The Annals of Statistics
1990, Vol. 18, No. 2, 742–757

CONSISTENCY FOR ACE-TYPE METHODS[1]

BY ROBERT KOYAK

The Johns Hopkins University

The ACE (alternating conditional expectations) algorithm developed by Breiman and Friedman is an iterative method for finding optimal transformations of variables in multiple regression. Recently, several authors have extended ACE to discriminant analysis, time series and principal components. The central idea of ACE and of each of these extensions is that an optimal transformation ϕ^* minimizes a squared error-related functional over a Hilbert space, subject to nonlinear functional constraints. An estimate $\hat{\phi}^{(N)}$ is obtained by minimizing an estimate of the functional, subject to estimates of the constraints, over a smoothness restricted class of transformations. Using the method of sieves, conditions are established for consistency of $\hat{\phi}^{(N)}$ in the $\mathbf{L}^2$ sense.

1. Introduction. Estimators are often derived as optimizers of an empirically derived quantity (i.e., the likelihood). If the parameter estimated in this way is in a finite-dimensional class, obtaining consistency results is usually not difficult. We extend results on consistency to a class of problems where the parameter under estimation is the class of minimizers of a squared error-derived functional defined on an infinite-dimensional space. The method of sieves (Grenander, Geman) provides a useful framework for this task.

1.1. *Least squares problems.* Let $\mathbf{X} = (X_1, \ldots, X_M)'$ denote a vector random variable with values in $\mathbf{R}^M$, and let F denote its distribution. Assume, for now, that all second moments $E[X_j^2]$ are finite, and that $\mathbf{X}$ is *nondegenerate*; i.e., the smallest eigenvalue of $U(\mathbf{X}) = E[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)']$ is strictly positive. Let $\mathbf{R}^{M,d}$ refer to the space of real-valued $M \times d$ matrices, and $\mathbf{C}^{M,M}$ the class of positive-definite symmetric matrices in $\mathbf{R}^{M,M}$. For any matrix A, we will take A_j to denote its jth column.

In multivariate analysis, inference is often made with respect to a matrix parameter $A^* \in \mathbf{R}^{M,d}$ which minimizes a functional of the form

$$\xi^2(A) = \sum_{j=1}^{d} E\left[A_j'(\mathbf{X} - \mu_X)\right]^2 = \text{trace } A'U(\mathbf{X})A, \tag{1.1}$$

over a specified subset $\mathbf{A}$ of $\mathbf{R}^{M,d}$ which satisfies the following conditions:

1. $\mathbf{A} = \mathbf{A}(U(\mathbf{X}))$; i.e., $\mathbf{A}$ depends on F, if at all, only through $U(\mathbf{X})$;
2. $\mathbf{A}(U)$ is a *continuous* set-valued function on $\mathbf{C}^{M,M}$;
3. For any compact subset $\mathbf{S} \subset \mathbf{C}^{M,M}$, $\cup\{\mathbf{A}(U)\colon U \in \mathbf{S}\}$ is compact;
4. $\mathbf{A}(U)$ excludes the zero matrix for every $U \in \mathbf{C}^{M,M}$.

Received September 1986; revised June 1989.

[1]Research supported by Office of Naval Research Contract N00014-79-C-0801, Air Force Office of Scientific Research Grant 82-0029C and NSF Grant MC-80-02698.

AMS 1980 *subject classifications.* Primary 62H12; secondary 62G05.

Key words and phrases. ACE, transformations, consistency, method of sieves.

For compactness and continuity, take Euclidean distance as a base metric on $\mathbf{R}^{M,d}$ and $\mathbf{R}^{M,M}$, and the Hausdorf metric as a distance between subsets of $\mathbf{R}^{M,d}$. If A^* is not unique, the parameter of interest is the class of minimizers $\mathbf{A}^*$.

A variety of problems in multivariate analysis can be described in this fashion. These include:

(i) *Least squares regression*. Let $M = p + 1$, and $\mathbf{X} = (Y, Z_1, \ldots, Z_p)'$. The parameter of interest is the coefficient vector b in the linear regression model $Y = b'\mathbf{Z} + \varepsilon$. In the framework of (1.1), take $d = 1$, and $\mathbf{A}$ the class of M vectors of the form $a' = (1, b')$, where $b \in \mathbf{R}^p$ satisfies $\operatorname{Var}(b'\mathbf{Z}) \leq \operatorname{Var}(Y)$.

(ii) *Canonical correlation*. Let $M = p + q$, $\mathbf{Y} = (Y_1, \ldots, Y_p)'$, $\mathbf{Z} = (Z_1, \ldots, Z_q)'$, and take $\mathbf{X}' = (Y', Z')$. Again $d = 1$, but now $\mathbf{A}$ is the class of M vectors taking the form (a', b'), with $a \in R^p$ and $b \in R^q$ satisfying $\operatorname{Var}(a'\mathbf{Y}) = 1$ and $\operatorname{Var}(b'\mathbf{Z}) = 1$. For k canonical coordinates, with $k \leq \min(p, q)$,take $d = k$ and $\mathbf{A}$ the class of $M \times k$ matrices A with $A_j' = (a_j', b_j')$ satisfying $\operatorname{Cov}(a_i'\mathbf{Y}, a_j'\mathbf{Y}) = \delta_{ij}$, $\operatorname{Cov}(b_i'\mathbf{Z}, b_j'\mathbf{Z}) = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta function.

(iii) *Principal components*. Let k be an integer between 1 and $M - 1$ (inclusive), which is the number of principal components desired in the analysis. Take $d = M$, and $\mathbf{A}$ the class of $M \times M$ projection matrices having rank equal to $M - k$. The matrix A^* achieving the minimum of (1.1) can be written in the form $I - B^*$, where B^* is the eigenprojection corresponding to the k largest eigenvalues of $\operatorname{Cov}(\mathbf{X})$. That is, $B^* = ZZ'$, where Z is the $M \times k$ matrix whose columns contain the normalized eigenvectors of $\operatorname{Cov}(\mathbf{X})$ corresponding to the k largest eigenvalues, and the principal components are the k-derived linear combinations $Z_j'\mathbf{X}$. This factorization of B^* is not unique, but is motivated by the heuristic consideration that $Z_j'\mathbf{X}$ should have maximum variance over all linear combinations of the form $a'\mathbf{X}$ with $a'a = 1$ and $\operatorname{Cov}(a'\mathbf{X}, Z_i'\mathbf{X}) = 0$ for every $i < j$.

Let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ denote a sample, construed as N independent replicates of $\mathbf{X}$. As an estimate of A^*, choose a matrix $\hat{A}^{(N)}$ which minimizes

$$\hat{\xi}_N^2(A) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left[A_j'(\mathbf{X}^{(i)} - \hat{\mu}) \right]^2 = \operatorname{trace} A'\hat{U}(\mathbf{X})A, \tag{1.2}$$

over $A \in \mathbf{A}(\hat{U}(\mathbf{X}))$. In particular, this is appropriate if $\mathbf{X}$ is multivariate normal, since it entails belief that the variables are interrelated in a linear fashion, and that an analysis based on the first two moments is reasonably informative. Observe that there is a continuous function g for which $A^* = g(U(\mathbf{X}))$, and $\hat{A}^{(N)} = g(\hat{U}(\mathbf{X}))$, where $\hat{U}(\mathbf{X})$ is the maximum likelihood estimate of $U(\mathbf{X})$ under normality. Although g is typically not bijective, under the reasoning of Zehna (1966) one would still designate $\hat{A}^{(N)}$ a "maximum likelihood" estimate of A^* for $\mathbf{X}$ multivariate normal.

Multivariate data analytic techniques based on normal theory persist in wide usage for a variety of reasons. Basic to their proper usage is the notion that the variables are linearly interrelated, or at least approximately so. A

violation of this aspect of the normality assumption is the easiest one to detect, and it is also the most discomforting. When this situation arises, a common tactic is to transform the variables to ammeliorate the effects of nonlinearity.

1.2. *ACE-type problems.* We now describe a framework for choosing linearizing transformations, so that estimation problems of the above form become meaningful. The transformations that we will consider are *coordinate-wise*; i.e., taking the form $\phi(\mathbf{X}) = (\phi_1(X_1), \ldots, \phi_M(X_M))'$, Borel measurable and satisfying

$$\begin{aligned} E\big[\phi_j(X_j)\big] &= 0, \\ E\big[\phi_j^2(X_j)\big] &= 1, \qquad j = 1, \ldots, M. \end{aligned} \tag{1.3}$$

Let Φ denote this class of transformations. For a given transformation $\phi \in \Phi$, let $A^*(\phi)$ denote the minimizer of

$$\xi^2(\phi, A) = \sum_{j=1}^{d} E\big[A_j'\phi(\mathbf{X})\big]^2 = \text{trace } A'\mathbf{U}(\phi)A, \tag{1.4}$$

where $\mathbf{U}(\phi) = E[\phi(\mathbf{X})\phi(\mathbf{X})']$, and minimization is over $A \in \mathbf{A}(\mathbf{U}(\phi))$. Define the functional

$$\zeta(\phi) = \xi(\phi, A^*(\phi)) = \inf_{A \in \mathbf{A}(\mathbf{U}(\phi))} \xi(\phi, A). \tag{1.5}$$

Let Φ^* denote the subset of Φ consisting of minimizers of $\zeta(\phi)$, elements of which we will call *optimal transformations*. Our use of the word "subset" is deliberate, chosen to acknowledge that optimal transformations are never unique, if for no other reason that $\zeta(\phi) = \zeta(-\phi)$ for every $\phi \in \Phi$. We wish to construct a class of estimates $\{\hat{\phi}^{(N)}\}$ based on the sample for which $\hat{\phi}^{(N)} \to \Phi^*$ in an appropriate sense.

The class of problems having this structure will be called *ACE-type problems* in the remainder of this paper. Here, ACE stands for *alternating conditional expectations*, a term used by Breiman and Friedman (1985) for their approach to nonparametric additive regression. Optimal transformations $\phi^* \in \Phi^*$ usually cannot be expressed in a closed form even when the distribution is known, but instead are obtained by iterative approximation. The structure of the algorithm varies with the application, but central to each of these is the use made of conditional expectations in an alternating variables scheme, thus motivating our choice of nomenclature. Optimal transformations are estimated from data by imitating the algorithm with F replaced by $\hat{F}_N$, and with conditional expectations replaced by data smooths, or some other form of estimation. Since we do not attempt the same level of unification with respect to applied aspects of these problems, we will not pursue this point further here, but instead refer the interested reader to Breiman and Friedman (1985) for a treatment of regression via ACE, and Koyak (1985) for the analogous treatment of principal components. Slightly outside of the present context are ACE applied to discriminant analysis [Breiman and Ihaka (1984)]

and time series [Owen (1983)]. Work of a similar spirit has a long history in the psychometric community, a flavor for which can be found in Gifi (1981).

REMARK. If $\mathbf{X}$ is multivariate normal, it is reasonable to ask whether the identity transformation is optimal for a given ACE-type problem. This can be answered affirmatively for the nonlinear extensions of the three examples cited in Section 1.1. For canonical correlation, this result is the solution to Kolmogorov's canonical problem, a proof of which can be found in Lancaster (1969), with multiple regression following as a simple extension. For principal components, a proof of optimality is given in Koyak (1987). Our definition of ACE-type problems, however, is too broad to regard this as a general property. It may seem desirable to include in the definition of an ACE-type problem the requirement that the (scaled) identity transformation be a member of Φ^*. We presently desist from doing so, because this issue does not arise in our discussion of consistency.

1.3. *Mathematical framework.* The following notation and concepts will be used in our subsequent discussion. Denote by $\mathbf{H}$ the space of $\mathbf{R}^1$-valued Borel-measurable functions of $\mathbf{X}$ having zero mean and finite variance. Endow $\mathbf{H}$ with the $\mathbf{L}^2$ inner product with respect to F and take $\|g\| = \langle g, g\rangle^{1/2}$. With this inner product, $\mathbf{H}$ is a Hilbert space, as are its subspaces $\mathbf{H}_j$, $j = 1, \ldots, M$, consisting of transformations which depend on X_j alone, d-fold Cartesian products of $\mathbf{H}$ with itself which we denote $\mathbf{H}^d$, and the Cartesian product of the $\mathbf{H}_j$ which we denote $\dot{\mathbf{H}}^M$. For an inner product on $\mathbf{H}^d$ we take the natural $\mathbf{L}^2$ extension $\langle \mathbf{g}, \mathbf{h}\rangle_d = \sum_{j=1}^d \langle g_j, h_j\rangle$, with $\|\mathbf{g}\|_d = \langle \mathbf{g}, \mathbf{g}\rangle_d^{1/2}$. At various times we will consider Hilbert spaces of transformations that are *not* centered with respect to F, endowed with the same inner products and norms, which we will denote $\mathbf{H}$, $\dot{\mathbf{H}}^M$, etc. For an element $\mathbf{g} \in \mathbf{H}^d$, we will often use functional notation for moments: $\mu(g_j) = E_F g_j(\mathbf{X})$, $\mu(\mathbf{g}) = (\mu(g_1), \ldots, \mu(g_d))'$, $\sigma^2(g_j) = \mathrm{Var}_F\, g_j(\mathbf{X})$, $U(\mathbf{g}) = \mathrm{Cov}_F\, \mathbf{g}(\mathbf{X})$. Empirical moments $\hat{\mu}(g_j)$, $\hat{\mu}(\mathbf{g})$, $\hat{\sigma}^2(g_j)$ and $\hat{U}(\mathbf{g})$ are obtained by replacing F by $\hat{F}_N$. If $\mathbf{g}$ is itself stochastic, expectations will be understood in this notation to be taken conditional on the data, with the argument $\mathbf{X}$ denoting an independent replicate.

1.4. *Assumptions on F.* Our framework for ACE-type problems imposes few conditions on the nature of $\mathbf{X}$; in particular, we do not require absolute continuity or even that $\mathbf{X}$ assume values in an ordered set. Let $\mathbf{U}(\Phi) = \{\mathbf{U}(\phi): \phi \in \Phi\}$. For the remainder of this paper we will adopt the following two assumptions.

ASSUMPTION 1. The closure of $\mathbf{U}(\phi)$ is contained in $\mathbf{C}^{M,M}$; i.e., none of its limit points are singular.

ASSUMPTION 2. Every weakly convergent sequence $\{\phi^{(n)}\}$ in Φ for which $\zeta(\phi^{(n)}) \to \inf_{\phi \in \Phi} \zeta(\phi)$ is strongly convergent.

Assumption 1 implies that there is no transformation $\phi \in \tilde{\mathbf{H}}^M$ for which $\sum_{j=1}^M \phi_j(X_j) = 0$ a.e. Assumption 2 gives sufficient insurance that Φ^* is not empty. This is not a trivial concern, for unless F has finite support, Φ is not strongly compact, and it is conceivable that there is no element of Φ which attains the infimum of $\zeta(\phi)$. For the applications cited in Section 1.1, a sufficient condition for optimal transformations to exist is that the conditional expectation operators

$$E\big[\phi_j(X_j)\big|X_i\big]: \mathbf{H}_j \to \mathbf{H}_i, \qquad i \neq j, \tag{1.6}$$

are compact [Breiman and Friedman (1985) and Koyak (1985, 1987)]. In the language of Lancaster (1969), this is equivalent to the property that each of the bivariate marginal distributions of F is ϕ^2-*finite*. We do not give a proof of sufficiency at the level of generality of this paper, but instead adopt the stronger regularity condition implied in Assumption 2.

2. The method of sieves in ACE-type problems. For any element $\phi \in \tilde{\mathbf{H}}^M$ and subset $\Psi \subset \tilde{\mathbf{H}}^M$, the distance from ϕ to Ψ is defined by $d(\phi, \Psi) = \inf_{\psi \in \Psi} \|\phi - \psi\|_M$. Let $\hat{\phi}^{(N)}$ denote an optimal transformation estimate based on $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$.

DEFINITION 2.1. $\hat{\phi}^{(N)}$ is a *consistent* estimate for Φ^* if $d(\hat{\phi}^{(N)}, \Phi^*) \to 0$ a.s. F as $N \to \infty$.

The remainder of this paper will be focused on obtaining consistent estimates for Φ^* by minimizing

$$\hat{\zeta}_N^2(\phi) = \inf_{A(U(\phi))} \operatorname{trace} A'\hat{U}(\phi)A, \tag{2.1}$$

over a subset of $\hat{\Phi}_N = \{\phi \in \tilde{\mathbf{H}}^M\colon \hat{\mu}(\phi_j) = 0,\ \hat{\sigma}(\phi_j) = 1,\ j = 1, \ldots, M\}$. Minimizing over the whole of $\hat{\Phi}_N$ will generally not lead to consistent estimates unless F has finite support, since the problem is otherwise underidentified regardless of the sample size. The success of this approach depends on restricting the class of potential minimizers, but in a way that the restriction dissipates as $N \to \infty$.

A convenient framework for these problems is the method of sieves, originally proposed by Grenander (1981) and explored by Geman (1981), Geman and Hwang (1982) and a number of other authors. A *sieve* is a sequence of restricted parameter spaces $\{\Phi^{(m)}\}$ which satisfies (i) $\Phi^{(m)} \subset \Phi$; and (ii) for every $\phi \in \Phi$, there exists a sequence $\{\phi^{(m)}\}$ with $\phi^{(m)} \in \Phi^{(m)}$ such that $\|\phi - \phi^{(m)}\|_M \to 0$. A simple example of a sieve, for F concentrated in the unit M-cube, takes $\Phi^{(m)}$ to consist of transformations having the form

$$\phi_j(x_j) = \sum_{r=1}^{m} a_{rj}\big(\cos(\pi r x_j) - \mu_{rj}\big), \qquad j = 1, \ldots, M, \tag{2.2}$$

where $\mu_{rj} = E[\cos(\pi r x_j)]$, and $a_{1j}, \ldots, a_{mj}$ are arbitrary constants subject to $\|\phi_j\| = 1$, $j = 1, \ldots, M$. Here, the sieve is *nested*, but this would not be the case if, for instance, ϕ_j were a step function with jumps at the points i/m,

$i = 1, \ldots, m - 1$.

A peculiarity of ACE-type problems is that the parameter space itself is unknown, since its definition depends on F. An effective sieve for estimation purposes is *stochastic*, reflecting dependence on the sample through centering and scaling. Let m_N denote an increasing sequence of positive integers such that $m_N \uparrow \infty$.

DEFINITION 2.2. A sequence of sets $\hat{\Phi}_N^{(m_N)} \subset \hat{\Phi}_N$, is a *stochastic sieve* for Φ if

(i) For every $\phi \in \Phi$ there exits a sequence $\{\hat{\psi}^{(N)} \in \hat{\Phi}_N^{(m_N)}\}$ such that $\|\hat{\psi}^{(N)} - \phi\|_M \to 0$ with probability 1; and

(ii) $\text{Prob}\{\inf_{\hat{\phi} \in \hat{\Phi}_N^{(m_N)}} \sigma(\hat{\phi}_j) = 0 \text{ i.o.}\} = 0$ for all j.

If $\phi \in \Phi$ has $\hat{\sigma}(\phi_j) > 0$ for all j, then $\hat{\phi} \in \hat{\Phi}_N$ with $\hat{\phi}_j = (\phi_j - \hat{\mu}(\phi_j))/\hat{\sigma}(\phi_j)$ will be called its *stochastic image*, with a similar designation for the *deterministic image* of an arbitrary element $\hat{\phi} \in \hat{\Phi}_N$. We likewise refer to deterministic and stochastic images of sets, with the understanding that an element for which an image does not exist is not represented in the image set. This implies that a deterministic image does not contain transformations ϕ for which ϕ_j is constant on the sample values for some j, and in this sense, a deterministic image may also be a random set. We will let $\text{Stoch}(\Psi)$ denote the stochastic image of a subset Ψ.

An important idea that will emerge in subsequent developments in that of a uniform strong law of large numbers of sample moments taken over a class of transformations. This is made somewhat difficult in the present context by the fact that the class of transformations is itself a random set. We work around this by seeking a deterministic sieve $\{\Psi^{(N)}\}$ that, in a certain sense, dominates $\{\hat{\Phi}_N^{(m_N)}\}$:

DEFINITION 2.3. A stochastic sieve $\{\hat{\Phi}_N^{(m_N)}\}$ is *uniformly* $\mathbf{L}^2$ *consistent* (UL2C) for Φ if there exists a sequence of sets $\{\Psi^{(N)}\}$ in Φ such that $\hat{\Phi}_N^{(m_N)} \subseteq \text{Stoch}(\Psi^{(N)})$, and the following two properties hold a.s. F as $N \to \infty$:

$$\text{(2.3a)} \quad \text{(i)} \qquad \sup_{\phi \in \Psi^{(N)}} \sum_{j=1}^{M} \left|\hat{\mu}(\phi_j)\right| \to 0;$$

$$\text{(2.3b)} \quad \text{(ii)} \qquad \sup_{\phi \in \Psi^{(N)}} \sum_{i,j}^{M} \left|\hat{\mu}(\phi_i \phi_j) - \mu(\phi_i \phi_j)\right| \to 0.$$

Let $\hat{\phi}^{(N)}$ denote a minimizer of $\hat{\zeta}_N$ in $\hat{\Phi}_N^{(m_N)}$, $\hat{\zeta}_N^* = \hat{\zeta}_N(\hat{\phi}^{(N)})$, and let $\zeta^* = \inf_{\phi \in \Phi} \zeta(\phi)$. The following result links UL2C sieves and consistent optimal transformation estimates.

THEOREM 2.1. *If* $\{\hat{\Phi}_N^{(m_N)}\}$ *is a* UL2C *sieve for* Φ, *the following statements are true a.s. F as* $N \to \infty$:

(i) $\hat{\zeta}_N^* \to \zeta^*$;

(ii) $d(\hat{\phi}^{(N)}, \Phi^*) \to 0$, *i.e.*, $\hat{\phi}^{(N)}$ *is consistent for* Φ^*.

PROOF. Let $f(U, A) = \text{trace } A'UA$, and $T^2(U) = \inf_{A \in \mathbf{A}(U)} f(U, A)$. Observe that $\zeta(\phi) = T(\mathbf{U}(\Phi))$, and $\hat{\zeta}_N(\phi) = T(\hat{U}(\phi))$. That T is continuous on $\mathbf{C}^{M,M}$ follows from the continuity of f on $\mathbf{R}^{M,M} \times \mathbf{R}^{M,d}$ and the Hausdorf continuity of $\mathbf{A}(U)$. In fact, T is uniformly continuous on any compact subset of $\mathbf{C}^{M,M}$ whose interior contains the closure of $\mathbf{U}(\phi)$; the existence of such a subset is assured by Assumption 1. Let $\hat{\psi}^{(N)} \in \hat{\Phi}_N^{(m_N)}$ denote a sequence for which $d(\hat{\psi}^{(N)}, \Phi^*) \to 0$ a.s., and let $\psi^{(N)}$ and $\phi^{(N)}$ denote their deterministic images. If $\Phi^{(N)}$ is the set indicated in Definition 2.3, then $\psi^{(N)}$ and $\phi^{(N)}$ are members of $\Phi^{(N)}$. We have

$$\text{(2.4a)} \quad \hat{\zeta}_N^* - \zeta^* = \hat{\zeta}_N^* - \hat{\zeta}_N(\hat{\psi}^{(N)}) + T\big(\hat{U}(\hat{\psi}^{(N)})\big) - T\big(U(\psi^{(N)})\big) + \zeta(\psi^{(N)}) - \zeta^*,$$

$$\text{(2.4b)} \quad \hat{\zeta}_N^* - \zeta^* = T\big(\hat{U}(\hat{\phi}^{(N)})\big) - T\big(U(\phi^{(N)})\big) + \zeta(\phi^{(N)}) - \zeta(\psi^{(N)}) + \zeta(\psi^{(N)}) - \zeta^*.$$

The following arguments hold with probability 1: From (2.3a) and (2.3b), it follows that $\hat{U}(\hat{\psi}^{(N)}) - U(\psi^{(N)}) \to 0$, and $d(\psi^{(N)}, \Phi^*) \le d(\hat{\psi}^{(N)}, \Phi^*) + \|\hat{\psi}^{(N)} - \psi^{(N)}\|_M \to 0$. Since $\hat{\phi}^{(N)}$ is a minimizer of $\hat{\zeta}_N$ in $\hat{\Phi}_N^{(m_N)}$, (2.4a) therefore implies $\limsup_{N\to\infty} \hat{\zeta}_N^* - \zeta^*) \le 0$, and (2.4b) implies $\liminf_{N\to\infty}(\hat{\zeta}_N^* - \zeta^*) \ge 0$, and (a) is proven. In fact, $\zeta(\phi^{(N)}) \to \zeta^*$. Let $d_N = d(\phi^{(N)}, \Phi^*)$. Since $d(\hat{\phi}^{(N)}, \Phi^*) \le d_N + \|\hat{\phi}^{(N)} - \phi^{(N)}\|_M$, it is sufficient to show that $d_N \to 0$ to prove (b). If d_N does not go to 0, there is a weakly convergent subsequence $\{\phi^{(N_1)}\}$ with d_{N_1} bounded away from 0. But since $\zeta(\phi^{(N_1)}) \to \zeta^*$, Assumption 2 implies that this limit is strong, and therefore a member of Φ^*.

3. Consistent sieves. A sieve is uniformly $\mathbf{L}^2$ consistent if $m_N \uparrow \infty$ at the right rate. What is right depends on the type of sieve used. We show uniform $\mathbf{L}^2$ consistency for two classes of sieves. The developments of this section parallel those of Geman (1981).

3.1. *Bounded regression sieves.* Let $\{\beta_{m,rj}\}$, $r = 1, \ldots, m$; $m = 1, 2, \ldots$; $j = 1, \ldots, M$, denote a class of scalar-valued functions such that $|\beta_{m,rj}(x_j)|$ is uniformly bounded in r, m, j and $x_j \in \mathbf{R}^1$, and such that for every $\phi \in \tilde{\mathbf{H}}^M$, a sequence $\phi^{(m)} \in \tilde{\mathbf{H}}^M$ can be found where

$$\text{(3.1)} \qquad \phi_j^{(m)} = \sum_{r=1}^{m} a_{rj}^{(m)}(\beta_{m,rj} - \mu_{m,rj}), \qquad j = 1, \ldots, M,$$

$\mu_{m,rj} = E[\beta_{m,rj}(X_j)]$, and $\|\phi - \phi^{(m)}\|_M \to 0$. The class of sets taking the form

$$\text{(3.2)} \quad \hat{\Phi}_N^{m_N} = \left\{ \hat{\phi} \in \hat{\Phi}_N : \hat{\phi}_j = \sum_{r=1}^{m_N} a_{rj}(\beta_{m_N,rj} - \hat{\mu}_{m_N,rj}),\ j = 1, \ldots, M \right\}$$

will be called a *bounded regression sieve* if it meets the requirements of Definition 2.1. Let $\lambda_{m,j}$ denote the smallest eigenvalue of the $m \times m$ covariance matrix of the $\{\beta_{m,rj}\}$ taken as functions of X_j. We adopt as an additional requirement on F that a nonnegative $\alpha \in \mathbf{R}^1$ can be found such that $\liminf_{m\to\infty} m^{\alpha}\lambda_{m,j} > 0$ for all j. The following result is proved in the Appendix.

THEOREM 3.1. *Let* $m_N = [N^{\tau}]$, *where* $0 < \tau < 1/(2 + 2\alpha)$. *Then*, $\{\hat{\Phi}_N^{(m_N)}\}$ *is a UL2C sieve for* Φ.

In fact, the proof of Theorem 3.1 permits a stronger statement. Let $0 < \gamma < \frac{1}{2} - \tau(1 + \alpha)$, and $g_N = N^{\gamma}$.

COROLLARY 3.1a. *Under the conditions of Theorem* 3.1:

$$\text{(3.3a)} \quad \text{(i)} \qquad \sup_{\phi\in\Psi^{(N)}} g_N \sum_{j=1}^{M} \left|\hat{\mu}(\phi_j)\right| \to 0;$$

$$\text{(3.3b)} \quad \text{(ii)} \qquad \sup_{\phi\in\Psi^{(N)}} g_N \sum_{i=1}^{M}\sum_{j=1}^{M} \left|\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)\right| \to 0$$

a.s. as $N \to \infty$, *where* $\Psi^{(N)}$ *is as in Definition* 2.3.

If $\|U\|$ denotes the Euclidean matrix norm of $U \in \mathbf{R}^{M,M}$, Corollary 3.1a asserts that

$$\text{(3.4)} \qquad \sup_{\phi\in\Psi^{(N)}} g_N \left\|\hat{U}(\hat{\phi}) - \mathbf{U}(\phi)\right\| \to 0 \quad \text{a.s.},$$

where $\hat{\phi}$ is the stochastic image of ϕ. Let $\tilde{\mathbf{U}}$ denote a compact subset of $\mathbf{C}^{M,M}$ whose interior contains $\mathbf{U}(\Phi)$, and suppose T is differentiable with respect to U, with $T_{ij} = \partial T/\partial U_{ij}$ bounded for all i, j, and $U \in \tilde{\mathbf{U}}$ (each of the examples cited in Section 1.1 satisfy this condition). For N sufficiently large, (3.4) and the intermediate value theorem assert that for each $\phi \in \Phi^{(m_N)}$, a matrix $\tilde{U}_N(\phi) \in \tilde{\mathbf{U}}$ can be found such that

$$\text{(3.5)} \quad T\big(\hat{U}(\hat{\phi})\big) - T(\mathbf{U}(\phi)) = \sum_{i=1}^{M}\sum_{j=1}^{M} \big(\hat{U}_{ij}(\hat{\phi}) - U_{ij}(\phi)\big)T_{ij}\big(\tilde{U}_N(\phi)\big).$$

For some $0 < \kappa < \infty$, we therefore have

$$\text{(3.6)} \qquad \left|T\big(\hat{U}(\hat{\phi})\big) - T(\mathbf{U}(\phi))\right| \le \kappa\left\|\hat{U}(\hat{\phi}) - \mathbf{U}(\phi)\right\|;$$

hence,

$$\text{(3.7)} \qquad \sup_{\phi\in\Psi^{(N)}} g_N\left|T\big(\hat{U}(\hat{\phi})\big) - T(\mathbf{U}(\phi))\right| \to 0 \quad \text{a.s.}$$

Let $\phi^{(N)} \in \Psi^{(N)}$ denote any sequence where $d(\phi^{(N)}, \Phi^*) \to 0$, and take $d_N \downarrow 0$ a positive sequence where $d_N \ge d(\phi^{(N)}, \Phi^*)$. Take $h_N = g_N/(g_N d_N + 1)$

the slower of the growth rates g_N and $1/d_N$. From (3.3a), (3.3b) and (3.7) we have the following result.

COROLLARY 3.1b. $h_N(\hat{\zeta}_N^* - \zeta^*) \to 0$ *a.s. as* $N \to \infty$.

For an example of a UL2C bounded regression sieve, assume that F is *concentrated and nonnegligible* in the unit M-cube; i.e., there exists a number $f_{\min} > 0$ such that $\text{Prob}(X_j \in S) \geq f_{\min} L(S)$ for all Borel sets S in the unit interval, where $L(S)$ denotes Lebesgue measure. Take $\delta_{m,rj}(x_j) = \cos(\pi r x_j)$, $\lambda_{m,j} \geq 2 f_{\min} > 0$, and Theorem 3.1 applies for any $\tau < 1/2$.

Other candidates for bounded regression sieves include Hermite functions, step functions and B-splines. It should be noted, however, that for the latter two sieves, $\lambda_{m,j} = 0$. As a result, Theorem 3.1 is not an effective vehicle for establishing UL2C rates for these sieves. By a modification to the proof of Theorem 3.1, it can be shown that, under the same assumptions applied to the cosine sieve, the UL2C property holds for equispaced knots in either the step function or B-spline sieve, if the distance between successive knots is on the order of $N^{-\tau}$, with $\tau < 1/3$. Results on these two sieves will be described in a forthcoming paper.

3.2. *Regularized sieves.* Consider the set of functions $\mathbf{W}_k^{(m)} \subset \mathbf{C}^{k-1}[0,1]$ for which $w \in \mathbf{W}_k^{(m)}$ implies w is $k-1$ times continuously differentiable; $w^{|k|}(x) = d^k w/dx^k$ is piecewise continuous, satisfying

$$\int_0^1 \left(w^{|k|}(x)\right)^2 dx \leq m; \tag{3.8}$$

and $w^{|p|}(0^+) = w^{|p|}(1^-) = 0$ for all odd $p \leq k-1$. We will say that an element $\phi \in \tilde{H}^M$ is k, m-*regularized* if each of its coordinates ϕ_j is in $\mathbf{W}_k^{(m)}$. If $\hat{\Phi}_N^{(m_N)}$ is chosen to be the set of all k, m_N-regularized transformations in $\hat{\Phi}_N$, it is clearly necessary that $m_N \to \infty$ if we wish to consider $\{\hat{\Phi}_N^{(m_N)}\}$ as a stochastic sieve for Φ. In this case, we will refer to $\{\hat{\Phi}_N^{(m_N)}\}$ as a k, m_N-regularized class.

This approach to smoothing is often associated with spline theory, a subject of considerable interest in statistics and applied mathematics. Wegman and Wright (1983) give a thorough survey of the spline literature in statistics, among which the contributions of Wahba and her colleagues are prominent. For $k = 1$, the endpoint constraints are absent, and one can argue as in de Boor (1978) that a minimizer $\hat{\phi}^{(N)}$ of $\hat{\zeta}_N$ in $\hat{\Phi}_N^{(m_N)}$ can be found such that the coordinates $\hat{\phi}_j^{(N)}$ are piecewise-linear functions, with derivative discontinuities occurring at the data values $X_j^{(1)}, \ldots, X_j^{(N)}$. In general, for *natural* endpoint constraints [imposed on the kth through $(2k-1)$th derivatives], $\hat{\phi}_j^{(N)}$ can similarly be chosen as a spline of order $2k-1$ with knots at the data values. However, for $k \geq 2$, our conditions are not natural, so an association of $\{\hat{\Phi}_N^{(m_N)}\}$ with smoothing splines may be misleading.

Fix $w \in \mathbf{W}_k^{(m)}$, and write $w(x) = \sum_{r=0}^{\infty} a_r \cos(\pi r x)$ to denote its Fourier series. Integrating in (3.8) by parts is easily seen to yield

$$\frac{1}{2}\pi^{2k} \sum_{r=1}^{\infty} a_r^2 r^{2k} \leq m. \tag{3.9}$$

This establishes a certain similarity between the k-regularized classes and bounded regression sieves with cosines for the $\beta_{m,rj}$, the distinction lying in the smallness constraints placed on the coefficients. It also seen that, for every N, the k, m_N-regularized sets $\hat{\Phi}_N^{(m_N)}$ are decreasingly nested in k.

Let $\tau_k = (2k - 1)/4k$; i.e., $\tau_1 = 1/4$, and $\tau_2 = 3/8$. We give the proof of the following result in the Appendix.

THEOREM 3.2. *Let F be concentrated in the unit M-cube with a density bounded away from both 0 and ∞. Choose $\tau \in (0, \tau_k)$. Take $m_N = [N^\tau]$, and let $\hat{\Phi}_N^{(m_N)}$ denote the set of k, m_N-regularized elements of $\hat{\Phi}_N$. Then, $\{\hat{\Phi}_N^{(m_N)}\}$ is a UL2C sieve for Φ.*

It is interesting to note that Geman (1981), in the context of estimating a regression function, and using $k = 1$, obtained an upper bound of $1/4$ for the growth rate of the sieve. Choose τ in accordance with Theorem 3.2, let

$$\gamma_k(\tau) = \frac{1}{2} - \frac{1}{4k} - \tau, \tag{3.10}$$

and choose $\gamma \in (0, \gamma_k(\tau))$. Take $g_N = N^\gamma$, and define h_N as preceding Corollary 3.1b. Let $\hat{\zeta}_N^*$ denote the minimum value of $\hat{\zeta}_N(\hat{\phi})$ over $\hat{\phi} \in \hat{\Phi}_N^{(m_N)}$, where $\{\hat{\Phi}_N^{(m_N)}\}$ is the k-regularized sieve of Theorem 3.2 or Corollary 3.2a.

COROLLARY 3.2a. *If $T(U)$ satisfies the differentiability conditions of Corollary* 3.1b, *then $h_N(\hat{\zeta}_N^* - \zeta^*) \to 0$ a.s. as $N \to \infty$.*

4. Discussion. An attractive feature of the method of sieves is that it permits an elegant, yet very general approach to establishing consistency for a large range of infinite-dimensional estimation problems. In addition to the class of nonlinear multivariate problems encompassed in our ACE formalism, this strategy has been successfully applied to nonparametric regression and maximum likelihood [Geman (1981) and Geman and Hwang (1982)], to estimating the drift of a diffusion [Geman (1982)] and to estimating the intensity function of a Poisson process [Karr, Miller and Snyder, (1986)], to name several applications. While our results required little in the way of problem specification, or assumptions on the parameter being estimated, a narrower approach may allow a more comprehensive treatment of consistency properties in certain problems. We cite the work of Stone (1985) in additive regression and related nonparametric modeling, and recent work of Burman (1986), who has obtained results similar to those of Stone in the regression ACE setting, as

cases in point. Stone (1986) generalizes the additive approach to multivariate nonparametric estimation within the broad scope of exponential families.

APPENDIX

The proofs of Theorems 3.1 and 3.2 will make use of the following variant of an inequality due to Hoeffding (1963).

Let $Z_1, \ldots, Z_N$ denote a sequence of independent random variables which satisfy $|Z_i| \le C$, and $E(Z_i) = 0$, $i = 1, \ldots, N$. Then, for every $\varepsilon > 0$,

$$\text{(A.1.1)} \qquad \text{Prob}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i\right| \ge \varepsilon\right) \le 2\exp\left(\frac{-N\min(\varepsilon^2, 1/4)}{2C^2}\right).$$

PROOF OF THEOREM 3.1. Our arguments are similar to Theorem 1 of Geman (1981). Positive constants whose values are not of importance will be denoted $c, c_1, c_2, \ldots$. We will adopt the following vector and matrix notation:

$$\mathbf{a}_j = (a_{1j}, \ldots, a_{m_N j})', \qquad \boldsymbol{\beta}_j = (\beta_{m_N,1j}, \ldots, \beta_{m_N,M_N j})', \qquad \tilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j - \boldsymbol{\mu}(\boldsymbol{\beta}_j),$$

$$\mathbf{V}_{m_N} = \text{Cov}_F\big(\boldsymbol{\beta}_j(X_j)\big), \qquad \hat{\mathbf{V}}_{m_N} = \text{Cov}_{\hat{F}_N}\big(\boldsymbol{\beta}_j(X_j)\big), \qquad j = 1, \ldots, M.$$

Norms applied to vectors and matrices are taken in the Euclidean sense. The notation $\|\cdot\|_\infty$ refers to the sup-norm of a function.

Let ϕ: $\hat{\phi}_j = \mathbf{a}_j'(\boldsymbol{\beta}_j - \hat{\mu}(\boldsymbol{\beta}_j))$ denote an element of $\hat{\Phi}_N^{(m_N)}$. Observe that

$$\begin{aligned} 1 = \hat{\sigma}^2\big(\hat{\phi}_j\big) &= \mathbf{a}_j'\hat{\mathbf{V}}_{m_N}\mathbf{a}_j \\ &\le \|\mathbf{a}_j\|^2\lambda_{\max}\big(\hat{\mathbf{V}}_{m_N}\big) \\ &\le cm_N\|\mathbf{a}_j\|^2, \end{aligned} \qquad \text{(A.1.2)}$$

where c does not depend on N due to the boundedness of the $\beta_{m_N,rj}$. Hence, $\|\mathbf{a}_j\|^2 \ge 1/cm_N$. Since $\sigma^2(\hat{\phi}_j) = \mathbf{a}_j'\mathbf{V}_{m_N}\mathbf{a}_j \ge \|\mathbf{a}_j\|^2\lambda_{m_N,j}$, we conclude that (ii) in Definition 2.2 is satisfied.

Consider the set $\Phi^{(m_N)}$ of transformations $\phi \in \Phi$ which have the form $\phi_j = \mathbf{a}_j'\tilde{\boldsymbol{\beta}}_j$, $j = 1, \ldots, M$. Observe that $\text{Stoch}(\Phi^{(m_N)}) = \hat{\Phi}_N^{(m_N)}$, and that $\{\Phi^{(m_N)}\}$ is a sieve for Φ. We shall proceed to show that (2.3b) holds by taking $\Psi^{(N)} = \Phi^{(m_N)}$; (2.3a) will follow a fortiori. Fix arbitrary $\varepsilon > 0$ (actually, taking $0 < \varepsilon < 1/4$ will be more convenient and will entail no loss of generality in the arguments to follow). By the first Borel–Cantelli lemma, to prove (2.3b) it is sufficient to establish that

$$\text{(A.1.3)} \qquad \sum_{N=1}^{\infty} \text{Prob}\left\{\sup_{\phi \in \Phi^{(m_N)}} \left|\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)\right| > \varepsilon\right\} < \infty$$

for all $i, j = 1, \ldots, M$; this follows from

$$\text{(A.1.4)} \qquad \begin{aligned} &\text{Prob}\left\{ \sup_{\phi \in \Phi^{(m_N)}} \sum_{i,j=1}^{M} \left|\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)\right| > \varepsilon \right\} \\ &\quad \le \sum_{i,j=1}^{M} \text{Prob}\left\{ \sup_{\phi \in \Phi^{(m_N)}} \left|\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)\right| > \varepsilon \right\}. \end{aligned}$$

An element $\phi \in \Phi^{(m_N)}$ with $\phi_j = \mathbf{a}_j'\tilde{\boldsymbol{\beta}}_j$ satisfies $\|\mathbf{a}_j\|^2 \le c_1 m_N^{\alpha}$. Write $\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j) = \mathbf{a}_i'\mathbf{Z}\mathbf{a}_j$, where $\mathbf{Z} = \hat{\boldsymbol{\mu}}(\tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_j') - \mu(\tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_j')$ is an $m_N \times m_N$ matrix of averages of N i.i.d. bounded random variables having mean 0. Using the Cauchy–Schwarz inequality, we have

$$\text{(A.1.5)} \qquad \begin{aligned} |\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)|^2 &\le \|\mathbf{a}_i\|^2\|\mathbf{a}_j\|^2\|\mathbf{Z}\|^2 \\ &\le c_1^2 m_N^{2\alpha} \sum_{r,s=1}^{m_N} Z_{rs}^2, \end{aligned}$$

where the right-hand side of (A.1.5) does not depend on the coefficient vectors $\mathbf{a}_j$. Using Hoeffding's inequality, we now obtain

$$\text{(A.1.6)} \qquad \begin{aligned} \text{Prob}\left\{ m_N^{2\alpha} \sum_{r,s=1}^{m_N} Z_{rs}^2 > \varepsilon^2 \right\} &\le \sum_{r,s=1}^{m_N} \text{Prob}\{|Z_{rs}| > \varepsilon m_N^{-1-\alpha}\} \\ &\le m_N^2 \exp\{-c_2\varepsilon^2 N m_N^{-2-2\alpha}\}. \end{aligned}$$

Taking $m_N \approx N^{\tau}$, it is readily seen that the last term in (A.1.6) is summable with respect to N provided that $0 < \tau < 1/(2 + 2\alpha)$, thus establishing (2.3b).

To complete the proof, we need to show that $\{\hat{\Phi}_N^{(m_N)}\}$ satisfies (i) in Definition 2.2. Choose arbitrary $\phi \in \Phi$. Since $\{\Phi^{(m_N)}\}$ is a sieve for Φ, we can find a sequence $\{\phi^{(N)} \in \Phi^{(m_N)}\}$ such that $\|\phi^{(N)} - \phi\|_M \to 0$. With probability 1, for N sufficiently large, $\phi^{(N)}$ will admit a stochastic image $\hat{\phi}^{(N)} \in \hat{\Phi}_N^{(m_N)}$ [this follows from (2.3a) and (2.3b)] which moreover satisfies $\|\hat{\phi}^{(N)} - \phi^{(N)}\|_M \to 0$ a.s. Hence, $\|\hat{\phi}^{(N)} - \phi\|_M \to 0$ a.s., and the proof is complete. □

PROOF OF COROLLARY 3.1a. Replace ε with $N^{-\gamma}\varepsilon$ in the proof of Theorem 3.1. □

PROOF OF THEOREM 3.2. Again we will denote positive constants whose values are not of importance by $c, c_1, c_2, \ldots$. Our strategy of proof is similar to that of Theorem 2 of Geman (1981).

Let $\hat{\phi} \in \hat{\Phi}_N^{(m_N)}$, with $\hat{\phi}_j = \sum_{r=1}^{\infty} a_{rj}(\cos(\pi r x_j) - \hat{\mu}_{rj})$ satisfying

$$\pi^{2k} \sum_{r=1}^{\infty} a_{rj}^2 r^{2k} \le 2m_N, \qquad j = 1, \ldots, M,$$

where $\hat{\mu}_{rj} = \hat{\mu}(\cos(\pi r X_j))$. We will show that $\{\hat{\Phi}_N^{(m_N)}\}$ satisfies (ii) of Definition 2.2. For an arbitrary integer $R > 0$, write $\hat{\phi}_j = \hat{\phi}_j^{(R)} + \hat{\theta}_j^{(R)}$, where $\hat{\phi}_j^{(R)}$ is the sum of the first R terms in the Fourier expansion of $\hat{\phi}_j$, and $\hat{\theta}_j^{(R)}$ denotes the

remainder. Observe that

$$1 = \hat{\sigma}\big(\hat{\phi}_j\big) \le \hat{\sigma}\big(\hat{\phi}_j^{(R)}\big) + \hat{\sigma}\big(\hat{\theta}_j^{(R)}\big). \tag{A.2.1}$$

Let $\beta_{rj} = \cos \pi r X_j$, $\boldsymbol{\beta}_{R,j} = (\beta_{1j}, \ldots, \beta_{Rj})'$, $\mathbf{V}_{R,j} = \boldsymbol{\mu}(\boldsymbol{\beta}_{R,j}\boldsymbol{\beta}'_{R,j})$, $\hat{\mathbf{V}}_{R,j} = \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_{R,j}\boldsymbol{\beta}'_{R,j})$ and $\mathbf{a}_{R,j} = (a_{1j}, \ldots, a_{Rj})'$. We claim that there is a positive constant c_1 such that $\|\mathbf{a}_{R,j}\| \ge c_1$ holds for all R a.s. Observe that

$$\begin{aligned} \hat{\sigma}^2\big(\hat{\phi}_j^{(R)}\big) &\le \mathbf{a}'_{R,j}\hat{\mathbf{V}}_{R,j}\mathbf{a}_{R,j} \\ &\le \|\mathbf{a}_{R,j}\|^2\big[\lambda_{\max}(\mathbf{V}_{R,j}) + \lambda_{\max}\big(\hat{\mathbf{V}}_{R,j} - \mathbf{V}_{R,j}\big)\big], \end{aligned} \tag{A.2.2}$$

where now $\lambda_{\max}(\cdot)$ denotes the spectral radius of its argument. Because the density is bounded from above, and $2\int_0^1 \cos \pi rx \cos \pi sx\, dx = \delta_{rs}$, we have $\lambda_{\max}(\mathbf{V}_{R,j}) \le c_2$ for all R.

Let $Z_{r,s} = \hat{\mu}(\beta_{rj}\beta_{sj}) - \mu(\beta_{rj}\beta_{sj})$. By Theorem 5.6.7 of Graybill (1983), we have $\lambda_{\max}(\hat{\mathbf{V}}_{R,j} - \mathbf{V}_{R,j}) \le \sup_{1 \le r \le R} \sum_{s=1}^{R} |Z_{r,s}|$. Take $R \approx N^{\xi}$ for $\xi > 0$ to be determined later. Since $Z_{r,s}$ is the mean of N i.i.d. uniformly bounded random variables, Hoeffding's inequality asserts that

$$\sup_{1 \le r,s \le R} |Z_{r,s}| \le c_3(\log N/N)^{1/2} \tag{A.2.3}$$

holds for all N a.s. for a constant c_3 that depends on ξ. Hence,

$$\begin{aligned} \hat{\sigma}^2\big(\hat{\phi}^{(R)}\big) &\le \|\mathbf{a}_{R,j}\|^2\big(c_2 + c_3 R(\log N/N)^{1/2}\big) \\ &\le c_4^2\|\mathbf{a}_{r,j}\|^2\big(1 + R(\log N/N)^{1/2}\big) \quad \text{a.s.} \end{aligned} \tag{A.2.4}$$

We also obtain

$$\begin{aligned} \hat{\sigma}^2\big(\hat{\theta}_j^{(R)}\big) &\le 2 \sum_{s=R+1}^{\infty} s^{-2k} \sum_{r=1}^{\infty} a_{rj}^2 r^{2k} \\ &\le c_5^2 \frac{N^{\tau}}{R^{2k-1}}. \end{aligned} \tag{A.2.5}$$

Combining (A.2.1), (A.2.4) and (A.2.5) yields

$$1 \le c_4\|\mathbf{a}_{R,j}\|\big(1 + R(\log N/N)^{1/2}\big)^{1/2} + c_5\big(N^{\tau}/R^{2k-1}\big)^{1/2} \quad \text{a.s.} \tag{A.2.6}$$

By choosing $\tau/(2k-1) < \xi < 1/2$, both $R(\log N/N)^{1/2} \to 0$ and $N^{\tau}/R^{2k-1} \to 0$, and $\|\mathbf{a}_{R,j}\| \ge c_1$ a.s. as claimed.

Since $\sigma^2(\hat{\phi}_j) \ge 2 f_{\min} \sum a_{rj}^2 \ge 2 f_{\min} c_1^2$, (ii) in Definition 2.2 is satisfied. It likewise follows that if ϕ is the deterministic image of an element $\hat{\phi} \in \hat{\Phi}_N^{(m_N)}$, then on a set with probability 1, ϕ_j will satisfy the boundedness condition (3.9) for a rate p_N taken proportional to m_N. Thus, if $\Phi^{(m)}$ denotes the class of k, m-regularized transformations in Φ, we have

$$\hat{\Phi}_N^{(m_N)} \subseteq \mathrm{Stoch}(\Phi^{(p_N)}). \tag{A.2.7}$$

By considering truncated cosine series, it is not hard to show that $\{\Phi^{(p_N)}\}$ is a sieve for Φ. Taking $\Phi^{(p_N)}$ to play the role of $\Psi^{(N)}$ in Definition 2.3, we will

proceed to show that this sieve satisfies (2.3a) and (2.3b). By arguments similar to those used in the proof of Theorem 3.1, this will be sufficient to complete the proof of Theorem 3.2.

Fixing i and j, we will show that (A.1.3) holds, with $\Phi^{(m_N)}$ replaced by $\Phi^{(p_N)}$, where $p_N \approx N^\tau$. Let $\eta > 0$ be some number to be determined later, and let $R \equiv R_N \approx N^\eta$. For $\phi \in \Phi^{(p_N)}$, we again write $\phi_j = \phi_j^{(R)} + \theta_j^{(R)}$. We then have

$$\begin{aligned}\left|\hat{\mu}(\phi_i\phi_j) - \mu(\phi_i\phi_j)\right| \le{}& \left|\hat{\mu}\left(\phi_i^{(R)}\phi_j^{(R)}\right) - \mu\left(\phi_i^{(R)}\phi_j^{(R)}\right)\right| \\ &+ \left|\hat{\mu}\left(\phi_i^{(R)}\theta_j^{(R)}\right)\right| + \left|\hat{\mu}\left(\theta_i^{(R)}\phi_j^{(R)}\right)\right| + \left|\hat{\mu}\left(\theta_i^{(R)}\theta_j^{(R)}\right)\right| \\ &+ \left|\mu\left(\phi_i^{(R)}\theta_j^{(R)}\right)\right| + \left|\mu\left(\theta_i^{(R)}\phi_j^{(R)}\right)\right| + \left|\mu\left(\theta_i^{(R)}\theta_j^{(R)}\right)\right|,\end{aligned} \tag{A.2.8}$$

where

$$\begin{aligned}\left\|\phi_j^{(R)}\right\|_\infty &\le 2\left(\sum_{r=1}^{\infty} b_{rj}^2 r^{2k}\right)^{1/2}\left(\sum_{r=1}^{\infty} r^{-2k}\right)^{1/2} \\ &\le c_4 p_N^{1/2},\end{aligned} \tag{A.2.9a}$$

$$\begin{aligned}\left\|\theta_j^{(R)}\right\|_\infty &\le 2\left(\sum_{r=1}^{\infty} b_{rj}^2 r^{2k}\right)^{1/2}\left(\sum_{r=R+1}^{\infty} r^{-2k}\right)^{1/2} \\ &\le c_5 p_N^{1/2} R_N^{-(2k-1)/2}.\end{aligned} \tag{A.2.9b}$$

If η is chosen so that $p_N R_N^{-(2k-1)/2} \to 0$ as $N \to \infty$, then the last six terms in the right-hand side sum in (A.2.8) will *surely* go to 0 uniformly. Assuming that this is the case, we will now find conditions for the first of these terms to go to 0 uniformly.

Let $\tilde{\boldsymbol{\beta}}_{R,j}$ denote the R-vector of functions $\cos \pi r X_j - \mu_{rj}$, $r = 1, \ldots, R$, and let $\mathbf{Z} = \hat{\boldsymbol{\mu}}(\tilde{\boldsymbol{\beta}}_{R,i}\tilde{\boldsymbol{\beta}}'_{r,j}) - \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}_{R,i}\tilde{\boldsymbol{\beta}}'_{r,j})$. We thus have $\hat{\mu}(\phi_i^{(R)}\phi_j^{(R)}) - \mu(\phi_i^{(R)}\phi_j^{(R)}) = \mathbf{b}'_{R,i}\mathbf{Z}\mathbf{b}_{R,j}$, where

$$\begin{aligned}\mathbf{b}'_{R,i}\mathbf{Z}\mathbf{b}_{R,j} &\le p_N\left(\sum_{r,s=1}^{R} (rs)^{-2k} Z_{rs}^2\right)^{1/2} \\ &\le p_N\left(\sum_{r=1}^{R} r^{-4k}\right)^{1/2}\left(\sum_{r,s=1}^{R} Z_{rs}^4\right)^{1/4} \\ &= c_6 p_N\left(\sum_{r,s=1}^{R} Z_{rs}^4\right)^{1/4}.\end{aligned} \tag{A.2.10}$$

The last term in (A.2.10) does not depend on the coefficients b_{rj}, and the Z_{rs} are averages of N i.i.d. bounded random variables having zero mean. Using

Hoeffding's inequality, we obtain

$$\text{(A.2.11)}\quad \text{Prob}\left\{p_N^4 \sum_{r,s=1}^{R} Z_{rs}^4 > \varepsilon^4\right\} \le \sum_{r,s=1}^{R} \text{Prob}\{|Z_{rs}| > \varepsilon p_N^{-1} R_N^{-1/2}\} \le R_N^2 \exp\{-c_7 \varepsilon^2 N p_N^{-2} R_N^{-1}\}.$$

In order for the rightmost term in (A.2.11) and for $p_N R_N^{(2k-1)/2}$ to both go to 0, the following conditions need to be satisfied:

$$\text{(A.2.12)}\quad \begin{aligned} &(1) && \tau, \eta > 0; \\ &(2) && 2\tau + \eta < 1; \\ &(3) && \tau - \frac{2k-1}{2}\eta < 0. \end{aligned}$$

It is easy to see that the largest τ in the closure of this admissible region is $\tau^* = (2k-1)/4k$, and the claim of the theorem is affirmed. □

PROOF OF COROLLARY 3.2a. Replace ε by $N^{-\gamma}\varepsilon$ in the proof of Theorem 3.2, treating τ as fixed. The conclusion follows from straightforward calculations leading to conditions similar to (A.2.12). □

Acknowledgments. I would like to thank the referees for making suggestions that led to significant improvements in both the substance and style of this paper. In particular, I would like to acknowledge the comments of one referee that greatly improved the result of Theorem 3.1.

REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619.

BREIMAN, L. and IHAKA, R. (1984). Nonlinear discriminant analysis via scaling and ACE. Technical Report 40, Dept. Statist., Univ. California, Berkeley.

BURMAN, P. (1986). Rates of convergence for the estimation of the optimal transformations in correlation and regression. Technical Report, Dept. Statist., Univ. California at Davis.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

GEMAN, S. (1981). Sieves for nonparametric estimation of densities and regressions. Reports in Pattern Analysis 99. Division App. Math., Brown Univ.

GEMAN, S. (1982). An application of the method of sieves: Functional estimator for the drift of a diffusion. In *Nonparametric Statistical Inference* (B. V. Gnedenko, M. L. Puri and I. Vincze, eds.) **1** 231–252. North-Holland, Amsterdam.

GEMAN, S. and HWANG, C. R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GIFI, A. (1981). *Nonlinear Multivariate Analysis*. Dept. Datatheorie, Univ. Leiden.

GRAYBILL, F. A. (1983). *Matrices with Applications to Statistics*, 2nd ed. Wadsworth, Belmont, Calif.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

HAUSDORF, F. (1962). *Set Theory*. Chelsea, New York.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KARR, A. F., MILLER, M. I. and SNYDER, D. L. (1986). Estimation of intensity functions of Poisson processes via the method of sieves, with application to positron emission tomography. Unpublished manuscript.

KOYAK, R. (1985). Optimal transformations for multivariate linear reduction analysis. Ph.D. thesis, Dept. Statist., Univ. California, Berkeley.

KOYAK, R. (1987). On measuring internal dependence in a set of random variables. *Ann. Statist.* **15** 1215–1228.

LANCASTER, H. O. (1969). *The Chi-Squared Distribution*. Wiley, New York.

OWEN, A. (1983). Optimal transformations for autoregressive time series models. Technical Report ORION020. Dept. Statist., Stanford Univ.

STONE, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

WEGMAN, E. J. and WRIGHT, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assoc.* **78** 351–365.

ZEHNA, P. W. (1966). Invariance of maximum likelihood estimators. *Ann. Statist.* **37** 744.

DEPARTMENT OF MATHEMATICAL SCIENCES
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218