Reprinted from:
STATISTICAL DECISION THEORY AND RELATED TOPICS, II
© 1977
ACADEMIC PRESS, INC.
NEW YORK SAN FRANCISCO LONDON

#455

LARGE SAMPLE PROPERTIES OF NEAREST NEIGHBOR DENSITY FUNCTION ESTIMATORS

By David S. Moore* and James W. Yackel
Purdue University

1. *Introduction.* Let $X_1, X_2, \ldots$ be iid random variables having unknown density function f with respect to Lebesgue measure λ on Euclidean p-space R^p. We wish to estimate f(z) for a given z. Let $\{k(n)\}$ be a sequence of positive integers satisfying

(1.1) $$k(n) \to \infty \text{ and } k(n)/n \to 0 \quad \text{as } n \to \infty.$$

Define R(n) as the distance from z to the k(n)th closest of $X_1, \ldots, X_n$, distance being measured in a norm $||\cdot||$ on R^p which generates the usual topology. Denote by S(r) the "sphere"

$$S(r) = \{x \text{ in } R^p\colon ||x-z|| \le r\}.$$

A nearest neighbor estimator of f(z) is

$$g_n(z) = \frac{k(n)/n}{\lambda\{S(R(n))\}}.$$

Note that $g_n(z)$ is simply empiric measure divided by Lebesgue measure for the region S(R(n)). This estimator is essentially due to Fix and Hodges [2], and was explicitly introduced and studied by Loftsgaarden and Quesenberry [5]. These and subsequent authors used the Euclidean norm, but for $p > 1$ other norms may be useful (e.g., squares about z rather than spheres are obtained from the "maximum component" norm), and proofs are unaffected by this generality. We have suppressed the dependence on z of R(n) and S(r), since in this paper we consider only results for a fixed z in R^p.

Loftsgaarden and Quesenberry proved consistency in probability of $g_n(z)$. Wagner [8] established almost sure (a.s.)

*Research of this author was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. 72-2350 B.

consistency under a condition equivalent to $k(n)/\log n \to \infty$. For the case $p = 1$, Moore and Henrichon [6] proved a.s. uniform consistency when $k(n)/\log n \to \infty$. (They state only convergence in probability, but an application of the Borel-Cantelli lemma shows that their proof yields a.s. convergence.) The local control of the estimation process which is a feature of the nearest neighbor estimator has been popular with practitioners, who have used g_n in discrimination and pattern recognition problems.

Sections 2 and 3 of this paper are devoted to g_n. In Section 2 we establish a.s. consistency under the condition $k(n)/\log\log n \to \infty$. Since $R(n)$ is a sample $k(n)/n$-tile, the study by Kiefer [3] of sample p_n-tiles for $p_n \to 0$ provides the tools needed for our result. What is more, it follows from Kiefer's work that $k(n)/\log\log n \to \infty$ is the weakest condition on $\{k(n)\}$ satisfying (1.1) which guarantees a.s. convergence of $g_n(z)$ to $f(z)$. Section 3 proves asymptotic normality of $g_n(z)$. The proof uses the standard device of restating an event defined in terms of an order statistic as an event given in terms of a binomial random variable. The limiting distribution derived in Section 3 is required in the more general study in Section 4.

Recently, the authors [7] observed that g_n could be viewed as the uniform kernel case of the general nearest neighbor density function estimator defined by

$$f_n(z) = \frac{1}{nR(n)^p} \sum_{i=1}^{n} K\left(\frac{z-X_i}{R(n)}\right)$$

where

(1.2) $K(u)$ is a bounded density on R^p

$K(u) = 0$ for $||u|| > 1$

Here the norm $||\cdot||$ must satisfy the additional restriction that

$$\lambda\{S(r)\} = cr^p \quad \text{where } c = \lambda\{S(1)\}.$$

This is the case for, e.g., the usual Euclidean norm and the maximum-component norm.

The estimator f_n is the analog of the Rosenblatt-Parzen class of bandwidth estimators defined by

$$\hat{f}_n(z) = \frac{1}{nr(n)^p} \sum_{i=1}^{n} K\left(\frac{z-X_i}{r(n)}\right)$$

where $\{r(n)\}$ is a sequence of positive bandwidths satisfying $r(n) \to 0$ and $nr(n)^p \to \infty$ as $n \to \infty$. In our earlier paper we showed, roughly speaking, that any consistency theorem (in probability or almost sure, pointwise or uniform) true for $\hat{f}_n$ remains true for f_n having the same kernel K and $k(n) \sim \alpha nr(n)^p$ for some $\alpha > 0$. This allows the large literature on consistency of $\hat{f}_n$ to be restated for f_n. See [7] for details and qualifications. Here we mention only that either by this consistency-equivalence result or directly from Kiefer's work it follows that the uniform kernel case of $\hat{f}_n$ is a.s. consistent when $nr(n)^p/\log \log n \to \infty$. This is the analog of the result of Section 2 below, and is similarly best possible and stronger than known results for general kernels. Thus Section 2 sets a goal for work on a.s. consistency of $\hat{f}_n$ or f_n, and the results of [7] show that attaining this goal for either of $\hat{f}_n$ or f_n is sufficient to reach it for both.

Sections 4 and 5 concern the general nearest neighbor estimator f_n. Section 4 establishes asymptotic normality. It is noteworthy that f_n does not have the same asymptotic variance as the matching bandwidth estimator $\hat{f}_n$. The nearest neighbor method is more efficient than the bandwidth method when f(z) is small, as intuition might suggest. Section 5 shows that weak consistency of f_n implies mean consistency, a supplement to the consistency results in [7].

2. *Almost sure consistency, uniform kernel case.* We make use of a lemma which extracts a very small portion of Theorem 6 of Kiefer [3].

LEMMA 1. *Let Z_n be a sample α_n-tile from n iid random variables uniformly distributed on (0,1). If $\alpha_n \to 0$ and $n\alpha_n/\log \log n \to \infty$, then $Z_n/\alpha_n \to 1$ a.s.*

Lemma 1 is applied to $g_n(z)$ by noting that if

$$H(r) = P[||X-z|| \leq r] = \int_{S(r)} f(x)dx \tag{2.1}$$

then $H(R(n))$ is the sample $k(n)/n$-tile from n iid uniform random variables. Here is our a.s. consistency result.

THEOREM 1. *Let* f *be continuous at* z, *and let* $\{k(n)\}$ *satisfy* (1.1) *and* $k(n)/\log\log n \to \infty$. *Then* $g_n(z) \to f(z)$ a.s.

PROOF. Lemma 1 states that

$$\frac{k(n)/n}{H(R(n))} \to 1 \quad \text{a.s.} \tag{2.2}$$

from which it follows that

$$H(R(n)) \to 0 \quad \text{a.s.} \tag{2.3}$$

We claim that

$$R(n) \to r_0 = \inf\{r: H(r) > 0\} \quad \text{a.s.} \tag{2.4}$$

For clearly $R(n) \geq r_0$ a.s. for each n, and if for some $\varepsilon > 0$, $R(n) \geq r_0 + \varepsilon$ for a sequence of n at a sample point ω, then $H(R(n)) \geq H(r_0 + \varepsilon) > 0$ for these n at ω. By (2.3), this can occur only on a set of ω having probability zero.

Applying the mean value theorem for integrals to (2.1), there exist λ_n satisfying

$$\inf_{S(R(n))} f(x) \leq \lambda_n \leq \sup_{S(R(n))} f(x)$$

such that

$$H(R(n)) = \lambda_n \, \lambda\{S(R(n))\}. \tag{2.5}$$

With (2.2), (2.5) implies that $g_n(z)/\lambda_n \to 1$ a.s. If $f(z) > 0$, then by (2.4), $R(n) \to 0$ a.s. and by continuity of f at z, $\lambda_n \to f(z)$ a.s. If $f(z) = 0$, then (2.3), (2.4) and (2.5) imply that $\lambda_n \to 0$ a.s. In either case, $g_n(z) \to f(z)$ a.s.

The proof of Theorem 1 amounts to observing that

$$g_n(z) - \frac{k(n)/n}{H(R(n))} f(z) = a_n f(z)$$

and applying (2.2) to a_n. From Kiefer's Theorem 6 it follows also that if $k(n)/\log\log n \to v$, $0 < v < \infty$, then $\underline{\lim}\, a_n$ and

$\overline{\lim}\, a_n$ are unequal, finite and positive. If $k(n)/\log\log n \to 0$, then $\overline{\lim}\, a_n = \infty$ and $\underline{\lim}\, a_n = 0$. Thus $g_n(z)$ is not a.s. consistent if $k(n)$ increases more slowly than is assumed in Theorem 1. Of course, $g_n(z)$ remains weakly consistent as long as (1.1) holds.

3. *Asymptotic normality, uniform kernel case.* Although the proof in this section is straightforward, both it and the proof of Section 4 require the assumption

$$(3.1) \qquad (k(n))^{\frac{1}{2}}|f(z_n)-f(z)| \to 0(P) \text{ when } ||z_n-z|| \leq R(n).$$

The assumption (3.1) connects $\{k(n)\}$ and the local behavior of f at the point z. It can be restated in more explicit form for specific norms $||\cdot||$. In particular, when $f(z) > 0$ and either the Euclidean norm or the maximum-component norm is used,

$$(3.2) \qquad \frac{k(n)/n}{cR(n)^p} \to f(z)(P) \qquad c = \lambda\{S(1)\}$$

(This is just weak consistency as proved in [5]), so then $R(n) = 0_p\{(k(n)/n)^{1/p}\}$ and (3.1) is implied by

$$(3.3) \qquad (k(n))^{\frac{1}{2}}|f(z_n)-f(z)| \to 0 \text{ when } ||z_n-z|| = 0\{(\frac{k(n)}{n})^{1/p}\}.$$

If the p first partial derivatives exist and are bounded near z, (3.3) in turn is satisfied when

$$k(n) = o\{n^{2/(p+2)}\}.$$

THEOREM 2. Let f be continuous at z, $f(z) > 0$, $\{k(n)\}$ satisfy (1.1), and let (3.1) hold. Then

$$\mathcal{L}\{(k(n))^{\frac{1}{2}}(g_n(z)-f(z))\} \to N(0,f^2(z))$$

PROOF. As in the proof of Theorem 1, we can write

$$g_n(z) = \frac{k(n)/n}{H(R(n))} f(z_n) \qquad \text{for some } z_n \text{ in } S(R(n)).$$

Now (1.1) and $f(z) > 0$ are sufficient for $R(n) \to 0(P)$ and $H(R(n))/(k(n)/n) \to 1$ (P) (see [5]). Therefore from

$$\frac{(k(n))^{\frac{1}{2}}}{f(z)}(g_n(z)-f(z)) = (k(n))^{\frac{1}{2}}\left(\frac{k(n)/n}{H(R(n))} - 1\right) + (k(n))^{\frac{1}{2}}\left(\frac{f(z_n)}{f(z)} - 1\right)\frac{k(n)/n}{H(R(n))}$$

and (3.1), we need only show that

$$\mathcal{L}\{(k(n))^{\frac{1}{2}}\left(\frac{k(n)/n}{H(R(n))} - 1\right)\} \to N(0,1).$$

Since $H(R(n))$ is the $k(n)$th order statistic of n iid uniform random variables $U_1,\ldots,U_n$ on $(0,1)$,

$$\begin{aligned} P_n(a) &= P[(k(n))^{\frac{1}{2}}\left(\frac{k(n)/n}{H(R(n))} - 1\right) \le a] \\ &= P[H(R(n)) \ge \frac{k(n)/n}{1+ak^{-\frac{1}{2}}}] \\ &= P[B_n < k(n)] \end{aligned}$$

where B_n is the number of $U_1,\ldots,U_n$ falling below $\pi_n = (k(n)/n)/(1+ak(n)^{-\frac{1}{2}})$ and has the binomial (n,π_n) distribution. By (1.1), $\pi_n \to 0$ and $n\pi_n \to \infty$, so that B_n is asymptotically normal. Writing

$$P_n(a) = P\left[\frac{B_n-n\pi_n}{\sigma_n} < \frac{k(n)-n\pi_n}{\sigma_n}\right]$$

where $\sigma_n = [n\pi_n(1-\pi_n)]^{\frac{1}{2}}$, and computing

$$\frac{k(n)-n\pi_n}{\sigma_n} \sim \frac{k(n)-n\pi_n}{(n\pi_n)^{\frac{1}{2}}} = a\left(\frac{k^{\frac{1}{2}}}{a+k^{\frac{1}{2}}}\right)^{\frac{1}{2}} \to a$$

we obtain $P_n(a) \to \Phi(a)$, Φ being the standard normal df. This completes the proof.

4. *Asymptotic normality, general case.* Recall that in order to formulate the general nearest neighbor estimator f_n, we require that the norm $||\cdot||$ satisfy

(4.1) $$\lambda\{S(r)\} = cr^p \quad \text{where } c = \lambda\{S(1)\}.$$

In this case, (3.1) is equivalent to the more usable condition (3.3).

THEOREM 3. Let f be continuous at z, f(z) > 0, K(u) satisfy (1.2) and {k(n)} satisfy (1.1). Let also (3.3) and (4.1) hold. Then

$$\mathcal{L}\{(k(n))^{\frac{1}{2}}(f_n(z)-f(z))\} \to N(0,cf^2(z)\int K^2(u)du)$$

The proof will be divided into several parts. First note that in

$$f_n(z) = \frac{1}{nR(n)^p}\sum_{i=1}^{n} K(\frac{z-X_i}{R(n)})$$

there are exactly k(n)-1 nonzero summands by (1.2), corresponding to the first k(n)-1 order statistics of $||X_i-z||$. Denote by $Y_1,\ldots,Y_{k(n)-1}$ the subsequence of $X_1,\ldots,X_n$ defined by

$$Y_1 = X_{i_1} \qquad i_1 = \min\{i: ||X_i-z|| < R(n)\}$$

$$Y_j = X_{i_j} \qquad i_j = \min\{i > i_{j-1}: ||X_i-z|| < R(n)\}$$

and let $K_{n,i} = K(\frac{z-Y_i}{R(n)})$ be the nonzero summands in f_n. Then the conditional distribution of $Y_1,\ldots,Y_{k(n)-1}$ given R(n) = r is that of k(n)-1 independent observations each having the density function

$$f(y)/P(S(r)) \quad \text{for } y \text{ in } S(r)$$

where

$$P(S(r)) = \int_{S(r)} f(x)dx.$$

Therefore the conditional distribution of $K_1,\ldots,K_{k(n)-1}$ given R(n) = r is that of k(n)-1 iid random variables having mean

$$E(r) = E[K_i|R(n) = r] = \int_{S(r)} K(\frac{z-y}{r})\frac{f(y)}{P(S(r))}\,dy$$

and variance

$$\sigma^2(r) = \int_{S(r)} K^2(\frac{z-y}{r})\frac{f(y)}{P(S(r))}\,dy - E^2(r).$$

By the (vector) change of variables u = (z-y)/r and the mean value theorem for integrals, we can write

$$(4.2) \qquad E(r) = \frac{\lambda_{1,r}}{P(S(r))} r^p \int_{S(1)} K(u)du = \frac{\lambda_{1,r}}{P(S(r))} r^p$$

$$= \frac{\lambda_{1,r}}{c\lambda_{2,r}}$$

and

$$(4.3) \qquad \sigma^2(r) = \frac{\lambda_{3,r}}{c\lambda_{2,r}} \int_{S(1)} K^2(u)du - \left(\frac{\lambda_{1,r}}{c\lambda_{2,r}}\right)^2$$

where

$$\inf_{S(r)} f(x) \le \lambda_{i,r} \le \sup_{S(r)} f(x).$$

We first consider the normalized sum

$$Z_n = \sum_{i=1}^{k(n)-1} \frac{K_i - E(R(n))}{(k(n))^{\frac{1}{2}}\sigma(R(n))}$$

LEMMA 2. <u>Under the conditions of Theorem</u> 3, <u>if</u> $K(u)$ <u>is not constant on</u> $S(1)$, <u>then</u>

$$\mathcal{L}\{Z_n\} \to N(0,1).$$

PROOF. If $F_n(x|r)$ is the conditional df of Z_n given $R(n) = r$, then by the remarks above, the Berry-Esseen theorem applies to give

$$(4.4) \qquad |F_n(x|r) - \Phi(x)| \le \frac{3M^3}{\sigma(r)(k(n))^{\frac{1}{2}}}$$

where $M = \sup|K(u)| < \infty$. Since (1.1) implies that $R(n) \to 0(P)$,

$$(4.5) \qquad \sigma^2(R(n)) \to \sigma^2 = c^{-1} \int_{S(1)} K^2(u)du - c^{-2} \qquad (P)$$

and $\sigma^2 > 0$ when $K(u)$ is not the uniform pdf on $S(1)$. Then if G_n is the df of $R(n)$ and $\delta > 0$,

$$|P[Z_n \le x] - \Phi(x)| \le \int |F_n(x|r) - \Phi(x)| dG_n(x)$$

$$\le \frac{3M^2}{\delta(k(n))^{\frac{1}{2}}} P[\sigma(R(n)) > \delta] + 2P[\sigma(R(n)) \le \delta]$$

and this with (4.5) establishes that $\mathcal{L}\{Z_n\} \to N(0,1)$.

PROOF OF THEOREM 3. We write

$$(4.6)\quad (k(n))^{\frac{1}{2}}(f_n(z)-f(z)) = \frac{k(n)\sigma(R(n))}{nR(n)^p} Z_n + (k(n))^{\frac{1}{2}}\left[\frac{k(n)E(R(n))}{nR(n)^p} - f(z)\right]$$

Since by (3.2) and (4.5),

$$\frac{k(n)\sigma(R(n))}{nR(n)^p} \to cf(z)\sigma \qquad (P)$$

The first term on the right in (4.6) has

$$N(0, cf^2(z)\textstyle\int K^2(u)du - f^2(z))$$

as its limit in law by Lemma 2. The second term on the right of (4.6) can be written as

$$\frac{k(n)/n}{R(n)^p}(k(n))^{\frac{1}{2}}(E(R(n))-c^{-1})+(k(n))^{\frac{1}{2}}\left(\frac{k(n)/n}{cR(n)^p} - f(z)\right)$$

$$= \frac{k(n)/n}{cR(n)^p}(k(n))^{\frac{1}{2}}\left(\frac{f(z_{1,n})}{f(z_{2,n})} - 1\right) + (k(n))^{\frac{1}{2}}(g_n(z)-f(z))$$

for large n, by (4.2) and continuity of f at z. Here $z_{i,n}$ lie in $S(R(n))$. By (3.1) and (3.2) applied to the first term and Theorem 2 applied to the second, this last expression has $N(0,f^2(z))$ as its limiting distribution. Moreover, it is asymptotically independent of Z_n. To see this, it is sufficient to show that

$$P[Z_n \le a \mid k^{\frac{1}{2}}(g_n-f) \ge b] = P[Z_n \le a \mid R(n) \ge (\frac{k/nc}{f+bk^{-\frac{1}{2}}})^{1/p}]$$

converges to $\Phi(a)$ for any b. That this is true follows from the argument used to prove Lemma 2. Theorem 3 now follows from (4.6).

Note that the bandwidth estimator $\hat{f}_n$ using the same kernel $K(u)$ and $r(n) = (k(n)/n)^{1/p}$ does not have the same limiting distribution as does f_n. Cacoullos [1] shows (under conditions which ask more of f and less of $\{k(n)\}$ than those of Theorem 3) that

$$\mathcal{L}\{(nr(n)^p)^{\frac{1}{2}}(\hat{f}_n(z)-f(z))\} \to N(0,f(z)\textstyle\int K^2(u)du).$$

Comparison of asymptotic variances shows that $\hat{f}_n(z)$ is less efficient at points z where f(z) is small, that is, where use of the fixed radii {r(n)} may result in few observations.

5. *Mean consistency.* Pointwise weak consistency results for f_n are available both by direct proof (for g_n) and by the consistency-equivalence results of [7]. It is easy to show that under quite general conditions, weak consistency of f_n implies mean consistency. This we now do.

THEOREM 4. *If K(u) is bounded, f is bounded in a neighborhood of z, and {k(n)} satisfies (1.1), then $f_n(z) \to f(z)(P)$ implies that $E[|f_n(z)-f(z)|] \to 0$.*

PROOF. We must show (Loeve (1963), p. 163) that

$$\lim_{a\to\infty} \int_{\{|f_n|>a\}} |f_n|\,dP \to 0$$

uniformly in n. Let M denote an arbitrary positive constant. Since K is bounded,

$$|f_n| \leq M \frac{k(n)/n}{R(n)^p}$$

and hence if $c(n) = (Mk(n)/an)^{1/p}$,

$$P(n,a) = \int_{\{|f_n|>a\}} |f_n|\,dP \leq M \frac{k(n)}{n} \int_{\{R(n)<c(n)\}} R(n)^{-p}\,dP.$$

But R(n) is the k(n)th order statistic from n observations on the df H(r) (see (2.1)). So

$$P(n,a) \leq M \frac{k}{n} k\binom{n}{k} \int_0^{c_n} r^{-p} H^{k-1}(r)[1-H(r)]^{n-k} dH(r)$$

$$\leq M \frac{k^2}{n} \binom{n}{k} \int_0^{c_n} r^{-p} H^{k-1}(r)\,dH(r)$$

From $H(r) = \lambda(r)cr^p$ for $\inf f \leq \lambda(r) \leq \sup f$, we see

$$P(n,a) \leq M \frac{k^2}{n} \binom{n}{k} \int_0^{c_n} H^{k-2}(r)\,dH(r)$$

$$= M \frac{k^2}{n} \binom{n}{k} \frac{1}{k-1} H^{k-1}(c_n)$$

$$\leq M \left(\frac{k}{n}\right)^k \binom{n}{k} \left(\frac{M}{a}\right)^{k-1}$$

after again substituting $H(r) = \lambda(r)cr^p$ in $H(c_n)$. Thus we must show that given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\left(\frac{k}{n}\right)^k \binom{n}{k} \delta^{k-1} < \varepsilon \quad \text{for all } n.$$

That this is true follows easily from

$$\left(\frac{k}{n}\right)^k \binom{n}{k} \delta^{k-1} < \delta^{k-1} \frac{k^k}{k!} \sim \frac{\delta^{k-1}}{e^k (2\pi k)^{\frac{1}{2}}} \quad \text{as } k \to \infty.$$

References

[1] Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* 18, 178-189.

[2] Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: consistency properties. USAF Sch. Aviation Medicine, Rep. 4, Proj. 21-49-004.

[3] Kiefer, J. (1972). Iterated logarithm analogues for sample quantiles when $p_n \downarrow 0$. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* 1, 227-244.

[4] Loève, M. (1963). *Probability Theory*, 3rd. Ed., D. Van Nostrand Company, Princeton, N. J.

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

[6] Moore, D. S. and Henrichon, E. G. (1969). Uniform consistency of some estimates of a density function. *Ann. Math. Statist.* 40, 1499-1502.

[7] Moore, D. S. and Yackel, J. W. (1976). Consistency properties of nearest neighbor density function estimators. *Ann. Statist.* to appear.

[8] Wagner, T. J. (1973). Strong consistency of a nonparametric estimate of a density function. *IEEE Trans. Systems, Man and Cybernetics* 3, 289-290.