The Annals of Mathematical Statistics
1969, Vol. 40, No. 4, 1496–1498

DENSITY ESTIMATION BY ORTHOGONAL SERIES[1]

BY GEOFFREY S. WATSON

There has been much discussion of density estimation by kernel methods (e.g., Whittle (1958), Parzen (1962), Watson and Leadbetter (1963)). Thus, given a random sample $x_1, \cdots, x_n$ from the density $f(x)$, the estimator

$$\hat{f}_n = 1/n \sum_{k=1}^{n} \delta_n(x - x_k) \tag{1}$$

has minimum integrated mean square error (M.I.S.E.) if

$$\phi_{\delta_n} = |\phi_f|^2/\{n^{-1}[1 + (n-1)|\phi_f|^2]\} \tag{2}$$

where $\phi_{\delta_n} = \int e^{ixt}\delta_n(x)\, dt$, $\phi_f = \int e^{ixt} f(x)dx$. Whittle introduced Bayesian arguments to find the optimum kernel; he assumed a covariance function for the values of $f(x)$ at different x values.

It is obvious that orthogonal series estimates could be used if it is assumed that $f(x) = \sum_0^\infty \alpha_m \varphi_m(x)$, where $(\varphi_m(x))$ is an orthonormal basis. Several papers (Cencov (1962), van Ryzin (1966), Schwartz (1967), and Kronmal and Tarter (1968)) have considered this possibility. The estimator that springs to mind is

$$f_n^*(x) = \sum_0^\infty \lambda_m(n)\, a_m \varphi_m(x) \tag{3}$$

where

$$a_m = n^{-1} \sum_{k=1}^{n} \varphi_m(x_k). \tag{4}$$

and the sequence $\{\lambda_m(n)\}$ is chosen to improve the properties of $f_n^*(x)$ e.g., to make $f_n^*(x)$ a M.I.S.E. estimator in its class. The papers of Cencov, van Ryzin and Schwartz use a special sequence $\{\lambda_m(n)\}$; they set $\lambda_m(n) = 1$, $m = 1, \cdots, M(n)$, $\lambda_m(n) = 0$, $m > M(n)$, and concern themselves, in part, with the determination of $M(n)$. Now

$$J = E \int (f(x) - f_n^*(x))^2\, dx = \sum_0^\infty E(\alpha_m - \lambda_m(n) a_m)^2$$
$$= \sum_0^\infty \{\alpha_m^2(1 - \lambda_m(n))^2 + n^{-1}\lambda_m^2(n) \operatorname{var}(\varphi_m(x))\}.$$

Hence

$$\lambda_m(n) = \alpha_m^2/[\alpha_m^2 + n^{-1} \operatorname{var}(\varphi_m(x))], \tag{5}$$

i.e.,

$$\lambda_m(n) = \{\alpha_m^2/E(\varphi_m^2)\}/\{n^{-1}[1 + (n-1)\, \alpha_m^2/E(\varphi_m^2)]\}. \tag{5'}$$

Received 31 May 1968.

[1] Research supported by the Office of Naval Research under contract NONR 4010(09) awarded to the Department of Statistics, The Johns Hopkins University. This paper in whole or in part may be reproduced for any purpose of the United States Government.

Thus

$$\text{(6)} \qquad \min J = \sum_0^\infty [\alpha_m^2(E(\varphi_m^2) - \alpha_m^2)]/[E(\varphi_m^2) + (m-1)\alpha_m^2]$$
$$= \sum_0^\infty \alpha_m \operatorname{var}(\varphi_m)/(\operatorname{var}(\varphi_m) + \alpha_m^2).$$

Since $\alpha_m^2/E(\varphi_m^2)$, like $|\phi_f|^2$, lies in (0, 1) the similarity of (2) and (5) is striking. As $n \to \infty$, $\lambda_m(n) \to 1$ for fixed m. Furthermore, if one defines a symmetric function $k_n(\cdot, \cdot)$ by

$$\text{(7)} \qquad k_n(x, y) = \sum_0^\infty \lambda_m(n)\varphi_m(x)\varphi_m(y),$$

then

$$\text{(8)} \qquad n^{-1} \sum_{k=1}^n k_n(x, x_k) = \sum_0^\infty \lambda_m(n) a_m \varphi_m(x).$$

We note that $k_n(x, y)$ tends (as $n \to \infty$) to the Dirac Delta function with representation $\sum \varphi_m(x)\varphi_m(y)$. The identity would then be complete if the formal manipulations above could be justified. They can be if we work on a finite interval and f and $\{\varphi_m\}$ are square integrable. The writer has used the series method, without the weights $\{\lambda_m(n)\}$, many times in connection with goodness-of-fit tests (e.g., Watson (1967a), (1967b)) where the range of the distribution is (0, 1) and $\varphi_m(x) = \exp(2\pi imx)$.

The choice (5) serves more as a standard than as a practical suggestion since $\alpha_m/E(\varphi_m^2)$ will rarely be known in practice. (The same is true of (2).) Since, from (5) with m fixed, $\lambda_m(n) \to 1$ as $n \to \infty$, there is a strong reason to use

$$\text{(9)} \qquad \lambda_m(n) = 1, \quad m = 0, \cdots, M(n)$$
$$= 0, \quad m > M(n)$$

in which case

$$\text{(10)} \qquad J = \sum_0^{M(n)} n^{-1} \operatorname{var}(\varphi_m) + \sum_{M(n)+1}^\infty \alpha_m^2.$$

The comparison of (6) and (10) suggests that the choice (9) will be nearly optimal when the $\alpha_m (m \leqq M(n))$ are large compared with $n^{-1} \operatorname{var}(\varphi_n)$, and the $\alpha_m (m > M(n))$ are negligible. As this should usually be the case for some $M(n)$, it is seen that our suggestion cannot lead to estimates much better than those based on (9), provided $M(n)$ is chosen to minimize (10). This leads to the estimators discussed by Cencov. To discuss our estimators further theoretically, classes of densities would have to be defined in which the asymptotic behavior of $\alpha_m^2/E(\varphi_m^2)$ is the same.

In practice both the series and kernel methods must be used experimentally (i.e., (1) must be tried with different $\delta_n(\cdot)$, (2) with different $\{\lambda_m(n)\}$, (9) with different $M(n)$) until the result seems acceptable. Thus there seems very little, in either theory or practice, to recommend one method more than the other.

REFERENCES

[1] Cencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet. Math.* **3** 1559–1562.

[2] KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

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

[4] SCHWARTZ, STUART C. (1967). Estimation of probability density by an orthogonal series. *Ann. Math. Statist.* **38** 1261–1265.

[5] VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhyā Ser. A* **28** 261–270.

[6] WATSON, G. S. (1967a). Another test for the uniformity of a circular distribution. *Biometrika* **54** 675–677.

[7] WATSON, G. S. (1967b). Some problems in the statistics of directions. To be published in *Bull. Inst. Internat. Statist.* **42**.

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

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