J. Austral. Math. Soc. (*Series A*) **57** (1994), 316–329

CONVERGENCE OF THE BACKFITTING ALGORITHM FOR ADDITIVE MODELS

CRAIG F. ANSLEY and ROBERT KOHN

(Received 14 June 1991; revised 21 February 1992)

Communicated by A. J. Pakes

Abstract

The backfitting algorithm is an iterative procedure for fitting additive models in which, at each step, one component is estimated keeping the other components fixed, the algorithm proceeding component by component and iterating until convergence. Convergence of the algorithm has been studied by Buja, Hastie, and Tibshirani (1989). We give a simple, but more general, geometric proof of the convergence of the backfitting algorithm when the additive components are estimated by penalized least squares. Our treatment covers spline smoothers and structural time series models, and we give a full discussion of the degenerate case. Our proof is based on Halperin's (1962) generalization of von Neumann's alternating projection theorem.

1991 *Mathematics subject classification* (*Amer. Math. Soc.*): 62G05, 62J02, 65D07, 65D10.
Keywords and phrases: Additive model, alternating projection, concurvity, nonparametric regression, penalized least squares, partial spline, trend, seasonal.

1. Introduction

We consider a model where the observations are a sum of m unknown functions plus noise and we wish to estimate the unknown functions by penalized least squares. Thus we have

$$y(i) = \sum_{j=1}^{m} f_i(x_{ij}) + e(i) \tag{1.1}$$

with the x_{ij}, $i = 1, \dots, n$, $j = 1, \dots, m$ being the design points and $e(i)$ the noise. Although it is usually very expensive computationally to simultaneously estimate $f_1, \dots, f_m$, if we fix any $m - 1$ of the components, then an $O(n)$ algorithm is generally available for estimating the remaining component. This suggests the

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

following iterative scheme for solving the penalized least squares problem. Starting with some initial estimates of $f_1, \dots, f_m$ (possibly zero), estimate f_1 by penalized least squares holding $f_2, \dots, f_m$ fixed at their current estimates, then estimate f_2 holding $f_1, f_3, \dots, f_m$ fixed at their current estimates, and continue this process until the estimates converge. This iterative method for estimating unknown components of an additive model is called backfitting and was used by Friedman and Stuetzle (1981) for projection pursuit regression, and by Wecker and Ansley (1982) and Buja, Hastie and Tibshirani (1989) for fitting additive spline models. Buja *et al.* (1989) analyzed the backfitting algorithm for a class of important problems and showed that it converged both for the regular case where there is only one solution and also for the degenerate case where there are multiple solutions.

This paper provides a geometric approach to the backfitting method for a general class of penalized least squares problems by showing that the solution to a penalized least squares problem is a projection in an appropriate space S say, and that backfitting corresponds to successive projections onto subspaces $M_1, \dots, M_m$ whose intersection is S. Therefore, backfitting corresponds to the alternating projection method proposed by von Neumann (1950) and extended by Halperin (1962). We can therefore apply von Neumann's convergence result on alternating projections to deduce the convergence of the backfitting method to the solution of the penalized least squares problem. von Neumann's alternating projection method, generalized by Halperin (1962), can be described as follows.

THEOREM 1.1. (Halperin, 1962) *Suppose that H is a Hilbert space, S is a subspace of H and P is the projection onto S. Let $M_1, \dots, M_m$ be subspaces of H so that $S = M_1 \cap \cdots \cap M_m$, and let T_j be the projection onto M_j, $j = 1, \dots, m$. Form $T = T_m T_{m-1} \cdots T_1$. Then T^N converges strongly to P as $N \to \infty$; that is $T^N(f) \to P(f)$ as $N \to \infty$ for any $f \in H$.*

The alternating projection method is useful when projection onto the space S is difficult, but projecting onto the subspaces M_j is relatively easy, so that the projection P can be obtained by sequentially applying the projections T_j. This is true, in particular, for the additive model (1.1).

The key to our approach is to treat the residuals as an extra component, thus reducing the penalized least squares problem to an interpolation problem whose solution is a projection in an appropriately defined Hilbert space. Our treatment is more general than that in Buja *et al.* (1989) as we prove convergence of the backfitting algorithm not only at the design points x_{ij} but also at all values of the arguments of the functions. In addition we allow the evaluation functionals $f_j(x_{ij})$ to be replaced by general linear functionals. This generalization is of more than just theoretical interest. Two examples requiring this extra generality are cubic spline smoothing with a periodic component which is discussed in Example 2.2 and the estimation of an additive model

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

with trend and seasonal components discussed in Example 3.1.

The main feature of our approach, however, is that it provides a geometric solution to penalized least squares making the existence, uniqueness and convergence results transparent. To simplify our discussion, we first give results for the non-degenerate case where there is a unique solution for each of the components. In Section 3 we extend the results to the degenerate case, where there may be multiple solutions. Buja *et al.* (1989) refer to this as the problem of concurvity. Related methods for the block iterative solution of equations are given by Kaczmarz (1937) and Elfving (1980).

2. Nondegenerate Case

Consider the following mathematical structure. Suppose that $F_1, \dots, F_m$ are linear spaces and $\langle\cdot,\cdot\rangle_{0j}$ is a semi-inner product on F_j, $j = 1, \dots, m$, with corresponding semi-norm $\|\cdot\|_{0j}$. By a semi-inner product we mean that $\langle\cdot,\cdot\rangle_{0j}$ is an inner product except that $\|f_j\|_{0j}$ can be zero with $f_j \neq 0$. For $i = 1, \dots, n$, let γ_{ij} be a linear functional in the space F_j, $j = 1, \dots, m$. Define F as the product space $F = F_1 \otimes \cdots \otimes F_m$, with typical element $f = (f_1, \dots, f_m)$ and for $i = 1, \dots, n$ let $\gamma_i(f) = \gamma_{i1}(f_1) + \cdots + \gamma_{im}(f_m)$. We will assume that

ASSUMPTION 2.1. For $f \in F$, if $\|f_j\|_{0j} = 0$ for $j = 1, \dots, m$ and $\gamma_i(f) = 0$ for $i = 1, \dots, n$, then $f = 0$.

Assumption 2.1 implies that

$$(2.1) \qquad \langle f_j, g_j\rangle_j = \langle f_j, g_j\rangle_{0j} + \sum_{i=1}^{n} \gamma_{ij}(f_j)\gamma_{ij}(g_j), \qquad f_j, g_j \in F_j$$

is a proper inner product for F_j, $(j = 1, \dots, m)$ and

$$(2.2) \qquad \langle f, g\rangle_F = \sum_{j=1}^{m} \langle f_j, g_j\rangle_{0j} + \sum_{i=1}^{n} \gamma_i(f)\gamma_i(g), \qquad f, g \in F$$

is a proper inner product for F. Let $\|f_j\|_j = \langle f_j, f_j\rangle_j^{1/2}$ and $\|f\|_F = \langle f, f\rangle_F^{1/2}$ be the corresponding norms. We assume further that

ASSUMPTION 2.2. F_j is a Hilbert space, that is, complete, under the inner product (2.1), $(j = 1, \dots, m)$ and F is a Hilbert space with inner product (2.2).

We observe

$$(2.3) \qquad y(i) = \sum_{j=1}^{m} \gamma_{ij}(f_j) + e(i) = \gamma_i(f) + e(i) \qquad (i = 1, \dots, n)$$

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

where $f_j \in F_j$ and $e(i)$ is the unobserved noise. We propose to estimate the unobserved function $f_1, \dots, f_m$ by the penalized least squares criterion

$$\text{(2.4)} \qquad \operatorname*{minimum}_{f \in F} \sum_{i=1}^{n} \{y(i) - \gamma_i(f)\}^2 + \sum_{j=1}^{m} \|f_j\|_{0j}^2.$$

EXAMPLE 2.1. Additive spline model. We consider the additive model (1.1) with the components $f_1, \dots, f_m$ defined on the interval [0, 1]. Let F_A be the space of functions on [0, 1] having square integrable second derivatives. Let X be the $n \times m$ matrix with x_{ij} in the ijth position and let $\iota = (1, \dots, 1)'$ be $n \times 1$. We assume that the $n \times (m+1)$ matrix $X_A = (X, \iota)$ has full column rank. We estimate the components $f_1, \dots, f_m$ in (1.1) by penalized least squares by minimizing

$$\text{(2.5)} \qquad \sum_{i=1}^{n} \left\{ y(i) - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2 + \sum_{j=1}^{m} \lambda_j^{-1} \int_0^1 \{f_j^{(2)}\}^2 \, dx_j$$

over $f_j \in F_A$, where the λ_j are given positive constants and $f^{(2)}(x) = d^2 f(x)/d^2 x$. The integrals in (2.5) represent roughness penalties. Let $\hat{f} = (\hat{f}_1, \dots, \hat{f}_m)$ minimize (2.5). Then each $\hat{f}_j$ is a cubic smoothing spline, that is $\hat{f}_j$ is a piecewise cubic. Spline smoothing is a popular statistical method of estimating a function when it is observed with noise; see Buja *et al.* (1989) for a discussion. Application of the backfitting algorithm is useful here because one dimensional smoothing is computationally straightforward, see for example Hutchinson and de Hoog (1985), whereas simultaneously smoothing all m functions is far more difficult. To express the minimization of (2.5) geometrically define

$$\langle g, h \rangle_{0A} = \int_0^1 g^{(2)}(x) h^{(2)}(x) \, dx.$$

This makes $\langle g, h \rangle_{0A}$ a semi-inner product with corresponding semi-norm $\|h\|_{0A} = \langle h, h \rangle_{0A}^{1/2}$. If $\|h\|_{0A} = 0$ then h is linear. For $j = 1, \dots, m$, take $F_j = F_A$ and define the semi-inner products $\langle f_j, g_j \rangle_j = \lambda^{-1} \langle f_j, g_j \rangle_{0A}$ and the linear functionals $\gamma_{ij}(f_j) = f_j(x_{ij})$, $i = 1, \dots, n$. Then (2.1) is a proper inner product for each F_j. Without further assumptions, however, (2.2) is only a semi-inner product for F as $\|f\|_F$ can be zero with f being nonzero. To see this note that if $\|f\|_F = 0$, then each f_j, $j = 1, \dots, m$, is a linear function which we can write as $f_j(x) = \alpha_j + \beta_j x$, with α_j and β_j constants. Because $\gamma_i(f) = \sum_j \gamma_{ij}(f_j) = 0$ we have that $\sum_j (\alpha_j + \beta_j x_{ij}) = 0$ for $i = 1, \dots, n$. By assumption the matrix X_A is of full column rank so that $\beta_1, \dots, \beta_m$ are zero and $\sum_j \alpha_j = 0$. By choosing the α_j so that not all of them are zero but their sum is zero we obtain an $f \in F$ such that $\|f\|_F = 0$ but $f \neq 0$. This degeneracy, which leads to multiple minima of (2.5) is called the concurvity problem by Buja *et al.* (1989) and is fully dealt with in the next section.

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

In most problems of practical interest it is easy to define the spaces F_j so that Assumptions 2.1 and 2.2 are satisfied and hence Theorem 2.1 below holds. We now show two ways of doing so for the cubic spline smoothing problem. Let F_B be the subspace of F_A consisting of all functions h such that $h(0) = 0$, and define $F_1 = F_A$ and $F_j = F_B$ for $j = 2, \dots, m$, with semi-inner products $\langle\, ,\rangle_{0A}$ for all spaces. It can be readily checked that (2.2) is now a proper inner product. A second way of redefining the spaces F_j is to take $F_j = F_B$ for $j = 1, \dots, m$ and define F_{m+1} as the space of constant functions on $[0, 1]$ with semi-inner product identically zero. Put $F = F_1 \otimes F_2 \otimes \cdots \otimes F_m \otimes F_{m+1}$ and minimize

$$\sum_{i=1}^{n}\left\{y(i) - \sum_{j=1}^{m} f_j(x_{ij}) - \mu\right\}^2 + \sum_{j=1}^{m} \lambda_j^{-1} \|f_j\|_{0A}^2$$

over $f_1, \dots, f_m \in F_B$ and $\mu = f_{m+1}$. It can again be readily checked that

$$\|f\|_F^2 = \sum_{j=1}^{m} \|f_j\|_{0A}^2 + \sum_{i=1}^{n}\left\{\sum_{j=1}^{m} f_j(x_{ij}) + \mu\right\}^2$$

defines a proper norm.

We now express the penalized least squares problem as an interpolation problem as in Weinert, Byrd and Sidhu (1980) and hence express the solution to (2.4) as a projection. Define the product space $H = \mathbb{R}^n \otimes F$ where $\mathbb{R}^n$ is n dimensional Euclidean space. For convenience we will write $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ and denote a typical element of H by (ϵ, f). Then define the linear functionals μ_i, $i = 1, \dots, n$ on H by $\mu_i(\epsilon, f) = \epsilon(i) + \gamma_i(f)$. It is now straightforward to check that H is a Hilbert space with inner product

$$\langle(\epsilon, f), (\xi, g)\rangle_H = \langle\epsilon, \xi\rangle_0 + \sum_{j=1}^{m}\langle f_j, g_j\rangle_{0j} + \sum_{i=1}^{n} \mu_i(\epsilon, f)\mu_i(\xi, g) \tag{2.7}$$

and corresponding norm

$$\|(f, \epsilon)\|_H^2 = \|\epsilon\|_0^2 + \sum_{j=1}^{m} \|f_j\|_{0j}^2 + \sum_{i=1}^{n}\{\mu_i(\epsilon, f)\}^2 \tag{2.8}$$

where $\langle\epsilon, \xi\rangle_0 = \sum_i \epsilon_i \xi_i$. Assumption 2.1 guarantees that (2.7) is a proper inner product.

Next we show that (2.4) is equivalent to solving an interpolation problem whose solution is obtained as a projection. With the norm (2.8) the functional μ_i is bounded, and thus there exists a $\rho_i \in H$, called the representer of μ_i, such that $\mu_i(\epsilon, f) = \langle(\epsilon, f), \rho_i\rangle$, $i = 1, \dots, n$. Let S be the subspace of H generated by $\rho_1, \dots, \rho_n$, and P be the projection operator onto S. Further, let

$$U = \left\{(\epsilon, f) : \mu_i(\epsilon, f) = y(i), \quad i = 1, \dots, n\right\}$$

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

and note that U is nonempty because $(y, 0) \in U$, where $y = \{y(1), \ldots, y(n)\}'$.

LEMMA 2.1.

(i) *Solving* (2.4) *is equivalent to solving the interpolation problem*

$$\text{(2.9)} \qquad \underset{(\epsilon, f)\in H}{\text{minimize}} \|(\epsilon, f)\|_H : \ y(i) = \mu_i(\epsilon, f), \quad i = 1, \ldots, n.$$

(ii) *For any* $(\epsilon, f) \in H$, $\mu_i P(\epsilon, f) = \mu_i(\epsilon, f)$.

(iii) *For any* $(\epsilon, f) \in U$, $P(\epsilon, f)$ *minimizes* (2.9), *and so the solution to* (2.9), *and hence the penalized least squares problem* (2.4), *exists (because U is nonempty) and is unique.*

(iv) *If* $\mu_i(\epsilon, f) = \mu_i(\xi, g)$, $i = 1, \ldots, n$, *then* $P(\epsilon, f) = P(\xi, g)$.

PROOF. For $(\epsilon, f) \in U$, the final term in (2.8) is $\sum_i y(i)^2$ and hence the solutions to (2.4) and (2.9) are the same. Parts (ii) and (iii) are immediate consequences of the projection theorem (see Weinert and Sidhu, 1978), and Part (iv) follows from Part (iii).

The backfitting algorithm applied to the penalized least squares problem proceeds as follows. Assign initial estimates $f_1^{(0)}, \ldots, f_m^{(0)}$ to $f_1, \ldots, f_m$; these can be the zero functions. At a typical point in the iteration suppose that we have the estimates $f_1^{(l)}, \ldots, f_{j-1}^{(l)}, f_j^{(l-1)}, \ldots, f_m^{(l-1)}$. We obtain $f_j^{(l)}$ by minimizing (2.4) with respect to f_j keeping $f_1, \ldots, f_{j-1}, f_{j+1}, \ldots, f_m$ fixed at their current estimates. We continue the iteration until convergence. The corresponding algorithm for the interpolation problem initializes $f_1, \ldots, f_m$ as above and in addition initializes $\epsilon_i^{(0)} = y(i) - \gamma_i(f^{(0)})$, $i = 1, \ldots, n$. A typical step of the backfitting algorithm can be written as the solution to the penalized least squares problem

$$\text{(2.10)} \qquad \underset{f_j \in F_j}{\text{minimize}} \sum_{i=1}^{n} \{y(i) - \mu_i(f)\}^2 + \|f_j\|_{0j}^2$$

where f_l is fixed for $l \neq j$.

To write the solution to (2.10) as a projection, we proceed exactly as for (2.4). Define the product space $H_j = \mathbb{R}^n \otimes F_j$ with typical element (ϵ, f_j), and define the linear functionals μ_{ij} in H_j by $\mu_{ij}(\epsilon, f_j) = \epsilon_i + \gamma_{ij}(f_j)$, $i = 1, \ldots, n$. Then H_j is a Hilbert space with inner product

$$\text{(2.11)} \qquad \langle(\epsilon, f_j), (\xi, g_j)\rangle_j = \sum_{i=1}^{n} \epsilon_i \xi_i + \langle f_j, g_j\rangle_{0j} + \sum_{i=1}^{n} \mu_{ij}(\epsilon, f_j)\mu_{ij}(\xi, g_j)$$

and corresponding norm

$$\text{(2.12)} \qquad \|(\epsilon, f_j)\|_j^2 = \sum_{i=1}^{n} \epsilon_i^2 + \|f_j\|_{0j}^2 + \sum_{i=1}^{n} \{\mu_{ij}(\epsilon, f_j)\}^2.$$

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

The functionals μ_{ij} are bounded in H_j. Now let ρ_{ij} be the representer of μ_{ij} in H_j, S_j be the subspace of H_j generated by $\rho_{1j}, \dots, \rho_{nj}$, and let P_j be the projection operator onto S_j. Finally, for given $f_l \in F_l$, $l \neq j$, define the subset U_j of H_j by

$$U_j = \left\{(\epsilon, f_j) \in H_j : \mu_{ij}(\epsilon, f_j) = y(i) - \sum_{l \neq j} \gamma_{il}(f_l), \quad i = 1, \dots, n\right\}$$

and note that U_j is nonempty because $(y, 0) \in U_j$. The following lemma is just a special case of Lemma 2.1.

LEMMA 2.2.

(i) *Solving* (2.10) *is equivalent to solving the interpolation problem*

$$\underset{(\epsilon, f_j) \in H_j}{\text{minimize}} \|(\epsilon, f)\|_j^2 \ : \ \mu_{ij}(\epsilon, f_j) = y(i) - \sum_{l \neq j} \gamma_{il}(f_l), \quad i = 1, \dots, n \tag{2.13}$$

(ii) *For any* $(\epsilon, f_j) \in H_j$, $\mu_{ij} P_i(\epsilon, f_j) = \mu_{ij}(\epsilon, f_j)$.

(iii) $P_j(\epsilon, f_j)$ *minimizes* (2.13) *for any* $(\epsilon, f_j) \in U_j$, *and so the solution to* (2.13), *and hence the step of the backfitting algorithm represented by the penalized least squares problem* (2.10), *exists (because U_j is nonempty) and is unique.*

Thus we have expressed the solutions to both the penalized least squares problem (2.4) and the steps of the backfitting algorithm (2.8) as projections in Hilbert spaces, but in different Hilbert spaces. To apply Theorem 1.1, we reexpress the steps of the backfitting algorithm as projections in H. First, for $j = 1, \dots, m$, write $P_j(\epsilon, f_j) = (\tilde{\epsilon}_j, \tilde{f}_j) \in H_j$, define the operator T_j in H by

$$T_j(\epsilon, f) = \left(\tilde{\epsilon}, f_1, \dots, f_{j-1}, \tilde{f}_j, f_{j+1}, \dots, f_m\right) \tag{2.14}$$

and the subset M_j of H by

$$M_j = \left\{(\epsilon, f) \in H : (\epsilon, f_j) \in S_j \quad \text{in} \quad H_j\right\}. \tag{2.15}$$

LEMMA 2.3.

(i) *Given* f_l, $l \neq j$, *suppose that* ϵ *and* f_j *are such that* $(\epsilon, f_j) \in U_j$, *that is,* $(\epsilon, f) = (\epsilon, f_1, \dots, f_m) \in U$. *Then* $T_j(\epsilon, f)$ *is the solution to the step of the backfitting algorithm given by* (2.10), *in the sense that* $\tilde{\epsilon}$ *and* $\tilde{f}_j$ *as defined in* (2.14) *are such that* $(\tilde{\epsilon}, \tilde{f}_j) \in H_j$ *solves* (2.13) *while* f_l *is fixed for* $l \neq j$.

(ii) *The set* M_j *is a Hilbert subspace of* H *and* T_j *is the projection operator onto* M_j $(j = 1, \dots, m)$.

(iii) *The space* $S = M_1 \cap \dots \cap M_m$.

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

The proof is given in the appendix.

Using Lemma 2.3, we can describe a typical cycle of the backfitting algorithm as taking $(\epsilon^{(l-1)}, f^{(l-1)}) \in U$ as estimates of (ϵ, f) from one step and applying to it the mapping $T = T_m T_{m-1} \cdots T_1$ to obtain the next estimate $(\epsilon^{(l)}, f^{(l)}) \in U$. Because $S = M_1 \cap \cdots \cap M_m$, and T_j is the projection operator onto M_j, $j = 1, \dots, m$ we obtain from Theorem 1.1,

THEOREM 2.1.

(i) *The penalized least squares problem* (2.4) *has a unique solution.*

(ii) *The linear operator T^N converges strongly to the projection P as $N \to \infty$, that is $\|(T^N - P)(\epsilon^{(0)}, f^{(0)})\| \to 0$ as $N \to \infty$. Thus the backfitting algorithm applied to the penalized least squares problem converges to the unique optimum solution from any initial $f^{(0)}$, because by writing $\epsilon_i^{(0)} = y(i) - \mu_i(f^{(0)})$, $i = 1, \dots, n$, we have $(\epsilon^{(0)}, f^{(0)}) \in U$.*

It is straightforward to extend our results to the case where some functionals are observed without error. Thus suppose that

$$y(i) = \gamma_i(f) = \sum_j \gamma_{ij}(f_j) \qquad (i = n+1, \dots, n+r) \tag{2.16}$$

and f is estimated by minimizing (2.4) subject to (2.16). Define $\mu_i(\epsilon, f) = \gamma_i(f)$, $i = n+1, \dots, n+r$, and replace (2.8) by

$$\|(\epsilon, f)\|_H^2 = \|\epsilon\|_0^2 + \sum_{j=1}^m \|f_j\|_{0j}^2 + \sum_{i=1}^{n+r} \{\mu_i(\epsilon, f)\}^2.$$

Then Theorem 2.1 still holds. We illustrate the usefulness of this extension by two examples.

EXAMPLE 2.2. Suppose that in Example (2.1) we know that the first component f_1 is periodic so that $f_1(0) = f_1(1)$ and $f_1^{(1)}(0) = f_1^{(1)}(1)$. Define the functionals $\gamma_{n+1}(f) = f_1(0) - f_1(1)$ and $\gamma_{n+2}(f) = f_1^{(1)}(0) - f_1^{(1)}(1)$ and the extra two observations $y(n+1) = \gamma_{n+1}(f)$ and $y(n+2) = \gamma_{n+2}(f)$ with $y(n+1) = y(n+2) = 0$. Then the estimate of f_1 is a periodic cubic spline. Periodic cubic splines are discussed by Cogburn and Davis (1974) and Wahba (1980). Both these papers use periodic splines to estimate the spectral density of a stationary process.

EXAMPLE 2.3. Suppose that in Example 2.1 the curves f_1 and f_2 coincide for $x \le x^*$ and for $x > x^*$ the curve f_2 branches out from f_1. For example f_1 may be a control curve and f_2 a treatment curve with the treatment applied at time $x = x^*$ so that f_1 and f_2 are distinct for $x > x^*$. We assume that at $x = x^*$ the two curves

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

coincide so that $f_1(x^*) = f_2(x^*)$. To enforce this restriction define the functional $\gamma_{n+1}(f) = f_1(x^*) - f_2(x^*)$ and let $y(n+1) = \gamma_{n+1}(f)$ with $y(n+1) = 0$. The functions f_1 and f_2 are now estimated as above. Silverman and Wood (1987) and Kohn and Ansley (1991) discuss the estimation of such branching curves by spline smoothing and give a number of examples.

3. Degenerate Case

Although the results in Section 2 are sufficient to solve most problems of practical interest, it is sometimes computationally convenient to relax Assumption 2.1 so that (2.2) is a semi-inner product in F. In this case (2.4) no longer has a unique solution and this is called the concurvity problem by Buja *et al.* (1989). For instance, in Example 2.1 we may want to take each F_j as F_A (the space of functions with square integrable second derivatives) because the algorithm we have available carries out unconstrained cubic spline smoothing. We showed in Example 2.1 that Assumption 2.1 does not hold with this choice of F_j. We now show that as long as (2.1) is a proper inner product for each space F_j $(j = 1, \dots, m)$ then the backfitting algorithm still converges to a solution of (2.4) and if $\hat{f}$ and $\hat{g}$ are two solutions then $\|\hat{f} - \hat{g}\|_F = 0$.

We replace Assumptions 2.1 and 2.2 with

ASSUMPTION 3.1. For $f_j \in F_j$, if $\|f_j\|_{0j} = 0$ and $\gamma_{ij}(f_j) = 0$ for $i = 1, \dots, n$, then $f_j = 0$ for $j = 1, \dots, m$. This is sufficient to ensure that (2.1) is an inner product on F_j.

ASSUMPTION 3.2. F_j is a Hilbert space with inner product (2.1) $(j = 1, \dots, m)$.

Now Assumptions 3.1 and 3.2 are not sufficient to ensure that (2.2) is a proper inner product on F. In general, it is only a semi-inner product. We accommodate this problem by considering the space F^* of equivalence classes of elements of F, where we say that $f, g \in F$ are equivalent if $\|f - g\|_F = 0$. Let $[f]$ be the equivalence class generated by f. If $[f] = [g]$ then

$$\|f_j - g_j\|_{0j} = 0 \quad (j = 1, \dots, m) \qquad \text{and} \qquad \gamma_i(f) = \gamma_i(g) \quad (i = 1, \dots, n).$$

It is clear that if $\hat{f}$ is a solution to (2.4) then so are all members of its equivalence class.

If $[f] = [g]$ then $\gamma_i(f) = \gamma_i(g)$ $(i = 1, \dots, n)$ showing that the linear functionals γ_i are well defined in F^*, and F^* is a Hilbert space with inner product (2.2). We now show that the results in Section 2 apply to the elements of F^* rather than F. First we define the space $H^* = \mathbb{R}^n \otimes F^*$ whose elements are equivalence classes of

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

elements in H. We write a typical element of H^* as $(\epsilon, [f])$ so that $(\epsilon, [f]) = (\epsilon, [g])$ if $[f] = [g]$. This means that (2.7) and (2.8) are a proper inner product and norm, respectively, in H^*. Let U^* be defined with respect to H^* in the same way that U is defined with respect to H. Then Lemma 2.1 holds without modification while Lemma 2.2 concerns $H_j = \mathbb{R}^n \otimes F_j$ and is not affected. To check that Lemma 2.3 holds, we first show that the linear operator T_j in (2.14) is well defined in H^*. This is obtained from the following lemma.

LEMMA 3.1. *Suppose that* $z = (z_1, \dots, z_m) \in F$ *and* $[z] = [0]$. *Then* $P_j(0, z_j) = (0, z_j)$ *in* H_j, $j = 1, \dots, m$.

PROOF. Let $y_{ij} = \gamma_{ij}(z_j)$, $i = 1, \dots, n$, $j = 1, \dots, m$. Because $[z] = [0]$, $\|z_j\|_{0j} = 0$, and thus $(0, z_j)$ is the solution to the minimization problem

$$\underset{(\epsilon, f_j) \in H_j}{\text{minimize}} \|(\epsilon, f_j)\|_j^2 : \mu_{ij}(\epsilon, f_j) = y_{ij}, \quad i = 1, \dots, n.$$

By Lemma 2.1 the solution is unique and $P_j(0, z_j) = (0, z_j)$, $j = 1, \dots, m$.

For $z \in F$, if $[z] = [0]$, then $T_j(0, z) = (0, z)$ by Lemma 3.1. Hence if $f, g \in F$ are equivalent, then $T_j(\epsilon, f) = T_j(\epsilon, g) + (0, f - g)$ showing that $T_j(\epsilon, f)$ and $T_j(\epsilon, g)$ are equivalent. This shows that T_j is well defined in H^*. Let M_j^* be the range of the projection T_j^*. Then Lemma (2.3) holds in H^* if M_j is replaced by M_j^*.

The operator T and the projection operator P are well defined in H^*. We can deduce the following results from Theorem (2.1). If $(\epsilon, f) \in U$ then any member of the equivalence class $P(\epsilon, [f])$ minimizes (2.4) and if $(\hat{\epsilon}, \hat{f})$ and $(\hat{\epsilon}, \hat{g})$ belong to $P(\epsilon, [f])$ then $[\hat{f} - \hat{g}] = [0]$. The iterates $T^N(\epsilon, [f])$ converge to $P(\epsilon, [f])$ so that $\|T^N(\epsilon, f) - (\hat{\epsilon}, \hat{f})\|_H \to 0$ as $N \to \infty$, where $(\hat{\epsilon}, \hat{f})$ is a solution to (2.4). This means that $\|T^N(\epsilon, f)\|_H$ is nonincreasing and tends to $\|(\hat{\epsilon}, \hat{f})\|_H$. Now, for any $N \geq 0$ either $T^{N+1}(\epsilon, f) = T^N(\epsilon, f)$ in which case we have converged, or $\|T_j T^N(\epsilon, f)\|_H < \|T^N(\epsilon, f)\|_H$ for some j in which case $\|T^{N+1}(\epsilon, f)\|_H < \|T^N(\epsilon, f)\|_H$. This shows that $T^N(\epsilon, f)$ converges to a unique element of F. Finally, if $z \in F$ and $[z] = 0$ then $T(0, z) = (0, z)$ so that if $[f] = [g]$ then $T^N(\epsilon, f) = T^N(\epsilon, g) + (0, f - g)$ showing that the separation between f and g is preserved by T^N. This discussion is summarized in the following theorem.

THEOREM 3.1. *Suppose Assumptions* 3.1 *and* 3.2 *hold. Then*

(i) *the penalized least squares problem* (2.4) *has a solution that is unique up to equivalence in* F^*, *that is, up to a displacement* $z = (z_1, \dots, z_m)$ *such that* $\|z_j\|_{0j} = 0$, $j = 1, \dots, m$ *and* $\gamma_i(z) = 0$, $i = 1, \dots, n$.

(ii) *The linear operator* T^N *in* H^* *converges strongly to the projection* P *as* $N \to \infty$, *that is* $\|(T^N - P)(\epsilon, [f])\| \to 0$ *as* $N \to \infty$ *for* $(\epsilon, f) \in U \subset H$.

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

(iii) *The backfitting algorithm converges to a unique element of F which is an optimum solution from any initial value $f \in F$. Equivalent solutions can be found from any optimum solution by displacements as in* (i).

The following corollary can be immediately deduced from Theorem 3.1 and is Theorem 9 of Buja *et al.* (1989).

COROLLARY 3.1. *Suppose that the functionals μ_{ij} are the evaluation functionals $\mu_{ij}(f_j) = f_j(x_{ij})$, $j = 1, \dots, m$ and $i = 1, \dots, n$. Let $f_j = \{f_j(x_{1j}), \dots, f_j(x_{nj})\}'$, and let $\|f_j\|_{0j} = f_j' A_j f_j$ where A_j is a positive-semidefinite matrix, $j = 1, \dots, m$. Assume that Assumption* 3.1 *holds. Then Theorem* 3.1 *holds.*

EXAMPLE 2.2. (continued). Consider again the additive model (1.1) estimated by cubic splines with no restrictions on initial conditions so that $F_j = F_A$ for all j. To satisfy Assumptions 3.1 and 3.2, we require only that the vector $X_j = (x_{1j}, \dots, x_{nj})'$ is linearly independent of the vector $\iota = (1, \dots, 1)'$ for each $j = 1, \dots, m$, so that each step of each iteration of the backfitting procedure has a unique solution. Then, from any set of starting values, the backfitting algorithm converges to a unique solution $\hat{f} = (\hat{f}_1, \dots, \hat{f}_m)'$ in F, with each $\hat{f}_j$ a cubic spline. From Example 2.1 and the discussion above, there are multiple solutions equivalent to f and given by $\hat{f} + z$ with $[z] = [0]$. Let $z = (z_1, \dots, z_m) \in F$. We showed in Section 2 that if $[z] = [0]$ then $z_j(x) = \alpha_j + \beta_j x_j$ with the constants α_j and β_j constrained by $\sum_j (\alpha_j + \beta_j x_j) = 0$ for $j = 1, \dots, m$. In particular $[z] = [0]$ if we take all the β_j to be zero and $\sum_j \alpha_j = 0$ without all the α_js being zero.

Our results also show that we can introduce redundant components into the model and still have convergence. For example, suppose that $x_{ij} = a + bx_{ij'}$, $j \neq j'$ and $i = 1, \dots, n$. Although one of f_j and $f_{j'}$ is redundant because the knots for the jth component are a linear combination of those for the j'th component, the backfitting algorithm will still converge to the estimates $\hat{f}_j$ and $\hat{f}_{j'}$ of f_j and $f_{j'}$ determined by the initial conditions. It is sometimes faster to apply the backfitting algorithm to subsets of components rather than individual components in which case the functionals γ_{ij} will not be evaluation functionals. An important application of this observation is smoothing with time series structural component models.

EXAMPLE 3.1. Time series structural component model. Suppose that we have quarterly data collected in time order so that

$$y(i) = t(i) + s(i) + f_2(x_{i2}) + e(i), \qquad i = 1, \dots, n$$

where $t(i)$ is the trend, $s(i)$ is the seasonal, and $f_2(x_{i2})$ is a function of a regressor of

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

interest. We define the semi-norms

$$\|t\|_{0t}^2 = \sum_{i=1}^{n}\{t(i) - t(i-1)\}^2, \qquad \|s\|_{0s}^2 = \sum_{i=1}^{n}\{s(i) - s(i-4)\}^2.$$

Following Kitagawa and Gersch (1984) we estimate t, s and f_2 by minimizing

$$\sum_{i=1}^{n}\{y(i) - t(i) - s(i) - f_2(x_{i2})\}^2 + \|t\|_{0t}^2 + \|s\|_{0s}^2 + \|f_2\|_{0A}^2 \tag{3.1}$$

over $t = \{t(0), \dots, t(n)\} \in \mathbb{R}^{n+1}$, $s = \{s(-3), \dots, s(n)\}' \in \mathbb{R}^{n+4}$, and $f_2 \in F_A$. Although we can minimize (3.1) by applying the backfitting algorithm to the individual components t, s, and f_2, it is computationally more efficient to form the composite component $f_1(i) = \{t(i), s(i)\}$, define the functionals $\gamma_{i1}(f_1) = t(i) + s(i)$ and $\gamma_{i2}(f_2) = f_2(x_{i2})$, $i = 1, \dots, n$ and apply the backfitting algorithm to the two component model $y(i) = \gamma_{1i}(f_1) + f_2(x_{i2}) + e(i)$. Thus t and s are estimated simultaneously as in Kitagawa and Gersch (1984). We note that the functionals γ_{i1} are not evaluation functionals.

Appendix: Proofs

We now prove Lemma 2.3 and Theorem 1.1. Halperin (1962) proves a stronger result than Theorem 1.1 but if we look carefully at his argument we obtain a very simple proof of Theorem 1.1 which we present for completeness.

PROOF OF LEMMA 2.3. Part (i) follows immediately from Lemma 2.2 and the definition of T_j. For Part (ii), note first that because P_j is a projection operator in H_j, T_j is a linear operator in H and $M_j = \{(\epsilon, f) \in H : T_j(\epsilon, f) = (\epsilon, f)\}$, so that M_j is a Hilbert subspace of H. Now by (2.14), for any $(\epsilon, f) \in H$, $T_j(\epsilon, f) \in M_j$ and $(\epsilon, f) - T_j(\epsilon, f) = (\epsilon - \tilde{\epsilon}, 0, \dots, f_j - \tilde{f}_j, 0, \dots, 0)$. Hence for any $(\xi, g) \in M_j$,

$$\langle(\epsilon, f) - T_j(\epsilon, f), (\xi, g)\rangle = \langle(\epsilon, f_j) - P_j(\epsilon, f_j), (\xi, g_j)\rangle_j = 0$$

so that $(\epsilon, f) - T_j(\epsilon, f) \in M_j^{\perp}$, the orthogonal complement of M_j, and T_j is the projection operator onto M_j.

To show that $S = M_1 \cap \dots \cap M_m$, note first that by (2.14) and Lemma 2.2, for any $(\epsilon, f) \in H$, $\mu_i T_j(\epsilon, f) = \mu_i(\epsilon, f)$, $i = 1, \dots, n$, and hence by Lemma 2.1 $PT_j(\epsilon, f) = P(\epsilon, f)$. Thus for $(\epsilon, f) \in M_j^{\perp}$, $P(\epsilon, f) = PT_j(\epsilon, f) = 0$, so that $(\epsilon, f) \in S^{\perp}$, and $M_j^{\perp} \subset S^{\perp}$, $j = 1, \dots, m$. Hence $S \subset M_1 \cap \dots \cap M_m$.

To complete the proof, we show that $S^{\perp} \subset M_1^{\perp} + \dots + M_m^{\perp}$. The required result follows because $M_1^{\perp} + \dots + M_m^{\perp} \subset (M_1 \cap \dots \cap M_m)^{\perp}$. Take $(\epsilon, f) \in S^{\perp}$. Then

$$\epsilon_i + \gamma_{i1}(f_1) + \dots + \gamma_{im}(f_m) = 0 \quad (i = 1, \dots, n). \tag{A.1}$$

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

For $j = 1, \dots, m$, define the element $(\epsilon^{(j)}, f^{(j)}) \in H$ such that $f_j^{(j)} = f_j$, $f_l^{(j)} = 0$ $(l \neq j)$ and $\epsilon_i^{(j)} = -\gamma_{ij}(f_j)$, $j = 1, \dots, n$, so that $(\epsilon^{(j)}, f_j^{(j)}) \in S_j^{\perp}$ in H_j and hence $(\epsilon^{(j)}, f^{(j)}) \in M_j^{\perp}$ in H. By (A.1), $(\epsilon, f) = \sum_j (\epsilon^{(j)}, f^{(j)})$, so that $(\epsilon, f) \in M_1^{\perp} + \dots + M_m^{\perp}$.

PROOF OF THEOREM 1.1. The proof of Theorem 1.1 depends on the following three results.

(i) $S = \{x : Tx = x\}$ and for any $x \in H$, either $Tx = x$ or $\|Tx\| < \|x\|$.
(ii) If $y \in \overline{R(I - T)}$, where $R(I - T)$ is the range of $I - T$, then $T^j y \to 0$ as $j \to \infty$.
(iii) $S^{\perp} \subset \overline{R(I - T)}$.

It follows from (iii) that we can write any $h \in H$ as $h = x + y$ with $x \in S$ and $y \in \overline{R(I - T)}$ so that from (i) and (ii), $T^N h = x + T^N y \to x$ as $N \to \infty$.

To show that (i) holds we note that for any $x \in H$ either $T_j x = x$ for $j = 1, \dots, m$, in which case $Tx = x$ and $x \in S$, or $\|T_j x\| < \|x\|$ for at least one j, and then $\|Tx\| < \|x\|$ and $x \notin S$. We show that (ii) holds for $y \in R(I - T)$ as the extension to its closure is straightforward. We show below that

$$\|x - Tx\|^2 \leq m(\|x\|^2 - \|Tx\|^2). \tag{A.2}$$

It follows that

$$\|T^j x - T^{j+1} x\|^2 \leq m(\|T^j x\|^2 - \|T^{j+1} x\|^2).$$

Because $\|T^j x\|$ is a nonincreasing sequence for a given x, it follows that $\|T^j x\|^2 - \|T^{j+1} x\|^2 \to 0$ as $j \to \infty$, and so $T^j (I - T)x \to 0$ implying that (ii) holds. To show that (A.2) holds, it is sufficient to consider the case $m = 2$ and write

$$x - Tx = (x - T_1 x) + (T_1 x - T_2 T_1 x)$$

so that

$$\|x - Tx\|^2 \leq (\|x - T_1 x\| + \|T_1 x - T_2 T_1 x\|)^2 \leq 2(\|x - T_1 x\|^2 + \|T_1 x - T_2 T_1 x\|^2)$$

Because T_1 and T_2 are projections, $\|x - T_1 x\|^2 = \|x\|^2 - \|T_1 X\|^2$ and $\|T_1 x - T_2 T_1 x\|^2 = \|T_1 x\|^2 - \|T_2 T_1 x\|^2$ and (A.2) follows.

We show that $R(I - T)^{\perp} \subset S$ from which (iii) follows. If $x \in R(I - T)^{\perp}$ then $\langle x, (I - T)x \rangle = 0$ so that $Tx = x$.

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press

References

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models (with discussion)', *Ann. Statist.* **17**, 453–555.

Cogburn, R. & Davis, H. T. (1974), 'Periodic spline and spectral estimation', *Ann. Statist.* **2**, 1108–1126.

Elfving, T. (1980), 'Block-iterative methods for consistent and inconsistent linear equations', *Numer. Math.* **35**, 1–12.

Friedman, J. H. & Stuetzle, W. (1981), 'Projection pursuit regression', *J. Amer. Statist. Assoc.* **76**, 817–823.

Halperin, I. (1962), 'The product of projection operators', *Acta Sci. Math.* **23**, 96–99.

Hutchinson, M. F. & de Hoog, F. R. (1985), 'Smoothing noisy data with spline functions', *Numer. Math.* **47**, 99–106.

Kaczmarz, S. (1937), 'Angenaherte auflosung von systemen linearer gleichungen', *Bull. Acad. Polon. Sci.* pp. 335–357.

Kitagawa, G. & Gersch, W. (1984), 'A smoothness priors-state space modeling of time series with trend and seasonality', *J. Amer. Statist. Assoc.* **79**, 378–389.

Kohn, R. & Ansley, C. F. (1991), 'A signal extraction approach to the estimation of treatment and control curves', *J. Amer. Statist. Assoc.* **86**, 1034–1041.

von Neumann, J. (1950), *Functional operators Vol.* II: *The Geometry of Orthogonal Spaces*, Vol. 22 of *Ann. of Math. Stud.*, Princeton University Press, Princeton.

Silverman, B. W. & Wood, J. T. (1987), 'The nonparametric estimation of branching curves', *J. Amer. Statist. Assoc.* **82**, 551–558.

Wahba, G. (1980), 'Automatic smoothing of the log periodogram', *J. Amer. Statist. Assoc.* **75**, 122–132.

Wecker, W. E. & Ansley, C. F. (1982), 'Nonparametric multiple regression by the alternating projection method', *in Proc. ASA Bus. Econ. Statist. Section*, pp. 311–316.

Weinert, H. L. & Sidhu, G. S. (1978), 'A stochastic framework for recursive computation of spline functions: Part I, Interpolating splines', *IEEE Trans. Inform. Theory* **24**, 45–50.

Weinert, H. L., Byrd, R. H. & Sidhu, G. S. (1980), 'A stochastic framework for recursive computation of spline functions: Part II, Smoothing splines', *J. Optim. Theory Appl.* **30**, 255–268.

Department of Accounting and Finance
University of Auckland
Private Bag, Auckland
New Zealand

Australian Graduate School
of Management
University of NSW
Kensington
New South Wales
Australia

https://doi.org/10.1017/S1446788700037721 Published online by Cambridge University Press