ADDITIVE INTERACTIVE REGRESSION MODELS:
CIRCUMVENTION OF THE CURSE OF DIMENSIONALITY

DONALD W.K. ANDREWS AND YOON-JAE WHANG
Yale University

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. A lower bound on 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.

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 have also been referred to in the literature as interaction spline models. Their estimation using splines has been analyzed by Barry [4,5], Wahba [23], Gu et al. [12], and Chen [7]. A special case of the AIR model is the additive regression model that has been considered by Orcutt et al. [17, p. 62], Stone [21], Hastie and Tibshirani [13,14], and Buja, Hastie, and Tibshirani [6]. 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 per-
mit reasonably accurate estimation unless the sample size is extremely large. This is illustrated by the fact that the fastest possible rate of convergence (in $L^q$-norm for $0 < q < \infty$) of estimators in such models is $n^{-2/(4+d)}$ when the regression function is assumed to be twice differentiable (see Stone [19,20]). This rate is very slow if $d$ is in the range of 5 to 15, 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 [21] 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 [7] 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 nonstochastic “tensor product design.” 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 [7] and the errors may be independent nonidentically distributed (i.n.i.d.). The criterion of performance used here for the rate of convergence results is mean average squared error (MASE) as in Chen [7]. In contrast, Stone [21] considers mean integrated squared error.

The “rate of convergence” results established here are one-sided. The results establish convergence at a particular rate, but do not prove that a faster rate is unattainable. Nevertheless, we refer to such results throughout as rate of convergence results.

We note that series estimators of AIR models have already been discussed in Andrews [2]. 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 parameters, large numbers of regressors, and large sample sizes. They can be computed using standard statistical software. 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 \( Y_i, U_i \in \mathbb{R}, x_i \in \mathcal{X} \subset \mathbb{R}^d \), 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 \( d!/(d!/(d - a)!) \), 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 \( x \), 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 (FFF) (see Gallant [10], 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 set of nonnegative integers. Let

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

\[
Y = (Y_1, \ldots, Y_n)', \quad U = (U_1, \ldots, U_n)', \quad \text{and}
\]

\[
Z = (Z_k(x_1), \ldots, Z_k(x_n))' \in \mathbb{R}^{n \times k+1}, \tag{2.3}
\]

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 \( 1 \) is a \( D \) dimensional vector of ones.

Let \( \theta \) be the vector of dimension \( k'1 \) 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 \( \theta \) is

\[
\hat{\theta} = (Z'Z)^{-1}Z'Y, \tag{2.4}
\]

where \( (\cdot)^{-1} \) denotes some \( g \) inverse. (Suitable choices of \( k \) yield \( Z'Z \) to be nonsingular, so the \( g \) inverse \( (\cdot)^{-1} \) is usually just a standard inverse.) The corresponding series estimator \( \check{g} \) of \( g \) is

\[
\check{g}(\cdot) = Z_k(\cdot)'\hat{\theta}. \tag{2.5}
\]

Various properties of \( \check{g} \) 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 \( \sum_{a=1}^{A} \sum_{b(a)} \); \( \forall a, b \) abbreviates \( \forall b = 1, \ldots, B(a), \forall a = 1, \ldots, A \); all limits are taken as \( n \to \infty \); \( a_n/b_n \) denotes that \( a_n/b_n \) is bounded away from zero and infinity over \( n \geq 1 \); and for any function \( g^* \) from \( X \) to \( R \), \( g^* \) 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 dimen-
sionality by AIR models. They also have implications for the rate of convergence results given in Sections 4 and 5 below.

Consider the following models:

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

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

where $Y_i = \sum_a \sum_b Y_{ab}$ and $U_i = \sum_a \sum_b U_{ab}$ for $i = 1, \ldots, n$. \{U_{ab}\} are mean zero, variance $\sigma_{ab}^2$ random variables, independent across $i, a, b$. \{x_{i}\} are nonrandom regressor vectors in $X \subset R^d$. 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

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

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 (nonasymptotic) sense, the estimator circumvents the curse of dimensionality.

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

$$\text{MASE}(\hat{g}, g) \leq D \sum_{a, b}^{A} \text{MASE}(\hat{g}_{ab}, g_{ab}),$$

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

Comment 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 $\sigma_{ab}^2$ does not depend on $i, a, b$ and the estimators $\hat{g}$ and $\hat{g}_{ab}$ $\forall a, b$ do not incur any bias.)

Comment 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. MASE(\(\hat{g}, g\)) and \(\sum_a \sum_b MASE(\hat{g}_{ab}, g_{ab})\) can be decomposed into squared bias and variance terms:

\[
\text{MASE}(\hat{g}, g) = \frac{1}{n} \| g - P_Z g \|^2 + \frac{1}{n} E \| P_Z U \|^2 \quad \text{and} \quad (3.4)
\]

\[
\sum_a \sum_b MASE(\hat{g}_{ab}, g_{ab}) = \frac{1}{n} \sum_a \sum_b \| g_{ab} - P_{Z_{ab}} g_{ab} \|^2
\]

\[
+ \frac{1}{n} \sum_a \sum_b E \| P_{Z_{ab}} U_{ab} \|^2, \quad (3.5)
\]

where \(U_{ab} = (U_{1ab}, \ldots, U_{nab})\), \(P_Z = Z (Z^T Z)^{-1} Z^T\), \(P_{Z_{ab}} = Z_{ab} (Z_{ab}^T Z_{ab})^{-1} Z_{ab}^T\), 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

\[
\| g - P_Z g \| = \left\| \sum_a \sum_b (g_{ab} - P_{Z_{ab}} g_{ab}) \right\|
\]

\[
\leq \sum_a \sum_b \| g_{ab} - P_{Z_{ab}} g_{ab} \| \leq \sum_a \sum_b \| g_{ab} - P_{Z_{ab}} g_{ab} \|^2 \quad (3.6)
\]

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

\[
\| g - P_Z g \|^2 \leq \left( \sum_a \sum_b \| g_{ab} - P_{Z_{ab}} g_{ab} \|^2 \right)^2 \leq D \sum_a \sum_b \| g_{ab} - P_{Z_{ab}} g_{ab} \|^2 \quad (3.7)
\]

using the Cauchy–Schwarz inequality.

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

\[
E \| P_Z U \|^2 = \text{tr}(P_Z \Omega) = \sum_a \sum_b \text{tr}(P_Z \Omega_{ab}) \leq \tau^* \sum_a \sum_b \text{tr} P_Z \leq D \tau^* k^* 1. \quad (3.8)
\]

On the other hand,

\[
\sum_a \sum_b E \| P_{Z_{ab}} U_{ab} \|^2 = \sum_a \sum_b \text{tr}(P_{Z_{ab}} \Omega_{ab}) \geq \tau^* k^* 1. \quad (3.9)
\]

Equations (3.4)–(3.9) combined 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 $X$ to $R$. For any $g_1 \in \mathcal{G}$, let $\|g_1\|_{q,\infty,X}$ denote the supremum Sobolev norm of derivative order $q$, for some $q \geq 0$. That is,

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

where

$$\lambda = (\lambda_1, \ldots, \lambda_d)' \in I_d^d, \quad |\lambda| = \sum_{j=1}^d \lambda_j, \quad \text{and}$$

$$D^\lambda g_1(x) = \frac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} \cdots \partial x_d^{\lambda_d}} g_1(x).$$

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

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

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

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 nonstochastic regressor vectors in $X \subset R^d$.

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

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

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 [9, 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, $X$ has closure that lies in $(0,2\pi)^d$, and $X$ has minimally smooth boundary.$^{34}$ Examples of sets with this property include all convex sets. Note that if $X$ is any bounded subset of $R^d$, the regressors can be rescaled such that $X$ has closure that lies in $(0,2\pi)^d$.

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 $X$ is a closed, bounded, connected subset of $R^d$. This follows from Theorem 3.2 of Powell [18, 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 \( \text{MASE}(\hat{g}, g) \to 0 \). (b) If \( k_{ab} \sim n^{-r_{ab}} \) for some \( 0 < r_{ab} < 1 \) \( \forall a, b \), then \( \text{MASE}(\hat{g}, g) = O(n^{-\gamma}) \), 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 \( \gamma \) in part (b) is \( r_{ab} = 1/(2\alpha_{ab} + 1) \). In this case, \( r = \min_{a, b} 2\alpha_{ab}/(2\alpha_{ab} + 1) \) \(< \min_{a, b} 2S(g_{ab})/(2S(g_{ab}) + a)) \).

Comment 1. Theorem 2(a) continues to hold if Assumption A.2 and \( k_{ab} \to \infty \) \( \forall a, b \) are replaced by

\[
\lim_{n \to \infty} \inf_{\theta_{k_{ab}} \in R_{k_{ab}}} \left\{ \sum_{c=1}^{k_{ab}} z_{abc}(\cdot) \theta_{k_{ab}}(\cdot) - g_{ab}(\cdot) \right\}_{0,\infty, X} = 0 \quad \forall a, b.
\]

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 \( a, b \).

Comment 2. Theorem 2(b) and (c) shows that the rate of convergence of \( \text{MASE}(\hat{g}, g) \) depends on \( \alpha_{ab} \) \( \forall a, b \). The latter do not depend on the dimension \( d \) of \( x \), 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. (On the other hand, the constant associated with the given rate of convergence may increase with \( d \), as indicated in Theorem 1.)

Comment 3. In a nonparametric regression model with regression function \( g_{ab} \), the square of the pointwise and \( L^2 \) optimal rates of convergence of a nonparametric estimator is \( n^{-2S(g_{ab})/(2S(g_{ab}) + a)} \), see Stone [19,20]. 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 \( \text{MASE}(\hat{g}, g) \) is arbitrarily close to this optimal rate.

Comment 4. Theorem 2(c) implies that an optimal truncation parameter \( k_{ab}^* \) grows at rate \( n^{ \bar{c} / (2\alpha_{ab} + 1)} \). In this case, \( k_{ab}^* = \sum_{a} \sum_{b} k_{ab}^* \) grows at rate \( n^{\delta} \) 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 \( \text{MASE}(\hat{g}, g) \) (see (3.4)). We have

\[
\frac{1}{n} E\|P_Z U\|^2 = \frac{1}{n} \text{tr}(P_Z \Omega) \leq \sup_{i \leq 1} \alpha_i^2 k^* k^*/n. \tag{4.3}
\]

Let \( g_{k_{ab}} \) denote the approximation to \( g_{ab} \) given in Assumption A.2 with \( m = k_{ab} \). Let \( g_{k_{ab}}^* \) denote the remainder function from approximating \( g_{ab} \) by \( g_{k_{ab}} \). That is, \( g_{k_{ab}}^* = g_{ab} - g_{k_{ab}} \). Let \( g_k = \sum_a \sum_b g_{k_{ab}} \) and \( g_k^* = \sum_a \sum_b g_{k_{ab}}^* \) be the analogous approximating and remainder functions for \( g \). Note that \( P_Z g_k = g_k \), since \( g_k \) is a linear combination of the columns of \( Z \). We now have
\[
\frac{1}{n} \| (I - P_Z) g \|^2 = \frac{1}{n} \| (I - P_Z) \hat{g}_k \|^2 \leq \frac{1}{n} \| \hat{g}_k \|^2 \leq \frac{1}{n} \left( \sum_a \sum_b \| \hat{g}_{k_{ab}} \|^2 \right)^2 \leq \frac{D}{n} \sum_a \sum_b \| \hat{g}_{k_{ab}} \|^2 \leq D \sum_a \sum_b \| \hat{g}_{k_{ab}}(\cdot) \|^2 \| \hat{g}_{k_{ab}} \|_{0, \infty, X}^2.
\]  
(4.4)

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

\[
\text{MASE}(\hat{g}, g) \leq \sup_{i \geq 1} \sigma_i^2 k^{1/2} + D \sum_a \sum_b \| \hat{g}_{k_{ab}}(\cdot) \|^2 \| \hat{g}_{k_{ab}} \|_{0, \infty, X}^2.
\]  
(4.5)

Theorem 2(a) follows from (4.5) and Assumption A.2, since the latter implies that \( \| g_{k_{ab}}(\cdot) \|_{0, \infty, X} = o(1) \) 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

\[
\text{MASE}(\hat{g}, g) \leq \sum_a \sum_b \left( O(n^{r_{ab} - 1}) + O(n^{-2r_{ab} + 1}) k_{ab}^{2r_{ab}} \| \hat{g}_{k_{ab}}(\cdot) \|^2 \| \hat{g}_{k_{ab}} \|_{0, \infty, X}^2 \right)
\]

\[= O(n^{-r}). \]  
(4.6)

Theorem 2(c) holds because \( r \) is minimized by taking \( r_{ab} \) such that \( 1 - r_{ab} = 2\sigma_{ab} r_{ab}, \ \forall a, b \). This yields \( r_{ab} = 1/(2\sigma_{ab} + 1), \ r = \min_a, b 2\sigma_{ab} / (2\sigma_{ab} + 1) \).

Since \( 0 \leq \sigma_{ab} < S(g_{ab})/a \) 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 \( CL (GC_L) \), generalized cross validation (GCV), and cross validation (CV). Let \( \kappa_n \) denote the collection of vectors from which the automatic method chooses when the sample size is \( n \). \( \kappa_n \) is a subset of \( \{ k \in I^2: k \leq n \} \). Let \( \hat{g}_k \) denote the estimator \( \hat{g} \) when \( \hat{g} \) is based on the truncation vector \( k \) and let \( \text{MASE}(k) \) denote \( \text{MASE}(\hat{g}_k, g) \).

It is shown below that under suitable assumptions each of the above automatic truncation methods is asymptotically optimal in the sense that

\[
\frac{\| \hat{g}_k - g \|^2}{\min_{k \in \kappa_n} \| \hat{g}_k - g \|^2} \to 1 \quad \text{and} \quad \frac{\text{MASE}(k)}{\min_{k \in \kappa_n} \text{MASE}(k)} \to 1,
\]  
(5.1)

where \( \hat{k} \) is the vector \( k \) chosen from \( \kappa_n \) by \( GC_L \), GCV, or CV. These results are obtained by applying results of Li [15] and Andrews [3].

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 \( \{ \kappa_n : n \geq 1 \} \) is such that there is a sequence \( \{ k_n \in \kappa_n \} \) for which \( k_n \sim n^{1/(2\alpha_{ab}+1)} \forall a, b \), 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 [16]. It is suitable when the errors are homoskedastic. This criterion has been generalized straightforwardly to the case of heteroskedastic errors by Andrews [3]. The generalized criterion is called generalized \( C_L \) (\( GC_L \)). It selects \( \hat{k} \), denoted by \( \hat{k}_M \), that achieves

\[
\min_{k \in \kappa} n^{-1} \| Y - \hat{g}_k \|^2 + 2n^{-1} \text{tr} Z_k (Z_k' Z_k)^{-1} Z_k' \Omega, \tag{5.3}
\]

where \( Z_k \) denotes the \( n \times k'1 \) matrix \( Z \) when the latter is based on the truncation vector \( k \).

We introduce the following assumptions:

Assumption A.3. \( \sup_{D \geq 1} E(U_i^D)^{4+\delta} < \infty. \)

Assumption A.4. Either (i) for each fixed \( k \in I_1^D \), \( \sum_{i=1}^\infty (g(x_i) - g_k(x_i))^2 = \infty \) or (ii) \( \min_{k \in \kappa} k'1 \rightarrow \infty. \)

Assumption A.5. Some sequence \( \{ k_n : n \geq 1 \} \) for which \( k_n \in \kappa_n \forall n \) satisfies \( k_n \sim n^{1/(2\alpha_{ab}+1)} \forall a, b \), where \( \alpha_{ab} \) 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 \( k_n'1 \rightarrow \infty \) in order to obtain a consistent estimator \( \hat{g} \) of \( g \) or (ii) one is forced to choose such a sequence by definition of \( \kappa_n \). In either case, \( \min_{k \in \kappa} \text{MASE}(k) = O(n^{-1}). \) Assumption A.5 requires that \( \kappa_n \) be defined so as not to exclude all sequences \( \{ k_n \in \kappa_n \} \) 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, \( \text{MASE}(\hat{k}_M) = O(n^{-r}) \) for \( r = \min_{a, b} 2\alpha_{ab}/(2\alpha_{ab} + 1). \)

Comment 1. In practice, the covariance matrix \( \Omega \) typically is unknown, so the \( GC_L \) criterion is infeasible. If the errors are homoskedastic, however, \( \Omega \) can be replaced by \( \hat{\Omega} = \text{diag}(\hat{\delta}^2, \ldots, \hat{\delta}^2) \), where \( \hat{\delta}^2 \) is any consistent estimator of \( \text{var}(U_i) \), and the results of Theorem 3 still hold (see Li [15], Corollary 2). On the other hand, if the errors are heteroskedastic and \( \Omega \) 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.

Comment 2. For trigonometric and FFF series, Theorem 3(b) yields \( \text{MASE}(\hat{k}_{\text{ab}}) = O(n^{-s}) \) 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 [3] provided Assumption A.4 implies \( \min_{k \in \kappa_n} n \text{MASE}(k) \to \infty \). The latter holds under Assumption A.4(ii), since

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

using (3.4). To show that it holds under Assumption A.4(i), suppose \( \lim_{n \to \infty} \inf_{k \in \kappa_n} n \text{MASE}(k) < \infty \). Then there exists a sequence \( \{ k_n \} \subset \kappa_n \) and a subsequence \( \{ n_{m} \} \) of \( \{ n \} \) such that \( \lim_{m \to \infty} n_{m} \text{MASE}(k_{n_m}) < \infty \). Since \( n_{m} \text{MASE}(k_{n_m}) \geq \inf_{i \geq 1} \sigma_i^2 \sum_{b} k_{n_m \alpha b} \) by (5.4), \( \lim_{m \to \infty} k_{n_{m} \alpha b} < B \forall a, b \) for some \( B < \infty \). Thus, by (3.4),

\[
\lim_{m \to \infty} n_{m} \text{MASE}(k_{n_m}) \geq \lim_{m \to \infty} \sum_{i=1}^{n_{m}} (g(x_i) - g_{k_{n_m}}(x_i))^2
\]

\[
\geq \lim_{m \to \infty} \min_{k \in \kappa_B} \sum_{i=1}^{n_{m}} (g(x_i) - g_{k}(x_i))^2, \tag{5.5}
\]

where \( \kappa_B \) is the set of vectors \( k \) in \( I_+^D \) such that \( k_{ab} \leq B \forall a, b \). Since \( \kappa_B \) 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 Whaba [8]. It selects \( \hat{k} \), denoted by \( \hat{k}_G \), that achieves:

\[
\min_{k \in \kappa_n} n^{-1} \| Y - \hat{g}_k \|^2 / (1 - k'1/n)^2. \tag{5.6}
\]

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

Assumption A.6. Some sequence \( \{ k_n \} \) for which \( k_n \in \kappa_n \forall n \) satisfies \( k_{nab}/n \to 0, k_{nab} \to \infty \forall a, b \) and Assumption A.2 holds for \( \{ k_n \} \).

Assumption A.7. \( \max_{k \in \kappa} k'1/n \leq \gamma \forall n \) for some \( \gamma < 1 \).

Assumption A.8. \( \sigma_i^2 = \sigma^2 \forall i \geq 1 \).
Assumption A.6 requires that $\kappa_n$ is defined such that some fixed sequence $\{k_n \in \kappa_n\}$ satisfies $\text{MASE}(k_n) \to 0$. This is not overly restrictive, since $\kappa_n$ 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 [3], Sec. 3). Thus, neither $GC_\gamma$ 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 $\hat{k} = \hat{k}_C$. (b) Under Assumptions A.1–A.8, $\text{MASE}(\hat{k}_C) = O(n^{-r})$ 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 [3], 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 [1], Stone [22], Geisser [1], and Wahba and Wold [24]. It selects $\hat{k}$, denoted by $\hat{k}_C$, that achieves

$$\min_{k \in \kappa_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_k(x_i))^2/(1 - m_i(k))^2,$$

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 $Z_k(Z'_kZ_k)^{-1}Z'_k$.

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

Assumption A.9. $\lim_{n \to \infty} \sup_{k \in \kappa_n} \lambda(Z_k(Z'_kZ_k)^{-1}Z'_k) < 1$.

Assumption A.10. $\lambda(Z_k(Z'_kZ_k)^{-1}Z'_k) \leq \Lambda k'1/n \forall k \in \kappa_n \forall n$ 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 $\leq 1$, since $Z_k(Z'_kZ_k)^{-1}Z'_k$ 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 $Z_k(Z'_kZ_k)^{-1}Z'_k$ 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, 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 [3], 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.

NOTES

1. In particular, Chen [7] assumes the regressors are of the form

\[ \{x_j = (x_{i1}, \ldots, x_{id}) | i = 1, \ldots, n_j, j = 1, \ldots, d \} \]

where \( n = \prod_{j=1}^d n_j \) and \( x_{mj} \) is determined by \( \int_{0}^{m} w_j(t) \, dt = m/n_j \) 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.

2. Normalization to achieve identification of the functions \( g_{ab} \) usually reduces the number of nonzero functions. In consequence, the number of nonzero functions is not necessarily increasing in \( A \). For example, if \( A = d \), the simplest normalization rule leaves only one function.

3. Fourier flexible form series expansions are comprised of a finite number of polynomials followed by an infinite number of trigonometric functions, e.g., see Gallant [10].

4. Let \( X \) be an open set in \( \mathbb{R}^d \) and let \( \partial X \) be its boundary. By definition, \( \partial X \) is minimally smooth if there exists an \( \epsilon > 0 \), an integer \( N \), an \( M > 0 \), and a sequence \( U_1, U_2, \ldots \) of open sets in \( \mathbb{R}^d \) such that (a) if \( x \in \partial X \), then \( S(x, \epsilon) \subset U_i \) for some \( j \), where \( S(x, \epsilon) \) is the sphere centered at \( x \) with radius \( \epsilon \), (b) no point of \( \partial X \) is contained in more than \( N \) of the \( U_i \)’s, and (c) \( \forall j \geq 1, \exists \) a special Lipschitz domain \( D_j \subset \mathbb{R}^d \), whose bound does not exceed \( M \), such that \( U_j \cup X = U_1 \cap D_j \). A set \( D \subset \mathbb{R}^d \) is a special Lipschitz domain if \( D = \{ (x, y) \in \mathbb{R}^d : y > \phi(x) \} \) for some function \( \phi : \mathbb{R}^d \to \mathbb{R} \) that satisfies \( |\phi(x) - \phi(x')| \leq M|x - x'| \) for all \( x, x' \in \mathbb{R}^d \). The smallest \( M \) for which the Lipschitz condition holds is called the bound of the special Lipschitz domain.

5. This holds provided Stone’s index \( p \) of smoothness of \( \varphi_{ab} \) is integer valued.

6. For comparison with the MASE criterion, we consider the square of the rate of convergence here, since the MASE criterion is based on squared errors rather than on some norm of the errors.

REFERENCES