The Annals of Statistics
1991, Vol. 19, No. 4, 1855–1868

INTERACTION SPLINE MODELS AND THEIR CONVERGENCE RATES[1]

BY ZEHUA CHEN

Australian National University

We consider interaction splines which model a multivariate regression function f as a constant plus the sum of functions of one variable (main effects), plus the sum of functions of two variables (two-factor interactions), and so on. The estimation of f by the penalized least squares method and the asymptotic properties of the models are studied in this article. It is shown that, under some regularity conditions on the data points, the expected squared error averaged over the data points converges to zero at a rate of $O(N^{-2m/(2m+1)})$ as the sample size $N \to \infty$ if the smoothing parameters are appropriately chosen, where m is a measure of the assumed smoothness of f.

1. Introduction. Consider a system $(y, \mathbf{x})$ which can be described by

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \dots, N, \tag{1}$$

where f is an unknown function, the $\mathbf{x}_i$'s are d-dimensional vectors of covariates, and the ε_i's are i.i.d. noise with mean 0 and variance σ^2. The objective is to estimate f from N pairs of observations $(y_i, \mathbf{x}_i)$. If d is large, there is a major difficulty: the curse of dimensionality. Roughly speaking, the curse of dimensionality refers to the fact that in a high-dimensional space, the amount of data required to achieve a desired accuracy is impossible to obtain in practice. To bypass this difficulty, efforts have been made in the literature mainly by reducing the dimensionality of model (1). Friedman and Stuetzle (1981) proposed projection pursuit regression which essentially models f as the sum of univariate functions on one-dimensional projection spaces, that is,

$$f(\mathbf{x}) = f_1(\mathbf{a}'_1\mathbf{x}) + f_2(\mathbf{a}'_2\mathbf{x}) + \cdots + f_m(\mathbf{a}'_m\mathbf{x}).$$

Stone (1985) proposed additive model methodology which models f as an additive function of the covariates, that is,

$$f(\mathbf{x}) = f_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d),$$

where $\int f_i(x_i)\,dx_i = 0$, and the boldface letter denotes a vector while the lightface letters denote the components of the vector [see also Buja, Hastie and Tibshirani (1989)].

Received February 1989; revised August 1990.

[1] Work supported in part by the Dean's Fellowship of the University of Wisconsin-Madison and by AFOSR Grant AFOSR-87-0171 when the author was a graduate student at the University of Wisconsin-Madison.

AMS 1980 *subject classifications*. 62H12, 62G05, 62G20.

Key words and phrases. Prediction mean squared error, reproducing kernel Hilbert space, kernel matrix.

These methods are successful in a variety of problems. But projection pursuit regression lacks a clear interpretation, which is a practical drawback. Additive models are not suitable when interactions among the covariates are present.

In this article, we consider an alternative method: interaction spline models. An interaction spline models f as a constant plus the sum of functions of one variable (main effects), plus the sum of functions of two variables (two-factor interactions), and so on. For example, a function of three covariates might be modeled as

$$f(x_1, x_2, x_3) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{23}(x_2, x_3),$$

where the components integrate to 0 on their domains.

The basic idea of interaction splines appeared in Barry (1983, 1986) and Wahba (1986) who coined the name "interaction spline models." The terms "main effects" and "interactions" are borrowed from the analysis of variance. Interaction spline models may be viewed as the analysis of variance generalized to continuous functions. They are capable of overcoming the difficulty caused by the curse of dimensionality while, at the same time, being more interpretable than projection pursuit regression and more flexible than additive models. Interaction splines have a strong potential for empirical modeling of responses to economic and medical variables, and represent a major advance over the usual parametric models.

Gu, Bates, Chen and Wahba (1989) developed algorithms for the computation of interaction splines while Gu and Wahba (1991) provided an elegant algorithm for simultaneously choosing smoothing parameters. Chen (1989) has proposed a procedure for model selection.

In this article, we study the asymptotic behavior of interaction splines. Let $\{\mathbf{x}_i: i = 1, \dots, N\}$ be the data points satisfying some regularity conditions. Suppose f belongs to the space of tensor products of Sobolev spaces of order m. Define the prediction mean squared error by

$$R_N = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right)^2,$$

where $\hat{f}$ is the estimate of f. We are going to show that, under the extra assumption that $2m > d$, the expected prediction mean squared error ER_N converges to 0 at a rate of $O(N^{-2m/(2m+1)})$.

The article is organized as follows. We describe interaction spline models and their estimation in Section 2 and state the result concerning the convergence rate in Section 3. Some technical details are then given in Section 4.

2. Interaction spline models. We describe the models to be studied in this section. For an interaction spline model, we must first construct a reproducing kernel Hilbert space (r.k.h.s.) to accommodate the underlying regression function f. This is done by forming the tensor product of Sobolev spaces of real-valued univariate functions. In this article, we confine the domain of f to the unit cube $[0, 1]^d$.

The Sobolev space of real-valued univariate functions with order m and domain [0, 1], denoted by $W_2^{(m)}$, is defined by

$$W_2^{(m)} = \{f | f^{(\nu)} \text{ abs. cont.}, \nu = 0, 1, \dots, m-1;\ f^{(m)} \in L_2\}.$$

Define an inner product on $W_2^{(m)}$ by

$$\langle f, g \rangle = \sum_{\nu=0}^{m-1} (L_\nu f)(L_\nu g) + \int_0^1 f^{(m)}(u) g^{(m)}(u)\, du,$$

where $L_\nu f = \int_0^1 f^{(\nu)}(u)\, du$. The space $W_2^{(m)}$, endowed with this inner product, is a r.k.h.s. with reproducing kernel (r.k.) given by

$$R(s,t) = \sum_{\nu=0}^{m} k_\nu(s) k_\nu(t) + (-1)^{m-1} k_{2m}([s-t]),$$

where k_ν is the νth normalized Bernoulli polynomial satisfying $L_\nu k_\mu = \delta_{\nu\mu}$, $\delta_{\nu\mu}$ being the delta function, and $[t]$ is the fractional part of t. See, for example, Craven and Wahba (1979).

The tensor product of two r.k.h.s.'s E and F, denoted by $E \otimes F$, is defined by

$$E \otimes F = \left\{ f(x_1, x_2) = \sum_{i=1}^{n} \phi_i(x_1) \psi_i(x_2): \right.$$
$$\left. \phi_i \in E, \psi_i \in F, i = 1, \dots, n; n = 1, 2, \dots \right\}.$$

The corresponding inner product on $E \otimes F$ is defined by

$$(2) \qquad \langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \langle \phi_i, \tilde{\phi}_j \rangle_E \langle \psi_i, \tilde{\psi}_j \rangle_F,$$

where $f(x_1, x_2) = \sum_{i=1}^n \phi_i(x_1)\psi_i(x_2)$ and $g(x_1, x_2) = \sum_{j=1}^m \tilde{\phi}_j(x_1)\tilde{\psi}_j(x_2)$. A particular function may admit many different representations but the inner product, $\langle f, g \rangle$, is independent of the particular representation chosen. See Aronszajn (1950), page 358.

If E and F admit orthogonal decompositions given by, respectively,

$$E = E_1 \oplus E_2 \oplus \cdots \oplus E_p,$$
$$F = F_1 \oplus F_2 \oplus \cdots \oplus F_q,$$

then $E \otimes F$ admits the orthogonal decomposition given by

$$E \otimes F = \sum_{i=1}^{p} \sum_{j=1}^{q} E_i \otimes F_j.$$

The following results can be found in Aronszajn (1950): The tensor product of r.k.h.s.'s is again a r.k.h.s. and the r.k. of the product space is the product of the r.k.'s of its factors. The direct sum of orthogonal reproducing kernel

Hilbert subspaces is again a r.k.h.s. and the r.k. of the sum is the sum of the r.k.'s of its components. The above definitions and results can be extended to the tensor product of d r.k.h.s.'s with $d > 2$.

Now we consider the space $\otimes^d W_2^{(m)}$, the d-fold tensor product of $W_2^{(m)}$. Let

$$W_0 = \operatorname{span}\{1\},$$
$$W_1 = \operatorname{span}\{k_\nu : \nu = 1, \dots, m-1\}$$

and

$$W_2 = \{f\colon f^{(\nu)} \text{ abs. cont.}, L_\nu f = 0, \nu = 0, 1, \dots, m-1;\ f^{(m)} \in L^2\}.$$

These spaces are mutually orthogonal reproducing kernel Hilbert subspaces of $W_2^{(m)}$ with their r.k.'s given by, respectively,

$$Q_{W_0}(t, s) \equiv 1,$$
$$Q_{W_1}(t, s) = \sum_{\nu=1}^{m-1} k_\nu(t) k_\nu(s)$$

and

$$Q_{W_2}(t, s) = k_m(t) k_m(s) + (-1)^{m-1} k_{2m}([t-s]).$$

It can be easily checked that $W_2^{(m)}$ admits the orthogonal decomposition

$$W_2^{(m)} = W_0 \oplus W_1 \oplus W_2.$$

In what follows we attach a covariate to each of the components of $W_2^{(m)}$ in its subscript to indicate that the elements in the component are functions of the covariate. If follows from the properties of the tensor product of r.k.h.s.'s that $\otimes^d W_2^{(m)}$ is a r.k.h.s. with an orthogonal decomposition

$$\bigotimes^d W_2^{(m)} = \sum_{\nu_1=0}^{2} \cdots \sum_{\nu_d=0}^{2} W_{\nu_1, x_1} \otimes \cdots \otimes W_{\nu_d, x_d}, \tag{3}$$

and that the r.k. of $\otimes^d W_2^{(m)}$ is given by

$$\begin{aligned} Q_W(\mathbf{t}, \mathbf{s}) &= \prod_{i=1}^{d} R(t_i, s_i) \\ &= \sum_{\nu_1=0}^{2} \cdots \sum_{\nu_d=0}^{2} Q_{W_{\nu_1}}(t_1, s_1) \cdots Q_{W_{\nu_d}}(t_d, s_d). \end{aligned}$$

Noting that $W_0 \otimes W_{\nu, x_i} = W_{\nu, x_i} \otimes W_0 = W_{\nu, x_i}$ for any ν and i, we can rewrite the decomposition (3) as

$$\begin{aligned} \bigotimes^d W_2^{(m)} = W_0 &\oplus \sum_{i=1}^{d} \left\{ \sum_{\nu_i=1}^{2} W_{\nu_i, x_i} \right\} \\ &\oplus \sum_{i<j} \left\{ \sum_{\nu_i=1}^{2} \sum_{\nu_j=1}^{2} W_{\nu_i, x_i} \otimes W_{\nu_j, x_j} \right\} \\ &\oplus \cdots \oplus \left\{ \sum_{\nu_1=1}^{2} \cdots \sum_{\nu_d=1}^{2} W_{\nu_1, x_1} \otimes \cdots \otimes W_{\nu_d, x_d} \right\}. \end{aligned} \tag{4}$$

The formation (4) shows explicitly the main-effect-and-interaction structure of $\otimes^d W_2^{(m)}$. The subspaces in the braces of the first sum are the spaces of main effects and those in the braces of the second sum are the spaces of two-factor interactions, and so forth.

We illustrate how to obtain the explicit form of the inner products on the component spaces of $\otimes^d W_2^{(m)}$ through the following example. Let f and g belong to $W_{2,x_1} \otimes W_{2,x_2}$ and admit the representations $f = \sum_{i=1}^n f_i^1 f_i^2$ and $g = \sum_{j=1}^m g_j^1 g_j^2$, respectively. By definition

$$\begin{aligned}
\langle f, g\rangle &= \sum_{i=1}^{n} \sum_{j=1}^{m} \langle f_i^1, g_j^1\rangle_1 \langle f_i^2, g_j^2\rangle_2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^1 \frac{\partial^m f_i^1}{\partial x_1^m} \frac{\partial^m g_j^1}{\partial x_1^m}\, dx_1 \int_0^1 \frac{\partial^m f_i^2}{\partial x_2^m} \frac{\partial^m g_j^2}{\partial x_2^m}\, dx_2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^1 \int_0^1 \frac{\partial^{2m}\left(f_i^1 f_i^2\right)}{\partial x_1^m\, \partial x_2^m} \frac{\partial^{2m}\left(g_j^1 g_j^2\right)}{\partial x_1^m\, \partial x_2^m}\, dx_1\, dx_2 \\
&= \int_0^1 \int_0^1 \frac{\partial^{2m} f}{\partial x_1^m\, \partial x_2^m} \frac{\partial^{2m} g}{\partial x_1^m\, \partial x_2^m}\, dx_1\, dx_2.
\end{aligned}$$

The components of $\otimes^d W_2^{(m)}$ in (3) will be referred to as the fundamental subspaces of $\otimes^d W_2^{(m)}$. These fundamental subspaces are then used to construct the space of regression functions in an interaction spline model. An interaction spline model might be stated as follows:

$$\text{(5)} \qquad \begin{aligned} &y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \dots, N, \\ &f \in H, \end{aligned}$$

where H is the direct sum of some fundamental subspaces of $\otimes^d W_2^{(m)}$. For example, for an additive spline model, $H = W_0 \oplus \sum_{i=1}^d \{\sum_{\nu_i=1}^2 W_{\nu_i, x_i}\}$. A specific example is the model containing the main effects of x_1, x_2 and x_3 and the two-factor interaction between x_1 and x_2, for which

$$H = W_0 \oplus \sum_{i=1}^{3} \left\{ \sum_{\nu_i=1}^{2} W_{\nu_i, x_i} \right\} \oplus \left\{ \sum_{\nu_1=1}^{2} \sum_{\nu_2=1}^{2} W_{\nu_1, x_1} \otimes W_{\nu_2, x_2} \right\}.$$

In a smoothing spline model [see, e.g., Craven and Wahba (1979)], the underlying regression function f is estimated by minimizing over $W_2^{(m)}$ the sum of squared residuals plus the penalty functional $\lambda \int_0^1 (f^{(m)}(t))^2\, dt$ which annihilates all the polynomials up to degree $m - 1$. Analogously, we estimate f in model (5) by minimizing over H the sum of squared residuals plus a penalty functional which annihilates all the tensor products of polynomials up to degree $m - 1$. An explanation for doing so from the Taylor expansion's point of view is given by Eubank (1988), Chapter 5, for smoothing spline models. The same explanation applies to the interaction spline models. In the remainder of this article, we pull out the tensor products of polynomials up to degree $m - 1$ from the fundamental subspaces in H and treat them separately as a space H_0. For the sake of convenience, we denote, in what follows, the fundamental subspaces in H with the tensor products of polynomials up to degree $m - 1$ pulled out by generic notations H_j, $j = 1, \dots, p$, p being the total number of fundamental spaces in H.

Let

$$H = H_0 \oplus H_1 \oplus \cdots \oplus H_p. \tag{6}$$

The f in model (5) is then estimated by the solution to the problem:

$$\text{Minimize } \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{p} f_j(\mathbf{x}_i) \right)^2 + \sum_{j=1}^{p} \lambda_j \| f \|_j^2, \tag{7}$$

$$\text{subject to } f_j \in H_j,\ j = 0, 1, \dots, p,$$

where $f = \sum_{j=0}^{p} f_j$ and $\| \cdot \|_j$ is the norm $\| \cdot \|$ on H restricted to H_j. Note that $\| f \|_j^2 = \| f_j \|_j^2 = \| f_j \|^2$.

The solution for (7) exists and is uniquely given by functions of the form

$$\hat{f}_0(\mathbf{x}) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\mathbf{x}),$$

$$\hat{f}_j(\mathbf{x}) = \sum_{i=1}^{N} (c_i/\lambda_j) Q_j(\mathbf{x}_i, \mathbf{x}), \qquad j = 1, \dots, p,$$

where M is the dimension of H_0, $\phi_\nu(\mathbf{x}) = k_{\nu_1}(x_1) k_{\nu_2}(x_2) \cdots k_{\nu_d}(x_d)$ for some $\nu_1, \nu_2, \dots, \nu_d$, $\nu = 1, \dots, M$, which span H_0, and $Q_j(\mathbf{t}, \mathbf{s})$ is the r.k. of H_j.

The coefficients $\mathbf{d} = (d_1, \dots, d_M)^T$ and $\mathbf{c} = (c_1, \dots, c_N)^T$ are obtained by solving

$$(Q_\lambda + NI)\mathbf{c} + T\mathbf{d} = \mathbf{y}, \tag{8}$$

$$T'\mathbf{c} = 0, \tag{9}$$

where

$$Q_\lambda = \sum_{j=1}^{p} (1/\lambda_j) Q_j,$$

$$Q_j = \big(Q_j(\mathbf{x}_i, \mathbf{x}_k) \big)_{N \times N}$$

and

$$T = \big(\phi_\nu(\mathbf{x}_i) \big)_{N \times M}.$$

For derivation of the above results, see Chen, Gu and Wahba (1989).

To illustrate the idea, consider the example of an interaction model with $m = 2$, $d = 2$. Then

$$\begin{aligned} p &= 3,\\ H_0 &= \operatorname{span}\{1, k_1(x_1), k_1(x_2), k_1(x_1)k_1(x_2)\},\\ H_1 &= W_{2,x_1},\\ H_2 &= W_{2,x_2},\\ H_3 &= \{W_{1,x_1} \otimes W_{2,x_2}\} \oplus \{W_{2,x_1} \otimes W_{1,x_2}\} \oplus \{W_{2,x_1} \otimes W_{2,x_2}\},\\ \|f_1\|^2 &= \int_0^1 \left(\frac{\partial^2 f_1}{\partial x_1^2}\right)^2 dx_1,\\ \|f_2\|^2 &= \int_0^1 \left(\frac{\partial^2 f_2}{\partial x_2^2}\right)^2 dx_2,\\ \|f_3\|^2 &= \int_0^1 \left(\int_0^1 \frac{\partial^3 f_3}{\partial x_1 \partial x_2^2}\, dx_1\right)^2 dx_2 + \int_0^1 \left(\int_0^1 \frac{\partial^3 f_3}{\partial x_1^2 \partial x_2}\, dx_2\right)^2 dx_1\\ &\quad + \int_0^1 \int_0^1 \left(\frac{\partial^4 f_3}{\partial x_1^2 \partial x_2^2}\right)^2 dx_1\, dx_2. \end{aligned}$$

In the above example, we have $M = 4$ and

$$\begin{aligned} \phi_1(\mathbf{x}) &= 1,\\ \phi_2(\mathbf{x}) &= k_1(x_1),\\ \phi_3(\mathbf{x}) &= k_1(x_2),\\ \phi_4(\mathbf{x}) &= k_1(x_1)k_1(x_2),\\ Q_1(\mathbf{x}, \mathbf{x}') &= k_2(x_1)k_2(x_1') - k_4([x_1 - x_1']),\\ Q_2(\mathbf{x}, \mathbf{x}') &= k_2(x_2)k_2(x_2') - k_4([x_2 - x_2']) \end{aligned}$$

and

$$\begin{aligned} Q_3(\mathbf{x}, \mathbf{x}') &= k_1(x_1)k_1(x_1')Q_2(\mathbf{x}, \mathbf{x}') + k_1(x_2)k_1(x_2')Q_1(\mathbf{x}, \mathbf{x}')\\ &\quad + Q_1(\mathbf{x}, \mathbf{x}')Q_2(\mathbf{x}, \mathbf{x}'). \end{aligned}$$

Let $\hat{f} = \sum_{j=0}^{p} \hat{f}_j$, and $\hat{\mathbf{f}} = (\hat{f}(\mathbf{x}_1), \dots, \hat{f}(\mathbf{x}_N))'$. Then $\hat{\mathbf{f}}$ can be expressed as the product of a matrix $A(\lambda)$, which will be referred to as the influence matrix, and the data vector $\mathbf{y}$, that is,

$$\hat{\mathbf{f}} = A(\lambda)\mathbf{y}.$$

Let

$$\begin{aligned} A_0(\lambda) &= Q_\lambda(Q_\lambda + NI)^{-1},\\ E(\lambda) &= N(Q_\lambda + NI)^{-1}T\left(T'(Q_\lambda + NI)^{-1}T\right)^{-1}T'(Q_\lambda + NI)^{-1}. \end{aligned}$$

It can be obtained from (8) and (9) that

$$A(\lambda) = A_0(\lambda) + E(\lambda).$$

It can also be proven [cf. Craven and Wahba (1979), page 400] that

$$\operatorname{tr} A^2(\lambda) \le \operatorname{tr} A_0^2(\lambda) + 3M, \tag{10}$$

where M is the rank of T.

The tensor product space $\otimes^d W_2^{(m)}$ satisfies natural identifiability conditions. That is, any fundamental component of a function in $\otimes^d W_2^{(m)}$ is integrated to 0 with respect to any of its arguments, for example, $\int_0^1 f_i(x_i)\,dx_i = 0$, $\int_0^1 f_{ij}(x_i, x_j)\,dx_i = 0$, for any $i, j = 1, \dots, d$, and so on, which is analogous to the ANOVA.

An additive model in the class of interaction spline models is a model with the space of regression functions $H = W_{2,x_1}^{(m)} \oplus \cdots \oplus W_{2,x_d}^{(m)}$. The smoothness properties imposed on $W_2^{(m)}$ are the usual assumptions placed on the additive components in an additive model. Thus the class of interaction spline models described in this article is at least as rich a class as the additive models in the literature.

Concerning the modeling of interactions, Breiman (1989) proposed to model a two-factor interaction, say $f(x_1, x_2)$, as $\phi(x_1)\psi(x_2)$ or $\sum_{j=1}^{J} \phi_j(x_1)\psi_j(x_2)$. In the framework of interaction spline models, an interaction is modeled as the sum of products of univariate functions, which coincides with Breiman's idea in the case of two-factor interaction.

The tensor product space $\otimes^d W_2^{(m)}[0,1]$ is different from the Sobolev space $W_2^{(dm)}([0,1]^d)$. The membership of $W_2^{(dm)}([0,1]^d)$ requires all the derivatives up to order dm while the membership of $\otimes^d W_2^{(m)}[0,1]$ requires only a part of them. Neither one of the two spaces completely contains the other. The reason we choose to use the tensor product space is because of mathematical convenience.

3. Convergence rate.

We give our main results in this section. Suppose $\{\mathbf{x}_i: i = 1, \dots, N\}$ is a tensor product design given by

$$\left\{\mathbf{x}_i = (x_{i_1,1}, x_{i_2,2}, \dots, x_{i_d,d}) \middle| i_k = 1, \dots, n_k;\ k = 1, \dots, d\right\}, \qquad N = n_1 \cdots n_d,$$

where

$$x_{j,k} = j/n_k, \qquad j = 1, \dots, n_k,\ k = 1, \dots, d.$$

Let ER_N denote the expectation of the prediction mean squared error R_N with respect to $\mathbf{y}$.

THEOREM 1. *Suppose* (5), (6) *and the above assumptions on the data points hold and* $2m > d$. *Then* ER_N *tends to* 0 *at a rate of* $O(N^{-2m/(2m+1)})$ *if the smoothing parameters* λ_j*'s are chosen as* $O(N^{-2m/(2m+1)})$.

Proof. First, we can write ER_N as the sum of a bias term and a variance term:

$$\begin{aligned} ER_N &= E\frac{1}{N}\|\mathbf{f} - A(\lambda)\mathbf{y}\|^2 \\ &= \frac{1}{N}\|\mathbf{f} - A(\lambda)\mathbf{f}\|^2 + \frac{\sigma^2}{N}\operatorname{tr} A^2(\lambda) \qquad (11) \\ &= \text{bias}_N^2(\lambda) + \text{var}_N(\lambda), \end{aligned}$$

say. Let $\lambda_{\max} = \max\{\lambda_j\}$ and $\lambda_{\min} = \min\{\lambda_j\}$. We have [cf. Craven and Wahba (1979), Lemma 4.1] that

$$\text{bias}_N^2(\lambda) \le \sum_{j=1}^{p} \lambda_j \|f_j\|^2 \le \lambda_{\max} \sum_{j=1}^{p} \|f_j\|^2 = O(\lambda_{\max}). \qquad (12)$$

We prove in the next section that under the assumption of the theorem,

$$\operatorname{tr} A_0^2(\lambda) \le O(\lambda_{\min}^{-1/2m}). \qquad (13)$$

It follows from (10), (11), (12) and (13) that

$$ER_N \le O(\lambda_{\max}) + O\left(\frac{1}{N\lambda_{\min}^{1/2m}}\right) + O\left(\frac{1}{N}\right). \qquad (14)$$

We can minimize the right-hand side of (14) by either making $\lambda_{\max} = \lambda_{\min}$ in the first term or making $\lambda_{\min} = \lambda_{\max}$ in the second term, and then minimizing with respect to $\lambda_{\min}$ or $\lambda_{\max}$. In either case, we obtain that if the smoothing parameters are chosen as $N^{-2m/(2m+1)}$, then

$$ER_N \le O(N^{-2m/(2m+1)}).$$

The proof is complete. □

Let the interaction spline model with space H be referred to as of order r if H contains at least one r-factor interaction component but no higher interaction components. The definition of the order of an interaction spline model coincides with the definition of model dimensionality in Stone (1985). Notice that the rate $O(N^{-2m/(2m+1)})$ in Theorem 1 depends on neither d, the dimension of the covariates, nor r, the order of the concerned model. This seems unreasonable at first since, in general, as r becomes larger the rate of convergence for an estimator of f becomes slower. However, this will not be surprising if we notice that the smoothness requirement is implicitly changed as r is changed. For example, if $r = 1$, the membership of H requires m derivatives for each main effect component. If $r = 2$, the membership of H requires $2m$ mixed derivatives for each two-factor interaction component and m derivatives for each main effect component, and so on.

Stone (1982) has shown that if a regression function f of d covariates is p-times differentiable, then the optimal rate of convergence for an estimator of f is $O(N^{-2p/(2p+d)})$ in an L^2-norm in the absence of any special restrictive structure for f. Stone (1985) shows that under certain regularity conditions, regression spline estimates of the main effect projections of f achieve a rate of convergence of $O(N^{-2p/(2p+1)})$ in an L^2-norm, regardless of the form of f. He further presented a heuristic dimensionality reduction principle: If a r-dimensional $(r < d)$ model can be assumed for f, then the optimal rate of convergence should be $O(N^{-2p/(2p+r)})$ instead of $O(N^{-2p/(2p+d)})$. Our rate of convergence has a similarity with Stone's dimensionality reduction principle. In the case of $r = 1$, the order of derivatives required for f in an interaction spline model is m, which is roughly the same as Condition 3 in Stone (1985), that is, $p = m$. The rate $O(N^{-2m/(2m+1)})$ matches Stone's except in a different sense of convergence. Notice that in an interaction spline model of order r, the order of derivatives required for f is rm, that is, $p = rm$. In this case, the rate $O(N^{-2m/(2m+1)}) = O(N^{-2p/(2p+r)})$. When the design points get denser and denser, the criterion R_N is compatible with the L^2-norm. Therefore we might hope that the convergence rate established in Theorem 1 is optimal.

4. Technical details.

We prove in this section that if the assumption of the theorem is true, then

$$\operatorname{tr} A_0^2(\lambda) \le O\big(\lambda_{\min}^{-1/2m}\big)$$

with $\lambda_{\min} = \min\{\lambda_j\}$.

Let Q_W denote the kernel matrix corresponding to the r.k. of the tensor product space $\otimes^d W_2^{(m)}$, that is ,

$$Q_W = \big(Q_W(\mathbf{x}_i, \mathbf{x}_j)\big)_{N\times N}.$$

Let $Q_H = \Sigma_{j=1}^p Q_j$, where Q_j's are the kernel matrices corresponding to the r.k.'s of the components H_j's in model (5). Let $Q_H(\mathbf{x}, \mathbf{x}')$ denote the r.k. of $H_1 \oplus \cdots \oplus H_p$. Since $H_1 \oplus \cdots \oplus H_p$ is a closed subspace of $\otimes^d W_2^{(m)}$, $Q_W(\mathbf{x}, \mathbf{x}') - Q_H(\mathbf{x}, \mathbf{x}')$ is the r.k. of the orthogonal complement of $H_1 \oplus \cdots \oplus H_p$ in $\otimes^d W_2^{(m)}$. Thus it follows from the properties of r.k. that $Q_H \preccurlyeq Q_W$ (where the notation $B \preccurlyeq C$ means that $C - B$ is nonnegative definite). Since $Q_\lambda = \Sigma_{j=1}^p \lambda_j^{-1} Q_j$, it is obvious that $Q_\lambda \preccurlyeq \lambda_{\min}^{-1} Q_H$.

Lemma 1.

$$\operatorname{tr} A_0^2(\lambda) \le \operatorname{tr}\big[Q_W(Q_W + N\lambda_{\min} I)^{-1}\big]^2.$$

Proof. Let $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_N$ and $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_N$ be the eigenvalues of Q_λ and Q_W, respectively. Then

$$\alpha_i \le \beta_i / \lambda_{\min}, \qquad i = 1, \dots, N,$$

since $Q_\lambda \preccurlyeq \lambda_{\min}^{-1} Q_H \preccurlyeq \lambda_{\min}^{-1} Q_W$. Now

$$\begin{aligned} \operatorname{tr} A_0^2(\lambda) &= \operatorname{tr}\left[Q_\lambda(Q_\lambda + NI)^{-1}\right]^2 \\ &= \sum_{i=1}^{N} \left(\frac{\alpha_i}{\alpha_i + N}\right)^2 \\ &\le \sum_{i=1}^{N} \left(\frac{\beta_i}{\beta_i + N\lambda_{\min}}\right)^2 \\ &= \operatorname{tr}\left[Q_W(Q_W + N\lambda_{\min} I)^{-1}\right]^2. \end{aligned} \tag{15}$$

□

In the condition about the design points $\{\mathbf{x}_i: i = 1, \dots, N\}$, we assume, without loss of generality, that $n_1 = n_2 = \cdots = n_d = n$. Let $\Sigma = (R(x_{i,1}, x_{j,1}))_{n \times n}$, the marginal kernel matrix corresponding to the r.k. of $W_2^{(m)}$. Suppose the data points $\{\mathbf{x}_i: i = 1, \dots, N\}$ are permuted appropriately. Then

$$Q_W = \underbrace{\Sigma \otimes \cdots \otimes \Sigma}_{d \text{ fold}},$$

where "$\otimes$" is the Kronecker product operator. Denote by $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n$ the eigenvalues of Σ. Then the eigenvalues of Q_W are given by

$$\mu_{i_1}\mu_{i_2} \cdots \mu_{i_d}, \qquad i_k = 1, \dots, n, \ k = 1, \dots, d.$$

For the above results, see Chen (1987).

Recall that

$$R(t,s) = \sum_{\nu=0}^{m-1} k_\nu(t)k_\nu(s) + k_m(t)k_m(s) + (-1)^{m-1}k_{2m}([t-s]).$$

Let K be the $n \times n$ matrix defined by

$$K = \left(k_m(x_{i,1})k_m(x_{j,1}) + (-1)^{m-1}k_{2m}([x_{i,1} - x_{j,1}])\right)_{n\times n},$$

and let J be the $n \times n$ matrix with its (i,ν)th element being $k_\nu(x_{i,1})$, $\nu = 0, \dots, m-1$, $i = 1, \dots, n$. Then

$$\Sigma = JJ' + K.$$

Under the assumptions about the design points, the eigenvalues of K have a rate of decay ni^{-2m}, that is,

$$\tau_i \sim ni^{-2m}, \qquad i = 1, \dots, n, \tag{16}$$

where τ_i is the ith largest eigenvalue of K, and "$\sim$" is read as "has the same order as." See Utreras (1983) and Chen (1987).

The following two lemmas from Stewart (1973) establish the relationship between the eigenvalues of Σ and those of K.

LEMMA 2. *Let A be an $n \times n$ symmetric matrix and let X be an $n \times l$ matrix with orthonormal columns. Let $B = X'AX$ and $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_l$ be eigenvalues of A and B, respectively. Then*

$$\alpha_{n-l+i} \leq \beta_i \leq \alpha_i, \qquad i = 1, \dots, l.$$

LEMMA 3. *Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$ be the eigenvalues of the symmetric matrices A, B and $C = A + B$, respectively. Then*

$$\alpha_i + \beta_n \leq \gamma_i \leq \alpha_i + \beta_1, \qquad i = 1, \dots, n.$$

LEMMA 4. *Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ and $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n$ be the eigenvalues of Σ and K respectively. Then*:

(i) $\tau_i \leq \mu_i$, $i = 1, \dots, n$;
(ii) $\mu_i \leq \tau_i + nc$, $i = 1, \dots, m$, *where c is a constant*;
(iii) $\mu_i \leq \tau_{i-m}$, $i = m + 1, \dots, n$.

PROOF. Let X be an $n \times (n - m)$ matrix such that $X'X = I$ and $X'J = 0$. Then we have

$$X'\Sigma X = X'(JJ' + K)X = X'KX.$$

Note that

$$JJ' = \mathbf{1}\mathbf{1}' + \mathbf{k}_1\mathbf{k}_1' + \cdots + \mathbf{k}_{m-1}\mathbf{k}_{m-1}',$$

where $\mathbf{k}_\nu = (k_\nu(x_{1,1}), k_\nu(x_{2,1}), \dots, k_\nu(x_{n,1}))'$. Hence the largest eigenvalue of JJ' is less than or equal to $\Sigma_{\nu=0}^{m-1}\Sigma_{i=1}^{n} k_\nu^2(x_{i,1})$, which is bounded by nc for some constant c since those k_ν's are uniformly bounded on $[0, 1]$. The lemma then follows from Lemmas 2 and 3. □

It follows from Lemma 4 and (16) that

$$\mu_i \sim n, \qquad\qquad i \leq m,$$

$$\mu_i \sim n(i - m)^{-2m} \sim ni^{-2m}, \qquad i > m.$$

We now group the $\mu_{i_1} \cdots \mu_{i_d}$'s according to the number of those i_k's which are less than or equal to m. For $0 \leq l \leq d$, those $\mu_{i_1} \cdots \mu_{i_d}$'s which have exactly l subscripts less than or equal to m are of the orders

$$N(i_1 i_2 \cdots i_{d-l})^{-2m}, \qquad i_k = m + 1, \dots, n, k = 1, \dots, d - l,$$

where $N = n^d$. Among all the $\mu_{i_1} \cdots \mu_{i_d}$'s there are $m^l\binom{d}{l}$ of them that have the same order $N(i_1 i_2 \cdots i_{d-l})^{-2m}$, $i_k > m$, $k = 1, \dots, d - l$, since there are $\binom{d}{l}$ possibilities for exactly l of $i_1, i_2, \dots, i_d$ to be less than or equal to m and for each possibility the l subscripts can change from 1 to m.

Let $c(l) = m^l\binom{d}{l}$. It follows from Lemma 1 and the argument above that

$$\begin{aligned}
\operatorname{tr} A_0^2(\lambda) &\le \sum_{i_1,\dots,i_d=1}^{n}\left(\frac{\mu_{i_1}\mu_{i_2}\cdots\mu_{i_d}}{\mu_{i_1}\mu_{i_2}\cdots\mu_{i_d}+N\lambda_{\min}}\right)^2\\
&= \sum_{l=0}^{d} c(l) \sum_{i_1,\dots,i_{d-l}=m+1}^{n}\left(\frac{N(i_1 i_2\cdots i_{d-l})^{-2m}}{N(i_1 i_2\cdots i_{d-l})^{-2m}+N\lambda_{\min}}\right)^2\\
&\le \sum_{l=0}^{d} c(l) \sum_{i_1,\dots,i_{d-l}=1}^{n}\left(\frac{N(i_1 i_2\cdots i_{d-l})^{-2m}}{N(i_1 i_2\cdots i_{d-l})^{-2m}+N\lambda_{\min}}\right)^2\\
&= \sum_{l=0}^{d} c(l) \sum_{i_1,\dots,i_{d-l}=1}^{n}\left(\frac{1}{1+\lambda_{\min}(i_1 i_2\cdots i_{d-l})^{2m}}\right)^2\\
&= \sum_{l=0}^{d} c(l) I_l,
\end{aligned}$$

where

$$\begin{aligned}
I_l &= \sum_{i_1=1}^{n}\cdots\sum_{i_{d-l}=1}^{n}\left(\frac{1}{1+\lambda_{\min}(i_1 i_2\cdots i_{d-l})^{2m}}\right)^2\\
&< \sum_{i_1=1}^{\infty}\cdots\sum_{i_{d-l}=1}^{\infty}\left(\frac{1}{1+\lambda_{\min}(i_1 i_2\cdots i_{d-l})^{2m}}\right)^2\\
&< \int_0^{\infty}\cdots\int_0^{\infty}\left(\frac{1}{1+\lambda_{\min}(x_1 x_2\cdots x_{d-l})^{2m}}\right)^2 dx_1\cdots dx_{d-l}\\
&= \lambda_{\min}^{-1/2m}\int_0^{\infty}\cdots\int_0^{\infty}\left(\frac{1}{1+(x_1 x_2\cdots x_{d-l})^{2m}}\right)^2 dx_1\cdots dx_{d-l}.
\end{aligned}$$

If $2m > d$, the above integral is finite for any $0 \le l \le d$. Since, d, m, M are fixed, we have obtained that $\operatorname{tr} A_0^2(\lambda) \le O(\lambda_{\min}^{-1/2m})$.

Acknowledgments. The author wishes to express his sincere appreciation for the guidance and the encouragement of Professor Grace Wahba. He is also very grateful to the referees for their helpful suggestions and comments. He is especially indebted to one of the referees who patiently corrected numerous grammatical errors in the early versions of the article and worked far beyond the duty of a referee.

REFERENCES

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

BARRY, D. (1983). Nonparametric Bayesian regression. Ph.D. dissertation, Dept. Statistics, Yale Univ.

BARRY, D. (1986). Nonparametric Bayesian regression. *Ann. Statist.* **14** 934–953.

BREIMAN, L. (1989). Comment on "Linear smoothers and additive models" by Buja, Hastie and Tibshirani. *Ann. Statist.* **17** 510–515.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

CHEN, Z. (1987). A stepwise approach for the purely periodic interaction spline models. *Comm. Statist. Theory Methods* **16** 877–895.

CHEN, Z. (1989). Fitting multivariate regression functions by interaction spline models. Technical Report CSTR-003-89, Statistics Research Section, Australian National Univ.

CHEN, Z. GU, C. and WAHBA, G. (1989). Comment on "Linear smoothers and additive models" by Buja, Hastie and Tibshirani. *Ann. Statist.* **17** 515–522.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. North-Holland, Amsterdam.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1991). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

GU, C. and WAHBA, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12**.

STEWART, G. W. (1973). *Introduction to Matrix Computation*. Academic, New York.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

UTRERAS, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math.* **42** 107–117.

WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. J. Boardman, ed.) 75–80. Amer. Statist. Assoc., Washington, D.C.

DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE, 0511