The Annals of Statistics
2003, Vol. 31, No. 1, 152–173

WAVELET THRESHOLD ESTIMATION FOR ADDITIVE REGRESSION MODELS[1]

BY SHUANGLIN ZHANG AND MAN-YU WONG

Michigan Technological University and
Hong Kong University of Science and Technology

Additive regression models have turned out to be useful statistical tools in the analysis of high-dimensional data. The attraction of such models is that the additive component can be estimated with the same optimal convergence rate as a one-dimensional nonparametric regression. However, this optimal property holds only when all the additive components have the same degree of "homogeneous" smoothness. In this paper, we propose a two-step wavelet thresholding estimation process in which the estimator is adaptive to different degrees of smoothness in different components and also adaptive to the "inhomogeneous" smoothness described by the Besov space. The estimator of an additive component constructed by the proposed procedure is shown to attain the one-dimensional optimal convergence rate even when the components have different degrees of "inhomogeneous" smoothness.

1. Introduction. Nonparametric regression models are flexible. They allow researchers to evaluate data without knowledge of the shape of the relationship between response and covariate(s). However, when the explanatory variables are multidimensional, these methods are less efficient. In particular, the rate of convergence for the standard estimator is usually poor, while no simple plot is available for model selection. An elegant solution to this problem is an additive model, which was proposed originally by Friedman and Stuetzle (1981) and popularized by Hastie and Tibshirani (1990). Under the additive model, the conditional expectation function of a dependent variable, Y, given the covariates $X_1, \ldots, X_d$ is expressed as a sum of d terms, that is,

$$E(Y|X_1 = x_1, \ldots, X_d = x_d) = g(x_1, \ldots, x_d) = g_1(x_1) + \cdots + g_d(x_d). \tag{1.1}$$

This model is easy to interpret and is much more flexible than a linear model.

In most previous papers using this model, the back-fitting technique based on the iterative smoothing process has been used as the main tool for estimating the additive model [Hastie and Tibshirani (1990)]. Although the back-fitting technique has proved to be very useful in both application and simulation, it is somewhat

Received February 1999; revised December 2001.

[1]Supported in part by NNSF Grant 10071011 of China and Natural Science Foundation of Heilongjiang Province, China.

AMS 2000 subject classifications. Primary 62G07; secondary 62G20.

Key words and phrases. Local polynomial estimation, wavelet estimation, optimal convergence rate, additive regression, threshold, Besov space.

difficult to analyze theoretically [see the recent work by Opsomer and Ruppert (1997)]. For this reason, Linton and Nielsen (1995), Tjøstheim and Auestad (1994), Linton (1996, 1997), Nielsen and Linton (1998) and Fan, Härdle and Mammen (1998) proposed the use of a direct method based on "marginal integration" for estimation. It is based on the fact that, up to a constant, $g_v(x_v)$ is equal to

$$E[g(X_1, \ldots, X_{v-1}, X_v, X_{v+1}, \ldots, X_d)].$$

Using this idea and weighted local linear fitting, Fan, Härdle and Mammen (1998) proved that an additive component can be estimated with the same asymptotic bias and variance as if the other components were known.

In a series of studies, Stone (1985, 1986, 1994) proved that an additive component can be estimated by the one-dimensional optimal convergence rate under the assumption that all additive components have the same degree of "homogeneous" smoothness in Hölder space by using a linear estimator of the form

$$\sum_{i=1}^{n} Y_i W_i(x_v, \mathbf{X}_1, \ldots, \mathbf{X}_n), \tag{1.2}$$

where $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ is an $R^d \times R$-dimensional random sample of size n.

The methods mentioned above, using kernels, local polynomials or splines directly, construct the estimator of the additive components. It is called a one-step method. The one-step method, such as the local polynomial method, implicitly assumes that all additive components possess the same degree of smoothness and hence that they can be approximated equally well. However, different additive components possess different degrees of smoothness. The one-step estimator of a component, $g_v(\cdot)$, cannot attain the optimal convergence rate as if $g_v(\cdot)$ were actually smoother than other components. This problem has been raised explicitly by Fan and Zhang (1999) in the context of varying coefficient models. Fan and Zhang (1999) have shown that the one-step method cannot be optimal when different coefficient functions possess different degrees of smoothness. Furthermore, the linear estimator in the form of (1.2) cannot capture the "inhomogeneous" smoothness. Donoho and Johnstone (1998) and Zhang, Wong and Zheng (2002) have demonstrated that, in a one-dimensional regression, no linear estimator can attain the optimal convergence rate in a ball of Besov space, $B^s_{p,q}$, for $p < 2$ when the error is measured by L_2-norm, while a suitable wavelet thresholding estimator can attain the optimal convergence rate. Donoho (1995), Donoho and Johnstone (1994, 1995), Donoho, Johnstone, Kerkyacharian and Picard (1995), Hall and Turlach (1997), Hall, Kerkyacharian and Picard (1998), Neumann and von Sachs (1995) and Zhang, Wong and Zheng (2002) have discussed the advantages of wavelet thresholding estimation for regression and density functions.

In this paper, a two-step procedure is proposed to construct an estimator for an additive component. The first step involves establishing the initial estimator, $\breve{g}_v(\cdot)$,

of the additive component, $g_v(\cdot)$, using local polynomial fitting along with "marginal integration." Such an initial estimator is usually undersmoothed so that its bias is small. Then, in the second step, we transform the initial estimator, $\breve{g}_v(\cdot)$, by finite discrete wavelet transformation and use the thresholding technique to estimate $g_v(\cdot)$. As pointed out by Donoho and Johnstone (1995), this step is simple and practical. It can be handled by an algorithm with functions in $O(n)$ operations for a sample of size n.

This procedure has the following advantages:

1. Using the proposed two-step procedure, we can construct an estimator that is adaptive to different degrees of smoothness, s_j, $j = 1, \ldots, d$, in different components. This answers the question implied by Fan and Zhang (1999) for varying coefficient models.
2. Even if the components have "inhomogeneous" smoothness in Besov space, with appropriate choices of the bandwidth in the first step and of the threshold in the second step, each additive component can be estimated with the one-dimensional optimal convergence rate, which is the same optimal rate as if the other components were known. This extends the results of Fan, Härdle and Mammen (1998) in Besov space in terms of convergence rate.

Intuitively, we use a two-step procedure instead of a one-step method to make the estimator adaptive to different degrees of smoothness in different components. We also use the wavelet thresholding estimator in the second step instead of a linear estimator to capture the "inhomogeneous" smoothness.

This paper is organized as follows. In Section 2, we introduce some basic concepts and properties of wavelet and Besov space in $[0, 1]$ intervals. Section 3 proposes the procedure of a two-step wavelet thresholding estimation. Asymptotic properties of the proposed estimator are presented in Section 4. Section 5 includes the proofs of the main results. All proofs of the lemmas are given in the Appendix.

2. Wavelet and Besov space of [0, 1] intervals. In this paper, we confine our attention to the wavelet basis of $[0, 1]$ intervals given by Cohen, Daubechies and Vial (1993), that is, the collection of $\{\phi_{J_0,k},\ k = 0, 1, \ldots, 2^{J_0} - 1;\ \psi_{j,k},\ j \geq J_0 \geq 0,\ k = 0, 1, \ldots, 2^j - 1\}$ forms an orthonormal basis of $L^2[0, 1]$. The wavelet series representation of a function, $f \in L^2[0, 1]$, is then

$$f = \sum_{k=0}^{2^{J_0}-1} \alpha_{J_0,k}\phi_{J_0,k} + \sum_{j \geq J_0} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

where the coefficients are

$$\alpha_{J_0,k} = \int_0^1 f(x)\phi_{J_0,k}(x)\,dx \quad \text{and} \quad \beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)\,dx. \tag{2.1}$$

We also denote $\beta_{J_0-1,k} = \alpha_{J_0,k}$ for simplicity of notation.

These basis functions are derived from Daubechies' orthonormal compact-supported wavelet [Daubechies (1992)] at the interior of the interval and with boundary correction at the "edges." This "boundary correction" affects only a fixed number of wavelet coefficients at each resolution level and does not alter the qualitative phenomenon we consider here. Thus, in the following discussion, we assume that $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ for all (j, k), where ϕ and ψ are compact-supported scaling and wavelet functions, respectively.

DEFINITION 1. A wavelet basis on $[0, 1]$ has regularity r if all functions used in the analysis are compactly supported and have r continuous derivatives.

In practical application, a wavelet transformation is usually carried out by the manipulation on the filter coefficients, that is, by a finite discrete wavelet transformation rather than by evaluating the wavelet and scaling functions explicitly. This transformation, along with a careful treatment of the boundary correction, has been described by Cohen, Daubechies and Vial (1993) and Donoho and Johnstone (1994). In the present paper, to focus on our main purpose, we employ the simple periodic version of a finite discrete wavelet transformation. This version yields an exact orthogonal transformation between the data and wavelet coefficients [Donoho and Johnstone (1994)].

Suppose that we have data $y = (y_i)_{i=0}^{N-1}$ with $N = 2^J$. For various combinations of the regularity of the wavelet basis, r, and the low-resolution cutoff, J_0, one may construct the finite discrete wavelet transformation matrix, $\mathcal{W}$, which is an $N \times N$ orthogonal matrix. This matrix yields a vector, w, from the empirical wavelet coefficients of y via $w = \mathcal{W}y$ and thus the inversion formula $y = \mathcal{W}^T w$.

For the vector, w, with $N = 2^J$ elements, we index it dyadically as follows:

$$w_{j,k}(j = J_0, \ldots, J-1;\ k = 0, \ldots, 2^j - 1)$$

corresponding to the wavelet coefficients, $\beta_{j,k}$, given in (2.1) and the remaining elements $w_{J_0-1,k}$ $(k = 0, \ldots, 2^{J_0-1})$ corresponding to $\alpha_{J_0,k}$ (or $\beta_{J_0-1,k}$) given in (2.1). To interpret these coefficients, denote $\mathcal{W}_{j,k}$ as the (j, k)th row of $\mathcal{W}$. The wavelet transformation and, thus, the inversion formula become

$$w_{j,k} = \sum_{i=0}^{N-1} y_i \mathcal{W}_{j,k}(i) \quad \text{and} \quad y_i = \sum_{j,k} w_{j,k} \mathcal{W}_{j,k}(i), \tag{2.2}$$

respectively. Furthermore, if y_k is the true wavelet coefficient corresponding to a scaling function of function f in resolution J, that is, $y_k = \alpha_{J,k} = \int \phi_{J,k}(x) f(x)\, dx$, then $w_{j,k}$ given by (2.2) is the true wavelet coefficient, that is, $\beta_{j,k}$, in resolution j given in (2.1).

DEFINITION 2 [Donoho and Johnstone (1998)]. Let $\alpha_{j_0,k}$ and $\beta_{j,k}$ be the wavelet coefficients of f corresponding to a wavelet basis with regularity r $(> s)$. Define the norm

$$\|f\|_{B^s_{p,q}} = \|\alpha_{j_0}\|_p + \left(\sum_{j \geq j_0} \left(2^{j(s+(1/2)-(1/p))} \|\beta_j\|_p\right)^q \right)^{1/q}$$

for $1 \leq p, q \leq \infty$, where $\|\alpha_{j_0}\|_p^p = \sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0,k}|^p$ and $\|\beta_j\|_p^p = \sum_{k=0}^{2^j-1} |\beta_{j,k}|^p$. The Besov space of $[0,1]$, denoted by $B^s_{p,q}[0,1]$, is the set of functions, $f : [0,1] \to R$ with $\|f\|_{B^s_{p,q}} < \infty$. A Besov ball, denoted by $B^s_{p,q}(H)$, is given by

$$B^s_{p,q}(H) = \left\{ f \in B^s_{p,q}[0,1] : \|f\|_{B^s_{p,q}} \leq H \right\}$$

for a finite constant H.

Besov space contains many traditional function spaces. We recall some injection results, from Donoho, Johnstone, Kerkyacharian and Picard (1996) and Triebel (1992):

$$\begin{aligned}
&B^s_{\infty,\infty}[0,1] = C^s[0,1] && \text{for } 0 < s < 1,\\
&B^s_{p,q}[0,1] \subset C^0[0,1] \subset L_\infty[0,1] && \text{for } s > \frac{1}{p},\\
&B^s_{p,q}[0,1] \subset B^{s'}_{p',q}[0,1] && \text{for } s' = s - \frac{1}{p} + \frac{1}{p'},\ p' \geq p,\\
&B^s_{p,q}[0,1] \subset B^{s'}_{p,q}[0,1] && \text{for } s \geq s',
\end{aligned}$$

where $C^s[0,1]$ and $C^0[0,1]$ denote Hölder and bounded continuous functional spaces of $[0,1]$.

Further, we denote

$$B(x,t) = \{y \in [0,1] : |x - y| \leq t\}, \qquad x \in [0,1],\ t > 0,$$

and

$$\theta_j^{(M)}(x) = \inf_P \sup_{y \in B(x, 2^{-j})} |f(y) - P(y)|$$

in which the function, f, is clearly understood, where inf is taken over all polynomials P with degree no greater than M; $\xi_m = 2^{-J}(m + 1/2)$ and $N = 2^J$ for an integer J. Using this notation, we present the following propositions.

PROPOSITION 1 [Propositions 3.4.2 and 3.4.3 in Triebel (1992)]. *For any function f, there exist a positive constant, C, an integer, j_0, independent of*

j and f, a function, f_1, and an optimal polynomial $P_x(y) = \sum_{\alpha=0}^{M} \frac{D^\alpha f_1(x)}{\alpha!} \times (y-x)^\alpha$, such that:

(a) $\sup_{y \in B(x,2^{-j})} |f(y) - P_x(y)| \le C, \theta_{j-J_0}^{(M)}(x)$;
(b) $D^\alpha f_1(z) \le C 2^{j\alpha} \theta_{j-j_0}^{(\alpha-1)}(x)$ *for* $z \in B(x, 2^{-j}),\ \alpha = 0, 1, \ldots, M$;
(c) $\theta_j^{(\alpha-1)}(x) \le C \sum_{l=0}^{j} 2^{-(j-l)\alpha} \theta_l^{(\alpha)}(x) + C 2^{-j\alpha} \int_{B(x,c)} |f(y)|\, dy$

for integers M and $j \ge 0$.

PROPOSITION 2 [Proposition 2.4 in Zhang, Wong and Zheng (2002)]. *If $f \in B_{p,q}^s(H)$, $s > 1/p$ and $M \ge [s]$, then, for any integer, l, satisfying $0 \le l \le J$,*

$$\sum_{m=0}^{2^J-1} |\theta_l^{(M)}(\xi_m)|^p \le C 2^{-lsp+J},$$

where the constant, C, does not depend on f, J and l.

Using Definition 2 and Propositions 1 and 2, we attain the following lemmas. Their proofs are given in the Appendix.

LEMMA 2.1. *For any constant, C', there exists a constant, C, such that*

$$\frac{1}{N} \sum_{m=0}^{N-1} \sup_{y, y' \in B(\xi_m, C'/N)} |P_{\xi_m}(y) - P_{\xi_m}(y')| \le C \frac{J}{N}$$

for all $f \in B_{p,q}^s(H)$ with $1 \le p, q \le \infty$, where $P_x(y)$ is the optimal polynomial of degree $M = [s]$ corresponding to f.

We allow

$$\beta = (\beta_{j,k}) = \mathcal{W}\alpha_J \quad \text{and} \quad \beta^* = (\beta_{j,k}^*) = \mathcal{W}\mathbf{f}/\sqrt{N},$$

where $\alpha_J = (\alpha_{J,k})_{k=0}^{N-1}$ and $\mathbf{f} = (f(\xi_k))_{k=0}^{N-1}$. For a smooth function f, the wavelet coefficient corresponding to the scaling function given in (2.1), $\alpha_{J,k}$, approximately equals $f(\xi_k)/\sqrt{N}$ and the wavelet coefficient corresponding to the mother wavelet function, $\beta_{j,k}$, approximately equals $\beta_{j,k}^*$. More specifically, we have:

LEMMA 2.2. *If $f \in B_{p,q}^s(H)$ with $1 \le p,\ q \le \infty$ and $s > \max\{\frac{1}{2}, \frac{1}{p}\}$ and the wavelet and scaling functions used are continuous, then there exists a constant, C, such that*

$$\sum_{k=0}^{N-1} \left(\alpha_{J,k} - \frac{1}{\sqrt{N}} f(\xi_k) \right)^2 \le C \frac{J^2}{N} \quad \text{and} \quad \sum_{j,k} (\beta_{j,k} - \beta_{j,k}^*)^2 \le C \frac{J^2}{N}.$$

3. Estimation procedure. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample from an $R^d \times R$ random vector $(\mathbf{X}, Y)$, where $\mathbf{X}_i$ and $\mathbf{X}$ are $(X_{i,1}, \ldots, X_{i,d})$ and $(X_1, \ldots, X_d)$, respectively. We consider the following additive model:

$$Y_i = g(\mathbf{X}_i) + \sigma(\mathbf{X}_i)e_i = \alpha + g_1(X_{i,1}) + \cdots + g_d(X_{i,d}) + \sigma(\mathbf{X}_i)e_i, \tag{3.1}$$

where $e_1, \ldots, e_n$ are independent of each other and independent of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ with $E(e_i) = 0$ and $\operatorname{Var}(e_i) = 1$. To ensure the identifiability of the functions, $g_v(\cdot)$, for $v = 1, \ldots, d$, we include the intercept α and assume $E(g_v(X_{i,v})) = 0$ for all v. Furthermore, we assume that $\mathbf{X}$ is distributed with density $f(x)$ on the interval $[0, 1]^d$.

As discussed in the Introduction, a one-step procedure cannot be optimal when the additive components have different degrees of smoothness and we do not know in advance which component is smoother. In the proposed two-step procedure, we construct the estimator of the additive component $g_v(\cdot)$ such that the estimator is adaptive to different degrees of smoothness in additive components and use the wavelet thresholding estimator in the second step to capture the "inhomogeneous" smoothness in the additive components. First, we introduce the following notation:

$$\begin{aligned}
N &= 2^J \asymp \frac{n}{\ln^2 n}, \\
A_x &= \{i : N^{1/d}|X_{i,j} - x_j| \le \tfrac{1}{2},\ j = 1, \ldots, d\}, \qquad |A_x| = \operatorname{card}(A_x), \\
\Lambda_x &= \{u' V_x^{-1} u \le K_A |A_x|^{-1},\ V_x > 0\}, \\
z_{i,j} &= N^{1/d}(X_{i,j} - x_j), \\
Z_{x,i} &= \big(1, z_{i,1}, \ldots, z_{i,1}^D, \ldots, z_{i,d}, \ldots, z_{i,d}^D\big)^T, \\
V_x &= \sum_{i \in A_x} Z_{x,i} Z_{x,i}^T, \\
u_{(dD+1)\times 1} &= (1, 0, \ldots, 0)', \\
g^{(v)}(x, \xi) &= g(x_1, \ldots, x_{v-1}, \xi, x_{v+1}, \ldots, x_d) \\
\xi_m &= 2^{-J}\big(m + \tfrac{1}{2}\big),
\end{aligned} \tag{3.2}$$

where $x = (x_1, \ldots, x_d)^T$, K_A is a positive constant, D is a nonnegative integer, and $a_n \asymp b_n$ means that $0 < \inf\lim \frac{a_n}{b_n} \le \sup\lim \frac{a_n}{b_n} < \infty$. With this notation, we construct the estimator of $g_v(\cdot)$ by the following two steps:

Step 1. We use the local polynomial with degree D and the idea of "marginal integration" to determine an initial estimator for $g_v(\cdot)$. For any given point, x_o, we approximate the function locally as

$$g(y_o) = \alpha + g_1(y_{o1}) + \cdots + g_d(y_{od}) \approx \eta_0 + \sum_{i=1}^{d} \sum_{j=1}^{D} \eta_{i,j}(y_{oi} - x_{oi})^j,$$

where $y_o = (y_{o1}, \ldots, y_{od})^T$ is in the neighborhood of $x_o = (x_{o1}, \ldots, x_{od})^T$. This leads to the following local least-squares problem: minimize

$$\sum_{i \in A_x} (Y_i - \eta^T Z_{x,i})^2,$$

where $\eta^T = (\eta_0, \eta_{1,1}, \ldots, \eta_{1,D}, \ldots, \eta_{d,1}, \ldots, \eta_{d,D})$. If $V_x > 0$ (that is, V_x is a positive definite matrix), then we derive the unique local polynomial estimator of $g(x)$, that is, the local least-squares estimator of η_0, as

$$\tilde{g}(x) = \sum_{i \in A_x} u' V_x^{-1} Z_{x,i} Y_i.$$

Since this estimator is not defined in the case of $\det(V_x) = 0$, we modify the estimator to

$$\hat{g}(x) = \sum_{i \in A_x} u' V_x^{-1} Z_{x,i} Y_i I_{\Lambda_x}. \tag{3.3}$$

Note that $g_v(x_v) = E(g^{(v)}(X, x_v)) - \alpha$. The idea of "marginal integration" yields the initial estimator of $g_v(x_v)$ as

$$\breve{g}_v(x_v) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(X_{i,1}, \ldots X_{i,(v-1)}, x_v, X_{i,(v+1)}, \ldots, X_{i,d}) - \overline{Y} \tag{3.4}$$

where $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Step 2. We use a finite discrete wavelet transformation and the threshold idea to determine the wavelet thresholding estimator of $g_v(\cdot)$. The procedure is as follows:

(a) Take the value of the initial estimator at equally spaced points $\xi_0, \ldots, \xi_{N-1}$. Treat $\breve{\mathbf{g}}_v = (\breve{g}_v(\xi_k))_{k=0}^{N-1}$ as our data and apply a finite discrete wavelet transformation to obtain N empirical wavelet coefficients, $w_{j,k}^{(v)}$,

$$w^{(v)} = (w_{j,k}^{(v)}) = \mathcal{W} \breve{\mathbf{g}}_v, \tag{3.5}$$

where $\mathcal{W}$ is a finite discrete wavelet transformation matrix with regularity, r, and low resolution cutoff, J_0, as given in Section 2.

(b) Choose threshold λ_j and apply either a hard or soft threshold. We attain empirical wavelet thresholding coefficients

$$\hat{w}^{(v)} = (\hat{w}_{jk}^{(v)}) \quad \text{with} \quad \hat{w}_{j,k}^{(v)} = \delta(w_{j,k}^{(v)}, \lambda_j), \tag{3.6}$$

where $\delta(\cdot, \cdot)$ denotes either the hard thresholding function, $\delta^{(h)}(y, \lambda) = y I_{\{|y| \geq \lambda\}}$, or the soft thresholding function, $\delta^{(s)}(y, \lambda) = \operatorname{sgn}(y)(|y| - \lambda)_+$.

(c) Invert the finite discrete wavelet transformation. We obtain the wavelet thresholding estimator of $g_v(\cdot)$ at points $\xi_k (k = 0, \ldots, N-1)$ denoted by

$$\hat{\mathbf{g}}_v = (\hat{g}_{vk})_{k=0}^{N-1} = \mathcal{W}^T \hat{w}^{(v)}. \tag{3.7}$$

(d) The wavelet thresholding estimator of $g_v(\cdot)$ at continuous points is thus given by the linear interpolation

$$\hat{g}_v(x) = \begin{cases} \hat{g}_{v0}, & \text{for } x \le \xi_0, \\ \hat{g}_{vm} + (\hat{g}_{v(m+1)} - \hat{g}_{vm})N(x - \xi_m), & \text{for } \xi_m \le x < \xi_{m+1}, \\ \hat{g}_{v(N-1)}, & \text{for } x \ge \xi_{N-1}. \end{cases} \tag{3.8}$$

REMARK 1. Theoretically, we can choose a constant, K_A, such that, for any integer h, $P(\bigcap_{x\in[0,1]} \Lambda_x) = 1 - O(n^{-h})$. This means that if we choose the bandwidth with an order of $\ln^2 n/n$ in the estimation of a local polynomial with degree D, we guarantee that, with probability approximately equal to 1, there are enough data in the local neighborhood of any point to fit the polynomial. For a given data set, there may not be enough data in the neighborhood of some points. In this case, we set the initial estimated value to be zero at these points. In practice, other methods, such as interpolation using the estimated values from nearby points, may be more appropriate. Moreover, a data-driven method for choosing the bandwidth needs further investigation.

4. Theoretical results. In this section, we show that, by the appropriate choice of threshold, λ_j, and the low resolution cutoff, J_0, the estimator of $g_v(\cdot)$ constructed above can attain the optimal one-dimensional convergence rate (i.e., the optimal rate as if we knew all other components) when the additive components have "inhomogeneous" smoothness described by Besov space and when we do not know in advance which component is smoother. Furthermore, we show that even if we do not know the degrees of smoothness in the additive components, the estimator constructed above can attain the optimal convergence rate up to a logarithmic factor by choosing suitable λ_j and J_0. We first assume some conditions under model (3.1).

CONDITION 1. The density function, $f(x)$, is continuous on interval $[0, 1]^d$ and bounded from zero and infinity; that is, there exist constants, k^* and k_*, such that

$$0 < k_* \le f(x) \le k^* < \infty \qquad \text{for } x \in (0, 1]^d.$$

CONDITION 2. The variance function, $\sigma(\cdot)$, in model (3.1), is bounded; that is, there is a constant, σ_0, such that $\sigma(x) \le \sigma_0$ for $x \in [0, 1]^d$.

CONDITION 3. $E|e_i|^l \le \frac{1}{2} l! H_0^{l-2}$ for $i = 1, \ldots, n,\ l \ge 3$ and a constant H_0.

CONDITION 4. $r \ge \max\{s_1, \ldots, s_d\}$ and $D \ge \max\{s_1, \ldots, s_d\}$, where r and D are the regularity of a wavelet basis and the degree of a local polynomial, respectively.

REMARK 2. The moment assumption in Condition 3 is not strong. For example, if e_i has a normal distribution, $N(0, 1)$, then $E|e_i|^l = (\frac{l}{2}-1)(\frac{l}{2}-2)\cdots\frac{1}{2}$. It is clear that there exists a constant, H_0, such that Condition 3 holds. In fact, most textbook distributions satisfy this condition.

For any v, let threshold $\lambda_j = K(j - J_0)_+\sqrt{N/n}$ where K is a constant and low resolution cutoff, J_0, has an order of $2^{J_0} \asymp n^{1/(2s_v+1)}$. Then, under Conditions 1, 2, 3 and 4, we have the following theorems.

THEOREM 1. *Let $1 \le p_t, q_t \le \infty$ and $s_t > \max\{\frac{d}{2}, \frac{d}{p_t}\}$ for $t = 1, \ldots, d$. Then, for any given v, there exist constants, K' and K_A^*, such that*

$$\sup_{g_t \in B^{s_t}_{p_t,q_t}(H_t):\ t=1,\ldots,d} E\left(\frac{1}{N}\sum_{m=0}^{N-1}\left(\hat{g}_v(\xi_m) - g_v(\xi_m)\right)^2\right) = O(n^{-2s_v/(2s_v+1)})$$

for $K \ge K'$ and $K_A \ge K_A^*$.

THEOREM 2. *Under the same conditions as Theorem 1, for any given v, there exist constants, K' and K_A^*, such that*

$$\sup_{g_t \in B^{s_t}_{p_t,q_t}(H_t):\ t=1,\ldots,d} E\int_0^1 \left(\hat{g}_v(x_v) - g_v(x_v)\right)^2 dx_v = O(n^{-2s_v/(2s_v+1)})$$

for $K \ge K'$ and $K_A \ge K_A^*$.

It is noted that the wavelet thresholding estimator of g_v given above depends on the smooth parameter, s_v. However, slightly modifying the wavelet thresholding estimator can render it adaptive, in the sense that it attains the optimal convergence rate up to a logarithmic factor without specifying the value of s_v. We apply the same estimation procedure but change the values of wavelet threshold, λ_j, and low resolution cutoff, J_0, to $\sqrt{t_n N/n}$ and a constant independent of n, respectively. Denote the wavelet thresholding estimator of g_v corresponding to such values of λ_j and J_0 by $\hat{g}_{vA}$. Then, under Conditions 1, 2, 3 and 4, the following theorem holds.

THEOREM 3. *If t_n satisfies the condition that $\frac{t_n}{\ln^2 n} \to \infty$, then, for any given v, there exists a constant, K_A^*, such that*

$$\begin{aligned}\sup_{g_t \in B^{s_t}_{p_t,q_t}(H_t):\ t=1,\ldots,d} & E\left(\frac{1}{N}\sum_{m=0}^{N-1}\left(\hat{g}_{vA}(\xi_m) - g_v(\xi_m)\right)^2\right) \\ &= O\left(\left(\frac{t_n}{n}\right)^{-2s_v/(2s_v+1)}\right)\end{aligned} \tag{4.1}$$

and

$$\sup_{g_t \in B^{st}_{p_t,q_t}(H_t):\ t=1,\dots,d} E\int_0^1 \bigl(\hat{g}_{vA}(x_v) - g_v(x_v)\bigr)^2\,dx_v = O\left(\left(\frac{t_n}{n}\right)^{-2s_v/(2s_v+1)}\right) \tag{4.2}$$

for all s_t, p_t *and* q_t *satisfying* $1 \le p_t, q_t \le \infty$, $\max\{\frac{d}{p_t}, \frac{d}{2}\} < s_t$ *for* $t = 1, \dots, d$, *and* $K_A \ge K_A^*$.

REMARK 3. If t_n satisfies the conditions that $\frac{t_n}{\ln^2 n} \to \infty$ and $\frac{t_n}{(\ln^2 n)^\beta} \to 0$ for any $\beta > 1$ (e.g., $t_n = \ln^2 n \cdot \ln(\ln n)$), then

$$\left(\frac{t_n}{n}\right)^{2s_v/(2s_v+1)} = o(\ln^2 n\, n^{-2s_v/(2s_v+1)}).$$

5. Proofs of theorems. In this section, we offer the proofs of Theorems 1, 2 and 3. For simplicity, we assume that $d = 2$ and $v = 1$; that is, we consider only the estimator of the first additive component in a two-dimensional additive model. Furthermore, without loss of generality, we assume $\alpha = 0$ and $\overline{Y} = 0$ in the model of (3.1).

In order to prove the theorems, we present several lemmas and their proofs are given in the Appendix. First of all, we introduce some notation:

$$\begin{aligned}
A_{x,\delta} &= \{i : \sqrt{N}|X_i - x| \le \delta\}, & B_m &= \left\{i : |X_{i,2} - \xi_m| \le \frac{1}{2N}\right\},\\
C_{m,l,i} &= u'(V_{x_{m,l}})^{-1} Z_{x_{m,l},i}, & \varepsilon_i &= \sigma(X_i)e_i,\\
\alpha_J &= (\alpha_{J,0}, \dots, \alpha_{J,2^J-1})^T, & \beta &= (\beta_{j,k}) = \mathcal{W}\alpha_J,\\
\mathbf{g}_1 &= \bigl(g_1(\xi_0), \dots, g_1(\xi_{N-1})\bigr)^T, & \beta^* &= (\beta^*_{jk}) = \mathcal{W}\mathbf{g}_1/\sqrt{N},\\
\hat{\beta} &= (\hat{\beta}_{jk}) = \hat{w}^{(1)}/\sqrt{N}, & \tilde{\beta} &= (\tilde{\beta}_{jk}) = w^{(1)}/\sqrt{N},\\
A_{m,l} &= A_{x_{m,l}},
\end{aligned} \tag{5.1}$$

where, for $x = (x_1, x_2)$ and $x' = (x_1', x_2')$, $|x - x'| \le \delta$ implies that $|x_1 - x_1'| \le \delta$ and $|x_2 - x_2'| \le \delta$; u, $V_{x_{m,l}}$ and $Z_{x_{m,l},i}$ are given in (3.2) for $x_{m,l} = (\xi_m, X_{l,2})$; $\alpha_{J,k} = \int \phi_{J,k}(x) g_1(x)$ is the true wavelet coefficient of $g_1(\cdot)$ corresponding to the scaling function; $w^{(1)}$ and $\hat{w}^{(1)}$ are given in (3.5) and (3.6), respectively.

Using this notation, we can write the initial estimator given in (3.4) as $\breve{g}_1(\xi_m) = \frac{1}{n}\sum_{l=1}^n \sum_{i \in A_{m,l}} C_{m,l,i} Y_i I_{\Lambda_{x_{m,l}}}$. From Lemma 2.2, we conclude that β^* and $\hat{\beta}$ are the asymptotic wavelet coefficients of $g_1(\cdot)$ and the estimator $\hat{g}_1(\cdot)$, respectively.

With the notation given in (3.2) and (5.1), we have the following lemmas.

LEMMA 5.1. (a) *For any constant δ, there exist constants $C_*(\delta)$ and $C^*(\delta)$ such that, for any $h > 0$,*

$$1 - P\left(C_*(\delta) \le \frac{N}{n}|A_{x,\delta}| \le C^*(\delta) \text{ for all } x \in [0,1]^2\right) = O(n^{-h}). \tag{5.2}$$

(b) *For any $h > 0$, there exists a constant, C_0, such that*

$$1 - P\left(|B_m| \le C_0\frac{n}{N},\ m = 1, \ldots, N\right) = O(n^{-h}). \tag{5.3}$$

LEMMA 5.2. *Under Condition* 1, *for any $h > 0$, there exists a constant, K_A^*, such that $1 - P(\Lambda) = O(n^{-h})$ for $K_A \ge K_A^*$, where $\Lambda = \bigcap_{x\in[0,1]^2} \Lambda_x = \{u'V_x^{-1}u \le K_A|A_x|^{-1},\ V_x > 0, \text{ for all } x \in [0,1]^2\}$.*

In the following discussion, we assume that the constant, K_A, satisfies the condition $K_A \ge K_A^*$, where K_A^* is given by Lemma 5.2. Denote

$$\begin{aligned} \Omega_1 &= \left\{C_* \le \frac{N}{n}|A_x| \le C^* \text{ for all } x \in [0,1]^2\right\}, \\ \Omega_2 &= \left\{|B_m| \le C_0\frac{n}{N},\ m = 1, \ldots, N\right\}, \\ \Omega &= \Lambda \cap \Omega_1 \cap \Omega_2, \end{aligned} \tag{5.4}$$

where C_0, $C_* = C_*(1/2)$ and $C^* = C^*(1/2)$ are given in Lemma 5.1. From Lemmas 5.1 and 5.2 , we have, for any $h > 0$,

$$1 - P(\Omega) = O(n^{-h}). \tag{5.5}$$

LEMMA 5.3. *Under Conditions* 1 *and* 4 , *if $g_t \in B_{p_t,q_t}^{s_t}(H_t)$ with $s_t > \max\{\frac{2}{p_t}, 1\}$ for $t = 1, 2$, then*

$$\begin{aligned} (\tilde{\beta}_{j,k} - \beta_{j,k}^*)I_\Lambda = \frac{1}{n\sqrt{N}} \sum_{m=1}^{N}\sum_{l=1}^{n}\left(\sum_{i\in A_{m,l}} C_{m,l,i}\varepsilon_i\right)\mathcal{W}_{j,k}(m)I_\Lambda \\ + \frac{a_{jk}}{n}\sum_{l=1}^{n} g_2(X_{l,2})I_\Lambda + o\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

for (j,k) in the domain, where a_{jk} is a constant with $a_{jk}^2 \le 1$ and $o(\cdot)$ is uniform for $g_t \in B_{p_t,q_t}^{s_t}(H_t)$ and $t = 1, 2$.

LEMMA 5.4. *Under the same conditions as in Lemma* 5.3 *and Conditions* 2 *and* 3 *stated in Section* 4, *for any* (j,k) *in the domain, then*

$$E|\tilde{\beta}_{j,k} - \beta^*_{j,k}|^\tau = O(n^{-\tau/2})$$

for $\tau \geq 2$.

LEMMA 5.5 [Petrov (1975), Zhang, Wong and Zheng (2002)]. *Let* Z_i *be a sequence of independent random variables such that*

$$E(Z_i) = 0, \quad E(Z_i^2) \leq \sigma_i^2 < \infty \quad \text{and} \quad |E(Z_i^h)| \leq \tfrac{1}{2} h! \sigma_i^2 H^{h-2}$$

for $i = 1, \ldots, n$, $h \geq 3$ *and a constant* H. *Then, for* $S = \sum_{i=1}^n Z_i$ *and* $B_n \geq \sum_{i=1}^n \sigma_i^2$, *the inequalities*

$$P(|S| > \lambda) \leq 2e^{-\lambda^2/(4B_n)}$$

for $0 \leq \lambda \leq \frac{B_n}{H}$ *and*

$$P(|S| > \lambda) \leq 2e^{-\lambda/(4H)}$$

for $\lambda > \frac{B_n}{H}$ *hold.*

In the remainder of this section, we present the proofs of three theorems. In the proofs, we denote $C, C_1, \ldots$ to be constants having different values at different places and use the Hölder-type inequality

$$\left(\sum_{i=1}^m |a_i|^q\right)^{1/q} \leq \left(\sum_{i=1}^m |a_i|^p\right)^{1/p}$$

for $p \leq q$ and

$$\left(\frac{1}{m}\sum_{i=1}^m |a_i|^q\right)^{1/q} \leq \left(\frac{1}{m}\sum_{i=1}^m |a_i|^p\right)^{1/p} \tag{5.6}$$

for $p \geq q$. Moreover, we use the following inequality [Petrov (1975)]. Assume $Z_1, \ldots, Z_n$ to be independent random variables with zero mean. Then, for any $\tau > 0$, there exists a constant, C, independent of τ such that

$$E\left|\sum_{i=1}^n Z_i\right|^\tau \leq C\left[\sum_{i=1}^n E|Z_i|^\tau + \left(\sum_{i=1}^n EZ_i^2\right)^{\tau/2}\right]. \tag{5.7}$$

PROOF OF THEOREM 1. Denote $\lambda_j^* = \lambda_j/\sqrt{N}$. Using the notation given above and noting that $\hat{\beta}$ and β^* are orthogonal transformations of $\hat{\mathbf{g}}_1/\sqrt{N} =$

$(\hat{g}_1(\xi_0)/\sqrt{N}, \ldots, \hat{g}_1(\xi_{N-1})/\sqrt{N})^T$ and $\mathbf{g}_1/\sqrt{N}$, respectively, it follows that

$$\begin{aligned} &E\left(\frac{1}{N}\sum_{m=0}^{N-1}\big(\hat{g}_1(\xi_m) - g_1(\xi_m)\big)\right)^2 \\ &\quad = E(\hat{\beta} - \beta^*)^T(\hat{\beta} - \beta^*) \\ &\quad = E\big[(\hat{\beta} - \beta^*)^T(\hat{\beta} - \beta^*)I_\Omega\big] + O\left(\frac{1}{n}\right) \\ &\quad = \sum_{j,k} E\big[\big(\delta(\tilde{\beta}_{j,k}, \lambda_j^*) - \beta_{j,k}^*\big)^2 I_\Omega\big] + O\left(\frac{1}{n}\right). \end{aligned}$$

From Lemma 2 of Delyon and Juditsky (1996), we note that there exists a constant, C, such that

$$\big(\delta(x, \lambda) - y\big)^2 \le C\big[\min\{|y|, \lambda\}^2 + (x - y)^2 I_{\{|x-y| \ge \lambda/2\}}\big],$$

for both hard and soft thresholding functions, $\delta(\cdot, \cdot)$. Using this result, we have

$$\begin{aligned} &E\left(\frac{1}{N}\sum_{m=0}^{N-1}\big(\hat{g}_1(\xi_m) - g_1(\xi_m)\big)\right)^2 \\ &\quad \le C\sum_{j,k}\Big[\min\{|\beta_{j,k}^*|, \lambda_j^*\}^2 \\ &\qquad\quad + E\big[(\tilde{\beta}_{j,k} - \beta_{j,k}^*)^2 I_\Omega I_{\{|\tilde{\beta}_{j,k} - \beta_{j,k}^*| I_\Omega \ge \lambda_j^*/2\}}\big]\Big] + O\left(\frac{1}{n}\right) \\ &\quad \le C\sum_{j,k}\Big[\min\{|\beta_{j,k}|, \lambda_j^*\}^2 + (\beta_{j,k} - \beta_{j,k}^*)^2 \\ &\qquad\quad + E\big[(\hat{\beta}_{j,k} - \beta_{j,k})^2 I_\Omega I_{\{|\hat{\beta}_{j,k} - \beta_{j,k}| I_\Omega \ge \lambda_j^*/2\}}\big]\Big] + O\left(\frac{1}{n}\right) \\ &\quad \triangleq I_1(n) + I_2(n) + I_3(n) + O\left(\frac{1}{n}\right). \end{aligned}$$

Intuitively, the sum of $I_1(n)$ and $I_2(n)$ is the bias and $I_3(n)$ is the variance. The orders of $I_1(n)$ and $I_2(n)$ can be deduced by the properties and the definition of the Besov space and the order of $I_3(n)$ can be obtained using the inequality given in Lemma 5.5.

Clearly, from Lemma 2.2, $I_2(n) = O(n^{-2s_1/(2s_1+1)})$. Noting that $\lambda_{J_0-1}^* = \lambda_{J_0}^* = 0$, it follows from Definition 2 that

$$\begin{aligned} I_1(n) &\le C\sum_{j,k}|\beta_{j,k}|^{p_1}(\lambda_j^*)^{2-p_1} \\ &\le C\sum_{j=J_0}^{J-1}\left(\frac{j - J_0}{\sqrt{n}}\right)^{2-p_1} 2^{-j(s_1p_1 + p_1/2 - 1)} \\ &= O(n^{-2s_1/(2s_1+1)}) \end{aligned}$$

for $p_1 \le 2$ and

$$I_1(n) \le C \sum_{j=J_0}^{J-1} \sum_{k=0}^{2^j-1} \beta_{j,k}^2 \le \sum_{j=J_0}^{J-1} 2^{j(1-2/p_1)} \left(\sum_{k=0}^{2^j-1} |\beta_{j,k}|^{p_1} \right)^{2/p_1}$$

$$= O(n^{-2s_1/(2s_1+1)})$$

for $p_1 \ge 2$ uniformly for $g_t \in B_{p_t,q_t}^{s_t}(H_t)$ and $t = 1, 2$.

Now, we need to prove $I_3(n) = O(n^{-2s_1/(2s_1+1)})$ only. For $\tau > 2$, we have, by using Lemma 5.4,

$$I_3(n) \le C \sum_{j,k} \big(E\big(|\tilde{\beta}_{j,k} - \beta_{j,k}^*|^\tau I_\Omega \big) \big)^{2/\tau} \big(P\big(|\tilde{\beta}_{j,k} - \beta_{j,k}^*| I_\Omega \ge \lambda_j^*/2 \big) \big)^{(\tau-2)/\tau}$$

$$\le \frac{C}{n} \sum_{j,k} \big(P\big(|\tilde{\beta}_{j,k} - \beta_{j,k}^*| I_\Omega \ge \lambda_j^*/2 \big) \big)^{(\tau-2)/\tau}.$$

Denote

$$S_1(n) = \frac{1}{n\sqrt{N}} \sum_{m=0}^{N-1} \sum_{l=1}^{n} \sum_{i \in A_{m,l}} C_{m,l,i} \varepsilon_i \mathcal{W}_{j,k}(m) I_\Omega$$

and

$$S_2(n) = \frac{a_{jk}}{n} \sum_{i=1}^{n} g_2(X_{i,2}) I_\Omega.$$

Then, it follows from Lemma 5.3 that

$$I_3(n) \le \frac{C}{n} \left[\sum_{j,k} \big(P\big(|S_1(n)| \ge \lambda_j^*/6 \big) \big)^{(\tau-2)/\tau} + \sum_{j,k} \big(P\big(|S_2(n)| \ge \lambda_j^*/6 \big) \big)^{(\tau-2)/\tau} \right].$$

Note that $g_2(\cdot)$ is bounded. Using Bernstein's inequality, it is easy to obtain

$$\frac{C}{n} \sum_{j,k} \big(P\big(|S_2(n)| \ge \lambda_j^*/6 \big) \big)^{(\tau-2)/\tau} = O(n^{-2s_1/(2s_1+1)}).$$

In the remainder, we prove that

$$\frac{C}{n} \sum_{j,k} \big(P\big(|S_1(n)| \ge \lambda_j^*/6 \big) \big)^{(\tau-2)/\tau} = O(n^{-2s_1/(2s_1+1)}). \tag{5.8}$$

Let $\eta_{mt} = \sum_{l \in B_t} \sum_{i \in A_{m,l}} C_{m,l,i} \varepsilon_i \mathcal{W}_{j,k}(m) I_\Omega$. Then

$$S_1 = \frac{1}{n\sqrt{N}} \sum_{m=0}^{N-1} \sum_{t=0}^{N-1} \eta_{mt}.$$

Since the η_{mt} are correlated, we decompose $S_1(n)$ into blocks each with independent items. Denote

$$V(l_1, l_2) = \sum_{m \in I_{l_1}} \sum_{t \in I_{l_2}} \eta_{mt},$$

where, for an integer l_i, $I_{l_i} = \{2\sqrt{N}(l_i - 1), 2\sqrt{N}(l_i - 1) + 1, \ldots, 2\sqrt{N}l_i - 1\}$. Then

$$\begin{aligned} S_1(n) &= \frac{1}{n\sqrt{N}} \sum_{l_1=1}^{\sqrt{N}/2} \sum_{l_2=1}^{\sqrt{N}/2} V(l_1, l_2) \\ &= \frac{1}{n\sqrt{N}} \sum_{l_1=1}^{\sqrt{N}/4} \sum_{l_2=1}^{\sqrt{N}/4} \big(V(2l_1, 2l_2) + V(2l_1 - 1, 2l_2) \\ &\qquad + V(2l_1, 2l_2 - 1) + V(2l_1 - 1, 2l_2 - 1)\big) \\ &\triangleq T_1(n) + T_2(n) + T_3(n) + T_4(n). \end{aligned} \tag{5.9}$$

In order to prove (5.8), we need only to prove

$$\frac{C}{n} \sum_{j,k} \big(P\big(|T_i(n)| \geq \lambda_j^*/24\big)\big)^{(\tau-2)/\tau} = O(n^{-2s_1/(2s_1+1)})$$

for $i = 1, \ldots, 4$. The proofs of these four terms are similar. We give the proof of the first term only. Note that $V(2l_1, 2l_2)$ $(l_1, l_2 = 1, \ldots, \sqrt{N}/4)$ are independent of each other and $E(V(2l_1, 2l_2)) = 0$. We use Lemma 5.5 to evaluate the upper bound of $P(|T_i(n)| \geq \lambda_j^*/24)$. Denote $E_X(\cdot)$ as the conditional expectation for given $X_1, \ldots, X_n$. If we use (5.7), it is not difficult to verify that

$$E_X\big(V(2l_1, 2l_2)\big)^2 \leq CN^{1/2}n \sum_{m \in I_{l_1}} w_{j,k}^2(m) \triangleq \sigma_{l_1,l_2}^2 \tag{5.10}$$

with

$$\sum_{l_1=1}^{\sqrt{N}/4} \sum_{l_2=1}^{\sqrt{N}/4} \sigma_{l_1,l_2}^2 \leq C_2 Nn \triangleq \mathbf{B}_n$$

and, for any $h > 0$,

$$\begin{aligned} E_X|V(2l_1, 2l_2)|^h &\leq (4N)^{h-1} \sum_{m \in I_{l_1}} \sum_{t \in I_{l_2}} E_x|\eta_{ml}|^h \\ &\leq \tfrac{1}{2}h!H_n^{h-2}\sigma_{l_1,l_2}^2, \end{aligned} \tag{5.11}$$

where $H_n = H_1\sqrt{nN}$ and H_1 is a constant. Let $\lambda = \sqrt{N}n\lambda_j^*/24$. Then, there exists a constant, K', such that

$$\lambda > \frac{\mathbf{B}_n}{H_n} \quad \text{and} \quad k^* = \frac{K}{24H_1}\frac{\tau - 2}{\tau} > \ln 2$$

for $K > K'$ and $j > J_0$. Using Lemma 5.5, we have, for $K > K'$,

$$\frac{1}{n}\sum_{j,k}\bigl(P\bigl(|T_1(n)| \geq \lambda_j^*/24\bigr)\bigr)^{(\tau-2)/\tau}$$
$$= O(n^{-2s_1/(2s_1+1)}) + \frac{1}{n}\sum_{j=J_0+1}^{J-1}\sum_{k=0}^{2^j-1} e^{-k^*(j-J_0)} = O(n^{-2s_1/(2s_1+1)}).$$

We have completed the proof of Theorem 1. □

PROOF OF THEOREM 2.

$$E\int_0^1 \bigl(\hat{g}_1(x) - g_1(x)\bigr)^2\,dx$$
$$= \sum_{m=0}^{N-2}\int_{\xi_m}^{\xi_{m+1}} \bigl(\hat{g}_1(x) - g_1(x)\bigr)^2\,dx$$
$$+ \left[\int_0^{1/2N} + \int_{1-1/2N}^1\right]\bigl(\hat{g}_1(x) - g_1(x)\bigr)^2\,dx$$
$$\triangleq T_1(n) + T_2(n).$$

We prove only $T_1(n) = O(n^{-2s_1/(2s_1+1)})$ here. Using the same argument, we get $T_2(n) = O(n^{-2s_1/(2s_1+1)})$.

Decompose $T_1(n)$ as

$$T_1(n) \leq C\sum_{m=0}^{N-2}\int_{\xi_m}^{\xi_{m+1}}\bigl[\bigl(\hat{g}_1(x) - \hat{g}_1(\xi_m)\bigr)^2$$
$$+ \bigl(\hat{g}_1(\xi_m) - g_1(\xi_m)\bigr)^2 + \bigl(g_1(\xi_m) - g_1(x)\bigr)^2\bigr]\,dx$$
$$\triangleq S_1(n) + S_2(n) + S_3(n).$$

Following the proof of Theorem 1, $S_2(n) = o(n^{-2s_1/(2s_1+1)})$. Denote $P_{\xi_m}(\cdot)$ as the optimal polynomial corresponding to $g_1(\cdot)$. Then,

$$S_3(n) \leq C\sum_{m=0}^{N-2}\int_{\xi_m}^{\xi_{m+1}}\bigl[\bigl(g_1(\xi_m) - P_{\xi_m}(\xi_m)\bigr)^2$$
$$+ \bigl(P_{\xi_m}(\xi_m) - P_{\xi_m}(x)\bigr)^2 + \bigl(P_{\xi_m}(x) - g_1(x)\bigr)^2\bigr]\,dx$$
$$\triangleq S_{31}(n) + S_{32}(n) + S_{33}(n).$$

If we use Lemma 2.1, it is easy to show that $S_{32}(n) = O(n^{-2s_1/(2s_1+1)})$. Furthermore, $S_{31}(n) \leq \frac{C}{N}\sum_{m=0}^{N-2}(\theta_J^{(M)}(\xi_m))^2$. Using Proposition 2 and inequality (5.6), we have

$$S_{31}(n) \leq \frac{C}{N}\left(\sum_{m=0}^{N-2}\bigl(\theta_J^{(M)}(\xi_m)\bigr)^{p_1}\right)^{2/p_1} \leq C2^{-J(2s_1-2/p_1+1)} = O(n^{-2s_1/(2s_1+1)})$$

for $p_1 \le 2$ and

$$S_{31}(n) \le C\left(\frac{1}{N}\sum_{m=0}^{N-2}\left(\theta_J^{(M)}(\xi_m)\right)^{p_1}\right)^{2/p_1} \le C2^{-2s_1J} = O(n^{-2s_1/(2s_1+1)})$$

for $p_1 > 2$. Similarly, we show that $S_{33}(n) = O(n^{-2s_1/(2s_1+1)})$ and thus $S_3(n) = O(n^{-2s_1/(2s_1+1)})$.

By the same argument as the proof for the order of $S_3(n)$, it follows $S_1(n) = O(n^{-2s_1/(2s_1+1)})$. We thus complete the proof of Theorem 2. □

PROOF OF THEOREM 3. Using the same notation as the proof of Theorem 1, we have

$$E\left(\frac{1}{N}\sum_{m=0}^{N-1}\left(\hat{g}_{1A}(\xi_m) - g_1(\xi_m)\right)\right)^2 \triangleq I_1(n) + I_2(n) + I_3(n) + O\left(\frac{1}{n}\right).$$

By the same argument as the proof of Theorem 1, we find

$$I_2(n) = O(n^{-2s_1/(2s_1+1)}) \quad \text{and} \quad I_3(n) = O(n^{-2s_1/(2s_1+1)}).$$

Define j_2 by

$$2^{j_2} \asymp \left(\frac{n}{t_n}\right)^{1/(2s_1+1)}.$$

We then decompose $I_1(n)$ as

$$\begin{aligned} I_1(n) &= \sum_{j=J_0}^{j_2}\sum_k \min\{\beta_{j,k}, \lambda_j^*\}^2 + \sum_{j=j_2}^{J-1}\sum_k \min\{\beta_{j,k}, \lambda_j^*\}^2 \\ &\triangleq I_{11}(n) + I_{12}(n). \end{aligned}$$

Clearly,

$$I_{11}(n) \le C\sum_{j=J_0}^{j_2}(\lambda_j^*)^2 2^j = O\left(\frac{t_n}{n}\right)^{2s_1/(2s_1+1)}.$$

Similar to the proof for the order of $I_1(n)$ in Theorem 1, we have

$$I_{12}(n) = O\left(\frac{t_n}{n}\right)^{2s_1/(2s_1+1)}.$$

We thus complete the proof of the first half of Theorem 3 [i.e., (4.1)]. Using the same argument as in the proof of Theorem 2, we prove the second half of Theorem 3 [i.e., (4.2)]. □

APPENDIX

Proofs of lemmas. In this Appendix, we briefly prove all lemmas. The detailed proofs can be found on the web page of the first author (http://www.math.mtu.edu/~shuzhang).

PROOF OF LEMMA 2.1. If $0 < s < 1$, then $M = 0$ and thus Lemma 2.1 is obviously true. Thus, we assume $s \geq 1$. Using Proposition 1, we have

$$\frac{1}{N}\sum_{m=1}^{N}\sup_{y,y'\in B(\xi_m,C'/N)}\left|P_{\xi_m}(y)-P_{\xi_m}(y')\right| \leq \frac{C}{N}\sum_{t=1}^{M}\sum_{m=1}^{N}\theta_J^{(M-t)}(\xi_m).$$

Noting $B_{p,q}^{s} \subset B_{p,q}^{M-t+1}$ and using Propositions 1 and 2, we can verify that

$$\frac{1}{N}\sum_{m=1}^{N}\theta_J^{(M-t)}(\xi_m) \leq C\frac{J}{N}$$

for any integer t satisfying $1 \leq t \leq M$. Then Lemma 2.1 follows. □

PROOF OF LEMMA 2.2. Let $M = [s]$. Let $P_{\xi_m}(\cdot)$ denote the optimal polynomial corresponding to $f(x)$. It follows from Proposition 2, Lemma 2.1 and inequality (5.6) that

$$\begin{aligned}
&\left(\alpha_{Jm}-\frac{1}{\sqrt{N}}f(\xi_m)\right)^2 \\
&\quad\leq\left[\frac{1}{\sqrt{N}}\int_{-L}^{L}|\phi(y)|\left(\left|f\left(\frac{y-1/2}{N}+\xi_m\right)-P_{\xi_m}\left(\frac{y-1/2}{N}+\xi_m\right)\right|\right.\right. \\
&\qquad\qquad+\left|P_{\xi_m}\left(\frac{y-1/2}{N}+\xi_m\right)-P_{\xi_m}(\xi_m)\right| \\
&\qquad\qquad\left.\left.+\left|P_{\xi_m}(\xi_m)-f(\xi_m)\right|\right)dy\right]^2 \\
&\quad\leq\frac{C}{N}\sum_{m=1}^{N}\left[\left(\theta_{J-j_0}^{(M)}(\xi_m)\right)^2+\left(\sup_{x,y\in B(\xi_m,C'/N)}\left|P_{\xi_m}(x)-P_{\xi_m}(y)\right|\right)^2\right] \\
&\quad=O\left(\frac{J^2}{N}\right).
\end{aligned}$$

Furthermore, since β and β^* are orthogonal transformations of α_J and $(f(\xi_m)/\sqrt{N})_{m=0}^{N-1}$, respectively, we have $\sum_{j,k}(\beta_{j,k}-\beta_{j,k}^*)^2 \leq C\frac{J^2}{N}$. □

PROOF OF LEMMA 5.1. By Bernstein's inequality [page 192 in Pollard (1984)], it is easy to show that, for any fixed x, there exist positive constants,

$C_1(\delta)$ and $C_2(\delta)$, independent of x such that, for any $h > 0$,

$$\begin{aligned} &P\left(\frac{N}{n}|A_{x,\delta}| \geq C_2(\delta)\right) = O(n^{-h}) \quad \text{and} \\ &P\left(\frac{N}{n}|A_{x,\delta}| \leq C_1(\delta)\right) = O(n^{-h}). \end{aligned} \tag{A.1}$$

Note that we can find discrete points $x_l, x_l' \in [0,1]^2$ such that $\bigcup_{x\in[0,1]^2}(\frac{N}{n}|A_{x,\delta}| \geq C) \subset \bigcup_l(\frac{N}{n}|A_{x_l,\delta_1}| \geq C)$ and $\bigcup_{x\in[0,1]^2}(\frac{N}{n}|A_{x,\delta}| \leq C) \subset \bigcup_l(\frac{N}{n}|A_{x_l,\delta_2}| \leq C)$ for choosing suitable constants for δ_1 and δ_2. Then, Lemma 5.1 follows from (A.1). □

PROOF OF LEMMA 5.2. Denote $V_{x,\delta} = \sum_{i\in A_{x,\delta}} Z_{x,i} Z_{x,i}'$. Using an argument similar to the proof of Lemma 1 in Zhang, Wong and Zheng (2002), we can show that, for any fixed x and constant δ, there exists a constant K_δ such that, for any $h > 0$,

$$P\big(\lambda_{\min}(V_{x,\delta}) \leq K_\delta |A_{x,\delta}|\big) = O(n^{-h}), \tag{A.2}$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a matrix, A. Similar to the idea used in the proof of Lemma 5.1, we may find some discrete points $x_l \in [0,1]^2$ such that $\bigcup_{x\in[0,1]^2}(\lambda_{\min}(V_{x,\delta}) \leq C|A_{x,\delta}|) \subset \bigcup_l(\lambda_{\min}(V_{x_l,\delta_1}) \leq C|A_{x_l,\delta_1}|)$. Then Lemma 5.2 follows from (A.2). □

PROOF OF LEMMA 5.3. We decompose $(\tilde{\beta}_{j,k} - \beta_{j,k}^*)I_\Lambda$ as

$$\begin{aligned} &(\tilde{\beta}_{j,k} - \beta_{j,k}^*)I_\Lambda \\ &\quad = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \mathcal{W}_{j,k}(m)\big(\breve{g}_1(\xi_m) - g_1(\xi_m)\big)I_\Lambda \\ &\quad = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \mathcal{W}_{j,k}(m)\frac{1}{n}\sum_{l=1}^{n}\sum_{i\in A_{m,l}} C_{m,l,i}\big[\big(g_1(X_{i,1}) - g_1(\xi_m)\big) \\ &\qquad\qquad + g_2(X_{i,2}) + \varepsilon_i\big]I_\Lambda \\ &\quad \triangleq I_1(n) + I_2(n) + \frac{1}{n\sqrt{N}} \sum_{m=0}^{N-1}\sum_{l=1}^{n}\sum_{i\in A_{m,l}} C_{m,l,i}\varepsilon_i \mathcal{W}_{j,k}(m)I_\Lambda. \end{aligned}$$

Denote $P_x(\cdot)$ as the optimal polynomial of degree M $(= [s] \leq D)$ corresponding to $g_1(\cdot)$. Note that $\hat{g}(x)$ is fitted with an additive polynomial of degree D in the neighborhood of x. Thus, $\sum_{i\in A_x} C_{m,l,i} P_x(X_{i,t}) = P_x(x_t)$. Using Propositions 1, 2

and inequality (5.6), we have

$$|I_1(n)| = \frac{1}{n\sqrt{N}} \sum_{m=0}^{N-1} |\mathcal{W}_{j,k}(m)| \sum_{l=1}^{n} \sum_{i \in A_{m,l}} \big|C_{m,l,i}\big[\big(g_1(X_{i,1}) - P_{\xi_m}(X_{i,1})\big) + \big(P_{\xi_m}(\xi_m) - g_1(\xi_m)\big)\big]\big| I_\Lambda$$

$$\leq \frac{1}{n\sqrt{N}} \sum_{m=0}^{N-1} |\mathcal{W}_{j,k}(m)| \sum_{l=1}^{n} \sum_{i \in A_{m,l}} |C_{m,l,i}| \theta_{J/2}^{(M)}(\xi_m) I_\Lambda$$

$$= o\left(\frac{1}{\sqrt{n}}\right).$$

Similarly, we show $I_2(n) \leq \frac{a_{jk}}{n} \sum_{l=1}^{n} g_2(X_{l,2}) I_\Lambda$ and thus complete the proof of Lemma 5.3. □

PROOF OF LEMMA 5.4. We prove only $E|(\tilde{\beta}_{j,k} - \beta_{j,k}^*) I_\Omega|^\tau = O(n^{-\tau/2})$ here. With reference to the notation in the proof of Theorem 1, it follows from Lemma 5.3 that

$$(\tilde{\beta}_{j,k} - \beta_{j,k}^*) I_\Omega = S_1(n) + S_2(n) + o\left(\frac{1}{\sqrt{n}}\right).$$

Using the same argument as the proof of Theorem 1 and using Equations (5.10) and (5.11) [noting that the proofs of (5.10) and (5.11) do not require Lemma 5.4], we have $E|S_1(n)|^\tau = O(n^{-\tau/2})$. Thus, Lemma 5.4 follows by noting the obvious fact that $E|S_2(n)|^\tau = O(n^{-\tau/2})$. □

Acknowledgments. The authors thank the Associate Editor and a referee for their constructive comments and Dr. Virginia Unkefer for her detailed corrections of English usage.

REFERENCES

COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228.

DONOHO, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26** 943–971.

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518.

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

HALL, P., KERKYACHARIAN, G. and PICARD, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26** 922–942.

HALL, P. and TURLACH, B. A. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Statist.* **25** 1912–1925.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

LINTON, O. B. (1996). Estimation of additive regression models with known links. *Biometrika* **83** 529–540.

LINTON, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–473.

LINTON, O. B. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–100.

NEUMANN, M. H. and VON SACHS, R. (1995). Wavelet thresholding: Beyond the Gaussian i.i.d. situation. *Wavelets and Statistics. Lecture Notes in Statist.* **103** 301–329. Springer, Berlin.

NIELSEN, J. P. and LINTON, O. B. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 217–222.

OPSOMER, J. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

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.

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

TJØSTHEIM, D. and AUESTAD, B. H. (1994). Nonparametric identification of nonlinear time series: Projections. *J. Amer. Statist. Assoc.* **89** 1398–1409.

TRIEBEL, H. (1992). *Theory of Function Spaces II*. Birkhäuser, Boston.

ZHANG, S., WONG, M.-Y. and ZHENG, Z. (2002). Wavelet threshold estimation of a regression function with random design. *J. Multivariate Anal.* **80** 256–284.

DEPARTMENT OF MATHEMATICAL SCIENCES
MICHIGAN TECHNOLOGICAL UNIVERSITY
HOUGHTON, MICHIGAN 49931
AND
DEPARTMENT OF MATHEMATICS
HEILONGJIANG UNIVERSITY
HARBIN 150080
CHINA

DEPARTMENT OF MATHEMATICS
HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
CLEAR WATER BAY, KOWLOON
HONG KONG
E-MAIL: mamywong@ust.hk