Statistics and Probability Letters 83 (2013) 2399–2406

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

$\mathscr{L}_1$-deficiency of the sample quantile estimator with respect to a kernel quantile estimator

Mu Zhao [a,b], Hongmei Jiang [a,*], Yong Zhou [c,d]

[a] *Department of Statistics, Northwestern University, Evanston, IL, 60208, USA*
[b] *School of Mathematical Sciences, Anhui University, Hefei, 230039, China*
[c] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China*
[d] *School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China*

A R T I C L E   I N F O

Article history:
Received 5 March 2013
Received in revised form 27 June 2013
Accepted 28 June 2013
Available online 3 July 2013

Keywords:
Sample quantile estimator
Kernel quantile estimator
$\mathscr{L}_1$-deficiency
Optimal bandwidth

A B S T R A C T

The performance of the sample quantile estimator versus a kernel quantile estimator under the criterion of mean integrated absolute error (MIAE) for randomly right-censored data is considered in this paper. We show that the so called $\mathscr{L}_1$-deficiency of the sample quantile estimator with respect to the kernel quantile estimator is convergent to infinity. The optimal bandwidth in the sense of MIAE is obtained. Some simulation studies and one real data analysis are used to illustrate the results.

1. Introduction

Let X be the random variable of interest with continuous distribution function (d.f.) F, and C be the right censoring random variable with continuous d.f. G. We assume that X and C are independent. In practice, the observed dataset usually consists of n independent pairs (Y_i, δ_i), $i = 1, \dots, n$, where $Y_i = X_i \wedge C_i = \min(X_i, C_i)$ and $\delta_i = I(X_i \le C_i)$. Let L denote the d.f. of Y. By assumption of independence, we have $1 - L = (1 - F)(1 - G)$.

A nonparametric maximum likelihood estimator of d.f. $F(t)$ proposed in Kaplan and Meier (1958) is defined as

$$\widehat{F}_n(t) = \begin{cases} 1 - \prod_{Y_{(i)} \le x} \left(\dfrac{n-i}{n-i+1} \right)^{\delta_{(i)}}, & \text{if } t < Y_{(n)}, \\ 1, & \text{if } t \ge Y_{(n)}, \end{cases} \tag{1.1}$$

where $Y_{(1)} \leqslant Y_{(2)} \leqslant \cdots \leqslant Y_{(n)}$ are the ordered statistics of $Y_1, Y_2, \dots, Y_n$, and $\delta_{(1)}, \delta_{(2)}, \dots, \delta_{(n)}$ are the corresponding δ_i. The p-quantile $(0 < p < 1)$ of d.f. F is defined as

$$Q(p) = \inf\{t : F(t) \geqslant p\},$$

which can be estimated by its empirical estimator

$$\xi_{np} = \widehat{F}_n^{-1}(p) = \inf\{x : \widehat{F}_n(x) \ge p\}.$$

* Corresponding author. Tel.: +1 847 467 1087.
E-mail address: hongmei@northwestern.edu (H. Jiang).

0167-7152/$ – see front matter
http://dx.doi.org/10.1016/j.spl.2013.06.035

Lo and Singh (1986), Major and Rejtö (1988), Gijbels and Veraverbeke (1988), Beirlant and Einmahl (1990), among many others, show that the estimators $\widehat{F}_n(t)$ and $\widehat{F}_n^{-1}(p)$, $(0 < p < 1)$ are strongly consistent and asymptotically normal. However, the following two facts that

- the d.f. F and $Q(p)$ themselves are absolutely continuous,
- the simple step estimators may experience a substantial lack of efficiency caused by the variability of individual order statistics,

motivate numerous authors to consider the kernel type Kaplan–Meier estimator and quantile estimator. See, for example, Reiss (1981), Ghorai and Susarla (1990) and Lemdani and Ould-Saïd (2001) for distribution function, and Padgett (1986), Sheather and Marron (1990), Xiang (1995) and Cheng and Peng (2002) for quantile functions. For the same reasons, similar estimators have been proposed for copulas and ROC curves. The kernel quantile estimator introduced by Parzen (1979) is of the form:

$$\widehat{Q}_n(p) = \frac{1}{h_n} \int_0^1 \widehat{F}_n^{-1}(u) K\left(\frac{u-p}{h_n}\right) du, \tag{1.2}$$

where $K(\cdot)$ is a kernel function and h_n is a bandwidth.

To compare two different estimating procedures, Hodges and Lehmann (1970) proposed the conception of *deficiency*. Assume a less effective procedure requires k_n observations to give equally good performance as a statistical procedure based on n observations. The ratio $e = \lim_{n\to\infty} n/k_n$ and difference $d = \lim_{n\to\infty}(k_n - n)$ are two natural quantities for this comparison. Here, e which is known as *asymptotic relative efficiency* is more popular due to its stability in large samples. However, when $e = 1$ as in many important statistical problems, the quantity d may become a useful measure. Hodges and Lehmann (1970) named d *asymptotic deficiency*. The deficiencies of the sample quantile estimator with respect to kernel quantile estimator under the criterions of equal MSE and covering probability have been established, for example, by Falk (1984, 1985), Xiang (1995) and Zhao et al. (2011). To derive a more robust estimator, Lemdani and Ould-Saïd (2003) considered the $\mathscr{L}_1$-deficiency of $\widehat{F}_n(t)$ with respect to its corresponding smoothed counterpart.

In this paper, we study the $\mathscr{L}_1$-deficiency of the sample quantile estimator with respect to the kernel quantile estimator for right censored data. This article is organized as follows. Some definitions and assumptions are introduced in Section 2. The main results are stated in Section 3. Some simulation results and a real data example are reported in Section 4. In Section 5, we give some concluding remarks. Proofs of the results are deferred to the Appendix.

2. Definitions and assumptions

For any $0 < p < 1$, define the *mean absolute error* (MAE) of ξ_{np} as:

$$\text{MAE}(\xi_{np}) = E|\xi_{np} - Q(p)|,$$

whereas, the *mean integrated absolute error* (MIAE) of ξ_{np} as:

$$\text{MIAE}(\xi_{np}) = \int_0^1 \text{MAE}(\xi_{np})dp.$$

Let $i(n, p)$ denote the number of observations needed to achieve the same (or better) performance for ξ_{np} as the smoothed counterpart based on n observations, that is,

$$i(n, p) = \min\{j = 1, 2, \ldots, \text{MAE}(\xi_{jp}) \leqslant \text{MAE}(\widehat{Q}_n(p))\}.$$

In the same way, let

$$i(n) = \min\{j = 1, 2, \ldots, \text{MIAE}(\xi_{jp}) \leqslant \text{MIAE}(\widehat{Q}_n(p))\}.$$

The quantities $i(n, p) - n$ and $i(n) - n$ are referred to as *relative deficiency* and *relative $\mathscr{L}_1$-deficiency* of ξ_{np} with respect to $\widehat{Q}_n(p)$, respectively.

In the following section, we will show that the smoothed estimator $\widehat{Q}_n(p)$ is better than sample quantile ξ_{np} in the sense of MIAE for some bandwidths. To prove our main results, we make use of the following assumptions.

A1. $Q(x)$ has a bounded $(r + 1)$-th derivative in a neighborhood of p with $0 < p < F(b_L)$, and $Q(p)$ is the unique solution of $F(x) = p$, where $b_L = \sup\{t : L(t) < 1\}$.
A2. $F(t)$ has a bounded second derivative on $[0, \tau]$ where τ is a constant such that $0 < \tau < b_L$.
A3. $K(x)$ is Lipschitz of order 1 and has compact support on $[-1, 1]$.
A4. $K(x)$ is a kernel function of $(r + 1)$-th order with $r \geq 2$. Hence, $\int_{-1}^1 K(x)dx = 1$, $\int_{-1}^1 x^j K(x)dx = 0$, for $j = 1, \ldots, r$, and $\int_{-1}^1 x^{r+1}K(x)dx = c_r$, where c_r is a constant.
A5. As $n \to \infty$, the sequence of bandwidths satisfies:

$$\frac{nh_n^4}{\log^3 n} \to \infty \quad \text{and} \quad nh_n^{2r+2} \to 0.$$