Convergence Rates for Parametric Components in a Partly Linear Model

Hung Chen

Annals of Statistics, Volume 16, Issue 1 (Mar., 1988), 136-146.

The Annals of Statistics
1988, Vol. 16, No. 1, 136–146

CONVERGENCE RATES FOR PARAMETRIC COMPONENTS IN A PARTLY LINEAR MODEL[1]

BY HUNG CHEN

State University of New York at Stony Brook

Consider the regression model $Y_i = X_i'\beta + g(t_i) + e_i$ for $i = 1, \ldots, n$. Here g is an unknown Hölder continuous function of known order p in R, β is a $k \times 1$ parameter vector to be estimated and e_i is an unobserved disturbance. Such a model is often encountered in situations in which there is little real knowledge about the nature of g. A piecewise polynomial g_n is proposed to approximate g. The least-squares estimator $\hat{\beta}$ is obtained based on the model $Y_i = X_i'\beta + g_n(t_i) + e_i$. It is shown that $\hat{\beta}$ can achieve the usual parametric rates $n^{-1/2}$ with the smallest possible asymptotic variance for the case that X and T are correlated.

1. Introduction. Consider the model given by

$$Y = X'\beta + g(T) + e, \tag{1}$$

where $X' = (x_1, \ldots, x_k)$ are explanatory variables that enter linearly, β is a $k \times 1$ vector of unknown parameters, T is another explanatory variable that enters in a nonlinear fashion, $g(\cdot)$ is an unknown smooth function of T in R, (X,T) and e are independent, and e is the random error with mean 0 and variance σ^2. Assumptions on the X and T will be introduced in Section 2. Suppose we are interested in estimating β only and treat $g(\cdot)$ as a nuisance parameter. In this paper, a technique of data smoothing is used to analyze the data, and it is shown that the convergence rate for the obtained estimator of β is $n^{-1/2}$ under suitable conditions on (X, T). The model defined in (1) belongs to the class of partial spline models, which has been recently proposed and applied by Engle, Granger, Rice and Weiss (1986) to study the effect of weather on electricity demand. For generalizations and related models, see also Wahba (1984). This model is much more flexible than the standard linear model since it combines both parametric and nonparametric components. It can be used in applications where one may believe, without knowing the parametric form of g, that the dependence of Y on X is linear but might suspect that the response Y is nonlinearly related to a particular independent variable T.

There has been a trend in the past few years to move away from the standard linear regression models and to model the dependence of Y on X in a more nonparametric fashion, e.g., as done by Stone (1982). However, it is well known that unrestricted multivariate nonparametric regression is subject to the "curse of dimensionality," fails to take advantage of structure in the phenomena being modeled and is hard to interpret. Stone (1985, 1986) proposes the additive

Received February 1986; revised April 1987.

[1]Research supported in part by Air Force Office of Scientific Research Grant AFOSR-ISSA-86-0025.

AMS 1980 *subject classifications.* Primary 62J05, 62J10; secondary 62G99.

Key words and phrases. Partially splined model, additive regression, semiparametric model.

regression model, which allows easier interpretation of the contribution of each explanatory variable and may be preferable to a fully nonparametric regression model for a moderate sample size if the model given by (1) is a good approximation and introduces a heuristic dimensionality reduction principle. According to that principle, the parametric components β could be estimated with the usual parametric convergence rates; and the overall regression function $X'\beta + g(T)$ should be estimable with the typical nonparametric convergence rates. However, we show that we may need to use different smoothing schemes in order to achieve such a goal for the model considered in this paper if X and T are not independent.

Engle, Granger, Rice and Weiss (1986), Green, Jennison and Scheult (1985), Wahba (1984) and others have studied the estimate of the regression function $X'\beta + g(T)$. Heckman (1986) and Chen (1985) proved that the estimate of β can achieve the convergence rate $n^{-1/2}$ if X and T are not related to each other. Rice (1986) obtains the asymptotic bias of a partial smoothing spline estimator of β due to the dependence between X and T and shows that it is not generally possible to attain $n^{-1/2}$ convergence rate for the parametric components β.

In this paper, an estimate of β, $\hat{\beta}$ is obtained by using a piecewise polynomial to approximate g. The convergence rate of $\hat{\beta}$ is shown to be $n^{-1/2}$ with the smallest possible variance even when X and T are dependent. Assumptions on the X and T will be introduced in the next section. During most of this paper we consider only real-valued T. Extensions to higher dimensions are indicated at the end of Section 2, in which our conditions and main results are formulated. The piecewise polynomial estimator for g is described in Section 2. The main proofs are presented in Section 3.

2. Statement of the main results. Let Y, $X = (x_1, \ldots, x_k)'$ and T be random variables such that T ranges over a nondegenerate compact 1-dimensional interval C, X is a $k \times 1$ vector in R^k and Y is real-valued. Without loss of generality, it can be assumed that C is the unit interval $[0,1]$. Let $\{X_i = (x_{i1}, \ldots, x_{ik})', T_i, Y_i, 1 \le i \le n\}$ denote a sample of size n from the model

$$Y_i = X_i'\beta + g(T_i) + e_i,$$

where the errors e_i are assumed to be independent and identically distributed with mean 0 and finite variance $\sigma^2 > 0$, (X_i, T_i) and e_i are independent, β is the $k \times 1$ vector of unknown parameters and g is an unknown smooth function. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{e} = (e_1, \ldots, e_n)'$, $\mathbf{1}_{n\times 1} = (1, \ldots, 1)'$, $\mathbf{g}(T) = (g(T_1), \ldots, g(T_n)'$ and $\mathbf{X} = (x_{ij})_{n\times k}$. Given $T = t$, set $\theta_i(t) = E(x_i|t)$ and $\Sigma_t = \mathrm{Cov}(X|t)$ for $1 \le i \le k$. We also assume that $\mathbf{1}_{n\times 1}$ is not in the space spanned by the column vectors of $\mathbf{X}$, so that the model is identifiable; that is, if $X_i'\beta_1 + g_1(T_i) = X_i'\beta_2 + g_2(T_i)$ for $1 \le i \le n$, then $\beta_1 = \beta_2$ and $g_1 = g_2$. If $\mathbf{1}_{n\times 1}$ is in the space spanned by the column vectors of $\mathbf{X}$, $\mathbf{X}a$, for some $a \in R^k$, is proportional to $\mathbf{1}_{n\times 1}$. This contradicts Condition 3 and the detail of this argument is shown by Lemma 3 in Section 3.

The following three conditions are sufficient for the statement of the main results.

CONDITION 1. The distribution of T is absolutely continuous and its density is bounded away from 0 and ∞ on C.

CONDITION 2. Let m, γ and M denote real constants such that $0 < \gamma \leq 1$ and $0 < M$; g is an m-times continuously differentiable function such that

$$\left|g^{(m)}(t') - g^{(m)}(t)\right| \leq M|t' - t|^{\gamma}, \text{ for } 0 \leq t, t' \leq 1.$$

Think of $p = m + \gamma$ as a measure of the smoothness of the function g.

CONDITION 3. There exist positive definite matrices Σ_{00} and Σ_{01} such that both $\Sigma_t - \Sigma_{00}$ and $\Sigma_{01} - \Sigma_t$ are nonnegative definite for all $t \in [0, 1]$.

If X and T are functionally related, it may not be possible to estimate β with rate $n^{-1/2}$. An example of such a case is that $g(T) = \alpha x_1$ for some unknown constant α. In practice, the implication of Condition 3 is that any linear combination of the components of X cannot be a function of T. The validation of Condition 3 can be checked by plotting the scatter diagrams of the components of X with respect to T.

First, we describe a piecewise polynomial estimator of g, which has been investigated by Tukey (1961), Major (1973), Chen (1986) and Stone (1985). Although a piecewise polynomial is not the most widely used estimator, it does allow us to see easily why the $n^{-1/2}$ convergence rate for β can be achieved. Given a positive integer M_n, the estimator has the form of a piecewise polynomial of degree m based on M_n intervals of length $1/M_n$, where the $(m + 1)M_n$ coefficients are chosen by the method of least squares on the basis of the data $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$, $1 \leq i \leq n$. Let $I_{n\nu}$, $1 \leq \nu \leq M_n$, denote the subintervals of $[0, 1]$ defined by $I_{n\nu} = [(\nu - 1)/M_n, \nu/M_n)$ for $1 \leq \nu < M_n$ and $I_{nM_n} = [1 - 1/M_n, 1]$. Let $\psi_{n\nu}$ denote the indicator function for the interval $I_{n\nu}$, so that $\psi_{n\nu}(t) = 1$ or 0 according to $t \in I_{n\nu}$ or $t \notin I_{n\nu}$. Consider the piecewise polynomial estimator of g of degree m given by

$$\hat{g}_n(t) = \sum_{\nu} \psi_{n\nu}(t)\hat{P}_{nm\nu}(t),$$

where $\{\hat{P}_{nm\nu}\}$ are polynomials of degree m chosen to minimize the residual sum of squares

$$\sum_i \left(Y_i - X_i\hat{\beta} - \hat{g}_n(T_i)\right)^2.$$

Set

$$\sigma_{ij} = \operatorname{Cov}\left(x_i - \theta_i(T), x_j - \theta_j(T)\right) = \operatorname{Cov}(x_i, x_j) - \operatorname{Cov}\left(\theta_i(T), \theta_j(T)\right),$$

for $1 \leq i, j \leq k$, and $\Sigma = (\sigma_{ij})_{k \times k}$.

THEOREM 1. *Suppose that Conditions* 1–3 *hold and that* $\lim_n n^{-\lambda}M_n = 0$ *for some* $\lambda \in (0, 1)$ *and* $\lim_n \sqrt{n}\, M_n^{-p} = 0$. *Then* $\sqrt{n}(\hat{\beta} - \beta)$ *converges to a* k*-variate normal distribution with mean* 0 *and the variance–covariance matrix* $\sigma^2\Sigma^{-1}$.

REMARK 1. An interesting question related to Theorem 1 is whether $\sigma^2\Sigma^{-1}$ is the smallest possible asymptotic variance. Recently, the Hájek–Le Cam convolution-type representation theorem has been further developed and applied by Begun, Hall, Huang and Wellner (1983) and Schick (1986) to a wide range of "regular" semiparametric models. If the definition of "regular" estimator $\hat{\beta}$ of β given by Schick (1986) is adopted, a slight modification of Theorem 3.1 of Begun, Hall, Huang and Wellner (1983) and the argument of Schick (1986) lead to the conclusion that $\sigma^2\Sigma^{-1}$ is the smallest possible asymptotic variance for all regular estimators of β when the random error e is normally distributed.

REMARK 2. If we put some smooth conditions on $\theta_i(T)$ such as $|\theta_i^{(m_1)}(t') - \theta_i^{(m_1)}(t)| \le M_1|t' - t|^{\gamma_1}$ *for* $0 \le t, t' \le 1$ and $1 \le i \le k$, where m_1, γ_1 *and* M_1 denote real constants such that $0 < \gamma_1 \le 1$ and $0 < M_1$, Theorem 1 holds for all (X, T) that satisfy Condition 3. However, the least-squares estimator of β proposed in this paper depends only on the smooth parameter p $(= m + \gamma)$.

REMARK 3. If we replace Condition 2 by assuming that $Eg^2(T)$ is finite, Theorem 1 is still true (for $p = 1$) since a function in L^2 can be arbitrarily well approximated in norm by a continuous function. However, the result holds only for a specific unknown function g.

A generalization of the model described in this paper is as follows:

$$Y = X'\beta + g(T) + e \quad \text{(model I)}$$

and

$$Y = \mu + X'\beta + \sum_{j=1,\,d} g_j(t_j) + e \quad \text{(model II)},$$

where $T = (t_1, \ldots, t_d) \in C$ in R^d, g is a smooth function in R^d and g_j are smooth functions in R that satisfy Condition 3 and $Eg_j(t_j) = 0$ for $1 \le j \le d$.

Model II is assumed to be additive in each variable t_j, whereas model I admits the interactions among t_j's. Theorem 1 can easily be extended to model I but the choice of M_n (the number of partitions at each coordinate) should satisfy $\lim_n n^{-\lambda}M_n = 0$ for some $\lambda \in (0,1)$ and $\lim_n \sqrt{n}\, M_n^{-dp} = 0$.

Although Theorem 1 still holds for model II, the estimator proposed in this paper does not take advantage of the additive structure on g. Alternatively, one can use the additive spline estimator, which is a smoothly joined piecewise polynomial proposed by Stone (1985), to estimate g. Let $\mathrm{ACE}(x|t_1, \ldots, t_d)$ be the function $\mu + \sum_{j=1,\,d} h_j(t_j)$ that minimizes $E(x - \mu + \sum_{j=1,\,d} h_j(t_j))^2$ for $\mu \in R$, $Eh_j(t_j) = 0$ and $Eh_j^2(t_j) < \infty$. In this case, Theorem 1 holds for those M_n that satisfy $\lim_n n^{-\lambda}M_n = 0$ for some $\lambda \in (0,1)$ and $\lim_n \sqrt{n}\, M_n^{-p} = 0$, but the covariance matrix is $\sigma^2\Sigma^{-1}$, where $\Sigma = (\sigma_{ij})_{k\times k}$ and

$$\sigma_{ij} = \mathrm{Cov}\big(x_i - \mathrm{ACE}(x_i|t_1, \ldots, t_d),\, x_j - \mathrm{ACE}(x_j|t_1, \ldots, t_d)\big).$$

The next two theorems are used to illustrate whether we can, by the proposed estimator, simultaneously estimate β and $g(t)$ with convergence rates $n^{-1/2}$ and

$n^{-p/(2p+1)}$. To avoid the unnecessary complexity in the proof, we will put an additional condition on $\theta_i(t)$.

CONDITION 4. Let M_2 be a positive constant such that

$$|\theta_i(t)| \le M_2, \quad \text{for all } t \in [0,1] \text{ and } 1 \le i \le k.$$

THEOREM 2. *Suppose that Conditions* 1–4 *hold and that* $\lim_n n^{-\lambda} M_n = 0$ *for some* $\lambda \in (0,1)$ *and* $\lim_n M_n = \infty$. *Then*

$$\hat{\beta} - \beta = O(M_n^{-p}) + O_p(n^{-1/2})$$

and

$$\hat{g}(t) - g(t) = O(M_n^{-p}) + O_p\left(\sqrt{n^{-1}M_n}\right), \quad \text{for any given } t \in [0,1].$$

REMARK 4. If $M_n = n^{1/(2p+1)}$, β and $g(T)$ can be estimated with convergence rate $n^{-p/(2p+1)}$, which is consistent with Theorem 1 of Stone (1985). However, Theorem 1 does not hold for this choice of M_n.

CONDITION 5. Let m_1, γ_1 and M_1 denote real constants such that $0 < \gamma_1 \le 1$ and $0 < M_i$; the θ_i's are m_1-times continuously differentiable functions such that

$$\left|\theta_i^{(m_1)}(t') - \theta_i^{(m_1)}(t)\right| \le M_1 |t' - t|\gamma_1, \quad \text{for } 0 \le t, t' \le 1.$$

Set $p_1 = m_1 + \gamma_1$ as a measure of the smoothness of the function θ_i.

THEOREM 3. *Suppose that Conditions* 1–5 *hold and that* $M_n = n^{1/(2p+1)}$ *and* $p_1 > 1/2$. *Then* $\sqrt{n}(\hat{\beta} - \beta)$ *converges to a k-variate normal distribution with mean* 0 *and the variance–covariance matrix* $\sigma^2\Sigma^{-1}$, *and* $\hat{g}(t) - g(t) = O_p(n^{-p/(2p+1)})$ *for any given* $t \in [0,1]$.

REMARK 5. Theorem 3 shows that we can estimate β and $g(T)$ with rates $n^{-1/2}$ and $n^{-p/(2p+1)}$, respectively, if the θ_i's are smooth ($p_1 > 1/2$). This explains the results obtained by Heckman (1986) and Chen (1985) since they assumed that those θ_i's are constant functions. Speckman (1986) also discusses this phenomenon and proposes the "partial kernel" estimator that can estimate β and $g(T)$ with rates $n^{-1/2}$ and $n^{-p/(2p+1)}$ simultaneously if g and the θ_i's satisfy Condition 2.

3. Proofs of the theorems. Throughout this section it is assumed that Conditions 1–3 hold, $\lim_n n^{-\lambda} M_n = 0$ for some $\lambda \in (0,1)$ and $\lim_n M_n = \infty$. Since any smooth function can be approximated well by a polynomial locally, $\hat{g}_n(t)$ $[= \Sigma_\nu \psi_{n\nu}(t)\hat{P}_{nm\nu}(t)]$ is used to approximate the function $g(t)$. For notational convenience, we write $(\hat{g}_n(T_1), \ldots, \hat{g}_n(T_n))'$ as $\mathbf{Z}\alpha$, where $\mathbf{Z}$ is an $n \times (m+1)M_n$ matrix and α is an $(m+1)M_n \times 1$ vector. Hence we need to find β and α to minimize $(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha)'(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\alpha)$.

LEMMA 1. *Suppose the minimization problem has a unique solution. Then* $\hat{\alpha} = A(\mathbf{Y} - \mathbf{X}\hat{\beta})$ *and* $\hat{\beta} = (\mathbf{X}'(I - P)\mathbf{X})^{-1}\mathbf{X}'(I - P)\mathbf{Y}$, *where* $A = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ *and* $P = \mathbf{Z}A$.

PROOF. See Theorem 3.7 of Seber (1976). □

LEMMA 2. *The minimization problem has a unique solution except on an event whose probability tends to* 0 *with* n.

PROOF. Since the density of T is bounded away from 0 and ∞, it follows from the Glivenko–Cantelli lemma and the fundamental theorem of algebra that $\mathbf{Z}$ is a full rank matrix except on an event whose probability tends to 0 with n. If we can show that $\mathbf{X}'(I - P)\mathbf{X}$ is a positive definite matrix, the proof of Lemma 2 will be complete.

Given $a = (a_1, \ldots, a_k)' \in R^k$ and $\Sigma_i a_i^2 = 1$, observe

$$E(a'X|T = t) = \sum_i a_i\theta_i(t) = \theta_a(t)$$

and

$$\operatorname{Var}(a'X|T = t) = a'\Sigma_t a.$$

It follows from Condition 3 that $a'\Sigma_t a$ is bounded away from 0 and ∞ for $0 \le t \le 1$. Since

$$a'\mathbf{X}'(I - P)\mathbf{X}a = (\mathbf{X}a - P\mathbf{X}a)'(\mathbf{X}a - P\mathbf{X}a)$$

can be thought of as the residual sum of squares after we regress $\mathbf{X}a$ on T [i.e., use a piecewise mth-degree polynomial to estimate $\theta_a(T) = E(a'X|T)$], it follows easily that $a'\mathbf{X}'(I - P)\mathbf{X}a$ tends to ∞ in probability as n goes to ∞ based on the proof of Theorem 1 in Stone (1982), on $a'\Sigma_t a$ being bounded away from 0 and ∞ for $0 \le t \le 1$, and on the general property of solution to least-squares problems. Hence $\mathbf{X}'(I - P)\mathbf{X}$ is a positive definite matrix. □

It follows from Lemma 1 that

$$\begin{aligned}\hat{\beta} - \beta &= (\mathbf{X}'(I - P)\mathbf{X})^{-1}\mathbf{X}'(I - P)\mathbf{Y} - \beta \\ &= (\mathbf{X}'(I - P)\mathbf{X})^{-1}\mathbf{X}'(I - P)(\mathbf{g}(T) + e) \qquad (2) \\ &= (\mathbf{X}'(I - P)\mathbf{X})^{-1}\mathbf{X}'(I - P)\mathbf{g}(T) + (\mathbf{X}'(I - P)\mathbf{X})^{-1}\mathbf{X}'(I - P)\,e.\end{aligned}$$

The purpose of the next lemma is to prove that the model described by (1) is identifiable.

LEMMA 3. $\mathbf{1}_{n\times 1}$ *is not in the space spanned by the column vectors of* $\mathbf{X}$ *except on an event whose probability tends to* 0.

PROOF. By the definition of $\mathbf{Z}$ and P, $\mathbf{1}_{n\times 1}$ is in the space spanned by the column vectors of P. It follows easily from the proof of Lemma 2 that

$a'\mathbf{X}'(I-P)\mathbf{X}a$ tends to ∞ for $a=(a_1,\ldots,a_k)'\in R^k$ and $\Sigma_i a_i^2=1$ as n goes to ∞ in probability. Hence $\mathbf{1}_{n\times 1}$ is not in the space spanned by the column vectors of $\mathbf{X}$. □

LEMMA 4. *$nM_r^{-1}P$ and $nM_n^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}$ are bounded in probability; that is, every element of $nM_n^{-1}P$ and $nM_n^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}$ is $O_P(1)$.*

This result is due to Stone (1980, 1982). □

Next, we are going to prove that

$$\mathbf{X}'(I-P)\mathbf{X}/n\to\Sigma,\quad \text{as } n\to\infty. \tag{3}$$

Set $\varepsilon_i=x_i-\theta_i(T)$, $\varepsilon_{ij}=x_{ij}-\theta_i(T_j)$, $\varepsilon_i=(\varepsilon_{i1},\ldots,\varepsilon_{in})'$, $\theta(T)=(\theta_1(T),\ldots,\theta_k(T))'$ and $\theta_i(T)=(\theta_i(T_1),\ldots,\theta_i(T_n))'$ for $1\le i\le k$ and $1\le j\le n$. Let $f(T)$ denote the density function of T. Given $\mathbf{a}=(a_1,\ldots,a_k)'\in R^k$ and $\Sigma_i a_i^2=1$, it follows from Conditions 1 and 3 that

$$\begin{aligned}a'\Sigma a&=\operatorname{Var}(a'(X-\theta(T)))\\&=\int_{[0,1]}\operatorname{Var}(a'X|t)f(t)\,dt>0.\end{aligned}$$

Hence Σ is a positive definite matrix.

LEMMA 5. *$\mathbf{X}'(I-P)\mathbf{X}/n\to\Sigma$ in probability as n tends to ∞.*

PROOF. Observe that

$$\begin{aligned}(\mathbf{X}'(I-P)\mathbf{X})_{ij}&=(\varepsilon_i+\theta_i(T))'(I-P)(\varepsilon_j+\theta_j(T))\\&=\varepsilon_i'(I-P)\varepsilon_j+\varepsilon_i'(I-P)\theta_j(T)\\&\quad+\theta_i(T)'(I-P)\varepsilon_j+\theta_i(T)'(I-P)\theta_j(T).\end{aligned}\tag{4}$$

Since P is a nonnegative definite matrix except on an event whose probability goes to 0 as n tends to ∞ and P is also an idempotent matrix, we get

$$\begin{aligned}E(|\varepsilon_i'P\varepsilon_i|\,|P)&=E(\varepsilon_i'P\varepsilon_i|P)\\&=E(\operatorname{tr}(P\varepsilon_i\varepsilon_i')|P)\\&=O_p(1)\operatorname{tr}(PE(\varepsilon_i\varepsilon_i'))\\&=[\operatorname{Var}(x_i-\theta_i(T))](m+1)M_nO_p(1)\\&=o_p(n)\end{aligned}$$

and

$$\begin{aligned}E(|\varepsilon_i'P\varepsilon_j|\,|P)&\le\{E(\varepsilon_i'P\varepsilon_i|P)+E(\varepsilon_j'P\varepsilon_j|P)\}/2\\&=o_p(n).\end{aligned}$$

It follows from the law of large numbers that $\varepsilon_i'\varepsilon_j/n$ converges to σ_{ij} in

probability. Consequently,

$$\varepsilon_i'(I-P)\varepsilon_j/n \to \sigma_{ij}, \quad \text{in probability as } n \to \infty. \tag{5}$$

If we can prove

$$n^{-1}\theta_j(T)'(I-P)\theta_j(T) \to 0, \quad \text{in probability for } 1 \le j \le k, \tag{6}$$

then the conclusion of Lemma 5 will hold. Since $I-P$ is an idempotent matrix, it follows from (5) and (6), the Markov inequality and the Cauchy–Schwarz inequality that

$$n^{-1}\varepsilon_i'(I-P)\theta_j(T) \to 0, \quad \text{for } 1 \le i, j \le k,$$

except on an event whose probability tends to 0 as n tends to ∞.

It follows from Condition 3 that $E\theta_j^2(T)$ is finite for $1 \le j \le k$. Choose $\varepsilon > 0$, and let $\tilde{\theta}_j(T)$ be a continuous function on $[0,1]$ such that

$$E|\theta_j(T) - \tilde{\theta}_j(T)|^2 \le \varepsilon.$$

Then there exists a positive constant B_1 such that

$$E|\theta_j(T)(I-P)\theta_j(T)| \le 2\left\{E|\tilde{\theta}_j(T)(I-P)\tilde{\theta}_j(T)| + nB_1\varepsilon\right\}.$$

Thus to prove that (6) holds, it suffices to verify that (6) holds with θ_j replaced by $\tilde{\theta}_j$. Since $(I-P)\tilde{\theta}_j(T)$ is the remainder term of $\tilde{\theta}_j(T)$ after using an mth-degree polynomial to approximate $\tilde{\theta}_i(T)$ at each subinterval $I_{n\nu}$, and $\tilde{\theta}_j(T)$ is a continuous function on $[0,1]$ and because of the property of solution to linear least-squares problems, there exists a positive constant B_n with $\lim_n B_n = 0$ such that

$$\left|\left((I-P)\tilde{\theta}_j(T)\right)_i\right| \le B_n, \quad \text{for } 1 \le j \le k \text{ and } 1 \le i \le n.$$

Hence (6) is true. □

Now let $d_{mn\nu}$ be the center of the interval $I_{n\nu}$. Let $P_{mn\nu}$ be the Taylor polynomial approximation of degree m to g about $d_{mn\nu}$. Since

$$\left|g^{(m)}(t') - g^{(m)}(t)\right| \le M|t'-t|^{\gamma}, \quad \text{for } t', t \in C,$$

there is a constant $B_2 > 0$ such that

$$|\psi_{n\nu}(t)(P_{mn\nu} - g(t))| \le B_2 M_n^{-p}, \quad \text{for all } \nu \text{ and } t \in C. \tag{7}$$

Since $\hat{g}_n(t)$ is a piecewise polynomial, there are at most $O(n/M_n)$ nonzero elements in each row of P. Observe

$$(I-P)\mathbf{g}(T) = (I-P)(\mathbf{g}(T) - \mathbf{Z}\alpha + \mathbf{Z}\alpha) = (I-P)(\mathbf{g}(T) - \mathbf{Z}\alpha);$$

based on (7) and Lemma 4 there is a constant $B_3 > 0$ such that

$$|((I-P)\mathbf{g}(T))_i| \le B_3 M_n^{-p}, \quad \text{for } 1 \le i \le n, \tag{8}$$

except for an event whose probability tends to 0 with n.

Observe that

$$\sqrt{n}\,(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P)\mathbf{g}(T)$$
$$= (\mathbf{X}'(I-P)\mathbf{X}/n)^{-1}n^{-1/2}\mathbf{X}'(I-P)(I-P)\mathbf{g}(T).$$

It follows from the above arguments and Lemma 5 that there is a constant $B_4 > 0$ such that

$$\left|\left(\sqrt{n}\,(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P)\mathbf{g}(T)\right)_\nu\right| \le B_4\sqrt{n}\,M_n^{-p}, \quad \text{for } 1 \le \nu \le k,$$

except for an event whose probability tends to 0 with n.

Next, we show that $\sqrt{n}\,(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P)\,e$ converges in distribution to a k-variate normal random variable with mean 0 and covariance matrix $\sigma^2\Sigma^{-1}$.

Let r_{ii} be the ith diagonal element of the projection matrix

$$(I-P)\mathbf{X}(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P).$$

According to Proposition 2.2 of Huber (1973), if we can prove

$$\max_i r_{ii} \to 0, \quad \text{as } n \to \infty,$$

then $\sqrt{n}\,(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P)\,e$ is asymptotically normal. Since the covariance matrix of $\sqrt{n}\,(\mathbf{X}'(I-P)\mathbf{X})^{-1}\mathbf{X}'(I-P)\,e$ is given by $n(\mathbf{X}'(I-P)\mathbf{X})^{-1}\sigma^2$, the asymptotic variance of $\sqrt{n}\,(\hat{\beta}-\beta)$ is $\sigma^2\Sigma^{-1}$ by Lemma 5.

PROOF OF THEOREM 1. Since $(\mathbf{X}'(I-P)\mathbf{X})/n$ converges to a positive definite matrix Σ by Lemma 5, it follows from Lemma 3 of Wu (1981) that

$$\max_i r_{ii} \to 0, \quad \text{as } n \to \infty.$$

Hence it follows from the above arguments that $\sqrt{n}\,(\hat{\beta}-\beta)$ converges in distribution to a normal random variable with mean 0 and covariance matrix $\sigma^2\Sigma^{-1}$ if $\lim_n \sqrt{n}\,M_n^{-p} = 0$. □

PROOF OF THEOREM 2. It follows from the proof of Theorem 1 that

$$\hat{\beta}-\beta = O(M_n^{-p}) + O_p(n^{-1/2}).$$

For any given $t \in [0,1]$, we can write $\hat{g}(t)$ as $z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Y}-\mathbf{X}\hat{\beta})$ by Lemma 1 for a suitable z. Hence

$$\hat{g}(t) - g(t) = z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Y}-\mathbf{X}\hat{\beta}) - g(t)$$
$$= z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\,e + z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}(T) - g(t)$$
$$+ z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\beta - \mathbf{X}\hat{\beta}).$$

It follows from the argument of Stone (1980) that

$$(9)\quad z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\,e + z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}(T) - g(t) = O(M_n^{-p}) + O_p\left((n/M_n)^{-1/2}\right).$$

Let $\mathbf{X}_i$ be the ith column vector of $\mathbf{X}$. Observe that

$$z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_i = z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\theta_i(T) + \varepsilon_i).$$

It follows from Condition 4 that the θ_i's are bounded and the argument of Stone (1980) that

$$z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_i = O(1) + O_p\big((n/M_n)^{-1/2}\big).$$

Hence

$$\begin{aligned} z(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\hat{\beta} - \beta) &= \big[O(1) + O_p\big((n/M_n)^{-1/2}\big)\big] \\ &\quad \times \big[O(M_n^{-p}) + O_p(n^{-1/2})\big] \\ &= O(M_n^{-p}) + O_p\big((n/M_n)^{-1/2}\big). \end{aligned} \tag{10}$$

[redacted]

$$\hat{g}(t) - g(t) = O(M_n^{-p}) + O_p\big((n/M_n)^{-1/2}\big).$$

This establishes Theorem 2. □

PROOF OF THEOREM 3. Observe that for $1 \le \nu \le k$,

$$\begin{aligned} (\mathbf{X}'(I - P)\mathbf{g}(T))_\nu &= (\theta_\nu(T) + \varepsilon_\nu)'(I - P)\mathbf{g}(T) \\ &= \theta_\nu(T)'(I - P)(I - P)\mathbf{g}(T) + \varepsilon_\nu'(I - P)\mathbf{g}(T). \end{aligned}$$

It follows from (8), Condition 5, the argument used to derive (8) and Lemma 4

[redacted]

that there is a constant $B_5 > 0$ such that

$$|\theta_\nu(T)'(I - P)(I - P)\mathbf{g}(T)| \le B_5 M_n^{-(p+p_1)}, \quad \text{for } 1 \le \nu \le k, \tag{11}$$

except for an event whose probability tends to 0 with n. We also get

$$\varepsilon_\nu'(I - P)\mathbf{g}(T) = O_p\big(n^{1/2}M_n^{-p}\big), \quad \text{for } 1 \le \nu \le k, \tag{12}$$

by (8) and Chebyshev's inequality. It follows from the proof of Theorem 1, the choice of M_n $(= n^{1/(2p+1)})$, (11) and (12) that $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribu-

CHEN, H. (1985). Data smoothing in analysis of covariance. Manuscript, Dept. Appl. Math. and Statistics, SUNY at Stony Brook.

CHEN, K. W. (1987). Asymptotically optimal selection of a piecewise polynomial estimator of a regression function. *J. Multivariate Anal.* **22** 230–244.

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.

GREEN, P., JENNISON, C. and SEHEULT, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47** 299–315.

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

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

MAJOR, P. (1973). On a nonparametric estimation of the regression function. *Studia Sci. Math. Hungar.* **8** 347–361.

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

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151.

SEBER, G. A. F. (1976). *Linear Regression Analysis*. Wiley, New York.

SPECKMAN, P. (1986). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B*. To appear.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

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.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

TUKEY, J. (1961). Curves as parameters, and touch estimation. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 681–694. Univ. California Press.

WAHBA, G. (1984). Cross validated 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.) 205–235. Iowa State Univ. Press, Ames, Ia.

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513.

DEPARTMENT OF APPLIED
MATHEMATICS AND STATISTICS
STATE UNIVERSITY OF NEW YORK
STONY BROOK, NEW YORK 11794-3600