Uniform Consistency of Some Estimates of a Density Function

by

D.S. Moore and E.G. Henrichon

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series No. 168
August, 1968

Uniform Consistency of Some Estimates of a Density Function

by

D. S. Moore and E. G. Henrichon

Purdue University

1. Introduction and summary. Let $X_1, \ldots, X_n$ be independent random variables identically distributed with absolutely continuous distribution function F and density function f. Loftsgaarden and Quesenberry [2] propose a consistent nonparametric point estimator $\hat{f}_n(z)$ of $f(z)$ which is quite easy to compute in practice. In this note we introduce a step-function approximation f_n^* to $\hat{f}_n$, and show that both $\hat{f}_n$ and f_n^* converge uniformly (in probability) to f, assuming that f is positive and uniformly continuous in $(-\infty, \infty)$. For more general f, uniform convergence over any compact interval where f is positive and continuous follows. Uniform convergence is useful for estimation of the mode of f, for it follows from our theorem (see [3], section 3) that a mode of either $\hat{f}_n$ of f_n^* is a consistent estimator of the mode of f. The mode of f_n^* is particularly tractable; it is applied in [1] to some problems in pattern recognition.

2. The result. Choose a non-decreasing sequence of positive integers, $\{k(n)\}$, such that $k(n) \to \infty$ but $k(n) = o(n)$. For any real number z, let $r_{k(n)}(z)$ be the distance from z to the $k(n)$th closest of the observations $X_1, \ldots, X_n$. Then the univariate form of the Loftsgaarden-Quesenberry estimator is

$$\hat{f}_n(z) = \{(k(n) - 1) / n\} \{1/2r_{k(n)}(z)\} .$$

We define also the random step-function f_n^* as follows: let $X_{1n} \leq X_{2n} \leq \cdots \leq X_{nn}$ be the order statistics from $X_1, \ldots, X_n$. Then

$$f_n^*(z) = 0 \qquad z < X_{1n} \quad \text{or} \quad z \geq X_{nn}$$
$$\qquad = \hat{f}_n(X_{in}) \qquad X_{in} \leq z < X_{i+1,n} \qquad i = 1, \ldots, n-1 \quad .$$

<u>THEOREM</u>. <u>If</u> $f(z)$ <u>is uniformly continuous and positive on</u> $(-\infty, \infty)$ <u>and</u> $(\log n) / k(n) \to 0$, <u>then for every</u> $\epsilon > 0$

(2.1) $$P[\sup_{-\infty < z < \infty} |\hat{f}_n(z) - f(z)| > \epsilon] \to 0$$

and

(2.2) $$P[\sup_{-\infty < z < \infty} |f_n^*(z) - f(z)| > \epsilon] \to 0 \quad .$$

<u>Proof</u>. We will abbreviate (2.1) by $\hat{f}_n \to f$ (UP) and denote convergence in probability by $a_n \to a(P)$. Define

$$U_{k(n)}(z) = F(z + r_{k(n)}(z)) - F(z - r_{k(n)}(z)).$$

We show first that

(2.3) $$\{n/(k(n)-1)\} \, U_{k(n)}(z) \to 1(UP) \quad .$$

By definition of $r_{k(n)}(z)$, the interval $[z - r_{k(n)}(z), z + r_{k(n)}(z)]$ contains exactly $k(n)$ observations, one of which falls at an endpoint of the interval. Suppose the order statistic X_{qn} is the lower endpoint. Then

$$(2.4)\qquad \sum_{j=1}^{k(n)-1} \{F(X_{q+j,n}) - F(X_{q+j-1,n})\} \leq U_{k(n)}(z)$$

$$\leq \sum_{j=1}^{k(n)} \{F(X_{q+j,n}) - F(X_{q+j-1,n})\}$$

with the conventions $F(X_{o,n}) = 0$ and $F(X_{n+1,n}) = 1$. Upper and lower bounds having the same distribution as those in (2.4) exist when X_{qn} is on upper endpoint. (It is stated in [2] that $U_{k(n)}$ has the beta distribution of one of the sums of elementary coverages in (2.4). This is false, since w.p.1 only one endpoint of the interval coincides with an observation; the modifications required to correct the proof of [2] are trivial.)

It is well known that

$$F(X_{1n}),\ F(X_{2n}) - F(X_{1n}), \ldots,\ 1-F(X_{nn})$$

have the same joint distribution as

$$Y_1 / S_{n+1}, \ldots, Y_{n+1} / S_{n+1} \quad ,$$

where $Y_1, \ldots, Y_{n+1}$ are independent exponential random variables with mean 1 and $S_{n+1} = Y_1 + \ldots + Y_{n+1}$. So the upper and lower bounds in (2.4) will converge to 1 (UP) if we can prove that

$$(2.5)\qquad \max_{0\leq i\leq n-k(n)+1} \left| \left\{\frac{1}{k(n)} \sum_{j=i+1}^{i+k(n)} Y_j \Big/ \frac{1}{n} S_{n+1}\right\} - 1 \right| \to 0 \ (P) \quad .$$

Since $n^{-1}S_{n+1} \to 1$ w.p.1 by the law of large numbers, (2.5) will follow if we can show that the sums $\{k(n)\}^{-1} \Sigma_{i+1}^{i+k(n)} Y_j$ are uniformly near 1 in probability. For any $\epsilon > 0$,

$$(2.6) \qquad P_n = P[\text{for some } i, \left|\sum_{j=i+1}^{i+k(n)} (Y_j-1)\right| > k(n)\,\epsilon]$$

$$\leq \sum_{i=1}^{n+1} P[\sum_{j=i+1}^{i+k(n)} (Y_j-1) > k(n)\epsilon] + \sum_{i=1}^{n+1} P[\sum_{j=i+1}^{i+k(n)} (Y_{j-1}) < -k(n)\epsilon] .$$

Using the fact that $P[X > 0] \leq E[e^{tX}]$ for any random variable X and $t > 0$ such that the right side is finite, we obtain

$$P[\sum_{j=i+1}^{i+k(n)} (Y_j-1) > k(n)\epsilon] \leq E[e^{t(\Sigma Y_j - k(n) - k(n)\epsilon)}]$$

$$= \{e^{-t(1+\epsilon)}/(1-t)\}^{k(n)} \qquad 0 < t < 1 .$$

(Recall that a sum of $k(n)$ Y_j's has the gamma distribution with parameter $k(n)$.) Choosing the minimizing value $t = 1 - (1+\epsilon)^{-1}$ gives the bound $\{(1+\epsilon)e^{-\epsilon}\}^{k(n)}$. A similar bound holds for each term of the second sum on the right side of (2.6). Therefore $P_n \leq (n+1)\, a(\epsilon)^{-k(n)}$, where $a(\epsilon) > 1$ for $\epsilon > 0$. Since $(\log n) / k(n) \to 0$, $P_n \to 0$ and (2.5) is proved.

It follows from (2.3) that $U_{k(n)} \to 0$ (UP) and hence, since f is everywhere positive, that $r_{k(n)} \to 0$ (UP).

To conclude (2.1) we need only (2.3) and the fact that $U_{k(n)}/2r_{k(n)} \to f$ (UP). Since f is uniformly continuous and $r_{k(n)} \to 0$ (UP), this is immediate from the estimate

$$(2.7) \qquad \left|\frac{U_{k(n)}(z)}{2r_{k(n)}(z)} - f(z)\right| = \left|\frac{1}{2r_{k(n)}(z)} \int_{z-r}^{z+r} [f(t)-f(z)]\,dt\right|$$

$$\leq \max \{|f(t)-f(z)| : z-r_{k(n)}(z) \leq t \leq z + r_{k(n)}(z)\} .$$

The argument for (2.2) is slightly longer. Let $i(z)$ be the index such that

$$X_{i(z),n} \leq z < X_{i(z)+1,n}$$

For any compact interval I, the probability that X_{1n} and X_{nn} fall outside I approaches 1 as $n \to \infty$, by positivity of f. Thus $i(z)$ is defined for all $z \in I$ with probability approaching 1 for large n. The Glivenko-Cantelli theorem and uniform continuity of F^{-1} on $[\alpha, 1-\alpha]$ for any $\alpha > 0$ give that

$$\sup_{z \in I} | X_{i(z),n} - z| \to 0 \ (P). \tag{2.8}$$

From (2.8) and the fact that $r_{k(n)} \to 0$ (UP), we can conclude by an estimate analogous to (2.7) that

$$\sup_{z \in I} \left| \frac{U_{k(n)}(X_{i(z),n})}{2r_{k(n)}(X_{i(z),n})} - f(z) \right| \to 0 \ (P)$$

and hence, using (2.3), that for any compact interval I and any $\epsilon > 0$,

$$\lim_{n \to \infty} P[\sup_{z \in I} | f_n^*(z) - f(z)| > \epsilon] = 0 \tag{2.9}$$

If we can establish that for any $\epsilon > 0$ there is a compact interval I_ϵ such that

$$(2.10)\qquad \lim_{n\to\infty} P[\sup_{z \notin I_\epsilon} |f_n^*(z) - f(z)| > \epsilon] = 0,$$

this with (2.9) will imply (2.2).

Since $f(z) \to 0$ as $z \to \pm\infty$, we can choose a compact interval $I^* = [a,b]$ such that $f(z) < \epsilon/2$ outside I^*. Then by (2.1), $\hat{f}_n(z) < \epsilon$ for all $z \notin I^*$ with probability approaching 1 as $n\to\infty$. Let $I_\epsilon = [a,b+c]$ for some $c > 0$. Then by (2.8) and the fact that $P[X_{1n} < a,\ X_{nn} > b+c] \to 1$, we have that

$$P[X_{i(z),n} \notin I^* \text{ for all } z \notin I_\epsilon \text{ with } X_{1n} \leq z < X_{nn}] \to 1.$$

Thus with probability approaching 1, $f_n^*(z)$ is either 0 or $\hat{f}_n(X_{in})$ for some $X_{in} \notin I^*$, for all $z \notin I_\epsilon$. This establishes (2.10).

References

[1] Henrichon, E.G. and Fu, K.S. (1968). On mode estimation for pattern recognition. Submitted to IEEE Trans. Information Theory.

[2] Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36 1049-1051.

[3] Parzen, Emanuel (1962). On estimation of a probability density function and mode. Ann. Math. Statist. 33 1065-1076.