A NONPARAMETRIC ESTIMATE OF A MULTIVARIATE DENSITY FUNCTION[1]

By D. O. Loftsgaarden and C. P. Quesenberry

Montana State College

1. Introduction and summary. Let $x_1, \cdots, x_n$ be independent observations on a p-dimensional random variable $X = (X_1, \cdots, X_p)$ with absolutely continuous distribution function $F(x_1, \cdots, x_p)$. An observation x_i on X is $x_i = (x_{1i}, \cdots, x_{pi})$. The problem considered here is the estimation of the probability density function $f(x_1, \cdots, x_p)$ at a point $z = (z_1, \cdots, z_p)$ where f is positive and continuous. An estimator is proposed and consistency is shown.

The problem of estimating a probability density function has only recently begun to receive attention in the literature. Several authors [Rosenblatt (1956), Whittle (1958), Parzen (1962), and Watson and Leadbetter (1963)] have considered estimating a univariate density function. In addition, Fix and Hodges (1951) were concerned with density estimation in connection with nonparametric discrimination. Cacoullos (1964) generalized Parzen's work to the multivariate case. The work in this paper arose out of work on the nonparametric discrimination problem.

2. Preliminaries and notation. Let $d(x, z)$ represent the p-dimensional Euclidean distance function $|x - z|$. If z is the point given in the first paragraph of the preceding section, a p-dimensional hypersphere about z of radius r will be designated by $S_{r,z}$, i.e. $S_{r,z} = \{x \mid d(x, z) \leqq r\}$. The volume or measure of the hypersphere $S_{r,z}$ will be called $A_{r,z}$. $A_{r,z}$ is equal to $2r^p\pi^{p/2}/p\Gamma(p/2)$. Using this notation and noting that $A_{r,z} \to 0$ if and only if $r \to 0$, we have

$$f(z_1, \cdots, z_p) = \lim_{r\to 0} P(S_{r,z})/A_{r,z}, \tag{2.1}$$

i.e. there exists an R such that if $r < R$ then

$$|P(S_{r,z})/A_{r,z} - f(z_1, \cdots, z_p)| < \epsilon, \tag{2.2}$$

for arbitrary $\epsilon > 0$.

3. A consistent density function estimator. According to (2.2), $P(S_{r,z})/A_{r,z}$ can be made as near $f(z_1, \cdots, z_p)$ as one chooses by letting r approach zero. $P(S_{r,z})$ is unknown since it depends on the density f being estimated. The approach used here is to find a good estimate for $P(S_{r,z})$ and substitute this in the expression $P(S_{r,z})/A_{r,z}$ to obtain an estimate for f at z.

Let $\{k(n)\}$ be a non-decreasing sequence of positive integers (this can be

Received 4 September 1964; revised 22 December 1964.

[1] Work supported by National Aeronautics and Space Administration Research Grant NsG-562.

generalized to more general $k(n)$ with minor difficulty) such that

$$\lim_{n\to\infty} k(n) = \infty \tag{3.1}$$

and

$$\lim_{n\to\infty} k(n)/n = 0.$$

Once $k(n)$ is chosen and a sample $x_1, \cdots, x_n$ is available, $r_{k(n)}$ is determined as follows. It is chosen as the distance from z to the $k(n)$th closest x_i to z as determined by the Euclidean distance function. For $S_{r_{k(n)},z}$ the hypersphere of radius $r_{k(n)}$ about z and $A_{r_{k(n)},z}$ its measure, $\hat{f}_n(z)$ is defined as follows:

$$\begin{aligned}\hat{f}_n(z) &= \{(k(n) - 1)/n\}\{1/A_{r_{k(n)},z}\} \\ &= \{(k(n) - 1)/n\}\{p\Gamma(p/2)/2r_{k(n)}^p\pi^{p/2}\}.\end{aligned} \tag{3.2}$$

There is a basic difference between this estimator and those proposed by most of the authors mentioned above. Here a specific number of observations $k(n)$ is given and the distance to the $k(n)$th from z is measured. This determines $r_{k(n)}$ in (3.2). Parzen in the first case he considers, for example, specifies a distance from z, say $h(n)$, and counts the number of observations falling within this distance from z. This is equivalent to fixing $r_{k(n)}$ in (3.2) and then determining $k(n)$.

THEOREM 3.1. *The density estimator $\hat{f}_n(z)$ as given in* (3.2) *is consistent.*

PROOF. This proof makes use of the theory of coverages [cf. Wald (1943), Tukey (1947), or Wilks (1962)]. The ordering function necessary for the application of this theory is $|x - z|$ as given earlier. Using this theory we find that $P(S_{r_{k(n)},z}) = U_{k(n)}$ has a beta distribution with parameters $k(n)$ and $n - k(n) + 1$.

The first step in this proof is to show that $U_{k(n)}/A_{r_{k(n)},z} \to_P f(z_1, \cdots, z_p)$. A simple application of the Tchebysheff inequality yields

$$U_{k(n)} \to_P 0. \tag{3.3}$$

However, this can happen only if $A_{r_{k(n)},z} \to_P 0$ which in turn occurs only if $r_{k(n)} \to_P 0$. Let R be as defined in (2.2). There exists an N such that for $n > N$, and for arbitrary $\eta > 0$,

$$P\{r_{k(n)} < R\} > 1 - \eta. \tag{3.4}$$

This is sufficient to imply that

$$U_{k(n)}/A_{r_{k(n)},z} \to_P f(z_1, \cdots, z_p). \tag{3.5}$$

Rewriting (3.5) gives

$$\{n/(k(n) - 1)\}U_{k(n)}/\{n/(k(n) - 1)\}A_{r_{k(n)},z} \to_P f(z_1, \cdots, z_p). \tag{3.6}$$

If $\{n/(k(n) - 1)\}U_{k(n)} \to_P 1$ then it follows that

$$\hat{f}_n(z) = \{(k(n) - 1)/n\}\{1/A_{r_{k(n)},z}\} \to_P f(z_1, \cdots, z_p). \tag{3.7}$$

Using the fact that $U_{k(n)}$ has a beta distribution along with a simple application of the Tchebysheff inequality yields

$$\{n/(k(n) - 1)\}U_{k(n)} \to_P 1. \tag{3.8}$$

4. Comments. It should be noted that the estimator (3.2) is particularly easy to compute in practice. A choice of $k(n)$ must be made subject to (3.1). On the basis of some empirical work a value of $k(n)$ near $n^{\frac{1}{2}}$ appears to give good results.

In the preceding, the Euclidean distance function was used only because it seems to be a natural choice in many estimation problems. Any other metric would serve equally as well.

REFERENCES

CACOULLOS, T. (1964). Estimation of a Multivariate Density. Technical Report No. 40, Department of Statistics, University of Minnesota.

FIX, E. and HODGES, J. L. JR. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

TUKEY, J. W. (1947). Non-parametric estimation II. Statistically equivalent blocks and tolerance regions—the continuous case. *Ann. Math. Statist.* **18** 529–539.

WALD, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Statist.* **14** 45–55.

WATSON, G. S. and LEADBETTER, M. R. (1963). On the estimation of the probability density, I. *Ann. Math. Statist.* **34** 480–491.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20** 334–343.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.