TI 2011-011/4
Tinbergen Institute Discussion Paper

Kernel-Smoothed Conditional Quantiles of Correlated Bivariate Discrete Data

Jan G. De Gooijer[1]
Ao Yuan[2]

[1] *University of Amsterdam, and Tinbergen Institute;*
[2] *National Human Genome Center, Howard University, Washington.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Kernel-Smoothed Conditional Quantiles of Correlated Bivariate Discrete Data*

Jan G. De Gooijer[1†] and Ao Yuan[2]

[1] Department of Quantitative Economics and Tinbergen Institute
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: j.g.degooijer@uva.nl

[2] Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC, USA
e-mail: ayuan@howard.edu

Abstract: Often socio-economic variables are measured on a discrete scale or rounded to protect confidentiality. Nevertheless, when exploring the effect of a relevant covariate on the whole outcome distribution of a discrete response variable, virtually all common quantile regression methods require the distribution of the covariate to be continuous. This paper departs from this basic requirement by presenting an algorithm for nonparametric estimation of conditional quantiles when both the response variable and the covariate are discretely distributed. Moreover, we allow the variables of interest to be pairwise correlated. For computational efficiency, we aggregate the data into smaller subsets by a binning operation, and make inference on the resulting prebinned data. Specifically, we propose two kernel-based binned conditional quantile estimators, one for untransformed discrete response data and one for rank-transformed response data. We establish asymptotic properties of both estimators. A practical procedure for jointly selecting band- and binwidth parameters is also presented. Simulation results show excellent estimation accuracy in terms of bias, mean squared error, and confidence interval coverage. Typically prebinning the data leads to considerable computational savings when large datasets are under study, as compared to direct (un)conditional quantile kernel estimation of multivariate data. With this in mind, we illustrate the proposed methodology with an application to a large real dataset concerning US hospital patients with congestive heart failure.

Key words and phrases: Binning, Bootstrap, Confidence interval, Jittering, Nonparametric.

*Running title: Conditional Quantiles of Correlated Discrete Data

†Corresponding author

1 Introduction

Nonparametric estimation of conditional cumulative distribution function (CDF) of a response variable given a covariate has been well studied for continuous data. Given a sample conditional CDF, sample conditional quantiles are often computed to characterize the distribution. There are situations, however, where it is necessary to calculate sample conditional quantiles from data having a completely discrete distribution. For example, in studying the relationship between monthly unemployment spell and experience-education profile (in years), policymakers want to know whether higher education reduces unemployment spell between the 25th and 75th quantiles of the response variable as a function of the covariate. Another example concerns the need for better management of hospital care by describing the shape of the conditional distribution of the length of hospital stay, a variable often considered as a measure of patients' recovery, given covariates like patients age, sex, gender, ethnicity, or severity of disease.

Methods for *un*conditional quantile estimation for discrete data have been proposed by González-Barrios and Rueda (2001), Chen and Lazar (2010), and Frydman and Simon (2007), etc. Machado and Santos Silva (2005) introduced a variant of quantile regression for mixed discrete-continuous variables. More recently, Li and Racine (2008) considered nonparametric estimation of conditional CDFs for mixed discrete (categorical)–continuous random variables. Clearly, the latter two approaches are restricted by the continuous assumption of the covariates. But, as with the examples above, in practice there may not exist continuously distributed covariates. In this paper we consider a setting where both the response variable and the covariate are assumed to be discretely distributed. Moreover, we allow the variables of interest to be pairwise correlated. For efficient and effective smoothing, we aggregate the data into smaller subsets by a binning operation, and make inference on the resulting prebinned data. This set-up may offer new insights as to the nature of the relationship between a response and potential covariates, possibly with interesting implications to the issue at hand.

As an illustration Figure 1 displays a scatterplot of a discrete-valued response variable "Length of hospital stay" (in days) versus "Severity of disease" (measured on a 7-point scale) for a dataset of $20,631$ patients; see Section 7 for more details. In addition Figure 1 shows some selected conditional percentiles using a "naive" method, i.e. assuming the data are from a continuous distribution. Kendall's rank correlation coefficient indicates, at a two-tailed significance level of 1%, that both variables are positively correlated. Note that the conditional percentiles diverge noticeably toward the high end of the severity scale. For instance the central

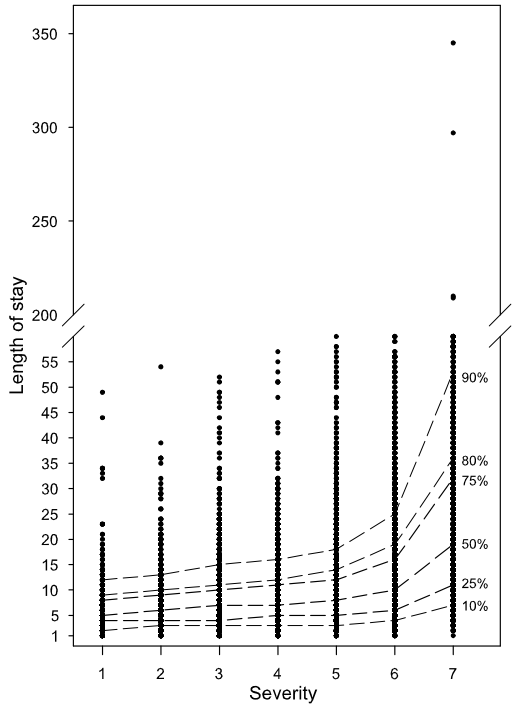

Figure 1. Some selected population conditional percentiles for the discrete-valued response variable "Length of hospital stay" versus the covariate "Severity of disease" based on 20,631 observations.

50% of the response, 25th through 75th percentile, at a value 5 of the covariate is equal to 8 days while this becomes 22 days when the covariate takes a value 7. On the other hand, when unconditional percentiles are computed, ignoring correlations between the response variable and the covariate, the central 50% of the response variable is equal to 10 days. Hence, conditional percentiles convey more information about the data than unconditional percentiles.

Note that the above dataset is large, which is common in many areas of science. However, analysing large sample sizes n often needs special statistical methods to reduce the amount of storage space and computing time. For instance, computing a standard kernel-smoothed density estimator at m evaluation points for continuously distributed data requires nm kernel evaluations; see, e.g., Fan and Marron, 1994. A number of methods are available to address this problem and most work with the idea of binning the data first. The binning operation will lead to substantial computational savings since calculations are based on the number of bins, rather than the number of data points with the total number of operations directly proportional to the expected number of nonempty bins. Moreover, it has been shown (cf., González-Manteiga,

Sánchez-Sellero and Wand 1996, and Holmström, 2000) that the estimation error of the binned kernel-smoothed density estimator is essentially the same as that of the standard kernel estimator for continuous data.

With a strong view toward analysing large datasets we propose a kernel-based, nonparametric, approach for conditional quantile estimation of correlated discretely distributed bivariate data by binning the discrete-valued response variable. Binning has been studied in the case of unconditional kernel-based density estimation and local polynomial kernel estimation of continuous data. Its application to estimating conditional quantiles for data from bivariate discrete distributions has not been a topic of research, as far as we are aware. Often binning is viewed as a discrete approximation to the continuum with only mass points on a regular grid. But in our case the discretization of the binning operation naturally parallels the type of data under study. Hence, the binning operation is well motivated.

With discretely distributed data, each bin defines a local neighborhood of the data in the sense that observations within a bin have similar probability mass function (pmf). The resulting conventional binned sample conditional CDF is purely discrete, jumping at the bin points. Thus asymptotic theory breaks down due to the fact that the sample conditional CDF is not absolutely continuous with respect to Lebesgue measure. Consequently, asymptotic properties of binned sample conditional quantiles cannot be obtained by Taylor series expansion. To circumvent this problem we propose a new conditional quantile estimator by linearly interpolating between the jumps of the conventional binned sample CDF. In fact, we consider two variants of this estimator: one for untransformed discrete data, and one that can be used for rank-transformed response data. In both cases we establish, under mild regularity conditions, asymptotic normality and consistency. In addition to these theoretical contributions, we also formulate a data-driven algorithm for jointly selecting the band- and binwidths of the proposed conditional quantile estimators.

The paper has seven remaining sections. In Section 2, we present the binning principle in its simplest form. In Section 3, we define two binned kernel-based conditional quantile estimators. Section 4 presents asymptotic properties of the proposed quantile estimators. Various practical issues related to the two estimators are discussed in Section 5. In Section 6 we provide numerical simulations of the performance of the estimators for correlated bivariate discrete random variables. Section 7 contains an empirical illustration. Finally, Section 8 presents some concluding remarks. We relegate all technical arguments to an Appendix.

2 Binning

Let $\{Z_1, \ldots, Z_n\}$ denote a set of observed random variables having a probability density function $f(z)$ where $z \in \mathbb{R}$ has bounded support on $[a, b]$ $(-\infty < a < b < \infty)$. To bin the data, we divide the support set into M intervals (bins, or alternatively referred to as grid sizes), $\{\mathcal{A}_j\}_{j=1}^{M}$, with $\mathcal{A}_j = [a_{j-1}, a_j)$ $(j = 1, \ldots, M-1)$, $\mathcal{A}_M = [a_{M-1}, a_M]$, and $a = a_0 < a_1 < \cdots < a_M = b$. For notational simplicity, we assume that the lengths of the intervals (binwidths), $\delta = (a_j - a_{j-1})$, are fixed across the intervals. We refer to $g_\ell \equiv \ell\delta$ $(\ell = 0, 1, \ldots, M)$ as the grid points. Then a binning rule in its simplest form may be represented by a sequence of weights $\{w_\ell(z, \delta); \ell = 0, 1, \ldots, M\}$, with $\sum_{\ell=0}^{M} w_\ell(z, \delta) = 1$, such that the data are replaced by the weights attached to each grid point according to a certain rule. Thus a new data value is created as $c_\ell(z) = \sum_{i=1}^{n} w_\ell(Z_i, \delta)$ (grid count) at grid point g_ℓ. One example of a binning rule, adopted in this paper, is linear binning (Jones and Lotwick, 1983) where

$$c_\ell(z) = \sum_{i=1}^{n} (1 - |\delta^{-1} Z_i - \ell|)^+, \tag{2.1}$$

with $x^+ = \max(0, x)$. Thus, it assigns a continuous value at l by the total weights attached to it. Data values Z_i closer to g_ℓ contribute more weight, while Z_i's with distance $> \delta$ contribute zero weight. Hall and Wand (1996) showed that, in terms of approximation error, linear binning is more accurate than simple binning, where each observation is assigned to its nearest grid point.

3 Binned Conditional Quantiles for Discrete Distributions

3.1 Untransformed data

Our analysis concerns a pair of random variables (X, Y) such that X (covariate) and Y (response) are discrete, taking values in the space $\mathbb{N}_1 \times \mathbb{N}_0$ from a joint discrete probability distribution function $F(x, y) = \Pr(X \le x, Y \le y)$, with $\mathbb{N}_1$ the set of positive integers and $\mathbb{N}_0$ the set of non-negative integers. Then it is well-known that the α-conditional quantile $(\alpha \in (0, 1))$ is defined as the value $\theta_\alpha(x)$ that solves $F(\theta_\alpha(x)|x) = \alpha$, where $F(y|x) = \Pr(Y \le y | X = x)$, or alternatively $\theta_\alpha(x) = \min_{\eta \in \mathbb{N}_1}\{F(\eta|x) \ge \alpha\}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote a random sample of size n from the distribution $F(x, y)$. Assume that the raw data $\{Y_i\}$ are arranged in increasing order, giving rise to the data $\{Y_{(1)}, \ldots, Y_{(n)}\}$ with $\{\tilde{X}_i\}$ denoting the corresponding rearranged X_i's.

To obtain a binned estimator of $F(y|x)$, the dataset $\{\tilde{X}_i, Y_{(i)}\}$ is split up into m disjoint subsets with m denoting the number of distinct values x_u^* $(u = 1, \ldots, m)$ taken on by the X_i's

$(i = 1, \ldots, n)$. Assume that each subset contains $n_u = \sum_{i=1}^{n} I(\tilde{X}_i = x_u^*)$ $(n_u > 0; u = 1, \ldots, m)$ observations, with $I(A)$ the indicator function for set $\{A\}$. In addition, let the grid count at grid point g_ℓ, conditioning on the covariate $X = x_u^*$, be given by

$$c_{u,\ell}(y|x) = \sum_{i=1}^{n} I(\tilde{X}_i = x_u^*)(1 - |\delta^{-1} Y_{(i)} - (\ell - 1)|)^+, \ (u = 1, \ldots, m; \ell = 0, 1, \ldots, M - 1). \quad (3.1)$$

Note that (3.1) is a rescaled version of (2.1) with $I(\tilde{X}_i = x_u^*)$ representing the conditioning on $X = x$ in $F(y|x)$. The rescaling allows for the case $c_{u,0}(y|x) = 0$. Given (3.1), and assuming $n_u \neq 0$, the relative cumulated grid counts at each grid point g_ℓ $(\ell = 0, \ldots, M)$ conditional on each value x_u^* $(u = 1, \ldots, m)$ can be computed.

Now the basic idea is to accumulate the bin weights in a matrix with each row corresponding to a bin and each column to a conditioning value, and then kernel smooth the rows. This requires, three steps. Summarize the relative cumulated grid counts $c_{u,\ell}(\cdot)$ into an $M \times m$ matrix $\boldsymbol{C}$, defined as follows

$$C_{v,u}(y|x) = \begin{cases} 0 & \text{if} \quad u = 1, \ldots, m; v = 1, \\ \frac{1}{n_u} \sum_{\ell=1}^{v-1} c_{u,\ell}(y|x) & \text{if} \quad u = 1, \ldots, m; v = 2, \ldots, M. \end{cases}$$

Next, compute an $m \times m$ kernel weighting matrix $\boldsymbol{W}$, defined such that its (u, j)th element is given by

$$W_{u,j}(h) = \frac{n_u K_d((x_j - x_u^*)/h)}{\sum_{u=1}^{m} n_u K_d((x_j - x_u^*)/h)},$$

where $K_d(\cdot)$ is a discrete kernel function on the ordered set $[1, \infty]$, with h $(h \in \mathbb{N}_1)$ the bandwidth. Finally, combine $\boldsymbol{C}$ and $\boldsymbol{W}$ into an $M \times m$ matrix $\boldsymbol{CW}$ with binned kernel-smoothed empirical conditional CDFs. The (v, j)th element $(v = 1, \ldots, M; j = 1, \ldots, m)$ of this product of matrices is given by

$$F_n(y|x) = \frac{1}{\sum_{u=1}^{m} n_u K_d((x_j - x_u^*)/h)} \sum_{u=1}^{m} n_u K_d((x_j - x_u^*)/h) C_{v,u}(y|x). \quad (3.2)$$

Note (3.2) may be considered as a binned discrete analogue of the Nadaraya-Watson conditional CDF estimator in the case (X, Y) is continuously distributed.

Using (3.2), it may be shown that the binned conditional quantile estimator is not consistent if α is at a plateau of $F(y|x)$, i.e. $F(y|x) = \alpha$ and $F(y|x)$ is flat in a right-neighborhoud of $\theta_\alpha(x)$; see Section 4. A way around this difficulty is to define a new conditional CDF, say $\tilde{F}(y|x)$, from $F(y|x)$ by interpolation at successive grid points, i.e.

$$\tilde{F}(y|x) = F(g_L|x) + \Big(\frac{y - g_L}{\delta}\Big)\Big[F(g_{L+1}|x) - F(g_L|x)\Big], \quad (3.3)$$

where L is chosen such that $g_L \leq y < g_{L+1}$. Consequently, the corresponding conditional quantile is defined as $\theta_\alpha^c(x) = \inf_{\eta\in\mathbb{R}}\{\tilde{F}(\eta|x) \geq \alpha\}$, where the superscript c highlights the fact that (3.3) is a continuous CDF (see Section 4 for our choice of interpolation).

Given (3.3), and making use of (3.2), an estimator $F_n(y|x)$ of $\tilde{F}(y|x)$ at a general point x_u^* $(u = 1, \ldots, m)$, can be obtained as a linear interpolation between successively grid points. More precisely, the binned kernel-smoothed empirical conditional CDF is given by

$$\tilde{F}_n(y|x_u^*) = F_n(g_L|x_u^*) + \left(\frac{y - g_L}{\delta}\right)\Big[F_n(g_{L+1}|x_u^*) - F_n(g_L|x_u^*)\Big]. \tag{3.4}$$

Hence the binned smoothed conditional quantile estimator of $\theta_\alpha(x)$ is defined by

$$\hat{\theta}_\alpha^c(x_u^*) = \begin{cases} \left(\frac{\alpha - F_n(g_L|x_u^*)}{F_n(g_{L+1}|x_u^*) - F_n(g_L|x_u^*)}\right)g_L + \left(\frac{F_n(g_{L+1}|x_u^*) - \alpha}{F_n(g_{L+1}|x_u^*) - F_n(g_L|x_u^*)}\right)g_{L+1}, & \text{if } n_u \neq 0, \\ \theta_0 & \text{if } n_u = 0, \end{cases} \tag{3.5}$$

where θ_0 is an arbitrary value, and where L is chosen such that $\sum_{v=1}^{L} F_n(g_v|x_u^*) \leq \alpha < \sum_{v=1}^{L+1} F_n(g_v|x_u^*)$.

The binned conditional quantile estimator $\hat{\theta}_\alpha(\cdot)$ of $F(y|x)$ can be obtained from (3.5) by using the transformation

$$\hat{\theta}_\alpha(x_u^*) = \lceil \hat{\theta}_\alpha^c(x_u^*) - 1 \rceil, \tag{3.6}$$

where $\lceil\cdot\rceil$ is the ceiling function that returns the smaller integer greater than, or equal to its argument; see Machado and Santos Silva (2005, Theorem 2). Clearly, the advantage of binning stems essentially from the fact that the grid counts need to be computed only once.

3.2 Rank-transformed data

In practice one may encounter a situation with highly skewed empirical conditional CDFs (see Section 7). This gives rise to many empty bins and a few bins containing a large proportion of the observations. A rank transformation of Y_i avoids this problem. Hence, we now introduce a binned conditional quantile estimator for rank-transformed discrete-valued data. The proposed estimator is similar in spirit to a kernel-smoothed binned conditional quantile estimator for bivariate continuously distributed random variables introduced by Magee, Burridge and Robb (1991).

Ranks can be assigned to data in several ways. Here, we rank the entire set of observations $\{Y_i\}$ from smallest to largest, giving rise to the dataset $\{R_1, \ldots, R_n\}$ with $\{\tilde{X}_i\}$ the corresponding rearranged X_i's. In the case of ties we order tied observations at random (random ranking

approach). Then, following the same setup as introduced in Subsection 3.1, the dataset $\{\tilde{X}_i, R_i\}$ is split up into m disjoint subsets each containing n_u $(u = 1, \ldots, m)$ observations. From the ranking of the response data we conclude that $g_0 = 0$ and $g_M = n + 1$. Then, similar to (3.1), we assume that the grid count at grid point g_ℓ, conditioning on the covariate $X = x_u^*$, is given by

$$\tilde{c}_{u,\ell}(y|x) = \sum_{i=1}^{n} I(\tilde{X}_i = x_u^*)(1 - |\delta^{-1}R_i - (\ell - 1)|)^+, \ (u = 1, \ldots, m; \ell = 0, 1, \ldots, M-1), \quad (3.7)$$

where $\delta = (n+1)/M$ denotes the corresponding binwidth. Thus, in analogy with (3.2), a typical element of the $M \times m$ matrix $\boldsymbol{CW}$ with binned rank-transformed empirical CDFs is given by

$$\begin{aligned} F_n^R(y|x) &= \frac{1}{\sum_{u=1}^{m} n_u K_d((j-u)/h)} \sum_{u=1}^{m} n_u K_d((j-u)/h) \tilde{C}_{v,u}(y|x), \qquad (3.8) \\ &\qquad\qquad (v = 1, \ldots, M; j = 1, \ldots, m), \end{aligned}$$

where

$$\tilde{C}_{v,u}(y|x) = \begin{cases} 0 & \text{if} \quad u = 1, \ldots, m; v = 1, \\ \frac{1}{n_u} \sum_{\ell=1}^{v} \tilde{c}_{u,\ell}(y|x) & \text{if} \quad u = 1, \ldots, m; v = 2, \ldots, M. \end{cases}$$

Using (3.8), the αth smoothed conditional quantile estimator $\tilde{\theta}_\alpha(x_u^*)$ for the binned rank-transformed data, at a general point x_u^* $(u = 1, \ldots, m)$, can be obtained as a linear interpolation between two known successive grid points. More precisely, the binned kernel-smoothed empirical conditional CDF is given by

$$\tilde{F}_n^R(y|x_u^*) = F_n^R(y|x_u^*) + \Big(\frac{y - g_L}{g_{L+1} - g_L}\Big)[F_n^R(y+1|x_u^*) - F_n^R(y|x_u^*)], \qquad (3.9)$$

where L is such that $g_L < y < g_{L+1}$, and $y+1$ denotes the smallest rank in the data bigger than y. Hence, the binned kernel-smoothed quantile estimator for the rank-transformed observations is given by

$$\tilde{\theta}_\alpha(x_u^*) = \begin{cases} \Big(\frac{\alpha - F_n^R(y|x_u^*)}{F_n^R(y+1|x_u^*) - F_n^R(y|x_u^*)}\Big) g_L + \Big(\frac{F_n^R(y+1|x_u^*) - \alpha}{F_n^R(y+1|x_u^*) - F_n^R(y|x_u^*)}\Big) g_{L+1} & \text{if } n_u \neq 0, \\ \theta_0 & \text{if } n_u = 0, \end{cases} \qquad (3.10)$$

where L is chosen such that $\sum_{v=1}^{L} F_n^R(y|x_u^*) \leq \alpha < \sum_{v=1}^{L+1} F_n^R(y|x_u^*)$, and θ_0 is an arbitrary value. Transforming (3.10) back into the original observations, and assuming $n_u \neq 0$, results in the following binned kernel-smoothed estimator of $\theta_\alpha(x_u^*)$:

$$\tilde{\theta}_\alpha^{\mathrm{Y}}(x_u^*) = [1 - (\tilde{\theta}_\alpha(x_u^*) - \lfloor \tilde{\theta}_\alpha(x_u^*) \rfloor)]\tilde{Y}_{(u, \lfloor \tilde{\theta}_\alpha(x_u^*) \rfloor)} + (\tilde{\theta}_\alpha(x_u^*) - \lfloor \tilde{\theta}_\alpha(x_u^*) \rfloor)\tilde{Y}_{(u, \lfloor \tilde{\theta}_\alpha(x_u^*) \rfloor + 1)}, \quad (3.11)$$

where $\lfloor \cdot \rfloor$ is the floor function returning the greatest integer less than or equal to its argument, and where $\{\tilde{Y}_{(u,1)}, \ldots, \tilde{Y}_{(u,n_u)}\}$ are the ordered (increasing order) values of $\{Y_i\}$ when $\tilde{X}_i = x_u^*$ $(i = 1, \ldots, n; u = 1, \ldots, m)$. Then, similar to the transformation in (3.6), the discrete conditional binned quantile estimator of $\theta_\alpha(\cdot)$ is given by

$$\tilde{\theta}_\alpha(x_u^*) = \lceil \tilde{\theta}_\alpha^{\mathrm{Y}}(x_u^*) - 1 \rceil. \tag{3.12}$$

The code for computing (3.6) and (3.12) is available upon request from first author.

4 Asymptotic Results

To study the asymptotic properties of $\hat{\theta}_\alpha(\cdot)$ we use the method of jittering; see, e.g., Machado and Santos Silva (2005). With this method, a continuous distribution is constructed which coincides with the discrete distribution up to interpolation. Asymptotic properties of the conditional quantile estimator in the continuous case are easily obtained using existing theory, and then asymptotic results for the discrete case follow from relationship (3.6). Asymptotic properties of (3.12) will be based on theorems for linear rank statistics given by Hájek, Šidák and Sen (1999).

4.1 Untransformed data

Without loss of generality, we assume that the support of $Y|x$ is composed of consecutive integers $\{0, 1, 2, \ldots\}$. Let $p_i = P(Y = y_i|x)$ and $P_j = P(Y \leq y_j|x) = \sum_{i \leq j} p_i$. We add small jitters e_i to Y_i, so that the data $Z_i = Y_i + e_i$ $(i = 1, \ldots, n)$ are of continuous type, and the order of the Y_i's is preserved, i.e., if $Y_i < Y_j$ then $Z_i < Z_j$. The distribution of e_i is chosen to be independent of the value of Y_i as $e_i \sim U[0, 1)$, the uniform distribution on $[0, 1)$. Let $F_J(z|x)$ be the distribution of the Z_i's, then we have the following fact

$$F_J(k|x) = F(k|x), \quad (k = 0, 1, 2, \ldots).$$

$F_J(\cdot|x)$ is a continuous distribution by linear interpolating $F(\cdot|x)$, and $\tilde{F}(y|x)$ given in (3.3) is exactly the conditional distribution function of the binned version based on the Z_i's, $F_n(y|x)$ in (3.4) is the corresponding empirical version, $\hat{\theta}_\alpha^c(\cdot)$ is the corresponding conditional quantile estimate, and the relationship between $\hat{\theta}_\alpha(\cdot)$ and $\hat{\theta}_\alpha^c(\cdot)$ is given in (3.6).

As in Cheng and Lazar (2010), we consider two cases.

(i) $P_{k-1} < \alpha < P_k$ for some $k \in \{1, 2, \ldots\}$. So α is not at a plateau of $\tilde{F}(\cdot|x)$. Using results for continuous distribution, we will have $\hat{\theta}_\alpha^c(\cdot) \overset{a.s.}{\to} \theta_\alpha^c(\cdot)$, and since the function $\lceil t - 1 \rceil$ is continuous at $t = \theta_\alpha^c(\cdot)$, consequently we have $\hat{\theta}_\alpha(\cdot) = \lceil \hat{\theta}_\alpha^c(\cdot) - 1 \rceil \overset{a.s.}{\to} \lceil \theta_\alpha^c(\cdot) - 1 \rceil = \theta_\alpha(\cdot)$.

(ii) If $\alpha = P_k$ for some k, then $\theta_\alpha(\cdot)$ is an integer, the function $\lceil t-1 \rceil$ has a jump at $t = \theta_\alpha(\cdot)$. Consequently, although $\hat{\theta}^c_\alpha(\cdot) \xrightarrow{a.s.} \theta^c_\alpha(\cdot)$, but $\lceil \hat{\theta}^c_\alpha(\cdot) - 1 \rceil \xrightarrow{a.s.} \lceil \theta^c_\alpha(\cdot) - 1 \rceil$ fails because of the discontinuity. So for such α, $\hat{\theta}_\alpha(\cdot)$ is not consistent for $\theta_\alpha(\cdot)$. As implied in Serfling (1980, p.77), in this case, for large n, $\hat{\theta}_\alpha(\cdot) \geq \theta_\alpha(\cdot) + 1$ with probability approximately 0.5.

The theoretical results are derived under the following assumptions:

A.1 $h = h_n \to 0$ and $\sum_{n\geq 1} \exp(-Cnh_n) < \infty$ for all $C > 0$.

A.2 $\delta = \delta_n \to 0$.

A.3 $K(\cdot)$ has bounded continuous third derivative.

A.4 $h = o(n^{-1/5})$ and $\delta = O((nh)^{-1/4})$.

Let $\xrightarrow{L_2}$ stands for convergence in squared mean, $\xrightarrow{D}$ for convergence in distribution, $B(p)$ be Bernoulli distribution with $p = P(X = 1)$, $p(x)$ be the pmf of X, and $p(x, y)$ the joint pmf of (X, Y).

Theorem 1. *Under A.1–A.3, as $n \to \infty$, we have*

$$\sup_y E(F_n(y|x) - F(y|x))^2 \to 0; \qquad \sup_y E(\tilde{F}_n(y|x) - \tilde{F}(y|x))^2 \to 0.$$

Theorem 2. *Under A.1–A.3, as $n \to \infty$, we have*

$$(i) \qquad \hat{\theta}^c_\alpha(x) \xrightarrow{L_2} \theta^c_\alpha(x), \qquad \forall \alpha \in (0,1);$$

$$(ii) \qquad \hat{\theta}_\alpha(x) \xrightarrow{L_2} \theta_\alpha(x), \qquad \forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}.$$

Theorem 3. *Assume that A.3 and A.4 hold. Then*

(i) $\forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}$, we have

$$\sqrt{nh}(\hat{\theta}^c_\alpha(x) - \theta^c_\alpha(x)) \xrightarrow{D} N(0, \sigma^2), \text{ with } \sigma^2 = \alpha(1-\alpha)p(x) \int K^2(v)dv / p^2(x, \theta^c_\alpha(x)). \tag{4.1}$$

(ii) If $\alpha = P_k$ for some $k \in \mathbb{N}_1$, we have

$$\sqrt{nh}(\hat{\theta}^c_\alpha(x) - \theta^c_\alpha(x)) \xrightarrow{D} N^+(0, \sigma^2_{k,+}) + N^-(0, \sigma^2_{k,-}), \tag{4.2}$$

where $\sigma^2_{k,+} = \alpha(1-\alpha)p_k \int K^2(v)dv/[p_k p(\theta^c_\alpha(x)|x)]^2$, $\sigma^2_{k,-} = \alpha(1-\alpha)p_{k-1} \int K^2(v)dv/[p_{k-1} p(\theta^c_\alpha(x)|x)]^2$, and $N^+(0, \sigma^2_1) + N^-(0, \sigma^2_2)$ denotes the two-piece normal distribution function $\Phi(\sigma_1 t) I(t >$

$0) + \Phi(\sigma_2 t)I(t < 0)$, *with* $\Phi(\cdot)$ *the distribution function of* $N(0,1)$.

(iii) $\forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}$, *or* $\alpha \in \{P_k : k \in \mathbb{N}_0\}$, *we have*

$$\hat{\theta}_\alpha(x) - \theta_\alpha(x) \xrightarrow{D} B(0); \quad \text{or} \quad \hat{\theta}^c_\alpha(x) - \theta^c_\alpha(x) \xrightarrow{D} B(0.5).$$

In practice, the unknown quantities p_k, $p(x)$ and $p(x, \theta^c_\alpha(x))$ for given α can be estimated by

$$\hat{p}_k = \sharp\{i : (X_i, Y_i) = (x,k)\}/n, \quad \hat{p}(x) = \sharp\{i : X_i = x\}/n, \quad \text{and} \quad p(x, \hat{\theta}^c_\alpha(x)),$$

where $\hat{\theta}^c_\alpha(x)$ is given by (3.5). Let $\hat{\sigma}^2$ be an estimator of σ^2 with $p(x)$ and $p(x, \theta^c_\alpha(x))$ replaced by $\hat{p}(x)$ and $p(x, \hat{\theta}^c_\alpha(x))$. Denote $\hat{\sigma}^2_{k,+}$ and $\hat{\sigma}^2_{k,+}$ accordingly. For fixed x, let n_x be the number of values x taken on by the X_i's, $n_{k|x}$ be the number of (Y_i, X_i)'s with $X_i = x$ and $Y_i = k$. Then we have the following corollary.

Corollary 1. *Assume A.1–A.4 hold, and that* $n_x \to \infty$ *and* $n_{k|x} \to \infty$ *as* $n \to \infty$, *then Theorem 3 still holds with* σ^2, $\sigma^2_{k,+}$ *and* $\sigma^2_{k,-}$ *replaced by* $\hat{\sigma}^2$, $\hat{\sigma}^2_{k,+}$ *and* $\hat{\sigma}^2_{k,+}$ *respectively.*

4.2 Rank-transformed data

Since $\tilde{F}^R_n(v|j) = F^R_n(v|j)$ for all v ($v = 1, \ldots, M; j = 1, \ldots, m$), so $\tilde{F}^R_n(\cdot|j) \to \tilde{F}(\cdot|j)$ iff $F^R_n(\cdot|j) \to F(\cdot|j)$. We impose the following conditions:

B.1 $nh \to \infty$, with $h > 0$.

B.2 $\delta \to 0$.

B.3 $\int K^2(t)dt < \infty$.

B.4 $\delta = o((nh)^{-1/2})$.

Theorem 4. *Under B.1–B.3, we have*

$$F^R_n(y|x) \xrightarrow{P} F(y|x); \qquad \tilde{F}^R_n(y|x) \xrightarrow{P} \tilde{F}(y|x).$$

Theorem 5. *Under B.1–B.3, we have*

$$(i) \qquad \tilde{\theta}^R_\alpha(x) \xrightarrow{P} \theta^c_\alpha(x), \qquad \forall \alpha \in (0,1);$$

$$(ii) \qquad \tilde{\theta}_\alpha(x) \xrightarrow{P} \theta_\alpha(x), \qquad \forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}.$$

Theorem 6. *Assume B.1, B.3 and B.4 hold. Then*

(i)

$$\sqrt{nh}(\tilde{F}_n^R(y|x) - \tilde{F}(y|x)) \xrightarrow{D} N(0, \sigma_1^2), \text{ with } \sigma_1^2 = \tilde{F}(y|x)(1 - \tilde{F}(y|x))p^{-1}(x) \int K^2(v)dv.$$

(ii) $\forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}$, *we have*

$$\sqrt{nh}(\tilde{\theta}_\alpha(x) - \theta_\alpha(x)) \xrightarrow{D} N(0, \sigma^2), \text{ with } \sigma^2 = (1-\alpha) \int K^2(v)dv / p^2(x, \theta_\alpha(x)).$$

(iii) If $\alpha = P_k$ *for some* $k \in \mathbb{N}_1$, *we have*

$$\sqrt{nh}(\tilde{\theta}_\alpha(x) - \theta_\alpha(x)) \xrightarrow{D} N^+(0, \sigma_{k,+}^2) + N^-(0, \sigma_{k,-}^2),$$

where $\sigma_{k,+}^2 = \alpha(1-\alpha) \int K^2(v)dv / [p_k p(\theta_\alpha^c(x)|x)]^2$, $\sigma_{k,-}^2 = \alpha(1-\alpha) \int K^2(v)dv / [p_{k-1} p(\theta_\alpha^c(x)|x)]^2$.
(iv) $\forall \alpha \in (0,1) \setminus \{P_k : k \in \mathbb{N}_0\}$, *or* $\alpha \in \{P_k : k \in \mathbb{N}_0\}$, *we have*

$$\tilde{\theta}_\alpha(x) - \theta_\alpha(x) \xrightarrow{D} B(0); \quad \text{or} \quad \tilde{\theta}_\alpha(x) - \theta_\alpha(x) \xrightarrow{D} B(0.5).$$

Proofs of these theorems may be found in the Appendix.

In practice, $\tilde{F}(y|x)$ is estimated by $\tilde{F}_n^R(y|x)$, σ_1^2 by $\hat{\sigma}_1^2$ with $\tilde{F}(y|x)$ and $p^{-1}(x)$ replaced by $\hat{p}^{-1}(x)$ and $\tilde{F}_n^R(y|x)$. Denote $\hat{\sigma}^2$ accordingly, with $\theta_\alpha(x)$ and $\theta_\alpha^c(x)$ replaced by $\tilde{\theta}_\alpha(x)$ and $\tilde{\theta}_\alpha^R(x)$, and $\hat{p}(x)$ and $p(x, \hat{\theta}_\alpha(x))$ as in Corollary 1. Then we have the following corollary.

Corollary 2. *Assume B.1–B.4 hold, and the conditions on* n_x *and* $n_{k|x}$ *of Corollary 1, then Theorem 6 still holds.*

5 Practical Issues

5.1 Kernel selection

It is well-known that the bias of kernel estimates of pmfs for cells near the boundary of the sample space can incur increased bias. Reduced bias may be achieved by using boundary corrected kernel estimators. For discrete data, a kernel function devised to correct boundary effects has been proposed by Rajagopalan and Lall (1995). The general form of this kernel is given by

$$K_d(u_j) = au_j^2 + b, \quad |u_j| \leq 1, \tag{5.1}$$

where $u_j = (i-j)/h$ $(i, j \in \mathbb{N}_1)$ with i the point at which the kernel function is evaluated. The parabolic shape of (5.1) was inspired by the so-called Epanechnikov kernel in the continuous

Table 1: Expressions for the coefficients a and b for the discrete quadratic kernel $K_d(u_j) = au_j^2+b$, $|u_j| \leq 1$, where $u_j = (i-j)/h$ with i the point at which the kernel is evaluated.

Region	Coefficients	
	a	b
$i > h+1$	$\frac{3h}{1-4h^2}$	$\frac{-3h}{1-4h^2}$
$1 < i \leq h+1$	$\frac{-6h^2}{(i-1+h)(i-2+h)(i-3+h)}$	$\frac{3(2-h+h^2-3i+i^2)}{(i-1-h)(i-2+h)(i-3+h)}$
$i = 1$ $(h > 1)$	$\frac{-6h}{h^2-1}$	$\frac{3}{h+1}$

Note: Based on Rajagopalan and Lall (1995), after correcting mistakes in the expressions for a and b for the regions $i > h+1$ and $1 < i \leq h+1$. Further, the expressions for a and b with $i = 1$ given by these authors, have been simplified.

case, which enjoys some optimality properties as opposed to other commonly used kernels. The coefficients a and b are functions of h. They can be obtained by solving the following conditions: i) $K_d(u_j) = 0$ for $|u_j| \geq 1$, ii) $\sum_{j=i-h}^{j=i+h} K_d(u_j) = 1$, and iii) $\sum_{j=i-h}^{j=i+h} K_d(u_j)u_j = 0$. Expressions for a and b are given in Table 1. Throughout the paper we adopt (5.1). But in the initial phase of the simulation study we experimented with (5.1), a triangular, and a rectangular kernel function. In the latter two cases, the bias of the conditional quantile estimators can be quite large for near-boundary bins if $n \leq 10,000$.

5.2 Grid size, band- and binwidth selection

To obtain good nonparametric estimates, we use the cross-validation (CV) principle for jointly selecting band- and binwidth parameters, adapting the CV-approach used by Magee, Burridge and Robb (1991). As an example, consider the case of estimating (3.12). Given a fixed value of δ a value of h can be obtained by minimizing the loss function $L(h) = \sum_{u=1}^{m}\sum_{i=1}^{n} I(\tilde{X}_i = x_u^*)\rho_\alpha(Y_i - \tilde{\theta}_\alpha^{(-i)}(x_u^*))$, where $\rho_\alpha(u) = \alpha u I(u > 0) + (\alpha - 1)uI(u < 0)$ is the so-called "check" function, and $\tilde{\theta}^{(-j)}(\cdot)$ is the "delete-one" conditional quantile estimate.

To avoid summing over n terms, it is convenient to subtract from $L(h)$ the quantity $L^* = \rho_\alpha(Y_i - \breve{\theta}_\alpha(x_u^*))$, where $\breve{\theta}_\alpha(x_u^*)$ is some quantile estimate at x_u^* that does not vary with h. Let n_{1u} (n_{4u}) be the number of observations $Y_i < \min\{\tilde{\theta}_\alpha(x_u^*), \breve{\theta}_\alpha(x_u^*)\}$ ($Y_i > \max\{\tilde{\theta}_\alpha(x_u^*), \breve{\theta}_\alpha(x_u^*)\}$), when $I(\tilde{X}_i = x_u^*) = 1$. Similarly, let n_{2u} and n_{3u} denote the number of observations falling in the interval $(\min\{\tilde{\theta}_\alpha(x_u^*), \breve{\theta}_\alpha(x_u^*)\}, \max\{\tilde{\theta}_\alpha(x_u^*), \breve{\theta}_\alpha(x_u^*)\})$ when respectively $\tilde{\theta}_\alpha(x_u^*) > \breve{\theta}_\alpha(x_u^*)$ and $\tilde{\theta}_\alpha(x_u^*) < \breve{\theta}_\alpha(x_u^*)$ with $I(\tilde{X}_i = x_u^*) = 1$. Then, rewriting $L(h) - L^*$ and replacing Y_i by

$\bar{Y}_u = \frac{1}{2}(\tilde{\theta}_\alpha(x_u^*) + \breve{\theta}_\alpha(x_u^*))$, gives the loss function

$$\sum_{u=1}^{m}[n_{1u}(1-\alpha)\{\tilde{\theta}_\alpha(x_u^*) - \breve{\theta}_\alpha(x_u^*)\} + n_{2u}I(\tilde{\theta}_\alpha(x_u^*) > \breve{\theta}_\alpha(x_u^*))(\alpha\breve{\theta}_\alpha(x_u^*) - \bar{Y}_u + (1-\alpha)\tilde{\theta}_\alpha(x_u^*))$$
$$+n_{3u}I(\tilde{\theta}_\alpha(x_u^*) < \breve{\theta}_\alpha(x_u^*))(-(1-\alpha)\breve{\theta}_\alpha(x_u^*) - \alpha\tilde{\theta}_\alpha(x_u^*) + \bar{Y}_u) + n_{4u}\alpha\{\breve{\theta}_\alpha(x_u^*) - \tilde{\theta}_\alpha(x_u^*)\}]. \quad (5.2)$$

Now a practical band- and binwidth approach involves the following steps:

1. Consider the CV criterion

$$\mathrm{CV}(\delta) = \sum_{i=1}^{n}\rho_\alpha(Y_{(i)} - \breve{\theta}_\alpha^{(-i)}(x_u^*)),$$

   where $\breve{\theta}_\alpha^{(-j)}(\cdot)$ is an unsmoothed estimator of $\theta_\alpha(\cdot)$ given the sample $\{(\tilde{X}_i, R_i)|\, 1 \le i \le n, i \ne j\}$. Select a prespecified range of grid sizes M, giving rise to a set of binwidth values $\mathcal{D}$. Then, the binwidth is chosen as

$$\delta_{\mathrm{CV}} = \arg\min_{\delta\in\mathcal{D}}\{CV(\delta)\}.$$

2. Given δ_{CV} obtained in Step 1, the bandwidth parameter is chosen as to minimize (5.2).

3. Repeat Steps 1 and 2 with h_{CV} obtained from Step 2, and by replacing the unsmoothed estimator of the CDF in Step 1 with the binned kernel-smoothed conditional CDF (3.9).

6 Simulations

In this section we evaluate via simulation the performance of $\hat{\theta}_\alpha(\cdot)$ using uncorrelated and positively correlated uniformly distributed discrete bivariate random variables X and Y, using a bivariate Gaussian copula with a given correlation parameter ρ. Assume that X (Y) is distributed over the support $S_x = \{1, 2, \ldots, I_x\}$ ($S_y = \{1, 2, \ldots, I_y\}$), with I_x (I_y) a positive integer, according to the pmf $p(x) = 1/I_x$ ($p(y) = 1/I_y$) for $x \in S_x$ ($y \in S_y$) and zero elsewhere. The support of (X, Y) is $S_x \times S_y$. Then it may be deduced from results in Nelsen (1987) that the maximum feasible correlation between X and Y is given by $(I_y/I_x)\{(I_x^2-1)/(I_y^2-1)\}^{1/2}$. Hence, with $X_i \sim U[1, 10]$ and $Y_i \sim U[1, 100]$, the correlation parameter becomes 0.995. For each choice of n we generate $R = 10,000$ random samples, and for each n conditioning is based on each $\tilde{X} = x_u^*$ value. Values for ρ are set at $\rho = 0$, 0.2, 0.4, and 0.6. In practice, positive rather than negative or zero correlations are often observed between pairs (X, Y). This setup resembles the empirical application in Section 7, with a small number of distinct covariates as

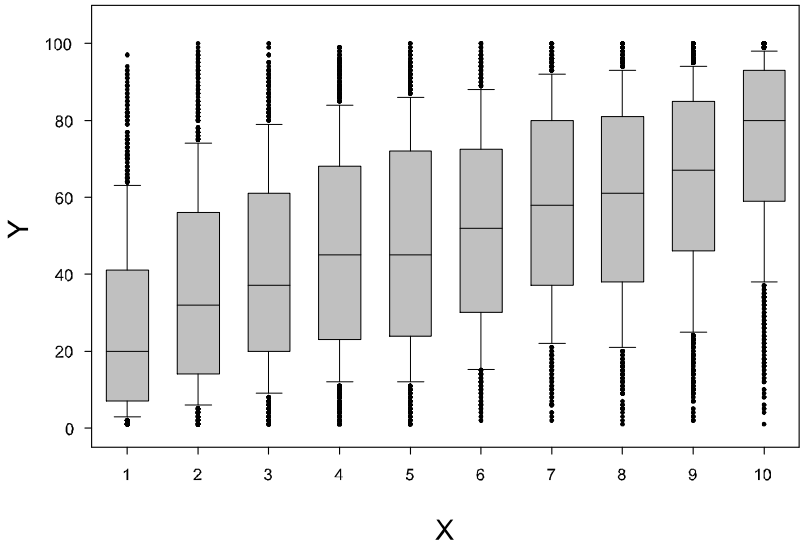

Figure 2. Boxplots of a simulated correlated dataset with $X \sim U[1, 10]$ *and* $Y \sim U[1, 100]$*;* $n = 10,000$*,* $\rho = 0.6$*.*

opposed to a relatively larger number of response values. Figure 2 shows boxplots of the conditional distribution of Y given values of X for a typical simulated dataset of size $n = 10,000$ with $\rho = 0.6$. It can be seen that the conditional distribution of Y is skewed to the right (left) for X taking values at the lower (higher) boundary region. For $\rho = 0.2$ skewness of the underlying conditional distribution is less pronounced, but still present.

6.1 Performance of the untransformed binned conditional quantile estimator

To abstract from the issue of binwidth selection, we evaluate the precision of $\hat{\theta}_\alpha(\cdot)$ by taking grid sizes $M = 30$, 40, and 50. So in Step 1 of the CV-algorithm δ will be held constant while in Step 2 we select the value of h_{CV} from the set $\{2, 3, \ldots, 20\}$. We only report simulation results for $\alpha = 0.5$ (median), because a pilot study showed that similar conclusions could be reached with other quantile values. Overall, the optimal CV-selected bandwidth value is not very sensitive to the choice of n with an maximum average (across all $R = 10,000$ replications) value of about 11 and a maximum average standard deviation of about 6. Assuming $\rho = 0$, $\theta_{0.5}(x_u^*) = 50$ for all values $x_u^* = \{1, 2, \ldots, 10\}$. For $\rho = 0.2$, 0.4 and 0.6, $\theta_{0.5}(\cdot)$'s are calculated on the basis of 10,000 independently sampled, but paired correlated uniformly distributed discrete random variables with samples of size 30,000. These "theoretical" quantiles are included in Table 3, columns 2 and 9.

Our measure of precision is the empirical mean squared error (MSE) of the untransformed

Table 2: Ratios of the empirical MSEs (averaged over $R = 10,000$ replications) of $\hat{\theta}_{0.5}(x_u^*)$ relative to the CQR estimator and, in parentheses, relative to the "naive" estimator; $X \sim U[1,10]$, $Y \sim U[1,100]$, $M = 40$.

x_u^*	$\rho = 0$			$\rho = 0.2$		
	$n = 5,000$	$n = 10,000$	$n = 20,000$	$n = 5,000$	$n = 10,000$	$n = 20,000$
1	2.44(0.95)	2.13(0.90)	1.81(0.85)	0.80(0.97)	0.49(0.95)	0.26(0.95)
2	3.24(0.94)	2.64(0.91)	2.14(0.86)	3.13(0.93)	2.58(0.89)	2.00(0.84)
3	4.21(0.94)	3.29(0.91)	2.39(0.86)	2.62(0.95)	1.73(0.94)	1.08(0.91)
4	5.20(0.94)	3.88(0.91)	2.62(0.85)	6.06(0.94)	4.92(0.90)	3.58(0.85)
5	6.05(0.94)	4.46(0.92)	2.62(0.85)	8.41(0.94)	6.94(0.91)	5.87(0.84)
6	5.90(0.95)	4.31(0.91)	2.67(0.85)	4.19(0.96)	2.68(0.91)	1.55(0.87)
7	4.96(0.94)	4.03(0.91)	2.47(0.86)	1.89(0.95)	1.14(0.90)	0.62(0.86)
8	4.15(0.93)	3.35(0.90)	2.37(0.86)	1.02(0.92)	0.61(0.87)	0.35(0.82)
9	3.21(0.93)	2.72(0.91)	2.05(0.85)	1.34(0.94)	0.80(0.92)	0.45(0.85)
10	2.45(0.94)	2.14(0.90)	1.79(0.86)	1.79(0.90)	1.39(0.85)	1.08(0.80)
	$\rho = 0.4$			$\rho = 0.6$		
1	0.30(0.96)	0.17(0.95)	0.09(0.93)	0.18(0.98)	0.10(0.99)	0.06(0.98)
2	1.05(0.86)	0.60(0.92)	0.34(0.86)	0.55(0.90)	0.31(0.83)	0.18(0.76)
3	0.87(0.93)	0.48(0.95)	0.26(0.92)	0.66(0.91)	0.36(0.84)	0.21(0.77)
4	1.77(0.96)	0.98(0.96)	0.54(0.94)	1.81(0.90)	1.01(0.86)	0.61(0.80)
5	9.31(0.95)	8.42(0.87)	8.06(0.80)	7.20(0.95)	4.31(0.92)	2.49(0.92)
6	4.21(0.90)	2.44(0.94)	1.40(0.94)	1.68(0.91)	0.98(0.86)	0.54(0.78)
7	0.73(0.84)	0.43(0.86)	0.24(0.78)	0.65(0.96)	0.33(0.94)	0.18(0.91)
8	0.44(0.95)	0.25(0.85)	0.14(0.79)	0.31(0.95)	0.16(0.93)	0.08(0.90)
9	0.48(0.94)	0.26(0.86)	0.15(0.81)	0.26(0.94)	0.14(0.91)	0.08(0.88)
10	0.65(0.92)	0.42(0.79)	0.28(0.70)	0.45(0.78)	0.32(0.70)	0.25(0.65)

conditional quantile estimates averaged over all replications. We consider the following two estimators: (a) the conditional quantile regression (CQR) estimator of Machado and Santos Silva (2005), for discrete-continuous data; (b) the unsmoothed "naive" conditional quantile estimator, i.e. grouping the data according to the values of the covariates and assuming that the underlying distribution for each resulting dataset is continuous. Thus the numbers in Table 2 represent the relative improvement attained by $\hat{\theta}_{0.5}(\cdot)$ relative to the CQR estimator and, in parentheses, the relative improvement attained by $\hat{\theta}_{0.5}(\cdot)$ relative to the "naive" estimator. To conserve space, Table 2 only contains results for $M = 40$.

The results show that across all covariates X and sample sizes n, the CQR estimator achieves the lowest empirical MSE when $\rho = 0$. However, when both ρ and n increase in value, the $\hat{\theta}_{0.5}(\cdot)$ estimator performs markedly better than the CQR estimator, apart for covariate values in the range [4, 6]. In particular, the MSE of the CQR estimator is dominated by increased positive bias when X takes values at the lower boundary region. As an example, Figure 3 shows the bias of both conditional quantile estimators for $n = 5,000$. We see that for all values ρ and $x_u^* \neq 5$, the absolute bias of the CQR estimator is substantially higher than the absolute bias of

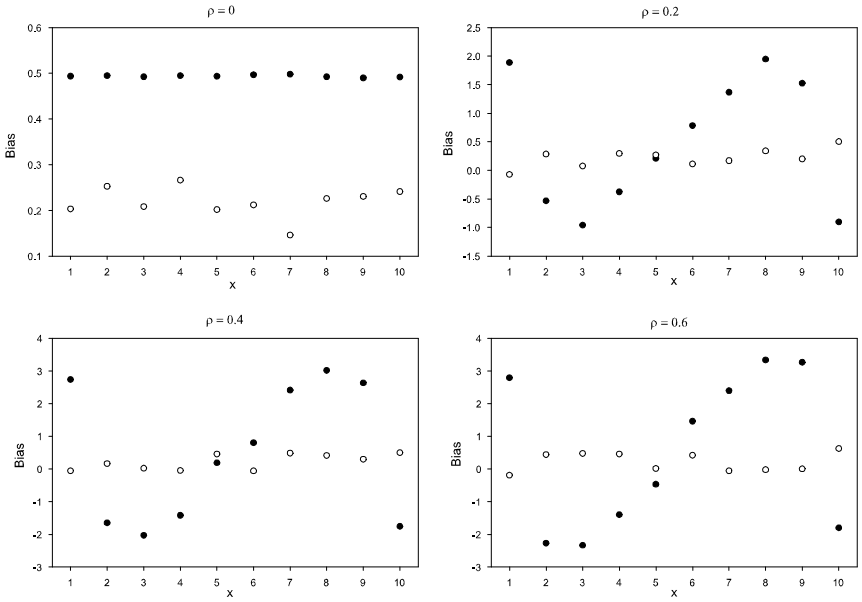

Figure 3. Bias of the CQR estimator (black circles) and the binned conditional quantile estimator $\hat{\theta}_{0.5}(\cdot)$ *(white circles) averaged over* $R = 10,000$ *replications;* $X \sim U[1,10]$*,* $Y \sim U[1,100]$*,* $M = 40$*,* $n = 5,000$*.*

$\hat{\theta}_{0.5}(\cdot)$. At $x_u^* = 5$, however, the difference between both estimators is minimal, suggesting that the CQR estimator is only feasible if the conditional distribution of Y, given values of X, is not too skewed.

A much clearer picture about the usefulness of $\hat{\theta}_{0.5}(\cdot)$ emerges from the relative improvement attained by this estimator relative to the infeasible "naive" estimator. In each case, the numbers in parentheses in Table 2 show that $\hat{\theta}_{0.5}(\cdot)$ has lower empirical MSE than the "naive" estimator. In cases where $n = 5,000$, the difference is minor across almost all values of ρ. But, as n increases, the difference is quite substantial. The largest gains are for $\rho = 0.4$, and 0.6 at $x_u^* = 10$, when $n = 20,000$. Thus meaningful improvements in estimation efficiency can be achieved by using the untransformed binned conditional quantile estimator when the data are correlated and the sample size is large.

6.2 Confidence intervals and coverage probabilities

Theorems 3 and 6 can be used to construct large-sample confidence interval (CIs) for $\theta_\alpha(\cdot)$ at a given confidence level ξ. Here we focus on the practicality of the asymptotic result (4.1). For $K(\cdot)$ we select the Gaussian kernel which is widely recognized as having good asymptotic efficiency compared to the optimal one, the Epanechnikov kernel. There is no simple rule for choosing the bandwidth h. Its choice depends on the sample size, the value of the covariate,

Table 3: Theoretical conditional quantile values $\theta_{0.5}(\cdot)$, averages of coverage probabilities of $\hat{\theta}_{0.5}(\cdot)$ based on Theorem 3(i) ("True"), averages of bootstrapped coverage probabilities of $\hat{\theta}_{0.5}(\cdot)$ with corresponding averages of upper- and lower CI limits in squared brackets; $X \sim U[1, 10]$, $Y \sim U[1, 100]$, $M = 40$, $R = 1,000$, $B = 500$, $1 - \xi = 0.95$.

x_u^*	$\theta_{0.5}(\cdot)$	$\rho = 0$						$\theta_{0.5}(\cdot)$	$\rho = 0.2$					
		$n = 5,000$			$n = 10,000$				$n = 5,000$			$n = 10,000$		
		True	Boot	$S(\cdot)$	True	Boot	$S(\cdot)$		True	Boot	$S(\cdot)$	True	Boot	$S(\cdot)$
1	50	0.959	0.955	[45, 53]	0.973	0.975	[46, 52]	37	0.976	0.950	[32, 39]	0.922	0.962	[33, 38]
2	50	0.957	0.969	[45, 53]	0.973	0.950	[46. 52]	42	0.965	0.966	[37, 45]	0.979	0.953	[39, 44]
3	50	0.962	0.957	[45, 53]	0.972	0.980	[46, 52]	45	0.965	0.971	[40, 48]	0.978	0.977	[41, 47]
4	50	0.957	0.958	[45, 53]	0.973	0.970	[46, 52]	47	0.961	0.958	[42, 50]	0.976	0.977	[43, 49]
5	50	0.955	0.955	[45, 53]	0.977	0.953	[46, 52]	49	0.959	0.964	[44, 52]	0.975	0.951	[45, 51]
6	50	0.960	0.960	[45, 53]	0.972	0.967	[46, 52]	51	0.956	0.955	[46, 54]	0.979	0.955	[47, 53]
7	50	0.959	0.961	[45, 53]	0.973	0.956	[46, 52]	53	0.960	0.961	[48, 56]	0.977	0.962	[50, 56]
8	50	0.960	0.969	[45, 53]	0.974	0.974	[46, 52]	55	0.964	0.972	[50, 58]	0.974	0.977	[51, 57]
9	50	0.957	0.951	[45, 53]	0.972	0.950	[46, 52]	58	0.965	0.968	[53, 61]	0.979	0.957	[55, 61]
10	50	0.958	0.962	[44, 53]	0.975	0.958	[46, 52]	63	0.972	0.974	[59, 66]	0.979	0.977	[60, 65]
		$\rho = 0.4$							$\rho = 0.6$					
1	27	0.977	0.974	[24, 30]	0.975	0.965	[24, 28]	21	0.967	0.963	[19, 23]	0.918	0.973	[18, 21]
2	36	0.932	0.979	[32, 40]	0.936	0.962	[32, 37]	32	0.954	0.961	[29, 35]	0.949	0.956	[29, 33]
3	41	0.973	0.975	[37, 45]	0.919	0.953	[38, 44]	38	0.927	0.968	[35, 42]	0.920	0.972	[35, 40]
4	45	0.968	0.972	[41, 49]	0.983	0.968	[40, 46]	43	0.973	0.969	[39, 47]	0.914	0.976	[40, 45]
5	48	0.962	0.967	[44, 52]	0.974	0.967	[45, 51]	48	0.974	0.962	[44, 52]	0.985	0.962	[45, 51]
6	52	0.963	0.953	[47, 55]	0.981	0.954	[47, 53]	52	0.971	0.973	[48, 56]	0.979	0.961	[49, 54]
7	55	0.962	0.968	[51, 59]	0.977	0.955	[52, 57]	57	0.980	0.968	[53, 61]	0.923	0.963	[53, 58]
8	59	0.969	0.976	[55, 63]	0.979	0.960	[56, 61]	62	0.934	0.952	[58, 65]	0.939	0.966	[59, 64]
9	64	0.929	0.983	[60, 68]	0.925	0.971	[61, 66]	68	0.959	0.956	[65, 71]	0.960	0.952	[65, 69]
10	73	0.967	0.966	[70, 76]	0.955	0.958	[70, 74]	79	0.940	0.953	[76, 81]	0.836	0.979	[77, 80]

and the shape of the empirical (a)symmetric distribution of the data. We simply replace h by $\hat{h} = Cn^{-1/5}\sigma_{Y|X}/\log(n)$ with $C = 0.01, 0.02, \ldots, 1$ a set of constants, and with $\sigma_{Y|X}$ the conditional variance of $Y|X = x$. This choice of h satisfies Condition A.4.

Note that for $\rho = 0$, we have $\sigma_{Y|X} = \sqrt{(I_y^2 - 1)/12}$. For $\rho \neq 0$, the "theoretical" values of $\sigma_{Y|X}$, $p(x)$ and $p(x, y)$ are, as before, based on $R = 10,000$ independently sampled, but paired correlated uniformly distributed discrete random variables with samples of size 30,000. Let z_ξ denote the $100(1 - \xi)$ percentile of the standard normal distribution. Now for each pair $(x_u^*, \theta_\alpha(x_u^*))$ $(u = 1, \ldots, 10)$, correlation parameter ρ, sample size n, and value C, the performance of the CIs are assessed by recording the proportion of times, using $R = 10,000$ replications, the theoretical conditional quantiles $\theta_\alpha(\cdot)$ are contained in the interval

$$\hat{\theta}_\alpha(x_u^*) \pm z_\xi \sqrt{\alpha(1-\alpha)p(x_u^*) \int K^2(v)dv / p^2(x_u^*, \theta_\alpha(x_u^*))} \Big/ \sqrt{n\hat{h}}. \tag{6.1}$$

Let $\hat{P}(\theta_\alpha(x_u^*))$ $(u = 1, \ldots, 10)$ denote the resulting proportion of times $\theta_\alpha(\cdot)$ falls in the interval (6.1) for all combinations of ρ, n, and C. Then, for each ρ and n, the "best" theoretical cov-

erage probability (denoted by "True" in Table 3) is chosen as $\min_{C\in[0.01,\,1.00]} \sum_{u=1}^{10} |\hat{P}(\theta_\alpha(x_u^*)) - 0.95|$, with $\alpha = 0.5$. The corresponding values of C vary between the ranges [0.37, 0.54] ($\rho = 0$, $n = 5,000$) and [0.63, 0.65] ($\rho = 0.6$, $n = 10,000$). On balance, the coverage probabilities are satisfactory, with values very close to the nominal level for all values of ρ and n. Based on the above results, we conclude that a value of C in the range [0.47, 0.57] may result in good theoretical coverage probabilities across all values of ρ and n.

An alternative approach to compute empirical coverage probabilities and CIs is to use the bootstrap. Guerra, Polansky and Schucany (1997) introduced this method for discrete univariate datasets. Here, as an example, we employ the sequence of bootstrap realizations of $\hat{\theta}_\alpha(x_u^*)$, $\{\hat{\theta}_\alpha^{\dagger b}(x_u^*)\}_{b=1}^B$, to estimate $P(\hat{\theta}_\alpha(x_u^*) \le \theta_\alpha(x_u^*))$. We can do so by defining the bootstrap statistic

$$\widehat{P}(\hat{\theta}_\alpha^\dagger(x_u^*) \le \theta_\alpha(x_u^*)) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}_\alpha^{\dagger b}(x_u^*) \le \hat{\theta}_\alpha(x_u^*)). \tag{6.2}$$

Thus a two-sided $(1-\xi)$ CI may be based on finding the largest $\theta_\alpha^U(x_u^*)$ and smallest $\theta_\alpha^L(x_u^*)$ in a fixed and countable set of estimated conditional quantiles such that both $P(\hat{\theta}_\alpha^\dagger(x_u^*) \le \theta_\alpha^L(x_u^*)) \le \xi/2$ and $P(\hat{\theta}_\alpha^\dagger(x_u^*) \le \theta_\alpha^U(x_u^*)) \ge 1 - \xi/2$ hold. In other words, given the discrete nature of the conditional quantiles, the set $S(x_u^*) = [\theta_\alpha^L(x_u^*), \theta_\alpha^U(x_u^*)]$ is the narrowest interval such that $P(\hat{\theta}_\alpha^\dagger(x_u^*) \in S(x_u^*)) \ge 1 - \xi$.

The bootstrapping takes place as follows. At each simulation run (i) generate a random sample (X_i^*, Y_i^*) of size n with replacement from the empirical joint distribution of (X, Y), (ii) compute $\hat{\theta}_\alpha(\cdot)$, (iii) compute (6.2) based on $B = 500$ bootstrap samples drawn from the original sample, (iv) calculate the lower and upper CI bounds, using a nominal level ξ and (6.2). Repeat this procedure over 1,000 independent runs. To save space, we only report results for the parameter configurations listed in Table 2.

Table 3 summarizes the averages of the empirical coverage probabilities obtained from (6.2) and averaged over 1,000 runs, and averages of the confidence sets $S(\cdot)$ (given in squared brackets). Recall that, by construction, we require that the empirical bootstrapped coverage probabilities are ≥ 0.95. Nonetheless, we may conclude that the empirical coverages are quite close to the nominal coverage probability. In addition, in all cases the CIs contain the true conditional quantile value. As expected, the length of $S(\cdot)$ decreases as n increases from $n = 5,000$ to $n = 10,000$, from a maximum average of about 9 to 6. In that case the minimum average length of $S(\cdot)$ goes down from about 8 to 3. If we increase n to 25,000, with $\rho = 0.6$, the average CI encompasses approximately 2 or 3 observations. Hence, the simulation evidence for $\hat{\theta}_{0.5}(\cdot)$ confirms the implications of the asymptotic results in Theorem 3. To provide a contrast, we also

computed CIs for the rank-transformed conditional quantile estimator $\tilde{\theta}_{0.5}(\cdot)$. In all cases the interval lengths are longer, containing on average 2 additional observations. From Theorem 6 we know that $\tilde{\theta}_{\alpha}(\cdot)$ is less efficient, in the sense that its variance is bigger than the asymptotic variance of $\hat{\theta}_{\alpha}(\cdot)$. Our simulation results confirm this result.

7 Empirical Illustration

We illustrate the method with an application using a large US database about hospital patients that have congestive heart failure and are transferred to a tertiary facility. About 96% of these patients are living in New York state. The total dataset contains information on all medical conditions being identified by the medical staff for each of the years 2003–2005 ($n = 20,631$ observations). We assume that the dataset has the nature of a population, and hence empirical conditional quantiles will be considered as population quantities. One key response variable Y is "Length of stay" (Los), measured in days. This variable is used for management of hospital care, quality control, appropriateness of hospital use, and hospital planning. A suitable covariate X is "Severity", measured on a 7-point scale. This latter variable is largely a function of how complex the initial condition of a patient is. Increments of one are added for each confounding condition presented. Of course, if two conditions are present and both are life threatening the result could be four, but this is unusual. Similarly, a secondary condition like diabetes may not raise the variable "Severity" if it is generally associated with the primary condition like a heart attack. Figure 4 shows conditional frequency distributions for each covariate. Skewness (kurtosis) ranges between 2.9 (15.6) at $x = 3$ and 6.0 (83.9) at $x = 4$. The probability masses are fairly concentrated around the values of the conditional median, with large outliers at $x = 6$ and $x = 7$.

The empirical illustration of the performance of $\hat{\theta}_{\alpha}(\cdot)$ and $\tilde{\theta}_{\alpha}(\cdot)$ is based on $R = 10,000$ random samples of size $n = 5,000$ drawn without replacement from the full dataset. For both $\hat{\theta}_{\alpha}(\cdot)$ and $\tilde{\theta}_{\alpha}(\cdot)$ the optimal bandwidth values were chosen using the CV-criterion in Subsection 5.2 with M running from 55 up to 200, and with h_{CV} in the range [2, 31]. The accuracy of each estimator is assessed by the average bias computed over the entire range of conditional quantile estimates. Table 4 shows these result for the first three conditional quartiles. Columns 2, 5, and 8 contain the "population" conditional quantiles. For this particular dataset these latter quantile estimates are the same as the values obtained by the "naive" quantile estimator, and plotted in Figure 1.

Note that across all values of x and α, the conditional quantile estimators perform well in

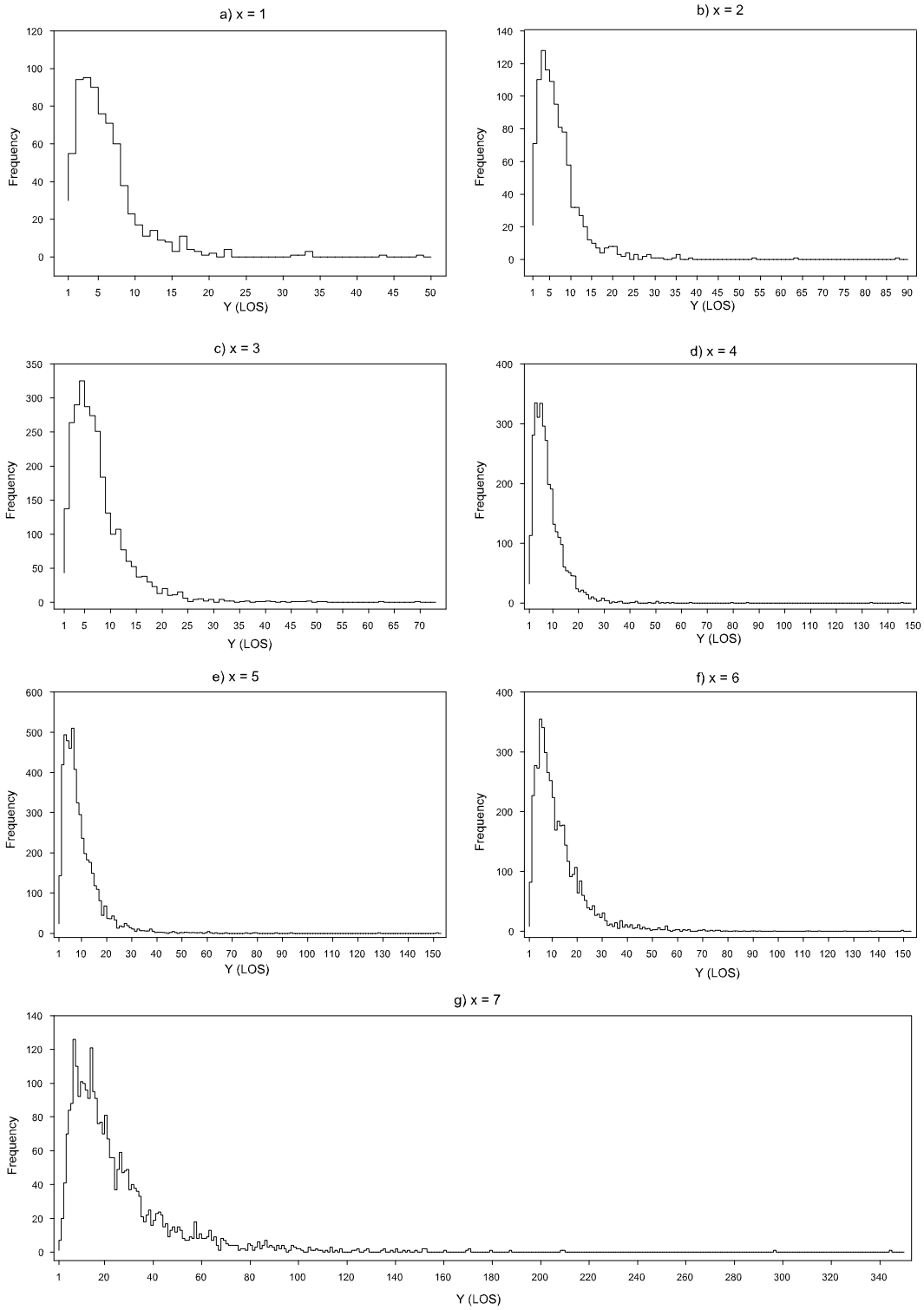

Figure 4. Conditional frequency distributions of Y "Length of stay" (LOS) as a function of the covariate X "Severity".

Table 4: Average bias for $\hat{\theta}_\alpha(\cdot)$ and, in parentheses, $\tilde{\theta}_\alpha(\cdot)$; $n = 5,000$.

x	$\theta_\alpha(\cdot)$	$\alpha = 0.25$	$\theta_\alpha(\cdot)$	$\alpha = 0.50$	$\theta_\alpha(\cdot)$	$\alpha = 0.75$
1	4	-0.658 (-0.506)	5	0.356 (0.439)	8	-0.015 (0.005)
2	4	-0.102 (0.002)	6	0.123 (0.166)	9	0.241 (0.247)
3	4	0.124 (0.180)	7	-0.338 (-0.105)	10	-0.045 (-0.046)
4	5	-0.439 (-0.173)	7	0.035 (0.040)	11	-0.049 (-0.029)
5	5	-0.047 (-0.000)	8	-0.213 (-0.055)	12	0.220 (0.243)
6	6	0.056 (0.054)	10	0.027 (0.045)	16	0.337 (0.303)
7	11	0.152 (0.227)	19	-0.251 (-0.287)	32	-0.260 (-0.829)

terms of the highest percentage of zero bias. The effects on the bias of the estimators vary substantially with different values of the covariate X and quartiles α, showing both underestimation and overestimation. As an example, consider the case $x = 1$. At $\alpha = 0.75$ we see that both conditional quantile estimators have good performance, but this is not the case at $\alpha = 0.25$. On the other hand, at $x = 7$ and $\alpha = 0.75$ the ranked-based conditional quantile estimator performs worse than the untransformed conditional quantile estimator. Clearly, the shape of the conditional distribution has an impact on the performance of the estimators. However, removing five outlying observations at $x = 7$ from the sample, we noticed no improvement in the bias results. Basically, the above observations also apply to samples of size $n = 10,000$. Although in the latter case the bias of the conditional quantile estimators is in general smaller than the bias of these estimators when $n = 5,000$.

When we estimated the full dataset, without resampling, both conditional quantile estimators showed zero bias results for almost all values of x and α. For $\hat{\theta}_\alpha(\cdot)$ the agreement between "population" and estimation results was less good at $x = 1$, $\alpha = 0.50$, at $x = 2$, $\alpha = 0.75$, and at $x = 6$, $\alpha = 0.75$. In all cases we noted a positive bias of one observation. For $\tilde{\theta}_\alpha(\cdot)$ we obtained a negative bias of one observation at $x = 1$, $\alpha = 0.25$, and at $x = 31$, $\alpha = 0.75$.

8 Conclusions

In this paper, we offer two kernel-based methods for estimating conditional quantiles of pairwise correlated and discretely distributed random variables. Both methods make use of the local structure through pre-binning the data. Since grid counts need to be computed only once for the life of a dataset, large gains in computational efficiencies can be achieved. In particular, jointly with the introduction of a practical band- and binwidth selection procedure, we showed that our kernel methods can be applied to the large datasets that appear in real-world problems.

The results from the simulations show that the untransformed binned conditional quan-

tile estimator achieves excellent estimation accuracy in terms of bias, MSE, and CI coverage. However, for long-tailed right skewed distributions the rank-transformed conditional quantile estimator also produces good nonparametric estimates. The rank-transform has the advantage of distributing the data more evenly across the set of bins. So the estimation error will be less large in some areas than with untransformed data, and hence will offset the increase in asymptotic variance of the rank-based conditional quantile estimator as opposed to the untransformed estimator. In addition, compared to the CQR estimator and the "naive" estimator the gains in using the binned conditional quantile estimators can be considerable when dealing with large correlated bivariate discrete datasets. Moreover, our explicit asymptotic CIs on both conditional quantile estimates can be used for quite quick computation of the intervals.

Appendix: Proofs

For ease of notation, we write $x \equiv x_u^*$ throughout the appendix.

Proof of Theorem 1. Let $\hat{F}_n(y|x)$ be the Nadaraya-Waston estimator of $\tilde{F}(y|x)$ based on the original data Z_i's, i.e.

$$\hat{F}_n(y|x) = \frac{\sum_{i=1}^n K((x-X_i)/h) I(Z_i \le y)}{\sum_{i=1}^n K((x-X_i)/h)}, \quad y \in \mathbb{R}.$$

Since $\tilde{F}(y|x) = F(y|x)$ and $\tilde{F}_n(y|x) = F_n(y|x)$ for all $y \in \mathbb{N}_1$, we have

$$(F_n(y|x) - F(y|x))^2 = (\tilde{F}_n(y|x) - \tilde{F}(y|x))^2$$

$$\le (\tilde{F}_n(y|x) - \hat{F}_n(y|x))^2 + 2|\tilde{F}_n(y|x) - \hat{F}_n(y|x)||\hat{F}_n(y|x) - F(y|x)| + (\hat{F}_n(y|x) - F(y|x))^2.$$

By (A.1) and Theorem 1 in Stute (1986), $\sup_y |\hat{F}_n(y|x) - F(y|x)| \stackrel{a.s.}{\to} 0$, which implies

$$\sup_y E(\hat{F}_n(y|x) - F(y|x))^2 \to 0, \quad \text{and} \quad \sup_y E\Big(|\tilde{F}_n(y|x) - \hat{F}_n(y|x)||\hat{F}_n(y|x) - F(y|x)|\Big) \to 0.$$

Note the definition of $c_{u,\ell}(z|x)$ does not depend on the ordered data, i.e.

$$c_{u,\ell}(z|x) = \sum_{i=1}^n I(X_i = x)(1 - |\delta^{-1} Z_i - \ell|)^+.$$

So,

$$(\tilde{F}_n(y|x) - \hat{F}_n(y|x))^2 = \frac{(\tilde{L}_n - \hat{L}_n)^2}{[(nh)^{-1}\sum_{i=1}^n K((x-X_i)/h)]^2}$$

$$:= \frac{\Big((nh)^{-1}\sum_{i=1}^n K((x-X_i)/h) C_{i,u}(z|x) - (nh)^{-1}\sum_{i=1}^n K((x-X_i)/h) I(Z_i \le y)\Big)^2}{[(nh)^{-1}\sum_{i=1}^n K((x-X_i)/h)]^2}.$$

Since $(nh)^{-1}\sum_{i=1}^n K((x-X_i)/h) \stackrel{a.s.}{\to} P(x) > 0$, so $(nh)^{-1}\sum_{i=1}^n K((x-X_i)/h) > P(x)/2$ (a.s.) for large n. In $\tilde{L}_n$, the X_i's are the original data, let L_n be the counter part of $\tilde{L}_n$ based on the binned X_i's with binwidth δ, then $E(\tilde{L}_n - \hat{L}_n)^2 \le E(L_n - \hat{L}_n)^2$.

Thus under A.2 and A.3, by the same way as in Theorem 1 of González-Manteiga (1996), we have

$$E(\tilde{F}_n(y|x) - \tilde{F}(y|x))^2 \le \frac{4}{P^2(x)} E(\tilde{L}_n - \hat{L}_n)^2 \le \frac{4}{P^2(x)} E(L_n - \hat{L}_n)^2 = \frac{4}{P^2(x)}[O(h^2) + o(\delta^2)],$$

the right hand side above is independent of y. These give the desired result.

Proof of Theorem 2. (i) For $\epsilon > 0$, we have

$$\tilde{F}(\theta_\alpha^c(x) - \epsilon|x) < \alpha < \tilde{F}(\theta_\alpha^c(x) + \epsilon|x).$$

Let S be the set of all events on which the results in Theorem 1 hold. Then for large n, on S we have

$$\tilde{F}_n(\theta_\alpha^c(x) - \epsilon|x) < \alpha < \tilde{F}_n(\theta_\alpha^c(x) + \epsilon|x).$$

Since $\tilde{F}(z|x) \ge t$ iff $z \ge \tilde{F}^{-1}(t|x)$, we have, on S for all large n,

$$\theta_\alpha^c(x) - \epsilon \le \hat{\theta}_\alpha^c(x) \le \theta_\alpha^c(x) + \epsilon.$$

Since $\hat{\theta}_\alpha^c(x)$ is bounded on S for large n, and $\epsilon > 0$ is arbitrary, the conclusion is true.

(ii) It is a consequence of (i), as in the discussion of (ii) before Condition A.1.

Proof of Theorem 3. (i) Let $\hat{F}_n(y|x)$ as given in the proof of Theorem 1. Note A.4 implies A.1 and A.2, so as in the proof of Theorem 1, $\tilde{F}_n(y|x) = \hat{F}_n(y|x) + o_P((hn)^{-1/2})$, thus we have

$$P\bigg(\sqrt{nh}(\hat{\theta}_\alpha^c(x) - \theta_\alpha^c(x)) \le t\bigg) = P\bigg(\hat{\theta}_\alpha^c(x) \le \theta_\alpha^c(x) + (nh)^{-1/2}t\bigg)$$

$$= P\bigg(\tilde{F}_n(\theta_\alpha^c(x) + (nh)^{-1/2}t|x) \ge \alpha\bigg) = P\bigg(\hat{F}_n(\theta_\alpha^c(x) + (nh)^{-1/2}t|x) \ge \alpha + o((hn)^{-1/2})\bigg)$$

$$= P\bigg(\frac{1}{nh}\sum_{i=1}^n K(\frac{x-X_i}{h}) I(Z_i \le \theta_\alpha^c(x) + (nh)^{-1/2}t) \ge \alpha \frac{1}{nh}\sum_{i=1}^n K(\frac{x-X_i}{h}) + o((hn)^{-1/2})\bigg)$$

$$= P\bigg(\frac{1}{n}\sum_{i=1}^n \frac{1}{h} K(\frac{x-X_i}{h})[I(Z_i \le \theta_\alpha^c(x) + (nh)^{-1/2}t) - \alpha] \ge o((hn)^{-1/2})\bigg)$$

$$:= P\bigg(\frac{1}{n}\sum_{i=1}^n A_{n,i} \ge o((hn)^{-1/2})\bigg).$$

Note by A.4, $nh \to \infty$ and $h = o((nh)^{-1/2})$. Also, $P(Z < \theta^c_\alpha(x)|x) = \alpha$, so

$$\begin{aligned} E(A_{n,i}) &= \frac{1}{h}\int\int K(\frac{y-x}{h})[I(z < \theta^c_\alpha(x) + (nh)^{-1/2}t) - \alpha]F(dy, dz) \\ &= \int\int K(v)[I(z < \theta^c_\alpha(x) + (nh)^{-1/2}t) - \alpha]p(x + hv, dz)dv \\ &= \int [I(z < \theta^c_\alpha(x) + (nh)^{-1/2}t) - \alpha]p(x, dz)(1 + O(h)) \\ &= P(Z < \theta^c_\alpha(x)|x)p(x) + p(x, \theta^c_\alpha(x))(nh)^{-1/2}t - \alpha p(x) + o((nh)^{-1/2}) \\ &= p(x, \theta^c_\alpha(x))(nh)^{-1/2}t + o((nh)^{-1/2}), \end{aligned}$$

also,

$$Var(A_{n,i}) = E(A^2_{n,i}) - E^2(A_{n,i}) \sim E(A^2_{n,i})$$

$$= h^{-1}\int\int K^2(v)[I(z < \theta^c_\alpha(x) + (nh)^{-1/2}t) - 2\alpha I(z < \theta^c_\alpha(x) + (nh)^{-1/2}t) + \alpha^2]p(x + hv, dz)dv$$

$$= \frac{1}{h}\alpha(1-\alpha)p(x)\int K^2(v)dv + o(h^{-1}) := \frac{1}{h}\sigma_0^2 + o(h^{-1}).$$

Thus by the Central limit theorem, we get

$$P\bigg(\frac{1}{n}\sum_{i=1}^n A_{n,i} \geq o((hn)^{-1/2})\bigg)$$

$$= P\bigg(-\sqrt{nh}\sigma_0^{-1}\frac{1}{n}\sum_{i=1}^n (A_{n,i} - E(A_{n,i})) \leq \sigma_0^{-1}p(x, \theta^c_\alpha(x))t + o(1)\bigg) \to \Phi(\sigma_0^{-1}p(x, \theta^c_\alpha(x))t),$$

where $\Phi(\cdot)$ is the distribution function of $N(0,1)$, i.e.

$$\sqrt{nh}(\hat{\theta}^c_\alpha(x) - \theta^c_\alpha(x)) \xrightarrow{D} N(0, \sigma^2)$$

with $\sigma^2 = \alpha(1-\alpha)p(x)\int K^2(v)dv/p^2(x, \theta^c_\alpha(x))$.

(ii) In this case, for $t > 0$, $E(A_{n,i}) = p_k p(\theta^c_\alpha(x)|x)(nh)^{-1/2}t + o((nh)^{-1/2})$, and $Var(A_{n,i}) \sim h^{-1}\alpha(1-\alpha)p_k\int K^2(v)dv$; and for $t < 0$, $E(A_{n,i}) = p_{k-1}p(\theta^c_\alpha(x)|x)(nh)^{-1/2}t$, and $Var(A_{n,i}) \sim h^{-1}\alpha(1-\alpha)p_{k-1}\int K^2(v)dv$. The rest of the proof is the same as in (i).

(iii) This is a direct result from the discussion before Theorem 1.

Proof of Corollary 1. By the given condition we have $\hat{p}_k \to p_k$ (a.s.), $\hat{p}(x) \to p(x)$ (a.s.) and by Theorem 2, $\hat{\theta}^c_\alpha(x) \xrightarrow{L_2} \theta^c_\alpha(x)$. The above convergences are all stronger than those in probability. Thus the results in Theorem 3 hold by Slutsky's Theorem.

Proof of Theorem 4. We only prove consistency for $F^R_n(y|x)$ to $F(y|x)$, that for $\tilde{F}^R_n(y|x)$ follows as the comment before Theorem 4. Rewrite $F^R_n(y|x) := F^R_n(v|j)$, for R_v being the

smallest rank such that $Y_{(R_v)} \leq y$, and with $\tilde{X}_{(R_v)} = j$, as the standard form of a linear rank statistic

$$F_n^R(v|j) = \sum_{i=1}^n c_{n,i} a(R_i)$$

with,

$$c_{n,i} = \sum_{u=1}^n \frac{1}{n_u} I(X_i = u) w_{n,u}, \quad w_{n,u} = \frac{K((j-u)/h)}{\sum_{u=1}^n K((j-u)/h)}, \quad a(R_i) = \sum_{\ell=1}^v (1 - |\delta^{-1} R_i - (\ell-1)|)^+.$$

In the above we used the fact that $\sum_{u=1}^m n_u K((j-u)/h) = \sum_{u=1}^n K((j-u)/h)$.

We will use the method and results in Hájek, Šidák and Sen (1999). Let

$$\overline{c}_n = \frac{1}{n} \sum_{i=1}^n c_{n,i}, \quad \overline{a} = \frac{1}{n} \sum_{i=1}^n a(i), \quad \sigma_a^2 = \frac{1}{n-1} \sum_{i=1}^n (a(i) - \overline{a})^2.$$

Note $(1 - |\delta^{-1} i - \ell|)^+ = [\delta - (\delta(\ell-1) - i)]/\delta$ if $\delta(\ell-2) \leq i < \delta(\ell-1)$; $= [\delta - (i - \delta(\ell-1))]/\delta$ if $\delta(\ell-1) \leq i < \delta\ell$; and $= 0$ elsewhere. So if $i \in ((k-1)\delta, (k+1)\delta)$ for some $k \leq v$, then $a(i) = \sum_{l=1}^v (1 - |\delta^{-1} i - \ell|)^+ = [(k\delta - i) + (i - (k-1)\delta)]/\delta = 1 = a^2(i)$, and $a(i) = 0$ for $i > v$. We have

$$\overline{a} = \frac{v}{n} + O(\frac{1}{n}), \qquad \sigma_a^2 = \frac{v}{n}(1 - \frac{v}{n}) + O(\frac{v}{n^2}),$$

$$\sum_{i=1}^n c_{n,i} = \sum_{u=1}^n w_{n,u} \frac{1}{n_u} \sum_{i=1}^n I(X_i = u) = \sum_{u=1}^n w_{n,u} = 1.$$

Since R_v is the smallest rank with $Y_{(R_v)} \leq y$ and with $\tilde{X}_{(R_v)} = x$, v is the number of the Y_i's with $Y_i \leq y$ and with $X_i = x$, so we have

$$\overline{a} = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq y | X_i = x) + O(1/n) \stackrel{a.s.}{\to} F(y|x), \quad \sigma_a^2 \stackrel{a.s.}{\to} F(y|x)(1 - F(y|x)).$$

Also, with $j = x$,

$$\begin{aligned} \sum_{i=1}^n (c_{n,i} - \overline{c})^2 &= \sum_{u=1}^n w_{n,u}^2 \frac{1}{n_u^2} (\sum_{i=1}^n I(X_i = u))^2 - \overline{c}^2 = \sum_{u=1}^n w_{n,u}^2 - 1/n \\ &= (nh)^{-1} \frac{(nh)^{-1} \sum_{u=1}^n K^2((j-u)/h)}{\left((nh)^{-1} \sum_{u=1}^n K((j-u)/h)\right)^2} - 1/n. \end{aligned}$$

By standard results on kernel estimation we have

$$(nh)^{-1} \sum_{u=1}^n K((j-u)/h) = p(x) + o(1), \quad (nh)^{-1} \sum_{u=1}^n K^2((j-u)/h) = p(x) \int K^2(v) dv + o(1).$$

So we get, as $1/n = o((nh)^{-1})$,

$$\sum_{i=1}^n (c_{n,i} - \overline{c})^2 = (nh)^{-1} \frac{p(x) \int K^2(v) dv + o(1)}{p^2(x) + o(1)} - 1/n = (nh)^{-1} p^{-1}(x) \int K^2(v) dv + o((nh)^{-1}).$$

Thus, by Theorem 3.3.3 in Hájek, Šidák and Sen (1999, p. 61),

$$E(F_n^R(v|j)) = E(\overline{a}\sum_{i=1}^{n} c_{n,i}) = E(\frac{v}{n}) + O(1/n) \to F(y|x),$$

and

$$Var(F_n^R(v|j)) = \sigma_a^2 \sum_{i=1}^{n}(c_{n,i} - \overline{c})^2 = (nh)^{-1}F(y|x)(1 - F(y|x))p^{-1}(x)\int K^2(v)dv + o((nh)^{-1}).$$

Now we have, $\forall \epsilon > 0$,

$$\begin{aligned} P\bigg(|F_n^R(y|x) - F(y|x)| \geq \epsilon\bigg) &\leq P\bigg(|F_n^R(y|x) - E[F_n^R(y|x)]| \geq \epsilon/2\bigg) \leq \frac{4}{\epsilon^2}Var(F_n^R(y|x)) \\ &\sim \frac{4}{\epsilon^2 nh}F(y|x)(1 - F(y|x))p^{-1}(x)\int K^2(v)dv \to 0. \end{aligned}$$

This complete the proof of Theorem 4.

Proof of Theorem 5. The proof is similar to that of Theorem 2.

Proof of Theorem 6. We use the notations in the proof of Theorem 4.

(i) Let

$$\varphi(t) = \begin{cases} 1, & t \leq \tilde{F}(y|x), \\ 0 & \text{otherwise.} \end{cases}$$

Then it is easy to check that

$$\lim_n \int_0^1 [a(1 + [tn]) - \varphi(t)]^2 dt = 0.$$

Also $\int_0^1 \varphi^2(t)dt < \infty$, $\int_0^1[\varphi(t) - \overline{\varphi}]^2 dt = \tilde{F}(y|x)(1 - \tilde{F}(y|x)) > 0$, where $\overline{\varphi} = \int_0^1 \varphi(t)dt = \tilde{F}(y|x)$. By definition we have $\tilde{F}_n^R(y|x) = F_n^R(y|x) + O(\delta)$, and $O(\delta) = o((nh)^{-1})$ by B.4. Thus by the liner rank statistic form of $F_n^R(y|x)$ as given in the proof of Theorem 4, Theorem 6.6.1 in Hájek, Šidák and Sen (1999, p. 194), and the relationship $\tilde{F}_n^R(y|x) = F_n^R(y|x) + o((nh)^{-1})$, we have that $\tilde{F}_n^R(y|x)$ is asymptotically normal (μ_n, σ_n^2) with $\mu_n = E\tilde{F}_n^R(y|x) = \overline{a} + o((nh)^{-1/2})$ and $\sigma_n^2 = [\sum_{u=1}^n (c_{n,u} - \overline{c})^2]\int_0^1[\varphi(t) - \overline{\varphi}]^2 dt$. Using results of $\overline{a}$ and $\sum_{u=1}^n (c_{n,u} - \overline{c})^2$ from the proof of Theorem 4, we get

$$\sqrt{nh}(\tilde{F}_n^R(y|x) - \tilde{F}(y|x)) \xrightarrow{D} N(0, \sigma_1^2), \text{ with } \sigma_{1u}^2 = \tilde{F}(y|x)(1 - \tilde{F}(y|x))p^{-1}(x)\int K^2(v)dv.$$

(ii) As in the proof of Theorem 3, here we have

$$P\bigg(\sqrt{nh}(\tilde{\theta}_\alpha(x) - \theta_\alpha(x)) \leq t\bigg) = P\bigg(\tilde{F}_n^R(\theta_\alpha(x) + (nh)^{-1/2}t|x) \geq \alpha\bigg).$$

As in the proof of Theorem 4, we have, with $p(y|x)$ being the density of $\tilde{F}(y|x)$,

$$\begin{aligned} E(\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x)) &= \tilde{F}(\theta_\alpha(x)+(nh)^{-1/2}t|x)+O(1/n) \\ &= \tilde{F}(\theta_\alpha(x)|x)+p(\theta_\alpha(x)|x)(nh)^{-1/2}t+o((nh)^{-1/2})+O(1/n) \\ &= \alpha+p(\theta_\alpha(x)|x)(nh)^{-1/2}t+o((nh)^{-1/2}). \end{aligned}$$

Also,

$$\begin{aligned} Var(\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x)) &= (nh)^{-1}\tilde{F}(\theta_\alpha(x)+(nh)^{-1/2}t|x) \\ &\quad \times(1-\tilde{F}(\theta_\alpha(x)+(nh)^{-1/2}t|x))p^{-1}(x)\int K^2(v)dv+o((nh)^{-1}) \\ &= (nh)^{-1}\alpha(1-\alpha)p^{-1}(x)\int K^2(v)dv+o((nh)^{-1}). \end{aligned}$$

Applying Theorem 6.6.1 in Hájek, Šidák and Sen (1999, p. 194) again to the linear rank statistic form of $\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x)$, as in the proof of Theorem 4, we get

$$P\bigg(\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x)\geq\alpha\bigg)=P\bigg(\frac{E(\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x))-\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t|x)}{\sqrt{Var(\tilde{F}_n^R(\theta_\alpha(x)+(nh)^{-1/2}t))}}$$

$$\leq\frac{p(\theta_\alpha(x)|x)t}{[\alpha(1-\alpha)p^{-1}(x)\int K^2(v)dv]^{1/2}}+o(1)\bigg)\to\Phi\bigg(\frac{p(x)p(\theta_\alpha(x)|x)}{[\alpha(1-\alpha)\int K^2(v)dv]^{1/2}}t\bigg),$$

or

$$\sqrt{nh}(\tilde{\theta}_\alpha(x)-\theta_\alpha(x))\xrightarrow{D}N(0,\sigma^2),\ \text{with}\ \sigma^2=\alpha(1-\alpha)\int K^2(v)dv/p^2(x,\theta_\alpha(x)).$$

(iii) and (iv) the proofs are similar to those of (ii) and (iii) in Theorem 3.

Proof of Corollary 2. Since by Theorems 4 and 5, $\tilde{F}_n^R(y|x)\xrightarrow{P}\tilde{F}(y|x)$, $\tilde{\theta}_\alpha^R(x)\xrightarrow{P}\theta_\alpha^c(x)$ and $\tilde{\theta}_\alpha(x)\xrightarrow{P}\theta_\alpha(x)$. The rest of the proof is similar to that of Corollary 1.

Acknowledgements The authors thank the associate editor and two referees for providing helpful comments.

References

Chen, J. and Lazar, N. (2010). Quantile estimation for discrete data via empirical likelihood. *J. Nonparametr. Stat.* **22**, 237-255.

Fan, J. and Marron, J.S. (1994). Fast implementation of nonparametric curve estimators. *J. Comput. Graph. Statist.* **13**, 35–56.

Frydman, H. and Simon, G. (2007). Discrete quantile estimation. Available at SSRN: http// ssm.com/abstract=1293142.

González-Barrios, J.M. and Rueda, R. (2001). On convergence theorems for quantiles. *Comm. Statist. Theory Methods* **30**, 943–955.

González-Manteiga, W., Sánchez-Sellero, C. and Wand, M.P. (1996). Accuracy of binned kernel functional approximations. *Comput. Statist. Data Anal.* **22**, 1–16.

Guerra, R., Polansky, A.M. and Schucany, W.R. (1997), Smoothed bootstrap confidence intervals with discrete data. *Comput. Statist. Data Anal.* **26**, 163–176.

Hájek, J., Šidák, Z., Sen, P.K. (1999). *Theory of Rank Tests* (second edition). Academic Press, San Diego.

Hall, P. and Wand, M.P. (1996). On the accuracy of binned kernel density estimators. *J. Multivariate Anal.* **56**, 165–184.

Holmström, L. (2000). The accuracy and computational complexity of a multivariate binned kernel density estimator. *J. Multivariate Anal.* **72**, 264–309.

Jones, M.C. and Lotwick, H.W. (1983). On the errors involved in computing the empirical characteristic function. *J. Statist. Comput. Simulation* **13**, 173–149.

Li, Q. and Racine, J.S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *J. Bus. Econom. Stat.* **26**, 423–434.

Machado, J.A. and Santos Silva, J.M.C. (2005). Quantiles for counts. *J. Amer. Statist. Assoc.* **100**, 1226–1237.

Magee, L., Burridge, J.B. and Robb, A.L. (1991). Computing kernel-smoothed conditional quantiles from many observations. *J. Amer. Statist. Assoc.* **86**, 673–677.

Nelsen, R.B. (1987). Discrete bivariate distributions with given marginals and correlation. *Comm. Statist. Simulation Comput.* **16**, 199–208.

Rajagopalan, R. and Lall, U. (1995). A kernel estimator for discrete distributions. *J. Nonparametr. Stat.* **4**, 409–426.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Stute, W. (1986). On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.* **11**, 891–901.