JOURNAL OF MULTIVARIATE ANALYSIS **58**, 162–181 (1996)
ARTICLE NO. 0045

Bivariate Tensor-Product B-Splines in a Partly Linear Model

Xuming He*

University of Illinois at Urbana-Champaign, Champaign, Illinois 61820

and

Peide Shi

Peking University, Beijing, People's Republic of China

In some applications, the mean or median response is linearly related to some variables but the relation to additional variables are not easily parameterized. Partly linear models arise naturally in such circumstances. Suppose that a random sample $\{(T_i,\ X_i,\ Y_i),\ i=1, 2, ..., n\}$ is modeled by $Y_i = X_i^T\beta_0 + g_0(T_i) + error_i$, where Y_i is a real-valued response, $X_i \in R^p$ and T_i ranges over a unit square, and g_0 is an unknown function with a certain degree of smoothness. We make use of bivariate tensor-product B-splines as an approximation of the function g_0 and consider M-type regression splines by minimization of $\sum_{i=1}^n \rho(Y_i - X_i^T\beta - g_n(T_i))$ for some convex function ρ. Mean, median and quantile regressions are included in this class. We show under appropriate conditions that the parameter estimate of β achieves its information bound asymptotically and the function estimate of g_0 attains the optimal rate of convergence in mean squared error. Our asymptotic results generalize directly to higher dimensions (for the variable T) provided that the function g_0 is sufficiently smooth. Such smoothness conditions have often been assumed in the literature, but they impose practical limitations for the application of multivariate tensor product splines in function estimation. We also discuss the implementation of B-spline approximations based on commonly used knot selection criteria together with a simulation study of both mean and median regressions of partly linear models. © 1996 Academic Press, Inc.

Received October 19, 1994; revised March 1996.

AMS 1991 subject classifications: primary 62G07; secondary 62G20.

Key words and phrases: B-spline functions, rate of convergence, mean regression, median regression, M-estimator, partly linear model, regression quantile.

* Xuming He is Associate Professor, Department of Statistics, University of Illinois, Champaign, IL 61820. Peide Shi is Associate Professor, Department of Probability and Statistics, Peking University, China. The research is partially supported by National Security Agency Grant MDA904-96-1-0011, University of Illinois' Research Board and the National Natural Sciences Foundation of China.

0047-259X/96 $18.00

Copyright © 1996 by Academic Press, Inc.
All rights of reproduction in any form reserved.

1. INTRODUCTION

There are obvious reasons for the popularity of linear regression among which is simplicity in computation and interpretation. This remains the case despite the fact that theory and methods in nonparametric regression have taken much of the spotlight in the statistical literature in recent years. When multiple predictor variables are included in the regression equation, the size of a data set is often too small to justify a nonparametric regression fit with reasonable precision. But this does not mean that a linear relationship is always sufficient. In some applications, the mean or median response is linearly related to some variables but the relation to additional variables are not easily parameterized. Partly linear models become natural choices in such applications: the linear model is minimally altered to allow one or a few of the independent variables to have complicated effects. Shiller (1984) considered an earlier cost curve study in the utility industry using a partly linear model. Engle, Granger, Rice and Weiss (1986) studied the highly nonlinear relationship between temperature and electricity usage where other related factors such as income and price are parameterized linearly in the model.

Suppose that a random sample $\{(T_i, X_i, Y_i)\}_1^n$ of (T, X, Y) is modeled by a partly linear model

$$Y_i = X_i^T \beta_0 + g_0(T_i) + e_i, \qquad 1 \leqslant i \leqslant n, \tag{1.1}$$

where Y is the real-valued response variable, $X \in R^p$, T ranges over $[0, 1]^q$ and the e_i's are random errors which are assumed to be independent of $\{(T_i, X_i)\}$ and of each other. The model consists of a p-dimensional parameter β_0 and an unspecified q-variate function g_0.

The increasing recognition of partly linear models has attracted a number of authors to study the asymptotic behavior of both the parameter and function estimates. One interesting question is the effect of infinite dimensional nuisance parameters on the estimates of linear coefficients. Ritov and Bickel (1990) showed that without regularity conditions there may not exist any sequence of estimators which is $n^{-\alpha}$ consistent for β. Information bounds for more general semiparametric models were further discussed in Bickel, Klaassen, Ritov and Wellner (1993).

The partial splines approach (cf. Wahba, 1984) uses a roughness penalty in minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \beta - g(T_i))^2 + \lambda J(g)$$

where λ is a smoothing parameter and $J(g)$ is a penalty function. In the case of $q=1$, a common choice is $J(g)=\int_0^1 (g^{(m)}(t))^2\,dt$ for some m. For this type of penalty function, the solution to the minimization problem lies in a natural spline space of dimension n.

Heckman (1986) showed that if the regressors X and T are not correlated, the parameter estimate of β_0 from partial splines is consistent and asymptotically normal. In a thought-provoking paper, however, Rice (1986) found that in general the asymptotic bias of the parameter estimates can dominate the variance unless the nonparametric function is undersmoothed. Thus routine use of cross-validation procedures for smoothing parameter selection becomes questionable if the parametric components are of interest. Quickly coming to the rescue are Speckman (1988) and Chen (1988) among others. Speckman (1988) developed an alternative method based on kernel smoothing and showed that optimal rates of convergence for both the parametric and nonparametric components can be achieved in a partly linear regression. His idea is to adjust variables for the covariates and then regress residuals on residuals. Chen (1988) established the same for piecewise polynomials. Chen and Shiau (1991) explored a two-stage spline smoothing method. Also see Eubank and Whitney (1989) for related work on rate of convergence.

All the above-mentioned references are concerned with the conventional mean regression where the use of the least squares facilitates computation and asymptotic analysis. In the present paper, we consider M-type regression. The M-type objective function is often used for robust estimation, but our motivation for using this general formulation is to treat mean regression, median regression and other quantile regression in one setting. Tensor product B-splines are used to estimate the nonparametric part of the model. We focus on how the parametric estimates behave when such nonparametric function estimation is involved. We allow dependence between X and T, and do not assume additivity for the function g_0.

Mean regression curves provide a grand summary for the averages, just as the mean does for a univariate distribution. Mosteller and Tukey (1977) advocated the use of regression quantiles for a more complete picture of the data at hand. Some interesting examples of quantile regression in linear models and nonparametric models can be found in Efron (1991), Hendricks and Koenker (1992), and He, Ng and Portnoy (1995).

The formulation of regression quantiles by Koenker and Bassett (1978) openened a new window for regression analysis. Instead of taking the squared distance in (1.1), we minimize

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - X_i^T\beta - g_n(T_i)) \tag{1.2}$$

where g_n is a function in a m-dimensional B-spline space and

$$\rho_\alpha(s) = |s| + (2\alpha - 1)s \tag{1.3}$$

is used to obtain an estimate of the α-th conditional quantile function.

Nonparametric estimates of conditional quantiles for a univariate regressor variable have been considered by Koenker, Ng and Portnoy (1994) for smoothing splines and by He and Shi (1994) using B-spline approximations. Chaudhuri (1991) proposed a local polynomial approximation for one or higher dimensional regressors. More recently, He (1996) proposed restricted regression quantiles to avoid the problem of quantile crossing.

By the nature of (1.3), the asymptotic analysis of quantile estimates is well typified by the median regression which corresponds to $\alpha = 1/2$. One added difficulty here comes from the discontinuity of the score function $\Psi(s) = \rho'_\alpha(s)$. In the present paper, we make use of the convexity of ρ_α.

Use of B-spline approximations in statistical models is not new. Agarwal and Studden (1980) derived asymptotic bias and variance for regression splines which are linear in the observations of the response variable, including in particular the least squares estimator and a bias minimizing estimator. Least squares spline in nonparametric regression was also considered in Chen (1991). Stone (1991) provided asymptotic results for log-spline response models. A recent article of Stone (1994) provided a firmer theoretical ground for use of polynomial splines and their tensor products in a variety of multivariate function estimation problems including regression, logistic regression, Poisson regression, log-linear models and proportional hazard models. However, asymptotic results on the parametric estimates outside the least squares universe are not yet available in the literature. The maximum likelihood approach used by Stone and some other authors is limited to densities of exponential families and does not cover the L_1-type loss function used for quantile regression.

The rest of the paper is organized as follows. Selection 2 introduces the M-type regression splines, and Section 3 is devoted to their asymptotic analysis. Most of our results are stated for the case of $q = 2$ where bivariate tensor-product splines will be used as a basis for function approximations. A recent paper of Shi and Li (1994) obtained similar asymptotic results for the special case of $q = 1$. Under suitable conditions on the distribution of (X, T) and on the score function Ψ, we show that the optimal rate of convergence for the function estimate and the root-n asymptotic normality of the parameter estimate $\hat{\beta}$ are obtained when the the number of knots per dimension is chosen to be in the order of $n^{1/(2r+q)}$, where q is the dimensionality of the nonparametric function and r is its degree of smoothness. The asymptotic normality result can be used for large sample tests of

linear hypotheses. Our result for optimal mean squared error for the nonparametric regression spline requires that the degree of smoothness of g_0 increases linearly with dimension in the form of $r > q/2$. This might limit the use of tensor-product B-splines in high dimensions. This type of condition has also been used by other authors in similar settings without much discussion. We conjecture that this smoothness requirement in necessary for the spline method, and a discussion is given in Section 4. In Section 5, we consider the practical problem of knot selection using the idea of cross validation and information-based criteria. Part of our simulation study for using bivariate tensor-products of quadratic splines is also reported in Section 5. The median regression spline appears very competitive with the least squares regression at normal errors and performs far better at heavier-tailed distributions. The large-sample tests of linear hypotheses on β_0 show satisfactory performance in our simulated example.

2. M-TYPE REGRESSION SPLINES

The degree of smoothness of the true regression function determines how well the function can be approximated. It can be formulated as follows.

Let C_r be the space of all bivariate functions $h(t_1, t_2)$ on $[0, 1]^2$ such that $D^{(u_1, u_2)}h = \partial^{u_1+u_2}h/\partial^{u_1}t_1\,\partial^{u_2}t_2$ is continuous and Lipschitz of order γ:

$$|D^{(u_1, u_2)}(T) - D^{(u_1, u_2)}(S)| \leqslant W_0\,|T-S|^{\gamma}$$

for any $T, S \in [0, 1]^2$ and $u_1 + u_2 \leqslant r - \gamma$, where W_0 is a finite constant.

CONDITION 1. $g_0 \in C_r$ for some $r \geqslant 1$.

The quantity r is the order of smoothness of the true regression function g_0. We shall use the normalized B-splines of order r associated with any quasi-uniform sequence of knots on $[0, 1]$. In practice, the value of r is unknown and has to be determined by specific considerations or by examination of the data. The choices of $r = 2$ (piecewise linear), 3 (quadratic) and 4 (cubic) are common. They avoid the oscillation problem often associated with higher order polynomials but preserve certain degree of smoothness.

For simplicity, we assume without loss of generality that all B-splines here are defined on an extended partition associated with a uniform partition of M_n knots for each regressor variable. Following Schumaker (1981), we denote by $B_j(t_i)$ $(j = 1, 2, ..., N = M_n + r,\ i = 1, 2)$ the B-spline basis functions on the ith component of T. Furthermore, define

$$B_{i_1, i_2}(t) = B_{i_1}(t_1)\,B_{i_2}(t_2) \tag{2.1}$$

for any $1 \leqslant i_1, i_2 \leqslant N$. Let $\pi(t)$ be the N^2-dimensional vector consisting of all product functions of the form (2.1). The B-spline estimate of the regression function is now given by the parameter $\hat{\beta}$ and $\hat{g}_n(t) = \pi^T(t)\hat{a}$ where $(\hat{\beta}, \hat{a})$ solves

$$\sum_{i=1}^{n} \rho(Y_i - X_i^T\beta - \pi(T_i)^T a) = minimum! \tag{2.2}$$

for a properly chosen loss function ρ.

The mean and median regression correspond to $\rho(s) = s^2$ and $\rho(s) = |s|$ respectively. To limit the influence of outlying observations in Y, a ρ function with a bounded derivative is often recommended.

There are several efficient algorithms to generate the B-spline basis functions for a given set of knots in each variable t_i. A direct recursion relationship (see Schumaker, 1981, p. 120) is often used to compute $B_j(\cdot)$. The line average algorithm described in Chui (1988, pp. 8–11) for uniform knots provides an efficient approximation scheme useful for displaying curves. Another type of algorithm is based on an explicit formulation for each polynomial piece of the B-spline, see Chui and Lai (1987) for details.

3. SOME ASYMPTOTIC RESULTS

In this section, we give sufficient conditions under which the B-spline estimates of both linear and nonparametric components converge at their best possible rates as the sample size goes to infinity. We also show that the parameter estimate $\hat{\beta}$ is asymptotically normal, so it is possible to make standard large sample inference on the parameter.

To make technicalities manageable, we prove all our results for nonstochastic and uniform knots. In practice, data-based adaptive knots are far more flexible and useful. Despite this limitation, we feel that results obtained for nonadaptive splines would shed light on the large sample behavior of adaptive estimates. A recent work of Chen and Shiau (1994) considered data-driven selection of smoothing parameters for partial regression models.

The distrubutional assumptions on (X, T) are given first.

CONDITION 2. The density function w of T is bounded away from zero and infinity, that is, there exist two positive constants b_1 and b_2 such that $b_1 \leqslant w(t) \leqslant b_2$ for all $t \in [0, 1]^2$.

CONDITION 3. $E(X) = 0$, $E|X|^3 < \infty$, and $\zeta(t) = E(X \mid T = t) \in C_r$. Also, $Var(X)$ and $\Sigma = Var(X - \zeta(T))$ are finite and nonsingular. Furthermore,

there exists a positive definite matrix Σ_0 such that $Var(X-\zeta(t)) < \Sigma_0$ for all $t \in [0, 1]^2$.

In the special case where X and T are independent, $\zeta(t)$ does not depend on t and Σ is the variance-covariance matrix of X. One common motivation of the ζ function is to think of X_i being related to T_i by a regression model $E(X \mid T=t) = \zeta(t)$. This is a sufficient but not necessary condition for most of the results obtained in this section.

Let $\Psi(s) = \rho'(s)$ be the derivative function of ρ. Our conditions on Ψ are formulated as follows.

CONDITION 4. $\rho(s)$ is a convex function with $E\Psi(e_1) = 0$, $E\Psi^2(e_1) < \infty$, and for some $b_3 > 0$,

$$E\Psi(e_1 + s) = b_3 s + o(s) \qquad \text{as} \quad s \to 0. \tag{3.1}$$

In some cases, we assume a stronger version of (3.1) as follows:

$$E\Psi(e_1 + s) = b_3 s + O(s^2) \qquad \text{as} \quad s \to 0. \tag{3.2}$$

CONDITION 5. There exist positive constants b_4, b_5 and b_6 such that

$$E(\Psi(e_1 + s) - \Psi(e_1))^2 \leqslant b_4 |s|,$$
$$|\Psi(v + s) - \Psi(v)| \leqslant b_5 \qquad \text{for all} \quad |s| \leqslant b_6 \text{ and } v \in R.$$

Verification of Conditions 4 and 5 is usually straightforward. The mean regression corresponds to zero mean and finite variance for the error variable e_1. If zero is the 100α percentile of the error variable e_1, we will be estimating the αth conditional quantile. In the latter case, we assume that the density function f of the error variable satisfies

CONDITION 6. $f(0) > 0$ and $f(s)$ is Lipschitz in a neighborhood of zero.

The dimensionality of the approximating B-spline space has to increase with n for asymptotic consistency. The number of knots must be properly chosen to balance the bias and variance.

THEOREM 3.1. *Assume Conditions* 1–5. *If the number of knots* $M_n \approx n^{1/(2r+2)}$ *and* $r > 1$, *then*

$$|\hat{\beta} - \beta_0| = O_p(n^{-r/(2r+2)})$$

and

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{g}_n(T_i) - g_0(T_i))^2 = O_p(n^{-r/(r+1)}). \tag{3.3}$$

Furthermore,

$$\|D^{(u_1, u_2)}\hat{g}_n - D^{(u_1, u_2)}g_0\| = O_p(n^{-(2r-m)/(2r+2)}) \tag{3.4}$$

where $u_1 + u_2 = m < r$, *and* $\|h\|^2 = \int_{[0,1]^2} h^2(t)\, f(t)\, dt$.

The nonparametric rate of convergence in (3.4) is optimal by Stone (1982) for the estimation of the regression function as well as its derivatives. However, we show under some additional conditions that the resulting parameter estimate $\hat{\beta}$ actually converges at the parametric rate of $n^{-1/2}$ with desirable asymptotic normality.

THEOREM 3.2. *Assume Conditions* 1–6 *with* (3.1) *replaced by* (3.2). *Suppose that* $r > 3$, *or* $r > 2$ *if* Ψ *is Lipschitz. We then have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \sigma^2 \Sigma^{-1}) \tag{3.5}$$

where $\sigma^2 = E\Psi^2(e_1)/b_3^2$.

Remark. If T is univariate, then the asymptotic normality holds for any $r > 3/2$, or $r > 1$ if Ψ is Lipschitz. In general, the smoothness conditions required are $r > 3q/2$ or $r > q$.

The asymptotic efficiency of the parameter estimate is determined by Ψ in the same way as that of the M-estimate of location. According to Chen (1988), the least squares method with $\Psi(s) = s$ gives the best possible asymptotic variance in the case of normal errors. For robustness considerations, Huber's score function $\Psi_c(x) = median\{-c, x, c\}$ would yield the minimax variance estimate when the error distribution is believed to be in a small neighborhood of Normality, see Huber (1981) for details.

The result of Theorems 3.2 can be applied directly to hypothesis testing. For example, consider testing the hypothesis that

$$A^T \beta_0 = 0 \tag{3.6}$$

where A is a known $p \times d_0$ matrix with rank d_0. Then, we have the following result.

COROLLARY 3.1. *If the conditions of Theorem* 3.2 *are satisfied, then under the null hypothesis* (3.6),

$$n\sigma^{-2}(A^T\hat{\beta})^T [A^T \hat{\Sigma}_n^{-1} A]^{-1} (A^T\hat{\beta}) \xrightarrow{d} \chi^2_{d_0},$$

where $\chi^2_{d_0}$ *is the chi-square distribution of degree* d_0, $Z_n^T = (\pi(T_1), ..., \pi(T_n))$ *and* $\hat{\Sigma}_n = n^{-1}(X_1, ..., X_n)(I - Z_n(Z_n^T Z_n)^{-1} Z_n^T)(X_1, ..., X_n)^T$.

4. B-SPLINE APPROXIMATION IN HIGHER DIMENSIONS

Theorems 3.1 and 3.2 can be generalized to higher dimensions with $q \geqslant 2$ without further technical complication. If the number of knots is chosen to be in the order of $n^{1/(2r+q)}$, the optimal rate of convergence for the MSE at $n^{-2r/(2r+q)}$ will follow under the smoothness condition of $r > q/2$. Even stronger conditions are needed to prove the asymptotic normality of $\hat{\beta}$.

This type of smoothness condition has also appeared in the literature without a good explanation. For example, Chen (1991) dealt with the least squares spline estimate under the assumption of $r > q$. Our approach weakens the condition to $r > q/2$, so does Stone (1994). A recent paper of Shen and Wong (194) indicated that in the univariate spline approximation the best rate of convergence cannot be achieved in the case of $r \leqslant 1/2$ (for $q = 1$), unless some restricted optimization is carried out. The phenomenon is expected to be the same in general dimensions when $r \leqslant q/2$.

Note that such smoothness conditions are not mandatory for the estimation problem at hand. For example, the piecewise polynomial approximation considered in Chen (1988) and Chaudhuri (1991) for the mean and median regression achieves the optimal nonparametric rate of convergence without imposing conditions on r. We naturally ask whether the same holds true for B-spline approximations.

We do not yet have a definite answer for the question. Whereas the deficiency in our proofs probably accounts for part of the problem, we have some reason to think that it is necessary for the smoothness r to increase linearly with dimension q in order to obtain the desirable rate of convergence in our case. The following is one plausible explanation.

The tensor-product B-splines have the needed approximation power, that is, the function $g_0(x)$ can be uniformly approximated by $\pi(x)^T \alpha_0$ for some α_0 with an error rate of $O(M_n^{-r})$. When $M_n \approx n^{1/(2r+q)}$, the "bias" vanishes at the desirable rate. What remains is essentially a linear regression problem with increasing dimensions. Without loss of generality, we consider a purely nonparametric model with g_0 in the approximating B-spline space so that there is no bias. The linear model can then be written as $Y_i = \pi(T_i)^T \alpha_0 + e_i$ with $P_n = M_n^q \approx n^{q/(2r+q)}$ parameters.

There have been intensive studies on the linear model with increasing number of parameters, see Huber (1981) and Portnoy (1984) among others. Since $\sum_{i=1}^n \pi(T_i)\,\pi(T_i)^T$ is on the order of n/P_n, we rewrite the linear model as $Y_i = z_i^T \theta_0 + e_i$ where $z_i = \sqrt{P_n}\,\pi(T_i)$ and $\theta_0 = P_n^{-1/2}\alpha_0$. The z_i's resemble the design considered in Portnoy (1984, p. 1301). Results of Huber and of Portnoy suggest that $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{P_n/n})$, that is, $\|\hat{\alpha} - \alpha_0\| = O_p(P_n/\sqrt{n})$ on the original scale. Since $P_n \approx n^{q/(2r+q)}$, it is necessary to require $r > q/2$ in our setup for the consistency of $\hat{\alpha}$.

Note that the analysis we do assumes that all basis functions for the tensor product splines are entered into the model. It is possible that adaptively choosing a suitable subset would be helpful in high dimensions. On the other hand, if one considers additive or interaction models, the effective dimensionality is reduced to q' if the model consists of interactions of up to order q'. Chen (1991) assumed $r > q$ for interaction models with $q' = 2$, but it can be weakened to $r > 2$ if bivariate tensor products are used.

5. IMPLEMENTATION AND SIMULATION

Implementation of the B-spline approximation requires knot selection. In this section, we consider using both uniform and non-uniform knots selected by an information-based criterion or by the idea of cross-validation. We allow for different number of knots for each component of T when uniform knots are used, but the same number is used when nonuniform knots are considered in the selection.

5.1. *Selection Criteria.* To compare two sets of prospective knots, we consider some commonly-used selection criteria.

To use the idea of cross validation, we first consider the case where ρ is everywhere differentiable. Let A be the set of knots under consideration, $\pi_A(\cdot)$ be the vector of B-spline basis functions and $u_i = (X_i^T, M_n^{1/2}\pi_A(T_i)^T)^T$. Also write $\theta_A = (\hat{\beta}^T, M_n^{-1/2}\hat{\alpha}^T)^T$, $r_i = Y_i - u_i^T\theta_A$, and $U_n = (u_1, ..., u_n)^T$. Furthermore, let $w_i = \Psi(r_i)/r_i$ if $r_i \neq 0$, and $w_i = 0$ if $r_i = 0$, and $D_w = diag(w_1, w_2, ..., w_n)$.

The M-estimator of (2.2) solves $\sum_i \Psi(r_i)u_i = 0$, or $\sum_i w_i u_i r_i = 0$. Thus θ_A satisfies a weighted least squares equation

$$\left(n^{-1}\sum_{i=1}^{n} w_i u_i u_i^T\right)\theta_A = n^{-1}\sum_{i=1}^{n} w_i Y_i u_i \tag{5.1}$$

and the GCV criterion (see Wahba, 1990) can be formulated as

$$GCV(\theta_A) = \frac{n^{-1}\sum w_i r_i^2}{(1 - n^{-1}\, trace(B_n))^2} \tag{5.2}$$

where $B_n = U_n(U_n^T D_w U_n)^{-1} U_n^T D_w$. A set of knots is preferred by GCV to another if it has a smaller GCV value.

In the case of quantile regression, ρ_α is not differentiable at zero. However, the number of zero residuals is in the order of $O(M_n) = o(n)$, and (5.1) holds up to a remainder of $O(M_n/n) = o(1)$. Thus, (5.2) can be used as a good approximation.

Since each set of knots determines an approximating model, information based criteria are natural choices. Parallel to the AIC function proposed by Akaike (1973), we select knots for low values of

$$AIC(\theta_A) = \ln\left(n^{-1}\sum_{i=1}^{n}\rho(r_i)\right) + \frac{2p}{n}. \tag{5.3}$$

where p is the dimensionality of the approximating model.

Hurvich and Tsai (1989) observed that AIC often leads to overfitting. They suggested to use a bias correction to AIC. A somewhat more dramatic correction is to minimize.

$$BIC(\theta_A) = \ln\left\{\sum_{i=1}^{n}\rho(r_i)\right\} + \rho \ln n/(2n), \tag{5.4}$$

see Schwarz (1978) for motivation. Koopersberg and Stone (1992) found the $\ln n$ factor works quite well in logspline density estimation.

Our experience shows that the information-based criteria generally perform a little better than cross validation for knot selection. The difference between AIC and BIC is very marginal for sample sizes between 50 and 150. Tighter control on the number of knots is obtained by BIC when the sample size is really large.

5.2. *Distribution of Knots.* Let $s_{j,1}, s_{j,2}, ..., s_{j,M_{nj}}$ be the B-spline knots used for the jth component of T, $j = 1, 2$. Uniform knots refer to the case of $s_{j,i} = i/M_{nj}$ $(i = 1, 2, ..., M_{nj})$. In this case, we only need to determine the sizes M_{nj} in the knot selection procedure. Uniform knots are usually sufficient when the function g_0 does not exhibit dramatic changes in its derivatives.

Non-uniform knots are desirable when the function has very different local behaviors in different regions. We adopt a stepwise strategy for knot placement and deletion, in a way similar to that of Friedman and Silverman (1989). The stepwise selection procedure works as follows.

We consider a subset of locations defined by the distinct values realized by the data set. Knots will be selected from this collection in order to follow the change in the curve while containing costs. Suppose that the first k knots $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$ have been selected, the additional knot $T_{i_{k+1}}$ is so chosen that $\{T_{i_1}, ..., T_{i_k}, T_{i_{k+1}}\}$ is preferred by the selection criterion being used to any other $\{T_{i_1}, ..., T_{i_k}, T_l\}$, $1 \leqslant l \leqslant n$. This knot placement procedures continues until no additional knot in the form of T_l is preferred by the selection criterion. Then we start a stepwise deletion by removing one knot at a time. If removal of one knot would be preferred by the selection criterion, we leave this one out. If there are several such candidates, leave out the one such that the resulting set of knots is the most preferred by the

selection criterion. This process continues until the current set of knots is preferred to any one-point deletion.

The stepwise knot placement procedure may also be applied separately to each component of T by searching over a fine grid of $[0, 1]$. This is much more time consuming even in the bivariate case, and in our experience does not improve on the results in any significant way.

5.3. *Simulation.* We ran a small simulation study for quadratic B-spline estimates of partly linear models with a bivariate function g_0. Both the mean square errors of the linear parameter estimate and the nonparametric function estimate as follows:

$$LE^2 = average\{(\hat{\beta} - \beta_0)' \Sigma^{-1} (\hat{\beta} - \beta_0)\} \tag{5.5}$$

and

$$NE^2 = average\left\{ n^{-1} \sum_i (\hat{g}_n(T_i) - g_0(T_i))^2 \right\}. \tag{5.6}$$

The following knot selection scheme is used in each example unless otherwise specified. Uniform knots are first selected by AIC. In the final stage, a stepwise deletion from the equally-spaced sets is performed to reduce the dimensionality of the fit whenever possible.

The first model we consider is

$$y = x_1 + 2x_2 + g(t_1, t_2) + e \tag{5.7}$$

where (x_1, x_2) has a standard bivariate normal distribution and (t_1, t_2) are uniformly distributed on the unit square. The error variable is $N(0, 1)$. We use the test function

$$GBCW(t_1, t_2) = \frac{40\ exp(8((t_1 - 0.5)^2 + (t_2 - 0.5)^2))}{\begin{pmatrix} exp(8((t_1 - 0.2)^2 + (t_2 - 0.7)^2)) \\ + exp(8((t_1 - 0.7)^2 + (t_2 - 0.2)^2)) \end{pmatrix}}$$

which has been used by several other authors including Breiman (1991), Friedman (1991), and Gu *et al.* (1989).

We take 1000 samples of size 150 in the experiment. The estimated LE and NE are given in Table 1. Also included are the average number of knots (M_i) used for each variable t_i and the dimensionality of the fitted model (d.f.). Results for using other selection criteria or uniform knots are similar.

A noticeable point is that the number of knots used in the B-spline approximation is very small. It averages about 2.5 for each dimension. In this case, the average number of parameters used in the linearized fitting is

TABLE I

Quadratic Spline for GBCW Function; Uniform Knots

	Median Reg					Mean Reg				
Distribution	*LE*	*NE*	M1	M2	d.f.	*LE*	*NE*	M1	M2	d.f.
Normal	0.1631	0.5368	2.7440	2.3370	20.7580	0.1358	0.4671	2.9030	2.4030	21.7720
Contaminated	0.1706	0.5957	2.4700	2.2120	19.0390	0.1928	0.6408	2.6950	2.3500	20.7360

only around 20, substantially smaller than the sample size $n = 150$. This offers some advantage over the method of penalized smoothing splines where the number of parameters is around n in the calculation.

It is not surprising that the L_2 method has a slightly better precision than the L_1 method for normal errors. But the advantage quickly disappears for heavier-tailed errors such as a contaminated normal $0.95N(0, 1) + 0.05N(0, 5^2)$, also considered in Table 1.

We examined the bias and variance for each component of $\hat{\beta}$. Variance is the dominating factor in the mean squared error. For the L_2 estimate, the component-wise variance in this example is about 0.009, compared to its asymptotic value of 0.007. The L_1-estimate has a component-wise variance of 0.013, compared to its asymptotic value of 0.010. At the sample size $n = 300$, the finite sample variances are nearly the same as their asymptotic ones.

In order to have some idea of how well the B-spline approximation competes with some well-known multivariate regression estimation technique, we simulated bivariate data from the model $y = g(t) + e$ and compared the B-spline based mean (L_2) and median (L_1) regression with the Friedman's MARS algorithm and Breiman's PI algorithm. The four test functions are

(1) EXP $exp(t_1\ sin(\pi t_2))$ with $t \in [-1, 1]^2$, $\sigma = 0.5$;

(2) SIN1 $3\ sin(t_1 t_2)$ with $t \in [-2, 2]^2$, $\sigma = 1$;

(3) GBCWwith $t \in [0, 1]^2$, $\sigma = 1$.

(4) SIN2 $\sin(2\pi t_1) t_2^2$ for $-1 \leqslant t_1 \leqslant 0$, $2t_1 t_2^2$ for $0 < t_1 \leqslant 1$, with $t_2 \in [-3, 3]$, $\sigma = 1$.

The first three are taken from the "benchmark" functions of Breiman (1991), and the fourth is constructed to have lower and uneven degree of smoothness. Three error distributions are considered. They are *Normal*: $N(0, \sigma^2)$ with the same variance used in Breiman; *Mixture*: $.95N(0, \sigma^2) + 0.05N(0, 25\sigma^2)$, and *Slash*: N(0, 0.01)/U(0, 1).

Table II gives the values of NE (the first number in each case) computed from 1000 replicates of sample size $n = 100$. The associated SD's (the second number in each case) are standard deviations of individual

TABLE II

Comparison: Tensor-product B-splines, MARS, and PIMPLE

Function	Distr.	L_2 Fit	L_1 Fit	MARS	PIMPLE
	Normal	0.231, 0.017	0.262, 0.022	0.288, 0.034	0.245, 0.184
	Mixture	0.340, 0.076	0.301, 0.054	0.405, 0.079	0.856, 14.7
EXP	Slash	18.7, 7546	0.538, 6.46	31.5, 18342	45.5, 35492
	Normal	0.453, 0.079	0.528, 0.098	0.775, 0.225	0.409, 0.086
	Mixture	0.682, 0.322	0.622, 0.238	1.02, 0.475	0.693, 0.439
SIN1	Slash	51.7, 75380	0.612, 6.46	28.7, 13851	44.4, 35429
	Normal	0.544, 0.072	0.609, 0.095	0.827, 0.261	0.558, 0.090
	Mixture	0.747, 0.327	0.691, 0.233	1.07, 0.530	0.822, 0.398
GBCW	Slash	31.1, 18135	0.689, 6.52	38.1, 28283	12424, ***
	Normal	0.805, 0.472	0.999, 0.723	0.963, 0.374	0.456, 0.151
	Mixture	0.992, 0.587	1.078, 0.746	1.224, 0.650	0.854, 0.761
SIN2	Slash	98.8, 304670	1.500, 14.51	135.99, 554882	8394, ***

NE's from simulated samples. When a number is greater than 10^6, it is replaced by *** in the table. A large SD indicates that the quality of fits would vary a lot from one sample to another. It is clear from the table that the tensor-product B-spline approximation is highly competitive and generally outperforms MARS. PIMPLE is an excellent performer with normal error, but it does a poor job if the error is heavy-tailed. Professors Friedman and Brieman kindly supplied their fortran codes of MARS and PIMPLE respectively. The calculations of the B-spline based mean and median regression were carried out by GAUSS programs on an IBM 486.

One referee pointed out that MARS was developed with applications with many more predictors in mind. The comparison we made was limited to two predictors. Therefore the results presented here do not imply that the tensor-product B-spline approximations are always preferred.

Finally, we consider the performance of approximate tests based on Corollary 3.1 in finite sample problems. The simple model investigated here is

$$Y = X\beta_0 + T\exp(T/3) + e, \tag{5.8}$$

where the random error e has distribution $N(0, 1)$, $X = X_0 + T/4$, X_0 and T are independent and distributed as $N(0, 1)$ and $U(-2, 2)$ respectively. We wish to test the null hypothesis of β_0. For a fixed level of $\alpha = 0.05$, sample sizes of 50, 100, and 150 are used. The type I and type II errors of the large-sample test based on Corollary 3.1 are estimated from 4000 Monte Carlo samples, and given in Tables III (A and B). Both the L_1 and L_2 methods provide rather reliable significance levels.

TABLE III

A. Estimated Type I Errors

L_1-Estimator			L_2-Estimator		
$n=50$	$n=100$	$n=150$	$n=50$	$n=100$	$n=150$
0.0473	0.0470	0.0433	0.0560	0.0545	0.05175

B. Estimated Type II Errors

	L_1-Estimator			L_2-Estimator		
β_0	$n=50$	$n=100$	$n=150$	$n=50$	$n=100$	$n=150$
0.3	0.6328	0.3688	0.1998	0.4658	0.1693	0.0563
0.5	0.2498	0.0348	0.0033	0.0910	0.0018	0.0003

APPENDIX: PROOFS OF MAIN RESULTS

In this section, we prove the results of Theorems 3.1 and 3.2 for any $q \geqslant 2$. We assume that $s_{j,i} = i/M_n$, $M_{nj} = M_n$, and $N_j = N$, for $i = 1, ..., M_n$, $1 \leqslant j \leqslant q$.

Let

$$
\begin{aligned}
Z_n &= (\pi(T_1), ..., \pi(T_n))^T, \\
G &= Z_n (Z_n^T Z_n)^- Z_n^T, \\
\hat{\Sigma}_n &= (X_1, ..., X_n)(I - G)(X_1, ..., X_n)^T,
\end{aligned}
\tag{6.1}
$$

where $(Z_n^T Z_n)^-$ stands for Moore Inverse of $Z_n^T Z_n$. Direct verifications show that $n^{-1}\hat{\Sigma}_n \to \Sigma$ almost surely. By arguments similar to Lemma 3.1 of He and Shi (1994) and Theorem 4 of Chen (1991), we see that the eigenvalues of $n^{-1}N^q Z_n Z_n^T$ and $n^{-1}\hat{\Sigma}_n$ are bounded away from zero and infinity. Without loss of generality, we assume that both eigenvalues are greater than or equal to a constant λ for all n. The following lemma justifies the approximation power of the tensor-product B-splines. Its proof follows readily from Theorem 12.7 of Schumaker (1981).

LEMMA 6.1. *Under Condition* 1, *there exist a constant* W_3 *depending only on* m, q *and* W_0 *such that*

$$\sup_{t \in [0,1]^q} |g_0(t) - \pi(t)^T a_0| \leqslant W_3 M_n^{-r} \tag{6.2}$$

where a_0 *is a* N^q*-dimensional vector depending on* g_0.

The following lemma is needed to prove consistency of the regression splines.

LEMMA 6.2. *Assume Conditions* 1 *and* 4 *and* $N = O(n^{1/(2r+q)})$. *We have as* $n \to \infty$,

(a) $\sup_{|a|=1, |\beta|=1} P(a^T \pi(t)\, x^T \beta) N^{q/2} = O(1)$,

(b) $\lim_{n \to \infty} \sup_{|a|=1, |\beta|=1} |(P_n - P)\, a^T \pi(t)(x - \zeta(t))^T \beta \mid N^{q/2} = 0$, *a.s.*

(c) $\lim_{n \to \infty} \sup_{|a|=1, |\beta|=1} |(P_n - P)\, a^T \pi(t)\, x^T \beta \mid N^{q/2} = 0$, *a.s.*

where Ph *and* $P_n h$ *denote the expectation under the probability distribution of* (X, T) *and the empirical distribution respectively.*

The proof of Lemma 6.2 is straightforward by the standard Cauchy–Schwartz inequality, convergence of $P_n xx^T$, and boundness of the eigenvalues of $n^{-1} N^q Z_n Z_n^T$.

Proof of Theorem 3.1. By Lemma 6.1, there exists $a^*(g_0)$ such that $R_n(t) = g_0(t) - \pi(t)^T a^*(g_0) = O(M_n^{-r})$ uniformly in t. Let $H_n^2 = diag(\hat{\Sigma}_n, N^q Z_n^T Z_n)$ be a symmetric and block-diagonal matrix and

$$\hat{\theta} \equiv \begin{pmatrix} \hat{\theta}_{1n} \\ \hat{\theta}_{2n} \end{pmatrix}$$

$$= H_n \begin{pmatrix} \hat{\beta} - \beta_0 \\ (\hat{a} - a^*(g_0)) N^{-q/2} + (Z_n^T Z_n) - \sum_{k=1}^{n} \pi(T_k)\, X_k^T (\hat{\beta} - \beta_0) N^{-q/2} \end{pmatrix}.$$

Since $|\hat{\Sigma}_n^{1/2}(\hat{\beta} - \beta_0)| \leqslant |\hat{\theta}|$ and $[n^{-1} \sum_{i=1}^n (\pi(T_i)^T (\hat{a} - a^*(g_0)))^2]^{1/2} \leqslant n^{-1/2} |\hat{\theta}| + \Lambda_n^{-1/2} |\hat{\beta} - \beta_0| \sup_{|a|=1, |\beta|=1} |n^{-1} \sum_{i=1}^n a^T \pi(T_i)\, X_i^T \beta N^{q/2}|$, it suffices to show that $|\hat{\theta}| = O_p(M_n^{q/2})$. The rest of the proof follows the same arguments as those of He and Shi (1994).

To obtain asymptotic normality of $\hat{\beta}$, we shall make use of stronger asymptotic linearization results than Lemmas 3.3 and 3.4 of He and Shi (1994). The key difference is that the following lemmas give the uniform linearization over a broader parameter space for $\theta_1 \in R^p$. This is where

stronger smoothness conditions of $r > q$ and $r > 3q/2$ are used. Since the proofs follow the same line as in He and Shi (1994), they are also omitted in the present paper.

Let

$$R_{ni} = g_0(T_i) - \pi(T_i)^T a_0, \qquad z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix}, \qquad z_{2i} = H_{2n}^- \pi(T_i) N^{q/2},$$

$$z_{1i} = H_{1n}^- \left(X_i - \sum_{k=1}^{n} \pi(T_k)^T (Z_n^T Z_n)^- \pi(T_i) X_k \right), \qquad i = 1, ..., n.$$

LEMMA 6.3. *Under the conditions of Theorem* 3.2, *we have for any* $L > 0$ *and* $M > 0$,

$$\sup_{|\theta_1| \leqslant M,\, |\theta_2| \leqslant LM_n^{q/2}} \left| \sum_{i=1}^{n} [\rho(e_i - z_{1i}^T\theta_1 - z_{2i}^T\theta_2 - R_{ni}) - \rho(e_i - z_{2i}^T\theta_2 - R_{ni}) + z_{1i}^T\theta_1 \Psi(e_i) - E_e(\rho(e_i - z_{1i}^T\theta_2 - R_{ni}) - \rho(e_i - z_{2i}^T\theta_2 - R_{ni}))] \right| = o_p(1), \tag{6.3}$$

where E_e *is the conditional expectation operator given* $(T_1, X_1), ..., (T_n, X_n)$.

LEMMA 6.4. *Under the conditions of Theorem* 3.2, *we have*

$$\sum_{i=1}^{n} E_e(\rho(e_i - z_{1i}^T\theta - z_{2i}^T\theta_2 - R_{ni}) - \rho(e_i - z_{2i}^T\theta_2 - R_{ni})) = \theta_1^T\theta_1 b_3/2 + r_n(\theta_1, \theta_2),$$

where $\sup_{|\theta_1| \leqslant M,\, |\theta_2| \leqslant LM_n^{q/2}} |r_n(\theta_1, \theta_2)| = o_p(1)$;

Proof of Theorem 3.2. Let $\hat\theta_n^T = (\hat\theta_{1n}^T, \hat\theta_{2n}^T)$ as in the proof of Theorem 3.1, and $\tilde\theta_1 = b_3^{-1} \sum_{i=1}^n z_{1i} \Psi(e_i)$. By the triangle inequality and Lemmas 6.3 and 6.4, we have for any $L > 0$, and $\delta > 0$,

$$\sup_{|\theta_1 - \tilde\theta_1| = \delta,\, \theta_1 \in R^p} I(|\tilde\theta_1| \leqslant L, |\hat\theta_{2n}| \leqslant LM_n^{q/2}) \left| \sum_{i=1}^{n} (\rho(e_i - z_{1i}^T\theta_1 - z_{2i}^T\hat\theta_{2n} - R_{ni}) - \rho(e_i - z_{1i}^T\tilde\theta_1 - z_{2i}^T\hat\theta_{2n} - R_{ni})) - \frac{b_3\delta^2}{2} \right|$$

$$\leqslant 2 \sup_{|\theta_1| \leqslant L+\delta,\, |\theta_2| \leqslant LM_n^{q/2}} \left| \sum_{i=1}^{n} (\rho(e_i - z_{1i}^T\theta_1 - z_{2i}^T\theta_2 - R_{ni}) - \rho(e_i - z_{2i}^T\theta_2 - R_{ni}) + z_{1i}^T\theta_1 \Psi(e_i)) - \frac{b_3\theta_1^T\theta_1}{2} \right| = o_p(1).$$

On the other hand, $P\{|\tilde{\theta}_1| \leqslant L\}$ and $P\{|\hat{\theta}_{2n}| \leqslant LM_n^{q/2}\}$ decrease to zero as $L \to \infty$. Therefore,

$$\lim_{n \to \infty} \mathscr{P} \left\{ \sup_{|\theta_1 - \tilde{\theta}_1| = \delta,\ \theta_1 \in R^p} \left| \sum_{i=1}^{n} (\rho(e_i - z_{1i}^T \theta_1 - z_{2i}^T \hat{\theta}_{2n} - R_{ni}) \right.\right.$$
$$\left.\left. - \rho(e_i - z_{1i}^T \tilde{\theta}_1 - z_{2i}^T \hat{\theta}_{2n} - R_{ni})) - \frac{b_3 \delta^2}{2} \right| \geqslant \varepsilon \right\} = 0 \tag{6.4}$$

for any $\varepsilon > 0$. By Corollary 25 of Eggleston (1958, p. 47), (6.4) implies that

$$\lim_{n \to \infty} P \left\{ \inf_{|\theta_1 - \hat{\theta}_1| \geqslant \delta} \sum_{i=1}^{n} (\rho(e_i - z_{1i}^T \theta_1 - z_{2i}^T \hat{\theta}_{2n} - R_{ni}) \right.$$
$$\left. > \sum_{i=1}^{n} \rho(e_i - z_{1i}^T \hat{\theta}_1 - z_{2i}^T \hat{\theta}_{2n} - R_{ni}) \right\} = 1.$$

By the definition of $\hat{\theta}_n$, we have $\hat{\theta}_{1n} = \tilde{\theta}_1 + o_p(1)$. The result then follows from the central limit theorem on $\tilde{\theta}_1$.

REFERENCES

Agarwal, G. G., and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pp. 267–281.

Bickel, P. J., Klaassen, C., Ritov. Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins Series in the Mathematical Sciences. Johns Hopkins Univ. Press, Baltimore/London.

Breiman, L. (1991). The II method for estimating multivariate functions from noisy data (with discussion), *Technometrics* **33** 125–143.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation, *Ann. Statist.* **19** 760–777.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

Chen, H. (1991). Polynomial splines and nonparametric regression. *J. Nonparametric Statist.* **1** 143–156.

Chen, H., and Shiau, J. G. (1991). A two-stage spline smothing method for partially linear models. *J. Statist. Planning & Inference* **25** 187–201.

Chen, H., and Shiau, J. G. (1994). Data-driven efficient estimators for a partially linear model. *Ann. Statist.* **22** 211–237.

Chen, Z. (1991). Interaction spline models and their convergence rates, *Ann. Statist.* **19** 1855–1868.

Chui, C.K. (1988). *Multivariate Splines*, CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Chui, C. K., and Lai, M. J. (1987). Computation of box splines and B-splines on triangulations of nonuniform rectangular partitions. *Approx. Theory Appl.* **3** 37–62.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1** 93–125.

Eggleston, H. G. (1958). *Convexity*, Cambridge Tracts in Mathematics and Mathematical Physics, Vol. 47. Cambridge University Press, Cambridge, MA.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

Eubank, R. L., and Whitney, P. (1989). Convergence rates for estimation in certain partially linear models. *J. Statistical Planning and Inference* **23** 33–43.

Friedman, J. H. (1991). Multivariate additive regression splines, (with discussion). *Ann. Statist.* **19** 1–67.

Friedman, J. H., and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–21.

Gu, C., Bates, D. M., Chen, Z., and Wahba, G. (1989). The computation of generalized gross-validation functions through Householder tridiagonalization with applications to the fitting of interaction spline models, *SIAM J. Matrix Anal. Appl.* **10** 457–480.

He, X. (1996). Quantile curves without crossing, preprint.

He, X., Ng, P., and Portnoy, S. (1995). Bivariate quantile smoothing splines, preprint.

He, X., and Shi, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions, *J. Nonparametric Statist.* **3** 299–308.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. B* **48** 244–248.

Hendricks, W., and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Amer. Statistical Assn.* **87** 58–68.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Hurvich, C. M., and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307.

Koenker, R., and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

Koopersberg and Stone (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1** 301–328.

Mosteller, F., and Tukey, J. W. (1977). *Data Analysis and Regression.* Addison–Wesley, Reading, MA.

Portnoy, S. (1984). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309.

Rice, J. (1986). Convergence rates for partially splines models, *Statist. Probab. Lett.* **4** 203–208.

Ritov, Y., and Bickel, P. J. (1990). Achieving information bounds in non- and semiparametric models, *Ann. Statist.* **18** 925–938.

Roberts, A. W., and Varberg, D. E. (1973). *Convex Functions*. Academic Press, New York/London.

Schumaker, L. L. (1981). *Spline Functions*. Wiley, New York.

Schwarz, G. (1978). Estiamting the dimension of a model. *Ann. Statist.* **6** 461–464.

Shen and Wang (1994). Convergence rates for Sieve estimators. *Ann. Statist.* **22**, in print.

Shi, P., and Li, G. (1994). On the rates of convergence of minimum L_1-norm estimators in a partly linear model. *Comm. in Statist. Theory and Methods* **23** 175–196.

Shiller, R. J. (1984). Smoothness priors and nonlinear regression. *J. Amer. Statist. Assoc.* **79** 609–615.

Speckman, P. (1988). Kernel smoothing in partial linear models, *J. Roy. Statist. Soc. B* **50** 413–436.

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

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

Stone, C. (1991). Asymptotic for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184.

Wahba, G. (1984). Cross validation spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory* (H. A. David and H. T. David, Eds.), pp. 205–235. Iowa State Univ. Press, Ames, IA.

Wahba, G. (1990). *Spline Methods for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.