OPTIMAL GLOBAL RATES OF CONVERGENCE IN PARTLY LINEAR MODELS

by

YOUNG K. TRUONG

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2174

February 1997
OPTIMAL GLOBAL RATES OF CONVERGENCE
IN PARTLY LINEAR MODELS

YOUNG K. TRUONG

The University of North Carolina at Chapel Hill

December 17, 1996

Abstract. Let \((X, Z, Y)\) denote a random vector such that \(Z\) and \(Y\) are real-valued, and \(X \in \mathbb{R}^d\). Consider the partly linear regression function \(E(Y \mid Z, X) = \beta Z + \phi(X)\), where \(\beta\) is an unknown parameter, and \(\phi(\cdot)\) is a smooth function on \(\mathbb{R}^d\). Suppose that \(\phi(\cdot)\) has a bounded \(k\)th derivative. Under appropriate conditions, it will be shown that the local polynomial based estimate achieve the usual optimal \(L_2\) and \(L_\infty\) rates of convergence given by \(n^{-k/(2k+d)}\) and \((\log n/n)^{k/(2k+d)}\), respectively.

Keywords. Partly linear models; semi-parametric models; nonparametric regression; local polynomial estimator, optimal rate of convergence.

\(^1\)This research was supported in part by National Science Foundation Grant DMS-9403800 and National Institute of Health Grant CA61937.
1 Introduction

Let $(X, Z, Y)$ denote a random vector such that $Z$ and $Y$ are real-valued, and $X \in \mathbb{R}^d$. In partly linear models, the regression function is given by

$$E(Y \mid Z, X) = \beta Z + \phi(X),$$

where $\beta$ is an unknown parameter and $\phi(\cdot)$ is a real-valued smooth function on $\mathbb{R}^d$.

There are many practical applications for partly linear models. In the classical analysis of covariance models, one is interested in estimating the 'treatment effect' $Z$ by linearly adjusting the covariates $X$ (Scheffé [11]). We use model (1.1) to adjust the unspecified covariate effect on the outcome while still maintaining an interest on the parametric inferences for $\beta$. In functional modeling involving a categorical variable $Z$ and continuous covariates $X$, it is more appealing to regress the response $Y$ on $Z$ linearly and then model the continuous covariate effects using functional approaches. For further specific applications of partly linear models, see Wahba [17], Engle et al. [4], Green et al. [7], Shiau et al. [12], and Speckman [13].

Partly linear models, on the other hand, has an interesting motivation for the theoretical development of functional modeling procedures. In fact, the use of additivity can be viewed as a way to alleviate the curse of dimensionality in functional modeling, see Stone [16]. Furthermore, in situations where it is appropriate to model the effect of $Z$ parametrically, one would expect to estimate $\beta$ with root-$n$ rate of convergence.

For non-random $X$ in (1.1), the issues on rates of convergence for estimating $\beta$ and $\phi(\cdot)$ have been addressed by Wahba [17], Engle et al. [4], Green et al. [7], Shiau et al. [12], Heckman [9], Rice [10], Chen and Shiau [2, 3] using smoothing spline approaches; by Speckman [13] using kernel methods; and by Eubank, Hart and Speckman [5] using a trigonometric series approach. For random covariate $X$, a piecewise constant approach was considered by Chen [1] where the asymptotic distribution for the estimate of
\( \beta \) and the pointwise rate of convergence for estimating \( \phi(\cdot) \) are established. These results have been extended to local linear methods by Hamilton and Truong [8]. In particular, they obtained the asymptotic normal distributions for local linear based estimates of \( \beta \) and \( \phi(\cdot) \).

The problem of demonstrating the achievability of the optimal global rates of convergence for functional estimate of \( \phi(\cdot) \) has not been addressed. Under appropriate conditions, the objective of the present paper is to establish that the local polynomial based estimate of \( \phi(\cdot) \) achieve the \( L_2 \) and \( L_\infty \) rates of convergence given by \( n^{-k/(2k+d)} \) and \( (\log n/n)n^{k/(2k+d)} \), respectively. According to Stone [15], these rates are optimal for estimates based on \( d \)-dimensional covariate component \( \phi(\cdot) \). This is an improvement of the rate of convergence for estimates using the original \( (d+1) \)-dimensional regressor, the faster rates is achieved by using the additivity and the parametric assumption in model (1.1).

The rest of the paper is organized as follows. Section 2 motivates and describes the local polynomial estimates in partly linear model (1.1). Asymptotic properties of the estimates will be given in Section 3. In particular, root-\( n \) consistency of the parametric estimate and the global rate of convergence will be discussed. Proofs are given in Section 4.

2 Method

The method of estimation can be motivated as follows. Suppose that

\[
Y = \beta Z + \phi + \varepsilon,
\]

where

\[
Y^\top = (Y_1, Y_2, \ldots, Y_n),
\]
\[
Z^\top = (Z_1, Z_2, \ldots, Z_n),
\]
\[
\phi^\top = (\phi(X_1), \phi(X_2), \ldots, \phi(X_n)),
\]
\( \phi \) is a smooth function on \( \mathbb{R}^d \), and \( \varepsilon^T = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \) is a random vector with mean zero. For a vector of numbers \( a \), let \( \text{ave}_i(a) \) denote the local average of \( a \) at \( X_i \) (see the definition of \( S(\cdot) \) given below). Then
\[
\text{ave}_i(Y) \approx \beta \text{ave}_i(Z) + \text{ave}_i(\phi).
\]
Since \( \phi \) is a smooth function, we have \( \text{ave}_i(\phi) \approx \phi(X_i) \). Thus
\[
Y_i - \text{ave}_i(Y) \approx \beta [Z_i - \text{ave}_i(Z)] + \phi(X_i) - \text{ave}_i(\phi) \approx \beta [Z_i - \text{ave}_i(Z)].
\]
This suggests that an estimate of \( \beta \) is obtained by regressing \( Y_i - \text{ave}_i(Y) \) on \( Z_i - \text{ave}_i(Z) \). Let \( \hat{\beta} \) denote such an estimate. Then \( \phi \) can be estimated by smoothing the response \( Y_i - \hat{\beta}Z_i \) using \( X_i \) as a predictor.

We now describe the method formally. Suppose (1.1) holds. Let \( x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \) and let \( \theta(x) \) denote the regression function of \( Z \) on \( X \), so that \( \theta(x) = E(Z \mid X = x) \). (In this paper, \( A^T \) denote the transpose of a matrix \( A \).) The smoother mentioned above will be described in terms of the local polynomial estimate of \( \theta(x) \). Let \( (X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n) \) denote a random sample from the distribution of \( (X, Z, Y) \). Let \( K(\cdot) \) denote a kernel function on \( \mathbb{R}^d \) and set \( K_h(x) = h^{-d}K(x/h) \), where \( h \) is a bandwidth. Given nonnegative integers \( s_1, \ldots, s_d \), set \( s = (s_1, \ldots, s_d) \), \( |s| = s_1 + \cdots + s_d \), \( s! = (s_1)! \cdots (s_d)! \), and \( x^s = x_1^{s_1} \cdots x_d^{s_d} \) for \( x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \). Also, set \( \mathcal{A} = \{s : |s| < k\} \). Given \( x \in \mathbb{R}^d \), let \( W_n(x), X_n(x) \) and \( A_n(x) \) be defined as follows: \( W_n(x) \) is the \( n \times n \) diagonal matrix given by \( W_n(x) = \text{diag}(K_h(X_1 - x), \ldots, K_h(X_n - x)) \); \( X_n(x) = (X_{nis}(x)) \) is the \( n \times \#(\mathcal{A}) \) matrix given by \( X_{nis}(x) = (X_i - x)^s/h^{|s|} \); \( A_n(x) = (A_{nis}(x)) = X_n^T(x)W_n(x)X_n(x) \), which is a \( \#(\mathcal{A}) \times \#(\mathcal{A}) \) matrix.

Let \( a = (1, 0, \ldots, 0)^T \) denote the \( \#(\mathcal{A}) \)-dimensional vector and let \( S^T(x) \) denote the \( 1 \times n \) row vector given by
\[
S^T(x) = a^T[A_n(x)]^{-1}X_n^T(x)W_n(x).
\]
Set \( Z_n = (Z_1, \ldots, Z_n)^T \). Then the local polynomial estimator of \( \theta(x) \) is given by
\[
\hat{\theta}(x) = a^T[A_n(x)]^{-1}X_n^T(x)W_n(x)Z_n = S^T(x)Z_n.
\]
See Stone [14, 15].

We next describe the estimates of $\beta$ and $\phi(x)$. Let $S$ denote the $n \times n$ matrix with the $j$th row equal to $S^T(X_j)$, $j = 1, \ldots, n$. Let $I = I_n$ denote the $n \times n$ identity matrix and $Y_n = (Y_1, \ldots, Y_n)^T$. Set

$$\hat{Z} = (I - S)Z_n \quad \text{and} \quad \hat{Y} = (I - S)Y_n.$$ 

Let $\hat{\beta}$ denote the solution to $\min_{\beta} (\hat{Y} - \beta \hat{Z})^T(\hat{Y} - \beta \hat{Z})$, so that $\hat{\beta} = (\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T \hat{Y}$. Set

$$\hat{\phi}(x) = a^T [A_n(x)]^{-1} X_n^T(x) W_n(x)(Y_n - \hat{\beta} Z_n) = S^T(x)(Y_n - \hat{\beta} Z_n).$$

We use $\hat{\beta}$ and $\hat{\phi}(x)$ to estimate $\beta$ and $\phi(x)$ of model (1.1). Note that the kernel-based method described by Speckman [13] is a special case of the above procedure.

3 Results

In this section conditions and results for achieving the optimal rates of convergence will be described. The first condition is required for bounding the bias terms of the local polynomial estimates.

**Condition 1** The functions $\phi(\cdot)$ and $\theta(\cdot)$ have bounded $k$th partial derivatives.

The next two conditions are required for bounding the variance terms of our estimates.

**Condition 2** The random vector $X^T = (X_1, \ldots, X_d)$ has a continuous density function $f(\cdot)$ supported on a compact set $C_f \subset \mathbb{R}^d$. Moreover, $f(\cdot)$ is bounded away from zero and infinity on $C_f$.

**Condition 3** $\sup_{z,x} \text{var}(Y \mid Z = z, X = x) < \infty$ and $\sup_{x} \text{var}(Z \mid X = x) < \infty$. 

The following condition is required for the kernel function. It is easily satisfied by taking products of symmetric univariate kernels with compact supports.

**Condition 4** The kernel function $K(\cdot)$ is a continuous kernel function on $\mathbb{R}^d$ with compact support such that $\int K(u) \, du = 1$. Also, all odd order moments of $K$ vanish; that is, $\int u_1^s u_2^t K(u) \, du = 0$ for non-negative integers $s, t$ such that their sum is odd. Moreover, $\int uu^\top K(u) \, du = \mu_d(K) I$, where $\mu_d(K) > 0$ and $I$ is the $A \times A$ identity matrix.

The $L_2$ rate of convergence is described in the following result. Its proof can be found in subsection 4.1.

**Theorem 1** Suppose $h \sim n^{-1/(2k+d)}$, $2k > d$ and that Conditions 1–4 hold. Then

$$\int |\hat{\phi}(x) - \phi(x)|^2 f(x) \, dx = O_p \left( n^{-2k/(2k+d)} \right).$$

The next condition is required for establishing the $L_\infty$ rates of convergence of $\hat{\phi}$. See Stone [15].

**Condition 5** For $t > 0$,

$$\sup_x E \left( \exp\{t[Z - \theta(X)]\} \mid X = x \right) < \infty, \quad (3.1)$$

and

$$\sup_{z, x} E \left( \exp\{t[Y - \beta Z - \theta(X)]\} \mid Z = z, X = x \right) < \infty. \quad (3.2)$$

Note that Condition 5 implies Condition 3. The $L_\infty$ rate of convergence is described in the next result, whose proof will be given in subsection 4.2.

**Theorem 2** Suppose $h \sim (\log n/n)^{1/(2k+d)}$, $2k > d$ and that Conditions 1–5 hold. Then

$$\sup_{x \in C_f} |\hat{\phi}(x) - \phi(x)| = O_p \left( (\log n/n)^{k/(2k+d)} \right).$$

6
Remarks

1. Under similar conditions, Chen [1] obtained the pointwise rate of convergence \( n^{-2k/(2k+d)} \) using piecewise constant estimates. This is further extended to asymptotic normal distribution based on more appealing local linear estimates by Hamilton and Truong [8]. That is, they showed that \( \sqrt{n}(\hat{\beta} - \beta) \) and \( \sqrt{n}\hat{h}(\hat{\phi} - E\hat{\phi}) \) are asymptotically normal for local linear based estimates \( \hat{\beta} \) and \( \hat{\phi} \).

2. For non-random designs, Speckman [13] showed that the \( L_2 \) rate of convergence was achieved by kernel estimates. This result is generalized to random vector of predictors \( X \) as described in Theorem 1. Moreover, this result shows that the rate is achieved by local polynomial estimates, and is obtained without imposing extra smoothness conditions on the density function of \( X \). See Stone [15] and Fan [6]. Additional smoothness conditions will be required by the kernel method for random \( X \).

3. According to Stone [15], the optimal \( L_2 \) and \( L_\infty \) rates of convergence for functional (nonparametric) estimates of \( E(Y \mid Z, X) \) are given respectively by \( n^{-k/(2k+d+1)} \) and \( (n^{-1}\log n)^{k/(2k+d+1)} \). Under the lower dimensional model (1.1), these rates have been improved and are given by \( n^{-k/(2k+d)} \) and \( (n^{-1}\log n)^{k/(2k+d)} \). These results and those given in Hamilton and Truong [8] provide a partial justification to the *dimension reduction principle* as discussed in Stone [16].

4 Proofs

Let \( \theta_k(\cdot; x) \) and \( \phi_k(\cdot; x) \) denote the \( (k-1)\)th-degree Taylor polynomials of \( \theta(\cdot) \) and \( \phi(\cdot) \) about \( x \), respectively. That is,

\[
\theta_k(\cdot; x) = \sum_{|s|<k} \frac{D^s\theta(x)}{s!} (-x)^s \quad \text{and} \quad \phi_k(\cdot; x) = \sum_{|s|<k} \frac{D^s\phi(x)}{s!} (-x)^s, \quad x \in C_f,
\]
where
\[ D^s = \frac{\partial |s|}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}. \]

Then
\[ \sum_{j=1}^n S(X_j, x)\theta_k(X_j; x) = \theta(x) \quad \text{and} \quad \sum_{j=1}^n S(X_j, x)\phi_k(X_j; x) = \phi(x). \quad (4.1) \]

We next describe a few results that will be used in the proofs.

**Lemma 1** Suppose Conditions 2 and 4 hold. Then there is a positive constant \( c_1 \) such that
\[ \lim_n P\left(n\{[A_n(x)]^{-1}\}_{st} \leq c_1 \quad \text{for } x \in C_f \text{ and } s, t \in A\right) = 1. \]

**Proof.** See Stone [15] or Lemma 1 of Hamilton and Truong [8]. \( \Box \)

It follows from Conditions 2 that there are positive constants \( c_2 \) and \( c_3 \) such that
\[ \lim_n P\left(c_2 \leq n^{-1} \sum_{i=1}^n K_h(X_i - x) \leq c_3, \quad x \in C_f\right) = 1. \quad (4.2) \]

Denote the \( i \)th entry of \( S(x) \) by \( S_i(x) \). The next result gives an upper bound for \( ||S(x)|| \) and \( \max_i |S_i(x)| \).

**Lemma 2** Suppose Conditions 2 and 4 hold. Then there are positive constants \( c_4 \) and \( c_5 \) such that
\[ \lim_n P(\Psi_n) = 1, \]

where
\[ \Psi_n = \left\{ \max_{x \in C_f} ||S(x)||^2 \leq c_4 n^{-1} h^{-d}, \max_{x \in C_f} \max_{1 \leq i \leq n} |S_i(x)| \leq c_5 n^{-1} h^{-d} \right\}. \]

**Proof.** See Lemma 2 of Hamilton and Truong [8]. \( \Box \)

The following result describes the rate of convergence of \( \hat{\beta} \), which will be required in the proof of Theorems 1 and 2.
Lemma 3 Suppose Conditions 1–4 hold. Then

\[ |\hat{\beta} - \beta| = O_p(n^{-1/2}). \]

Proof. See Theorem 1 of Hamilton and Truong [8]. Note that the condition \( 2p > d \) is required to achieve the indicated rate of convergence. \( \square \)

4.1 Proof of Theorem 1

Set \( \Phi_n = [\phi(X_1), \ldots, \phi(X_n)]^T \) and \( \Theta_n = [\theta(X_1), \ldots, \theta(X_n)]^T \). Write

\[
\hat{\phi}(x) - \phi(x) = S^T(x)(Y_n - \beta Z_n - \Phi_n) + [S^T(x)\Phi_n - \phi(x)] - (\hat{\beta} - \beta)S^T(x)Z_n. \quad (4.3)
\]

Let \( E_n(V) \) denote the the conditional expectation of a random variable \( V \) given \( X_1, \ldots, X_n \) and \( B_1, \ldots, B_n \). It follows from Condition 3 and Lemma 2 that there are positive constants \( c_6 \) and \( c_7 \) such that

\[
E_n|S^T(x)(Y_n - \beta Z_n - \Phi_n)|^2 \leq c_6 n^{-1} h^{-d}, \quad x \in C_f \text{ on } \Psi_n, \quad (4.4)
\]

\[
E_n|S^T(x)(Z_n - \Theta_n)|^2 \leq c_7 n^{-1} h^{-d}, \quad x \in C_f \text{ on } \Psi_n. \quad (4.5)
\]

From Lemma 1, (4.1) and Condition 1,

\[
\sup_{x \in C_f} |S^T(x)\Phi_n - \phi(x)| = O_p(h^k), \quad (4.6)
\]

\[
\sup_{x \in C_f} |S^T(x)\Theta_n - \theta(x)| = O_p(h^k). \quad (4.7)
\]

It follows from (4.5), (4.7) and Condition 1 that there is a positive constant \( c_{10} \) such that

\[
E_n|S^T(x)Z_n|^2 = E_n|S^T(x)[Z_n - \Theta_n] + S^T(x)[\Theta_n - \theta(x)] + \theta(x)|^2 \leq c_{10}, \quad x \in C_f \text{ on } \Psi_n.
\]

Hence, by Lemmas 2 and 3,

\[
\int |(\hat{\beta} - \beta)S^T(x)Z_n|^2 f(x) \, dx
\]

\[
= (\hat{\beta} - \beta)^2 \int |S^T(x)Z_n|^2 f(x) \, dx = O_p(n^{-1}). \quad (4.8)
\]

The desired result follows from (4.3), (4.4), (4.6) and (4.8). \( \square \)
4.2 Proof of Theorem 2

We begin with two lemmas.

**Lemma 4** Suppose Conditions 1-4, (3.1) hold and that $1/\gamma h^d_n = o(n^{-\epsilon})$ for some $\epsilon > 0$. Then there is a positive constant $c_{11}$ such that

$$
\lim_n P \left( \sup_{x \in C_f} |S^T(x)(Z_n - \Theta_n)| \geq c_{11} \sqrt{\gamma (\log n) / nh^d} \right) = 0.
$$

**Proof.** Choose $\gamma > 0$. It follows from (3.1) and Markov's inequality that

$$
\max_i |Z_i - x_i\theta| = O_p(n^{-\gamma}). \quad (4.9)
$$

We may assume that $C_f = [-1/2, 1/2]^d$. Let $\lambda$ be a positive integer such that $L_n = n^\lambda$. Let $W_n$ be the collection of $(2L_n + 1)^d$ points in $[-1/2, 1/2]^d$ each of whose coordinates is of the form $j/(2L_n)$ for some integer $j$ such that $|j| \leq L_n$. Then $[-1/2, 1/2]^d$ can be written as the union of $(2L_n)^d$ subcubes, each having length $1/(2L_n)$ and all of its vertices in $W_n$. For each $x \in [-1/2, 1/2]^d$ there is a subcube $Q_w$ with center $w$ such that $x \in Q_w$. Let $C_n$ denote the collection of centers of these subcubes. Write

$$
S^T(x)(Z_n - \Theta_n) - S^T(w)(Z_n - \Theta_n)
= a^T [A_n^{-1}(x) - A_n^{-1}(w)]X(w)W(w)(Z_n - \Theta_n).
$$

Observe that

$$
\max_{w \in C_n} \sup_{x \in Q_w} \max_{i \in A} |[X_{nis}(x) - X_{nis}(w)]K_h(X_i - w)|
= O\left(n^{-\lambda h^{-1-d}}\right). \quad (4.11)
$$

Thus, for sufficiently large $\lambda$,

$$
X(x)W(x) - X(w)W(w) = X(x)W(x) - X(x)W(w)
+ X(x)W(w) - X(w)W(w)
= O_p\left(n^{-\lambda h^{-1-d}}\right).
$$
We conclude from Lemma 1 and (4.9) that, for \( \lambda \) sufficiently large,

\[
\max_{w \in C_n} \sup_{x \in Q_w} |a^T A_n^{-1}(x)[X(x)W(x) - X(w)W(w)](Z_n - \Theta_n)| = O_p(\sqrt{1/nh^d}). \tag{4.12}
\]

It follows from Condition 4 and (4.11) that, for \( \lambda \) sufficiently large,

\[
\max_{w \in C_n} \sup_{x \in Q_w} \max_{s,t \in A} |A_{nst}(x) - A_{nst}(w)| = o_p\left(n^\gamma h^{-d}\right). \tag{4.13}
\]

We conclude from Lemma 1, (4.2) and (4.13) that, for \( \lambda \) sufficiently large,

\[
\max_{w \in C_n} \sup_{x \in Q_w} \max_{s,t \in A} |(A_n^{-1}(x))_{st} - (A_n^{-1}(w))_{st}| = o_p\left(n^\gamma (nh^d)^{-1}\right)
\]

and hence that

\[
\max_{w \in C_n} \sup_{x \in Q_w} |a^T [A_n^{-1}(x) - A_n^{-1}(w)]X_n^T(w)W_n(w)(Z_n - \Theta_n)|
\]

\[
= O_p\left(\sqrt{1/nh^d}\right). \tag{4.14}
\]

By (4.10), (4.12) and (4.14) that

\[
\max_{w \in C_n} \sup_{x \in Q_w} |[S^T(x) - S^T(w)](Z_n - \Theta_n)| = o_p\left(\sqrt{1/nh^d}\right). \tag{4.15}
\]

For each fixed \( \lambda > 0 \), it follows from Lemma 2, (3.1) and Markov's inequality that there is a positive constant \( c_{12} \) such that

\[
\lim_{n \to \infty} P\left(\max_{w \in C_n} |S^T(w)(Z_n - \Theta_n)| \geq c_{12}\sqrt{(\log n)/nh^d}\right) = 0. \tag{4.16}
\]

(See Stone [15].) The desired result follows from (4.15) and (4.16). \( \square \)

The next result follows similarly from the above argument.

**Lemma 5** Suppose Conditions 1-4, (3.2) hold and that \( 1/nh_n^d = o(n^{-\epsilon}) \) for some \( \epsilon > 0 \). Then there is a positive constant \( c_{13} \) such that

\[
\lim_{n \to \infty} P\left(\sup_{x \in C_n} |S^T(x)(Y_n - \beta Z_n - \Phi_n)| \geq c_{13}\sqrt{(\log n)/nh^d}\right) = 0.
\]
We now give a proof of Theorem 2. Note that $|S^T(x)Z_n| \leq |S^T(x)(Z_n - \Theta_n)| + |S^T(x)[\Theta_n - \theta(x)]| + |\theta(x)|$. Hence by Lemmas 3, 4 and (4.7),

$$\sup_{x \in C_f} |(\hat{\beta} - \beta)S^T(x)Z_n| = O_p(n^{-1}). \quad (4.17)$$

The conclusion of Theorem 2 follows by applying Lemma 5, (4.6) and (4.17) to (4.3). □

References


12


Department of Biostatistics
CB# 7400, McGavran-Greenberg Hall
The University of North Carolina at Chapel Hill
Chapel Hill NC 27599-7400