Univariate Density Estimation by Orthogonal Series

H. D. Brunk

Biometrika, Vol. 65, No. 3 (Dec., 1978), 521-528.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197812%2965%3A3%3C521%3AUDEBOS%3E2.0.CO%3B2-O

Biometrika is currently published by Biometrika Trust.

Biometrika (1978), **65**, 3, *pp.* 521–8
With 1 *text-figure*
Printed in Great Britain

Univariate density estimation by orthogonal series

BY H. D. BRUNK
Department of Statistics, Oregon State University, Corvallis

SUMMARY

Orthogonal series estimators of univariate densities are proposed that are derived from and motivated by kernel estimators optimal in Whittle's (1958) sense. A preliminary fit of the data from within a one or two parameter class of densities plays the role of a prior mean density. The ratio of the true density and the prior mean density is assumed to have a series expansion in terms of functions orthogonal with respect to the prior mean density. Coefficients of terms of the series are given a joint prior distribution according to which they are independent, with zero means.

Some key words: Kernel estimator; Nonparametric density estimation; Orthogonal series estimator.

1. INTRODUCTION

The vast literature on estimation of univariate densities may be divided roughly into two categories: parametric and nonparametric. A parametric method typically uses a class of models capable of being described in terms of a few parameters. In using a nonparametric method, the investigator envisages a much richer class of models. The method studied in the present paper belongs to the latter category, though many parameters are involved in its description. Excellent reviews of nonparametric estimation of densities are to be found in papers by Wegman (1972a, b) and by Tarter & Kronmal (1976), which also contains references to the very substantial literature on kernel estimators and on the subclass of estimators based on expansions in orthogonal series.

In stark contrast to criteria for asymptotic optimality of sequences of estimators, the criterion introduced by Whittle (1958) is directly applicable to samples of arbitrary size. Whittle's outstanding paper has not escaped notice; see, for example, Dickey (1969) and Fellner (1974). The present paper presents a way of implementing Whittle's approach via orthogonal series representations of densities. The method is directed to two kinds of problem: (a) the estimation of a density at a point, and (b) the study of the shape of the graph of the density. To achieve a third possible goal, estimation of the 'tails' of the density, would appear to require relatively stronger assumptions, and is not considered here. As to (a), it may often happen that the investigator is confident that the density can be fitted adequately near the point in question within a class of densities described in terms of only a few parameters; in such case one should expect little improvement from the use of methods discussed here, though they are simple and easy to apply. As to (b), various admirable methods of smoothing histograms are available; see, for example, Boneva, Kendall & Stefanov (1971), Good & Gaskins (1971), Wahba (1976) and unpublished work of Leland Stewart on Bayesian estimation using splines. But the method studied here represents a refinement of existing methods of estimation using orthogonal series, of wide applicability, that the writer finds especially appealing on both aesthetic and practical grounds.

2. WHITTLE ESTIMATORS

Let f_0 be a given density, thought of as an initial guess or estimate, prior to the observations, of an unknown density f. Let $\{\phi_r(x)\}$ be a sequence of functions orthonormal with respect to f_0, with $\phi_0(x) \equiv 1$. Assume that there is a sequence $\beta_0, \beta_1, \ldots$ of real numbers such that

$$f(x) = f_0(x) \sum_{r=0}^{\infty} \beta_r \phi_r(x). \tag{1}$$

In order that $\int f(x)\,dx = 1$, we must have $\beta_0 = 1$. Let $\tilde{x}_1, \ldots, \tilde{x}_n$ be a simple random sample from the distribution having density f; here and elsewhere a tilde will be used, where feasible because of the absence of another superior marking, to denote a quantity modelled as random. Whittle (1958) noted that the estimator

$$\hat{f}(x) = f_0(x)\{1 + \sum_{r=1}^{\infty} \hat{\beta}_r \phi_r(x)\}, \tag{2}$$

where

$$\hat{\beta}_r = \left(\frac{n}{n-1+\pi_r}\right) \bar{\phi}_r, \tag{3}$$

$$\bar{\phi}_r = n^{-1} \sum_{j=1}^{n} \phi_r(\tilde{x}_j) \quad (r = 1, 2, \ldots), \tag{4}$$

and π_r is a prior precision, $\pi_r = \{\text{var}(\tilde{\beta}_r)\}^{-1}$ $(r = 1, 2, \ldots)$, is a kernel estimator optimal in his sense, described below. Virtually the same estimator was also obtained as an approximate posterior mode by Brunk & Pierce (1974).

A kernel estimator (Rosenblatt, 1956) is of the form

$$\hat{f}(x) = \sum_{j=1}^{n} K_n(x, \tilde{x}_j)/n, \tag{5}$$

where $K_n(.,.)$ is a prescribed function of two real arguments. If f is the density of a distribution over an interval I, Whittle assumes the ordinates $\{\tilde{f}(x) : x \in I\}$ given a prior distribution. For given n and fixed x in I, the function $K_n(x, .)$ and the corresponding estimator $\hat{f}(x)$ are optimal in Whittle's sense if $\hat{f}(x)$ provides the minimum value

$$D^2(x) = \min E\{\hat{f}(x) - \tilde{f}(x)\}^2, \tag{6}$$

where $\hat{f}$ ranges over the class of kernel estimators. Here the expectation is calculated according to the joint distribution of $\tilde{x}_1, \ldots, \tilde{x}_n$ and $\{\tilde{f}(y) : y \in I\}$. Only first and second moments of $\{\tilde{f}(y) : y \in I\}$ enter in the computation of (6).

We note in passing that when only a single observation x_1 is made, the solution kernel $K_1(.,.)$ and the estimate $\hat{f}$ are given by

$$\hat{f}(x) = K_1(x, x_1) = E\{\tilde{f}(x)\tilde{f}(x_1)\}/E\{\tilde{f}(x_1)\}. \tag{7}$$

3. PRIOR DISTRIBUTION OF COEFFICIENTS

In the present adaptation of Whittle's method we assume that f has an expansion (1), and impose a prior distribution on $\{\tilde{f}(y) : y \in I\}$ by prescribing a joint distribution for the coefficients $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$. We note that for fixed k, $\beta_1, \ldots, \beta_k$ have an interpretation in terms of best fit: $1 + \Sigma \beta_r \phi_r$, where the summation is over $r = 1, \ldots, k$, gives the 'best' fit to f/f_0 in the class of linear combinations g of $\{\phi_0, \phi_1, \ldots, \phi_k\}$, in the sense of minimizing $\int\{(f/f_0) - g\}^2 f_0$. But since the functions $\{\phi_r\}$ are orthonormal with respect to f_0, each individual coefficient has a similar interpretation: for each r, $\beta_r \phi_r$ gives the best fit to f/f_0 in the class of multiples of ϕ_r, independently of values of β_s for $s \neq r$. This suggests a joint prior distribution for $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ according to which they are independent. Further, we shall have $E\{\tilde{f}(x)\} = f_0(x)$ for each x, provided that we set $E(\tilde{\beta}_r) = 0$ $(r = 1, 2, \ldots)$. The specification of the moments of first and second order of $\{\tilde{\beta}_r; r = 1, 2, \ldots\}$ is completed by specifying the variances, $\sigma_r^2 = \text{var}(\tilde{\beta}_r)$ $(r = 1, 2, \ldots)$. In subsequent discussion, the particular functions $\{\phi_r\}$ used are ordered so that smoothness decreases as r increases. An opinion that f is smooth is expressed by requiring that coefficients β_r of large index r be near 0. Thus we shall choose prior variances so that σ_r^2 decreases to 0 as $r \to \infty$. Thus henceforth

$$\beta_0 = 1, \quad E(\tilde{\beta}_r) = 0, \quad \text{cov}(\tilde{\beta}_r, \tilde{\beta}_s) = \delta_{rs}/\pi_r \quad (r, s = 1, 2, \ldots), \tag{8}$$

where $\delta_{rr} = 1$, $\delta_{rs} = 0$ $(r \neq s)$, and where $\pi_r = 1/\sigma_r^2$. In this context, Whittle's formulae (1958, p. 342) yield an optimal estimator $\hat{f}(x)$ given by (2), (3) and (4). It is remarkable that the right-hand member of (3), which with (2) provides the minimum value (6) for given, fixed x, does not depend on x.

It should be observed that the truncated series estimator

$$\hat{f}_1(x) = f_0(x) \sum_{r=0}^{m} \bar{\phi}_r \phi_r(x)$$

is optimal for an investigator for whom $E(\tilde{\beta}_r) = 0$ $(r = 1, 2, \ldots)$, $\sigma_r^2 = 1$ $(r = 1, \ldots, m)$ and $\sigma_r^2 = 0$ for $r > m$. One would ordinarily expect somewhat better results using a more realistic prior distribution than this will appear to be in most situations.

One should note also that prior distributions of the kind described may give at least a small positive probability that the random density f will take negative values. In particular, Whittle (1958, p. 343) shows that if $\tilde{f}(x)$ is to be positive for all x with probability 1, one must prescribe $\pi_r \geqslant 1$ $(r = 1, 2, \ldots)$; but these conditions may not be sufficient. In practice, if the observed data are reasonably consistent with the prior mean f_0, and if reasonably large precisions π_r are specified, one will not expect negative estimates $\hat{f}(x)$.

Typically, many increasing sequences $\{\pi_r\}$ will appear to the investigator to be reasonable specifications of prior precisions. Some checks of reasonableness are available. He may examine the corresponding specification of covariances of $\{\tilde{f}(x) : x \in I\}$, for which Whittle gives the formula

$$\text{cov}\{\tilde{f}(x), \tilde{f}(y)\} = f_0(x) f_0(y) \sum_{r=1}^{\infty} \sigma_r^2 \phi_r(x) \phi_r(y). \tag{9}$$

The corresponding kernel,

$$K_n(x, y) = f_0(x) \left[1 + \sum_{r=1}^{\infty} \left\{\frac{n}{(n-1+\pi_r)}\right\} \phi_r(x) \phi_r(y)\right], \tag{10}$$

may be more readily interpreted. In particular, if one imagines a sample x_1 of size 1 observed, then from (7) we have

$$\hat{f}(x) = K_1(x, x_1) = E\{\tilde{f}(x) \tilde{f}(x_1)\}/f_0(x_1) = f_0(x) [1 + \Sigma \sigma_r^2 \phi_r(x) \phi_r(x_1)].$$

One sometimes determines prior distributions on parameters by expressing them in terms of hyperparameters, which in turn are given prior distributions. It is interesting that in the present context a rather simple and appealing prior distribution of this kind leads to specifications of the form $\text{var}(\tilde{\beta}_r) = c\rho^{-r}$ $(r = 1, 2, \ldots)$. The model to be described involves hyperparameters $\tilde{u} = (\tilde{u}_1, \tilde{u}_2, \ldots)$, interpreted as logarithms of ratios of variances of successive parameters $\tilde{\beta}_r$. Given $\tilde{u} = u$, let $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ be independent with

$$E(\tilde{\beta}_r \mid \tilde{u} = u) = 0, \quad E(\tilde{\beta}_r^2 \mid \tilde{u} = u) = \exp\left(-\sum_{j=1}^{r} u_j\right) \quad (r = 1, 2, \ldots).$$

Let λ and a_r be positive $(r = 1, 2, \ldots)$. Let $\tilde{u}_1, \tilde{u}_2, \ldots$ be independent, gamma-distributed random variables, the density of $\tilde{u}_r$, evaluated at $y > 0$, being proportional to

$$\exp\{(a_r - 1)\log y - \lambda y\}.$$

Then

$$\operatorname{var}(\tilde{\beta}_r) = E\{E(\tilde{\beta}_r^2 \mid \tilde{u})\} = E\left\{\exp\left(-\sum_{j=1}^{r} \tilde{u}_j\right)\right\} = \exp\left[\left(\sum_{j=1}^{r} a_j\right)\log\{\lambda/(\lambda+1)\}\right] \quad (r = 1, 2, \ldots).$$

If $a_2 = a_3 = \ldots = a$, then $\pi_r = \{\operatorname{var}(\tilde{\beta}_r)\}^{-1} = c\rho^r$ $(r = 1, 2, \ldots)$, where $\rho = \{(\lambda+1)/\lambda\}^a > 1$ and $c = \{\lambda/(\lambda+1)\}^{a_1 - a}$.

4. Precision of estimators

For a hypothetical investigator whose prior precisions are given by a sequence $\{\pi_r\}$, Whittle (1958, p. 342) gives the following formula for expected squared error at x, using $\hat{f}$ as given by (2) and (3):

$$D^2(x) = \{f_0(x)\}^2 \sum_{r=1}^{\infty} \frac{\pi_r - 1}{\pi_r(\pi_r + n - 1)} \{\phi_r(x)\}^2. \tag{11}$$

One can make a crude approximation to the posterior distribution of $\tilde{f}(x)$ as follows, assuming n large, and $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ independent and approximately normally distributed, with $E(\tilde{\beta}_r) = 0$, $\operatorname{var}(\tilde{\beta}_r) = \sigma_r^2 = 1/\pi_r$ $(r_r^2 = 1, 2, \ldots)$. We have

$$\bar{\phi}_r = n^{-1} \sum_{j=1}^{n} \phi_r(\tilde{x}_j),$$

so that

$$\{\operatorname{cov}(\bar{\phi}_r, \bar{\phi}_s)\}_{\tilde{\beta}=\beta} = n^{-1}\{\operatorname{cov}\{\phi_r(\tilde{x}), \phi_s(\tilde{x})\}\}_{\tilde{\beta}=\beta},$$

where $\tilde{x}$ has density $f = f_0 \Sigma \beta_r \phi_r$. Since the functions $\{\phi_r\}$ are orthonormal with respect to f_0, we have

$$\{\operatorname{cov}(\bar{\phi}_r, \bar{\phi}_s)\}_{\tilde{\beta}=0} = n^{-1}\delta_{rs}.$$

For the trigonometric series discussed below one has, in fact,

$$E\{\operatorname{cov}(\bar{\phi}_r, \bar{\phi}_s) \mid \tilde{\beta}\} = n^{-1}\delta_{rs}.$$

According to the proposed rough approximation, the $\bar{\phi}_r$ are independent observations on normal distributions $N(\beta_r, 1/n)$, respectively. The $\{\tilde{\beta}_r\}$ would then be independent according to their joint posterior distribution, with

$$E(\tilde{\beta}_r \mid \bar{\phi}_r) = n\bar{\phi}_r/(n + \pi_r), \quad \operatorname{var}(\tilde{\beta}_r \mid \bar{\phi}_r) = 1/(n + \pi_r) \quad (r = 1, 2, \ldots).$$

Set $\bar{\phi} = (\bar{\phi}_1, \bar{\phi}_2, \ldots)$. For fixed x, $\tilde{f}(x)$ is normal with

$$E\{\tilde{f}(x) \mid \bar{\phi}\} = f_0(x)\left[1 + \sum_{r=1}^{\infty}\{n/(n + \pi_r)\}\bar{\phi}_r \phi_r(x)\right] \simeq \hat{f}(x),$$

$$\operatorname{var}\{f(x) \mid \bar{\phi}\} = \{f_0(x)\}^2 \sum_{r=1}^{\infty}\{1/(n + \pi_r)\}\phi_r^2(x). \tag{12}$$

The sampling expectation of the squared error is the sum of two terms, the variance and the square of the bias, $B(x)$. If

$$\hat{f}(x) = f_0(x)\left\{1+\sum_{r=1}^{\infty}\lambda_{nr}\bar{\phi}_r\,\phi_r(x)\right\},$$

these are given approximately by

$$\operatorname{var}\{\hat{f}(x)\} \simeq n^{-1}\{f_0(x)\}^2\sum_{r=1}^{\infty}\{\lambda_{nr}\,\phi_r(x)\}^2, \tag{13}$$

$$B(x) \simeq f_0(x)\sum_{r=1}^{\infty}(1-\lambda_{nr})\,\beta_r\,\phi_r(x). \tag{14}$$

A crude estimate of the bias is obtained by replacing β_r by $\lambda_{nr}\bar{\phi}_r$:

$$\hat{B}(x) = f_0(x)\sum_{r=1}^{\infty}(1-\lambda_{nr})\,\lambda_{nr}\,\bar{\phi}_r\,\phi_r(x). \tag{15}$$

Watson (1969) discusses, briefly, estimators of the form $\Sigma_r\,\lambda_{nr}\,\bar{\phi}_r\,\phi_r(x)$. Such estimators are readily motivated as representing an attempt to damp out oscillations of the partial sums of the series $\Sigma_r\,\bar{\phi}_r\,\phi_r(x)$. The multipliers $\lambda_{nr} = n/(n-1+\pi_r)$ that appear in (13) have desirable features from this point of view: λ_{nr} increases to 1 as $n\rightarrow\infty$ for fixed r, while λ_{nr} decreases to 0 as $r\rightarrow\infty$ for fixed n, if π_r increases to ∞.

When one wishes to apply the method in order to estimate a particular density, the question of greatest interest is: is there a reasonable way to select prior precisions for that particular situation, so that one is assured of an acceptably smooth estimate, while at the same time there is sufficient latitude for the estimate to differ appreciably from f_0 if the data suggest it? The principal focus thus is on the individual situation. But it may be in order also to make a remark or two about average performance, in terms, for example, of expected integrated squared error. The author made some calculations in connexion with a few examples, using trigonometric series on $[0, 1]$, and exponentially increasing precisions $\{\pi_1, \pi_2, \ldots\}$. The results suggest that the best one can do with Whittle estimators may be only slightly better than the best one can do using a truncated series; for results with truncated series, see Kronmal & Tarter (1968). But to use one term too many or one term too few in the series may imply a substantial increase in expected integrated squared error, while for the Whittle estimators the expected integrated squared error appears relatively insensitive to change in the rate of exponential increase of the precisions. Results of these calculations are available from the author on request.

5. Selection of a specific estimator

The choice of f_0 available for use with well-known standard orthonormal systems is limited. But one can easily adapt a device suggested by Neyman (1937) in a somewhat similar context, and also by Fellner (1974), to obtain a system orthonormal with respect to an arbitrarily prescribed f_0. Let $\tilde{x}$ have unknown density f, and let f_0 be an initial guess at f. Set

$$F_0(x) = \int_{-\infty}^{x} f_0(t)\,dt,$$

and make the probability integral transformation, $\tilde{z} = F_0(\tilde{x})$. Then the range of $\tilde{z}$ is $[0, 1)$. The density, p, of $\tilde{z}$ is

$$p(z) = f\{F_0^{-1}(z)\}/f_0\{F_0^{-1}(z)\}, \tag{16}$$

and an initial guess is $p_0(z) = 1$ $(0 \leqslant z \leqslant 1)$. With p estimated by $\hat{p}$, the corresponding estimate of f is

$$\hat{f}(x) = \hat{p}\{F_0(x)\} f_0(x). \tag{17}$$

If $\hat{p}$ is a Whittle estimator of p, that is, if it is a kernel estimator minimizing $E\{\hat{p}(z) - \tilde{p}(z)\}^2$ for each z, then also $\hat{f}$ minimizes $E\{\hat{f}(x) - \tilde{f}(x)\}^2$ in the class of kernel estimators $\hat{f}$, for each x. There is, of course, a change of variable determined by (16) in the correlation:

$$\text{corr}\,\{\tilde{f}(x), \tilde{f}(y)\} = \text{corr}\,[\tilde{p}\{F_0(x)\}, \tilde{p}\{F_0(y)\}].$$

Further $\hat{f}$ is the kernel estimator

$$\hat{f}(x) = n^{-1} \sum_{j=1}^{n} K_n(x, x_j),$$

where

$$K_n(x, y) = f_0(x)\, k_n\{F_0(x), F_0(y)\},$$

$k_n(.,.)$ being the kernel associated with $\hat{p}$.

The author has done some simulation experiments, estimating a density positive on $[0, \infty]$; details are available on request. In his limited experience he has gained reasonable confidence in his ability to select reasonable precisions $\{\pi_1, \pi_2, \ldots\}$ only when f_0 fits the data rather well. He therefore proposes an *ad hoc* estimator, guided by Bayesian considerations outlined above.

Step 1. Choose a parametric family involving one or two parameters; from this family, select one as f_0 that fits the data rather well.

Step 2. Calculate the estimate $\hat{f}$, using $\pi_r = 5^r$ $(r = 1, 2, \ldots)$.

In the simulation examples this choice of $\{\pi_1, \pi_2, \ldots\}$, together with a choice of f_0 as described in step 1 above, has yielded a smoothly decreasing f.

The possibility of using the data to help suggest values of π_r was considered. Estimates of prior variances were calculated in a number of simulation experiments using various f and f_0, for $r = 1, \ldots, 50$. Typically, these estimates were small, the largest smaller than 0·1; and they appeared to vary randomly about zero. In several cases isotonic regression was used to fit the estimates by a nondecreasing sequence of numbers, which was then smoothed by eye. In each case the estimates $\hat{f}$ then computed lacked a reasonable degree of smoothness; they reflected the vagaries of the histogram more faithfully than did the estimates $\hat{f}$ that used more rapidly decreasing prior variances. In all simulation experiments carried out, the use of prior precisions $\pi_r = 5^r$ $(r = 1, 2, \ldots)$ yielded reasonably smooth estimates $\hat{f}$ that were monotonically decreasing, or nearly so, while substantially less rapidly increasing π_r did not consistently do so.

Some appreciation of the relative contributions of histogram and prior mean can be gained from examination of Fig. 1(a), (b) and (c). Each shows a true density f, positive on $[0, \infty]$, a grouped histogram of a sample of 300, a prior mean density f_0, and the Whittle estimate using $\pi_r = 5^r$ $(r = 1, 2, \ldots)$.

A further small simulation study of the precision of the estimator $\hat{f}(0)$ was undertaken. For this study 25 pseudorandom samples of size $n = 300$ were taken from the 'folded normal' distribution with density

$$f(x) = \sqrt{(2/\pi)}\, \sigma^{-1} \exp\left(-\tfrac{1}{2}x^2/\sigma^2\right) \quad (x \geqslant 0)$$

with $\sigma = 39{\cdot}4953$, so that $f(0) = 1/49{\cdot}5 = 0{\cdot}0202$. The same logistic density was used as prior

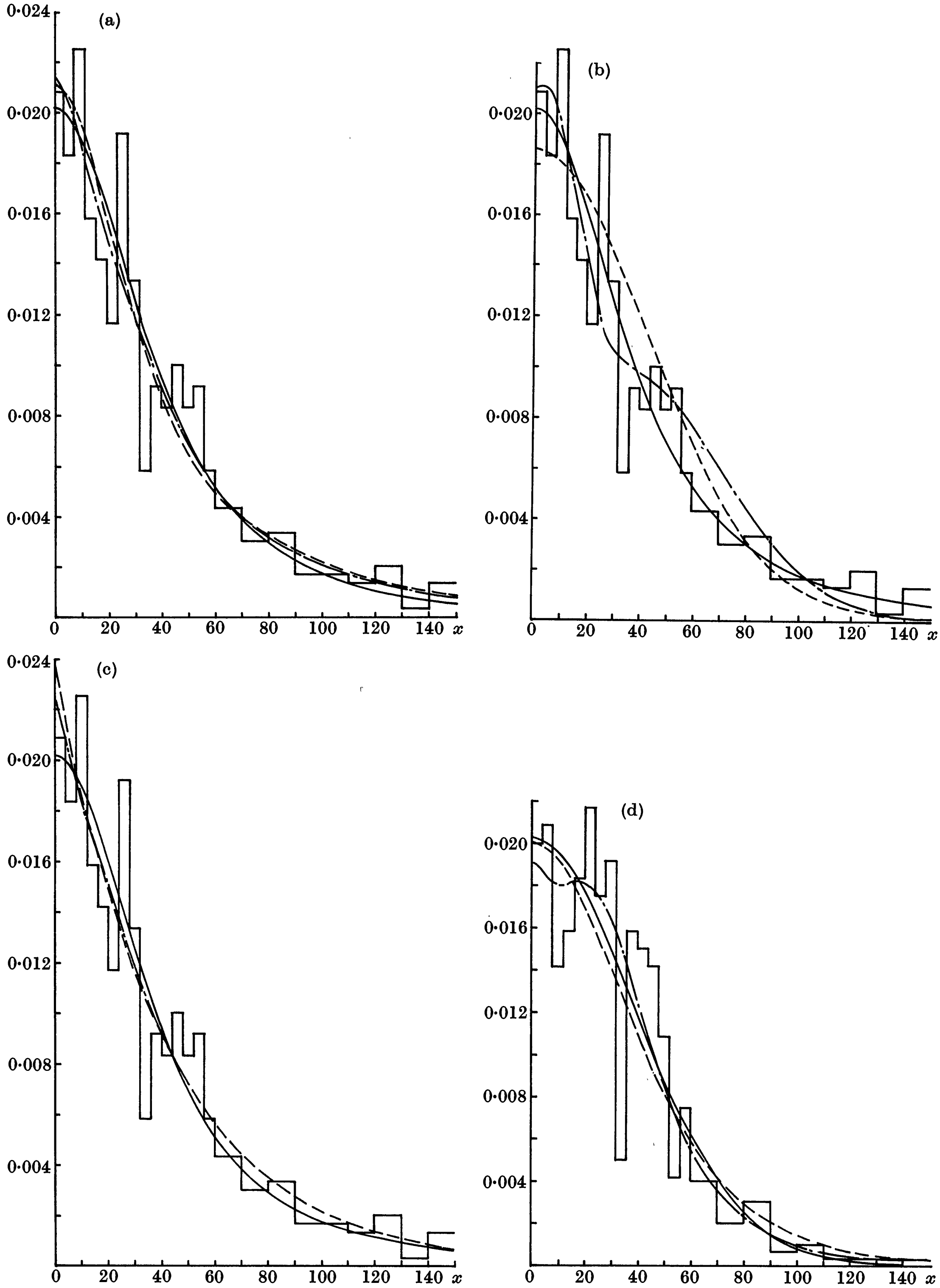

Fig. 1. Whittle estimates of folded scaled t densities and folded normal densities. Solid curve, population density; rectangular graph, histogram of grouped sample of 300; dashed curve, prior mean density; short- and long-dashed curve, Whittle estimate. (a) Population, folded scaled t_2; prior, mixture of folded logistics. (b) Population, folded scaled t_2; prior, folded normal (with same median as data). (c) Population, folded scaled t_2; prior, exponential (with same median as data). (d) Population, folded normal; prior, folded logistic.

mean in all cases, with prior precisions $\pi_r = 5^r$ ($r = 1, 2, \ldots$). The empirical distribution function of the 25 values of $\hat{f}(0)$, plotted on normal probability paper, appeared rather concave. A chi-squared goodness-of-fit test of normality, based on 6 groups, 3 degrees of freedom, yielded a value of about 3·4, indicating no significant difference from normality. The empirical distribution function of $\log \hat{f}(0)$, plotted on normal probability paper, approximated a straight line somewhat better. The sample mean of the 25 values of $\hat{f}(0)$ was 0·0198. The sample estimate of the standard deviation of the distribution of $\hat{f}(0)$ was 0·0034. The estimate of bias furnished by (15) was $-0{\cdot}00718$, suggesting that the contribution of bias to $E\{\hat{f}(0)-\tilde{f}(0)\}^2$ was small. For comparison, formula (13) yields 0·0026 for the standard deviation of $\hat{f}(0)$, which is of the same order of magnitude as the above sample estimate, 0·0034. Formula (12) yields 0·0029 for the posterior standard deviation of $\tilde{f}(0)$, and (11) yields 0·0028 for the root of the prior expected squared error, $[E\{\hat{f}(0)-\tilde{f}(0)\}^2]^{\frac{1}{2}}$.

Since the same prior mean was used for each of the 25 samples of 300, one has here an opportunity to observe results using $\pi_r = 5^r$ when the prior mean is not fitted to the data. The true density, the prior mean, the histogram and the estimate $\hat{f}$ are shown in Fig. 1(d) for one of the 25 samples.

The author acknowledges many useful comments on earlier versions of this paper, in particular by Søren Johansen, F. Tom Lindstrom, Donald A. Pierce and Patrick C. Pointer. Research for this paper was supported in part by the National Science Foundation.

REFERENCES

BONEVA, L. J., KENDALL, D. G. & STEFANOV, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst. *J. R. Statist. Soc.* B **33**, 1–71.

BRUNK, H. D. & PIERCE, D. A. (1974). Estimation of discrete multivariate densities for computer-aided differential diagnosis of disease. *Biometrika* **61**, 493–9.

DICKEY, J. M. (1969). Smoothing by cheating. *Ann. Math. Statist.* **40**, 1477–82.

FELLNER, W. H. (1974). Heuristic estimation of probability densities. *Biometrika* **61**, 485–92.

GOOD, I. J. & GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255–77.

KRONMAL, R. A. & TARTER, M. E. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Am. Statist. Assoc.* **63**, 925–52.

NEYMAN, J. (1937). 'Smooth' test for goodness of fit. *Skand. Aktuar.* **20**, 144–99.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–7.

TARTER, M. E. & KRONMAL, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *Am. Statistician* **30**, 105–12.

WAHBA, G. (1976). Histosplines with knots which are order statistics. *J. R. Statist. Soc.* B **38**, 140–51.

WATSON, G. S. (1969). Density estimation by orthogonal series. *Ann. Math. Statist.* **40**, 1496–8.

WEGMAN, E. J. (1972a). Nonparametric probability density estimation: I. A summary of available methods. *Technometrics* **14**, 513–46.

WEGMAN, E. J. (1972b). Nonparametric probability density estimation. II. A comparison of density estimation methods. *J. Statist. Comp. and Simul.* **1**, 225–45.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. R. Statist. Soc.* B **20**, 334–43.

[*Received July* 1977. *Revised January* 1978]