J. R. Statist. Soc. B (1978),
40, *No.* 3, *pp.* 364–372

Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression

By GRACE WAHBA

University of Wisconsin at Madison

(Received February 1978. Final revision July 1978)

SUMMARY

Spline and generalized spline smoothing is shown to be equivalent to Bayesian estimation with a partially improper prior. This result supports the idea that spline smoothing is a natural solution to the regression problem when one is given a set of regression functions but one also wants to hedge against the possibility that the true model is not exactly in the span of the given regression functions. A natural measure of the deviation of the true model from the span of the regression functions comes out of the spline theory in a natural way. An appropriate value of this measure can be estimated from the data and used to constrain the estimated model to have the estimated deviation. Some convergence results and computational tricks are also discussed.

Keywords: SPLINE SMOOTHING; IMPROPER PRIORS; NONPARAMETRIC REGRESSION; MODEL ERRORS

1. INTRODUCTION

CONSIDER the model

$$Y(t_i) = g(t_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n; \quad t_i \in \mathscr{T}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \sim \mathscr{N}(0, \sigma^2 I_{n\times n})$ and $g(\cdot)$ is some "smooth" function defined on some index set $\mathscr{T}$. When $\mathscr{T}$ is an inverval of the real line, cubic polynomial smoothing splines are well known to provide an aesthetically satisfying method for estimating $g(\cdot)$, from a realization $\mathbf{y} = (y_1, \ldots, y_n)'$ of $\mathbf{Y} = (Y(t_1), \ldots, Y(t_n))'$. See Rowlands, Liber and Daniel (1974) for a very useful example. Splines are an appealing alternative to fitting a specified set of m regression functions, for example polynomials of degree less than m, when one is uncertain that the true curve $g(\cdot)$ is actually in the span of the specified regression functions. Kimeldorf and Wahba (1970a, b, 1971) explored certain relationships between Bayesian estimation and spline smoothing. In this note we provide a somewhat different formulation and generalization of the result in Kimeldorf and Wahba (1971). Here we prove that polynomial spline (respectively generalized spline) smoothing is equivalent to Bayesian estimation with a prior on g which is "diffuse" on the coefficients of the polynomials of degree $<m$ (respectively specified set of m regression functions), and "proper" over an appropriate set of random variables not including the coefficients of the regression functions. Since Gauss Markov estimation is equivalent to Bayesian estimation with a prior diffuse over the coefficients of the regression functions, this result leads to the conclusion that spline smoothing is a (the ?) natural extension of Gauss–Markov regression with m specified regression functions. We claim that spline smoothing is an appropriate solution to the problem arising when one wants to fit a given set of regression functions to the data but one also wants to "hedge" against model errors, that is, against the possibility that the true model g is not exactly in the span of the given set of regression functions. We show that the spline smoothing approach leads to a natural measure of the deviation of the true g from the span of the regression functions and, furthermore, a good value of this measure can be estimated from the data. The estimated value of the measure is then used to control the deviation of the estimated g.

From another point of view this measure can be viewed as the "bandwidth parameter" which controls the "smoothness" of the estimated g, and so in this approach to non-parametric (or semi-parametric) regression, a good value of the bandwidth parameter can be estimated from the data.

Smith (1973) introduced uncertainty about a particular form of regression model in a Bayesian context, though he did not assume a non-parametric form for the regression function. The present work is philosophically close to that of Blight and Ott (1975), who adopted a Bayesian approach to estimating g, and part of this work may be considered to be a generalization of theirs. O'Hagan (1978) also treats the model (1.1) from a Bayesian viewpoint, but the details of his approach appear to be somewhat different. The model of Young (1977) can also be seen, in part, to be a special case of our generalized spline model with $m = 1$, although our approach diverges from Young's at the point where he introduces priors on his "hyperparameters". None of these works provide the feature of estimating the bandwidth parameter from the data. The present set-up is briefly mentioned in my discussion to O'Hagan's paper, where it is observed that O'Hagan's experimental design criteria (for the choice of $t_1, \ldots, t_n$) can be formulated in the context of the approach in the present paper.

Other approaches to the estimation of g in the model (1.1) have been made by Priestly and Chao (1972), Benedetti (1977), Clark (1977) and Stone (1977). Priestly and Chao, and Benedetti use kernel non-parametric regression to estimate g and provide mean square error convergence rates. For the polynomial smoothing splines considered here, integrated m.s.e. convergence rates of the estimated g to the true g, as $\max_i |t_{i+1}-t_i| \to 0$, have been recently found by Craven and Wahba (1977) and are quoted in Section 5 for comparison with Priestly and Chao's, and Benedetti's results.

In Section 4 we make some remarks concerning the efficient computation of generalized splines.

We note that the method for estimating the "bandwidth" parameter of this paper can also be used in connection with certain density and log spectral density estimates, see Wahba (1978a, b).

Other recent related work is Silverman (1978a) who provides a different approach to estimating the bandwidth parameter in the density estimation context, and (1978b) provides a spline estimate of the log density ratio, and Leonard (1978) who develops density estimates from a Bayesian point of view.

2. Polynomial Splines as Posterior Means with a Partially Improper Prior on the Polynomials of Degree Less than m

Let $\mathscr{T} = [0,1]$. Given data $\{y(t_1), \ldots, y(t_n)\}$, $0 < t_1 \ldots t_n < 1$, the smoothing polynomial spline of degree $2m-1$ to the data, call it $g_{n,\lambda}$, is defined as the solution to the minimization problem: Find $g \in W_2^{(m)}$: $\{g: g, g', \ldots, g^{(m-1)}$ abs. cont. $g^{(m)} \in \mathscr{L}_2[0,1]\}$ to minimize

$$n^{-1}\sum_{j=1}^{n}(g(t_j)-y_j)^2+\lambda\int_0^1 (g^{(m)}(u))^2\,du, \tag{2.1}$$

where $y_j = y(t_j)$, and λ is to be chosen. If y cannot be interpolated exactly by some polynomial of degree less than m, then the solution is well known to be unique, and to be a polynomial spline of degree $2m-1$ (see Schoenberg, 1964), that is, it is piecewise a polynomial of degree $2m-1$ in each interval $[t_i, t_{i+1}]$, $i = 1, 2, \ldots, n-1$, with the pieces joined so that the resulting function has $2m-2$ continuous derivatives. An efficient computational algorithm for the cubic polynomial smoothing spline ($m = 2$) is given by Reinsch (1967) and code is available in the IMSL library (1977). We show that the spline solution $g_{n,\lambda}$ to the minimization problem of (2.1) is a Bayesian estimate for g with a "partially diffuse" prior; the quantity $J = \int_0^1 (g_{n,\lambda}^{(m)}(u))^2\,du$ is a natural measure of the deviation of $g_{n,\lambda}$ from the span of the polynomials of degree less than m, and furthermore a good value of J can be estimated from the data.

Theorem 1. Let $g(t)$, $t\in[0,1]$ have the prior distribution which is the same as the distribution of the stochastic process $X_\xi(t)$, $t\in[0,1]$,

$$X_\xi(t) = \sum_{j=1}^{m} \theta_j \phi_j(t) + b^{\frac{1}{2}} Z(t), \tag{2.2}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \sim \mathcal{N}(0, \xi I_{m\times m})$, $\phi_j(t) = t^{j-1}/(j-1)!$, $j = 1, 2, \ldots, m$, b is fixed, and $Z(t)$ is the m-fold integrated Wiener process (Shepp, 1966),

$$Z(t) = \int_0^t \frac{(t-u)^{m-1}}{(m-1)!}\, dW(u). \tag{2.3}$$

Then

(i) The polynomial spline $g_{n,\lambda}(\cdot)$ which is the minimizer of (2.1) has the property

$$g_{n,\lambda}(t) = \lim_{\xi\to\infty} E_\xi\{g(t) \mid \mathbf{Y} = \mathbf{y}\} \tag{2.4}$$

with $\lambda = \sigma^2/nb$, where E_ξ is expectation over the posterior distribution of $g(t)$ with the prior (2.2). ($\xi = \infty$ corresponds to the "diffuse" prior on $\boldsymbol{\theta}$.)

(ii) Suppose $\mathbf{y}$ cannot be interpolated exactly by some polynomial of degree less than m. Then $\lim g_{n,\lambda}(\cdot)$, as $\lambda\to\infty$, is the polynomial of degree $m-1$ best fitting the data in a least squares sense, $\lim g_{n,\lambda}(\cdot)$, as $\lambda\to 0$, is that function in $W_2^{(m)}$ which minimizes $\int_0^1 (g^{(m)}(u))^2\, du$ subject to the condition that it interpolates $\mathbf{y}$, and $J(\lambda) = \int_0^1 (g_{n,\lambda}^{(m)}(u))^2\, du$ is a monotone strictly decreasing function of λ.

(iii) Let loss be measured by the mean square prediction error $R(\lambda)$ given by

$$R(\lambda) = n^{-1} \sum_{j=1}^{n} (g(t_j) - g_{n,\lambda}(t_j))^2.$$

Define $\hat{R}(\lambda)$ by

$$\hat{R}(\lambda) = n^{-1}\{\| (I - A(\lambda))\, \mathbf{y}\|^2 + \sigma^2 \operatorname{tr} A^2(\lambda) - \sigma^2 \operatorname{tr} (I - A(\lambda))^2\},$$

where $A(\lambda)$ is the symmetric non-negative definite matrix satisfying

$$\mathbf{g}_{n,\lambda} = A(\lambda)\, \mathbf{y},$$

where

$$\mathbf{g}_{n,\lambda} = (g_{n,\lambda}(t_1), \ldots, g_{n,\lambda}(t_n))'.$$

If $\mathbf{g} = (g(t_1), \ldots, g(t_n))'$ is viewed as fixed, and expectation taken with respect to $\boldsymbol{\varepsilon}$, then

$$E\hat{R}(\lambda) = ER(\lambda)$$

so that an optimum λ for squared error of prediction loss may be estimated from the data by minimizing $\hat{R}(\lambda)$.

Before giving the proof we discuss the meaning of this Theorem. We interpret (i) and (ii) as saying that estimation with the polynomial spline $g_{n,\lambda}$ should be viewed as a (the ?) natural extension of Gauss–Markov estimation with polynomial regression functions (i.e. estimation with $g_{n,\infty}$). This is because the Gauss–Markov regression estimate can be obtained as the posterior mean of g when g has a prior diffuse on the coefficients of the polynomials; $g_{n,\lambda}$, $\lambda<\infty$ is obtained as the posterior mean of g when g has a diffuse prior on the coefficients of the polynomials modified by the addition of $b^{\frac{1}{2}} Z(\cdot)$ to the prior specification, $b>0$.

In practice $\lambda = \sigma^2/nb$ is not generally known, so that it is fortunate that λ can be estimated from the data via (iii). If σ^2 is not known an estimate of λ which minimizes $ER(\lambda)$ asymptotically for large n for fixed $g\in W_2^{(m)}$ can be obtained by using the method of generalized cross-validation (GCV) as described in Craven and Wahba (1977).

Proof of Theorem 1. Part (ii) is well known, see Schoenberg (1964), Reinsch (1967, 1971), Anselone and Laurent (1968), Kimeldorf and Wahba (1970b). To prove (i) we use Lemma 5.1 of Kimeldorf and Wahba (1971) where an explicit formula for $g_{n,\lambda}$ is given. It is

$$g_{n,\lambda}(t) = (\phi_1(t), \ldots, \phi_m(t))(T' M^{-1} T)^{-1} T' M^{-1} y + (Q_{t_1}(t), \ldots, Q_{t_n}(t)) M^{-1}(I - T(T' M^{-1} T)^{-1} T' M^{-1}) y, \tag{2.5}$$

where T is the $n \times m$ matrix of rank m with jkth entry $\phi_k(t_j)$, $M = n\lambda I_{n\times n} + Q_n$, Q_n is the $n \times n$ matrix with jkth entry $Q(t_j, t_k)$ and $Q_{t_i}(t) \equiv Q(t_i, t)$, where

$$Q(s,t) = \int_0^1 \frac{(s-u)_+^{m-1}}{(m-1)!} \frac{(t-u)_+^{m-1}}{(m-1)!} du.$$

(We remark that $Q(s,t) = EZ(s)Z(t)$.) With the prior of (2.2) it is easily seen that the prior covariances $E\mathbf{Y}' X_\xi(t)$ and $E\mathbf{Y}'\mathbf{Y}$ are

$$E\mathbf{Y}' X_\xi(t) = \xi(\phi_1(t), \ldots, \phi_m(t) T' + b(Q_{t_1}(t), \ldots, Q_{t_n}(t)),$$

$$E\mathbf{Y}\mathbf{Y}' = \xi TT' + bQ_n + \sigma^2 I.$$

Setting $\lambda = \sigma^2/nb$, $\eta = \xi/b$ and $M = Q_n + n\lambda I$ gives

$$E\{X_\xi(t) \mid \mathbf{Y} = \mathbf{y}\} = (\phi_1(t), \ldots, \phi_m(t)) \eta T'(\eta TT' + M)^{-1} \mathbf{y} + (Q_{t_1}(t), \ldots, Q_{t_n}(t))(\eta TT' + M)^{-1} \mathbf{y}. \tag{2.6}$$

By comparing (2.5) and (2.6), it remains only to show that

$$\lim_{\eta\to\infty} \eta T'(\eta TT' + M)^{-1} = (T' M^{-1} T)^{-1} T' M^{-1} \tag{2.7}$$

and

$$\lim_{\eta\to\infty} (\eta TT' + M)^{-1} = M^{-1}(I - T(T' M^{-1} T)^{-1} T' M^{-1}). \tag{2.8}$$

Now, it can be verified that

$$(\eta TT' + M)^{-1} = M^{-1} - M^{-1} T(T' M^{-1} T)^{-1}\{I + \eta^{-1}(T' M^{-1} T)^{-1}\}^{-1} T' M^{-1}, \tag{2.9}$$

and expanding in powers of η and letting $\eta \to \infty$ completes the proof of (2.7) and (2.8). Part (iii) appears in Craven and Wahba (1977), but since the proof is immediate we give it here: We have

$$ER(\lambda) = En^{-1} \| A(\lambda)\mathbf{y} - \mathbf{g}\|^2 = n^{-1}\{\| (I - A(\lambda))\mathbf{g}\|^2 + \sigma^2 \operatorname{tr} A^2(\lambda)\}$$

and (iii) follows from

$$E\| (I - A(\lambda))\mathbf{y}\|^2 = \| (I - A(\lambda))\mathbf{g}\|^2 + \sigma^2 \operatorname{tr}(I - A(\lambda))^2.$$

We remark that $A(\lambda)$ is obtained from (2.5) and is

$$A(\lambda) = T(T' M^{-1} T)^{-1} T' M^{-1} + Q_n M^{-1}(I - T(T' M^{-1} T)^{-1} T' M^{-1}).$$

Craven and Wahba (1977) and Utreras (1978) both give some aesthetically very pleasing plots of $g_{n,\hat\lambda}$, where $\hat\lambda$ is chosen by the GCV method, and $m = 2$. (The GCV method chooses λ to minimize $V(\lambda) = \| (I - A(\lambda))y\|^2 / [\operatorname{tr}(I - A(\lambda))]^2$.) Both reports demonstrate nicely how well $g_{n,\hat\lambda}$ recovers g. If σ^2 is known accurately, the minimizer of $\hat R(\lambda)$ can be expected to behave much like $\hat\lambda$. In Craven and Wahba, the algorithm of Reinsch (1967) is used to compute $g_{n,\lambda}$. Utreras (1978) gives approximate expressions for the eigenvalues of $A(\lambda)$ in the large n, equally spaced data case which can considerably simplify the calculation of $\hat R(\lambda)$ or $V(\lambda)$.

We remark that m as well as λ can be estimated from the data by minimizing $\hat R$ (or V) as a function of both these parameters.

Numerical experiments in estimating m as well as λ have been performed in connection with the log spectral density estimates of Wahba (1978b) and it was found that a modest improvement in mean square error can sometimes be made by estimating m, instead of using $m = 2$, the cubic spline case.

3. Generalized Splines as Posterior Means with a Partially Improper Prior

We now consider the general case where polynomials on [0, 1] are replaced by some real-valued functions $\{\phi_j(\cdot)\}_{j=1}^m$ defined on some arbitrary index set $\mathcal{T}$. For example, $\mathcal{T}$ may be a square or sphere. We require only that the $n \times m$ matrix with jkth entry $\phi_k(t_j)$ be of rank m. Families of extensions of Gauss–Markov estimates analogous to the smoothing polynomial spline will be found. These estimates will be generalized splines.

A very general form of Theorem 1, for these essentially arbitrary $\mathcal{T}$ and $\{\phi\}$ can be stated in the context of reproducing kernel Hilbert spaces (r.k.h.s.). We have concluded from the work of Parzen (1961, 1970), that r.k.h.s. is in fact a natural setting for analysing arbitrary Gaussian stochastic processes with continuous time parameter. Thus, we beg the reader's indulgence while we give a definition of a generalized spline as the solution to a minimization problem in r.k.h.s. Then we proceed to the general form of Theorem 1.

We note (Aronszajn, 1950), that a (real) r.k.h.s. $\mathcal{H}$ is a Hilbert space of real-valued functions on $\mathcal{T}$ with the property that, for each fixed $t_* \in \mathcal{T}$, the linear functional which maps $g \in \mathcal{H}_K$ to $g(t_*)$ is a continuous linear functional. Then, by the Riesz representation theorem (Akhiezer and Glazman, 1961, p. 33), there exists an element, call it δ_{t_*} in $\mathcal{H}$ such that $\langle g, \delta_{t_*} \rangle = g(t_*)$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. We can associate with $\mathcal{H}$ the so-called reproducing kernel (r.k.) $K(s,t), s, t \in \mathcal{T}$, defined by $K(s,t) = \langle \delta_s, \delta_t \rangle$, clearly $\delta_s(t) = K(s,t)$. The kernel $K(s,t)$ is always positive definite (since $\|\Sigma \alpha_i \delta_{s_i}\|^2 \geqslant 0$) and so there always exists a Gaussian stochastic process with K as its covariance. We will denote by $\mathcal{H}_K$ the r.k.h.s. with r.k. $\mathcal{H}$, and let the inner product in $\mathcal{H}_K$ be $\langle \cdot, \cdot \rangle_K$.

We now let $\mathcal{H}_K$ be any r.k.h.s. of real-valued functions on $\mathcal{T}$ which contain the $\{\phi_j\}$. (Construction of such $\mathcal{H}_K$ when the $\{\phi_j\}$ are any extended Tchebychev system of functions on [0, 1] may be found in Kimeldorf and Wahba (1971).) It is not hard to show that $\mathcal{H}_K$ has a representation as the direct sum of span $\{\phi_j\}$ and $\mathcal{H}_Q$, the r.k.h.s. with r.k. $Q(s,t), s, t \in \mathcal{T}$ given by (see Wahba, 1973)

$$Q(s,t) = K(s,t) - \sum_{i,j=1}^{m} \phi_i(s) k_{ij} \phi_j(t),$$

where k_{ij} is the ijth entry of the inverse of the (necessarily strictly positive definite) matrix with ijth entry $\langle \phi_i, \phi_j \rangle_K$. Let P_Q be the orthogonal projection operator in $\mathcal{H}_K$ onto $\mathcal{H}_Q$. (That is, $I - P_Q$ is the orthogonal projection in $\mathcal{H}_K$ onto span $\{\phi_j\}$.) The analogue of $\int_0^1 (g^{(m)}(u))^2 du$ is $\|P_Q g\|_K^2$, and this is, of course, a measure of the deviation of g from span $\{\phi_j\}$, being the distance in $\mathcal{H}_K$ from g to span $\{\phi_j\}$.

Suppose $\mathbf{y}$ is not in the span of the vectors $\{\boldsymbol{\phi}_j\}_{j=1}^m$, where $\boldsymbol{\phi}_j = (\phi_j(t_1), \ldots, \phi_j(t_n))'$. Then (Anselone and Laurent, 1968; Kimeldorf and Wahba, 1971) there is a unique solution, call it $g_{n,\lambda}$, to the minimization problem: Find $g \in \mathcal{H}_K$ to minimize

$$n^{-1} \sum_{j=1}^{n} (g(t_j) - y_j)^2 + \lambda \|P_Q g\|_K^2. \tag{3.1}$$

We shall call any $g_{n,\lambda}$ obtained as a solution of this minimization problem a generalized smoothing spline, or, consistent with the terminology in Anselone and Laurent, just a smoothing spline.

Theorem 2. Let $g(t)$, $t\in\mathscr{T}$ have the prior distribution which is the same as the distribution of the stochastic process $X_\xi(t)$,

$$X_\xi(t)=\sum_{j=1}^{m}\theta_j\phi_j(t)+b^{\frac{1}{2}}Z(t),\quad t\in\mathscr{T},\tag{3.2}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)\sim\mathscr{N}(0,\xi I_{m\times m})$, b is fixed $\geqslant 0$ and $Z(t)$ is a zero mean Gaussian stochastic process with $EZ(s)Z(t)=Q(s,t)$. Then:

(i) The generalized spline $g_{n,\lambda}$ which is the minimizer of (3.1) has the property

$$g_{n,\lambda}(t)=\lim_{\xi\to\infty}E_\xi\{g(t)\,|\,\mathbf{Y}=\mathbf{y}\},$$

with $\lambda=\sigma^2/nb$, where E_ξ is expectation over the posterior distribution of $g(t)$ with the prior (3.2).

(ii) Suppose $\mathbf{y}$ is not in the span of the $\{\boldsymbol{\phi}_j\}$. Then $\lim_{\lambda\to\infty}g_{n,\lambda}(\cdot)$ is that element in span $\{\phi_j(\cdot)\}$ best fitting the data in a least squares sense. If Q_n, the $n\times n$ matrix with ijth entry $Q(t_i,t_j)$, is of full rank, $\lim_{\lambda\to 0}g_{n,\lambda}(\cdot)$ is that function in $\mathscr{H}_K$ which minimizes $\|P_Q g\|_K^2$ subject to the conditions that it interpolate the data, and $J(\lambda)=\|P_Q g_{n,\lambda}\|_K^2$ is a monotone decreasing function of λ.

(iii) Let

$$R(\lambda)=n^{-1}\sum_{j=1}^{n}(g(t_j)-g_{n,\lambda}(t_j))^2,$$

and define $\hat{R}(\lambda)$ by

$$\hat{R}(\lambda)=n^{-1}\{\|(I-A(\lambda))\mathbf{y}\|^2+\sigma^2\operatorname{tr}A^2(\lambda)-\sigma^2\operatorname{tr}(I-A(\lambda))^2\},$$

where $A(\lambda)$ is the symmetric, non-negative definite matrix satisfying

$$\mathbf{g}_{n,\lambda}=A(\lambda)\mathbf{y}.$$

If $\mathbf{g}$ is viewed as fixed, then

$$ER(\lambda)=E\hat{R}(\lambda)$$

so that an optimum λ for squared error of prediction loss may be estimated from the data by minimizing $\hat{R}(\lambda)$.

We remark that the function $g_{n,\lambda}$ is given by (2.5), with $Q(s,t)=EZ(s)Z(t)$, similarly, the matrix $A(\lambda)$ is as in Section 2.

Proof of Theorem. Beginning with Lemma 5.1 of Kimeldorf and Wahba (1971), the proof parallels directly the proof of Theorem 1, and is omitted.

4. Representations of $g_{n,\lambda}$ for Efficient Computing

We believe smoothing splines to be appropriate for solving a wide variety of practical problems, in practice, including smoothing surfaces, once efficient numerical algorithms are developed. If $\mathscr{H}_K$ is a space of periodic functions on $[0,1]$ or a tensor product of periodic spaces on $[0,1]\times\ldots\times[0,1]$, and the $\{t_i\}$ are equally spaced or the tensor product of equally spaced points then computing problems are readily solved. (See Wahba, 1977, for a computed example.) In general, however, the efficient computation of $g_{n,\lambda}$ presents challenges, if n is very large, as would usually be the case if $\mathscr{T}$ is a rectangle in d-space. It will probably be necessary to choose Q with computational ease—an important consideration.

Equation (2.5) will generally not be the best representation for computing $g_{n,\lambda}$. We discuss some other representations for $g_{n,\lambda}$ chosen with efficient computing in mind. We assume below that Q_n is of full rank. Since

$$T'M^{-1}(I-T(T'M^{-1}T)^{-1}T'M^{-1})=O_{m\times n}$$

it is clear that $g_{n,\lambda}$ has a representation

$$g_{n,\lambda} = \sum_{i=1}^{m} \theta_i \phi_i + \sum_{i=1}^{n-m} c_i h_i, \tag{4.1}$$

where $\boldsymbol{\theta}$ and $\mathbf{c} = (c_1, \ldots, c_{n-m})$ are vectors of constants, and

$$h_i(\cdot) = \sum_{j=1}^{n-m} b_{ij} Q_{tj}(\cdot),$$

where the $(n-m)\times n$ dimensional matrix B with ijth entry b_{ij} satisfies $BT = O_{(n-m)\times m}$ but is otherwise arbitrary.

We will demonstrate shortly that $\mathbf{c}$, $\boldsymbol{\theta}$ and $A(\lambda)\mathbf{y} = \mathbf{g}_{n,\lambda}$ satisfy

$$(\Sigma_h + n\lambda BB')\mathbf{c} = B\mathbf{y}, \tag{4.2}$$

$$T\boldsymbol{\theta} = y - MB'\mathbf{c} \tag{4.3}$$

and

$$\mathbf{g}_{n,\lambda} \equiv A(\lambda)\mathbf{y} = y - n\lambda B'\mathbf{c}, \tag{4.4}$$

where Σ_h is the $(n-m)\times(n-m)$ dimensional matrix with jkth entry $\langle h_i, h_j\rangle_Q$. One attempts to choose B so that $\{h_j\}$, B and Σ_h have convenient properties for computing, and then to obtain $\mathbf{c}$, $\boldsymbol{\theta}$, $\mathbf{g}_{n,\lambda}$ and $g_{n,\lambda}(\cdot)$ from (4.1) to (4.4) by first solving the linear system (4.2). In the polynomial spline case, by choosing the entries in B corresponding to divided differences, one can obtain Σ_h and B both banded matrices and an efficient code results (see Reinsch, 1967; Anselone and Laurent, 1968). The span of the $\{h_j\}$ can be constructed from B-splines, which are nice hill-like functions (see Curry and Schoenberg, 1966; deBoor, 1972).

Equation (4.2) can be shown to be equivalent to Anselone and Laurent, equations (8.26) and (9.1). However, we provide a direct proof of (4.2) using (2.5) without the elegant but lengthy machinery of their work. We must show that

$$(h_1, \ldots, h_{n-m})(\Sigma_h + n\lambda BB')^{-1} B\mathbf{y} \equiv (Q_{t_1}, \ldots, Q_{t_n})(M^{-1} - M^{-1}T(T'M^{-1}T)^{-1}T'M^{-1})\mathbf{y}. \tag{4.5}$$

Now since $\langle Q_{t_i}, Q_{t_j}\rangle_K = Q(t_i, t_j)$, we have that $\Sigma_h = BQ_nB'$ and so the left-hand side of (4.5) is given by

$$(Q_{t_1}, \ldots, Q_{t_n})B'(BMB')^{-1}B\mathbf{y}. \tag{4.6}$$

However,

$$B'(BMB')^{-1}B \equiv M^{-1} - M^{-1}T(T'M^{-1}T)^{-1}T'M^{-1} \tag{4.7}$$

as can be seen by observing that the $n\times n$ matrix $X = \left[\begin{array}{c} T' \\ \hline BM \end{array}\right]$ is of full rank and

$$X\{M^{-1} - M^{-1}T(T'M^{-1}T)^{-1}TM^{-1}\}X' = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & BMB' \end{array}\right) = X\{B'(BMB')^{-1}B\}X'. \tag{4.8}$$

Equations (4.3) and (4.4) follow immediately from (4.2) and (3.3).

5. Convergence Properties of $g_{n,\lambda}$

In the case of polynomial splines with $\mathscr{T} = [0, 1]$ the mean square error convergence properties (of $ER(\lambda)$) are known from Craven and Wahba (1977) and we give them here for comparison purposes. We have, from Theorem 1,

$$ER(\lambda) = En^{-1}\sum_{i=1}^{n}(g(t_i) - g_{n,\lambda}(t_i))^2 \equiv n^{-1}\{\|(I - A(\lambda))\mathbf{g}\|^2 + \sigma^2 \operatorname{tr} A^2(\lambda)\}.$$

Using Lemmas 4.1 and 4.3 of Craven and Wahba it can be shown (ignoring terms of $o(1)$), that an upper bound on $ER(\lambda)$ is given by

$$ER(\lambda) \leqslant \lambda \int_0^1 (g^{(m)}(u))^2 \, du + \frac{c}{n\lambda^{1/2m}},$$

where

$$c = \sigma^2 \max_i [n(t_{i+1} - t_i)]^{1/2m} \int_0^\infty \frac{dx}{(1+x^{2m})^2}.$$

This bound is minimized for $\lambda = \text{const} \, n^{-2m/(2m+1)}$ and so

$$\min_\lambda R(\lambda) \leqslant O(n^{-2m/(2m+1)}).$$

We remark on the comparison between this rate and that obtained by Priestly and Chao (1972) and Benedetti (1977) for kernel type non-parametric regression estimates. They obtain mean square error at a point convergence rates for their estimate, call it $\hat{g}$, of the form

$$E(g(t) - \hat{g}(t))^2 = O(n^{-2m/(2m+1)})$$

under the assumption that $g^{(m)}(\cdot)$ is well defined and bounded at t. Their rates and ours are not *directly* comparable since we assume $g \in W_2^{(m)}$, and compute an estimate of *integrated* mean square error. However, as in the case of density estimation (see Wahba, 1975a, 1976) it appears that the same convergence rates under identical assumptions will obtain if the method is matched to m *and* the bandwidth parameter is chosen optimally.

ACKNOWLEDGEMENTS

We thank a referee for suggesting the identity (2.9), which considerably shortened the original proof of Theorem 1.

This research was supported by the U.S. Army under Contract No. DAAG29–77–G–0207.

REFERENCES

AKHIEZER, N. I. and GLAZMAN, I. M. (1961). *Theory of Linear Operators in Hilbert Space.* Translated from the Russian by Merlynd Nestel, p. 33. New York: Ungar.

ANSELONE, P. M. and LAURENT, P. J. (1968). A general method for the construction of interpolating or smoothing spline-functions. *Numer. Math.*, **12**, 66–82.

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**, 337–404.

BENEDETTI, JACQUELINE K. (1977). On the nonparametric estimation of regression functions. *J. R. Statist. Soc.* B, **39**, 248–253.

BLIGHT, B. J. N. and OTT, C. (1975). A Bayesian approach to model inadequacy for polynomial regression. *Biometrika*, **62**, 79–88.

CLARK, R. M. (1977). Non-parametric estimation of a smooth regression function. *J. R. Statist. Soc.* B, **39**, 107–113.

CRAVEN, P. and WAHBA, G. (1977). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. University of Wisconsin, Statistics Department Technical Report No. 445, October 1977. To appear in *Numer. Math.*

CURRY, H. B. and SCHOENBERG, I. J. (1966). On Polya frequency functions IV; the fundamental spline functions and their limits. *J. Analyse Math.*, **7**, 71–107.

DE BOOR, C. (1972). On calculating with B-splines. *J. Approximation Theory*, **6**, 50–62.

HOUSEHOLDER, A. (1964). *The Theory of Matrices in Numerical Analysis.* New York: Blaisdell.

INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES, INC., MANUAL (1977). Subroutine ICSSCU.

KIMELDORF, G. and WAHBA, G. (1970a). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495–502.

—— (1970b). Spline functions and stochastic processes. *Sankhyā*, A, **32**, 173–180.

—— (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. and Applic.*, **33**, 82–95.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with Discussion). *J. R. Statist. Soc.* B, **40**, 113–146.

O'HAGAN, A. (1978). Curve fitting and optimal design for prediction (with Discussion). *J. R. Statist. Soc.* B, **40**, 1–42.

PARZEN, E. (1961). An approach to time series analysis. *Ann. Math. Statist.*, **32**, 951–989.

—— (1970). Statistical inference on time series by RKHS methods, Proceedings of the 12th Biennial Seminar of the Canadian Mathematical Congress, pp. 1–37.

PRIESTLY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. R. Statist. Soc.* B, **34**, 385–392.

REINSCH, C. H. (1967). Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

—— (1971). Smoothing by spline functions II. *Numer. Math.*, **16**, 451–454.

ROWLANDS, R. E., LIBER, T. and DANIEL, I. M. (1974). Stress analysis of anisotropic laminated plates. *J. Amer. Inst. Aeronautics and Astronautics*, **12**, 7, 903–908.

SCHOENBERG, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.* (USA), **52**, 947–950.

SHEPP, L. A. (1966). Radon–Nikodym derivatives of Gaussian measures. *Ann. Math. Statist.*, **37**, 321–354.

SILVERMAN, B. W. (1978a). Choosing the window width when estimating a density. *Biometrika*, **65**, 1–11.

—— (1978b). Density ratios, empirical likelihood and cot death. *Appl. Statist.*, **27**, 26–33.

SMITH, A. F. M. (1973). Bayes estimates in one-way and two-way models. *Biometrika*, **60**, 319–329.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

UTRERAS, F. (1978). Sur le choix paramètre d'ajustement dans le lissage par fonctions spline. No. 296, Seminaire d'Analyse Numérique, Mathématiques Appliquées, Université Scientifique et Médicale de Grenoble.

WAHBA, G. (1973). A class of approximate solutions to linear operator equations. *J. Approxim. Theory*, **9**, 61–77.

—— (1975a). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.*, **3**, 15–29.

—— (1975b). A canonical form for the problem of estimating smooth surfaces. University of Wisconsin–Madison, Department of Statistics, Technical Report No. 420.

—— (1976). Histosplines with knots which are order statistics. *J. R. Statist. Soc.* B, **38**, 140–151.

—— (1977). Optimal smoothing of density estimates. In *Classification and Clustering* (J. Van Ryzin, ed.), pp. 423–458. New York: Academic Press.

—— (1978a). Data-based optimal smoothing of orthogonal series density estimates. University of Wisconsin–Madison, Department of Statistics, Technical Report No. 509, (submitted).

—— (1978b). Automatic smoothing of the log periodogram. University of Wisconsin–Madison, Department of Statistics. Technical Report No. 536 (submitted).

YOUNG, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika*, **64**, 309–317.