COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station

New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 925

ADDITIVE INTERACTIVE REGRESSION MODELS: CIRCUMVENTION OF THE CURSE OF DIMENSIONALITY

by

Donald W.K. Andrews and Yoon-Jae Whang

September, 1989

ABSTRACT

This paper considers series estimators of additive interactive regression (AIR) models. AIR models are nonparametric regression models that generalize additive regression models by allowing interactions between different regressor variables. They place more restrictions on the regression function, however, than do fully nonparametric regression models. By doing so, they attempt to circumvent the curse of dimensionality that afflicts the estimation of fully nonparametric regression models.

In this paper, we present a finite sample bound and asymptotic rate of convergence results for the mean average squared error of series estimators that show that AIR models do circumvent the curse of dimensionality. The rate of convergence of these estimators is shown to depend on the order of the AIR model and the smoothness of the regression function, but not on the dimension of the regressor vector. Series estimators with fixed and data–dependent truncation parameters are considered.

JEL Classification Number: 211.

Keywords: Additive interactive regression model, cross–validation, curse of dimensionality, generalized cross–validation, mean average squared error, nonparametric estimation, nonparametric regression, series estimator.

1. INTRODUCTION

This paper considers series estimators of *additive interactive regression* (AIR) models. The paper focuses on the extent to which these estimators circumvent the "curse of dimensionality" that afflicts estimators of fully nonparametric regression models.

AIR models are also known in the literature as *interaction spline* models. Their estimation using splines has been analyzed by Barry (1983, 1986), Wahba (1986), Gu, Bates, Chen, and Wahba (1988), and Chen (1988). A special case of the AIR model is the additive regression model that has been considered by Orcutt *et al.* (1961, p. 62), Stone (1985), Hastie and Tibshirani (1986, 1987), and Buja, Hastie, and Tibshirani (1989). AIR models allow for interactions between the elements of the regressor vector. Such interactions are precluded in additive regression models.

When the number of regressors d is large, fully nonparametric regression models do not place enough restrictions on the regression function to permit reasonably accurate estimation unless the sample size is extremely large. This is illustrated by the fact that the fastest possible rate of convergence of estimators in such models is $n^{-2/(4+d)}$ when the regression function is assumed to be twice differentiable (see Stone (1980, 1982)). This rate is very slow if d is in the range of five to fifteen, which is quite common in econometrics. Furthermore, if one is interested in estimating derivatives of the regression function, then the best possible rate is even slower.

The difficulty in estimating fully nonparametric regression models is that one has to estimate a high dimensional surface when d is large. Additive regression and AIR models allow one to replace the estimation of such a surface with the estimation of several low dimensional surfaces. This yields large efficiency gains if d is large and the true regression function is of the additive or AIR form. In consequence, AIR models appear to be well suited to many econometric applications, since many econometric applications have too many regressor variables for fully nonparametric regression methods to be effective.

Stone (1985) has shown that it is possible to achieve the same rate of convergence in an additive regression model with any number of regressors as in a nonparametric regression model with only one regressor. Based on this result, an obvious speculation is that it is possible to achieve the same rate of convergence in an AIR model with an arbitrary number of regressors, but with interactions between at most A of these regressors, as in a nonparametric regression model with A regressors. If true, one can say that AIR models circumvent the curse of dimensionality, since the rate of convergence of an estimator of an AIR model is not necessarily related to the dimension of the regressor vector.

In fact, Chen (1988) has established the above result for a particular form of AIR model using spline estimators. The model he considers is one in which the errors are independent and identically distributed and the regressors are from a non-stochastic "tensor product design."[2] For many applications, however, this regressor design is too restrictive.

In this paper, we establish the above result for a general class of AIR models using series estimators. The regressors are not restricted as in Chen (1988) and the errors may be independent non–identically distributed (inid). The criterion of performance used here for the rate of convergence results is mean average squared error (MASE) as in Chen (1988). In contrast, Stone (1985) considers mean integrated squared error.

We note that series estimators of AIR models have already been discussed in Andrews (1989a). The latter paper gives conditions under which such estimators are pointwise consistent and asymptotically normal.

Regarding the comparison of series and spline estimators of AIR models, little research has been conducted. Series estimators are much more tractable computationally, especially when there are multiple smoothing paramters, large numbers of regressors, and large sample sizes. On the other hand, spline estimators have the attribute of being solutions to an explicit variational problem and have a Bayesian interpretation.

The remainder of this paper is organized as follows: Section 2 defines AIR models and series estimators of these models. Section 3 presents a finite sample result in which the MASE of a series estimator of an AIR model of order A is bounded by the sum of MASEs of series estimators of several fully nonparametric regression models each with regressor vector of dimension $\leq A$ ($\leq d$). Section 4 states rate of convergence results for series estimators of AIR models when the estimators are based on fixed truncation sequences. These results illustrate the circumvention of the curse of dimensionality by AIR models. Section 5 discusses the asymptotic optimality of several data–dependent truncation procedures and the rate of convergence of series estimators defined using these procedures. Specifically, generalized C_L, generalized cross–validation, and cross–validation are considered.

2. SERIES ESTIMATORS OF AIR MODELS

An AIR model is defined by

$$Y_i = g(x_i) + U_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $Y_i, U_i \in R$, $x_i \in \mathcal{X} \subset R^d$, and $EU_i = 0$ and where $g(\cdot)$ is known to be of the form

$$g(x_i) = \sum_{a=1}^{A} \sum_{b=1}^{B(a)} g_{ab}(x_i). \tag{2.2}$$

Here, $g_{ab}(x_i)$ is an unknown function that depends on only " a " ($\leq d$) elements of x_i for each $b = 1, \ldots, B(a)$. For example, one might have $g_{1b}(x_i) = g^*_{1b}(x_{i1})$ and $g_{2b}(x_i) = g^*_{2b}(x_{i1}, x_{i2})$, where $x_i = (x_{i1}, \ldots, x_{id})'$.

The order of an AIR model is given by A. If $A = 1$, the model is an additive regression model. If $A > 1$, the AIR model allows for interactions between regressors. For example, a second order AIR model allows for interactions between (some) pairs of regressors, but not between triplets. If $A = d$, the model is a fully nonparametric

regression model. When $A < d$, the model (2.1)–(2.2) imposes restrictions on the regression function $g(\cdot)$ that should permit more efficient estimation of it than in a fully nonparametric regression model.

Note that our attention here is on the estimation of $g(\cdot)$ rather than the component functions $\{g_{ab}\}$. Clearly, normalization conditions need to be added to identify the functions $\{g_{ab}\}$, if the estimation of these functions is of interest. Also note that A and B(a) are positive integers not exceeding d and d!/(a!(d–a)!) respectively.

A series estimator of $g(\cdot)$ is constructed using a series approximation $\sum_{c=1}^{k_{ab}} z_{abc}(x_i)\theta_{abc}$ of each function $g_{ab}(x_i)$, where $\{\theta_{abc}\}$ are unknown coefficients to be estimated, $z_{abc}(\cdot)$ is a known function that depends on the same elements of x_i as does $g_{ab}(\cdot)$ for all $c = 1, \ldots, k_{ab}$, and k_{ab} is a truncation parameter. Examples of approximating functions $z_{abc}(\cdot)$ include: trigonometric, Fourier flexible form (see Gallant (1981, p. 219)), and polynomial functions. The truncation parameter k_{ab} implicitly depends on the sample size n. It may be fixed, as in Sections 3 and 4 below, or data-dependent, as in Section 5.

Let I_+ denote the sets of non–negative integers. Let

$$D = \sum_{a=1}^{A} B(a),\ k = (k_{11}, \ldots, k_{1B(1)}, k_{21}, \ldots, k_{AB(A)})' \in I_+^D,$$

$$Y = (Y_1, \ldots, Y_n)',\ U = (U_1, \ldots, U_n)', \text{ and} \tag{2.3}$$

$$Z = (Z_k(x_1), \ldots, Z_k(x_n))' \in R^{n \times k}, \underset{\sim}{1}$$

where the i–th row of the matrix Z, $Z_k(x_i)$, is given by the elements of $\{z_{abc}(x_i)$: $c = 1, \ldots, k_{ab}; b = 1, \ldots, B(a); a = 1, \ldots, A\}$ and $\underset{\sim}{1}$ is a D dimensional vector of ones.

Let θ be the $k' \underset{\sim}{1}$–vector with elements given by $\{\theta_{abc} : c = 1, \ldots, k_{ab};$ $b = 1, \ldots, B(a);\ a = 1, \ldots, A\}$. The least squares (LS) estimator of θ is

$$\hat{\theta} = (Z'Z)^{-}Z'Y\,, \tag{2.4}$$

where $(\cdot)^{-}$ denotes some g–inverse. The corresponding series estimator $\hat{g}$ of g is

$$\hat{g}(\cdot) = Z_k(\cdot)'\hat{\theta}\,. \tag{2.5}$$

Various properties of $\hat{g}$ are investigated in Sections 3–5 below.

For notational simplicity, we adopt the following conventions in the remainder of the paper: $\sum_a \sum_b$ abbreviates $\sum_{a=1}^{A} \sum_{b=1}^{B(a)}$; $\forall a, b$ abbreviates $\forall b = 1, \ldots, B(a)$, $\forall a = 1, \ldots, A$; all limits are taken as $n \to \infty$; $a_n \sim b_n$ denotes that a_n/b_n is bounded away from zero and infinity over $n \geq 1$; and for any function g^* from $\mathcal{X}$ to R, $\underset{\sim}{g}^*$ denotes the n–vector $(g^*(x_1), \ldots, g^*(x_n))'$.

3. AIR MODELS VERSUS FULLY NONPARAMETRIC REGRESSION MODELS

In this section, we relate the finite sample MASE of series estimators of AIR models with those of fully nonparametric regression models. The results have immediate implications regarding the circumvention of the curse of dimensionality by AIR models. They also have implications for the rate of convergence results given in Sections 4 and 5 below.

Consider the following models:

$$Y_i = \sum_a \sum_b g_{ab}(x_i) + U_i \ \ (= g(x_i) + U_i)\,,\ i = 1, \ldots, n\,,\ \text{and} \tag{3.1}$$

$$Y_{iab} = g_{ab}(x_i) + U_{iab}\,,\ i = 1, \ldots, n\,,\ \forall a, b\,, \tag{3.2}$$

where $Y_i = \sum_a \sum_b Y_{iab}$ and $U_i = \sum_a \sum_b U_{iab}$ for $i = 1, \ldots, n$. $\{U_{iab}\}$ are mean zero, variance σ^2_{iab} random variables, independent across i, a, b. $\{x_i\}$ are non–random regressor vectors in $\mathcal{X} \subset R^d$. The fully nonparametric regression models of (3.2) are

considered for theoretical purposes only. They generate the AIR model of (3.1) by summation.

Let $\hat{g}$ be a series estimator of g in (3.1) based on the series functions $\{z_{abc}(\cdot) : \forall c = 1, \ldots, k_{ab}; \forall a, b\}$. For each a, b, let $\hat{g}_{ab}$ be the corresponding series estimators of g_{ab} based on the functions $\{z_{abc}(\cdot) : \forall c = 1, \ldots, k_{ab}\}$. The *mean average squared error* (MASE) of $\hat{g}$ is defined to be

$$\text{MASE}(\hat{g},g) = n^{-1}E\|\hat{g}-g\|^2 \left[= n^{-1} \sum_{i=1}^{n} (\hat{g}(x_i) - g(x_i))^2 \right], \tag{3.3}$$

where $\|\cdot\|$ denotes the Euclidean norm. The MASE of $\hat{g}_{ab}$ for estimating g_{ab} is defined analogously $\forall a, b$.

We show that the MASE of $\hat{g}$ in the AIR model (3.1) can be bounded above by a constant times the sum over a, b of the MASE of $\hat{g}_{ab}$ in the fully nonparametric regression model of (3.2). The latter MASE does not depend on the dimension of x_i, but rather, on the number, a, of elements of x_i upon which g_{ab} depends $\forall a, b$. Thus, the bound on the MASE of the series estimator $\hat{g}$ in the AIR model is independent of the dimension of x_i. In this (non–asymptotic) sense, the estimator circumvents the curse of dimensionality.

THEOREM 1. *Let* $\hat{g}(\cdot)$ *and* $\hat{g}_{ab}(\cdot)$ *be as defined above. Assume* $0 < \tau_* = \inf_{i,a,b} \sigma^2_{iab} \leq \tau^* = \sup_{i,a,b} \sigma^2_{iab} < \infty$. *Then,*

$$\text{MASE}(\hat{g},g) \leq D \frac{\tau^*}{\tau_*} \sum_a \sum_b \text{MASE}(\hat{g}_{ab}, g_{ab}),$$

where $D = \sum_{a=1}^{A} B(a)$.

Comments. 1. The upper bound in Theorem 1 is sharp. That is, if D is replaced by any $D' < D$, then the inequality does not necessarily hold. (To see this, consider the case where σ^2_{iab} does not depend on i, a, b and the estimators $\hat{g}$ and $\hat{g}_{ab}$ $\forall a, b$ do not incur any bias.)

2. Although Theorem 1 is a finite sample result, it has clear implications for asymptotic rate of convergence results for series estimators of AIR models. In particular, it points out those characteristics of an AIR model and its estimator that serve to determine the estimator's rate of convergence.

Proof of Theorem 1. $\mathrm{MASE}(\hat{g},g)$ and $\sum_a \sum_b \mathrm{MASE}(\hat{g}_{ab}, g_{ab})$ can be decomposed into squared bias and variance terms:

$$\mathrm{MASE}(\hat{g},g) = \tfrac{1}{n}\|g - P_Z g\|^2 + \tfrac{1}{n}E\|P_Z U\|^2 \quad \text{and} \tag{3.4}$$

$$\sum_a \sum_b \mathrm{MASE}(\hat{g}_{ab},g_{ab}) = \tfrac{1}{n}\sum_a \sum_b \|g_{ab} - P_{Z_{ab}} g_{ab}\|^2 + \tfrac{1}{n}\sum_a \sum_b E\|P_{Z_{ab}} U_{ab}\|^2 , \tag{3.5}$$

where $U_{ab} = (U_{1ab}, \ldots, U_{nab})'$, $P_{Z_{ab}} = Z_{ab}(Z'_{ab}Z_{ab})^{-}Z'_{ab}$, and Z_{ab} is the $n \times k_{ab}$ matrix whose i–th row is $(z_{ab1}(x_i), \ldots, z_{abk_{ab}}(x_i))$.

First, we compare the bias terms. We have

$$\|g - P_Z g\| = \|\sum_a \sum_b (g_{ab} - P_Z g_{ab})\| \le \sum_a \sum_b \|g_{ab} - P_Z g_{ab}\| \le \sum_a \sum_b \|g_{ab} - P_{Z_{ab}} g_{ab}\| \tag{3.6}$$

using the fact that Z_{ab} consists of columns of Z. Therefore,

$$\|g - P_Z g\|^2 \le (\sum_a \sum_b \|g_{ab} - P_{Z_{ab}} g_{ab}\|)^2 \le D \sum_a \sum_b \|g_{ab} - P_{Z_{ab}} g_{ab}\|^2 \tag{3.7}$$

using the Cauchy–Schwartz inequality.

Next, we compare the variance terms. Let $\Omega = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ and $\Omega_{ab} = \mathrm{diag}\{\sigma_{1ab}^2, \ldots, \sigma_{nab}^2\}$. Note that $\Omega = \sum_a \sum_b \Omega_{ab}$. We have

$$E\|P_Z U\|^2 = \mathrm{tr}(P_Z \Omega) = \sum_a \sum_b \mathrm{tr}(P_Z \Omega_{ab}) \le \tau^* \sum_a \sum_b \mathrm{tr}\, P_Z = D\tau^* k' 1 . \tag{3.8}$$

On the other hand,

$$\sum_a \sum_b E\|P_{Z_{ab}} U_{ab}\|^2 = \sum_a \sum_b \mathrm{tr}(P_{Z_{ab}} \Omega_{ab}) \geq \tau_* k' \underset{\sim}{1} . \tag{3.9}$$

Equations (3.4)–(3.9) combine to give the desired result. □

4. RATES OF CONVERGENCE

Next we present rate of convergence results for series estimators of AIR models. This section considers estimators based on fixed truncation parameters. The results show that the rates of convergence depend on the order A of the AIR model and the smoothness of the functions g_{ab}, but not on the dimension d of x_i.

Let $\mathcal{G}$ be a class of differentiable functions from $\mathcal{X}$ to R. For any $g_1 \in \mathcal{G}$, let $\|g_1\|_{q,\infty,\mathcal{X}}$ denote the *supremum Sobolev norm* of derivative order q, for some $q \geq 0$. That is,

$$\|g_1\|_{q,\infty,\mathcal{X}} = \sum_{\lambda:|\lambda|\leq q} \sup_{x\in\mathcal{X}} |D^\lambda g_1(x)| , \tag{4.1}$$

where $\lambda = (\lambda_1, \ldots, \lambda_d)' \in I_+^d$, $|\lambda| = \sum_{j=1}^d \lambda_j$, and $D^\lambda g_1(x) = \dfrac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} \cdots \partial x_d^{\lambda_d}} g_1(x)$.

If partial derivatives of $g_1(x)$ do not exist up to order q, then $\|g_1\|_{q,\infty,\mathcal{X}} = \infty$.

Define the *Sobolev smoothness index* of a function $g_1 \in \mathcal{G}$ to be

$$S(g_1) = \max\{v \geq 0 : \|g_1\|_{v,\infty,\mathcal{X}} < \infty\} . \tag{4.2}$$

We consider the AIR model defined by (2.1) and (2.2) and introduce the following assumptions:

ASSUMPTION A.1. $\{U_i : i \geq 1\}$ *are mean zero square integrable* rv's *with* $0 < \inf_{i\geq 1} \sigma_i^2 \leq \sup_{i\geq 1} \sigma_i^2 < \infty$ *and* $\{x_i : i \geq 1\}$ *are non–stochastic regressor vectors in* $\mathcal{X} \subset R^d$.

ASSUMPTION A.2. *For each* a , b , $\{z_{abc}(\cdot) : c \geq 1\}$ *satisfies: For all* $m \geq 1$, *there exists* $\theta_m = (\theta_{m1}, \ldots, \theta_{mm})' \in R^m$ *(which depends on* a , b *in general) such that*

$$m^{\alpha_{ab}} \left\| \sum_{c=1}^{m} z_{abc}(\cdot)\theta_{mc} - g_{ab}(\cdot) \right\|_{0,\infty,\mathcal{X}} \to 0 \quad as \quad m \to \infty$$

for some $0 \leq \alpha_{ab} < S(g_{ab})/a$.

If the regressors $\{x_i\}$ are random, they can be conditioned on. In this case, Assumption A.1 and the following results hold conditionally on $\{x_i\}$ for any sequence $\{x_i\}$.

By Corollary 1 of Edmunds and Moscatelli (1977, p. 28), Assumption A.2 holds for all $0 \leq \alpha_{ab} < S(g_{ab})/a$ $\forall a, b$, if the series functions are trigonometric or Fourier flexible form functions, $\mathcal{X}$ has closure that lies on $(0, 2\pi)^d$, and $\mathcal{X}$ has minimally smooth boundary. See Edmunds and Moscatelli (1977) for the definition of a minimally smooth boundary. Examples of sets with this property include all convex sets. Note that if $\mathcal{X}$ is any bounded subset of R^d , the regressors can be rescaled such that $\mathcal{X}$ has closure that lies on $(0, 2\pi)^d$.

When $A = 1$ (additive model), Assumption A.2 holds for all $0 \leq \alpha_{ab} < S(g_{ab})$ $\forall a, b$, if the series functions are polynomials and $\mathcal{X}$ is a closed, bounded, connected subset of R^d . This follows from Theorem 3.2 of Powell (1981, p. 26).

THEOREM 2. *Suppose Assumptions* A.1 *and* A.2 *hold.*

(a) *If* $k_{ab}/n \to 0$ *and* $k_{ab} \to \infty$ $\forall a, b$, *then* $MASE(\hat{g},g) \to 0$.

(b) *If* $k_{ab} \sim n^{r_{ab}}$ *for some* $0 < r_{ab} < 1$ $\forall a, b$, *then* $MASE(\hat{g},g) = O(n^{-r})$, *where* $r = \min_{a,b} \min\{1 - r_{ab}, 2\alpha_{ab}r_{ab}\}$.

(c) *The choice of* r_{ab} *that maximizes the rate of convergence* r *in part* (b) *is* $r_{ab} = 1/(2\alpha_{ab}+1)$. *In this case,* $r = \min_{a,b} 2\alpha_{ab}/(2\alpha_{ab}+1)$ $(< \min_{a,b} 2S(g_{ab})/(2S(g_{ab})+a))$.

Comments. 1. Theorem 2(a) continues to hold if Assumption A.2 and $k_{ab} \to \infty$ $\forall a, b$ are replaced by $\lim_{n\to\infty} \inf_{\theta_{k_{ab}} \in R^{k_{ab}}} \left\| \sum_{c=1}^{k_{ab}} z_{abc}(\cdot)\theta_{k_{ab}c} - g_{ab}(\cdot) \right\|_{0,\infty,\chi} = 0$ $\forall a, b$. The latter condition covers the case in which $g_{ab}(\cdot)$ is a finite linear combination of $\{z_{abc}(\cdot) : c \geq 1\}$ and $k_{ab} \not\to \infty$ for some a, b.

2. Theorem 2(b) and (c) shows that the rate of convergence of MASE($\hat{g}$,g) depends on α_{ab} $\forall a, b$. The latter do not depend on the dimension d of x_i, but rather, on the smoothness of g_{ab} and on the number of variables upon which g_{ab} depends. The latter may be very much smaller than d. In consequence, the curse of dimensionality, as measured by the asymptotic MASE criterion, is circumvented by series estimators of AIR models.

3. In a nonparametric regression model with regression function g_{ab}, the pointwise and L^q optimal rates of convergence of a nonparametric estimator are $n^{-2S(g_{ab})/(2S(g_{ab})+a)}$, see Stone (1980, 1982).[3] The slowest such rate over the functions g_{ab} $\forall a, b$ is n^{-v}, where $v = \min_{a,b} 2S(g_{ab})/(2S(g_{ab}) + a)$. Theorem 2(c) and the discussion above show that for trigonometric and FFF series, r_{ab} can be chosen such that the rate of convergence of MASE($\hat{g}$,g) is arbitrarily close to this optimal rate.

4. Theorem 2(c) implies that an optimal truncation parameter k^*_{ab} grows at rate $n^{1/(2\alpha_{ab}+1)}$. In this case, $k^{*\prime}\underset{\sim}{1} = \sum_a \sum_b k^*_{ab}$ grows at rate n^{δ} for $\delta = \max_{a,b} 1/(2\alpha_{ab} + 1)$ $(> \max_{a,b} a/(2S(g_{ab}) + a))$.

Proof of Theorem 2. First, we determine bounds on the variance and squared bias of MASE($\hat{g}$,g) (see (3.4)). We have

$$\frac{1}{n} E\|P_Z U\|^2 = \frac{1}{n} \operatorname{tr}(P_Z \Omega) \leq \sup_{i\geq 1} \sigma_i^2 k^{\prime}\underset{\sim}{1}/n . \tag{4.3}$$

Let $g_{k_{ab}}$ denote the approximation of g_{ab} given in Assumption A.2 with $m = k_{ab}$. Let $g^r_{k_{ab}}$ denote the remainder function from approximating g_{ab} by $g_{k_{ab}}$. That is, $g^r_{k_{ab}} = g_{ab} - g_{k_{ab}}$. Let $g_k = \sum_a \sum_b g_{k_{ab}}$ and $g^r_k = \sum_a \sum_b g^r_{k_{ab}}$ be the analogous approximating and remainder functions for g. Note that $P_Z \underset{\sim}{g}_k = \underset{\sim}{g}_k$, since $\underset{\sim}{g}_k$ is a linear combination of the columns of Z. We now have

$$\begin{aligned} \frac{1}{n}\|(I - P_Z)\underset{\sim}{g}\|^2 &= \frac{1}{n}\left\|(I - P_Z)\underset{\sim}{g}^r_k\right\|^2 \le \frac{2}{n}\left\|\underset{\sim}{g}^r_k\right\|^2 \\ \le \frac{2}{n}\left[\sum_a \sum_b \|\underset{\sim}{g}^r_{k_{ab}}\|\right]^2 &\le \frac{2D}{n}\sum_a \sum_b \left\|\underset{\sim}{g}^r_{k_{ab}}\right\|^2 \le 2D \sum_a \sum_b \left\|g^r_{k_{ab}}(\cdot)\right\|^2_{0,\infty,\gamma}. \end{aligned} \tag{4.4}$$

Combining (3.4), (4.3), and (4.4) gives

$$\mathrm{MASE}(\hat{g},g) \le \sup_{i \ge 1} \sigma_i^2 \, k' \underset{\sim}{1}/n + 2D \sum_a \sum_b \left\|g^r_{k_{ab}}(\cdot)\right\|^2_{0,\infty,\gamma}. \tag{4.5}$$

Theorem 2(a) follows from (4.5) and Assumption A.2, since the latter implies that $\|g^r_{k_{ab}}(\cdot)\|_{0,\infty,\gamma} = o(1)$ if $k_{ab} \to \infty$.

Theorem 2(b) follows from (4.5), Assumption A.2, and $k_{ab} \sim n^{r_{ab}}$, since the latter imply that

$$\mathrm{MASE}(\hat{g},g) \le \sum_a \sum_b \left[O\left[n^{r_{ab}-1}\right] + O\left[n^{-2\alpha_{ab}r_{ab}}\right] k_{ab}^{2\alpha_{ab}} \left\|g^r_{k_{ab}}(\cdot)\right\|^2_{0,\infty,\gamma}\right] = O(n^{-r}). \tag{4.6}$$

Theorem 2(c) holds because r is minimized by taking r_{ab} such that $1 - r_{ab} = 2\alpha_{ab}r_{ab}$, $\forall a, b$. This yields $r_{ab} = 1/(2\alpha_{ab} + 1)$ and $r = \min_{a,b} 2\alpha_{ab}/(2\alpha_{ab} + 1)$. Since $0 \le \alpha_{ab} < S(g_{ab})/a$ by Assumption A.2, we obtain $r < \min_{a,b} 2S(g_{ab})/(2S(g_{ab}) + a)$. □

5. AUTOMATIC TRUNCATION METHODS

In this section, we consider three automatic (i.e., data–driven) methods of determining the truncation vector k: generalized C_L (GC_L), generalized cross–validation (GCV), and cross–validation (CV). Let $\mathcal{K}_n$ denote the collection of vectors from which the automatic method chooses when the sample size is n. $\mathcal{K}_n$ is a subset of $\{k \in I_+^D : k'\underset{\sim}{1} \leq n\}$. Let $\hat{g}_k$ denote the estimator $\hat{g}$ when $\hat{g}$ is based on the truncation vector k and let $MASE(k)$ denote $MASE(\hat{g}_k, g)$. It is shown below that under suitable assumptions each of the above automatic truncation methods is asymptotically optimal in the sense that

$$\frac{\|\hat{g}_k - g\|^2}{\min_{k \in \mathcal{K}_n} \|\hat{g}_k - g\|^2} \xrightarrow{p} 1 \text{ and} \tag{5.1}$$

$$\frac{MASE(\hat{k})}{\min_{k \in \mathcal{K}_n} MASE(k)} \to 1, \tag{5.2}$$

where $\hat{k}$ is the vector k chosen from $\mathcal{K}_n$ by GC_L, GCV, or CV. These results are obtained by applying results of Li (1987) and Andrews (1989b).

The optimality results (5.1) and (5.2) imply that one does as well asymptotically in terms of average squared error and MASE using the automatic truncation procedure $\hat{k}$ as one would do if one knew the true function g (but one was restricted to the use of the linear estimators $\hat{g}_k$). A consequence of (5.2) is that provided $\{\mathcal{K}_n : n > 1\}$ is such that there is a sequence $\{k_n \in \mathcal{K}_n\}$ for which $k_{nab} \sim n^{1/(2\alpha_{ab}+1)}$ $\forall a, b$, the rate of convergence to zero of $MASE(\hat{k})$ is at least n^{-r} for r as in Theorem 2(c). Furthermore, this convergence rate is obtained without the use of knowledge of $\{\alpha_{ab} : \forall a, b\}$. Note that the latter usually depends on the smoothness of g_{ab} $\forall a, b$, which typically is unknown.

5.1. Generalized C_L

The C_L criterion is a generalization of the well–known C_p criterion and is due to Mallows (1973). It is suitable when the errors are homoskedastic. This criterion has been generalized straightforwardly to the case of heteroskedastic errors by Andrews (1989b). The generalized criterion is called *generalized* C_L (GC_L). It selects $\hat{k}$, denoted by $\hat{k}_M$, that achieves

$$\min_{k \in \mathcal{K}_n} n^{-1}\|Y - \hat{g}_k\|^2 + 2n^{-1} \operatorname{tr} Z_k(Z_k'Z_k)^{-}Z_k'\Omega , \tag{5.3}$$

where Z_k denotes the $n \times k'\underset{\sim}{1}$ matrix Z when the latter is based on the truncation vector k.

We introduce the following assumptions:

ASSUMPTION A.3. $\sup_{i \geq 1} EU_i^{4D+4} < \infty$.

ASSUMPTION A.4. *Either* (i) *for each fixed* $k \in I_+^D$, $\sum_{i=1}^{\infty} (g(x_i) - g_k(x_i))^2 = \infty$ *or* (ii) $\min_{k \in \mathcal{K}_n} k'\underset{\sim}{1} \to \infty$.

ASSUMPTION A.5. *Some sequence* $\{k_n : n \geq 1\}$ *for which* $k_n \in \mathcal{K}_n$ $\forall n$ *satisfies* $k_{nab} \sim n^{1/(2\alpha_{ab}+1)}$ $\forall a, b$, *where* α_{ab} *is as in Assumption* A.2.

Assumption A.4 is such that either (i) one needs to choose a truncation sequence $\{k_n\}$ such that $k_n'\underset{\sim}{1} \to \infty$ in order to obtain a consistent estimator $\hat{g}$ of g or (ii) one is forced to choose such a sequence by definition of $\mathcal{K}_n$. In either case, $\min_{k \in \mathcal{K}_n} \text{MASE}(k) \neq O(n^{-1})$. Assumption A.5 requires that $\mathcal{K}_n$ be defined so as not to exclude all sequences $\{k_n \in \mathcal{K}_n\}$ that yield fast rates of convergence of $\hat{g}$.

THEOREM 3. (a) *Under Assumptions* A.1, A.3, *and* A.4, GC_L *is asymptotically optimal in the sense that* (5.1) *and* (5.2) *hold with* $\hat{k} = \hat{k}_M$.

(b) *Under Assumptions* A.1–A.5, $MASE(\hat{k}_M) = O(n^{-r})$ *for* $r = \min_{a,b} 2\alpha_{ab}/(2\alpha_{ab} + 1)$.

Comments. 1. In practice, the covariance matrix Ω typically is unknown, so the GC_L criterion is infeasible. If the errors are homoskedastic, however, Ω can be replaced by $\hat{\Omega} = \text{diag}\{\hat{\sigma}^2, \ldots, \hat{\sigma}^2\}$, where $\hat{\sigma}^2$ is any consistent estimator of $\text{Var}(U_i)$, and the results of Theorem 3 still hold (see Li (1987, Corollary 2)). On the other hand, if the errors are heteroskedastic and Ω is unknown, no feasible version of GC_L is available (for which the results of Theorem 3 hold). In this case, the CV criterion discussed below needs to be used instead.

2. For trigonometric and FFF series, Theorem 3(b) yields $MASE(\hat{k}_M) = O(n^{-s})$ for all $s < v$, where v is defined in Theorem 2 Comment 3 and corresponds to the slowest optimal rate of convergence over g_{ab} $\forall a, b$.

Proof of Theorem 3. Theorem 3(a) holds by Corollary 2.1* of Andrews (1989b) provided Assumption A.4 implies $\min_{k \in \mathcal{K}_n} nMASE(k) \to \infty$. The latter holds under Assumption A.4(ii), since

$$MASE(k) \geq \frac{1}{n} E\|P_Z U\|^2 = \frac{1}{n} \text{tr}\, P_Z \Omega \geq \inf_{i \geq 1} \sigma_i^2 \, k' \underset{\sim}{1}/n \tag{5.4}$$

using (3.4). To show that it holds under Assumption A.4(i), suppose $\underline{\lim}_{n \to \infty} \inf_{k \in \mathcal{K}_n} nMASE(k) < \infty$. Then there exists a sequence $\{k_n \in \mathcal{K}_n\}$ and a subsequence $\{n_m\}$ of $\{n\}$ such that $\overline{\lim}_{m \to \infty} n_m MASE(k_{n_m}) < \infty$. Since $n_m MASE(k_{n_m}) \geq \inf_{i \geq 1} \sigma_i^2 \sum_a \sum_b k_{n_m ab}$ by (5.4), $\overline{\lim}_{m \to \infty} k_{n_m ab} < B$ $\forall a, b$ for some $B < \infty$. Thus, by (3.4),

$$\underset{m\to\infty}{\overline{\lim}}\, n_m \mathrm{MASE}(k_{n_m}) \geq \underset{m\to\infty}{\overline{\lim}} \sum_{i=1}^{n_m} (g(x_i)-g_{k_{n_m}}(x_i))^2 \geq \underset{m\to\infty}{\overline{\lim}} \min_{k\in\mathcal{K}_B} \sum_{i=1}^{n_m} (g(x_i)-g_k(x_i))^2, \quad (5.5)$$

where $\mathcal{K}_B$ is the set of vectors k in I_+^D such that $k_{ab} \leq B$ $\forall a, b$. Since $\mathcal{K}_B$ is a finite set, Assumption A.4(i) implies that the right–hand side of (5.5) is infinite, which yields a contradiction.

Theorem 3(b) follows from Theorems 2 and 3(a). □

5.2. Generalized Cross–validation

The GCV criterion was introduced by Craven and Wahba (1979). It selects $\hat{k}$, denoted by $\hat{k}_G$, that achieves:

$$\min_{k\in\mathcal{K}_n} n^{-1}\|Y-\hat{g}_k\|^2/(1-k'\underset{\sim}{1}/n)^2 . \quad (5.6)$$

The following assumptions are used to ensure the asymptotic optimality of GCV:

ASSUMPTION A.6. *Some sequence* $\{k_n\}$ *for which* $k_n \in \mathcal{K}_n$ $\forall n$ *satisfies* $k_{nab}/n \to 0$, $k_{nab} \to \infty$ $\forall a, b$, *and Assumption* A.2 *holds for* $\{k_n\}$.

ASSUMPTION A.7. $\max_{k\in\mathcal{K}_n} k'\underset{\sim}{1}/n \leq \gamma$ $\forall n$ *for some* $\gamma < 1$.

ASSUMPTION A.8. $\sigma_i^2 = \sigma^2$ $\forall i \geq 1$.

Assumption A.6 requires that $\mathcal{K}_n$ is defined such that some fixed sequence $\{k_n \in \mathcal{K}_n\}$ satisfies $\mathrm{MASE}(k_n) \to 0$. This is not overly restrictive, since $\mathcal{K}_n$ needs to be redefined if it is violated. Assumption A.7 is easy to verify (or to impose) and is not restrictive. On the other hand, the homoskedasticity Assumption A.8 is restrictive. Unless the errors are homoskedastic, the GCV criterion is not asymptotically optimal in general (see Andrews (1989b, Sec. 3)). Thus, neither GC_L nor GCV is both feasible and asymptotically optimal in the case of heteroskedastic errors.

THEOREM 4. (a) *Under Assumptions* A.1, A.3, A.4, *and* A.6–A.8, GCV *is asymptotically optimal in the sense that* (5.1) *and* (5.2) *hold with* $\hat{k} = \hat{k}_G$.

(b) *Under Assumptions* A.1–A.8, $\text{MASE}(\hat{k}_G) = O(n^{-r})$ *for* $r = \min_{a,b} 2\alpha_{ab}/(2\alpha_{ab} + 1)$.

Proof of Theorem 4. Theorem 4(a) holds by Theorem 3.1* of Andrews (1989b), since it is straightforward to show that the given assumptions imply those of Theorem 3.1*. Theorem 4(b) holds by Theorems 2 and 4(a) of this paper. □

5.3. Cross–validation

The CV criterion was first analyzed by Allen (1974), Stone (1974), Geisser (1975), and Wahba and Wold (1975). It selects $\hat{k}$, denoted by $\hat{k}_C$, that achieves:

$$\min_{k \in \mathcal{K}_n} n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}_k(x_i))^2/(1 - m_i(k))^2 , \tag{5.7}$$

where $\hat{g}_k(x_i)$ is the i–th element of $\hat{g}_k$ and $m_i(k)$ is the i–th diagonal element of $Z_k(Z_k'Z_k)^{-}Z_k'$.

Let $\bar{\lambda}(A)$ denote the largest diagonal element of the matrix A . The following assumptions are used to ensure that CV is asymptotically optimal:

ASSUMPTION A.9. $\overline{\lim}_{n\to\infty} \sup_{k \in \mathcal{K}_n} \bar{\lambda}(Z_k(Z_k'Z_k)^{-}Z_k') < 1$.

ASSUMPTION A.10. $\bar{\lambda}(Z_k(Z_k'Z_k)^{-}Z_k') \leq \Lambda k' \underset{\sim}{1}/n \quad \forall k \in \mathcal{K}_n \quad \forall n$ *for some constant* $0 < \Lambda < \infty$.

Assumption A.9 requires the self–weights, $\{m_i(k) : i \leq n\}$, to be bounded away from one. (They are necessarily ≤ 1 , since $Z_k(Z_k'Z_k)^{-}Z_k'$ is a projection matrix.) This condition is not overly restrictive, since it is easy to impose and its failure indicates potentially extreme overfitting of the model. Assumption A.10 prohibits highly unbalanced designs. It is equivalent to requiring the ratio of the maximum to the average diagonal

element of $Z_k(Z_k'Z_k)^{-}Z_k'$ to be bounded above by some $\Lambda < \infty$. It too is not overly restrictive and is easy to impose, since Z_k is observed.

THEOREM 5. (a) *Under Assumptions* A.1, A.3, A.4, A.6, A.9, *and* A.10, CV *is asymptotically optimal in the sense that* (5.1) *and* (5.2) *hold with* $\hat{k} = \hat{k}_C$.

(b) *Under Assumptions* A.1–A.6, A.9, *and* A.10, $\text{MASE}(\hat{k}_C) = O(n^{-r})$ *for* $r = \min_{a,b} 2\alpha_{ab}/(2\alpha_{ab} + 1)$.

Comment. Theorem 5 shows that CV is both feasible and asymptotically optimal when the errors are heteroskedastic. It is the only one of the three criteria considered that has this property.

Proof of Theorem 5. Theorem 5(a) holds by Theorem 4.2* of Andrews (1989b), since it is straightforward to show that the given assumptions imply those of Theorem 4.2*. Theorem 5(b) holds by Theorems 2 and 5(a) of this paper. □

FOOTNOTES

[1]The authors gratefully acknowledge the financial support of the Alfred P. Sloan Foundation and the National Science Foundation through a Research Fellowship (to the first author) and grant numbers SES–8618617 and SES–8821021 respectively.

[2]In particular, Chen (1988) assumes the regressors are of the form

$$\{x_i = (x_{i_1 1}, \ldots, x_{i_d d}) | i_j = 1, \ldots, n_j, j = 1, \ldots, d\},$$

where $n = \prod_{j=1}^{d} n_j$ and $x_{m,j}$ is determined by $\int_0^{x_{m,j}} w_j(t)dt = \frac{m}{n_j}$ for $m = 1, \ldots, n_j$, $j = 1, \ldots, d$, and $\{w_j : j = 1, \ldots, d\}$ are functions on [0,1] that are bounded above and away from zero.

[3]This holds provided Stone's index p of smoothness of g_{ab} is integer valued.

REFERENCES

Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16 (1974): 125–127.

Andrews, D. W. K. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Cowles Foundation Discussion Paper No. 874R, Yale University, 1989a.

Andrews, D. W. K. Asymptotic optimality of generalized C_L, cross–validation, and generalized cross–validation in regression with heteroskedastic errors. Cowles Foundation Discussion Paper No. 906, Yale University, 1989b.

Barry, D. *Nonparametric Bayesian Regression.* Ph.D. Thesis, Department of Statistics, Yale University, 1983.

Barry, D. Nonparametric Bayesian regression. *Annals of Statistics* 14 (1986): 934–953.

Buja, A., T. Hastie & R. Tibshirani. Linear smoothers and additive models. *Annals of Statistics* 17 (1989): 453–510.

Chen, Z. Interaction spline models and their convergence rates. Unpublished manuscript, Department of Statistics, University of Wisconsin, Madison, 1988.

Craven, P. & G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross–validation. *Numerische Mathematik* 31 (1979): 377–403.

Edmunds, B. J. & V. B. Moscatelli. Fourier approximation and embeddings in Sobolev space. *Dissertationes Mathematicae* 145 (1977): 1–46.

Gallant, A. R. On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form. *Journal of Econometrics* 15 (1981): 211–245.

Geisser, S. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70 (1975): 320–328.

Gu, C., D. M. Bates, Z. Chen, & G. Wahba. The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. Technical Report No. 823, Department of Statistics, University of Wisconsin, Madison, 1988.

Hastie, T. & R. Tibshirani. Generalized additive models. *Statistical Science* 1 (1986): 295–318.

Hastie, T. and R. Tibshirani. Generalized additive models: Some Applications. *Journal of the American Statistical Association* 82 (1987): 371–386.

Li, K.–C. Asymptotic optimality for C_p, C_L, cross–validation, and generalized cross–validation: Discrete index set. *Annals of Statistics* 15 (1987): 958–975.

Mallows, C. L. Some comments on C_p. *Technometrics* 15 (1973): 661–675.

Orcutt, G. H., M. Greenberger, J. Korbel, & A. M. Rivlin. *Microanalysis of Socioeconomic Systems: A Simulation Study.* New York: Harper, 1961.

Powell, M. J. D. *Approximation Theory and Methods.* Cambridge, England: Cambridge University Press, 1981.

Stone, C. J. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* 8 (1980): 1348–1360.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10 (1982): 1040–1053.

Stone, C. J. Additive regression and other nonparametric models. *Annals of Statistics* 13 (1985): 689–705.

Stone, M. Cross–validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society,* Series B 36 (1974): 111–147.

Wahba, G. Partial and interaction spline models for the semiparametric estimation of functions of several variables. In T. J. Boardman (ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface.* Washington, D.C.: American Statistical Association, 1986.

Wahba, G. & S. Wold. A completely automatic French curve: Fitting spline functions by cross–validation. *Communications in Statistics* 4 (1975): 1–17.