

Smooth nonparametric estimation of a quantile function under right censoring using beta kernels

CHANSEOK PARK¹

Department of Mathematical Sciences, Clemson University, Clemson, SC 29634

Short Title: Smooth nonparametric estimation of a quantile function

Smooth nonparametric estimators of a quantile function based on symmetric kernels suffer from spill-over around the boundaries which leads to boundary bias. Based on a beta probability density as a kernel function, a new nonparametric estimator of a quantile function under right censoring is proposed. This new quantile estimator is free of spill-over around the boundaries. Asymptotic properties of the proposed quantile estimator are studied. An illustrative example and Monte Carlo simulation results are presented to compare the proposed method with existing estimators and show the substantial improvement.

KEY WORDS: Kernel density estimation; Quantile; Percentile; Censoring; Asymmetric kernel; Bandwidth

1 INTRODUCTION

Let X_1, X_2, \dots, X_n be independent and identically distributed (iid) with distribution function $F(x)$ and let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ denote the corresponding order statistics. A version of the left-continuous quantile function is defined as $Q(p) = \inf\{t : F(t) \geq p\}$ for $0 < p < 1$. Then a conventional estimator of $Q(p)$ is the p th sample quantile of $F(\cdot)$ given by $X_{([np])}$ where $[np]$ is the largest integer not exceeding

¹Corresponding author. Email: cspark@ces.clemson.edu

np . For more details about this sample quantile, good references are David (1981) and Galambos (1978).

It is widely known that the sample quantile suffers from a substantial lack of efficiency. To remedy this problem, many authors proposed smooth alternatives to the sample quantile using kernel-type estimators. Early work on kernel-type estimators of the quantile function include Nadaraya (1964) and Parzen (1979). Reiss (1980) showed that the asymptotic relative deficiency of the sample quantile with respect to a linear combination of finitely many order statistics diverges to infinity as the sample size increases. Falk (1984) also examined the asymptotic relative deficiency of the sample quantile compared to kernel-type quantile estimators. Yang (1985) studied the asymptotic properties of kernel-type quantile estimators. Padgett (1986) extended the previous works to handle right-censored data. All of these results are based on symmetric kernel functions. Since the domain of the quantile function is a bounded interval $(0, 1)$, using symmetric kernel functions can lead to boundary bias or spill-over effects. Boundary bias is due to inappropriate weights of kernel functions around the boundaries of the quantile function when fixed symmetric kernels are used.

Chen (1999, 2000) proposed the use of beta kernel estimators for density functions and regression curves in order to avoid boundary bias. Considering that the support of a beta probability density function matches the domain of the quantile function, it is appropriate to incorporate a beta probability density into smooth non-parametric quantile function estimators. In this paper, we propose a new kernel-type quantile estimator under right censoring based on the beta probability density function, which is free of spill-over effects.

2 THE PRODUCT-LIMIT QUANTILE ESTIMATOR

In this section, we briefly introduce the product-limit quantile estimator under right censoring. Let T_1, T_2, \dots, T_n be iid true lifetimes of n individuals from the distribution $F(\cdot)$ with the pdf $f(\cdot)$. We assume that the lifetimes can possibly be censored on the right by iid random variables U_1, U_2, \dots, U_n from the distribution $H(\cdot)$ which are independent of the T_i 's. So, we only observe the right-censored data $X_i = \min(T_i, U_i)$ with the censoring indicator variable which is defined as

$$\Delta_i = \begin{cases} 1 & \text{if } T_i \leq U_i \\ 0 & \text{if } T_i > U_i \end{cases}.$$

The distribution of each X_i is then given by $G = 1 - (1 - F)(1 - H)$. To estimate the cdf of the true lifetime with the censored samples (X_i, Δ_i) , the product limit (PL) estimator, originally attributed to Kaplan and Meier (1958), is popularly used. Let $\hat{F}_n(\cdot)$ denote the empirical cdf $F(\cdot)$ with the empirical survival $\hat{S}_n(t) = 1 - \hat{F}_n(t)$. Then the product limit (PL) estimator of $S(t) = 1 - F(t)$ is defined as

$$\hat{S}_n(t) = \begin{cases} 1 & \text{if } t < X_{(1)} \\ \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1} \right)^{\Delta_{(i)}} & \text{if } X_{(k-1)} \leq t < X_{(k)}, \quad k = 2, \dots, n. \\ 0 & \text{if } t \geq X_{(n)} \end{cases}$$

Using this, the empirical cdf is given by $\hat{F}_n(t) = 1 - \hat{S}_n(t)$ and the empirical PL quantile function is defined as $\hat{Q}_n(p) = \hat{F}_n^{-1}(p) = \inf\{t : \hat{F}_n(t) \geq p\}$.

3 SMOOTH QUANTILE ESTIMATORS BASED ON KERNEL DENSITIES

In this section, we briefly describe the kernel-type quantile function estimator under right censoring which is proposed by Padgett (1986). Then we provide the proposed estimator along with its asymptotic properties.

Based on symmetric kernel functions, Padgett (1986) proposed the kernel-type quantile function defined as

$$\text{KQ}_n(p) = \int_0^1 \hat{F}_n^{-1}(t) \frac{1}{h_n} K\left(\frac{t-p}{h_n}\right) dt.$$

It is easily seen that the above becomes

$$\text{KQ}_n(p) = \sum_{i=1}^n X_{(i)} \int_{P_{i-1}}^{P_i} \frac{1}{h_n} K\left(\frac{t-p}{h_n}\right) dt = \sum_{i=1}^n X_{(i)} \left[L\left(\frac{P_i-p}{h_n}\right) - L\left(\frac{P_{i-1}-p}{h_n}\right) \right] dt, \quad (1)$$

where $L(u) = \int_{-\infty}^u K(t) dt$, $P_i = \hat{F}_n(X_{(i)})$ ($i = 1, 2, \dots, n$), and $P_0 = 0$. For more details, see Padgett (1986).

We propose the use of the nonparametric quantile estimator based on the beta probability density

$$\text{BQ}_n(p) = \int_0^1 \hat{F}_n^{-1}(t) \cdot K_\beta(t; p/b_n + 1, (1-p)/b_n + 1) dt,$$

where $K_\beta(t; a, b)$ is the beta probability density given by

$$K_\beta(t; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1} (1-t)^{b-1},$$

where $0 < t < 1$. For brevity, we denote $K_\beta(t; p) = K_\beta(t; p/b_n + 1, (1-p)/b_n + 1)$. Note that $K_\beta(t; p)$ has the mode at $t = p$. Using an approach similar to that used to derive (1), it is immediate that

$$\text{BQ}_n(p) = \sum_{i=1}^n X_{(i)} \left[L_\beta(P_i; p) - L_\beta(P_{i-1}; p) \right],$$

where $L_\beta(t; p)$ is the cdf of the beta pdf $K_\beta(t; p)$.

Theorem 1. Suppose that $F(\cdot)$ is twice differentiable at $\xi_p = Q(p)$ with $f(\xi_p) > 0$ and $\{b_n\}$ is such that $b_n \rightarrow 0$ as $n \rightarrow \infty$. Let $T_{H(F^{-1})} = \inf\{t : H(F^{-1}(t)) = 1\}$ and $p_0 = \min\{1, T_{H(F^{-1})}\}$. Then for $0 < p < p_0$, we have

$$\begin{aligned} \sqrt{n}\{\text{BQ}_n(p) - Q(p)\} &= -\sqrt{n}\frac{\hat{F}_n(\xi_p) - p}{f(\xi_p)} + n^{1/2}b_n(1 - 2p)Q'(p) + \frac{1}{2}n^{1/2}b_np(1 - p)Q''(p) \\ &\quad + o(n^{1/2}b_n) + o_p(1). \end{aligned}$$

Proof. Analogous to the setup in the beginning of the proof of Theorem 1 of Yang (1985), we write

$$\sqrt{n}\{\text{BQ}_n(p) - Q(p)\} = d_n(p) + A_n + B_n,$$

where

$$\begin{aligned} d_n(t) &= \sqrt{n}\{\hat{Q}_n(t) - Q(t)\} \\ A_n &= \int_0^1 \{d_n(t) - d_n(p)\}K_\beta(t; p)dt \\ B_n &= \sqrt{n} \int_0^1 \{Q(t) - Q(p)\}K_\beta(t; p)dt. \end{aligned}$$

From an almost-sure Bahadur-type representation (Bahadur, 1966) established by Cheng (1984), we have

$$d_n(p) \stackrel{a.s.}{=} -\sqrt{n}\frac{\hat{F}_n(\xi_p) - p}{f(\xi_p)} + O(n^{-1/4}(\log n)^{3/4}),$$

for $0 < p < p_0$; see also the equation (8.1.27) of Csörgő (1983). We write

$$A_n = a_1 + a_2 + a_3,$$

where for $0 < \delta < 1 - p$

$$\begin{aligned} a_1 &= \int_0^{p-\delta} \{d_n(t) - d_n(p)\}K_\beta(t; p)dt \\ a_2 &= \int_{p-\delta}^{p+\delta} \{d_n(t) - d_n(p)\}K_\beta(t; p)dt \\ a_3 &= \int_{p+\delta}^1 \{d_n(t) - d_n(p)\}K_\beta(t; p)dt. \end{aligned}$$

Then we have

$$a_1 \leq \sup_{t \in (0, p-\delta)} \left(|d_n(t) - d_n(p)| \int_0^{p-\delta} K_\beta(t; p) dt \right) = o_p(1),$$

and similarly $a_3 = o_p(1)$. It is easily seen that

$$a_2 \leq \sup_{t \in (p-\delta, p+\delta)} |d_n(t) - d_n(p)|.$$

Using an argument similar to that used in the proof of Lemma 2 of Lio *et al.* (1986), we have

$$a_2 \leq \sup_{t \in (p-\delta, p+\delta)} |d_n(t) - d_n(p)| = o_p(1)$$

and thus $A_n = o_p(1)$.

Using a Taylor series expansion of $Q(t)$ about p gives

$$Q(t) - Q(p) = Q'(p)(t - p) + \frac{1}{2}Q''(p)(t - p)^2 + R_n(p),$$

where the error term $R_n(p)$ is expressed as an integral (see Theorem 9.29 of Apostol (1974))

$$R_n(t) = \frac{1}{2} \int_p^t (t - u)^2 Q'''(u) du = \int_p^t (t - u) \{Q''(u) - Q''(p)\} du.$$

Following the same argument in the Appendix of Chen (2000), we have

$$\int_0^1 R_n(t) K_\beta(t; p) dt = o(b_n).$$

We also have

$$\begin{aligned} \int_0^1 t K_\beta(t; p) dt &= \frac{p + b_n}{1 + 2b_n} = p + (1 - 2p)b_n + O(b_n^2) \\ \int_0^1 t^2 K_\beta(t; p) dt &= \frac{(p + b_n)(p + 2b_n)}{(1 + 2b_n)(1 + 3b_n)} = p^2 + p(3 - 5p)b_n + O(b_n^2). \end{aligned}$$

Using the above results, we have

$$B_n = n^{1/2} b_n (1 - 2p) Q'(p) + \frac{1}{2} n^{1/2} b_n p (1 - p) Q''(p) + o(n^{1/2} b_n)$$

which completes the proof. □

Corollary 2. *In addition to the conditions in the above theorem, suppose that $\{b_n\}$ is such that $n^{1/2}b_n \rightarrow 0$ as $n \rightarrow \infty$. Then $\sqrt{n}\{\text{BQ}_n(p) - Q(p)\}$ converges in distribution to the normal distribution with mean zero and variance $\sigma^2(p)$, where*

$$\sigma^2(p) = \frac{(1-p)^2}{f^2(\xi_p)} \int_0^p \frac{du}{(1-u)^2\{1-H(F^{-1}(u))\}}.$$

Proof. Since $n^{1/2}b_n \rightarrow 0$ as $n \rightarrow \infty$, it suffices to show that $\sqrt{n}\{\hat{F}_n(\xi_p) - p\}/f(\xi_p)$ converges in distribution to the normal with mean zero and variance $\sigma^2(p)$. This is immediate by using Corollary 6.1 of Burke *et al.* (1981) or Theorem 8.1.1 of Csörgő (1983). \square

It is of interest to notice that the above normal distribution result is essentially the same as that of Padgett (1986) using the symmetric kernel approach.

4 AN ILLUSTRATIVE EXAMPLE

We illustrate the proposed quantile estimator with the comparison to the symmetric-kernel-based quantile estimator of Padgett (1986). To compare our result, we used the same data set analyzed by Padgett (1986) who explicitly provides the raw data set in Table 7. We briefly explain how he obtained the quantile estimator in his example. Then we compare our quantile estimates.

Padgett (1986) pointed out that it was difficult to find an optimal bandwidth in the sense of the (approximate) mean integrated square error due to mathematical difficulties arising from censoring. Thus, the bootstrap technique was used to determine the bandwidth h_n . A bootstrap sample $\{(X_1^*, \Delta_1^*), (X_2^*, \Delta_2^*), \dots, (X_n^*, \Delta_n^*)\}$ is randomly drawn from $\{(X_1, \Delta_1), (X_2, \Delta_2), \dots, (X_n, \Delta_n)\}$ with replacement and each element having probability $1/n$ of being sampled. Using 300 bootstrap samples of the raw data at each value of p and h_n , the bootstrap mean square errors

(MSEs) of $KQ_n(p)$ are obtained. The bootstrap bandwidth is obtained by choosing the value of h_n which gives the minimum bootstrap MSEs. From his example, the bootstrap estimates of h_n were $h_n^* = 0.30, 0.26, 0.34, 0.34, 0.40$ and 0.47 at $p = 0.1, 0.25, 0.50, 0.75, 0.90$ and 0.95 , respectively. Based on these results, a value of h_n was constructed as follows: $h_n = 0.28$ for $0 < p \leq 0.25$, $h_n = 0.34$ for $0.25 < p < 0.90$ and $h_n = 0.40$ for $0.90 \leq p \leq 1$. It is noted that he recommended that more smoothing is required for larger quantiles due to heavy right-censoring and small sample size of the example. Thus, for the quantile estimator at a large p , the bandwidth h_n is larger as the bootstrap estimate provides. In Figure 1, we drew $KQ_n(p)$ versus p (dashed curve). For more details, see Padgett (1986).

Similarly, we obtain our proposed quantile estimates as follows. First, using the aforementioned bootstrap technique, the bandwidth estimates were determined as $b_n^* = 0.01, 0.16, 0.85, 0.06, 0.37$ and 0.01 at $p = 0.1, 0.25, 0.50, 0.75, 0.90$ and 0.95 , respectively. Based on these estimates, we constructed a value of b_n as follows: $b_n = 0.01$ for $0 < p \leq 0.10$, $b_n = 0.16$ for $0.10 < p \leq 0.40$, $b_n = 0.85$ for $0.40 < p \leq 0.60$, $b_n = 0.40$ for $0.60 < p \leq 0.90$ and $b_n = 0.01$ for $0.90 < p \leq 1$. In order to make the comparison of the results of two methods more transparent, we superimposed the $BQ_n(p)$ versus p (solid curve) with the empirical PL quantile $\hat{Q}_n(p)$ (dotted step function) in Figure 1. From the figure, one can clearly see that the proposed quantile estimator is better near the boundaries than KQ_n . Especially when p is close to 0 or 1, the symmetric-kernel-based estimator is significantly small than the empirical PL quantile $\hat{Q}_n(p)$ due to the spill-over effects.

We calculate the MSE of the estimators with respect to the sample quantile as follows:

$$\begin{aligned} \text{MSE}(KQ_n) &= \frac{1}{n} \sum_{i=1}^n \left\{ KQ_n(\hat{F}_n(X_{(i)})) - X_{(i)} \right\}^2 = 0.25719 \\ \text{MSE}(BQ_n) &= \frac{1}{n} \sum_{i=1}^n \left\{ BQ_n(\hat{F}_n(X_{(i)})) - X_{(i)} \right\}^2 = 0.02179 \end{aligned}$$

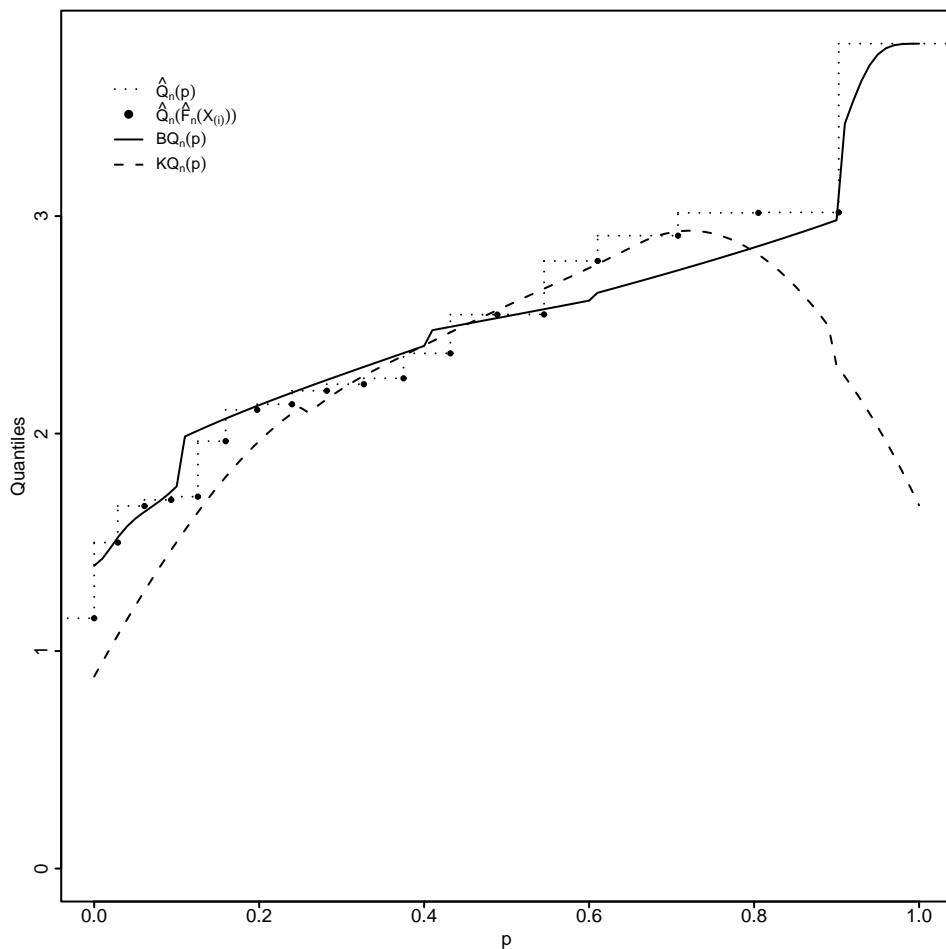


Figure 1: Quantile estimators under consideration.

The results also clearly indicate that the proposed method outperforms the symmetric kernel approach of Padgett (1986) overall.

5 SIMULATION

In order to examine the performance of the proposed method, we use a small Monte Carlo simulation with 10,000 runs. We compare the performance of the proposed method with the sample quantile and the method of Padgett (1986).

We generated a random sample of size $n = 50$ from the exponential distribution $F(t) = 1 - e^{-t}$ which is censored on the right with the censoring model $H(t) = 1 - e^{-t}$. Then we calculated the simulated MSEs of the sample quantile, the proposed quantile method and the quantile function of Padgett (1986) for $p = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ with various bandwidths (0.01, 0.02, 0.05, 0.1, 0.2, 0.5).

To help compare the simulated MSEs, we provide the simulated relative efficiency (SRE) based on the MSE which is defined as

$$\text{SRE}(Q_n(p)) = \frac{\text{MSE}(\hat{F}_n^{-1}(p))}{\text{MSE}(Q_n(p))},$$

where Q_n is either BQ_n or KQ_n . The results are presented in Table 1. From the simulation results, the proposed method using the beta kernel clearly outperforms the method of Padgett (1986) using the symmetric kernel. It is also noteworthy that the SREs for the above two methods are greater than one in most cases, which justifies the use of kernel smoothing instead of the sample quantile.

Table 1: Simulated relative efficiency under consideration

b_n, h_n	0.01		0.02		0.05		0.1		0.2		0.5	
	BQ_n	KQ_n	BQ_n	KQ_n	BQ_n	KQ_n	BQ_n	KQ_n	BQ_n	KQ_n	BQ_n	KQ_n
p												
0.05	1.101	0.997	1.236	0.987	1.781	0.996	3.472	1.024	8.442	1.206	0.933	3.120
0.10	1.134	1.026	1.238	1.046	1.617	1.053	2.585	1.069	6.570	1.144	2.801	2.139
0.25	1.048	1.002	1.117	1.005	1.343	1.013	1.813	1.031	3.121	1.082	4.663	1.622
0.50	1.105	1.020	1.159	1.036	1.276	1.075	1.415	1.132	1.580	1.253	1.721	1.580
0.75	1.880	1.017	2.090	1.035	2.593	1.088	3.304	1.164	4.400	1.285	5.186	4.368
0.90	1.849	1.012	2.059	1.022	2.734	1.053	3.926	1.131	5.613	2.765	4.884	3.345
0.95	1.782	1.002	2.014	1.007	2.563	1.039	3.225	2.078	3.605	3.130	2.437	1.223

We investigated the performance of the proposed method for various bandwidths. However, it should be noted that it is not appropriate to compare the two different

methods with the same bandwidths $b_n = h_n$. Thus we generated the pilot samples from the same exponential distribution to obtain the bootstrap bandwidths. Using 200 bootstrap samples of the pilot sample at each value of $b_n = 0.01(0.01)1.0$ and $h_n = 0.01(0.01)1.0$, the bootstrap mean square errors of BQ_n and KQ_n are obtained, respectively. We repeated this Monte Carlo experiment with 100 runs to obtain the simulated bootstrap bandwidths. Using these simulated bootstrap bandwidths, we performed a Monte Carlo simulation with 10,000 runs again. The results are presented in Table 2 which also indicates that the proposed method outperforms.

Table 2: Simulated bootstrap bandwidths and simulated relative efficiency under consideration

p	0.05	0.10	0.25	0.50	0.75	0.90	0.95
<i>Bootstrap bandwidths</i>							
b_n^*	0.0129	0.0149	0.0310	0.5381	0.1209	0.0206	0.0328
h_n^*	0.1242	0.1786	0.2722	0.7512	0.3428	0.0894	0.0199
<i>Simulated relative efficiency</i>							
BQ_n	4.836	5.496	4.179	1.734	5.141	3.679	2.012
KQ_n	0.991	1.039	1.008	1.588	1.192	1.023	1.019

6 CONCLUDING REMARKS

In this paper, we developed a new nonparametric estimator of a quantile function based on a beta probability density function. This new method offers clear advantages over the symmetric kernel approach and this was demonstrated through an example and a Monte Carlo simulation. Any kind of kernel approach requires an optimal bandwidth and because this application involves censoring, it is very difficult to obtain an optimal bandwidth. In order to calculate an optimal bandwidth, we used the bootstrap technique at a given point. Challenging future work would involve

developing methods for determining a global or local optimal bandwidth without the use of bootstrapping.

REFERENCES

- Apostol, T. M. (1974). *Mathematical Analysis*. Addison-Wesley Pub. Co., Reading, Massachusetts, 2nd edition.
- Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*, **37**, 577–580.
- Burke, M. D., Csörgő, S., and Horváth, L. (1981). Strong approximation of some estimates under random censorship. *Zeitschrift Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **56**, 87–112.
- Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, **31**, 131–145.
- Chen, S. X. (2000). Beta kernel smoothers for regression curves. *Statistica Sinica*, **10**, 73–91.
- Cheng, K.-F. (1984). On almost sure representation for quantiles of the product limit estimator with applications. *Sankhya Series A*, **46**, 426–443.
- Csörgő, M. (1983). *Quantile Processes with Statistical Applications*. SIAM, Philadelphia, PA.
- David, H. A. (1981). *Order Statistics*. John Wiley & Sons, New York.
- Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Annals of Statistics*, **12**, 261–268.

- Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. Wiley, New York.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.
- Lio, Y. L., Padgett, W. J., and Yu, K. F. (1986). On the asymptotic properties of a kernel type quantile estimator from censored samples. *Journal of Statistical Planning and Inference*, **14**, 169–177.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its applications*, **10**, 186–190.
- Padgett, W. J. (1986). A kernel-type estimator of a quantile function from right-censored data. *Journal of the American Statistical Association*, **81**, 215–222.
- Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105–131.
- Reiss, R.-D. (1980). Estimation of quantiles in certain nonparametric models. *Annals of Statistics*, **8**, 87–105.
- Yang, S.-S. (1985). A smooth nonparametric estimation of a quantile function. *Journal of the American Statistical Association*, **80**, 1004–1011.