathbf{X}_i - \mathbf{x}$, $\boldsymbol{\mu}_n(\mathbf{x}) = [\delta_n^{-[\mathbf{r}]}\mathbf{x}^{\mathbf{r}},\ \mathbf{r} \in A]$, a $n(A) \times 1$ vector and $K_{\delta_n}(.) = K(./\delta_n)$, with $K(.)$ usually chosen to be some symmetric density function in $R^p$ and a smoothing parameter

$\delta_n \to 0$ as $n \to \infty$. For more details on this estimation method, please refer to Chaudhuri (1991), Chaudhuri, Doksum and Samarov (1997) and Kong et al. (2010).

We now describe how to combine this idea with the RDW weighting scheme for estimation with censored data. Let us start with the ideal situation where $F_0(.|.)$ is known. In this situation, we write

$$w_{i0}(\tau) \equiv w_i(\tau|F_0) = \begin{cases} 1, & d_i = 1 \text{ or } F_0(C_i|\mathbf{X}_i) > \tau; \\ \frac{\tau - F_0(C_i|\mathbf{X}_i)}{1 - F_0(C_i|\mathbf{X}_i)}, & d_i = 0 \text{ and } F_0(C_i|\mathbf{X}_i) < \tau. \end{cases}$$

For any fixed $\mathbf{x} \in R^p$, estimates of $[D^{\mathbf{r}} Q_\tau(\mathbf{x}),\ \mathbf{r} \in A]$ could then be obtained by minimizing the following weighted function with respect to $\boldsymbol{\beta} \in R^{n(A)}$:

$$\sum_{i=1}^{n} K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \Big[ w_{i0}(\tau) \rho_\tau(Y_i - \boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})) + (1 - w_{i0}(\tau)) \rho_\tau(Y^{+\infty} - \boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})) \Big], \tag{8}$$

where $Y^{+\infty}$ could be any value that is large enough to exceed all $\boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}}),\ i = 1, \cdots, n$. Since the sub-gradient of (8) only depends on the signs of the residuals, we may simply set $Y^{+\infty} = +\infty$; in practice one could choose $Y^{+\infty} = 100 \max\{Y_1, \cdots, Y_n\}$ as suggested in Wang et al. (2009). Since $0 < \tau < 1$, $\rho_\tau(s) \to \infty$ as $|s| \to \infty$, $\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x})$ always exists. Propositions 1 and 2 in the supplementary material discuss the implicit form of the solution(s) to problem (8), as well as the sufficient and necessary conditions for a unique solution.

With $F_0(.|.)$ being unknown, a preliminary nonparametric estimator $\hat{F}_n(.|.)$ can be used as a substitute for $F_0(.|.)$ in the definition of $w_{i0}(\tau)$. Let us write $w_i(\tau|\hat{F}_n)$ as $w_{in}(\tau)$. Consequently, for any $\mathbf{x} \in \mathcal{D}$, $\tau \in (0, 1)$, an estimate of the $n(A) \times 1$ vector

$$\boldsymbol{\beta}_{n\tau}(\mathbf{x}) \overset{def}{=} [\delta_n^{[\mathbf{r}]} D^{\mathbf{r}} Q_\tau(\mathbf{x}) / \mathbf{r}!,\ \mathbf{r} \in A]$$

is thus given by the minimum of

$$\sum_{i=1}^{n} K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \Big[ w_{in}(\tau) \rho_\tau(Y_i - \boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})) + (1 - w_{in}(\tau)) \rho_\tau(Y^{+\infty} - \boldsymbol{\beta}^\top \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})) \Big], \tag{9}$$

as a function of $\boldsymbol{\beta} \in R^{n(A)}$. Denote it as $\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x})$.

A commonly used estimator of $F_0(.|\mathbf{X}_i)$ is the generalized K-M estimator (Gonzalez-Manteigaa and Cadarso-Suarez, 1994) defined as

$$\hat{F}_{KM}(t|\mathbf{x}) = 1 - \prod_{j=1}^{n} \Big\{ 1 - \frac{B_{nj}(\mathbf{x})}{\sum_{k=1}^{n} I(Y_k \geq Y_j) B_{nk}(\mathbf{x})} \Big\}^{b_j(t)}, \tag{10}$$

where $b_j(t) = I(Y_j \leq t, d_j = 1)$, and $\{B_{nk}(\mathbf{x}),\ k = 1, \cdots, n\}$ is a sequence of non-negative weights adding up to 1. In Wang et al. (2009), the Nadaraya-Waston type of weights (local constant weights) were used, with a univariate $\mathbf{X}$ in mind. For a multivariate $\mathbf{X}$, the so-called 'curse of dimensionality' calls for a greater smoothing parameter. Therefore, to keep the approximation bias under control we need to engage 'higher order' weights, such as the local polynomial 'equivalent kernel/weight' (Fan and Gijbels, 1996), possibly derived from a different kernel density function $\tilde{K}(.)$ in $R^p$. For some positive integer $\kappa_1$ that depends on the smoothness of $F_0(t|\mathbf{x})$ as a function of $\mathbf{x}$, let $\tilde{A} = \{\mathbf{r} : [\mathbf{r}] \leq \kappa_1\}$ , $h_n \to 0$ $(n \to \infty)$, which is another smoothing parameter. For any $\mathbf{x} \in R^p$, write $\tilde{\boldsymbol{\mu}}_n(\mathbf{x}) = [h_n^{-[\mathbf{r}]}\mathbf{x}^{\mathbf{r}},\ \mathbf{r} \in \tilde{A}]$, and

$$B_{nk}(\mathbf{x}) = h_n^{-p}\underline{e}_1{}^\top \mathbf{M}_n^{-1}(\mathbf{x})\tilde{\boldsymbol{\mu}}_n(\mathbf{X}_{k\mathbf{x}})\tilde{K}_{h_n}(\mathbf{X}_{k\mathbf{x}}), \quad k = 1, \cdots, n, amy \tag{11}$$

where $\underline{e}_1$ is the $n(\tilde{A}) \times 1$ vector $(1, 0, \cdots, 0)^\top$, and

$$\mathbf{M}_n(\mathbf{x}) = \frac{1}{nh_n^p}\sum_{k=1}^{n} \tilde{K}_{h_n}(\mathbf{X}_{k\mathbf{x}})\tilde{\boldsymbol{\mu}}_n(\mathbf{X}_{k\mathbf{x}})\{\tilde{\boldsymbol{\mu}}_n(\mathbf{X}_{k\mathbf{x}})\}^\top.$$

Note that as a distribution function, $\hat{F}_{KM}(t|\mathbf{x})$ is not smooth; in fact it has jumps with a magnitude of $B_{ni}(\mathbf{x})/\{1 - \hat{G}_n(Y_i|\mathbf{x})\}$ at points $t = Y_i$ if $d_i = 1$, where $\hat{G}_n(.|.)$ is the generalized K-M estimator of $G(.|.)$; see also Gonzalez-Manteigaa et al. (1994). This could cause problems since, it requires among other things that the preliminary estimator should be smooth enough so that one could apply empirical process results regarding the stochastic equicontinuity properties of the non-smooth objective function of preliminary (nonparametric) estimators. We therefore suggest a smoothing procedure to be carried out on $\hat{F}_{KM}(.|\mathbf{x})$ to obtain the following smoothed generalized K-M estimator:

$$\hat{F}_n^S(t|\mathbf{x}) = \int \hat{F}_{KM}(s|\mathbf{x})\bar{K}((s-t)/h_{1n})/h_{1n}ds, \tag{12}$$

where $\bar{K}(.)$ is a univariate symmetric kernel function and $h_{1n}$ is a smoothing parameter. Since $\hat{F}_{KM}(s|\mathbf{x})$ is absolutely integrable, it is a standard result in real analysis that if $\bar{K}(.)$ infinitely differentiable, such the normal kernel, then so is $\hat{F}_n^S(t|\mathbf{x})$ with respect to $t$.[2] Therefore, from this point on, $\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x})$ will stand for the minimum of (9), where the weight function $w_{in}(\tau)$ is derived from this smoothed generalized K-M estimator $\hat{F}_n^S(.|\mathbf{x})$. Nevertheless, we would like to point out that this smoothing of the K-M estimator is purely for technical purposes; our experience suggests that numerically, it makes little difference, if any, and it is not sensitive to the choice of the smoothing parameter $h_{1n}$.

# 3 Notations and assumptions

Let $q$ be a positive integer and let $\chi$ be a bounded set in $R^q$. For some constants $M > 0$ and $s > 0$, we will say that a function $m(.) : R^q \to R$ belongs to $\mathcal{C}_M^s(\chi)$, if

$$\max_{[\mathbf{r}]\leq\lfloor s\rfloor} \sup_{\mathbf{z}\in\chi} |D^{\mathbf{r}} m(\mathbf{z})| + \max_{[\mathbf{r}]=\lfloor s\rfloor} \sup_{\mathbf{z},\mathbf{z}'\in\chi} \frac{|D^{\mathbf{r}} m(\mathbf{z}) - D^{\mathbf{r}} m(\mathbf{z}')|}{\|\mathbf{z}-\mathbf{z}'\|^{s-\lfloor s\rfloor}} \leq M,$$

where $\mathbf{r} = (r_1, \cdots, r_q)$ is a vector of non-negative integers, $\lfloor s\rfloor$ denotes the greatest integer that is (strictly) smaller than $s$, and $\|.\|$ stands for the Euclidean norm of a vector. $\mathcal{C}_M^s(\chi)$ is essentially the class of functions on a bounded set $\chi$ in $R^q$, with uniformly bounded partial derivatives up to order $\lfloor s\rfloor$, whose highest partial derivatives are Lipschitz of order $s - \lfloor s\rfloor$. Such a class of functions is of particular interest in empirical processes: the $\epsilon-$covering numbers with respect to the supremum norm grow sufficiently slowly for the empirical processes indexed by this class of functions to possess some 'uniform' asymptotic properties (van der Vaart and Wellner, 1996). More details on this can be found in the proof of Proposition 5 in the supplementary material.

For any $\mathbf{t} \in [-1, 1]^p$, let denote by $\boldsymbol{\mu}(\mathbf{t})$, the $n(A) \times 1$ vector $[\mathbf{t}^{\mathbf{r}} : \mathbf{r} \in A]$. We can therefore define the following

$$\Sigma(A) = \int_{[-1,1]^p} K(\mathbf{t})\boldsymbol{\mu}(\mathbf{t})\boldsymbol{\mu}(\mathbf{t})^\top d\mathbf{t}, \quad \boldsymbol{\gamma}(A) = \int_{\mathbf{t}\in[-1,1]^p} \mu(\mathbf{t}) d\mathbf{t};$$

$\Sigma(\tilde{A})$ is similarly defined. Note that the vectors $A$ and $\tilde{A}$ have been defined in paragraphs preceding equations (7) and (11), respectively. It is assumed throughout this paper that both $\Sigma(A)$ and $\Sigma(\tilde{A})$ are invertible. Let $f_{\mathbf{X}}(.)$ be the probability density function of $\mathbf{X}$. For any $\mathbf{x} \in R^p$, let $f_0(.|\mathbf{x})$ and $g(.|\mathbf{x})$ denote the probability density functions of $T$ and $C$, respectively, conditional on $\mathbf{X} = \mathbf{x}$. We make the following assumptions, in which the constants $s_k,\ k = 1, \cdots, 4$, are to be specified later, and $M$ stands for a generic positive constant whose value may vary from assumption to assumption.

[A1] $\mathbf{X}$ has a compact support $\mathcal{D} \subset R^p$ and $f_{\mathbf{X}}(.) \in \mathcal{C}_M^{s_1}(\mathcal{D})$, for some $s_1 > 0$.

[A2] There exist some $s_2 > 0$ and $0 < \tau^* < 1/2$, such that the quantile function $Q_\tau(.) \in \mathcal{C}_M^{s_2}(\mathcal{D})$, for all $\tau \in [\tau^*, 1-\tau^*]$.

[A3] The kernel functions $K(.)$ is a symmetric probability density function on $R^p$ with finite second moments and bounded first order derivatives.

[A4] For the censoring variable $C$, the conditional distribution function $G(Q_\tau(\mathbf{x})|\mathbf{x})$ is uniformly bounded away from one for all $\mathbf{x} \in \mathcal{D}$ and $\tau \in [\tau^*, 1-\tau^*]$; $G(t|\mathbf{x})$ is also Holder-continuous over $\{(\mathbf{x}, t) : \mathbf{x} \in \mathcal{D}, Q_{\tau^*}(\mathbf{x}) < t < Q_{1-\tau^*}(\mathbf{x})\}$.

[A5] There exists some $s_3 > 0$, such that $F_0(t|\mathbf{x})$, as a function of $\mathbf{x}$, belongs to $\mathcal{C}_M^{s_3}(\mathcal{D})$, uniformly in $t$ within the region $\{(\mathbf{x}, t) : \mathbf{x} \in \mathcal{D}, Q_{\tau^*}(\mathbf{x}) < t < Q_{1-\tau^*}(\mathbf{x})\}$; $f_0(t|\mathbf{x})$ is Holder continuous in $t \in [Q_{\tau^*}(\mathbf{x}), Q_{1-\tau^*}(\mathbf{x})]$ uniformly for all $\mathbf{x} \in \mathcal{D}$, and $f_0(Q_\tau(\mathbf{x})|\mathbf{x})$ is uniformly bounded away from zero for all $\mathbf{x} \in D$ and $\tau \in [\tau^*, 1-\tau^*]$.

[A6] $\tilde{K}(.)$ is a symmetric probability density function on $R^p$ with finite second moments and bounded first order derivative; the kernel function $\bar{K}(.) \in \mathcal{C}_M^{s_4}([0,1])$ for some $s_4 > 0$.

[A7] The smoothing parameters $h_n,\ h_{1n}$ are chosen such that $h_n \to 0,\ h_{1n} \to 0$, $nh_n^p/\log n \to \infty$, $nh_n^{p+4s_3/3}/\log n < \infty$ and $h_{1n}^2/h_n^{s_3} < \infty$.

[A1]-[A3] are standard assumptions in nonparametric polynomial smoothing. Specifically, under [A2] with $k = \lfloor s_2 \rfloor$ and small enough $\delta_n$, for any given $\mathbf{x} \in D$ and $\mathbf{t} \in [-1,1]^p$, the difference between $Q_\tau(\mathbf{x} + \mathbf{t}\delta_n)$ and the corresponding $k$th order Taylor series expansion around $\mathbf{x}$, we have

$$Q_{n\tau}(\mathbf{x} + \mathbf{t}\delta_n, \mathbf{x}) = \beta_{n\tau}^\top(\mathbf{x})\boldsymbol{\mu}(\mathbf{t}), \tag{13}$$

and

$$r(\mathbf{t}\delta_n, \mathbf{x}) \stackrel{def}{=} Q_\tau(\mathbf{x} + \mathbf{t}\delta_n) - Q_{n\tau}(\mathbf{x} + \mathbf{t}\delta_n, \mathbf{x}) = O(\delta_n^{s_2}), \tag{14}$$

uniformly in $\mathbf{t} \in [-1,1]^p$ and $\mathbf{x} \in D$. [A4] and [A5] together ensure the identifiability of the model; this is a nonparametric extension of condition A3 in Wang et al. (2009), where they studied the estimation of parameter $\boldsymbol{\beta}_\tau$ in linear CQR at a prescribed level $\tau$. Their condition A3 is such that for $\boldsymbol{\beta}$ in the neighbourhood of $\boldsymbol{\beta}_\tau$, $E[\mathbf{X}\mathbf{X}^\top f_0(\mathbf{X}^\top\boldsymbol{\beta}|\mathbf{X})\{1 - G(\mathbf{X}^\top\boldsymbol{\beta}|\mathbf{X})\}]$ must be positive definite. A sufficient but not necessary condition for this to be true is that both $f_0(\mathbf{X}^\top\boldsymbol{\beta}|\mathbf{X})$ and $1 - G(\mathbf{X}^\top\boldsymbol{\beta}|\mathbf{X})$ are uniformly bounded away from zero for $\mathbf{X}$ in some subset of its support $\mathcal{D}$ with positive measure.

[A6] is imposed so that $\hat{F}_n^S(.|.) \in \mathcal{C}_M^{s_4}((0,1))$, a condition as discussed in Section 1 required to ease the study of the asymptotics of the estimators obtained from minimizing (9), where the weights are derived from this $\hat{F}_n^S(.|.)$. [A5] and [A7] are necessary for the asymptotic representation of the generalized K-M estimator; see Lemma 1 below for more details.

# 4 Convergence rate and asymptotic representation

We first present some results concerning the smoothed generalized K-M estimator $\hat{F}_n^S(.)$ of (12). Let $\tau_n = (n\delta_n^p/\log n)^{-1/2}$ and $\tilde{\tau}_n = (nh_n^p/\log n)^{-1/2}$.

**Lemma 1.** *Under Assumptions [A5] - [A7] with $\kappa_1 = \lfloor s_3 \rfloor$, we have with probability one,*

$$\hat{F}_n^S(t|\mathbf{x}) - F(t|\mathbf{x}) = O(\tilde{\tau}_n), \tag{15}$$

$$\hat{F}_n^S(t|\mathbf{x}) - F(t|\mathbf{x}) = \tfrac{1}{n}\sum_{k=1}^{n} \tilde{B}_{h_n}(\mathbf{X}_k;\mathbf{x})\varphi(Y_k, d_k, t, \mathbf{x}) + O(\tilde{\tau}_n^{3/2}) \tag{16}$$

*uniformly in $\mathbf{x} \in \mathcal{D}$ and $t$, where for $k = 1, \cdots, n$,*

$$\begin{aligned}
\tilde{B}_{h_n}(\mathbf{X}_k;\mathbf{x}) &= h_n^{-p}\underline{e}_1{}^\top \Sigma^{-1}(\tilde{A})\tilde{\boldsymbol{\mu}}_n(\mathbf{X}_{k\mathbf{x}})\tilde{K}_{h_n}(\mathbf{X}_{k\mathbf{x}})/f_{\mathbf{X}}(\mathbf{x}), \\
\varphi(Y_k, d_k, t, \mathbf{x}) &= \{1 - F_0(t|\mathbf{x})\}\Big[\frac{I(Y_k \le t, d_k = 1)}{\{1 - F_0(Y_k|\mathbf{x})\}\{1 - G(Y_k|\mathbf{x})\}} \\
&\qquad - \int_0^{\min(Y_k,t)} \frac{f_0(s|\mathbf{x})ds}{\{1 - F_0(s|\mathbf{x})\}^2\{1 - G(s|\mathbf{x})\}}\Big].
\end{aligned}$$

Regarding the uniform Bahadur type representation for $\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x})$, we have

**Theorem 1.** *Suppose [A1]- [A7] hold with $s_j > 0$, $j = 1, 2, 3, 4$, $\alpha = (p+1)/s_4 < 1$, $k = \lfloor s_2 \rfloor$, and the smoothing parameters $\delta_n$, $h_n$ are chosen such that $\delta_n^{s_2}/\tau_n < \infty$ and $\tilde{\tau}_n^{1-\alpha}/(\delta_n^{2p\alpha}\log n) < \infty$. Then*

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x}) - \boldsymbol{\beta}_{n\tau}(\mathbf{x}) &= \frac{\Sigma_{n\tau}^{-1}(\mathbf{x})}{n\delta_n^p}\sum_i K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})\Big[w_{i0}I\{Y_i \le Q_{n\tau}(\mathbf{X}_i, \mathbf{x})\} - \tau\Big] \\
&\quad -(1-\tau)\frac{\Sigma_{n\tau}^{-1}(\mathbf{x})}{n\delta_n^p}\sum_{j=1}^{n} E_i\Big[\tilde{B}_{h_n}(\mathbf{X}_j;\mathbf{X}_i)K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})\Phi(\mathbf{X}_i, Y_j, d_j|\tau)\Big] \\
&\quad +R_n(\mathbf{x}|\tau),
\end{aligned} \tag{17}$$

*where*

$$R_n(\mathbf{x}|\tau) = O_p\Big(\tau_n^{3/2} + \tau_n(\delta_n^{p\alpha}\log n/\tilde{\tau}_n^{1-\alpha})^{-1/2}\Big),$$

*uniform in $\mathbf{x} \in \mathcal{D}$ and $\tau \in [\tau^*, 1-\tau^*]$,*

$$\begin{aligned}
\Sigma_{n\tau}(\mathbf{x}) &= E_i\Big[\{1 - G(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)\}f_0(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})^\top\Big], \\
\Phi(\mathbf{X}_i, Y_j, d_j|\tau) &= E\Big[\frac{I(C_i \le Q_\tau(\mathbf{X}_i))}{1 - F_0(C_i|\mathbf{X}_i)}\varphi(Y_j, d_j, C_i, \mathbf{X}_i)\Big|\mathbf{X}_i\Big] \\
&\qquad -\{\frac{g}{f_0}\}(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)\varphi(Y_j, d_j, Q_\tau(\mathbf{X}_i), \mathbf{X}_i),
\end{aligned}$$

*$E_i(.)$ stands for expectation taken with respect to $\mathbf{X}_i$, and the expectation in the definition of $\Phi(\mathbf{X}_i, Y_j, d_j|\tau)$ is with respect to the conditional distribution of $C_i$ given $\mathbf{X}_i$.*

The LBT estimator of Kong et al. (2013), denoted as $\hat{\mathbf{c}}_{n\tau}(\mathbf{x})$, is defined as the minimum of the following empirical objective function

$$\sum_{i=1}^{n} \frac{d_i}{1 - \hat{G}(Y_i|\mathbf{X}_i)}K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\rho_\tau(Y_i - \mathbf{c}^\top\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})), \tag{18}$$

with respect to $\mathbf{c} \in R^{n(A)}$, where $\hat{G}(.|.)$ is the generalized K-M estimator of $G(.|.)$. Its Bahadur representation as given in Theorem 4.2 of Kong et al. (2013) is such that

$$\begin{aligned}\hat{\mathbf{c}}_{n\tau}(\mathbf{x}) - \boldsymbol{\beta}_{n\tau}(\mathbf{x}) &= \frac{\tilde{\Sigma}_{n\tau}^{-1}(\mathbf{x})}{n\delta_n^p} \sum_i \frac{d_i}{1 - G(Y_i|\mathbf{X}_i)} K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}}) \Big[ I(Y_i \leq Q_{n\tau}(\mathbf{X}_i, \mathbf{x})) - \tau \Big] \\ &\quad - \frac{\Sigma_{n\tau}^{-1}(\mathbf{x})}{n\delta_n^p} \sum_j E_i \Big[ \tilde{B}_{h_n}(\mathbf{X}_j; \mathbf{X}_i) K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}}) T_n(\mathbf{x}, \zeta_i, \zeta_j) \Big] \\ &\quad + R_n(\mathbf{x}|\tau), \end{aligned} \tag{19}$$

where $E_i(.)$ stands for expectation taken with respect to $(\mathbf{X}_i, Y_i)$, and

$$\begin{aligned}\tilde{\Sigma}_{n\tau}(\mathbf{x}) &= E_i \Big[ f_0(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i) K_{\delta_n}(\mathbf{X}_{i\mathbf{x}}) \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}}) \boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})^\top \Big], \\ T_n(\mathbf{x}, \zeta_i, \zeta_j|\tau) &= [I(T_i \leq Q_{n\tau}(\mathbf{X}_i, \mathbf{x})) - \tau] \varphi(Y_j, 1 - d_j, Y_i, \mathbf{X}_i).\end{aligned}$$

Compare (19) with (17) and we could come to the following conclusions.

(A) In either representation, the first term, referred to as the 'staple' term, stands for the leading stochastic error of the estimator when $F_0(.|.)$ is known. The difference between the two 'staple' terms reflects the difference between the general principles based on which the respective target functions are defined and in particular, the way the two types of weights are formed.

(B) The second term in either representation is referred to as the 'correction' term. Since $E[\Phi(\mathbf{X}_i, Y_j, d_j|\tau)|\mathbf{X}_i] = 0$, it reflects the asymptotic error induced by a preliminary nonparametric (K-M) estimator being used as a substitute for the true but unknown conditional distribution function when the latter is unavailable. It converges at the same speed as the corresponding 'staple' term or the preliminary K-M estimator, whichever is the higher; see (A.34)–(A.36) in the supplementary material.

These two observations facilitate the study of the asymptotic variance ('efficiency') of these two competing estimators. Apparently, compared to the 'staple' term, the 'correction' term in either presentation has a more complicated form yet is comparatively 'negligible', provided that the preliminary K-M estimator converges fast enough. For illustration purposes, our focus will be on this special case where for both estimators the 'staple' term dominates over other terms in the same representation.

**Corollary 1.** Suppose conditions are all met for both (17) and (19) to hold true with

$\delta_n = o(h_n)$. Then we have

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{\beta}_{n\tau}(\mathbf{x})) &= \frac{\Sigma^{-1}(A)]\tilde{\Sigma}(A)\Sigma^{-1}(A)}{\delta_n^p f_0^2(Q_\tau(\mathbf{x})|\mathbf{x})}\{1-G(Q_\tau(\mathbf{x})|\mathbf{x})\}^{-2}\{\tau(1-\tau)[1-G(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)] \\
&\quad +(1-\tau)^2 E\Big[I(C_i \le Q_\tau(\mathbf{X}_i))\frac{F_0(C_i|\mathbf{X}_i)}{1-F_0(C_i|\mathbf{X}_i)}\Big]\}(1+o(1)) \\
&\le \frac{\Sigma^{-1}(A)]\tilde{\Sigma}(A)\Sigma^{-1}(A)}{\delta_n^p f_0^2(Q_\tau(\mathbf{x})|\mathbf{x})}\frac{\tau(1-\tau)}{\{1-G(Q_\tau(\mathbf{x})|\mathbf{x})\}^2}, \\
\mathrm{Cov}(\hat{\mathbf{c}}_{n\tau}(\mathbf{x})) &= \frac{\Sigma^{-1}(A)]\tilde{\Sigma}(A)\Sigma^{-1}(A)}{\delta_n^p f_0^2(Q_\tau(\mathbf{x})|\mathbf{x})} E\Big\{\frac{[\tau - I(T\le Q_\tau(\mathbf{x}))]^2}{1-G(T|\mathbf{x})}\Big\}(1+o(1)),
\end{aligned}
$$

where $\tilde{\Sigma}(A) = \int K^2(\mathbf{t})\boldsymbol{\mu}(\mathbf{t})\boldsymbol{\mu}(\mathbf{t})^\top d\mathbf{t}$.

It is immediately clear that the covariances of these two estimators are given by the same (positive definite) matrix but multiplied by different constants: in the case of the RDW estimator, this constant is bounded from above by $\tau(1-\tau)/\{1-G(Q_\tau(\mathbf{x})|\mathbf{x})\}^2$; while for the LBT estimator, the constant given by $E\{[\tau - I(T \le Q_\tau(\mathbf{x}))]^2/(1-G(T|\mathbf{x}))\}$, assuming it is finite. In fact, the proof of (19) is done under this very assumption, namely that $E[(1-G(T|\mathbf{x}))^{-1}] < \infty$which imposes restrictions on $f_0(t|\mathbf{x})$ and $G(t|\mathbf{x})$ over the entire domain, especially near the right tail: as $t$ increases, $f_0(t|\mathbf{x})$ must decrease more quickly compared to the rate at which $1-G(t|\mathbf{x})$ decreases. In contrast, for the RDW estimator, condition [A4] concerns only $G(t|\mathbf{x})$ and with $t$ confined within the interval $[Q_{\tau^*}(\mathbf{x}), Q_{1-\tau^*}(\mathbf{x})]$.
We further have the following remarks with regards to the results given in Theorem 1.

**Remark 1.** When projected to a parametric set-up, the result is consistent with what is given Wang et al. (2009), where the 'staple' term is of order $O_p(n^{-1/2})$, the K-M estimator converges at a nonparametric whence slower rate and the estimator is therefore root-$n$ consistent. We would also like to point out that despite the fact that in some cases the 'correction' term is negligible relative to the 'staple' term, we choose to keep this term visible in the representation instead of indiscriminately sweeping it into the remainder term $R_n(\mathbf{x})$. This is because in some applications which involve averaging $\hat{\boldsymbol{\beta}}_{n\tau}(\mathbf{x})$ over $\mathbf{x}(\in \mathcal{D})$, such as the average derivative estimator (Chaudhuri et al., 1997), both terms will be of order $O_p(n^{-1/2})$ and thus play equally important roles in the asymptotics of the resulting estimator. Please refer to Section 5.1 of Kong et al. (2013) for more details on such examples.

**Remark 2.** If the smoothing parameter $\delta_n$ is allowed to go to $\infty$, the kernel weight $(n\delta_n^p)^{-1}K_{\delta_n}(.)$ assigned to all observations are eventually identical; since they must also add up to one (asymptotically at least), each observation must be given an equal weight of $n^{-1}$. It thus reduces to the case where $Q_\tau(.)$ is indeed a polynomial function, i.e. $Q_\tau(.) \equiv Q_{n\tau}(., \mathbf{x})$. The above result consequently coincides with that in Wang et al. (2009) for linear censored quantile regression.

**Remark 3.** As remarked earlier, the 'correction' term has expectation zero and a variance at most comparable to that of the 'staple' term. Therefore, the order of the 'optimal' bandwidth which minimizes the mean square error, either pointwise or globally, solely depends on the 'staple' term, which has a variance of order $O((n\delta_n^p)^{-1})$. As for its expectation, it follows from equality (A.18) in the supplementary material that

$$
\begin{aligned}
&\delta_n^{-p} E\{K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})[w_{i0}(\tau) I(Y_i \leq Q_{n\tau}(\mathbf{X}_i, \mathbf{x})) - \tau]\} \\
&= \delta_n^{-p} E\{K_{\delta_n}(\mathbf{X}_{i\mathbf{x}})\boldsymbol{\mu}_n(\mathbf{X}_{i\mathbf{x}})[1 - G(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)] f_0(Q_\tau(\mathbf{X}_i)|\mathbf{X}_i)\{Q_{n\tau}(\mathbf{X}_i, \mathbf{x}) - Q_\tau(\mathbf{X}_i)\}\} \\
&\quad + O(\delta_n^{2s_2}),
\end{aligned}
$$

where under [A2] and [A6], the first term on the right hand side is of order $O(\delta_n^{s_2})$ uniformly in $\mathbf{x} \in \mathcal{D}$. Thus the optimal bandwidth is of order $n^{-1/(p+2s_2)}$, consistent with what is already known in local polynomial smoothing (Masry, 1996).

**Remark 4.** In applications, we have control over the remainder term through proper choice of $h_n$ and $\delta_n$. For example, if the generalized K-M estimator converges at a comparable rate (i.e., $\tilde{\tau}_n = O(\tau_n)$), then through the use of an infinitely smooth kernel $\bar{K}(.)$ in (12) (i.e. $s_4 = \infty$), $R_n(.)$ in (17) would be of an order infinitely close to $O_p((n\delta_n^p/\log n)^{-3/4})$, the optimal rate according to Kong et al. (2010) with uncensored data.

# 5 Applications to semi-parametric CQR models

Applications of uniform Bahadur representation like that given in Theorem 1 include but are not limited to: (a) estimation of semiparametric models where the parameters of interest are explicit or implicit functionals of nonparametric regression functions and their derivatives, such as the Average Derivative Estimator (ADE) for single-index quantile regression models (Ichimura and Lee, 2010; Kong and Xia, 2012) with censored data; (b) estimation of structured nonparametric models like the additive models (Linton, Sperlich and Van Keilegom, 2007; Horowitz and Lee, 2005). Applying standard results on U-processes like those in Arcones (1995), Kong et al. (2013) derived the asymptotics of these estimation procedures based upon the LBT estimator defined as in (18). Parallel results could be established in a similar manner if the RDW estimator is used in place of the LBT estimator. Again, we would like to emphasize that since the LBT estimator is likely to have a much larger variance than the RDW estimator as implied by Corollary 1, estimation procedures based on the RDW estimator will again be more 'efficient' than their counterparts based on the LBT estimator. For the rest of this section, we will demonstrate another use of the uniform Bahadur representation for the purpose of sufficient dimension reduction (SDR) in regression.

The paradigm of SDR combines the idea of dimension reduction with the concept of sufficiency, and aims to generate a lower-dimensional summary plot without appreciable loss of information. Consider the following definition of SDR in regression formulated in Cook (2007), where $T$ and $\mathbf{X}$ stand for the usual scalar dependent variable and the vector of covariates, respectively. Let $\mathbf{B}$ be a $p \times q$ ($q \leq p$) (constant) orthonormal matrix. The space $\mathcal{S}(\mathbf{B})$ spanned by the columns of $\mathbf{B}$ is said to be the SDR subspace if the conditional distribution $F(.|\mathbf{B}^\top\mathbf{X})$ of $T$ given $\mathbf{B}^\top\mathbf{X}$ is identical to that of $T$ given $\mathbf{X}$, i.e.

$$F_0(.|\mathbf{X}) = F(.|\mathbf{B}^\top\mathbf{X}) \quad \text{almost surely.} \tag{20}$$

In other words, $\mathbf{B}^\top\mathbf{X}$ captures all the information relevant to regressing $T$ on $\mathbf{X}$. Under quite general conditions, the minimal SDR subspace exists and is given by

$$\mathcal{S}_0 = \cap\{\mathcal{S}(\mathbf{B}) : \text{model (20) holds for } \mathbf{B}\}.$$

We refer to it as the central subspace (CS) and its orthogonal basis $\beta_{01}, \cdots, \beta_{0q}$ as the dimension reduction directions. Let $\mathbf{B}_0 = (\beta_{01}, \cdots, \beta_{0q})$ and consequently we have

$$F_0(.|\mathbf{X}) = F(.|\mathbf{B}_0^\top\mathbf{X}) \quad \text{almost surely.} \tag{21}$$

For more detail and existing research on dimension reduction, we refer to Cook (2007) and the references therein.

The idea behind the composite quantile approach to dimension reduction of Kong and Xia (2014) is as follows. Consider the following outer-product of gradients (OPG) matrix associated with level $\tau$:

$$\Sigma(\tau) = E\{\nabla Q_\tau(\mathbf{X})[\nabla Q_\tau(\mathbf{X})]^\top\}, \quad \text{where } \nabla Q_\tau(\mathbf{X}) = \frac{\partial Q_\tau(\mathbf{X})}{\partial \mathbf{X}}.$$

What then follows from identity (21) is that for any $\tau \in (0,1)$, $\mathcal{S}(\Sigma(\tau)) \subseteq \mathcal{S}(\mathbf{B}_0)$. Indeed, it is easy to come up with examples where the above inequality holds strictly for at least one $\tau \in (0,1)$. This implies that OPG matrices evaluated at any finite number of quantiles may not always be able to reveal the full picture about the CS; what is needed is the so-called composite OPG matrix:

$$\Sigma = \int_0^1 \Sigma(\tau) d\tau; \tag{22}$$

indeed it is proved that $\mathcal{S}(\Sigma) = \mathcal{S}(\mathbf{B}_0) = \mathcal{S}_0$. Thus the search for the CS boils down to the estimation of matrix $\Sigma$, which in turn translates into the problem of estimating $\nabla Q_\tau(\mathbf{x})$ for any given $\tau$ and $\mathbf{x} = \mathbf{X}_i$, $i = 1, \cdots, n$. Suppose $\hat{\nabla} Q_\tau(\mathbf{x})$ denotes a certain estimate, an estimate of $\Sigma$ could then be formed as

$$\hat{\Sigma} = \int_0^1 n^{-1} \sum_{j=1}^n \hat{\nabla} Q_\tau(\mathbf{X}_j) \{\hat{\nabla} Q_\tau(\mathbf{X}_j)\}^\top d\tau. \tag{23}$$

Exactly the same line of arguments would apply in describing a composite quantile approach to dimension reduction for censored data. We already discussed in Section 2 how to obtain an estimator for the gradient vector $\nabla Q_\tau(.)$. Now what makes a representation like (17) particularly relevant in this context is that it could be simply plugged into (23) in place of $\hat{\nabla} Q_\tau(.)$ and the asymptotics of $\hat{\Sigma}$ could be subsequently established in a way similar to that in Kong and Xia (2014). Note that this plug-in of representations like (17) into an expression which involves integration over $\tau$ and the summation across $\mathbf{x}$ is 'legitimate' only because we have a 'uniform' control over the remainder term, i.e. it converges to zero at a speed which is uniform both in $\mathbf{x}$ and in $\tau$.

# 6 Simulation study

We examine the finite sample performance of the newly proposed RDW estimator, the LBT estimator of Kong et al. (2013), and finally the naive method where $Y_i$ takes the place of $T_i$ as if no censoring has occurred.

We will use the Epanechnikov kernel. As for bandwidth selection, regularity conditions in Theorem 1 provide general guidelines as to the speed at which the bandwidths should go to zero and in particular, about how the choice of $\delta_n$ also affects the choice of $h_n$. To reduce the computational effort, we choose $\delta_n$ by applying the 'rule-of-thumb' of Yu and Jones (1999), which relates the optimal bandwidth for conditional quantile regression to that for conditional mean regression via the following identity

$$h_\tau = h_{mean} \{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))\}^{1/5},$$

where $h_{mean}$ is the optimal bandwidth for local linear smoothing estimator in conditional mean regression, and $h_\tau$ is that for quantile regression at level $\tau$. Functions $\phi(.)$ and $\Phi(.)$ are respectively the probability density and distribution functions of the standard normal distribution.

Consider the following two (log-transformed) survival models

$$\text{Model (A): } T = X_1X_2 + (0.1 + X_1^2)^{1/2}\varepsilon$$
$$\text{Model (B): } T = \exp\{(X_1 - X_2)/\sqrt{2}\} + (0.1 + X_3)\varepsilon.$$

In both models, $T$ is subject to censoring from random variable

$$C = X_1 + X_2 + c + \epsilon,$$

where $\epsilon \sim N(0,1)$ is independent of the covariate $\mathbf{X} = (X_1, X_2, X_3)^\top$, and $c$ is a constant which dictates the expected censoring rate $P(T > C)$.

For model (A), the conditional $\tau$th quantile of $T$ is obviously given by

$$Q_\tau(\mathbf{X}) = X_1X_2 + z_\tau\sqrt{0.1 + X_1^2},$$

where $z_\tau$ is the $\tau$th quantile of $N(0,1)$. The covariates are generated according to $X_1 = U_0 + U_1$, $X_2 = U_0 + U_2$, where $U_0, U_1$ and $U_2$ are independent uniform(-1.5,1.5) random variables. We calculate estimates of $Q_\tau(u_i, u_j)$ for all possible combinations of $(u_i,\ u_j)$ with $u_i,\ u_j \in \{-2, -1.6, \cdots, 1.6, 2\}$ and define the estimation error as

$$\sum_{i=1}^{11}\sum_{j=1}^{11} |Q_\tau(u_i, u_j) - \hat{Q}_\tau(u_i, u_j)|/121.$$

A summary of the simulation results for this model is given in Table 1.

For model (B), $X_1$ and $X_2$ are both set as $N(0,1)$ with $corr(X_1, X_2) = 0.5$ and $X_3 \sim$ Uniform$(0,1)$, independent of $(X_1, X_2)$. Since the conditional quantile function is given by

$$Q_\tau(\mathbf{X}) = \exp\{(X_1 - X_2)/\sqrt{2}\} + (0.1 + X_3)z_\tau,$$

the average gradient vector evaluated at quantile level $\tau$ is thus

$$E\Big[\frac{\partial Q_\tau(\mathbf{X})}{\partial \mathbf{X}}\Big] = (c_1, -c_1, z_\tau)^\top$$

where $c_1 = E[\exp\{(X_1 - X_2)/\sqrt{2}\}]/\sqrt{2}$. Consequently, $\theta_\tau = (c_1, -c_1, z_\tau)^\top \ /\sqrt{2c_1^2 + z_\tau^2}$ is the direction of the corresponding average gradient vector and in this case we define the estimation error as

$$(1 - |\theta_\tau^\top * \hat{\theta}_\tau|)^{1/2}.$$

A summary of the simulation results for this model is given in Table 2.

We could draw the following conclusions based on these simulation results. Within any given method, there isn't any noticeable disparity among its performance across different

Table 1. Average estimation errors [ and their interquartile range] based on 100 simulation runs, for various combinations of quantile level $\tau$, sample size $n$ and censoring rate: Model (A)

| $n$ | censoring rate | $\tau$ | Naive | LBT | RDW |
|---|---|---|---|---|---|
| 200 | 60% | 0.30 | 1.3210[0.4086] | 1.0930[0.3679] | 0.8474[0.2201] |
| | | 0.50 | 1.3774[0.4649] | 1.1894[0.3154] | 0.8901[0.2634] |
| | | 0.80 | 1.6175[0.4624] | 1.5885[0.3580] | 1.2875[0.3063] |
| | 40% | 0.30 | 0.9924[0.2393] | 0.8803[0.1955] | 0.7508[0.1851] |
| | | 0.50 | 1.0069[0.2334] | 0.9446[0.2359] | 0.7804[0.1944] |
| | | 0.80 | 1.1948[0.2292] | 1.2136[0.2033] | 0.9963[0.2232] |
| | 15% | 0.30 | 0.7495[0.1692] | 0.7184[0.1586] | 0.6661[0.1738] |
| | | 0.50 | 0.7563[0.1613] | 0.7379[0.1689] | 0.6554[0.1927] |
| | | 0.80 | 0.8719[0.1731] | 0.8950[0.1775] | 0.7726[0.1501] |
| 500 | 60% | 0.30 | 1.2187[0.2565] | 0.9233[0.2223] | 0.7307[0.1962] |
| | | 0.50 | 1.2721[0.2655] | 1.0035[0.2280] | 0.7891[0.1671] |
| | | 0.80 | 1.4934[0.2511] | 1.3479[0.2438] | 1.1112[0.2347] |
| | 40% | 0.30 | 0.9111[0.1438] | 0.7821[0.1667] | 0.6589[0.1598] |
| | | 0.50 | 0.9404[0.1660] | 0.8472[0.2139] | 0.6696[0.1938] |
| | | 0.80 | 1.0915[0.1833] | 1.0401[0.1686] | 0.7577[0.2007] |
| | 15% | 0.30 | 0.6626[0.1171] | 0.6192[0.1494] | 0.5720[0.1439] |
| | | 0.50 | 0.6763[0.1335] | 0.6334[0.1534] | 0.5507[0.1492] |
| | | 0.80 | 0.7474[0.1407] | 0.7301[0.1459] | 0.5982[0.1683] |

Table 2. Average estimation errors [ and their interquartile range] based on 100 simulation runs, for various combinations of quantile level $\tau$, sample size $n$ and censoring rate: Model (B)

| $n$ | censoring rate | $\tau$ | Naive | LBT | RDW |
|---|---|---|---|---|---|
| 200 | 60% | 0.30 | 0.2836[0.1618] | 0.2369[0.1679] | 0.1503[0.1363] |
| | | 0.50 | 0.3394[0.2078] | 0.2865[0.1668] | 0.1780[0.1315] |
| | | 0.80 | 0.3180[0.2211] | 0.3127[0.2376] | 0.1679[0.1501] |
| | 40% | 0.30 | 0.1996[0.1341] | 0.1576[0.1193] | 0.1114[0.0829] |
| | | 0.50 | 0.2173[0.1679] | 0.1902[0.1616] | 0.1366[0.1116] |
| | | 0.80 | 0.2045[0.1452] | 0.1942[0.1372] | 0.1212[0.0886] |
| | 15% | 0.30 | 0.1070[0.0745] | 0.0941[0.0663] | 0.0886[0.0801] |
| | | 0.50 | 0.1067[0.0886] | 0.1019[0.0838] | 0.1004[0.0823] |
| | | 0.80 | 0.1104[0.0981] | 0.1096[0.0868] | 0.0899[0.0720] |
| 500 | 60% | 0.30 | 0.2906[0.0925] | 0.2180[0.1286] | 0.1113[0.1164] |
| | | 0.50 | 0.3383[0.1564] | 0.2424[0.1824] | 0.1024[0.0846] |
| | | 0.80 | 0.2544[0.1435] | 0.2313[0.1904] | 0.1105[0.0677] |
| | 40% | 0.30 | 0.1883[0.0916] | 0.1239[0.0896] | 0.0683[0.0656] |
| | | 0.50 | 0.2059[0.1398] | 0.1440[0.1171] | 0.0823[0.0810] |
| | | 0.80 | 0.1451[0.0902] | 0.1270[0.0928] | 0.0739[0.0619] |
| | 15% | 0.30 | 0.0951[0.0737] | 0.0724[0.0671] | 0.0603[0.0508] |
| | | 0.50 | 0.0842[0.0709] | 0.0748[0.0596] | 0.0674[0.0627] |
| | | 0.80 | 0.0826[0.0461] | 0.0788[0.0583] | 0.0650[0.0587] |

quantile levels, although in the case of the LBT estimator, when the sample size is small, its performance deteriorates dramatically as $\tau$ increases from 0.5 to 0.8. This could be attributed to the fact the LBT estimator is more susceptible to the accuracy of the preliminary K-M estimators, as discussed in Section 1. The second observation is that as the sample size increases or as the censoring rates decreases, we see a reduction in the estimation error for all three estimation methods, which is just as expected. Across the competing methods, the LBT estimator fares better than the naive estimation procedure in most cases, although to a lesser extent in the case of model (A). The RDW estimator outperforms the other two by a great margin in all cases, more so when the censoring rate is high. It is also the most stable, judging by the interquartile range of the estimation error. This is also in line with the observations by Wang et al. (2009) in linear CQR models.

## Notes

[1]Note that the conditional independence does not imply unconditional independence and vice versa; see, for example, Phillips (1988) and Su and White (2008). Here we assume the conditional independence simply because it is more common in practice.

[2]A detailed proof can be found at www.math.lsa.umich.edu/∼mityab/teaching/m395f10/solutions9.pdf, pp. 5, Problem 9.8.

REFERENCES


Arcones, M.A. (1995) A Bernstein-type inequality for U-statistics and U-processes. *Statistics and Probability Letters* 22, 239-247.

Babu, G. (1989) Strong representations for LAD estimators in linear models. *Probability Theory and Related Fields* 83, 547-558.

Bahadur, R.R. (1966) A note on quantiles in large samples. *Annals of Mathematical Statistics* 37, 577-580.

Bang, H. & Tsiatis, A.A. (2002) Median regression with censored cost Data. *Biometrics* 58, 643-649.

Bickel, P. & Rosenblatt, M. (1973) On some global measures of the deviation of density function estimators. *Annals of Statistics* 1, 1071-1095.

Buchinsky, M. & Hahn, J. (1998) An alternative estimator for the censored quantile regression Model. *Econometrica* 66, 653-671.

Carroll, R. J. (1978) On almost sure expansions for M-estimates. *Annals of Statistics* 6, 314-318.

Chen, S., Dahl, G.B. & Khan, S. (2005) Nonparametric identification and estimation of a censored location-regression model. *Journal of the American Statistical Association* 100, 212-221

Chaudhuri, P. (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics* 19, 760-777.

Chaudhuri, P., Doksum, K. & Samarov, A. (1997) On average derivative quantile regression. *Annals of Statistics* 25, 715-744.

Chernozhukov, V. & Hong, H. (2002) Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association* 97, 872-882.

Cook, R.D. (2007) Fisher Lecture: dimension reduction in regression (with discussion). *Statistical Science* 22, 1-26.

Dabrowska, D.M. (1987) Nonparametric regression with censored survival time data. *Scandinavian Journal of Statistics* 14, 181-192.

Dabrowska, D.M. (1992) Nonparametric quantile regression with censored data. *Sankhya* Ser. A 54, 252-259.

Heuchenne, C. & I. Van Keilegom (2007) Location estimation in nonparametric regression with censored data. *Journal of Multivariate Analysis* 98, 1558-1582.

Heuchenne, C. & I. Van Keilegom (2008) Estimation in nonparametric location-scale regression models with censored data. *Annals of the Institute of Statistical Mathematics* 62, 439-463.

De Gooijer, J.G. & D. Zerom (2003) On additive conditional quantiles with high dimensional covariates. *Journal of the American Statistical Association* 98, 135-146.

Efron, B. (1967) The two sample problem with censored Data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 831-853. University of California Press.

El Ghouch, A. & Van Keilegom, I. (2009) Local linear quantile regression with dependent censored data. *Statistica Sinica* 19, 1621-1640.

Fan, J. & Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications.* Chapman and Hall.

Gonzalez-Manteigaa, W. & Cadarso-Suarez, C. (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Journal of Nonparametric Statistics* 4, 65-78.

Guerre, E. & Sabbah, C. (2009) Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory* 28, 87-129.

He, X. & Shao, Q.M. (1996) A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Annals of Statistics* 24, 2608-2630.

Heuchenne, C. & I. Van Keilegom (2007). Location estimation in nonparametric regression with censored data. *Journal of Multivariate Analysis* 98, 1558-1582.

Heuchenne, C. & I. Van Keilegom (2008) Estimation in nonparametric location-scale regression models with censored data. *Annals of the Institute of Statistical Mathematics* 62, 439-463.

Ichimura, H. & Lee, S. (2010) Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics* 159, 252-266.

Honoré, B. Khan, S. & Powell, J.L. (2002) Quantile regression under random censoring. *Journal of Econometrics* 109, 67-105.

Horowitz, J.L. & Lee, S. (2005) Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association* 100, 1238-1249.

Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica* 46, 33-50.

Koenker, R. & Geling, O. (2001) Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association* 96, 458-468.

Kong, E., Linton, O. & Xia, Y. (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* 26, 1529-1564.

Kong, E. & Xia, Y. (2012) A single-index quantile regression model and its estimation. *Econometric Theory* 28, 730-768.

Kong, E., Linton, O. & Xia, Y. (2013) Global Bahadur representation for nonparametric censored regression quantiles and its applications. *Econometric Theory* 29, 941-968.

Kong, E. & Xia, Y. (2014) An adaptive composite quantile approach to dimension reduction. *Annals of Statistics* 42, 1657-1688.

Lewbel, A. & Linton, O. (2002) Nonparametric censored and truncated regression. *Econometrica* 70, 765-779.

Lindgren, A. (1997) Quantile regression with censored data using generalized $L_1$ minimization. *Computational Statistics and Data Analysis* 23, 509-524.

Linton, O., Mammen, E., Nielsen, J.P. & Van Keilegom, I. (2011) Nonparametric regression with filtered Data. *Bernoulli* 17, 60-87.

Linton, O., Sperlich, S. & Van Keilegom, I. (2008) Estimation of a semiparametric transformation model by minimum distance. *Annals of Statistics* 36, 686-718.

Mack, Y. P. & Silverman, B.W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields* 61, 405-415.

Masry, E. (1996) Multivariate local polynomial regression for time series: uniform strong convergence and rates. *Journal of Time Series Analysis* 17, 571-599.

Martinsek, A. (1989) Almost sure expansion for M-estimators and S-estimators in regression. *Technical Report*, Department of Statistics, University of Illinois.

Peng, L. & Huang, Y. (2008) Survival analysis with quantile regression models *Journal of the American Statistical Association* 103, 637-649.

Philips, P.C.B. (1988) Conditional and unconditional statistical independence. *Journal of Econometrics* 38, 341-348.

Portnoy, S. (1997) Local asymptotics for quantile smoothing splines. *Annals of Statistics* 25, 414-434.

Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators. *Econometric Theory* 7, 186-199.

Portnoy, S. (2003) Censored regression quantiles. *Journal of the American Statistical Association* 98, 1001-1012.

Powell, J.L. (1984) Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303-325.

Powell, J.L. (1986) Censored regression quantiles. *Journal of Econometrics* 32, 143-155.

Su, L. & White, H. (2008) A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* 24, 829-864.

Van Keilegom, I. & Veraverbeke, N. (1998) Bootstrapping quantiles in a fixed design regression model with censored data. *Journal of Statistical Planning and Inference* 69, 115-131.

Wang, H. & Wang, L. (2009) Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104**, 1117-1128.

Wu, W.B. (2005) On the Bahadur representation of sample quantiles for dependent sequences. *Annals of Statistics* 33, 1934-1963.

Yu, K. & Jones, M. C. (1998) Local linear quantile regression. *Journal of the American Statistical Association* 93, 228-238.

Zhou, Z. & Wu, W.B. (2009) Local linear quantile estimation for non-stationary time series. *Annals of Statistics* 37, 2696-2729.