A difference based approach to the semiparametric partial linear model

Probability Surveys
ISSN: 1549-5787
arXiv: math.PR/0000000

A Difference Based Approach to the Semiparametric Partial Linear Model

Lie Wang*,

Department of Mathematics, Massachusetts Institute of Technology,
e-mail: liewang@math.mit.edu

Lawrence D. Brown† and T. Tony Cai‡

Department of Statistics, The Wharton School, University of Pennsylvania,
e-mail: lbrown@wharton.upenn.edu; tcai@wharton.upenn.edu

Abstract: A commonly used semiparametric partial linear model is considered. We propose analyzing this model using a difference based approach. The procedure estimates the linear component based on the differences of the observations and then estimates the nonparametric component by either a kernel or a wavelet thresholding method using the residuals of the linear fit. It is shown that both the estimator of the linear component and the estimator of the nonparametric component asymptotically perform as well as if the other component were known. The estimator of the linear component is asymptotically efficient and the estimator of the nonparametric component is asymptotically rate optimal. A test for linear combinations of the regression coefficients of the linear component is also developed. Both the estimation and the testing procedures are easily implementable. Numerical performance of the procedure is studied using both simulated and real data. In particular, we demonstrate our method in an analysis of an attitude data set as well as a data set from the Framingham Heart Study.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: Asymptotic efficiency, difference-based method, kernel method, wavelet thresholding method, partial linear model, semiparametric model.

*Supported in part by NSF Grant DMS-1005539.
†Supported in part by NSF Grant DMS-0707033.
‡Supported in part by NSF FRG Grant DMS-0854973.

1. Introduction

Semiparametric models have received considerable attention in statistics and econometrics. In these models, some of the relations are believed to be of certain parametric form while others are not easily parameterized. In this paper, we consider the following semiparametric partial linear model

$$Y_i = a + X_i'\beta + f(U_i) + \epsilon_i, \quad i = 1, ..., n, \tag{1}$$

where $X_i \in I\!R^p$, $U_i \in I\!R$, β is an unknown vector of parameters, a is the unknown intercept term, $f(\cdot)$ is an unknown function and ϵ_i's are independent and identically distributed random noise with mean 0 and variance σ^2 and are independent of (X_i', U_i).

Literature Review

The semiparametric partial linear model has been extensively studied and several approaches have been developed to construct the estimators. A penalized least-squares method was used in for example [36, 14, 9]. A kernel smoothing approach was introduced in [32]. A partial residual method was proposed for example in [10]. And a profile likelihood approach was used in [31] and [6]. The test of significance of partial linear model was discussed in [16, 25, 39]. Moreover, the estimation of the nonparametric component is discussed in [7, 33, 18, 20]. The issue of achieving the information bound in this and other non- and semiparametric models has been examined by [28] and extensively discussed in [1].

In this article, a difference based estimation method is considered. The estimation procedure is optimal in the sense that the estimator of the linear component is asymptotically efficient, see for example [29], and the estimator of the nonparametric component is asymptotically minimax rate optimal. [27] introduced a first-order differencing estimator in a nonparametric regression model for estimating the variance of the random errors. [19, 26] extended the idea to higher-order differences for efficient estimation of the variance in such a setting. [22] used differencing for testing between a parametric model and a nonparametric alternative.

In particular, [38, 39] introduced the differencing method to semiparametric regression with the focus on estimating the linear component. By using higher-order differences [38, 39] showed that the bias induced from the presence of the nonparametric component can be essentially eliminated. He constructed an estimator of the linear component and showed it to be asymptotically efficient under the condition that the nonparametric function f is fixed (for all n) and has a bounded first derivative. See also [15, 23].

Main Results

In this paper, instead of focusing on the linear component as in [38, 39], we treat estimation of both the linear and the nonparametric components. We extend the

results in[38, 39] to general smoothness classes for the nonparametric component and the condition on nonparametric component is weakened. In addition, our results hold uniformly over such classes and so enable traditional asymptotic minimax conclusions. They also show what minimal smoothness assumptions are needed. Moreover, we consider the hypotheses testing problem of the linear coefficients and an F statistics is constructed. We show that asymptotic power of the F test is the same as if the nonparametric component is known. We also consider adaptive estimation of the nonparametric function f using wavelet thresholding. It is interesting to note that although the differences are correlated the correlation should be ignored and the linear regression coefficient vector β should be estimated by the ordinary least squares estimator instead of a generalized least squares estimator which takes into account the correlations among the differences. If the correlation structure is incorporated in the estimation, the resulting generalized least squares estimator will not be optimal (in most cases, even not consistent).

Estimation Procedure

The procedure begins by taking differences of the ordered observations (ordered according to the values of U_i). Let $d_t, t = 1, 2, \cdots, m+1$ be an order m difference sequence that satisfies $\sum_t d_t = 0$ and $\sum_t d_t^2 = 1$. For $i = 1, 2, \cdots, n-m$, let $D_i = \sum_{t=1}^{m+1} d_t Y_{i+m+1-t}$. Then D_i can be seen as the mth order difference of Y_i. The goal of this step is to eliminate the effect of the nonparametric component f. Now the problem reduces to the standard multiple linear regression problem. We then estimate the linear regression coefficients β by the ordinary least squares estimator based on the differences. Both the intercept a and unknown function f can be estimated based on the residual of the linear fit under certain identifiability assumptions.

We estimate the nonparametric function f by both kernel and wavelet thresholding methods. The results show that under certain conditions both the linear and nonparametric components are estimated as well as if the other component were known. We also derive a test for linear combinations of the regression coefficients of the linear component. The test is fully specified and the test statistic is shown to asymptotically have the usual F distribution under the null hypothesis.

Both the estimation and the testing procedures are easily implementable. Numerical performance of the estimation procedure is studied using both simulated and real data. The simulation results are consistent with the theoretical findings.

The paper is organized as follows. Section 2 considers the simpler case where X_i does not depend on U_i to illustrate the whole procedure. In Section 3 treats the general case where U_i are possibly correlated with the X_i and the main results are given. The testing problem is considered in Section 4. A simulation study is carried out in Section 5 to study the numerical performance of the

procedure. Real data applications are also discussed. The proofs are contained in Section 6.

2. Independent Case

In this section, we consider a simple version of the semiparametric partial linear model (1) where X_i does not depend on U_i. In section 3 we will consider the setting where X_i may depend on U_i. We shall always assume that X_i are random vectors. For the nonparametric component U_i, either $U_i = i/n$ or U_i are i.i.d. random variables on $[0, 1]$ and independent of X_i. In the second case, we also assume the density function of U_i is bounded away from 0. Assumptions on the function f are needed to make the model identifiable. Here we assume $\int_0^1 f(u)du = 0$ for the case where $U_i = i/n$; and assume $E(f(U_i)) = 0$ for the case where U_i are random variables.

Let $X_i = (X_{i1}, X_{i2}, \cdots .X_{ip})'$ be $p-$dimensional independent random vectors with a non-singular covariance matrix Σ_X. Define the Lipschitz ball $\Lambda^\alpha(M)$ in the usual way:

$$\begin{aligned}\Lambda^\alpha(M) &= \{g : \text{for all } 0 \le x, y \le 1,\ k = 0, ..., \lfloor\alpha\rfloor - 1, \\ &\qquad |g^{(k)}(x)| \le M, \text{ and } |g^{(\lfloor\alpha\rfloor)}(x) - g^{(\lfloor\alpha\rfloor)}(y)| \le M\,|x - y|^{\alpha'}\}\end{aligned}$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$. Suppose $f \in \Lambda^\alpha(M)$ for some $\alpha > 0$. Then the partial linear model (1) can be written as

$$Y_i = a + X_i'\beta + f(U_i) + \epsilon_i = a + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + f(U_i) + \epsilon_i \quad (2)$$

Here we assume the error terms ϵ_i, $i = 1, 2, \cdots, n$, are i.i.d. random variables with finite variance σ^2. The goal is to estimate the coefficient vector β, the intercept a, and the unknown function f. This will be done through a difference based estimation.

Suppose a difference sequence $d_1, d_2, \cdots, d_{m+1}$ satisfies $\sum_{i=1}^{m+1} d_i = 0$ and $\sum_{i=1}^{m+1} d_i^2 = 1$. Such a sequence is called an mth order difference sequence. Moreover, for $k = 1, 2, \cdots, m$ let $c_k = \sum_{i=1}^{m+1-k} d_i d_{i+k}$. Suppose

$$\sum_{k=1}^{m} c_k^2 = O(m^{-1}) \quad \text{as } m \to \infty. \quad (3)$$

One example of a sequence that satisfies these conditions is

$$d_1 = \sqrt{\frac{m}{m+1}}, d_2 = d_3 = \cdots = d_{m+1} = -\sqrt{\frac{1}{m(m+1)}}. \quad (4)$$

Remark 1 The asymptotic results in the theorems to follow require that the order $m \to \infty$ and that (3) be satisfied. However, even the simple choice of $m = 2$ seems to yield quite satisfactory performance as attested by the simulations in Section 5.

Remark 2 The asymptotic results like those in Theorem 1-5 are valid when X depends on n (say $X = X(n)$) under the condition that the multivariate sample CDF of $(X(n), U)$ converges to that which would occur as a limit in the setting of (3). We omit the details.

Remark 3 The case where U_i is multi-dimensional is much more involved than the one dimensional case since it is not easy to take difference. To use the difference based method in a high dimensional space, we need to carefully define the difference sequence $\{d_t\}$, see for example [4] and the references therein about the difference in high dimensional space. In this article, we only consider the one dimensional case.

We now consider the difference based estimator of β. Let $D_i = \sum_{t=1}^{m+1} d_t Y_{i+m+1-t}$, for $i = 1, 2, \cdots, n-m-1$. Then

$$D_i = Z_i'\beta + \delta_i + w_i, \quad i = 1, 2, \cdots, n-m-1, \tag{5}$$

where $Z_i = \sum_{t=1}^{m+1} d_t X_{i+m+1-t}$, $\delta_i = \sum_{t=1}^{m+1} d_t f(U_{i+m+1-t})$, and $w_i = \sum_{t=1}^{m+1} d_t \epsilon_{i+m+1-t}$. Written in matrix form, (5) becomes

$$D = Z\beta + \delta + w$$

where $D = (D_1, D_2, \cdots, D_{n-m-1})'$, $w = (w_1, w_2, \cdots, w_{n-m-1})'$ and Z is a matrix whose ith row is given by Z_i'. Let Ψ denote the $(n-m-1)\times(n-m-1)$ covariance matrix of w given by

$$\Psi_{i,j} = \begin{cases} 1 & \text{for } i = j \\ c_{|i-j|} & \text{for } 1 \le |i-j| \le m \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

In (5), δ_i are the errors related to the nonparametric component f in (1) and w_i are the random errors which are correlated, and have the covariance matrix $\Psi = (\Psi_{i,j})$ given by (6). For estimating the linear regression coefficient vector β, we use

$$\widehat{\beta} = (Z'Z)^{-1}Z'D. \tag{7}$$

Although not entirely intuitive, it is important in this step to ignore the correlation among the w_i and use the ordinary least squares estimate. If instead a generalized least squares estimator is used, i.e. $(Z'\Psi^{-1}Z)^{-1}Z'\Psi^{-1}D$, which incorporates the correlation structure, the optimality results in Theorem 1 below and Theorem 5 in the next section will not generally be valid.

Theorem 1 *Suppose $\alpha > 0$, $m \to \infty$, $m/n \to 0$, and that condition (3) is satisfied. Then the estimator $\widehat{\beta}$ given in (7) is asymptotically efficient, i.e.*

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{L} N(0,\ \sigma^2\Sigma_X^{-1}).$$

Remark 4 Using the generalized least squares method ($(Z'\Psi^{-1}Z)^{-1}Z'\Psi^{-1}D$) on the differences is similar to applying the ordinary least squares regression of Y on X in the original model (1). This would cause significant bias due to the presence of f. See Section 5 for numerical comparison.

Remark 5 Our proof shows that $\sqrt{n}(\hat{\beta}-\beta) \sim N(0, \sigma^2(1+2\sum_{k=1}^m c_k^2)\Sigma_X^{-1})$. This means condition (3) is necessary for this procedure to be asymptotically optimal. The factor $(1+2\sum_{k=1}^m c_k^2)$ describes the inefficiency that results from choice of a particular m and corresponding $\{c_1, ..., c_m\}$. It can perhaps best be recorded on a scale of relative values for the resulting standard deviations: rel.SD $= (1+2\sum_{k=1}^m c_k^2)^{-1/2}$. See Table 1 for a few such values for $\{c_k\}$ as in (4) and for the optimal $\{c_k\}$ of [19]. Note that even modest values of m yield quite high relative standard deviations.

m	1	2	4	8	16
$\{c_k\}$ from (4)	.816	.885	.933	.963	.980
Optimal $\{c_k\}$	.816	.894	.943	.970	.985

TABLE 1
Values of relative standard deviation for various m and $\{c_k\}$.

Remark 6 A similar result has been derived in [38, 39] under stronger conditions, where $\alpha \geq 1$.

A natural estimator of the intercept a is $\hat{a} = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\hat{\beta})$. Since $a = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\beta) - \frac{1}{n}\sum_{i=1}^n f(U_i)$, it follows that $\hat{a} - a = \frac{1}{n}\sum_{i=1}^n X_i'(\beta - \hat{\beta}) + \frac{1}{n}\sum_{i=1}^n f(U_i)$.

It can be seen that when $U_i = \frac{i}{n}$, $\frac{1}{n}\sum_{i=1}^n f(U_i) = O(n^{-\alpha})$. So when $\alpha > 1/2$ this term is negligible as compared with $\frac{1}{n}\sum_{i=1}^n X_i'(\beta - \hat{\beta})$. Therefore when $\alpha > 1/2$ the asymptotic property of $\hat{a}$ is purely driven by $\sum_{i=1}^n X_i'(\beta - \hat{\beta})$. Since $\hat{\beta}$ is an efficient estimator of β, $\hat{a}$ is also an efficient estimator of a. We thus have the following result.

Theorem 2 *When $U_i = i/n$ and $\alpha > 1/2$, $\hat{a}$ is an efficient estimator of a, i.e.*

$$\sqrt{n}(\hat{a} - a) \xrightarrow{L} N(0, \sigma^2).$$

Remark 7 For the case where $\alpha \leq 1/2$ and the U_i are i.i.d. random variables, it can be seen from the previous discussion that $\hat{a}$ is asymptotically normal, but the asymptotic variance may depend on f and the distribution of U_i.

Once we have the estimator $\hat{\beta}$ and $\hat{a}$, they can be plugged back into the original model (1) to remove the effect of the linear component. For $i = 1, 2, \cdots, n$, the residuals of the linear fit are

$$r_i = Y_i - \hat{a} - X_i'\widehat{\beta} = f(U_i) + a - \hat{a} + X_i'(\beta - \widehat{\beta}) + \epsilon_i.$$

The nonparametric component f can then be estimated by the Gasser-Mueller estimator based on r_i. Let $k(x)$ be a kernel function satisfying $\int k(x)dx =$

1 and has $\lfloor\alpha\rfloor$ vanishing moments. Take $h = n^{-1/(1+2\alpha)}$ and let $K_{i,h}(u) = \frac{1}{h}\int_{(U_i+U_{i-1})/2}^{(U_i+U_{i+1})/2} k(\frac{u}{h})du$ for $i = 1, 2, \cdots, n$. The estimator $\widehat{f}$ is then given by

$$\widehat{f}(u) = \sum_{i=1}^{n} K_{i,h}(u)r_i = \sum_{i=1}^{n} K_{i,h}(u)(Y_i - \hat{a} - X_i'\widehat{\beta}). \tag{8}$$

Theorem 3 *For each $\alpha > 0$, the estimator $\widehat{f}$ given in (8) satisfies*

$$\sup_{f\in\Lambda^\alpha(M)} E\left[\int(\widehat{f}(x) - f(x))^2dx\right] \le Cn^{-2\alpha/(1+2\alpha)}$$

for some constant $C > 0$. Moreover, for any $x_0 \in (0, 1)$,

$$\sup_{f\in\Lambda^\alpha(M)} E\left[(\widehat{f}(x_0) - f(x_0))^2\right] \le Cn^{-2\alpha/(1+2\alpha)}.$$

Theorem 3 is a standard results. It shows that the estimator $\hat{f}$ given in (8) attains the optimal rate of convergence over the Lipschitz ball $\Lambda^\alpha(M)$ under both the global and local losses for the semiparametric problem.

The kernel estimator constructed above enjoys desirable optimal rate properties. However, it relies on the assumption that the smoothness parameter α is given which is unrealistic in practice. It is thus important to construct estimators that automatically adapt to the smoothness of the unknown function f. We shall now introduce a wavelet thresholding procedure for adaptive estimation of the nonparametric component f.

Wavelet Thresholding Method

We work with an orthonormal wavelet basis generated by dilation and translation of a compactly supported mother wavelet ψ and a father wavelet ϕ with $\int\phi = 1$. A wavelet ψ is called *r-regular* if ψ has r vanishing moments and r continuous derivatives.

For simplicity in exposition, in the present paper we use periodized wavelet bases on $[0, 1]$. Let $\phi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty}\phi_{j,k}(x - l), \psi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty}\psi_{j,k}(x - l)$, for $t \in [0, 1]$. where $\phi_{j,k}(x) = 2^{j/2}\phi(2^jx - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx - k)$. The collection $\{\phi^p_{j_0,k},\ k = 1, \ldots, 2^{j_0};\ \psi^p_{j,k},\ j \ge j_0 \ge 0, k = 1, ..., 2^j\}$ is then an orthonormal basis of $L^2[0, 1]$, provided the primary resolution level j_0 is large enough to ensure that the support of the scaling functions and wavelets at level j_0 is not the whole of $[0, 1]$. The superscript "p" will be suppressed from the notation for convenience. An orthonormal wavelet basis has an associated orthogonal Discrete Wavelet Transform (DWT) which transforms sampled data into the wavelet coefficients. See [12, 34] for further details.

Wavelet thresholding methods have been well developed for nonparametric function estimation. One of the best known wavelet thresholding procedures is Donoho-Johnstone's VisuShrink ([13]). We shall now develop a wavelet thresholding procedure for the nonparametric component f in the semiparametric model similar to the VisuShrink for nonparametric regression.

Estimation of Nonparametric Component

For simplicity, here we suppose $n = 2^J$ for some integer J. The procedure begins by applying the discrete wavelet transformation to the residuals of the linear fit, $r = (r_1, r_2, \cdots, r_n)$. Let $v = n^{-\frac{1}{2}} W \cdot r$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Then v can be written as

$$v = (\tilde{v}_{j_0,1}, \cdots, \tilde{v}_{j_0,2^{j_0}}, v_{j_0,1}, \cdots, v_{j_0,2^{j_0}}, \cdots, v_{J-1,1}, \cdots, v_{J-1,2^{J-1}})' \qquad (9)$$

where $\widetilde{v}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $v_{j,k}$ ($j = j_0, \cdots, J-1, k = 1, \cdots, 2^j$) are empirical wavelet coefficients at level j which represent fine structure at scale 2^j. For convenience, we use (j,k) to denote the number $2^j + k$. Then the empirical wavelet coefficients can be written as

$$\tilde{v}_{j_0,k} = \xi_{j_0,k} + \tilde{\tau}_{j_0,k} + n^{-\frac{1}{2}} \tilde{z}_{j_0,k} \quad \text{and} \quad v_{j,k} = \theta_{j,k} + \tau_{j,k} + n^{-\frac{1}{2}} z_{j,k}.$$

where $\xi_{j_0,k}$ and $\theta_{j,k}$ are the wavelet coefficients of f, $\tau_{j,k}$ and $\tilde{\tau}_{j_0,k}$ denote combination of approximation error and the transformed linear residuals $n^{-\frac{1}{2}} W \cdot (X(\beta - \hat{\beta}) + a)$, and $z_{j,k}$ and $\tilde{z}_{j_0,k}$ are the transformed noise, i.e. $W \cdot \epsilon$. Our goal now is to estimate the wavelet coefficients $\xi_{j_0,k}$ and $\theta_{j,k}$.

For any y and $t \geq 0$, define the soft threshold function $\eta_t(y) = sgn(y)(|y| - t)_+$. Let J_1 be the largest integer satisfying $2^{J_1} \leq J^{-3} 2^J$, then the $\theta_{j,k}$ are estimated by

$$\hat{\theta}_{j,k} = \begin{cases} \eta_\lambda(v_{j,k}) & \text{if } j_0 \leq j \leq J_1 \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

where $\lambda = \sigma \sqrt{\frac{2 \log n}{n}}$. The coefficients of the father wavelets $\phi_{j_0,k}$ at the lowest resolution level are estimated by the corresponding empirical coefficients, $\hat{\xi}_{j_0,k} = \tilde{v}_{j_0,k}$. Write the estimated wavelet coefficients as

$$\hat{v} = (\hat{\xi}_{j_0,1}, \cdots, \hat{\xi}_{j_0,2^{j_0}}, \hat{\theta}_{j_0,1}, \cdots, \hat{\theta}_{j_0,2^{j_0}}, \cdots, \hat{\theta}_{J-1,1}, \cdots, \hat{\theta}_{J-1,2^{J-1}})'.$$

The estimate of f at the equally spaced sample points U_i is then obtained by applying the inverse discrete wavelet transform (IDWT) to the denoised wavelet coefficients. That is, $\{f(\frac{i}{n}) : i = 1, \cdots, n\}$ is estimated by $\widehat{f} = \{\widehat{f(\frac{i}{n})} : i = 1, \cdots, n\}$ with $\widehat{f} = n^{\frac{1}{2}} W^{-1} \cdot \hat{v}$. The estimate of the whole function f is given by

$$\widehat{f}(t) = \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t).$$

We have the following theorem.

Theorem 4 *Suppose the wavelet is r-regular and the moment generating function of ϵ_i exist in a neighborhood of the origin. Then for all $0 < \alpha \leq r$ the*

wavelet thresholding estimator $\hat{f}$ defined in (4) satisfies

$$\sup_{f\in\Lambda^\alpha(M)} E\left[\int(\widehat{f}(x)-f(x))^2dx\right] \le C\left(\frac{n}{\log n}\right)^{-2\alpha/(1+2\alpha)}$$

for some constant $C>0$. Moreover, for any $x_0\in(0,1)$,

$$\sup_{f\in\Lambda^\alpha(M)} E\left[(\widehat{f}(x_0)-f(x_0))^2\right] \le C\left(\frac{n}{\log n}\right)^{-2\alpha/(1+2\alpha)}.$$

Remark 8 Similar result for estimating the nonparametric component using wavelet thresholding method has been derived in [18]. In [18] the linear component and nonparametric component were estimated simultaneously but the estimation of the linear coefficients did not achieve the asymptotic efficiency.

Comparing the results in Theorem 4 with the minimax rate given in (3), the estimator $\hat{V}$ is adaptive to within a logarithmic factor of the minimax risk under both the global and local losses. Furthermore, it is not difficult to show that the extra logarithmic factor is necessary under the local loss. See, for example [2].

3. Dependence case

We now turn to the random design version of the partial linear model (1) where both X_i and U_i are assumed to be random and need not be independent of each other. Note that asymptotical efficiency in this setting has been discussed, for example, in [29]. Again let X_i be p dimensional random vectors. Let U_i be random variables on $[0,1]$ and suppose that (X_i', U_i), $i=1,...,n$, are independent with an unknown joint density function $g(x,u)$. Assume the ϵ_i are independent of (X_i', U_i). Let $h(U)=E(X|U)$ and $S(U)=E(X'X|U)$. Suppose $f(u)\in\Lambda^\alpha(M_f)$, and $h(u)\in\Lambda^\gamma(M_h)$ for some $\alpha>0$ and $\gamma>0$. (When X is a vector, we assume each coordinate of $h(u)$ satisfies this Lipschitz property.) Similar to the previous case, to make the model identifiable, assume $E(f(U_i))=0$. Moreover, suppose the marginal density of U is bounded away from 0, i.e. there exists a constant $c>0$ such that $\int g(x,u)dx\ge c$ for any $u\in[0,1]$.

Suppose $U_{(1)}\le U_{(2)}\le\cdots\le U_{(n)}$ are the order statistics of the U_i's and $X_{(i)}$ and $Y_{(i)}$ are the corresponding X and Y. Note that $X_{(i)}$'s are not the order statistics of X_i's, but the X associated with $U_{(i)}$. Similar to the independent case, we take the m-th order differences $D_i=\sum_{t=1}^{m+1} d_t Y_{i+m+1-t} = Z_i'\beta+\delta_i+w_i$, where $Z_i=\sum_{t=1}^{m+1} d_t X_{(i+t-1)}$, $\delta_i=\sum_{t=1}^{m+1} d_t f(U_{(i+t-1)})$, and $w_i=\sum_{t=1}^{m+1} d_t\epsilon_{(i+t-1)}$. Again we estimate the linear regression coefficient vector β by

$$\hat{\beta}=(Z^TZ)^{-1}Z^TD. \tag{11}$$

Theorem 5 *When $\alpha + \gamma > 1/2$ and $S(u) > 0$ for every u, the estimator $\widehat{\beta}$ is asymptotically efficient, i.e.*

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{L} N(0, \sigma^2 \Sigma_*^{-1}),$$

where $\Sigma_* = E[(X_1 - E(X_1|U_1))(X_1 - E(X_1|U_1))']$.

Remark 9 We can see from this theorem that we do not always need $\alpha > 1/2$ to ensure the asymptotic efficiency. We only need one of the two functions $f(u)$ and $h(U) = E(X|U)$ to have minimal smoothness. Theorem 1 can be considered to be a special case where γ is infinity.

Remark 10 [38] obtained similar results for the partial linear model (1) under the conditions that both f and h have bounded first derivatives and hence satisfy the conditions with $\alpha = 1$ and $\gamma = 1$. In this case the condition $\alpha + \gamma > 1/2$ of Theorem 5 is obviously satisfied.

When $\alpha > 1/2$ we can use the same procedure as in the previous section to efficiently estimate the intercept a. i.e. $\hat{a} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})$. Also, the asymptotic variance of $\hat{a}$ depends on the joint distribution of X and U.

Once we have an estimate of β, we can then use the same procedure to estimate $f(u)$ as in the fixed design case. Similarly, the estimator also attains the optimal rate of convergence over the Lipschitz ball $\Lambda^\alpha(M)$ under both the global and local losses.

The proof of Theorem 5 is given in Section 6. The following lemma is one of the main technical tools. It is useful in the development of the test given in Section 4.

Lemma 1 *Under the assumptions of Theorem 5, we have that as n goes to infinity,*

$$\frac{Z'Z}{n} \xrightarrow{D} \Sigma_*, \quad \frac{Z'\Psi Z}{n} \xrightarrow{D} (1 - \sum_{k=1}^{m} c_k^2)\Sigma_*, \text{ and } \frac{Z'\delta\delta'Z}{n} = O_p(n^{-2\alpha}) + O_p(n^{1-2\alpha-2\gamma})$$

where $\delta = (\delta_1, \delta_2, \cdots, \delta_{n-m})$ is given by $\delta_i = \sum_{t=1}^{m+1} d_t f(U_{(i+t-1)})$.

4. Testing the Linear Component

In this section, we consider the problem of testing the null hypothesis that the linear regression coefficients satisfy certain linear constraints. That is, we wish to test

$$H_0 : C\beta = 0 \quad \text{against} \quad H_1 : C\beta \neq 0,$$

where C is an $r \times p$ matrix with $\text{rank}(C) = r$. A special case is testing the hypothesis $H_0 : \beta_{i_1} = \cdots = \beta_{i_r} = 0$. In this section, we shall assume the errors ϵ_i are independent and identically distributed $N(0, \sigma^2)$ variables.

4.1. Fixed Design or Independent Case

We divide the testing problem into two cases. We first consider the case where $U_i = i/n$ (fixed design) or the U_i's are random but independent of the X_i's. From the previous sections, we know that asymptotically in this case the estimator $\hat{\beta}$ of the linear regression coefficient vector β satisfies $\sqrt{n}(\hat{\beta}-\beta) \sim N(0, \sigma^2\Sigma_X^{-1})$. This means asymptotically $\sqrt{n}(C\hat{\beta} - C\beta) \sim N(0, \sigma^2 C\Sigma_X^{-1}C')$. So our test statistic will be based on $\frac{n}{\sigma^2}\hat{\beta}'C'(C\Sigma_X^{-1}C')^{-1}C\hat{\beta}$. Both the covariance matrix Σ_X and the error variance σ^2 are unknown in general and thus need to be estimated. It follows from Lemma 1 that $Z'Z/n \xrightarrow{a.s.} \Sigma_X$ in this case. If σ^2 is given, the test statistic $\frac{1}{\sigma^2}\hat{\beta}'C'(C(Z'Z)^{-1}C')^{-1}C\hat{\beta}$ has a limiting χ^2 distribution with r degrees of freedom.

To estimate the error variance σ^2, set $H = I - Z(Z'Z)^{-1}Z'$. We shall use $\hat{\sigma}^2 = \frac{\|HD\|_2^2}{n-m-p}$ to estimate σ^2. Note that $\|HD\|_2^2 = w'Hw + 2w'H\delta + \delta'H\delta$. Suppose $\alpha > 1/2$. Then it is easy to see that $\delta'H\delta \xrightarrow{a.s.} 0$ as $n \to \infty$. Since $w'H\delta|\delta \sim N(0, 2\sigma^2\delta'H\delta)$, we know that $w'H\delta \xrightarrow{a.s.} 0$. Here we also assume that the first term of the difference sequence satisfies that $1 - d_1^2 = O(m^{-1})$ (the sequence given in (4) satisfies this condition). It can be shown that $\sigma^{-2}w'Hw$ is approximately distributed as chi-squared with $n - m - p$ degrees of freedom.

Theorem 6 *Suppose $\alpha > 1/2$ and $1 - d_1^2 = O(m^{-1})$. For testing $H_0 : C\beta = 0$ against $H_1 : C\beta \neq 0$, where C is an $r \times p$ matrix with $\mathrm{rank}(C) = r$, the test statistic*

$$F = \frac{\hat{\beta}'C'(C(Z'Z)^{-1}C')^{-1}C\hat{\beta}/r}{\hat{\sigma}^2}$$

asymptotically follows the $F(r, n-m-p)$ distribution under the null hypothesis. Moreover, the asymptotic power of this test (at local alternatives) is the same as the usual F test when f is not present in the model (1).

Hence an approximate level α test of $H_0 : C\beta = 0$ against $H_1 : C\beta \neq 0$ is to reject the null hypothesis H_0 when the test statistic $F \geq F_{r,n-m-p;\alpha}$ where $F_{r,n-m-p;\alpha}$ is the α quantile of the $F(r, n-m-p)$ distribution.

Remark 11 [39] considered the testing problem. A χ^2 statistic was derived under the condition that σ^2 is known.

4.2. General Random Design Case

We now turn to the test problem in the general random design case where U_i are random and correlated with X_i. Again suppose that (X_i', U_i), $i = 1, ..., n$, are independent with an unknown joint density function $g(x, u)$. We will show that the same F test also works in this case. Notice that in this case, asymptotically

$$\sqrt{n}(C\hat{\beta} - C\beta) \xrightarrow{L} N(0, \sigma^2 C\Sigma_*^{-1}C'),$$

where $\Sigma_* = E[(X_1 - E(X_1|U_1))(X_1 - E(X_1|U_1))']$. Lemma 1 shows that $Z'Z/n$ converges to Σ_*. Based on this observation and the discussion given in Section 4.1, we have the following theorem.

Theorem 7 *Suppose $\alpha > 1/2$ and $1 - d_1^2 = O(m^{-1})$. For testing $H_0 : C\beta = 0$ against $H_1 : C\beta \neq 0$, where C is an $r \times p$ matrix with $\mathrm{rank}(C) = r$, the test statistic*

$$F = \frac{\hat{\beta}' C' (C(Z'Z)^{-1}C')^{-1} C\hat{\beta}/r}{\hat{\sigma}^2}$$

asymptotically follows the $F(r, n-m-p)$ distribution under the null hypothesis.

5. Numerical Study

The difference based procedure for estimating the linear coefficients and the unknown function introduced in the previous sections is easily implementable. In this section we investigate the numerical performance of the estimator using both simulations and analysis of real data.

5.1. Simulation

We first study the effect of the unknown function f on the estimation accuracy of the linear component and then investigate the effect of the order of the difference sequence. In the first simulation study, we take $n = 500$, $U_i \overset{iid}{\sim} \mathrm{Uniform}(0, 1)$, $a = 0$ and consider the following four different functions,

$$f_1(x) = \begin{cases} 3 - 30x & \text{for } 0 \le x \le 0.1 \\ 20x - 1 & \text{for } 0.1 \le x \le 0.25 \\ 4 + (1 - 4x)18/19 & \text{for } 0.25 < x \le 0.725 \\ 2.2 + 10(x - 0.725) & \text{for } 0.725 < x \le 0.89 \\ 3.85 - 85(x - 0.89)/11 & \text{for } 0.89 < x \le 1 \end{cases}$$

$f_2(x) = 1 + 4(e^{-550(x-0.2)^2} + e^{-200(x-0.8)^2} + e^{-950(x-0.8)^2})$ and $f_3(x) = \sum h_j(1 + |\frac{x-x_j}{w_j}|)^{-4}$, where

$$\begin{aligned} (x_j) &= (0.10, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81), \\ (h_j) &= (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2), \\ (w_j) &= (0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005). \end{aligned}$$

And $f_4(x) = \sqrt{x(1-x)}\sin(\frac{2.1\pi}{x+0.05})$. The test functions f_3 and f_4 are the Bumps and Doppler functions given in [13]. When we do simulation, we will normalize these functions to make them have unit variance. We also consider the case where $f \equiv 0$ for comparison. The errors ϵ_i are generated from the standard normal distribution. For X_i and β, we consider two cases: Case (1). $p = 1$, $X_i \sim N(4U_i, 1)$, $\beta = 1$; Case (2). $p = 3$, $X_i \sim N((U_i, 2U_i, 4U_i^2), I_3)$, $\beta = (2, 2, 4)'$ where I_3 denotes the 3×3 identity matrix.

We first examine the effect of the unknown function f on the estimation of the linear component. In this part, the difference sequence in equation (4) with $m = 2$ is used. The mean squared errors (MSEs) of the estimator $\hat{\beta}$ is calculated over 200 simulation runs. We also consider the case where the presence of f is completely ignored and we directly run least squares regression of Y on X in model (1). The results are summarized in Table 2. The numbers insides the parentheses are the MSEs of the estimate when the nonparametric component is ignored. By comparing the MSEs in each row, it can be easily seen that we can estimate the linear coefficients nearly as well as if f were known. On the other hand, if f is simply ignored and β is estimated by applying the least squares regression of Y on X directly, the estimator is highly inaccurate. The mean squared errors are between 2 to over 600 times as large as those of the corresponding estimators based on the differences.

	$f \equiv 0$	f_1 (ignored)	f_2 (ignored)	f_3 (ignored)	f_4 (ignored)
Case (1)	0.0028	0.0028 (1.970)	0.0028 (0.054)	0.0034 (0.013)	0.0033 (0.011)
Case (2)	0.0027	0.0023 (0.705)	0.0023 (0.025)	0.0037 (0.009)	0.0032 (0.007)

TABLE 2

The MSEs of estimate $\hat{\beta}$ over 200 replications with sample size $n = 500$. The numbers insides the parentheses are the MSEs of the estimate when the nonparametric component is ignored.

For estimating the nonparametric function f, we use a kernel method with the Parzen's kernel. The bandwidth was selected by cross validation, see for example [21, 30]. For comparison, we also carried out the simulation in the case where $\beta = 0$. The mean squared error of the estimated f is summarized in Table 3. It can be seen that the MSEs in each column are close to each other and hence the performance of our estimator $\hat{f}$ does not depend sensitively on the structure of X and β.

	$f \equiv 0$	f_1	f_2	f_3	f_4
$\beta = 0$	0.02778	0.09201	0.22323	0.76808	0.35976
Case (1)	0.03199	0.10286	0.25008	0.78666	0.35950
Case (2)	0.02940	0.09765	0.25199	0.80667	0.37725

TABLE 3

The MSEs of the estimate $\hat{f}$ over 200 replications with sample size $n = 500$.

We now consider the effect of the order of the difference sequence m on the estimation accuracy. In this study, different combinations of the function f and the Cases (1) and (2) yield basically the same results. As an illustration of this, we focus on Case (2) and $f = f_2$. We compare four different values of m: 2, 4, 8, 16. The difference sequence in equation (4) was used in each case. We summarize in Table 4 the mean and standard deviation of the estimate $\hat{\beta}$ and the average MSE of the estimate $\hat{f}$. By comparing the means and standard deviations in each row we can see that the performance of the estimator does not depend significantly on m.

	$m = 2$	$m = 4$	$m = 8$	$m = 16$
Mean(sd) of $\hat{\beta}_1$	2.003(0.056)	2.007(0.053)	1.996(0.049)	1.996(0.051)
Mean(sd) of $\hat{\beta}_2$	2.001(0.051)	2.002(0.051)	2.0006(0.048)	1.999(0.050)
Mean(sd) of $\hat{\beta}_3$	4.001(0.055)	4.002(0.050)	3.999(0.045)	3.989(0.049)
MSE of $\hat{f}$	0.2540	0.2378	0.2400	0.2483

TABLE 4
The mean and standard deviation of the estimate $\hat{\beta}$ and the average MSEs of the estimate $\hat{f}$ over 200 replications with sample size $n = 500$.

Next, we consider the test of linear coefficient. In this study, we focus on case (2) with two different sets of linear coefficients. One of them is $\beta = (2, 2, 4)'$, the other one is $\beta = (0, 0, 4)'$. The hypothesis that will be tested is $H_0 : \beta_1 = \beta_2 = 0$. The total number of rejects (at level 0.05) over 200 runs and the mean value of F statistics are summarized in Table 5. We also compare the F statistics with its nominal distribution for the case $\beta = (0, 0, 4)'$ and $f = f_2$. The empirical cumulative distribution function and the quantile-quantile plot are plotted in figure 1. It can be seen that the F statistics fit the distribution very well and F test performs as if the nonparametric component is known.

	$f \equiv 0$	f_1	f_2	f_3	f_4
$\beta = (0, 0, 4)'$	12 (1.1608)	13 (1.2680)	14 (1.2284)	18 (1.1414)	14 (1.2317)
$\beta = (2, 2, 4)'$	200 (4021.3)	200 (3973.8)	200 (3891.0)	200 (2874.8)	200 (3714.5)

TABLE 5
The total number of rejects of F test over 200 replications with sample size $n = 500$ at level 0.05. The numbers insides the parentheses are the mean value of F statistics.

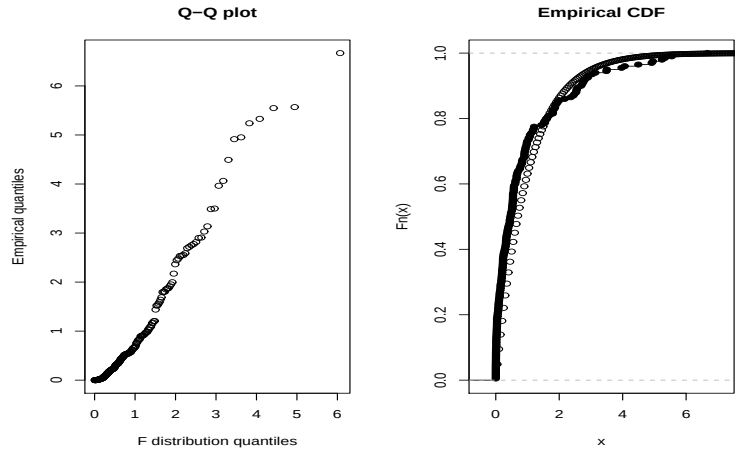

FIG 1. *QQ-plot and the plot of empirical cdf of the F statistics. On the right plot, the dot line is the plot of the true cdf.*

5.2. Application to Attitude Data

We now apply our estimation and testing procedures to the analysis of the attitude data. This data set was first analyzed in [8] using multiple linear regression and variable selection. This data set was from a study of the performance of supervisors and was collected from a survey of the clerical employees of a large financial organization. This survey was designed to measure the overall performance of a supervisor, as well as questions that related to specific characteristic of the supervisor. The numbers give the percent proportion of favorable responses to seven questions in each department. Seven variables, Y (over all rating of the job being done by supervisor), X_1 (raises based on performance), X_2 (handle employee complaints), X_3 (does not allow special privileges), X_4 (opportunity to learn new things), X_5 (rate of advancing to better job), and U (too critical to poor performances) are considered here. The goal is to understand the effect of variables ($X_1, ..., X_5$ and U) on Rating (Y). Figure 2 plots each independent variable against the response Y. We can see that the effect of U on Y is not linear, while the effect of other variables are roughly linear. So we employ the following model,

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + f(U) + \epsilon. \tag{12}$$

Using the estimation procedure discussed in Section 3 with $m = 2$, the linear

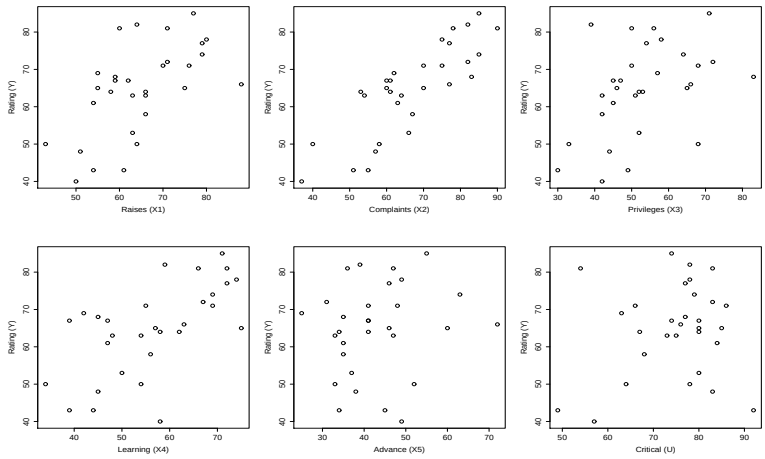

FIG 2. *Plot s of the individual explanatory variables against the response variable.*

component in the model (12) is estimated as $18.1127 - 0.0208X_1 + 0.6130X_2 - 0.1207X_3 + 0.5043X_4 - 0.3747X_5$. The F statistic and the p value for testing each coefficient $H_{i0} : \beta_i = 0$ against $H_{i1} : \beta_i \neq 0$ are given in Table 6. The p-values for β_1, β_3 and β_5 are exceedingly large.

	Estimated coefficient	F statistic	p value
X_1	-0.0208	0.0026	0.9597
X_2	0.6130	3.9270	0.0600
X_3	-0.1207	0.2142	0.6478
X_4	0.5043	2.5216	0.1259
X_5	-0.3747	1.2175	0.2813

TABLE 6
The estimated coefficients of the linear component and the significance tests.

We thus perform the simultaneous F test to test the hypothesis $H_0 : \beta_1 = \beta_3 = \beta_5 = 0$ against H_1 : at least one of them is nonzero. The value of the F statistic is 2.1577 and the p value is 0.1206. In comparison, the value of the F statistic for the global hypothesis $H_0 : \beta_1 = \cdots = \beta_5 = 0$ is 18.4038 and the p value is less than 0.0001. The results show that we fail to reject the hypothesis $H_0 : \beta_1 = \beta_3 = \beta_5 = 0$. We can thus refine the linear component by using only Learning (X_2) and Complaints (X_4) as independent variables. In this case, the estimated linear component is $16.3467 + 0.6725X_2 + 0.2068X_4$. The F value for this model is 34.3635 and p value is less than 0.0001.

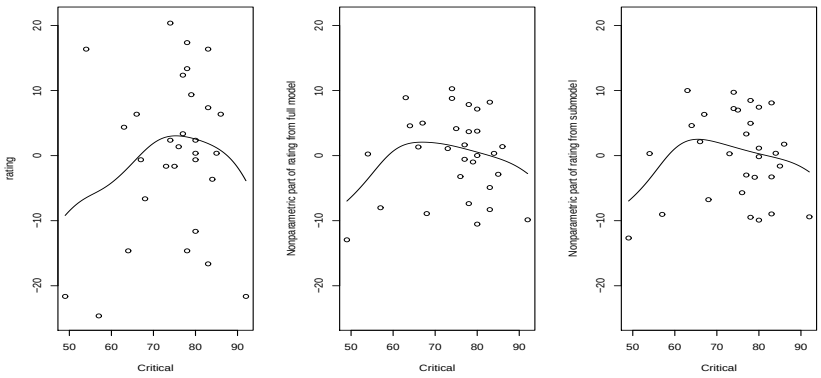

FIG 3. *Kernel estimates of the nonparametric component f. The points are the residuals of respective linear fits.*

We can then estimate the nonparametric component of the effect of Critical (U). For this, we run kernel estimation using the residuals of the linear fits as we did in Section 5.1. Figure 3 shows the nonparametric fits. The left panel plots the estimate of f under the model (12) but we ignore the linear component, the middle panel plots the estimate of f under the model (12) with all linear variables and the right panel plots $\hat{f}$ with the variables X_2 and X_4 in the linear part. We can see that the plot on the left panel is quite different from the other two. And the two plots on middle and right are similar since including a small number of additional non-significant variables does not have a large effect on the

estimates of the remaining parts of the model. Moreover, we test the significance of the nonparametric function, i.e. H_0: $f(u) = a+bu$ for some constants a, b. We follow the test procedure described in [17]. The p-value of the likelihood ratio test is 0.043, which shows the nonparametric function is significant. Actually, we have significant result with p-value 0.0259 when we fit a quadratic function to the nonparametric component.

Note that in [8], the standard multiple linear regression was used to model the relationship between the response and the explanatory variables. The linear model failed to detect the relation of the variable U and Y, and it concluded that variable U did not have significant effect on Y.

6. Proofs

We shall prove Theorems 1, 3, 4, ,5 and 6. The proof of Theorem 7 is similar to that of Theorem 3 and Theorem 6, respectively. We will first prove some technical lemmas.

6.1. Technical Lemmas

Lemma 2 *Under the assumptions of Theorem 1,*

$$\sqrt{n}(Z'Z)^{-1}Z'w \xrightarrow{L} N(0, (1+O(m^{-1}))\sigma^2\Sigma_X^{-1}).$$

Proof. The asymptotic normality of $\sqrt{n}(Z'Z)^{-1}Z'w$ follows from the Central Limit Theorem and the fact that $\sum_{k=1}^{m} c_k^2 = O(m^{-1})$. It is known that $E(\sqrt{n}(Z'Z)^{-1}Z'w) = 0$ and $Var((Z'Z)^{-1}Z'w|Z) = (Z'Z)^{-1}Z'\Psi Z(Z'Z)^{-1}$. Note that with $m = o(n)$, so $\frac{1}{n}(Z'Z) \xrightarrow{a.s.} E[(\sum_{t=1}^{m+1} d_t X_{i+m+1-t})(\sum_{t=1}^{m+1} d_t X'_{i+m+1-t})] = \Sigma_X$. Also with $m = o(n)$, for any $k = 1, 2, \cdots, m$,

$$\frac{1}{n}\sum_{i=1}^{n-m-1-k}(\sum_{t=1}^{m+1} d_t X_{i+m+1-t})(\sum_{t=1}^{m+1} d_t X'_{i+k+m+1-t}) \xrightarrow{a.s.} c_k\Sigma_X.$$

This implies $nVar((Z'Z)^{-1}Z'w|Z) \xrightarrow{a.s.} \sigma^2\Sigma_X^{-1}\Sigma_X(1+2\sum_{k=1}^{m} c_k^2)\Sigma_X^{-1} = (1+O(\frac{1}{m}))\sigma^2\Sigma_X^{-1}$. So $\sqrt{n}(Z'Z)^{-1}Z'w \xrightarrow{L} N(0, (1+O(\frac{1}{m}))\sigma^2\Sigma_X^{-1})$. ■

Lemma 3 *Under the assumptions of Theorem 1,*

$$nE\left[((Z'Z)^{-1}Z'\delta)((Z'Z)^{-1}Z'\delta)'\right] = O(\Big(\frac{m}{n}\Big)^{2(\alpha\wedge 1)})\,(\Sigma_X)^{-1}$$

Proof.

Note that $E(\sum_{i=1}^{n-m-1}(\sum_{t=1}^{m+1} d_t X'_{i+m+1-t})\delta_i) = 0$, so $E\{(Z'Z)^{-1}Z'\delta\} = 0$. Now

$$E\left[(\sum_{i=1}^{n-m-1} Z_i\delta_i)(\sum_{i=1}^{n-m-1} Z_i'\delta_i)\right] = (\sum_{i=1}^{n-m-1}\delta_i^2 - c_k\sum_{j=1}^{n-m-2}\delta_j\sum_{l=1}^{m}\delta_{j+m})\Sigma_X.$$

When $m = o(n)$, since $f \in \Lambda^\alpha(M)$, $|\delta_i| < M(\frac{m}{n})^{(\alpha\wedge 1)}$. So

$$\left|\sum\nolimits_{i=1}^{n-m-1} \delta_i^2 - \frac{1}{m}\sum\nolimits_{j=1}^{n-m-2}\delta_j \sum\nolimits_{l=1}^{m}\delta_{j+m}\right| = O(n^{1-2(\alpha\wedge 1)}m^{2(\alpha\wedge 1)}).$$

Also we know that $\frac{1}{n}Z'Z \xrightarrow{a.s.} \Sigma_X$, therefore, as $n \to \infty$,

$$nVar((Z'Z)^{-1}Z'\delta) = \frac{nO(n^{1-2(\alpha\wedge 1)}m^{2(\alpha\wedge 1)})}{n^2}(\Sigma_X)^{-1} = O(\Big(\frac{m}{n}\Big)^{2(\alpha\wedge 1)})(\Sigma_X)^{-1}.$$

■

The following lemma bounds the difference between the DWT of a sampled function and the true wavelet coefficients. See, for example, [3].

Lemma 4 *Let $\xi_{J,k} = \langle f, \phi_{J,k}\rangle$ and $n = 2^J$. Then for some constant $C > 0$,*

$$\sup_{V\in\Lambda^\beta(M)}\sum_{k=1}^{n}(\xi_{J,k} - n^{-\frac{1}{2}}V(\frac{k}{n}))^2 \le Cn^{-(2\alpha\wedge 1)}.$$

The following lemma is from [5].

Lemma 5 *Let $y = \theta + Z$, where θ is an unknown parameter and Z is a random variable with $EZ = 0$. Then*

$$E(\eta(y,\lambda) - \theta)^2 \le \theta^2 \wedge (4\lambda^2) + 2E(Z^2 I(|Z| > \lambda)).$$

Lemma 6 *Under the assumptions of Theorem 5, $Var(\frac{1}{n}\sum_{i=1}^{n-1} X_{(i)}X'_{(i+1)}) \to 0$ as $n \to \infty$. Here the variance and the limitation are both entry-wise.*

Proof. First we have

$$\begin{aligned} & Var(\frac{1}{n}\sum_{i=1}^{n-1} X_{(i)}X'_{(i+1)}) \\ = \; & \frac{1}{n^2}\sum_{i=1}^{n-1} Var(X_{(i)}X'_{(i+1)}) + \frac{2}{n^2}\sum_{i=1}^{n-2} Cov(X_{(i)}X'_{(i+1)}, X_{(i+1)}X'_{(i+2)}) \\ & + \frac{2}{n^2}\sum_{i+1<j} Cov(X_{(i)}X'_{(i+1)}, X_{(j)}X'_{(j+1)}). \end{aligned}$$

Here the covariance of two matrix means the covariance of corresponding entries. Let $\eta_i = h(U_{(i+1)}) - h(U_{(i)})$ and $H_i = h'(U_{(i)})h(U_{(i)})$ for $i = 1, 2, \cdots, n-1$. Note that when $\gamma > 0$, $\eta_i = O(n^{-\gamma})$. Then for $i+1 < j$, $Cov(X_{(i)}X'_{(i+1)}, X_{(j)}X'_{(j+1)}) = Cov(H_i, H_j) + O(n^{-\gamma})$. Hence,

$$\frac{2}{n^2}\sum_{i+1<j} Cov(X_{(i)}X'_{(i+1)}, X_{(j)}X'_{(j+1)}) = Var(\frac{1}{n}\sum_{i=1}^{n} H_i) + O(n^{-\gamma}).$$

Also, it is easy to see that $\frac{1}{n^2}\sum_{i=1}^{n-1} Var(X_{(i)}X'_{(i+1)}) = O(\frac{1}{n})$ and

$$\frac{2}{n^2}\sum_{i=1}^{n-2} Cov(X_{(i)}X'_{(i+1)}, X_{(i+1)}X'_{(i+2)}) = O(\frac{1}{n}).$$

Putting these together, the lemma is proved. ■

Remark 12 By the same calculation, we actually can prove that for any fixed integer $k > 0$, $Var(\frac{1}{n}\sum_{i=1}^{n-k} X_{(i)}X'_{(i+k)})$ goes to zero as n goes to infinity.

6.2. Proof of Lemma 1.

It follows from Lemma 6 and the fact $m = o(n)$ that

$$\lim_{n\to\infty} Var(\frac{Z'Z}{n}) = \lim_{n\to\infty} Var(\frac{Z'\Psi Z}{n}) = \lim_{n\to\infty} Var(\frac{Z'\delta\delta'Z}{n}) = 0.$$

So we only need to check the limit of the expectation. First note that

$$E(Z_iZ'_i) = \sum_{t=1}^{m+1} d_t^2 E(Var(X_{(i+t-1)}|U)) + [\sum_{t=1}^{m+1} d_t h(U_{(i+t-1)})]'[\sum_{t=1}^{m+1} d_t h(U_{(i+t-1)})],$$

$$E(Z_iZ'_{i+j}) = \sum_{t=1}^{m+1-j} d_{t+j}d_t E(Var(X_{(i+j+t-1)}|U)) + [\sum_{t=1}^{m+1} d_t h(U_{(i+t-1)})]'[\sum_{t=1}^{m+1} d_t h(U_{(i+j+t-1)})].$$

This implies

$$\lim_{n\to\infty} E(\frac{Z'Z}{n}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-m-1} E(Z_iZ'_i) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E[Var(X_i|U)] = \Sigma_*.$$

Here we use the fact that, since $h(U)$ has $\gamma > 0$ derivatives. Similarly, $\lim_{n\to\infty} \frac{1}{n}E(Z'\Psi Z) = (1 - \Sigma c_k^2)\Sigma_*$. Finally for the third equation, let $dh_i = \sum_{t=1}^{m+1} d_t h(U_{(i+t)})$

$$\begin{aligned}\frac{1}{n}E(Z'\delta\delta'Z) &= \frac{1}{n}\sum_{i=1}^{n} E[Var(X'_{(i)}|U)(\sum_t d_t^2\delta_{i-t}^2 + \sum_{t=1}^{m+1}\sum_{k=1}^{m+1} d_t d_{t+k}\delta_{i-t}\delta_{i+k-t})] \\ &\quad + \frac{1}{n}[\sum_{i=1}^{n}\delta_i E(dh_i)]'[\sum_{i=1}^{n}\delta_i E(dh_i)]\end{aligned}$$

Since $\sum_t d_t^2\delta_{i-t}^2 + \sum_{t=1}^{m+1}\sum_{k=1}^{m+1} d_t d_{t+k}\delta_{i-t}\delta_{i+k-t}$ is of order $n^{-2\alpha}$ for any i, the first part of the above expression is of order $n^{-2\alpha}$. And $\delta_i E(dh_i)$ is of order $n^{-\alpha-\gamma}$ for any i, so the second part of the above expression is of order $n^{1-2\alpha-2\gamma}$. This implies

$$\frac{1}{n}E(Z'\delta\delta'Z) = O(n^{-2\alpha}) + O(n^{1-2\alpha-2\gamma}). \quad ∎$$

6.3. Proofs of Theorems

Theorem 1 now follows from Lemmas 2 and 3. Note that $\sqrt{n}(\widehat{\beta}-\beta) = \sqrt{n}(Z'Z)^{-1}Z'(w+\delta)$. Lemma 3 implies $\sqrt{n}(Z'Z)^{-1}Z'\delta \xrightarrow{P} 0$. This together with Lemma 2 yield Theorem 1. ∎

For Theorem 3 ,we shall only prove the convergence rate under the pointwise squared error loss, the rate of convergence under the global mean integrated squared error risk can be derived using a similar line of argument.

Let $\hat{f}_1(u) = \sum_{i=1}^n K_{i,h}(u)(f(U_i)+\epsilon_i)$ and $\hat{f}_2(u) = \sum_{i=1}^n K_{i,h}(u)X_i(\beta - \hat{\beta}) + \hat{a} - a$, then $\hat{f}(u) = \hat{f}_1(u) + \hat{f}_2(u)$. In $\hat{f}_1$ there is no linear component. From the standard nonparametric regression results we know that for any x_0, $\sup_{f\in\Lambda^\alpha(M)} E\left[(\hat{f}_1(x_0) - f(x_0))^2\right] \le Cn^{-2\alpha/(1+2\alpha)}$ for some constant $C > 0$. Note that $\sum_{i=1}^n K_{i,h}^2(u) = O(\frac{1}{nh}) = O(n^{-2\alpha/(1+2\alpha)})$. So

$$E(\hat{f}_2(x_0)^2) = E\left[(\sum_{i=1}^n K_{i,h}X_i(\beta-\hat{\beta}))^2\right] \le \sum_{i=1}^n K_{i,h}^2 E(X_i(\beta-\hat{\beta}))^2 = O(n^{-2\alpha/(1+2\alpha)})$$

Hence the Theorem is proved. ∎

For Theorem 4, note that the maximum approximation error is of order $n^{-(2\alpha\wedge 1)}$ and is negligible relative to the minimax risk in Theorem 4. Now we shall only prove the upper bound for the integrated squared error. The case of local pointwise error is similar. Note that

$$E\left[\int(\widehat{f}(x)-f(x))^2dx\right] = E\sum_k(\hat{\xi}_{j_0,k}-\xi_{j_0,k})^2 + E\sum_{j=j_0}^{J_1}\sum_k(\hat{\theta}_{j,k}-\theta_{j,k})^2 + \sum_{j>J_1}\sum_k\theta_{j,k}^2. \tag{13}$$

Note that for $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$, there exists a constant $C > 0$ such that for all $j \ge j_0, 1 \le k \le 2^j$

$$\sup_{f\in\Lambda^\alpha(M)} |\theta_{j,k}| \le C2^{-j(\alpha+1/2)}. \tag{14}$$

See [12]. Hence $\sup_{f\in\Lambda^\alpha(M)}\sum_{j>J_1}\sum_k\theta_{j,k}^2 \le C2^{-J_1 2\alpha} = o(n^{-2\alpha/(1+\alpha)})$.

This means the third term in equation (13) is negligible. So we just need to focus on the first two terms. We know that $E(\hat{\xi}_{j_0,k} - \xi_{j_0,k})^2 \le 2E(\tilde{\tau}_{j_0,k}^2) + \frac{2}{n}E(\tilde{z}_{j_0,k}^2)$ and $E(\hat{\theta}_{j,k} - \theta_{j,k})^2 \le 2E(\tau_{j,k}^2) + 2E(\eta_\lambda(\theta_{j,k} + n^{-1/2}z_{j,k}) - \theta_{j,k})^2$. Putting them together, we have

$$\begin{aligned} & E[\int(\widehat{f}(x)-f(x))^2dx] \\ = \; & 2E\sum_{j,k}(\eta_\lambda(\theta_{j,k}+n^{-1/2}z_{j,k})-\theta_{j,k})^2 + \frac{2}{n}E\sum_k(\tilde{z}_{j_0,k}^2) \\ & +2E(\|X(\hat{\beta}-\beta)+(\hat{a}-a)\cdot 1_n\|_2^2). \end{aligned}$$

where 1_n denotes the n dimensional column vector of 1. From Theorem 1 and Theorem 5, we know that $E(\|X(\hat{\beta}-\beta)\|_2^2) = \frac{1}{n}E(X\Sigma^{-1}X') = O(\frac{1}{n})$ and $E(\hat{a}-a)^2 = O(n^{-1}\wedge n^{-2\alpha})$. This means the third term is negligible relative to the minimax rate. Also we know that there are only a fixed number of terms in the sum of the second term, so this term is also negligible. From now on, we just need to focus on the first term.

From lemma 5, $E\sum_{j,k}(\eta_\lambda(\theta_{j,k}+n^{-1/2}z_{j,k})-\theta_{j,k})^2 \le \sum\theta_{j,k}^2\wedge(4\lambda^2) + \sum\frac{2}{n}E(z_{j,k}^2 I(|z_{j,k}|\ge n^{1/2}\lambda)) \triangleq S_1+S_2$. By the same argument as the proof of Theorem 1 in [5], we can show that S_2 is negligible as compared to results of the Theorem 4.

For S_1, it can be seen that when $j \ge \frac{J-\log_2 J}{2\alpha+1}$, equation (14) yields $\theta_{j,k}^2\wedge(4\lambda^2) \le \theta_{j,k}^2 \le C2^{-j(1+2\alpha)}$ and when $j \le \frac{J-\log_2 J}{2\alpha+1}$, $\theta_{j,k}^2\wedge(4\lambda^2) \le 4\lambda^2 \le C\frac{\log n}{n}$ for some constant $C>0$. This means $S_1 \le \sum_{j\le\frac{J-\log_2 J}{2\alpha+1}} C2^j(\frac{\log n}{n}) + \sum_{j>\frac{J-\log_2 J}{2\alpha+1}} C2^{-2j\alpha} \le C(\frac{\log n}{n})^{2\alpha/(1+2\alpha)}$. Hence the Theorem is proved. ∎

Now we will prove Theorem 5. With Lemmas 1 and 2, the proof of Theorem 5 is now straightforward. Note that $\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}(Z'Z)^{-1}Z'w + \sqrt{n}(Z'Z)^{-1}Z'\delta$. It follows from the same argument as in the proof of Lemma 2 that $\sqrt{n}(Z'Z)^{-1}Z'w \sim N(0, n(Z'Z)^{-1}Z'\Phi Z(Z'Z)^{-1})$. The first two equalities of Lemma 1 yields

$$\lim_{n\to\infty} n(Z'Z)^{-1}Z'\Phi Z(Z'Z)^{-1}) = \{\Sigma_*\}^{-1}$$

and the third equality of Lemma 1 shows that, when $\alpha+\gamma > 1/2$, $\sqrt{n}(Z'Z)^{-1}Z'\delta$ is small and negligible. Theorem 5 now follows. ∎

Next we will prove Theorem 6. Let L be a $(n-m)\times n$ matrix given by

$$\mathrm{L}_{i,j} = \begin{cases} d_{j-i+1} & \text{for } 0\le j-i\le m \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

Moreover, let J be another $(n-m)\times n$ matrix given by $J_{i,i}=1$ for $i=1,2,\cdots,n-m$ and 0 otherwise. Then $w = L\epsilon = J\epsilon + (L-J)\epsilon = w_1+w_2$ where $w_1 = J\epsilon$ and $w_2 = (L-J)\epsilon$. We can see that $w_1 \sim N(0,\sigma^2 I_{n-m})$ and $w_2 \sim N(0,\sigma^2(L-J)(L-J)')$. Note that $d_1^2 = 1-O(m^{-1})$, hence each entry of the covariance matrix of w_2 is of order m^{-1}. So w_2 goes to 0 in probability as n goes to infinity and is negligible as compared to w_1.

We know that $\hat{\beta} = \beta+(Z'Z)^{-1}Z'w = \beta+(Z'Z)^{-1}Z'w_1+(Z'Z)^{-1}Z'w_2$. Hence under H_0, $C\hat{\beta} = C(Z'Z)^{-1}Z'w_1 + C(Z'Z)^{-1}Z'w_2$, where $C(Z'Z)^{-1}Z'w_1 \sim N(0,\sigma^2C(Z'Z)^{-1}C')$.

On the other hand, $w'Hw = w_1'Hw_1 + 2w_1'Hw_2 + w_2'Hw_2$. It is easy to check that H is a projection of rank $n-m-p$, i.e., $H^2=H$ and $\mathrm{rank}(H) = n-m-p$. Hence $w_1'Hw_1 = (Hw_1)'(Hw_1)$ follows a χ^2 distribution with $n-m-p$ degrees of freedom. Now we know that $\frac{1}{\sigma^2}w_1'Z(Z'Z)^{-1}C'(C(Z'Z)^{-1}C')^{-1}C(Z'Z)^{-1}Z'w_1$ follows $\chi^2(r)$ and $w_1'Hw_1$ follows $\chi^2(n-m-p)$ and they are independent. This means $\frac{\hat{\beta}'C'(C(Z'Z)^{-1}C')^{-1}C\hat{\beta}/r}{\hat{\sigma}^2}$ asymptotically follows $F(r,n-m-p)$ distribution.

By the same argument as in the proof of Theorem 1, we can prove that the asymptotic power of this test (at local alternatives) is the same as the usual F test when f is not present in the model (1). ∎

References

[1] Bickel, P. J., Klassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The John Hopkins University Press, Baltimore.

[2] Brown, L.D. and Low M.G. (1996). A Constrained Risk Inequality with Applications to Nonparametric Functional Estimations. *Ann. Statist.* **24**, 2524-2535.

[3] Cai, T. and Brown, L.D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26**, 1783-1799.

[4] Cai, T., Levine, M., and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression. *Journal of Multivariate Analysis,* **100**, 126-136.

[5] Cai, T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.* **36**, 2025C2054.

[6] Carroll, R. J., Fan, J., Gijbels,I., and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

[7] Chang, X., and Qu, L. (2004). Wavelet estimation of partially linear models. *Comput. Statist. Data Anal.* **47**, 31-48.

[8] Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* New York: Wiley.

[9] Chen, H. and Shiau, J. H. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Inference* **27**, 187-201.

[10] Cuzick, J. (1992). Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B* **54**, 831-843.

[11] Dawber, T. R., Meadors G. F., and Moore F. E. J. (1951). Epidemiological approaches to heart disease: the Framingham Study. *Amer. J. Public Health* **41**, 279-86.

[12] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

[13] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425-55.

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

[15] Fan, J., and Huang, L. (2001). Goodness-of-Fit Tests for Parametric Regression Models. *J. Amer. Statist. Assoc.* **96**, 640-652.

[16] Fan, J., and Zhang, J. (2004). Sieve empirical likelihood ratio tests for nonparametric functions. *Ann. Statist.* **32**, 1858-1907.

[17] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193.

[18] Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statist. Comput.* **17**, 293-310.

[19] Hall, P., Kay, J. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521-528.
[20] Hamilton, S., and Truong, Y. (1997). Local linear estimation in partly linear models. *J. Multivariate Anal.* **60**, 1C19.
[21] Härdle, W. (1991). *Smoothing Techniques*. Berlin: Springer-Verlag.
[22] Horowitz, J. and Spokoiny, V. (2001). An adaptive rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69**, 599-631.
[23] Lam, C., and Fan, J. (2007). Profile-Kernel Likelihood Inference With Diverging Number of Parameters. *Ann. Statist.* , to appear.
[24] Liang, H., Härdle, W. and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **27**, 1519-1535.
[25] Müller, M. (2001). Estimation and testing in generalized partial linear modelsA comparative study. *J. Statist. Comput.* **11**, 299-309.
[26] Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). On difference based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc. B* **67**, 19-41.
[27] Rice, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.
[28] Ritov, Y., and Bickel, P. J. (1990). Achieving information bounds in semi and non parametric models. *Ann. Statist.* **18**, 925-938.
[29] Robinson, P. M. (1988). Root-N consistent semiparametric regression. *Econometrica.* **56**, 931-954.
[30] Scott, D. W. (1992). *Multivariate Density Estimation.* New York: Wiley.
[31] Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditional parametric models. *Ann. Statist.* **20**, 1768-1802.
[32] Speckman, P. (1988). Kernel smoothing in partial linear models. *Jour. Roy. Statist. Soc.Ser. B* **50**, 413-436.
[33] Schimek, M. (2000). Estimation and inference in partially linear models with smoothing splines. *J. Statist. Plann. Inference* **91**, 525C540.
[34] Strang, G. (1989). Wavelet and dilation equations: A brief introduction. *SIAM Rev.* **31**, 614-627.
[35] Tjøstheim, D., and Auestad, B. (1994). Nonparametric Identification of Nonlinear Time Series: Projection. *J. Amer. Statist. Assoc.* **89**, 1398-1409.
[36] Wahba, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. *Statistics: An Appraisal. Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory* (H. A. David, ed.). Iowa State Univ. Press.
[37] Wang, L., Brown, L.D., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation on nonparametric regression. *Ann. Statist.* **36**, 646C664.
[38] Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters* **57**, 135-143.
[39] Yatchew, A. (2003). *Semiparametric regression for the applied econometri-*

cian. New York: Cambridge University press.