The Annals of Statistics
1980, Vol. 8, No. 6, 1307–1325

ASYMPTOTIC INTEGRATED MEAN SQUARE ERROR USING LEAST SQUARES AND BIAS MINIMIZING SPLINES[1]

BY GIRDHAR G. AGARWAL AND W. J. STUDDEN

Indian Statistical Institute and Purdue University

Let S_k^d be the set of dth order splines on [0, 1] having k knots $\xi_1 < \xi_2 \ldots < \xi_k$. We consider the estimation of a sufficiently smooth response function g, using n uncorrelated observations, by an element s of S_k^d. For large n and k we have discussed the asymptotic behavior of the integrated mean square error (IMSE) for two types of estimators: (i) the least squares estimator and (ii) a bias minimizing estimator. The asymptotic expression for IMSE is minimized with respect to three variables. (i) the allocation of observation (ii) the displacement of knots $\xi_1 < \cdots < \xi_k$ and (iii) number of knots.

1. Introduction. Suppose a functional relationship

$$\eta = g(x)$$

exists between a response η and an independent variable x, where x lies in the interval [0, 1]. The problem to be considered is to estimate g using n measurements of η. At each x_i, $i = 1, \ldots, r$, $n_i = n\mu_i$ measurements are taken. The probability measure assigning mass μ_i to the point $x_i(\sum \mu_i = 1)$ is referred to as the design and will be denoted by $\mu^{(n)}$. In observing the response η we assume that an additive experimental error, denoted by ϵ, exists so that, for each observation y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, r$; we can write

$$y_{ij} = \eta(x_i) + \varepsilon_{ij} = g(x_i) + \varepsilon_{ij}.$$

We assume that ε_{ij} are uncorrelated and identically distributed with mean zero and an unknown common variance σ^2 independent of x.

If it is known that the true functional relationship $\eta = g(x)$ has a certain form depending on a few parameters, then the problem is usually to estimate these parameters. If the form of the true functional relationship is unknown, the problem is to approximate the function $g(x)$ by some graduating function. In this paper we are interested in the latter problem. In the absence of the knowlege of the true functional relationship, it has been a common practice to use a polynomial as an approximating function. But when the degree of polynomial is high, a number of unpleasant features begin to appear, one of which is the high oscillatory behavior of the approximating polynomial. Spline functions, to be defined presently, are considerably less oscillatory. As an example see Jupp (1978) where he has fitted a polynomial of degree 9 as well as a cubic spline to the data of world sugar prices over a 31 year period taken from Guest (1961, page 194). The improvement in the fit to the data achieved by cubic splines is somewhat obvious since it shows less oscillation compared to the polynomial fit. Furthermore, the behavior of a polynomial in an arbitrary small region defines, through the concept of analytic continuity, its behavior elsewhere. This seems to manifest itself in situations where the function g, behaving poorly in a small region, gives rise to a polynomial approximation behaving poorly everywhere. On the other hand, the spline functions possess the property of having local behavior that is less dependent on their behavior elsewhere. Because of these properties spline functions are more and more being used in the exploration of response curves

Received August 1978; revised August 1979.

[1] This research was supported in part by NSF Grant No. MCS76-08235 A01 and MCS-7901707.

AMS 1970 *subject classifications.* Primary 62K05, 62J05, 62F10; secondary 41A50, 41A60.

Key words and phrases: Optimal design, least square estimator, bias minimizing estimator, approximation, *B*-splines, Bernoulli polynomial, L_2-projection operator.

for physical processes. Low order splines are commonly used in geophysics in the form of layered earth models (for example, see Vozoff and Jupp (1975) and Jupp and Stewart (1974)). In astrophysics, Holt (1974) has used piecewise linear splines to model the radiation profile of the sun's atmosphere. Lawton, Sylvestre, and Maggio (1972) have also used linear splines as "empirical functions" in approximating shape invariant models. These kinds of models arise in the studies of hearing response, or EKG's in the human population, or when one measures spectrophotometric curves from sampled product, or observes reaction curves in designed chemical experiments. Wold (1971, 1974) has used spline functions in the analysis of response curves in pharmacokinetics.

We assume that the function $g(x)$, defined on $[0, 1]$, is such that $g \in C^d[0, 1]$, i.e., g has d continuous derivatives. Here the function $g(x)$ will be approximated by a function $s(x)$ in the class S_k^d. The set S_k^d is the collection of all polynomial splines of order d (degree $d - 1$) having k knots $\xi_1 < \xi_2 < \cdots < \xi_k$ in the interior of the interval $[0, 1]$. That is, $s(x)$ is a polynomial of degree at most $d - 1$ on each interval (ξ_i, ξ_{i+1}) and belongs to $C^{d-1}[0, 1]$. For $d = 1$, S_k^d consists of functions which are constant on each interval (and suitably defined at each ξ_i). For $d = 2$, S_k^d consists of functions s which are linear on each interval (ξ_i, ξ_{i+1}) and continuous on $(0, 1)$. For general d, the function $s(x) \in S_k^d$ has the representation

$$s(x) = \textstyle\sum_{i=1}^{k+d} \theta_i N_i(x) \tag{1.1}$$

where N_i's are normalized B-splines. The polynomial splines and their B-spline basis will be discussed further in Section 2.

Let $\bar{y}_i$ denote the average of the n_i observations taken at x_i. Estimates which are linear in $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_r)'$ will be used in nearly all cases. Thus the vector of parameters $\theta = (\theta_1, \ldots, \theta_{k+d})'$ will be estimated by

$$\hat{\theta} = C\bar{y} \tag{1.2}$$

where C is a $(k + d) \times r$ matrix. As our criterion for the goodness of the estimate we shall use an integrated mean square error (IMSE); the integration being taken with respect to a measure λ which has a continuous strictly positive density with respect to Lebesgue measure. Our estimate is then

$$N'(x)\hat{\theta} = N'(x)C\bar{y}, \tag{1.3}$$

where $N(x) = (N_1(x), N_2(x), \ldots, N_{k+d}(x))'$. The mean value of $N'(x)\hat{\theta}$ is $N'(x)Cg_r$ where $g_r = (g(x_1), \ldots, g(x_r))'$. The variance is

$$E(N'(x)\hat{\theta} - N'(x)Cg_r)^2 = (\sigma^2/n)N'(x)CD^{-1}(\mu^{(n)})C'N(x),$$

where $D(\mu^{(n)})$ is an $r \times r$ diagonal matrix with diagonal elements $\mu_1, \ldots, \mu_r$. The mean square error is then variance plus squared bias and the integrated mean square error is

$$\text{IMSE} = V + B = (\sigma^2/n)\text{tr } CD^{-1}(\mu^{(n)})C'M(\lambda) + \int_0^1 (g(x) - N'(x)Cg_r)^2 \, d\lambda(x) \tag{1.4}$$

where $M(\lambda)$ is the $(k + d) \times (k + d)$ matrix

$$M(\lambda) = \int_0^1 N(x)N'(x) \, d\lambda(x). \tag{1.5}$$

Note that V and B denote the integrated variance and integrated squared bias respectively.

The IMSE involves three variables (i) the design $\mu^{(n)}$ (ii) the knots $\xi_1, \xi_2, \ldots, \xi_k$ and (iii) the estimate or choice of C. It is difficult to minimize the IMSE given in (1.4) directly with respect to these variables. The approach used is to first consider the asymptotic behavior of the IMSE for large n and k under some regularity conditions and then perform the minimization.

The purpose of studying (1.4) is to attempt to utilize the choice of these three "variables".

With regard to the choice of C we study mainly the least squares estimator (LSE). Some consideration is given to a bias minimizing estimator (BME) which resembles the estimator minimizing the total IMSE for known g. Further comments on the BME are given in Section 4.

An explanation of how the design and the knots are chosen is given after Theorem 3.2. We have as yet not exploited the choice of the order d of the splines.

In Sections 3 and 4 we have discussed the asymptotic behavior of the IMSE for two kinds of estimators, namely, the least squares estimator (LSE) and a bias minimizing estimator (BME). An example is given in Section 5 to illustrate the behavior of the procedure (for choosing the design and knots) which is proposed in Section 3. In order to facilitate the presentation of the results, we have deferred the proofs of all theorems to Section 6.

The main idea for the approach used here is from Dodson (1972), Rice (1969), and Burchard (1974) where nonstatistical approaches were used. Further discussion of the results can be found in Agarwal (1978), and Agarwal and Studden (1978a).

2. Splines and *B*-splines.

Let

$$(\xi_0 =)0 < \xi_1 < \cdots < \xi_k < 1(= \xi_{k+1}) \tag{2.1}$$

be a subdivision of the interval $[0, 1]$ by k distinct points. These points are the "knots" of the spline function which is defined as follows: a spline function, $s \in S_k^d$, is a function which (i) in each open interval (ξ_{i-1}, ξ_i) for $i = 1, \ldots, k + 1$ is a polynomial of degree $\leq (d - 1)$, (ii) has $(d - 2)$ continuous derivatives in the open interval $(0, 1)$.

For each (fixed) set of knots of the form (2.1), the class S_k^d of such splines is a linear space of functions of dimension $(k + d)$. A basis for this linear space is provided by B-splines, or basic splines (Curry and Schoenberg (1966)). As well as being a powerful theoretical tool in spline theory, these elementary spline functions provide stable methods for computing with spline functions (see deBoor (1972) and (1978)). One of the desirable properties of B-splines is that their support consists of a small fixed, finite number of intervals between knots.

For $d = 1$, the $N_i(x)$ are simply the indicator functions on the intervals $(\xi_{i-1}, \xi_i]$. For $d = 2$ the support consists of two consecutive intervals (except for the first and last function) and on these intervals is given by

$$\begin{aligned} N_{i+1}(x) &= (x - \xi_{i-1})/(\xi_i - \xi_{i-1}) \qquad \xi_{i-1} < x < \xi_i \\ &= (\xi_{i+1} - x)/(\xi_{i+1} - \xi_i) \qquad \xi_i < x < \xi_{i+1}. \end{aligned}$$

For equally spaced knots the N_i are proportional to the density of the sum of d uniform random variables on (0, 1) appropriately scaled and translated.

Explicit expressions for the B-splines will not be needed. For completeness we give a precise definition and list some of their properties below.

We write Π for the nondecreasing sequence $\{t_i\}_1^{k+2d}$ obtained from $\{\xi_i\}_0^{k+1}$ by repeating ξ_0 and ξ_{k+1} each exactly d times. The B-spline basis for the family S_k^d is formed by the following $k + d$ normalized B-splines

$$N_i(x) = (t_{i+d} - t_i)[t_i, \ldots, t_{i+d}](t - x)_+^{d-1} \tag{2.2}$$

$i = 1, \ldots, k + d$, where $[t_i, \ldots, t_{i+d}]\phi$ denotes dth order divided differences on the $(d + 1)$ points $t_i, \ldots, t_{i+d}$ of the function ϕ, and a_+^n means a^n if $a > 0$ and zero otherwise. For two or more than two coincident t_i's, the differences in (2.2) are taken to be confluent divided differences (cf. Milne-Thomson (1951)). The N_i are, apart from a constant factor, the B-splines of Curry and Schoenberg (1966).

The N_i defined in (2.2) satisfy

$$0 < N_i(x) \leq 1 \quad \text{for} \quad x \in (t_i, t_{i+d}) \quad \text{and} \quad N_i(x) = 0 \quad \text{otherwise}; \tag{2.3}$$

$$\begin{aligned} &\{N_i\}_{i=j}^{j+l} \text{ is linearly independent over the interval } [t_{j+d-1}, t_{j+l+1}] \\ &\text{for any } l \geq d - 1 \text{ and any } 1 \leq j \leq k + d - l; \end{aligned} \tag{2.4}$$

$$\{N_i\}_{i=1}^{k+d} \quad \text{spans} \quad S_k^d; \tag{2.5}$$

$$\textstyle\sum_{i=1}^{k+d} N_i(x) = 1 \quad \text{for all} \quad x; \tag{2.6}$$

$$\int_0^1 N_i(x) = (t_{i+d} - t_i)/d, \qquad i = 1, \ldots, k + d. \tag{2.7}$$

For (2.3), (2.5), (2.6) and (2.7) see Schoenberg (1966). DeBoor and Fix (1973) proved (2.4).

For $d > 1$, $N_i(x)$, as given by (2.2), are well defined continuous functions. For $d = 1$, (2.2) makes sense only for $x \neq t_j$, $1 \leq j \leq k + 2d$, because of the jump discontinuity of $(t - x)_+^0$ at $t = x$. So in this case we assume the definition (2.2) to be augmented by the (admittedly arbitrary) demand that $N_i(x)$ be right continuous everywhere. Thus for $d = 1$, we let

$$\begin{aligned} N_i(x) &= 1, \qquad t_i \leq x < t_{i+1} \\ &= 0, \qquad \text{otherwise.} \end{aligned}$$

3. Asymptotic value of IMSE for LSE. In considering the asymptotic behavior of the IMSE, we shall be concerned with the sequences $T_k = \{\xi_0, \xi_1, \ldots, \xi_k, \xi_{k+1}\}$ of knots defined by

$$\int_0^{\xi_i} p(x)\,dx = i/(k+1), \qquad i = 0, 1, \ldots, k + 1. \tag{3.1}$$

where $p(x)$ is a positive continuous density on [0, 1]. Sacks and Ylvisaker (1970) call the sequence $\{T_k, k \geq 1\}$ so defined as a regular sequence generated by $p\{\mathrm{RS}(p)\}$.

It will be convenient to introduce the following notation: for each fixed k, and $i = 1, \ldots, k + 1$ let

$$\delta_i = \xi_i - \xi_{i-1},\ \delta = \max \delta_i, \qquad \text{and} \qquad \Delta = \delta/\min_i \delta_i.$$

Letting $0 < p_{\min} = \min_x p(x) \leq \max_x p(x) = p_{\max}$, we see that

$$\Delta \leq p_{\max}/p_{\min}. \tag{3.2}$$

Also in view of the definition of t_j's in terms of ξ_i's we see that

$$\{\max_i (t_{i+d} - t_i)/\min_i (t_{i+d} - t_i)\} < d\Delta \leq dp_{\max}/p_{\min}. \tag{3.3}$$

In this section we discuss the asymptotic behavior of the ISME when the estimator used is the least squares estimator (LSE).

In the classical problem of regresson theory, the analytic form of the function $g(x)$ is supposed to be known. In our case g would be assumed to be of the form $g(x) = \sum_{i=1}^{k+d} \theta_i N_i(x)$. The estimator $\hat{\theta} = C\bar{y}$ is restricted to be unbiased. The unbiasedness of $\hat{\theta} = C\bar{y}$ restricts C so that

$$CF' = I$$

where F is the $(k + d) \times r$ matrix $F = (N(x_1), \ldots, N(x_r))$ and I is the $(k + d) \times (k + d)$ identity matrix. The quantity V in (1.4) is then minimized by the usual least squares estimate

$$C = M^{-1}(\mu^{(n)})FD(\mu^{(n)}). \tag{3.4}$$

Here $\mu^{(n)}$ represents the design measure placing mass μ_i on x_i, $i = 1, \ldots, r$, $M(\mu^{(n)})$ is the $(k + d) \times (k + d)$ matrix $\int N(x)N'(x)\,d\mu^{(n)}(x)$. The estimator $\hat{\theta} = C\bar{y}$ can then be represented by

$$\hat{\theta}_{\mathrm{LSE}} = C\bar{y} = M^{-1}(\mu^{(n)}) \int N(x)\bar{y}_x\,d\mu^{(n)}(x) \tag{3.5}$$

where $\bar{y}_x$ is the average of the observations taken at the point x. The LSE estimator gives a value of V and B as follows:

$$V = V(n, k) = (\sigma^2/n)\mathrm{tr}\, M^{-1}(\mu^{(n)})M(\lambda) \tag{3.6}$$

and

$$B = B(n, k) = \int_0^1 (g(x) - g_{n,k}(x))^2\,d\lambda(x) \tag{3.7}$$

where $g_{n,k}(x) = N'(x)M^{-1}(\mu^{(n)}) \int N(y)g(y)\, d\mu^{(n)}(y)$.

In order to make concrete asymptotic statements about V and B we assume (i) $\mu^{(n)}$ converges to a design measure μ, where μ has density h, continuous and positive and (ii) the rate of convergence is determined by k or rather by $\delta = \max(\xi_i - \xi_{i-1})$. More specifically we let $n = n_k$ be such that

$$\sup_x |H_{n_k}(x) - H(x)| = o(k^{-1}), \qquad k \to \infty, \tag{3.8}$$

where H_{n_k} and H are the cumulative distribution function of $\mu^{(n_k)}$ and μ respectively.

THEOREM 3.1. *Let $g \in C^d\,[0, 1]$, μ and λ have continuous strictly positive densities h and f respectively. If $\{T_k\}$ is* RS(p) *and condition* (3.8) *is satisfied by the designs $\mu^{(n_k)}$, then as $k \to \infty$,*

(A) $V \approx (k\sigma^2/n_k) \int (f(x)/h(x))p(x)\, dx$

(B) $B \approx (b/k^{2d}) \int \{(g^{(d)})(x))^2/(p(x))^{2d}\}f(x)\, dx$

where the symbol $\approx$ indicates that the ratio tends to one. The constant b equals $|B_{2d}|/(2d)!$, where B_{2d} is the 2*d*th *Bernoulli number (see Nörlund* (1924) *or Ghizzetti and Ossicini* (1970)).

The above theorem says that

$$\text{IMSE} \approx (k\sigma^2/n_k) \int (f(x)/h(x))p(x)\, dx + (b/k^{2d}) \int \{(g^{(d)}(x))^2/(p(x))^{2d}\}f(x)\, dx. \tag{3.9}$$

This asymptotic value depends upon (i) $h(x)$, the allocation of observations, (ii) k, the number of knots and (iii) $p(x)$, the displacement of knots. The results of minimizing the asymptotic value of the IMSE in (3.9) are given in the following theorem.

THEOREM 3.2. *The* IMSE *given in* (3.9) *is absolutely minimized by h, p and k given as follows*:

$$h(x) = \alpha_{g,f}\{(f(x))^{2d+1}(g^{(d)}(x))^2\}^{1/(4d+1)}, \tag{3.10}$$

$$p(x) = \beta_{g,f}\{f(x)(g^{(d)}(x))^4\}^{1/(4d+1)}, \tag{3.11}$$

$$k = \beta_{g,f}^{-1}\{(2bdn/\sigma^2)\alpha_{g,f}\}^{1/(2d+1)}, \tag{3.12}$$

where $\sigma_{g,f}^{-1} = \int_0^1 \{(f(x))^{2d+1}(g^{(d)}(x))^2\}^{1/(4d+1)}\, dx$, *and* $\beta_{g,f}^{-1} = \int_0^1 \{f(x)(g^{(d)}(x))^4\}^{1/(4d+1)}\, dx$.

For a proof of this theorem we refer to the proof of a theorem in Section 3 of Agarwal and Studden (1978b) in which parallel results are proved for the case $f(x) \equiv 1$.

The knot displacement in (3.11) indicates that the knots should be placed where $f(x)(g^{(d)}(x))^4$ is large. Using (3.10) and (3.11) (or going back to (3.9)) we see that $h\ \alpha\ (fp)^{1/2}$ so that h is usually more dispersed than p.

Equation (3.12) indicates that k is decreasing in σ and of order $\mathrm{n}^{1/(2d+1)}$. For example for $d = 2$ this gives $n\alpha k^5$. This indicates that there should be many more observations than knots.

The minimized value for the asymptotic expression for $V + B$ is used as follows. First, for a given n we iteratively choose the number and placement of the knots. Second, the function g is estimated sequentially and at each stage a better design or allocation of future observations is chosen. This is illustrated for the case $d = 2$.

For a given set of observations on g we estimate σ^2 and $g^{(2)}$. The number of knots $\hat{k}$ is calculated from (3.12) and their displacement is determined by $\hat{p}$ in (3.11). Then σ^2 and $g^{(2)}$ are reestimated and the knots are readjusted. For a fixed n, two or possibly three iterations seem to be sufficient. In choosing the observations sequentially, we attempt to choose future observations so that the allocation of the total set resembles the design $\hat{h}$ given by (3.10) with the most recent estimate of $g^{(2)}$. The above cycle is then repeated. A brief illustration of the above procedure is given by an example in Section 5.

REMARK 3.1. We have proposed choosing k and $\xi_1, \ldots, \xi_k$ and the design by estimating g. Other methods for choosing k and the knots $\xi_1, \ldots, \xi_k$ are certainly possible. One such method is the technique of cross-validation proposed and studied extensively by Wahba. See, for example, Wahba (1977) or Golub, Heath and Wahba (1979) or references therein.

REMARK 3.2. In minimizing the asymptotic expression for IMSE $= V + B$ we showed $k = O(n^{1/(2d+1)})$. When this is inserted into the expression for the IMSE in (3.9) we find that IMSE $= O(n^{-2d/(2d+1)})$. (If n and k are chosen so that $n_k = O(k^{2d+1})$, we should also keep in mind that condition (3.8) should be satisfied by the design $\mu^{(n_k)}$.) This is the same rate obtainable for smoothing splines (see Wahba (1978)). This rate appears to be the best possible as indicated by recent results of C. J. Stone concerning rates of convergence for nonparametric estimators. This gives considerable support for using splines in practical nonparametric work because they achieve the best possible convergence results and have in addition some nice properties as indicated in Section 1.

REMARK 3.3. We have imposed rather strong conditions on the functions g and f. These, of course, can be weakened somewhat. In particular g is assumed to be in C^d. That is, $g^{(d)}$ is assumed to be continuous. The rates involved seem to be attainable if $g^{(d-1)}$ is only assumed to be absolutely continuous and $g^{(d)}$ is in L_2. However, the constants involved in the bias term B would appear to change. Numerous results, concerning the rates, are available in the literature on spline theory without the constants. Exact constants are given, for example, in Barrow and Smith (1978a) and they also assume $g \in C^d$. Although the exact constants are desirable in minimizing the IMSE, they do not, however, appear to be overly important. For example, if the bias is multiplied by a factor of c then k changes by a factor of $c^{1/(2d+1)}$ which, say, is 1.58 if $c = 10$ and $d = 2$.

4. Asymptotic value of IMSE for BME. The IMSE, given in (1.4), is minimized if

$$E(C\bar{y}) = tM^{-1}(f)\left(\int_0^1 N(x)g(x)f(x)\,dx\right) \tag{4.1}$$

where $M(f) = M(\lambda) = \int N(x)N'(x)f(x)\,dx$, $t = q/((\sigma^2/n) + q)$, and $q - g_r'D(\mu^{(n)})g_r = \int g^2(x)\,d\mu^{(n)}(x)$. This minimization is easy to show and is given in Agarwal and Studden (1978a). Some discussion of the matrix C which actually minimizes the IMSE is given in Remark 4.2 below. The factor t does not seem to be important unless n is small relative to g^2 and q. We have not attempted to use t in our estimation of g. The expression (4.1) with $t = 1$ actually minimizes the bias B. Various authors, for example, Box and Draper (1959) and Karson, Manson and Hader (1969) have proposed attaching more importance to the bias part B.

In general a matrix C cannot be found for which (4.1) holds even if $t = 1$, so instead we try to find a C^* such that

$$E(C^*\bar{y}) \simeq M^{-1}(f)\int_0^1 N(x)g(x)f(x)\,dx. \tag{4.2}$$

The asymptote is in the sense that $\| E(C^*\bar{y}) - M^{-1}(f) \int N(x)g(x)f(x)\,dx\|$ goes to zero as n (number of observations) tends to infinity, where the vector norm $\| a \| =_{\text{def}} (a'a)^{1/2}$. We should emphasize here that k is fixed.

Let $L'(x) = (L_1(x), \ldots, L_r(x))$, where $L_i(x)$, $i = 1, \ldots, r$ are the normalized B-spline (Section 2) of order 2 with knots at the observation points x_i, $i = 2, \ldots, r - 1$. The $L_i(x)$ is a "roof-like" function which has a value one at x_i, goes linearly to zero at adjacent knots x_{i-1} and x_{i+1} and then remains zero. Let us define $\tilde{g}(x) = \sum_{i=1}^r g(x_i)L_i(x)$. Since $L_i(x_j) = \delta_{ij}$; $i, j = 1, \ldots r$, $\tilde{g}(x)$ interpolates g at x_i, $i = 1, \ldots, r$. As an approximation to g, the function $\tilde{g}$ satisfies the following two properties (e.g., see Prenter 1975).

(i) If g is continuous then $\tilde{g}$ converges to g as r (or n) $\to \infty$ in such a way that $\eta = \max_i (x_i - x_{i-1})$ tends to zero.

(ii) If g is twice continuously differentiable, then $\| g - \tilde{g} \|_\infty = max_{x\in[0,1]} \, | g(x) - \tilde{g}(x) | \le \alpha \| g^{(2)} \|_\infty \eta^2$, where α is a constant independent of n.

Now if we take

$$C^* = M^{-1}(f) \int_0^1 N(x)L'(x)f(x)\,dx \tag{4.3}$$

then in view of (i) and (ii) we see that C^* of (4.3) satisfies (4.2). Hence our "bias minimizing" estimate (BME) is defined as

$$\hat{\theta}_{\text{BME}} = C^*\bar{y} = M^{-1}(f)\left(\int_0^1 N(x)L'(x)f(x)\,dx\right)\bar{y}. \tag{4.4}$$

In the following theorem we shall find certain asymptotic expressions for the IMSE using the estimator $\hat{\theta}_{\text{BME}}$. For simplicity the design $\mu^{(n)}$ is assumed to have weight μ_i on x_i given by

$$\mu_i = \int_0^1 L_i(x)h(x)\,dx, \qquad i = 1, \ldots, r \tag{4.5}$$

for some continuous strictly positive density $h(x)$.

THEOREM 4.1. *If the estimator $\hat{\theta}_{\text{BME}}$, given in* (4.4), *is used, and the design is chosen using* (4.5) *and $\{T_k\}$ is* RS(p), *then*

(A) $\lim_{k\to\infty} \lim_{n\to\infty} (nV/k\sigma^2) = \int (f(x)/h(x))p(x)\,dx$

(B) $\lim_{k\to\infty} \lim_{n\to\infty} k^{2d}B = (|B_{2d}|/(2d)!) \int \{(g^{(d)}(x))^2/(p(x))^{2d}\}f(x)\,dx.$

REMARK 4.1. In Theorem 4.1 the limits are taken in such a way that n approaches infinity before k. Presumably the limits may be taken together as was done in Theorem 3.1 provided a condition like equation (3.8) holds. The condition which is imposed in equation (4.5) on the design measure $\mu^{(n)}$ does not seem to be necessary. It is imposed in order to obtain the same asymptotic value for the variance term as that obtained by the LSE. In the proof of Lemma 6.11 below it seems that something of the form

$$\mu_i = h(x_i)I_i(1 + o(1)) \tag{4.6}$$

where $I_i = \int L_i(x)\,dx$ is necessary. For a general sequence $\mu^{(n)}$ one could use a slightly different set of points $z_1, z_2, \ldots, z_r$, instead of $x_1, x_2, \ldots, x_r$, on which to base our L_i functions so that (4.6) holds. In this case however one has futher difficulties in showing the bias term involving (4.4) (or slight modifications) behaves properly.

REMARK 4.2. The matrix C which actually minimizes the IMSE for known g is given by

$$C_g = M^{-1}(f)sg_r'A^{-1} \tag{4.7}$$

where $s = \int Ngf\,dx$, $A = [g_rg_r' + (\sigma^2/n)D_n^{-1}]$ and $D_n = D(\mu^{(n)})$. Instead of starting with (4.1) in search of a good estimator one might try estimating g by some preliminary means and inserting this value into the matrix C_g from (4.7). It turns out that the resulting estimator is effectively (4.1) again with an estimate of quantity t.

Using (4.7) the minimum IMSE reduces to

$$\int g^2f\,dx - ts'M^{-1}(f)s. \tag{4.8}$$

With $t = 1$ this is the minimum of the bias term. This gives some support to placing more stress on the bias term if t is near 1. In evaluating the performance of any estimator the quantity (4.8) (the unachievable minimum IMSE) might be used as a comparison.

5. An example. In the example below the integrating measure f was taken to be $f \equiv 1$ and the function g was estimated by linear splines, i.e., $d = 2$. The function g was taken to be

$$g(x) = [0.01 + (2x - 0.3)^2]^{-1} + [0.0144 + (2x - 1.2)^2]^{-1}.$$

The choice of the above function is from Ichida, Kiyono and Yoshimoto (1977). For this function $g^{(2)}(x)$ varies considerably in the interval [0, 1]. The data errors were simulated by adding to $g(x_i)$ a number sampled from the normal distribution with mean zero and variance one. To begin with we took three equally spaced knots and took five observations at each knot and the end points. We performed three cycles with the number of observations $n = 75$, 150 and 250 respectively. This was done for both estimates namely $\hat{\theta}_{\text{LSE}}$ and $\hat{\theta}_{\text{BME}}$. Recall that a number of internal iterations within each cycle (for fixed n) were done to select the number and position of the knots. The linear spline obtained at the end of cycle zero, cycle one and cycle two for $\hat{\theta}_{\text{LSE}}$ and $\hat{\theta}_{\text{BME}}$ are shown in Figure 1 and Figure 2 respectively. The breaks (joints) in the graph are the knots. We can see that the linear spline fits obtained by the two estimates improved at each cycle. Once the fit started getting better and hence the estimates of g and $g^{(2)}$ were improved, the knots were chosen at the points where $g^{(2)}$ was large. Also the design (not shown in the graph) was more dispersed than the knots. Both of these things were expected (see the comments after Theorem 3.2 in Section 3).

At the start (cycle zero), the fit was bad. Actually, the bias was very large compared to the variance. Both the estimators tried to reduce the bias very fast. The BME shows some superiority over the LSE in the sense that it reduces the integrated squared bias (B) faster than the LSE does.

If in this example, we had chosen a larger σ^2 so that the integrated variance (V) is larger than B then the present set up does not show a strong case for the LSE minimizing V faster than the BME does. This is apparently due to the fact that after two or three cycles and a few observations have been taken, these observations are considerably dispersed (at least in this example), so that the LSE and the BME operate on V in a similar manner.

The details of the termination criterion, estimation of $g^{(2)}$, etc., are omitted here. A more complete report on this procedure and its use is available in Agarwal and Studden (1978b). The Fortran program based on this algorithm has been used on a few other examples. Further work especially using $d > 2$ and more detailed comparisons between estimators and with other procedures seems appropriate.

6. Proofs of theorems.

PROOF OF THEOREM 3.1A. We first prove the following important result which will be used in the proof of this theorem as well as in the proof of Theorem 4.1.

THEOREM 6.1. *Let $M(\phi)$ be the $(k + d) \times (k + d)$ matrix*

$$M(\phi) = \int_0^1 N(x)N'(x)\phi(x)\,dx. \tag{6.1}$$

If ϕ and ψ and p are continuous strictly positive functions defined on [0, 1] *and $\{T_k\}$ is* RS(p) (*see* (3.1)) *then as $k \to \infty$,*

$$\operatorname{tr} M^{-1}(\phi)M(\psi) \approx k \int_0^1 (\psi(x)/\phi(x))p(x)\,dx. \tag{6.2}$$

PROOF. Let us write

$$M(\phi) = M_0 D(\phi) - E(\phi) \tag{6.3}$$

where M_0 is given by (6.1) with $\phi \equiv 1$, $D(\phi)$ is the diagonal matrix with elements $\phi(\zeta_i)$, $i = 1$, $2, \ldots, k + d$ and the error term $E(\phi)$ is defined through (6.3). The points $\zeta_1 < \zeta_2 < \cdots < \zeta_{k+d}$ and $(k + d)$ arbitrary points in [0, 1] such that

$$\zeta_i \in \overline{\text{support}\ \ N_i}, \qquad i = 1, \ldots, k + d. \tag{6.4}$$

If we define

$$\zeta_i = \left(\textstyle\sum_{l=i+1}^{i+d-1} t_l\right)/(d - 1), \qquad i = 1, \ldots k + d,$$

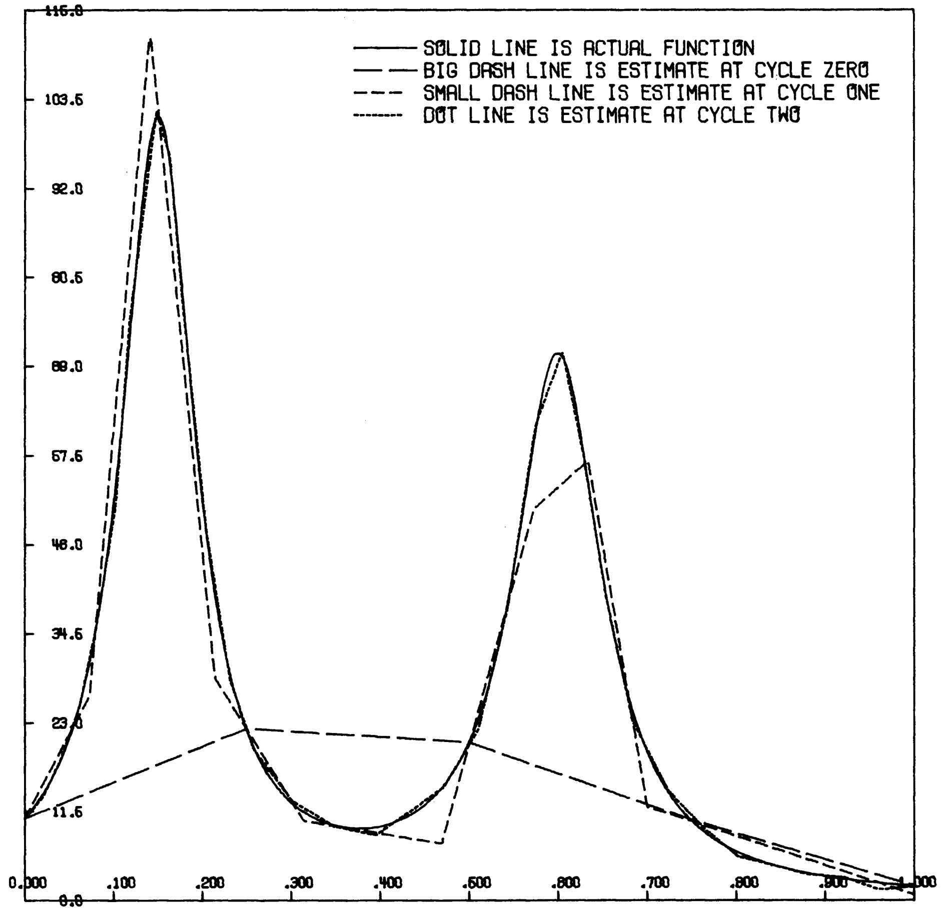

FIG. 1. *Function $g(x)$ and the* LSE *at end of cycles zero, one and two.*

we can see that these ξ_i's satisfy (6.4). Schoenberg (1966) calls these points "nodes" and has used them in another context. Using (6.1) and (6.3), we can write

$$M^{-1}(\phi)M(\psi) = [I - U]^{-1}[D^{-1}(\phi)D(\psi) - V] \tag{6.5}$$

where $U = D^{-1}(\phi)M_0^{-1}E(\phi)$, $V = D^{-1}(\phi)M_0^{-1}E(\psi)$ and I is $(k + d) \times (k + d)$ identity matrix. We want to expand $(I - U)^{-1}$ as a power series. This can be done if the matrix norm of U is less than one. In the following lemma we find $\| U \| =_{\text{def}} \max_x (\| Ux \| / \| x \|)$, where vector norm $\| X \| = (x'x)^{1/2}$.

LEMMA 6.2. *$\| U \| < \alpha\omega(\phi, \delta)$, where α is a constant independent of k and $\omega(\phi, \delta)$ is the modulus of continuity of ϕ and $\delta = \max_{1 \le i \le k+1} (\xi_i - \xi_{i-1})$.*

PROOF OF LEMMA. The proof consists of bounding the norms of M_0^{-1} and $E(\phi)$. Since M_0 is a positive definite matrix, it is easy to see that

$$\| M_0^{-1} \| = (1/l_{\min}), \tag{6.6}$$

where $l_{\min}$ is the smallest latent (or characteristic) root of M_0 given by

$$l_{\min} = \min_x \{(x'M_0x)/(x'x)\}. \tag{6.7}$$

Now to find an upper bound on $\| M_0^{-1} \|$ we use an inequality of deBoor (1973, page 273). The

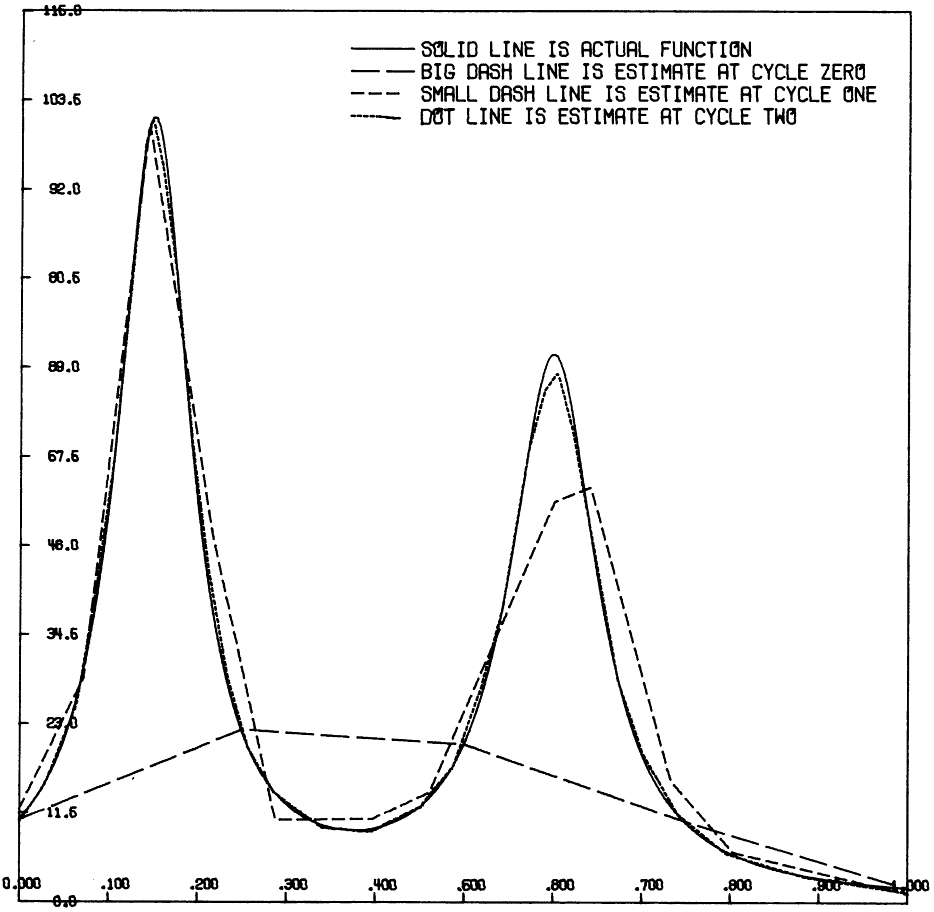

FIG. 2. *Function $g(x)$ and the* BME *at end of cycles zero, one and two.*

inequality states that

$$(6.8) \qquad \rho^2(\gamma'\gamma) \le (\gamma' A\gamma) \le (\gamma'\gamma) \qquad \text{for all} \qquad \gamma \in IR^{k+d}$$

where ρ is a constant independent of k and depends only on d, and matrix A, called a Gram matrix by deBoor, is related to matrix M_0 by

$$(6.9) \qquad M_0 = DAD$$

where D is the diagonal matrix with diagonal elements $\{(t_{i+d} - t_i)/d\}^{1/2}$, $i = 1, \ldots, k + d$. Using (6.6), (6.7), (6.8) and (6.9), we can show that

$$(6.10) \qquad \| M_0^{-1} \| \le (d/\rho^2)\{\min_{1\le i\le k+d} (t_{i+d} - t_i)\}^{-1}.$$

Now we shall find an upper bound on $\| E(\phi) \|$. First of all since $E(\phi) = [e_{ij}(\phi)]$ is a $(k + d) \times (k + d)$ band matrix of bandwidth $d - 1$ (i.e., $e_{ij}(\phi) = 0$ if $|i - j| > d - 1$; $i, j = 1, \ldots, k + d$) it is easy to check that

$$(6.11) \qquad \| E(\phi) \| \le (2d - 1)^{1/2}\{\max_{1\le i\le k+d} \textstyle\sum_{j=1}^{k+d} e_{ij}^2(\phi)\}^{1/2}.$$

In view of (6.4), we have for $i = 1, \ldots, k + d$,

$$(6.12) \qquad \textstyle\sum_{j=1}^{k+d} | e_{ij}(\phi) | < d\, \omega(\phi, \delta) \sum_{j=1}^{k+d} \displaystyle\int N_i(x)N_j(x)\, dx = \omega(\phi, \delta)(t_{i+d} - t_i).$$

The equality in (6.12) follows from (2.6) and (2.7). The equations (6.11) and (6.12) give

$$\| E(\phi) \| \le (2d-1)^{1/2}\omega(\phi, \delta) \max_{1\le i\le k+d} (t_{i+d} - t_i). \tag{6.13}$$

Finally since $\| D^{-1}(\phi) \| = \{\min_i \phi(\zeta_i)\}^{-1}$, we may combine (6.10) and (6.13), to obtain

$$\begin{aligned} \| U \| &= \| D^{-1}(\phi) M_0^{-1} E(\phi) \| \\ &\le \| D^{-1}(\phi) \| \, \| M_0^{-1} \| \, \| E(\phi) \| \\ &\le \frac{(2d-1)^{1/2}\, d}{\rho^2} \, \frac{\max_i (t_{i+d} - t_i)}{\min_i (t_{i+d} - t_i)} \, \frac{\omega(\phi, \delta)}{\min_i (\phi(\zeta_i))}. \end{aligned}$$

In view of the quasiuniformity condition and the fact that ϕ is bounded below, it follows that

$$\| U \| < \alpha\omega(\phi, \delta),$$

where the constant α does not depend on k. This proves the lemma.

Now since $\omega(\phi, \delta) \to 0$ as $k \to \infty$ (or $\delta \to 0$) we can make $\omega(\phi, \delta) < 1/\alpha$ and hence $\| U \| < 1$. We can then invert $(I - U)$ using a power series expansion,

$$\begin{aligned} (I - U)^{-1} &= I + U + U^2 + \cdots \\ &= I + W, \quad \text{say,} \end{aligned}$$

where $W = \sum_{j=1}^{\infty} U^j$. Therefore, from (6.5) and the above expansion,

$$\begin{aligned} \operatorname{tr} M^{-1}(\phi) M(\psi) &= \operatorname{tr} D^{-1}(\phi) D(\psi) - \operatorname{tr} V \\ &\quad + \operatorname{tr} W D^{-1}(\phi) D(\psi) - \operatorname{tr} VW. \end{aligned} \tag{6.14}$$

Now using the definition of the nodes ζ_i's and the mean value theorem in the expression (3.1), we see that the first term on the right of (6.14) divided by k (or $k + d$) will converge to the integral term in (6.2). Therefore Theorem 6.1 will be proved if we show that, as $k \to \infty$,

(i) tr $V = o(k)$

(ii) tr $WD^{-1}(\phi)D(\psi) = o(k)$

(iii) tr $VW = o(k)$.

Since $V = D^{-1}(\phi) M_0^{-1} E(\psi)$, from Lemma 6.2 we get

$$\| V \| < \beta\omega(\psi, \delta)$$

where β is a constant independent of k and $\omega(\psi, \delta)$ is the modulus of continuity of ψ. Also $| \operatorname{tr} V | < | (k + d) \| V \|$, where $(k + d)$ is the order of matrix V, hence (i) holds.

Using the matrix norm properties, namely

$$\| S + T \| \le \| S \| + \| T \|$$

and

$$\| ST \| \le \| S \| \, \| T \|$$

we can show that

$$\begin{aligned} \| W \| &\le \| U \| / (1 - \| U \|) \\ &\le \alpha\omega(\phi, \delta)/(1 - \alpha\omega(\phi, \delta)) \qquad \text{by Lemma 6.2.} \end{aligned} \tag{6.15}$$

Now since

$$\| \operatorname{tr} W D^{-1}(\phi) D(\psi) \| < (k + d) \, \| W \| \, \| D^{-1}(\phi) \| \, \| D(\psi) \|,$$

the relation (ii) holds in view of (6.15) and the fact that ϕ is bounded below and ψ is bounded above.

The proof for relation (iii) follows from the proof of (i) and (ii). □

Now we come back to the proof of Theorem 3.1A. We can write

$$\begin{aligned}(n_k V/\sigma^2) &= \operatorname{tr} M^{-1}(\mu^{(n_k)})M(\lambda)\\ &= \operatorname{tr} M^{-1}(\mu)M(\lambda) + \operatorname{tr}\,[M^{-1}(\mu^{(n_k)}) - M^{-1}(\mu)]M(\lambda).\end{aligned}$$

Since the design measure μ and the integrating measure λ have continuous strictly positive densities h and f respectively, in view of the above Theorem 6.1, the first factor on the right of the above expression has the asymptotic value $k\int(f(x)/h(x)p(x)\,dx$. To complete the proof of the theorem, we need to show that

$$\operatorname{tr}[M^{-1}(\mu^{(n_k)}) - M^{-1}(\mu)]M(\lambda) = o(k) \text{ as } k \to \infty. \tag{6.16}$$

Introducing $E(\mu^{(n_k)}) = M(\mu) - M(\mu^{(n_k)})$ we have

$$M^{-1}(\mu^{(n_k)}) - M^{-1}(\mu) = M^{-1}(\mu^{(n_k)})E(\mu^{(n_k)})M^{-1}(\mu),$$

and therefore

$$|\operatorname{tr}\{M^{-1}(\mu^{(n_k)}) - M^{-1}(\mu)\}M(\lambda)| \le (k+d)\|M^{-1}(\mu^{(n_k)})\|\,\|E(\mu^{(n_k)})\|\,\|M^{-1}(\mu)\|\,\|M(\lambda)\|.$$

The relation (6.16) will hold when we show that the following are true:

Lemma 6.3. $\|M(\lambda)\| = O(\delta)$.

Lemma 6.4. $\|E(\mu^{(n_k)})\| = O(\delta)$.

Lemma 6.5. $\|M^{-1}(\mu)\| = O(\delta^{-1})$.

Lemma 6.6. $\|M^{-1}(\mu^{(n_k)})\| = O(\delta^{-1})$.

Proof of Lemma 6.3. The matrix $M(\lambda) = M(f) = \int N(x)N'(x)f(x)\,dx$ is a band matrix. Therefore, as before (see (6.11)), we see that

$$\|M(\lambda)\| \le (2d-1)^{1/2}\{\max_i \textstyle\sum_{j=1}^{k+d} m_{ij}^2(\lambda)\}^{1/2},$$

where $m_{ij}(\lambda)$ are the elements of $M(\lambda)$. Using equations (2.6) and (2.7), we see that

$$\|M(\lambda)\| \le (2d-1)^{1/2} \max_x f(x)\delta.$$

This proves the lemma.

Proof of Lemma 6.4. We use condition (3.8) in finding the norm of $E(\mu^{(n_k)}) = [e_{ij}]$. This being a band matrix, we have

$$\|E(\mu^{(n_k)})\| \le (2d-1)^{1/2}\{\max_i \textstyle\sum_{j=1}^{k+d} e_{ij}^2\}^{1/2}. \tag{6.17}$$

Using integration by parts and the fact that $N_i(t_i) = N_i(t_{i+d}) = 0$ (see (2.3)), we can easily check that

$$e_{ij} = \int_{t_i}^{t_{i+d}} (H(x) - H_{n_k}(x))(N_i(x)N_j'(x) + N_j(x)N_i'(x))\,dx. \tag{6.18}$$

Using a recurrence formula (deBoor 1978, page 138) relating the derivatives of B-splines with the lower order B-splines and the fact that the knot sets determined by p are quasiuniform, we can find a constant c independent of k and i so that

$$\|N_i^{(j)}\|_\infty \le ck^j, \qquad j = 0, 1, \cdots, d-2. \tag{6.19}$$

(Note: $c = 1$ for $j = 0$, see (2.3)).

Condition (3.8) says that there exists a sequence $\{\epsilon_k\}_{k=1}^{\infty}$ tending to zero such that

$$|H_{n_k}(x) - H(x)| \le \epsilon_k/k \qquad \text{for all } x. \tag{6.20}$$

Equations (6.18), (6.19) and (6.20) can be combined to get $|e_{ij}| \leq 2\epsilon_k c\delta$ for all i and j, and therefore

$$\sum_{j=1}^{k+d} |e_{ij}| \leq 2(2d-1)\epsilon_k c\delta. \tag{6.21}$$

In the above, the sum has at most $(2d-1)$ nonzero elements since $E(\mu^{(n_k)})$ is a band matrix. The proof of the lemma now follows from (6.17) and (6.21).

Proof of Lemma 6.5. Recall that $M(\mu) = M(h) = \int N(x)N'(x)h(x)\,dx$. Let us use the representation (6.3) with ϕ replaced by h, i.e.,

$$M(h) = M_0 D(h) - E(h).$$

Now

$$M^{-1}(h) = (I-U)^{-1}D^{-1}(h)M_0^{-1}$$

where $U = D^{-1}(h)M_0^{-1}E(h)$. The rest of the proof follows in a manner similar to the proofs of Theorem 6.1 and Lemma 6.1. Actually, we can show that

$$\|M^{-1}(h)\| \leq \text{const.}\{(1-\alpha\omega(h,\delta))(\min_i h(\zeta_i))(\min_i(t_{i+d}-t_i))\}^{-1}.$$

The quasiuniformity condition (3.3) and the fact that h is bounded below implies that

$$\|M^{-1}(h)\| = O(\delta^{-1}).$$

Proof of Lemma 6.6. Writing $M(\mu^{(n_k)})$ as $M(\mu) + (M(\mu^{(n_k)}) - M(\mu))$, we have

$$M^{-1}(\mu^{(n_k)}) = [I - U_{n_k}]^{-1}M^{-1}(\mu), \tag{6.22}$$

where $U_{n_k} = M^{-1}(\mu)E(\mu^{(n_k)})$. Lemma 6.4 and Lemma 6.5 give $\|U_{n_k}\| = o(1)$, that is $\|U_{n_k}\| < \beta_k$, where $\beta_k \to 0$, as $k \to \infty$. For sufficiently large k, we can make $\|U_{n_k}\| < 1$ and then invert $(I - U_{n_k})$ using a power series expansion,

$$(I - U_{n_k})^{-1} = \sum_{j=0}^{\infty} U_{n_k}^j.$$

Using properties of the matrix norm (see proof of Theorem 6.1), we find that $\|(I - U_{n_k})^{-1}\| < (1-\beta_k)^{-1}$. The lemma now follows in view of equation (6.22) and Lemma 6.5 which gives a bound on the norm of $M^{-1}(\mu)$.

This completes the proof of Theorem 3.1A.

Before proving Theorem 3.1B, let us introduce some notation and describe two important results of Barrow and Smith (1978a and 1978b). These will be used in the proof of this theorem. Let $L_2^\nu = \{\psi | \int_0^1 \psi^2(x)\,d\nu(x) < \infty\}$ denote the L_2 space corresponding to measure ν with norm $\|\cdot\|_\nu$ and let P_k^ν denote the orthogonal projection operator from L_2^ν to S_k^d. The omission of the index ν will correspond to Lebesgue measure. Also, whenever $\nu = \mu^{(n)}$ we will simply use P_k^n.

Lemma 6.7. (*Barrow and Smith* 1978a). *If $g \in C^d[0,1]$, p is continuous and strictly positive and $\{T_k\}$ is* RS(p), *then*

$$\lim_{k\to\infty} k^{2d}\|g - P_k g|^2 = (|B_{2d}|/(2d)!) \int \{(g^{(d)}(x))^2/(p(x))^{2d}\}\,dx. \tag{6.23}$$

Note that the right side of the above expression differs from the asymptotic expression for $k^{2d}B$ by a factor of f in the integrand and that using the LSE with design $\mu^{(n)}$, the bias B is given by $\|g - P_k^n g\|_\lambda^2$ (see also (3.7)). It turns out that under the regularity condition (3.8), the projection $P_k^n g$ is asymptotically independent of $\mu^{(n)} = \mu^{(n_k)}$ as the number of knots $k \to \infty$. The error function $g - P_k^n g$ on each interval (ξ_i, ξ_{i+1}) begins to look proportional to a scaled version of the dth Bernoulli polynomial $B_d(x)$. A detailed discussion of polynomials $B_d(x)$ can be found in Schoenberg (1969), Ghizzetti and Ossicini (1970) or Nörlund (1924). We shall mention some of their properties momentarily.

To exploit the idea that locally the error $g - P_k g$ looks approximately like a Bernoulli polynomial, Barrow and Smith define a sequence of operators Q_k, such that $Q_k g \in S_k^d$ and is "close" to $P_k g$ in the sense that

$$\lim_{k\to\infty} k^d\{\|g - Q_k g\| - \|g - P_k g\|\} = 0. \tag{6.24}$$

Let

$$g(x) = \sum_{j=0}^d g^{(j)}(\xi_i)(x - \xi_i)^j/j! + o(\delta^d) = \bar{g}(x) + o(\delta^d)$$

and denote $g^{(j)}(\xi_i)/j!$ by g_i^j. The spline $Q_k g = \sum_{l=1}^{k+d} a_l N_l$ is essentially characterized by the requirement that on every dth interval (ξ_i, ξ_{i+1})

$$(\bar{g} - Q_k g)(x) = g_i^d B_d((x - \xi_i)/\delta_{i+1})\delta_{i+1}^d. \tag{6.25}$$

Due to the fact that $Q_k g$ must have $d - 2$ continuous derivatives at each ξ_i, the above equation cannot be made to hold on every interval (ξ_i, ξ_{i+1}) but only on every dth interval. For example if $d = 2$ we approximate g by a continuous broken line segment. The error $g - P_k g$ is approximately $g^{(2)}(\xi_i)$ times a scaled version of $B_2(x) = x^2 - x + 1/6$. One considers approximately the best line segment on every second interval and then joins the ends of these line segments on the intervals between. The polynomials $B_d(x)$ on $(0, 1)$ have leading coefficients one, satisfy

$$B_d^{(i)}(0) = B_d^{(i)}(1), \qquad i = 0, 1, \cdots, d - 2$$

and minimize $\int_0^1 B_d^2(x)\,dx$.

The coefficients a_l for $Q_k g = \sum_{l=1}^{k+d} a_l N_l$ can be determined explicitly by setting $\phi_{l,d}(s) = \prod_{r=1}^{d-1}(s - t_{l+r})$, $\gamma_{l,i}^j = ((-1)^j j!/(d - 1)!)\,\phi_{l,d}^{(d-1-j)}(\xi_i)$ and

$$a_l = \sum_{j=0}^{d-1} \gamma_{l,i}^j\left(g_i^j - g_i^d \delta_{i+1}^{d-j}\binom{d}{j}B_{d-j}\right), \qquad l = i + 1, \cdots, i + d. \tag{6.26}$$

By taking $i \equiv 0 \pmod d$ for sufficiently many i, all of the coefficients a_l can be determined. For $d = 2$ these coefficients turn out to be

$$a_{l+1} = g(\xi_l) - g^{(2)}(\xi_l)(\delta_l^2/12) + o(\delta^2).$$

Barrow and Smith (1978b) have shown that the operator Q_k, defined by the above scheme, satisfy (6.24) and the following.

LEMMA 6.8 (*Barrow and Smith* 1978b). *Let* $g \in C^d[0, 1]$, *and* $\bar{\xi} \in [0, 1)$. *Let* j *be chosen so that* $\xi_j \le \bar{\xi} < \xi_{j+1}$ *and let* $\delta_{j+1} = \xi_{j+1} - \xi_j$. *Let*

$$R_k(y, \bar{\xi}) = k^d(g - Q_k g)(\xi_j + y\delta_{j+1}), \qquad y \in [0, 1) \tag{6.27}$$

and

$$K(y, \bar{\xi}) = (g^{(d)}(\bar{\xi})/(p(\bar{\xi}))^d)(B_d(y)/d!).$$

Then there exists a sequence of positive constants $\{\epsilon_k\}_{k=1}^\infty$ *tending to zero and which may be chosen independently of* $\bar{\xi}$ *such that*

$$\|R_k(\cdot, \bar{\xi}) - K(\cdot, \bar{\xi})\|_\infty = \max_y |R_k(y, \bar{\xi}) - K(y, \bar{\xi})| < \epsilon_k.$$

As indicated above, this lemma says, in essence, that for k sufficiently large, the error function $g - Q_k g$ is nearly equal (in a sup norm) to a properly scaled Bernoulli polynomial on each subinterval (ξ_j, ξ_{j+1}).

PROOF OF THEOREM 3.1B. Let us recall that

$$B = \|g - P_k^n g\|_\lambda^2.$$

Since $Q_k g \in S_k^d$ we can write

$$g - P_k^n g = g - Q_k g - P_k^n(g - Q_k g).$$

Therefore, the proof of the theorem will be completed when we show that,

LEMMA 6.9. $\lim_{k\to\infty} k^{2d}\| g - Q_k g\|_\lambda^2 = C_f$, *where*

$$C_f = \{|B_{2d}|/(2d)!\} \int_0^1 \{(g^{(d)}(x))^2/(p(x))^{2d}\} f(x)\, dx.$$

LEMMA 6.10. $\lim_{k\to\infty} k^{2d}\| P_k^n(g - Q_k g)\|_\lambda^2 = 0$.

PROOF OF LEMMA 6.9. Let us consider, using (6.27)

$$k^{2d}\| g - Q_k g\|_\lambda^2 = k^{2d} \int_0^1 ((g - Q_k g)(x))^2 f(x)\, dx$$

$$= \textstyle\sum_{j=0}^k \delta_{j+1} \displaystyle\int_0^1 R_k^2(y, \xi_j) f(\xi_j + y\delta_{j+1})\, dy.$$

By Lemma (6.8), this equals

$$\textstyle\sum_{j=0}^k \delta_{j+1}[(g^{(d)}(\xi_j))^2/(p(\xi_j))^{2d}] \displaystyle\left\{\int_0^1 (B_d(y)/d!)^2 f(\xi_j + y\delta_{j+1})\, dy + \beta_{j,k}\right\}$$

where $|\beta_{j,k}| < \alpha\epsilon_k$, for some constant α which depends only on d, g and p. We also note that $f(\xi_j + y\delta_{j+1}) = f(\xi_j) + \gamma_j$, where $|\gamma_j| < \omega(f, \delta)$. Hence we have

$$k^{2d}\| g - Q_k g\|_\lambda^2 = \left(\int_0^1 (B_d(y)/d!)^2\, dy\right) \textstyle\sum_{j=0}^k \delta_{j+1}((g^{(d)}(\xi_j))/(p(\xi_j))^d)^2 f(\xi_j) + o(1).$$

Let $k \to \infty$ in such a way that $\delta = \max \delta_j \to 0$, then

$$\lim_{k\to\infty} k^{2d}\| g - Q_k g\|_\lambda^2 = C_f. \tag{6.28}$$

We use here the fact that

$$\int_0^1 (B_d(y)/d!)^2\, dy = |B_{2d}|/(2d)!$$

(see Ghizzetti and Ossicini (1970)). This proves the lemma.

PROOF OF LEMMA 6.10. Denoting $\| P_k^n(g - Q_k g)\|_\lambda$ by A, we have

$$A = \int_0^1 \{N'(x)M^{-1}(\mu^{(n_k)}) \int_0^1 N(y)(g - Q_k g)(y)\, d\mu^{(n_k)}(y)]^2 f(x)\, dx$$

$$= a'M^{-1}(\mu^{(n_k)})M(f)M^{-1}(\mu^{(n_k)})a,$$

where the $(k + d) \times 1$ vector a is given by

$$a = \int N(y)((g - Q_k g)(y))\, d\mu^{(n_k)}(y).$$

Using matrix norm properties, we see that

$$A \le \| a\|^2 \| M^{-1}(\mu^{(n_k)})\|^2 \| M(f)\|. \tag{6.29}$$

We have already found the bounds on the norm of the matrices $M(f)$ (or $M(\lambda)$) and $M^{-1}(\mu^{(n_k)})$ in Lemma 6.3 and Lemma 6.6. Here we shall find a bound on the norm of the vector a. The ith element ($i = 1, \cdots, k + d$) of this vector is,

$$a_i = \int_{t_i}^{t_{i+d}} N_i(x)((g - Q_k g)(x))\, d\mu^{(n_k)}(x)$$

$$= \textstyle\sum_{l=i}^{i+d-1} \displaystyle\int_0^1 N_i(t_l + y\delta_{l+1})\{(g - Q_k g)(t_l + y\delta_{l+1})\}\, d\mu^{(n_k)}(t_l + y\delta_{l+1}).$$

Using (6.27), we can write

$$k^d a_i = \textstyle\sum_{l=i}^{i+d-1} \displaystyle\int_0^1 N_i(t_l + y\delta_{l+1}) R_k(y, t_l)\, d\mu^{(n_k)}(t_l + y\delta_{l+1}).$$

Using Lemma 6.8, we get

$$k^d a_i = \textstyle\sum_{l=i}^{i+d-1} \displaystyle\int_0^1 N_i(t_l + y\delta_{l+1})\{\rho_l B_d(y) + \beta_k\}\, d\mu^{(n_k)}(t_l + y\delta_{l+1}),$$

where $\rho_l = g^{(d)}(t_l)/\{(p(t_l))^d d!\}$, and $\{\beta_k\}_{k=1}^{\infty}$ is a sequence of positive numbers tending to zero. Writing $\mu^{(n_k)}$ as $\mu + (\mu^{(n_k)} - \mu)$, we can see that

$$k^d a_i = \textstyle\sum_{l=i}^{i+d-1} \rho_l \displaystyle\int_{t_l}^{t_{l+1}} N_i(x) B_d((x - t_l)/\delta_{l+1})\, d\mu^{(n_k)}(x) + o(\delta)$$

$$(6.30) \qquad = \textstyle\sum_{l=i}^{i+d-1} \rho_l \displaystyle\int_{t_l}^{t_{l+1}} N_i(x) B_d((x - t_l)/\delta_{l+1})\, d\mu(x)$$

$$+ \textstyle\sum_{l=i}^{i+d-1} \rho_l \displaystyle\int_{t_l}^{t_{l+1}} N_i(x) B_d((x - t_l)/\delta_{l+1})\, d(\mu^{(n_k)} - \mu)(x) + o(\delta).$$

Now we show that the second factor on the right in (6.30) is $o(\delta)$. Since the sum involved in this factor is over d terms, it is enough to show that each term in this sum is $o(\delta)$. Using integration by parts, we can easily check that the lth term (except for the quantity ρ_l) in the second factor equals

$$N_i(t_{l+1}) B_d(H_{n_k} - H)(t_{l+1}) - N_i(t_l) B_d(H_{n_k} - H)(t_l)$$

$$(6.31) \qquad - \int_{t_l}^{t_{l+1}} (H_{n_k} - H)(x)\{\delta_{l+1}^{-1} N_i(x) B'_d((x - t_l)/\delta_{l+1})$$

$$+ N'_i(x) B_d(x - t_l)/\delta_{l+1})\}\, dx.$$

In the above, we have used $B_d(o) = B_d(1) = B_d$, the dth Bernoulli number. Using the upper bounds on (i) $\| N_i^{(j)} \|_\infty$ given in (6.19), (ii) $\| H_{n_k} - H \|_\infty$ given in (6.20) and the fact that the dth Bernoulli polynomial is bounded above by a number independent of k, we find that the quantity (6.31) is of order $o(\delta)$. Now,

$$k^d a_i = \textstyle\sum_{l=i}^{i+d-1} \rho_l \displaystyle\int_{t_l}^{t_{l+1}} N_i(x) B_d((x - t_l)/\delta_{l+1}) h(x)\, dx + o(\delta)$$

where h is the density of measure μ. On the interval (t_l, t_{l+1}), $h(x) = h(t_l) + \gamma_l$, where $|\gamma_l| < \omega(h, \delta)$, the modulus of continuity of h. We can write

$$(6.32) \qquad k^d a_i = \textstyle\sum_{l=i}^{i+d-1} \rho_l h(t_l) \displaystyle\int_{t_l}^{t_{l+1}} N_i(x) B_d((x - t_l)/\delta_{l+1})\, dx + o(\delta).$$

In the proof of their Lemma 2, Barrow and Smith (1978a) have an expression similar to the first factor on the right in (6.32) with $h \equiv 1$. They show, using the continuity of $g^{(d)}$ and the condition (3.3), that this quantity is of order $o(\delta)$. Since h is continuous on [0, 1], we can similarly show that the first factor in (6.32) is of order $o(\delta)$ and therefore $a_i = o(\delta^{d+1})$. Now we have

$$\| a \|^2 = \sum_{i=1}^{k+d} a_i^2 = o(\delta^{2d+1}).$$

Also by Lemma 6.3, $\| M(f) \| = 0(\delta)$ and by Lemma 6.6, $\| M^{-1}(\mu^{(n_k)}) \| = 0(\delta^{-1})$, therefore (6.29) gives $A = o(\delta^{2d})$. This completes the proof of the lemma, and hence also the proof of Theorem 3.1B.

Proof of Theorem 4.1B. Since the estimator $\hat{\theta}_{\mathrm{BME}}$, given in (4.4), satisfies (4.2), we can easily check that the bias term is asymptotically minimized, i.e.,

$$\lim_{k\to\infty} \lim_{n\to\infty} k^{2d} B = (| B_{2d} | / (2d)!) \int \{(g^{(d)}(x))^2 / (p(x))^{2d}\} f(x)\, dx.$$

Proof of Theorem 4.1A. With the choice of $\hat{\theta}_{\mathrm{BME}}$, as given in (4.4), it is easy to check that the integrated variance V is given by

$$\begin{aligned}(nV/\sigma^2) &= \operatorname{tr} C^* D^{-1}(\mu^{(n)}) C^{*\prime} M(f) \\ &= \operatorname{tr} M^{-1}(f) \left(\int N(x) L'(x) f(x)\, dx \right) D^{-1}(\mu^{(n)}) \left(\int N(y) L'(y) f(y)\, dy \right)' \qquad (6.33) \\ &= \operatorname{tr} M^{-1}(f) \left(\int \int N(x) N'(y) \sum_{i=1}^{r} \frac{L_i(x) L_i(y)}{\mu_i} f(x) f(y)\, dx\, dy \right).\end{aligned}$$

For the proof of the theorem we need the following lemma.

Lemma 6.11. *Let $u(x)$ and $v(x)$ be continuous functions defined on* [0, 1]. *If $\eta = \max_i (x_i - x_{i-1}) \to 0$ as $n \to \infty$, we have*

$$\lim_{n\to\infty} \int_0^1 \int_0^1 u(x) v(y) \sum_{j=1}^{r} \frac{L_j(x) L_j(y)}{\mu_j}\, dx\, dy = \int_0^1 \frac{u(x) v(x)}{h(x)}\, dx. \qquad (6.34)$$

The proof of the lemma is deferred until the end of this section. Assuming for the present the truth of Lemma 6.11 we complete the proof of the Theorem 4.1A. Let $n \to \infty$ in (6.33) and then use (6.34) to get

$$\lim_{n\to\infty} (nV/\sigma^2) = \operatorname{tr} M^{-1}(f) M(f^2/h)$$

where $M(f^2/h) = \int N(x) N'(x) (f^2(x)/h(x)/h(x))\, dx$. If we take $\phi \equiv f$ and $\psi \equiv f^2/h$ in Theorem 6.1, we then see that

$$\lim_{k\to\infty} (k^{-1} \operatorname{tr} M^{-1}(f) M(f^2/h)) = \int (f(x)/h(x)) p(x)\, dx$$

which completes the proof of the theorem. □

Proof of Lemma 6.11. Let us denote by I, the double integral on left of (6.34). Since $L_j(x)$ has support on the interval (x_{j-1}, x_{j+1}), we can express the integral I as

$$I = \sum_{j=1}^{r} \mu_j^{-1} \int_{x_{j-1}}^{x_{j+1}} \int_{x_{j-1}}^{x_{j+1}} u(x) v(y) L_j(x) L_j(y)\, dx\, dy,$$

where $x_0 = x_1 = 0$ and $x_{r+1} = x_r = 1$. By use of the mean value theorem, we get

$$I = \sum_{j=1}^{r} \left\{ \mu_j^{-1} u(x_j) v(x_j) \int_{x_{j-1}}^{x_{j+1}} \int_{x_{j-1}}^{x_{j+1}} L_j(x) L_j(y)\, dx\, dy \right\} + (1/4) \sum_{j=1}^{r} \mu_{j=1}^{-1} \gamma_j (n_j + \eta_{j+1})^2$$

where $| \gamma_j | < \alpha(\omega(u, \eta) + \omega(v, \eta))$ where α depends only on u and v, and $\omega(u, \eta)$ and $\omega(v, \eta)$ are the modulus of continuity of u and v. From (4.5), for $1 \leq j \leq r$,

$$\mu_j = (½)h(x_j)(\eta_j + \eta_{j+1})(1 + \tau_j)$$

where $\eta_j = x_j - x_{j-1}, j = 2, \cdots, r, \eta_1 = \eta_{r+1} = 0$, and $| \tau_j | < \rho\omega(h, \eta)$ where constant ρ depends only on h. Therefore now I equals

$$\sum_{j=1}^{r} \{u(x_j)v(x_j)/h(x_j)\}\{(\eta_j + \eta_{j+1})/2\} + o(1).$$

Now the proof of lemma follows since this sum is a Riemann sum for the integral on right of (6.34).

REFERENCES

[1] AGARWAL, G. G. (1978). Asymptotic design and estimation in polynomial spline regression. Technical Report No. 78-11, Purdue Univ.

[2] AGARWAL, G. G. and STUDDEN, W. J. (1978a). Asymptotic design and estimation using linear splines. *Comm. Statist.—Simulation Comput.* **B7(4)** 309–319.

[3] AGARWAL, G. G. and STUDDEN, W. J. (1978b). An algorithm for selection of design and knots in the response curve estimation by spline functions. Technical Report No. 78-15, Purdue Univ.

[4] BARROW, D. L. and SMITH, P. W. (1978a). Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quart. Appl. Math.* **36**, 293–304.

[5] BARROW, D. L. and SMITH, P. W. (1979). Efficient L_2 approximation by splines. *Numer. Math.* **33** 101–114.

[6] BOX, G. E. P. and DRAPER, N. R. (1959). A basis for selection of response surface design. *J. Amer. Statist. Assoc.* **54** 622–654.

[7] BURCHARD, H. G. (1974). Splines (with optimal knots) are better. *Applicable Anal.* **3** 309–319.

[8] CURRY, H. B. and SCHOENBERG, I. J. (1966). On Polya frequency functions IV: the fundamental spline functions and their limits. *J. Analyse Math.* **17** 71–107.

[9] DEBOOR, C. (1972). On calculating with *B*-splines. *J. Approximation Theory* **6** 50–62.

[10] DEBOOR, C. (1973). The quasi-interpolant as a tool in elementary spline theory. In *Approximation Theory* (G. G. Lorentz, ed.) Academic, New York.

[11] DEBOOR, C. (1978). *A Practical Guide to Splines.* Springer, New York.

[12] DEBOOR, C. and FIX, G. J. (1973). Spline approximation by quasiinterpolants. *J. Approximation Theory* **8** 19–45.

[13] DODSON, D. S. (1972). Optimal order approximation by polynomial spline function. Unpublished Ph.D. thesis, Depart. Comput. Sci., Purdue Univ.

[14] GHIZZETTI, A. and OSSICINI, A. (1970). *Quadrature Formulae.* Academic, New York.

[15] GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

[16] GUEST, P. G. (1961). *Numerical Methods of Curve Fitting.* Cambridge Univ.

[17] HOLT, J. N. (1974). A self-limiting inversion technique for centre to limb data. *Astronom. and Astrophys.* **30** 185–188.

[18] ICHIDA, K., KIYONO, T. and YOSHIMOTO, F. (1977). Curve fitting by a one-pass method with a piecewise cubic polynomial. *ACM Trans. Math. Software* **3** 164–174.

[19] JUPP, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM J. Numer. Anal.* **15** 328–343.

[20] JUPP, D. L. B. and STEWART, I. C. F. (1974). A piecewise exponential model for seismic well logging data. *J. Internat. Assoc. Mathematical Geol.* **6** 33–45.

[21] KARSON, M. J., Manson, A. R. and HADER, R. J. (1969). Minimum bias estimation and experimental design for response surfaces. *Technometrics* **11** 461–475.

[22] LAWTON, W. H., SYLVESTRE, E. A. and MAGGIO, M. S. (1972). Self modeling nonlinear regression. *Technometrics* **14** 513–532.

[23] MILNE-THOMSON, L. M. (1951). *The Calculus of Finite Differences.* Macmillan, London.

[24] NÖRLUND, N. E. (1942). *Vorlesungen Über Differenzenrechnung, Verlag Von Julius.* Springer, Berlin

[25] PRENTER, P. M. (1975). *Splines and Variational Methods*: Wiley, New York.

[26] RICE, J. R. (1969). On the degree of convergence of nonlinear spline approximation. In *Approximation Theory with Special Emphasis on Spline Functions.* (I. J. Schoenberg, ed.) Academic, New York.

[27] SACKS, J. and YLVISAKER, D. (1970). Design for regression problems with correlated errors III. *Ann. Math. Statist.* **41** 2057–2074.

[28] SCHOENBERG, I. J. (1966). On spline functions. MRC Technical Summary Report No. 625, Univ. Wisconsin-Madison.

[29] SCHOENBERG, I. J. (1969). Monosplines and quadrature formulae. In *Theory and Application of Spline Functions.* (T. N. E. Greville, ed.) Academic, New York.

[30] VOZOFF, K. and JUPP, D. L. B. (1975). Joint inversion of geophysical data. *Geophys. J. Roy. Astronom. Soc.* **42** 977–991.

[31] WAHBA, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Proc. Symp. Appl. Statist.* (P. R. Krishnaiah, ed.) North Holland, New York.

[32] WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

[33] WOLD, S. (1971). Analysis of kinetic data by means of spline functions. *Chem. Scripta* **1** 97–102.

[34] WOLD, S. (1974). Spline functions in data analysis. *Technometrics* **16** 1–11.

INDIAN STATISTICAL INSTITUTE
203 B. T. ROAD
CALCUTTA 700035
INDIA

DEPARTMENT OF STATISTICS
MATHEMATICAL SCIENCES BLDG.
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907