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 $x_1,\ldots,x_n$ be n independent observations on a p-dimensional random variable $X = (X_1,\ldots,X_p)$ with absolutely continuous distribution function $F(x_1,\ldots,x_p)$. The problem considered here is the estimation of the probability density function

$$f(x_1,\ldots,x_p) = \frac{\partial^p F(x_1,\ldots,x_p)}{\partial x_1,\ldots,\partial x_p} \tag{1.1}$$

at a point $z = (z_1,\ldots,z_p)$ where f is positive and continuous.

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 f(x). In addition, Fix and Hodges (1951) were concerned to a limited degree with density estimation and Cacoullos (1964) generalized Parzen's work to a p-variate density $f(x_1,\ldots,x_p)$. These authors were principally motivated by the problem of estimating the hazard, or conditional rate of failure, function $f(x)/\{1 - F(x)\}$, where F(x) is the distribution function corresponding to f(x). An exception to this statement was Fix and Hodges (1951) who were concerned with nonparametric discrimination. The work in this paper also arose out of work on the nonparametric discrimination problem.

A problem which arises in one approach to density estimation is choosing a neighborhood about the point z where the density is being

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

estimated. Parzen (1962), e.g., chooses an h(n) which is a function of n and this h(n) determines the size of a neighborhood about z. This h(n) is a constant for each n and does not depend on the observations $x_1,\ldots,x_n$. In this paper a random neighborhood which is a function of n and of the observations $x_1,\ldots,x_n$ is chosen. Fix and Hodges (1951) use a somewhat similar approach in one special case which they consider but all other authors mentioned above use an approach similar to Parzen's. An estimate based on this random neighborhood is proposed and shown to be consistent. A special case is considered which illustrates how estimates can be obtained in certain cases other than those explicitly considered here.

The efficiency of the estimate proposed herein for finite sample size n is not studied in this paper. However, further work in this area is planned and will involve some numerical comparisons with other estimates.

2. Preliminaries and notation. Let $x_1,\ldots,x_n$ be a sample of n p-dimensional observations on a random variable $X = (X_1,\ldots,X_p)$. An observation x_i on X is $x_i = (x_{1i},\ldots,x_{pi})$. Assume X has an absolutely continuous distribution function $F(x_1,\ldots,x_p)$. The corresponding density function $f(x_1,\ldots,x_p)$, as given in (1.1), exists almost everywhere. An estimate is desired for the density f at a point $z = (z_1,\ldots,z_p)$ where f is positive and continuous.

Let d(X,Z) represent the p-dimensional Euclidean distance function $|X-Z|$. A p-dimensional sphere about z of radius r will be designated by $S_{r,z}$, i.e. $S_{r,z} = \{x|d(x,z) \leq r\}$. The volume or measure of the sphere $S_{r,z}$ will be called $A_{r,z}$, and $A_{r,z}$ is equal to $2r^p\pi^{p/2}/p\Gamma(p/2)$. Briefly

(2.1) $$d(X,Z) = |X-Z|,$$

(2.2) $$S_{r,z} = \{x|d(x,z) \leq r\} \text{ and}$$

[redacted]

(2.3) $$A_{r,z} = \text{Measure of } S_{r,z} = 2r^p\pi^{p/2}/p\Gamma(p/2).$$

Using this notation and noting that $A_{r,z} \longrightarrow 0$ if and only if $r \longrightarrow 0$, we have

(2.4) $$f(z_1,\ldots,z_p) = \lim_{r\longrightarrow 0} P\left\{X \in S_{r,z}\right\}/A_{r,z},$$

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

(2.5) $$\left|P\left\{X \in S_{r,z}\right\}/A_{r,z} - f(z_1,\ldots,z_p)\right| < \epsilon,$$

for arbitrary $\epsilon > 0$.

In the preceding paragraph the Euclidean distance function is used. There and in the work which follows any other metric could just as well have been used. The Euclidean distance function is being used simply because it seems the natural one to use here.

3. A consistent density estimator. According to (2.5), $P\left\{X \in S_{r,z}\right\}/A_{r,z}$ can be made as near $f(z_1,\ldots,z_p)$ as one chooses by letting r approach zero. $P\left\{X \in S_{r,z}\right\}$ is unknown since it depends on the density f being estimated. Therefore if a good estimate of $P\left\{X \in S_{r,z}\right\}$ can be found it can be substituted in the expression $P\left\{X \in S_{r,z}\right\}/A_{r,z}$ and this should provide a good estimate of the density f at z. This is the approach which will be used here.

Let $\{k(n)\}$ be a non-decreasing sequence of positive integers (this can be generalized to more general k(n) with minor difficulty) such that

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

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

These are very important conditions in the work of this paper. They will implicitly control the size of the neighborhoods being chosen about the point z in such a way that the proposed estimate of f(z) will be consistent. If $[\cdot]$ is the greatest integer function then $k(n) = [bn^{\delta}]$ where b is a constant and $0 < \delta < 1$ will satisfy the conditions in (3.1). In view of the way in which k(n) will be used in this paper we will also require that $k(n) \leq n$ for all n. This will restrict the choice of b somewhat. Further work is planned which should help in the problem of choosing of k(n) in practice.

The discussion in this paragraph is based on the theory of coverages (cf. Wald (1943), Tukey (1947), or Wilks (1962)). Let $x_1,\ldots,x_n$ be a sample of size n from a p-dimensional distribution of the type discussed earlier, i.e. $x_i = (x_{1i},\ldots,x_{pi})$ for $i = 1,\ldots,n$. An ordering function $\varphi(x) = |x-z|$ is introduced where $|x-z|$ is as defined in (2.1). Then $w = \varphi(x)$ is a random variable which has a continuous distribution function, say H(w). Consider the new random variables $w_1, w_2,\ldots,w_n$ where $w_i = \varphi(x_i)$ for $i = 1,\ldots,n$. Let the ordered w_i's be $w_{(1)},\ldots,w_{(n)}$. The coverages are $c_1 = H(w_{(1)})$, $c_2 = H(w_{(2)}) - H(w_{(1)}),\ldots,c_{n+1} = 1 - H(w_{(n)})$. The corresponding p-dimensional sample blocks $B_p^{(1)}, \ldots, B_p^{(n+1)}$ are the disjoint parts that the p-dimensional space is divided into by the ordering curves $\varphi(x) = w_{(i)}$ for $i = 1,\ldots,n$. $P\{B_p^{(i)}\} = c_i$ for $i = 1,\ldots,n$. The distance from z to the x_i closest to it is $w_{(1)}$. Therefore $B_p^{(1)}$ consists of those points inside a p-dimensional sphere about z of radius $w_{(1)}$. This sphere is the ordering curve $\varphi(x) = w_{(1)}$. $B_p^{(k)}$ consists of those points which are inside a p-dimensional sphere of radius $w_{(k)}$ about z but which are not in $B_p^{(1)},\ldots,B_p^{(k-1)}$ for $k = 1,\ldots,n$. $B_p^{(n+1)}$ consists of those points

outside a p-dimensional sphere of radius $w_{(n)}$ about z. For convenience we now set $w_{(k)} = r_k$ for k=1,...,n. The sum of the first k blocks, $B_p^{(1)} + \ldots + B_p^{(k)}$, consists of those points inside a sphere of radius r_k about z, viz. $S_{r_k,z}$. The sum of the corresponding coverages $c_1+\ldots+c_k$ is equal to $P\{X \in S_{r_k,z}\}$ which we set equal to U_k. By the theory in the references given earlier in this paragraph U_k has a Beta distribution with (k, n-k+1) degrees of freedom.

It is convenient to think of the k in the preceding paragraph as the k of (3.1) as this is the way we will use it in the sequel. It should be recalled that

$$\begin{aligned} EU_k &= k/(n+1) \\ \text{var } (U_k) &= k(n-k+1)/(n+1)^2\,(n+2) \end{aligned} \tag{3.2}$$

We now define an estimate for the density f at the point $z = (z_1,\ldots,z_p)$, where f is positive and continuous. Put

$$\begin{aligned} \hat{f}_n(z) &= [k/(n+1)]\ [1/A_{r_k,z}] \\ &= [k/(n+1)]\ [p\Gamma(p/2)/2r_k^p\pi^{p/2}], \end{aligned} \tag{3.3}$$

where r_k is as defined above. In other words, once k(n) is chosen and a sample $x_1,\ldots,x_n$ is available one determines r_k as the distance to the k^{th} closest x_i to z and then proceeds to compute $\hat{f}_n(z)$ as given in (3.3). It should be noted that this estimate is particularly easy to obtain in practice.

THEOREM 3.1. The density estimator $\hat{f}_n(z)$ as given in (3.3) is consistent.

Proof. The first step in this proof is to show that $f(z_1,\ldots,z_p)$ can be approximated by $P\{X \in S_{r_k,z}\}/A_{r_k,z}$. This is done by showing that $P\{X \in S_{r_k,z}\}/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p)$.

$P\{X \in S_{r_k,z}\}$ is equal to U_k. Therefore using (3.2) and the Tchebysheff inequality we have for arbitrary $\epsilon > 0$

$$P\{|U_k - k/(n+1)| \geq \epsilon\} \leq \operatorname{var}(U_k)/\epsilon^2 \tag{3.4}$$
$$= k(n-k+1)/(n+1)^2(n+2)\epsilon^2.$$

Using the conditions (3.1), the right hand side of (3.4) is seen to converge to zero. Thus for large n the right hand side of (3.4) can be made arbitrarily small. That is $U_k - k/(n+1) \xrightarrow{P} 0$. Using (3.1) again gives $k/(n+1) \longrightarrow 0$. Combining these two results gives

$$U_k = P\{X \in S_{r_k,z}\} \xrightarrow{P} 0. \tag{3.5}$$

However, this can happen only if the measure of $S_{r_k,z}$, viz. $A_{r_k,z}$, converges in probability to zero, by the continuity assumptions. This in turn can occur if and only if $r_k \xrightarrow{P} 0$.

Let R be as defined in (2.5). Since $r_k \xrightarrow{P} 0$, there exists an N such that if $n>N$, and for arbitrary $\eta > 0$

$$P\{r_k < R\} > 1 - \eta. \tag{3.6}$$

Using (2.5) and (3.6) the following statement can be made.

If $n > N$

$$(3.7) \qquad P\left\{\left|P\left\{X \in S_{r_k,z}\right\}/A_{r_k,z} - f(z_1,\ldots,z_p)\right| < \epsilon\right\} > 1-\eta,$$

where ϵ is as defined previously in (2.5). Thus

$$(3.8) \qquad P\left\{X \in S_{r_k,z}\right\}/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p).$$

This concludes the first part of the proof.

The concluding portion of the proof goes as follows.

By (3.8) $U_k/A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p)$ or rewriting this

$$(3.9) \qquad [(n+1)/k]U_k/[(n+1)/k]A_{r_k,z} \xrightarrow{P} f(z_1,\ldots,z_p).$$

If it can be shown that the numerator of (3.9), viz. $[(n+1)/k]U_k$, converges in probability to 1, then it will follow that the denominator, viz. $[(n+1)/k]A_{r_k,z}$, will converge in probability to $1/f(z_1,\ldots,z_p)$.

This last statement is equivalent to

$$(3.10) \qquad [k/(n+1)]\ [1/A_{r_k,z}] \xrightarrow{P} f(z_1,\ldots,z_p).$$

This is the desired conclusion of the theorem. It remains then to show that $[(n+1)/k]U_k \xrightarrow{P} 1$.

Consider $[(n+1)/k]\ U_k$.

$$E[((n+1)/k)U_k] = 1$$

(3.11)

$$\text{and} \qquad \text{var}\ [((n+1)/k)U_k] = [1/k][(n-k+1)/(n+2)].$$

The variance in (3.11) approaches zero by (3.1). Using the Tchebysheff inequality and for arbitrary $\epsilon > 0$

$$(3.12) \qquad P\left\{\left| [(n+1)/k]U_k - 1\right| \geq \epsilon\right\} \leq \text{var}\,[((n+1)/k)U_k]/\epsilon^2 = [1/k]\,[(n-k+1)/(n+2)\epsilon^2].$$

Since for large n the right hand side of (3.12) can be made arbitrarily small we have

$$(3.13) \qquad [(n+1)/k]U_k \xrightarrow{P} 1.$$

Thus (3.10) follows from the argument above and the theorem is proved.

It was mentioned earlier that the neighborhoods determined about the point z would be random. It can now be noted that the neighborhoods about z are essentially determined by r_k which is indeed a random variable. It depends on the observations $x_1,\ldots,x_n$ which seems to be a desirable feature.

4. A modification of $\hat{f}_n(z)$ for a special case. In this section we consider modifying $\hat{f}_n(z)$ in order to use it at a point which does not satisfy the continuity assumptions of the previous work. A special problem will be considered but a similar technique can be applied to many other cases. Fix and Hodges (1951) have encountered the particular problem considered here in their work on nonparametric discrimination.

Let X be a univariate random variable with $f(x) = 0$ for $x < a$, and f is positive and continuous on $[a,\infty)$. An estimate for f(a) is desired. By definition f(a) is

$$(4.1) \qquad f(a) = \lim_{h \to 0} \frac{(F(a+h) - F(a))}{h} = \lim_{h \to 0} \frac{F(a+h)}{h} .$$

This statement corresponds to (2.4) in the general case. In this case h is the length of the interval $[a, a+h)$. There is no central interval about a where f is continuous as f is continuous on the right only at the point a. The theory of coverages used in the development of $\hat{f}_n(z)$ in the previous section required central intervals or neighborhoods about the point z. Thus some slight modifications are required.

The observations $x_{(1)},\ldots,x_{(n)}$ are ordered as to distance from $z = a$ since X is univariate and since $f(x) = 0$ for $x < a$ which implies that all $x_{(i)} \geq a$. Therefore the ordering function $d(x,z)$ is not necessary here. Consider the blocks $B_1^{(1)} = [a,x_{(1)}]$, $B_1^{(2)} = (x_{(1)},x_{(2)}],\ldots,$ $B_1^{(n+1)} = (x_{(n)},\infty)$.

Corresponding to these blocks are coverages $c_1 = F(x_{(1)})$, $c_2 = F(x_{(2)}) - F(x_{(1)}),\ldots,c_{n+1} = 1 - F(x_{(n)})$. Now

$$U_k = c_1+\ldots+ c_k$$

$$= F(x_{(k)}) \tag{4.2}$$

and

$$B_1^{(1)} +\ldots+ B_1^{(k)} = [a,x_{(k)}].$$

The measure of this last sum of blocks is $x_{(k)} - a$. The variable U_k has the same distribution as it had previously. Therefore the procedure now in getting an estimate $\hat{f}_n(a)$ for $f(a)$ is the same as that in developing $\hat{f}_n(z)$ in the general case. This leads us to define

$$\hat{f}_n(a) = [k/(n+1)]\ [1/(x_{(k)} - a)],$$

which is a consistent estimate for $f(a)$.

REFERENCES

Cacoullos, T. (1964). Estimation of a Multivariate Density. University of Minnesota, Dept. of Statist., Technical Report No. 40.

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

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., Series B 20 334-343.

Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.