韓國數學敎育學會誌<數學敎育>
1992. 12. 제 31권 제 2 호 133-140

Journal of the Korea Society of Mathematical Education
December 1992 Vol. XXXI, No. 2, 133-140

A STUDY ON KERNEL ESTIMATION OF A SMOOTH DISTRIBUTION FUNCTION ON CENSORED DATA

EUN SOOK JEE

ABSTRACT

The problem of estimating a smooth distribution function F at a point τ based on randomly right censored data is treated under certain smoothness conditions on F . The asymptotic performance of a certain class of kernel estimators is compared to that of the Kap lan-Meier estimator of $F(\tau)$. It is shown that the relative deficiency of the Kaplan-Meier estimator of $F(\tau)$ with respect to the appropriately chosen kernel type estimator tends to infinity as the sample size n increases to infinity. Strong uniform consistency and the weak convergence of the normalized process are also proved.

1. INTRODUCTION

Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be independently and identically distributed random variables with distribution functions F and G respectively. Let $\bar{F}= 1-F$ and $\bar{G} = 1-G$ denote the corresponding survival functions. Define $Z_i = X_i \wedge Y_i$, $\delta_i = 1$ if $X_i \leq Y_i$ and $= 0$ if $X_i > Y_i$, for $i = 1, \cdots, n$, where $\wedge$ denotes the minimum. Let $H(\cdot) = p(Z \leq \cdot)$ denote the distribution function of Z. It is easy to show that $1 - H = \bar{H} = \bar{F}\bar{G}$. The most well known estimator of F based on $(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$ is the Kaplan-Meier (1958) estimator. In recent years there has been a great deal of work on the large sample properties of this estimator. These large sample properties hold under very general conditions. However in some situations it may be assumed that F has a density, say,f . In such situations one would like to construct an estimator which is differentiable a.e. In the case of no censoring, this problem has been studied by Reiss(1981), Yamato(1972), Winter(1973, 1979), Azzalini(1981), Singh et al. (1983), Hill(1985), Falk(1983) and Mammitzsch(1984). Reiss(1981) has shown that a properly chosen kernel estimator is better than the usual empirical estimator with respect to relative deficiency. In this paper we define a smooth estimator of F based on a kernel function and the Kaplan-Meier estimator. The leading term of the mean square error of the proposed estimator at a fixed point τ is

* Associate Professor, Department of Mathematics , Kwang Woon University.
** The Research of this paper was supported by a prant from the Kwang Woon University.

calculated. Using this leading term it is shown that the relative deficiency of the Kaplan-Meier estimate tends to infinity as the sample size n increases to infinity. Uniform strong consistency and the weak convergence of the standardized process on a finite interval are also proved. Based on the randomly censored data, a kernel estimator of the density f can be defined as

$$\hat{f}_n(\tau) = \int \varepsilon_n^{-1} K((\tau - x/\varepsilon_n)\, d\hat{F}_{KM}(x) \tag{1-1}$$

where $K(\cdot)$ is a kernel function and $\hat{F}_{KM}$ is the Kaplan-Meier estimate defined as

$$1 - \hat{F}_{KM}(\tau) = \prod_j [\frac{N + (z_j)}{1 + N + (z_j)}]^{\delta_j I(Z_j \leq \tau)} \tag{1-2}$$

where

$$N^+(\cdot) = \sum_{j=1}^{n} I(z_j > 0) \text{ and } I(A) \text{ indicator of the set} A \tag{1-3}$$

Based on the estimator $\hat{f}_n$ an estimator of F can be defined as

$$\hat{F}_n(\tau) = \int_{-\infty}^{\tau} \hat{F}_n(x)\, dx \tag{1-4}$$

The estimator $\hat{F}_n$ can also be expressed as

$$\hat{F}_n(\tau) = \int K((\tau - y)/\varepsilon_n\, d\hat{F}_{KM}(y) \tag{1-5}$$

$$= \int \varepsilon_n^{-1} K((\tau - y)/\varepsilon_n\, d\hat{F}_{KM}(y)\, dy \tag{1-6}$$

where

$$K(x) = \int_{-\infty}^{x} K(u)\, du \tag{1-7}$$

Define $\hat{\bar{F}}_{KM} = 1 - \hat{F}_{KM}$, $T_F = Sup\{t|\bar{F}(t) > 0\}$

$$T_G = Sup\{t|\bar{G}(t) > 0\} \text{ and } g^{(l)}(x) = l^{th} \text{ derivative of } g$$

In the following section it will be assumed that the kernel K belongs to a class of kernels, K_m, of all Borel measurable bounded functions on the real line with the following properties

$$(i) \int K(x)\,dx = 1,$$

$$(ii) \int x^i K(x)\,dx = 0, \text{ for } i = 1. \cdots, m \text{ and}$$

$$(iii) M_{m+1} = \int |x|^{m+1}|K(x)|\,dx < \infty$$

2. Uniform Consistency and Weak Convergence

In this section uniform strong consistency results with rates and the weak convergence of the process $\{\sqrt{n}(\hat{F}_n - F)|x \leq T < T_F\}$ to the Gaussian process are proved.

Theorem 2.1. *Define* $T_n = sup\{t|\bar{H}(t) \geq 6\,(loglogn/(2n))^{1/2}\}$. *Suppose*

$$(i) k \in K_m,\ m \geq 0,$$

$$(ii) \lim_{n\to\infty} \varepsilon_n^{m+1}(n/loglogn)^{1/2} = 0, \quad \text{and}$$

$$(iii) either \quad \sup_x |f^{(m)}(x)| < \infty \quad or \quad \int |f^{(m)}(x)| < \infty$$

Then

$$\lim(sup|\hat{F}_n - F|/\bar{F}(T_n)) \leq \frac{52}{15}\int |k(t)|\,dt \quad a.s$$

Corollary 2.1. *Suppose the kernel K is a density satisfying the conditions of Theorem 2.1. Then*

$$\lim(sup|\hat{F}_n - F|/\bar{F}(T_n)) \leq \frac{52}{15} \quad a.s$$

proof of Theorem 2.1. Define

$$F_n(x) = \int \varepsilon_n^{-1} k((x-y)/\varepsilon_n)F(y)dy \quad \text{and}$$

$$B(F_n) = F_n(x) - F(x) \tag{2-1}$$

Then

$$sup|\hat{F}_n(x) - F_n(x)| \leq sup|\hat{F}_{KM}(x) - F(x)| \int |K(t)|dt \tag{2-2}$$

$$\text{and } B(F_n(x)) = \int K(t)(F(x - ht) - F(x))dt \tag{2-3}$$

Now using Taylor's expansion with remainder term in derivative form (2.3) can be expressed as

$$|B(F_n(x))| = |\int K(t)\frac{(\varepsilon_n t)^{m+1}}{(m+1)!}F^{(m+1)}(\xi(x.t))dt| \tag{2-4}$$

$$\leq \frac{(\varepsilon_n)^{m+1}}{(m+1)!}sup|f^{(m)}| \int |t|^{m+1}|K(t)|dt \tag{2-5}$$

$$= \varepsilon_n^{m+1} M_{m+1} sup|f^{(m)}| \tag{2-6}$$

On the other hand using Taylor's expansion with the integral form of the remainder, (2.3) can also be bounded as

$$|B(F_n(x))| = |\int K(t)\frac{1}{m!}\int_{x-\varepsilon_n t}^{x} F^{(m+1)}(u)(x-u)^m dudt| \tag{2-7}$$

$$\leq \varepsilon_n^{m+1} M_{m+1} \int |f^{(m)}(u)|du \tag{2-8}$$

Hence it follows from (2.6) and (2.8) that if $sup|f^{(m)}| < \infty$ or if

$$f^{(m)} \in L_1 \quad sup|B(F_n(x))| = O(\varepsilon_n^{m+1}) \tag{2-9}$$

Hence

$$sup|\hat{F} - F| \leq sup|\hat{F}_n - F| + sup|B(F_n)| \tag{2-10}$$

$$\leq sup|\hat{F}_{KM} - F| \int |K(t)| + O(\varepsilon_n^{m+1}) \tag{2-11}$$

This together with condition (ii) of Theorem 2.1 imply that

$$\lim(sup|\hat{F}_n - F|/\bar{F}(T_n)) \leq \lim sup|\hat{F}_{KM} - F|/\bar{F}(T_n) \int |K(t)|dt \ a.s$$

$$\leq \frac{52}{15}\int |K(t)dt, \tag{2-12}$$

where the second inequality follows from **Theorem 1** in Csőrgö and Horváth (1983). This completes the proof. □

We now establish the weak convergence of the process $\{\sqrt{n}(_n(x) - F(x)|x < Y_F\}$ Recall that

$$\hat{F}_n(x) - F(x) = I_n(x) + B(F_n(x)) \tag{2-13}$$

$$\text{where } \hat{F}_n(x) - F_n(x) = I_n(x) \tag{2-14}$$

and $F_n(x)$ and $B(F_n(x))$ are defined in (2.1) and (2.3) respectively.

Since $B(F_n(x))$ is non-stochastic it is enough to show that

$(i)\ \sqrt{n}\sup_x |B(F_n(x))| \Rightarrow 0$ as $n \Rightarrow \infty$ and

$(ii)\ \{\sqrt{n}I_n(x)|x \leq T\}$ converges weakly to a zero mean Gaussian process.

The next lemma establishes the tightness of the process. Throughout assume that

$$K(t) = 0 \text{ if } |t| > 1 \tag{2-15}$$

Lemma 2.1. *if* $\sqrt{n} \in \varepsilon_n^2 \to 0$ *as* $n \to \infty$ *then*

$(i)\ P(\sqrt{n}I_n(-\infty) > \varepsilon) < n$, *for all* $n \geq 1$

(ii) *for every* $\varepsilon > 0$ *and* $n > 0$, *there exist a* $\delta, 0 < \delta < 1$, *and an* n_0 *such that*

$$P(\sup_{|x-y|<\delta; x,y \leq T} \sqrt{n}|I_n(x) - I - n(y)| \geq \varepsilon) < n \text{ for all } n \geq n_0, \text{ and}$$

$(iii)\ \{\sqrt{n}(\hat{F}_n(x) - F(x))|x \leq T\}$ *is tight*

proof. Let n be large enough so that $\bar{H}(T_1) > 0$ with $T_1 > Y + \varepsilon_n$ in the details to follow. Let $\Phi(\delta) = \{(x,y)|x,y \leq T_1, |x-y| \leq \delta\}$ Since $I_n(-\infty) = 0$ for all n , (i) follows trivially. To show (ii) first recall that

$$I_n(x) = \int K(t)(\hat{F}_{KM}(x - \varepsilon_n t) - F(x - \varepsilon_n t))\, dt \tag{2-16}$$

Now observe that for every $\varepsilon > 0$ $0 < \delta < 1$, $\varepsilon^* \|k\|_1 = \varepsilon$, and $\|k\|_1 = \int |k(t)|\, dt$,

$$\begin{aligned} &P\{\sup_{|x-y|\leq\delta; x,y\leq T_1} \sqrt{n}|I_n(x) - I_n(y)| \geq \varepsilon\} \\ &\leq \{\sup_{|x-y|\leq\delta; x.y\leq T_1} |\sqrt{n}(\hat{F}_{KM}(x) - F(x)) - \sqrt{n}(\hat{F}_{KM}(y) - F(y))| \geq \varepsilon^*\} \end{aligned} \tag{2-17}$$

Let $\{W(x) : x \leq T_1\}$ denote a mean zero Gaussian process with the same covariance function as the Kaplan-Meier process. For given ε and η , choose $\delta = \delta(\varepsilon, \eta)$, $0 < \delta < 1$, such that

$$P(sup\{|W(x) - W(y)| : (x, y) \in \Phi(\delta)\} \geq \varepsilon^*) < \frac{\eta}{2} \tag{2-18}$$

Now fix this δ and choose n_0 such that

$$P\{\sup_{(x,y)\in\Phi(\delta)} |\sqrt{n}(\hat{F}_{KM}(x) - F(x)) - \sqrt{n}(\hat{F}_{KM}(y) - F(y))| \geq \varepsilon^*\} < \eta \tag{2-19}$$

for all $n \leq n_0$

Observe thatn_0 depends only on η and δ . Hence it follows from (2-17) and (2-19) that

$$P\{\sqrt{n}sup|I_n(x) - I_n(y)| : (x, y) \in \Phi(\delta)\} \geq \varepsilon\} < \eta \text{ for all} n \geq n_0$$

This completes the proof of (ii).
To show (iii), observe that

$$\sup_x \sqrt{n}|B(F_n(x))| \leq O(\sqrt{n}\varepsilon_n^{m+1}) \to 0 \qquad \text{for } m = 1$$

Since $\hat{F}_n(\cdot)$ is continuous, (i) and (ii) of this lemma together imply (iii) (see Theorem 8.2 in Billingsley 1968). This completes the proof of the lemma. □

In order to show that the finite dimensional distributions of the process $\sqrt{n}(\hat{F}_n - F)$ converge to the appropriate multivariate normal distributions we will use the following representation of the Kaplan-Meier process due to Gardiner and Susarla (1983).

Theorem 2.2. *Assuming F to be continuous one can write*

$$\hat{F}_{KM}(x) - F(x) = \frac{1}{n}\sum_{i=1}^{n} \xi(Z_i, \delta_i, x) + r_n(x), \tag{2-20}$$

where

$$\xi(Z, \delta, T) = \bar{F}(t)[g(zt) - (\bar{H}(t))^{-1} I(Z, \delta)], \tag{2-21}$$

$$g(u) = \int_0^u (\bar{H}(s))^{-2} \bar{G}(s) dF \tag{2-22}$$

$$\sup_{t \leq T_1} |r_n(t)| = O_p(n^{-1/2}) \tag{2-23}$$

Theorem 2.3. *Assume that $f^{(m)}$ is integrable for $m \geq 1$. Then for $t_1 \leq \cdots \leq t_p \leq T$*

$$\sum_{j=1}^{p} l_j \sqrt{n}(\hat{F}_n(t_j) - F(t_j)) \to N(0, l' \sum l) \quad \text{if } \sqrt{n}\varepsilon_n^m \to 0 \text{ as } n \to \infty$$

where $l' = (l_1, \cdots, l_p)$ and $\sum = (\sigma(t_i, t_j))$,

$$\sigma(t_i, t_j) = \bar{F}(t_i)\bar{F}(t_j) \int_0^{t_i \wedge t_j} (\bar{H}(u))^{-2} d\bar{H}_1(u), d\bar{H}_1 = \bar{G}dF \tag{2-24}$$

proof. The proof follows from the representation given in the previous theorem and the central limit theorem. The following theorem now follows from the Lemma 2.1 and Theorem 2.3. □

Theorem 2.4. *Let $\sigma(x, y)$ be as defined in (2.24). If $\sqrt{n}\varepsilon_n^m \to 0$ as $n \to \infty$ and $|f^{(m)}|$ is integrable for some positive integer $m \geq 1$, then*

$$\{\sqrt{n}(\hat{F}_n(x) - F(x)) : x \leq T\} \to \{W(x) : x \leq T\}$$

where $W(\cdot)$ is a zero mean Gaussian process with covariance function $\sigma(x, y)$.

References

[1] *Azzalini A (1981) A note on estimation of a distribution function and quantiles by a kernel method. Biometrika 68:326-328*

[2] *Billingsley P (1968) Convergence of probability measures. John Wiley and Sons*

[3] *Csőrgö S Horvath L (1983) The rate of strong uniform consistency for product limit estimators. Z Wahrsch verw Gebiete 62:411-426*

[4] *Falk M (1983) Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. Statistica Neerlandica 37:73-83*

[5] *Gardiner J, Susarla V (1983) Sequential estimation of the population mean with randomly censored data. Sequential Analysis 2:203-225*

[6] *Hill P (1985) Kernel estimation of a distribution function. Commun Statist-Theor Meth 14(3):605-620*

[7] *Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Amer Statist Assoc 53:457-481*

[8] *Major P, Rejto L (1988) Strong embedding of the estimator of the distribution function under random censorship. Annals of Statistics 16:1113-1132*

[9] *Mammitzsch V. (1984) On the asymptotically optimal solution within a certain class of kernel type estimators. Statistics Decisions 2:247-255*

[10] *Reiss R-D (1981) Nonparametric estimation of smooth distribution functions. Scand J Statist 8:116-119*

[11] *Singh R.S., Gasser T, Prasad B (1983) Nonparametric estimation of distribution functions. Commun Statist-Theor Meth 12(18):2095-2108*

[12] *Winter B.B. (1973) Strong uniform consistency of integals of density estimators. Canad J Statist 1:247-253*

[13] *Winter B.B. (1979) Convergence rates of perturbed empirical distribution functions. J Appl prob 16:163-173*

[14] *Yamato H. (1972) Some statistical properties of estimators of density and distribution functions. Bull Math Statist 14:113-131*

DEPARTMENT OF MATHEMATICS, KWANG WOON UNIVERSITY, SEOUL 139-701, KOREA