Econometric Theory, **14**, 1998, 295–325. Printed in the United States of America.

CONSISTENT SPECIFICATION TESTING WITH NUISANCE PARAMETERS PRESENT ONLY UNDER THE ALTERNATIVE

MAXWELL B. STINCHCOMBE
University of Texas

HALBERT WHITE
University of California, San Diego

The nonparametric and the nuisance parameter approaches to consistently testing statistical models are both attempts to estimate topological measures of distance between a parametric and a nonparametric fit, and neither dominates in experiments. This topological unification allows us to greatly extend the nuisance parameter approach. How and why the nuisance parameter approach works and how it can be extended bear closely on recent developments in artificial neural networks. Statistical content is provided by viewing specification tests with nuisance parameters as tests of hypotheses about Banach-valued random elements and applying the Banach central limit theorem and law of iterated logarithm, leading to simple procedures that can be used as a guide to when computationally more elaborate procedures may be warranted.

1. INTRODUCTION

In testing whether or not a parametric statistical model is correctly specified, there are a number of apparently distinct approaches one might take. The nonparametric approach compares a nonparametric estimator of the object of interest (say a conditional mean or density) to a parametric estimator (e.g., Eubank and Spiegelman, 1990; Eubank and LaRiccia, 1992; Gozalo, 1993; Härdle and Mammen, 1993; Hong and White, 1995; Leung and Yu, 1995; or Zheng, 1996). The nuisance parameter approach tests whether a statistic depending on a "nuisance parameter present only under the alternative" is zero for all values of the nuisance parameter, as is true under the null (e.g., Davies, 1977, 1987; Bierens, 1990). There are a variety of other possibilities and variants as well (e.g., Eubank and Hart, 1992; Eubank and LaRiccia, 1992; Blum, Kiefer, and Rosenblatt, 1961;

White's participation was supported by NSF grants SES-9209023 and SBR-9511253. We are grateful to Rob Engle, Clive Granger, and the participants of the UCSD econometrics workshops for helpful comments. Address correspondence to: Halbert White, Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA.

 0266-4666/98 $9.50

Ghorai, 1980; Holst and Rao, 1980; Robinson, 1991; Schweizer and Wolff, 1976, 1981; Wolff, 1981; or Zheng, 1994).

Our purpose here is to unify the apparently disparate nonparametric and nuisance parameter approaches to testing models consistently for arbitrary misspecification (i.e., with power approach one asymptotically for all deviations from the null). The insight providing this unification is that, fundamentally, all the different tests, and in particular the two of direct interest to us, are based on estimates of topological "distances" between a restricted (e.g., parametric) model and an unrestricted model. In this context, the notion of weak denseness or weak denseness of a span in the space containing the object of interest plays the central role. Further, verifying weak denseness is often quite easy. As we shall see, the two forms of the tests are distinct because one estimates the topological distance directly in the nonparametric approach and indirectly in the nuisance parameter approach.

By identifying the topological basis for a test and applying the notion of weak denseness appropriately, the fundamental relations between many of the different specifications testing approaches can be appreciated. As just one example, Eubank and Hart (1992) base a nonparametric specification test for a regression model on the number of included terms appearing in an auxiliary regression. The estimated residuals of the original regression are the dependent variable, and Fourier series terms in the original explanatory variables are the regressors. They reject the null hypothesis of correct specification if the number of included Fourier terms is greater than zero (including a penalty for the number of terms to control overfitting). This amounts to direct use of a nonvector space topology on the space of possible conditional expectations of the residual given the explanatory variable—estimating "distance" from zero in a topology where "distance" is the number of terms in the Fourier series representing the function. It is the weak denseness of Fourier series that makes this test consistent.

Thus, one can develop a taxonomy for specification testing that classifies tests according to the object tested (e.g., a density or an expectation), the topology forming the basis for the test (e.g., a normable or a nonnormable vector space topology or a nonvector space topology), whether or not the distance is estimated directly or indirectly, and other details of the testing procedure. We leave this exercise to other work but note that an informal survey reveals combinations of features that have not yet been proposed for testing and which therefore constitute interesting research opportunities.

An additional benefit to our approach is that it permits us to see the unity underlying procedures directed toward testing models of distribution on the one hand and regression models (i.e., models of conditional expectation) on the other. Specifically we shall see how such apparently disparate procedures as the Kolmogorov–Smirnov or Cramér–von Mises tests for the difference of distributions and Bierens's (1990) test for regression model misspecification have a common origin. This unity suggests multivariate analogs of the Kolmogorov–Smirnov and Cramér–von Mises tests that may have useful power.

More broadly, our unification of the nonparametric and nuisance parameter approaches to testing model specification permits us to obtain considerable extensions of the nuisance parameter approach, both in terms of the procedures themselves and their domain of applicability. Further, by putting such problems in the context of testing hypotheses about Banach-valued random variables we can obtain some simple test procedures that do not require computing the complicated null distributions that can arise using such statistics and which can be used as a simple guide as to when more elaborate computations may be warranted.

The plan of the paper is as follows. In Section 2, we discuss Bierens's (1990) nuisance parameter approach to consistent testing of regression models to expose certain fundamental issues and provide useful background for subsequent developments. As a foretaste of the results to come, we present an extension of Bierens's results establishing the existence of a broad range of similar procedures with identical properties. Section 3 addresses the topological underpinnings of the nuisance parameter approach for both regression models and probability models and unifies it with the nonparametric approach. In this context, we also will observe interesting connections between Bierens's approach and the approximation capabilities of artificial neural networks (Hornik, Stinchcombe, and White, 1989; White, 1989a). In Section 4 we show how the nuisance parameter approach to specification testing applies to testing probability models generally. Section 5 lays out the Banach random variable testing approach and obtains some simple but useful statistical procedures.

2. NUISANCE PARAMETER CONSISTENT TESTING OF REGRESSION MODELS

For the sake of explicitness, all random variables are assumed to be defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and $\sigma(X) \subset \mathcal{F}$ denotes the minimal σ-field making the random variable $X: \Omega \to \mathbb{R}^k$ measurable. Also for explicitness, parametrize the class of functions (or model) $\mathcal{S} := \{f(\cdot, \theta): \mathbb{R}^k \to \mathbb{R} | \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$. The model is correctly specified for $E(Y|X)$ when $f(X, \theta_0)$ is a version of $E(Y|X)$ for some $\theta_0 \in \Theta$. Let $\{\hat{\theta}_n: \Omega \to \Theta\}$ be a sequence of estimators consistent for a "pseudotrue" value $\theta^* \in \Theta$ regardless of the correctness of $\mathcal{S}$, such that $\theta^* = \theta_0$ when $\mathcal{S}$ is correctly specified. For example, $\hat{\theta}_n$ can be a nonlinear least-squares estimator from a random sample of size n on (Y, X) (e.g., White, 1981).

With $\epsilon := Y - f(X, \theta^*)$, the correct specification of $\mathcal{S}$ for $E(Y|X)$ is equivalent to $e := E(\epsilon|X) = 0$ a.s. Now, under suitable conditions on Y and f, e is an integrable function of X, that is, $e \in L^p(X) := L^p(\Omega, \sigma(X), P)$ for some $p \in [1, \infty]$. For $1/p + 1/q = 1$, $L^p(X)$ and $L^q(X)$ are a dual pair, that is, $g \in L^p(X)$ is equal to 0 if and only if $\langle g, h \rangle := \int gh \, dP = 0$ for all $h \in L^q(X)$, and $h \in L^q(X)$ is equal to 0 if and only $\langle g, h \rangle = 0$ for all $g \in L^p(X)$. Thus, the existence of a test function $h \in L^q(X)$ such that $\langle e, h \rangle \neq 0$ is sufficient to conclude that $e \neq 0$. The law of iterated expectations gives $\langle \epsilon, h \rangle \equiv \langle e, h \rangle$. Consequently, if for some test function h we

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

have evidence that $E(h(X)\epsilon) \neq 0$, then we have evidence of misspecification of $\mathcal{S}$ for $E(Y|X)$.

It is this fact that ensures the power against various misspecifications of the Hausman (1978) test, because the Hausman test implicitly uses a particular choice for h (see White, 1994, Ch. 10). Nevertheless, as Holly (1982) and Bierens (1982) showed, the Hausman test can fail to have power against potentially important departures from H_0, the null hypothesis of correct specification. The same is true of the conditional moment specification tests of Newey (1985), Tauchen (1985), and White (1987, 1994), which also rely on various specific choices of test function h to obtain power against a range of misspecifications.

One way to ensure power against arbitrary misspecification (provided that Y has finite variance) is to choose $h = e$ as the test function because $\|e\|^2 = \langle e, e\rangle = 0$ if and only if $e := E(\epsilon|X) = 0$ a.s. Although e is unknown a priori, consistent nonparametric estimators $\tilde{e}$ of e come arbitrarily close to e, ensuring the utility of the nonparametric approaches to testing regression models of specification (e.g., Eubank and Spiegelman, 1990; Hong and White, 1995). The nuisance parameter approach ensures power through essentially the same considerations.

Recall that the existence of a test function h such that $\langle e, h\rangle \neq 0$ is evidence of misspecification. Suppose that $\mathcal{H} \subset L^q(X)$ satisfies

$$(\forall g \in L^p(X))(\forall h \in L^q(X))(\forall \delta > 0)(\exists h' \in \operatorname{sp}\mathcal{H})[|\langle g, h\rangle - \langle g, h'\rangle| < \delta], \tag{1}$$

where $\operatorname{sp}\mathcal{H}$ is the span of $\mathcal{H}$. Using the linearity of $\langle g, \cdot\rangle$ and the definition of the span, it is immediate that $e = 0$ if and only if $\langle e, h\rangle = 0$ for all $h \in \mathcal{H}$. Now, the weak topology on $L^q(X)$ is by definition the weakest (or smallest) topology making all of the continuous linear functions of the form $g \mapsto \langle f, g\rangle$, $f \in L^p(\mu)$, continuous. Directly from this definition, we see that a class $\mathcal{H}$ satisfies (1) if and only if $\operatorname{sp}\mathcal{H}$ is dense in the weak topology in $L^q(X)$. It is this weak denseness that is at the basis of the consistency of the nuisance parameter approach.

For example, Bierens (1990) showed that the class of functions $\mathcal{H}_{\exp} = \{h_\tau | h_\tau(x) = \exp(x'\tau), \tau \in \mathbb{R}^k\}$ has the property that whenever $e \neq 0$, there is an $h \in \mathcal{H}_{\exp}$ with $\langle \epsilon, h\rangle \neq 0$. This leads to a consistent test because $\operatorname{sp}\mathcal{H}$ is weakly dense. For $p, q \in (1,\infty)$, the weak closure and the norm closure of convex sets are equal (Dunford and Schwartz, 1958, V.2.14, p. 418), so that weak denseness of $\operatorname{sp}\mathcal{H}_{\exp}$ is the same as norm denseness in these spaces. The nonparametric approach works because it is based on a class of functions that comes arbitrarily close to any function; the nuisance approach works because it is based on a class of functions with a span that comes arbitrarily close to any function.

A remarkable feature of Bierens's approach is that a smooth random choice of τ for the function $h_\tau(X) = \exp(X'\tau)$ will deliver a consistent test. To understand how this is possible, we introduce some convenient terminology adapted from the theory of topological vector spaces (Dunford and Schwartz, 1958, V.3.1, p. 418).

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

DEFINITION 2.1. *Let $\mathcal{H} \subset L^q(X)$, $q \in [1,\infty]$. For given nonzero $e \in L^p(X)$, $1/p + 1/q = 1$, $h \in \mathcal{H}$ reveals e if $\langle e, h\rangle \neq 0$. If for every nonzero $e \in L^p(X)$ there exists a revealing $h \in \mathcal{H}$, then $\mathcal{H}$ is totally revealing. Further, if for every nonzero $e \in L^p(X)$, all but a negligible set (defined precisely below) of $h \in \mathcal{H}$ are revealing, then $\mathcal{H}$ is generically totally revealing.*

Note that $\mathcal{H}$ is totally revealing if and only if its span, $\operatorname{sp}\mathcal{H}$, is. We do not distinguish between a class and its span when discussing totality. When the function e or the collection $\mathcal{E}$ is clear from the context, we speak of h as revealing and $\mathcal{H}$ as totally or generically totally revealing. Bierens's key result is as follows.

THEOREM 2.2 (Bierens, 1990, Theorem 1). *Suppose that ϵ is a random scalar with $E|\epsilon| < \infty$ and that X is a bounded random $k \times 1$ vector, $k \in \mathbb{N}$, such that $E(\epsilon|X) \neq 0$. For $\tau \in \mathbb{R}^k$, let $h_\tau(X) = \exp(X'\tau)$. Then there is a subset S of $\mathbb{R}^k$ such that for all $\tau \in S$, $\langle h_\tau(X), \epsilon\rangle \neq 0$. Further, S^c, the complement of S, has Lebesgue measure zero and is not dense in $\mathbb{R}^k$.*

A continuity argument shows that the closure of S^c has empty interior, that is, that it is negligible, so that $\mathcal{H}_{\exp}$ is generically totally revealing. If τ is chosen according to a smooth distribution, then the probability that $\tau \in S^c$ is equal to 0.

Bierens presents Theorem 2.2 as a "fundamental fact" but does not provide much insight into its genesis. However, from the previous discussion we know that $\mathcal{H}$ is revealing if and only if $\operatorname{sp}\mathcal{H}$ is weakly dense in $L^q(X)$ (and strongly dense for $q \in (1,\infty)$). This means that the topological basis of Bierens's approach is a check of whether or not e is in the weak neighborhood of 0 of the form $\{g : |\langle g, h_\tau\rangle| < r\}$ where r is determined by considerations involving the size of the test. Note that this is only implicitly an estimation of e because $\langle \epsilon, h\rangle = \langle e, h\rangle$ for all h.

Another remarkable aspect of Theorem 2.2 is that it makes no use of properties of X other than boundedness. From this we know that $\operatorname{sp}\mathcal{H}_{\exp}$ must be weakly dense in $L^q(X)$ for any (bounded) random vector X, a very strong property. But this just deepens the mystery: for example, what is the role played by the $\exp(\cdot)$ function? Bierens makes fundamental use of its properties in proving his result, but would other functions also work? Recent results for artificial neural networks provide an interesting answer to this question, considerably extending Bierens's result.

THEOREM 2.3. *Let ϵ and X be as in Theorem* 2.2. *Let $\mathcal{H}_G := \{h_\tau | h_\tau(x) = G(\tilde{x}'\tau), \tau \in \mathbb{R}^{k+1}\}$, where $\tilde{x} := (1, x')'$ and G is analytic.*[1] *Then $\mathcal{H}_G$ is generically totally revealing if and only if G is nonpolynomial.*

Theorem 2.3 does not provide insight into how and why Bierens's approach works but only suggests that something deeper is at work than is revealed by Bierens's result or its proof. The results of Section 3 provide the desired insight, but first we complete our discussion of Bierens's approach.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Bierens's result is contained as a special case of Theorem 2.3.[2] Further, there is a strong sense in which the exponential function is not special: the class of nonpolynomial analytic functions is dense in $C(\mathbb{R})$ in the compact-open topology (uniform denseness in $C(B)$ for every compact $B \subset \mathbb{R}$)—a dense collection of functions has the same property as Bierens's family $\mathcal{H}_{\exp}$.

Bierens implements his test with the sample analog of $E(h_\tau(X)\epsilon)$,

$$\hat{M}_n(\tau) := n^{-1} \sum_{i=1}^{n} h_\tau(X_i)\hat{\epsilon}_i, \tag{2}$$

where $\hat{\epsilon}_i := Y_i - f(X_i, \hat{\theta}_n)$, $i = 1,2,\ldots$, and $\{Z_i := (Y_i, X_i')'\}$ is a random sample drawn from the joint distribution of $Z = (Y, X')'$. It follows that for given $\tau \in \mathbb{R}^k$, $n^{1/2}\hat{M}_n(\tau) \Rightarrow N(0, s^2(\tau))$ under H_0 and standard regularity conditions, where

$$s^2(\tau) := \operatorname{var}[(h_\tau(X_i) - b^*(\tau)A^{*-1}\nabla f(X_i, \theta^*))\epsilon_i], \tag{3}$$

$b^*(\tau) := E[h_\tau(X_i)\nabla f(X_i, \theta^*)]$, $A^* := E(\nabla f(X_i, \theta^*)\nabla f(X_i, \theta^*)' - \epsilon\nabla^2 f(X, \theta^*))$, where ∇ and ∇^2 are the gradient and Hessian operators with respect to θ, yielding a $p \times 1$ vector and $p \times p$ matrix, respectively.

Given a consistent estimator $\hat{s}_n^2(\tau)$ for $s^2(\tau)$ and appropriate regularity conditions, it follows that for each τ outside of a negligible subset of $\mathbb{R}^k$, $\hat{W}_n(\tau) := n\hat{M}_n(\tau)^2/\hat{s}_n^2(\tau) \Rightarrow \chi_1^2$ under H_0; in contrast, $\hat{W}_n(\tau)/n \to \eta(h_\tau) > 0$ a.s. under H_a. Thus, a consistent test can be obtained by selecting τ at random, as proposed in Bierens (1987, 1988). However, random selection of the nuisance parameter τ introduces a degree of arbitrariness into both the size and power of the test.

To avoid this difficulty, Bierens (1990) proposed choosing τ to maximize $\hat{W}_n(\tau)$ over a hypercube $T \subset \mathbb{R}^k$. If $\mathcal{H} \subset L^q(X)$ is any norm bounded set with weakly dense span, then $\|g\|_{\mathcal{H}} := \sup\{|\langle g, h\rangle| : h \in \mathcal{H}\}$ defines a norm on $L^p(X)$. This type of norm gives rise to what is called a polar topology (Robertson and Robertson, 1973, III.2), and this is the topological basis of Bierens's second type of test. Again, no direct estimation of e need be made because $\langle \epsilon, h\rangle \equiv \langle e, h\rangle$.

Denote the maximizing value as $\hat{\tau}_n$. The asymptotic distribution of $\hat{W}_n(\hat{\tau}_n)$ is no longer χ_1^2 but, as Bierens showed, is instead a somewhat complex distribution associated with the extremum of the square of a particular Gaussian process. To avoid having to compute this distribution, Bierens proposed the following device: choose $\gamma > 0$, $\rho \in (0,1)$, and $\tau_0 \in T$ independently of the sample, and put

$$\tilde{\tau}_n = \begin{cases} \tau_0 & \text{if } \hat{W}_n(\hat{\tau}_n) - \hat{W}_n(\tau_0) \le \gamma n^\rho, \\ \hat{\tau}_n & \text{if } \hat{W}_n(\hat{\tau}_n) - \hat{W}_n(\tau_0) > \gamma n^\rho. \end{cases} \tag{4}$$

Bierens (1990, Theorem 4) showed that $\hat{W}_n(\tilde{\tau}_n) \Rightarrow \chi_1^2$ under H_0, whereas $\hat{W}_n(\tilde{\tau}_n)/n \to \sup_{\tau \in T}\eta(h_\tau) > 0$ a.s. under H_A. The test is thus consistent, and the power of the test is asymptotically not dependent on the random choice τ_0.

As should be expected, finite-sample results can be highly sensitive to choice of γ, ρ, and τ_0 so that different researchers can arrive at different conclusions

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

about H_0 for the same model and data. Hansen (1996) provided a direct Monte Carlo–based method for computing the distribution of $\hat{W}_n(\hat{\tau}_n)$, avoiding this undesirable property. In Section 5, we discuss a unified approach to hypothesis testing with such statistics and provide simple bounds on $\hat{W}_n(\hat{\tau}_n)$.

By contrast, the nonparametric approach to specification testing typically uses tests based directly on a nonparametric estimator of e, obtained, for example, by solving

$$\min_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \theta(X_i))^2 \tag{5}$$

with

$$\Theta_n = \left\{ \theta : \mathbb{R}^d \to \mathbb{R} \,|\, \theta(x) = \sum_{j=1}^{p_n} \beta_j \psi_j(x), \quad \beta_j \in \mathbb{R}, \quad \text{and} \quad \psi_j : \mathbb{R}^d \to \mathbb{R} \right\}, \tag{6}$$

where p_n is chosen to grow at an appropriate rate with the sample size n and $\{\psi_j\}$ is a sequence of basis functions such as Fourier series, Eubank and Speckman's (1990) polynomial-trigonometric series, Gallant's (1981) flexible Fourier form, as considered by Hong and White (1995), or splines as in Cox and Koh (1989) and Cox, Koh, Wahba, and Yandell (1988). With an indirect approach, e need not be estimated.

3. REVEALING TEST FUNCTIONS AND DUALITY

From the preceding section, we see that totally revealing classes of test functions play a central role in the nuisance parameter approach. In this section, we explore the connections between these classes and duality, providing the topological underpinnings for our unification of the nuisance parameter and the nonparametric approach.

We first consider the properties of totally revealing classes of test functions. For this, we use the spaces $L^p(X)$ of $\sigma(X)$-measurable functions, $p \in [1,\infty]$. By change of variable, we can equivalently use $L^p(\mu) := L^p(\mathbb{R}^k, \mathcal{B}^k, \mu)$ where $\mu(A) := P(X^{-1}(A))$ and $\mathcal{B}^k$ is the Borel σ-field on $\mathbb{R}^k$. For our later discussions of probability and conditional probability models, it is preferable to take μ to be a signed measure.[3] Thus, $L^p(\mu) = L^p(\mathbb{R}^k, \mathcal{B}^k, \mu)$ denotes the set of $\mathcal{B}^k$-measurable, real-valued functions f on $\mathbb{R}^k$ with the property that $\|f\| := [\int_\Omega |f(r)|^p \, d\nu\mu(r)]^{1/p} < \infty$, where $\nu\mu$ is the variation of μ, that is, the unique finite, countably additive positive measure defined by $\nu\mu(A) = \sup \sum_{j=1}^{J} |\mu(A_j)|$, $A \in \mathcal{B}^k$, where the supremum is taken over finite measurable partitions of A. When μ is positive, for example, a probability measure, then $\nu\mu = \mu$.

The space $L^p(\mu)$ is endowed with the metric $d(f_1, f_2) = \|f_1 - f_2\|$, and the corresponding topology is called the norm topology. The weak topology was introduced previously. Polar topologies are also of interest. Given a collection of norm bounded sets $\mathcal{A}$ of $L^q(\mu)$, the $\mathcal{A}$-polar topology on $L^p(\mu)$ is the weakest (or

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

smallest) topology making the functions (seminorms) $\sup\{|\langle g,h\rangle| : h \in A\}$ continuous for each $A \in \mathcal{A}$. When $\mathcal{A}$ is the set of all singleton sets in $L^q(\mu)$, the $\mathcal{A}$-polar topology is the weak topology; when $\mathcal{A}$ consists of only one set and that set is the unit ball in $L^q(\mu)$, the $\mathcal{A}$-polar topology is the norm topology.[4] A topology is separated (or Hausdorff) if points are closed sets. In particular, having the origin be a closed set is crucial to testing, hence the interest in whether or not a topology is separated, or separated at 0, as in Eubank and Hart (1992).

THEOREM 3.1. *Suppose that* $p \in [1,\infty]$ *and that* q *satisfies* $1/p + 1/q = 1$. *Let* $\mathcal{H}$ *be a norm bounded subset of* $L^q(\mu)$. *The following statements are equivalent*:

(a) $\mathcal{H}$ is totally revealing.
(b) $\operatorname{sp}\mathcal{H}$ *is weakly dense in* $L^q(\mu)$.
(c) *The* $\{\mathcal{H}\}$*-polar topology is separated.*

If in addition $p \in (1,\infty)$, *then the following statement is equivalent to* (a)–(c):

(d) $\operatorname{sp}\mathcal{H}$ *is norm dense in* $L^q(\mu)$.

Because μ is often unknown, it is important to ensure that $\operatorname{sp}\mathcal{H}$ is weakly dense in $L^p(\mu)$ for *any* μ. For this, we assume (throughout) that there is a compact set B such that $\nu\mu(B) = \nu\mu(\mathbb{R}^k)$. In the regression context, this is the assumption that the explanatory variables, X, must be bounded; in the probability model context, this is the assumption that the distributions have bounded support. This is without loss of generality in the following sense: we can always homeomorphically embed $\mathbb{R}^k$ in $(-1,+1)^k$. This retains all of the conditioning information in the regression context and all of the differences in distributions in the probability model context.

We let $M_b(B)$ denote the set of bounded measurable functions on B. Let $C(B) \subset M_b(B)$ be the set of continuous (hence bounded) functions on B. Endow $M_b(B)$ and its subsets with the uniform topology. Uniform denseness and uniform closure refer to denseness and closure in this topology. We now strengthen the concept of totality.

DEFINITION 3.2. *We say that* $\mathcal{H} \subset M_b(B)$ *is comprehensively revealing if it is totally revealing for* $L^p(\mu)$ *for every* $p \in [1,\infty]$ *and every finite signed measured* μ *supported on* B.

A key result regarding comprehensiveness is the following theorem.

THEOREM 3.3.

(a) *For* $\mathcal{H} \subset C(B)$, $\mathcal{H}$ *is comprehensively revealing if and only if* $\operatorname{sp}\mathcal{H}$ *is uniformly dense in* $C(B)$.
(b) *For* $\mathcal{H} \subset M_b(B)$, $\mathcal{H}$ *is comprehensively revealing given any of the following conditions*:
    (i) $\operatorname{sp}\mathcal{H}$ *contains a subalgebra of* $C(B)$ *containing the constants and separating points*;

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

(ii) sp $\mathcal{H}$ *contains a subalgebra* $\mathcal{A}$ *of* $M_b(B)$ *containing the constants, and the minimal* σ*-field making all functions in* $\mathcal{A}$ *measurable is* $\mathcal{B}^k$;
(iii) *The uniform closure of* sp $\mathcal{H}$ *contains a comprehensively revealing set.*

Theorem 3.3 permits many choices for comprehensively revealing $\mathcal{H}$. For example, $\mathcal{H}$ can be any basis for the algebraic polynomials, Bernstein polynomials, Chebyshev polynomials, trigonometric polynomials, Fourier series, B-splines, etc. Such choices form the basis for the nonparametric approach to specification testing. Theorem 3.3 thus provides the desired unification. Further, we can extend Bierens's approach by considering classes $\mathcal{H}$ of the form

$$\mathcal{H}_G = \{h : \mathbb{R}^k \to \mathbb{R} \,|\, h(x) = G[A(x)], A \text{ affine}\} \quad \text{where } G : \mathbb{R} \to \mathbb{R}, \tag{7}$$

as $\mathcal{H}_G$ preserves the appealing simplicity of Bierens's class $\mathcal{H}_{\exp}$. We say that G is totally revealing, comprehensively revealing, etc., whenever $\mathcal{H}_G$ is.

To contrast the requirements for a comprehensively revealing class with those given subsequently for a generically comprehensively revealing class, we introduce the following linear spaces of functions:

$$\Sigma(G,T) := \left\{ g : B \to \mathbb{R} \,|\, g(x) = \beta_0 + \sum_{j=1}^{r} \beta_j G(\tilde{x}'\tau_j),\ \beta_0, \beta_j \in \mathbb{R},\right.$$
$$\left. \tau_j \in T \subset \mathbb{R}^{k+1},\ j = 1, \ldots, r,\ r \in \mathbb{N} \right\}, \tag{8}$$

where B is compact and supports $\nu\mu$. The Σ in $\Sigma(G,T)$ suggests the span; the weights τ_j of the affine combination $A_j(x) = \tilde{x}'\tau_j$ are restricted to $T \subset \mathbb{R}^{k+1}$. When $T = \mathbb{R}^{k+1}$, $\Sigma(G,\mathbb{R}^{k+1}) = \text{sp}\{\mathcal{H}_G, \mathbf{1}\}$ where $\mathbf{1}$ is the function always equal to 1. Provided $G \neq 0$, this equals sp $\mathcal{H}_G$ because we can set $\tau_1 = (b, \mathbf{0}')'$ where $b \in \mathbb{R}$ satisfies $G(b) \neq 0$. Thus $\mathcal{H}_G$ is comprehensively revealing if and only if $\Sigma(G,\mathbb{R}^{k+1})$ is comprehensively revealing.

The question of when $\Sigma(G,\mathbb{R}^{k+1})$ is comprehensively revealing has been intensely investigated in the artificial neural network literature. This interest arises because the functions in $\Sigma(G,\mathbb{R}^{k+1}) = \text{sp}\,\mathcal{H}_G$ are the "output functions" of a leading class of artificial neural networks, the "single hidden layer feedforward networks" (e.g., Rumelhart, Hinton, and Williams, 1986). It suffices here to find conditions on G such that $\Sigma(G,\mathbb{R}^{k+1})$ is uniformly dense in $C(B)$ for any compact B.

DEFINITION 3.4. *A measurable function* $g : \mathbb{R} \to \mathbb{R}$ *has a nice interval if for some* $a < b$, g *is Riemann integrable in* $[a,b]$ *and nonpolynomial. If* g *is also continuous on* $[a,b]$, *then* g *has a very nice interval.*

In reading the first part of the following result, bear in mind that $G \in \Sigma(G,\mathbb{R})$ and that if the domains and ranges match, the composition of an affine function with an affine function is affine.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

LEMMA 3.5. *Suppose* $\Sigma(G,\mathbb{R})$ *contains a function* g *with a nice interval. Then* $\mathcal{H}_G$ *is comprehensively revealing. If* G *is continuous,* $\mathcal{H}_G$ *is comprehensively revealing if and only if* G *has a very nice interval.*

Thus, the only continuous choices for G that are *not* comprehensively revealing are the polynomials. Bierens's (1990) choice $G(a) = \exp(a)$ and White's (1989b) choice $G =$ logistic cumulative distribution function (c.d.f.) are clearly continuous and clearly not polynomials, whereas Hansen's (1990) choice $G(a) = 1_{(0,\infty)}(a)$ has a nice interval.

We now turn to issues of genericity. For T a nonempty subset of $\mathbb{R}^{k+1}$, let $\mathcal{H}_G(T)$ denote the set of functions of the form $x \to G(\tilde{x}'\tau)$ with $\tau \in T$.

DEFINITION 3.6. *We say that* $\mathcal{H}_G$ *is generically comprehensively revealing if for all* T *with nonempty interior, the uniform closure of* $\operatorname{sp}\mathcal{H}_G(T)$ *contains* $C(B)$ *for* ***every*** *compact* B.

It is straightforward to show that this implies that the set $\mathcal{S}^c_{\mathcal{H}_G}$ as defined in Section 2 is negligible regardless of the (signed) measure μ underlying the expectation. The difference between $\operatorname{sp}\mathcal{H}_G(T)$ and $\Sigma(G,T)$ is that $\operatorname{sp}\mathcal{H}_G(T)$ might not contain the constant functions. We saw that this difference could not arise when $T = \mathbb{R}^{k+1}$; the same result holds here.

LEMMA 3.7. *The class* $\mathcal{H}_G$ *is generically comprehensively revealing if and only if for every* T *with nonempty interior,* $\Sigma(G,T)$ *is uniformly dense in* $C(B)$ *for every compact* B.

Thus, for given G, the difference between $\mathcal{H}_G$ being comprehensively revealing and $\mathcal{H}_G$ being generically comprehensively revealing hinges on whether only $\mathbb{R}^{k+1}$ or alternatively an arbitrarily chosen (small) set T in $\mathbb{R}^{k+1}$ with nonempty interior can deliver the uniform denseness of $\Sigma(G,T)$. This seems a rather strong condition, having a "universe in a grain of sand" flavor. Nevertheless, Bierens proves that $G = \exp$ has this property, but it can be shown, for example, that Hansen's $G(a) = 1_{(0,\infty)}(a)$ does not. The real analytic functions have perhaps surprising properties in this regard.

THEOREM 3.8. *Let* G *be real analytic. Then* $\mathcal{H}_G$ *is comprehensively revealing if and only if it is generically comprehensively revealing.*

In view of Lemma 3.5, we have the following.

COROLLARY 3.9. *Let* G *be real analytic. Then* $\mathcal{H}_G$ *is generically comprehensively revealing* (*hence comprehensively revealing*) *if and only if* G *is not a polynomial.*

This covers Bierens's choice $G = \exp$ and White's choice $G =$ logistic c.d.f. Theorem 2.3 follows immediately from this result. It is, however, not necessary that G be real analytic.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

THEOREM 3.10. *Suppose that* $\mathrm{sp}\{D^\alpha G(a),\ 0 \le \alpha < \infty | a \in O\}$ *is dense in* $C(\mathbb{R})$ *for any nonempty open subset* O *of* $\mathbb{R}$ *for* G *infinitely differentiable. Then for any* $T \subset \mathbb{R}^{k+1}$ *with nonempty interior,* $\Sigma(G,T)$ *is uniformly dense in* $C(B)$ *for any compact* B*, so* $\mathcal{H}_G$ *is generically comprehensively revealing.*

A nonanalytic function satisfying this condition is the normal c.d.f. or density.

Thus, there is a large variety of choices besides $G = \exp$ that share the appealing features of Bierens's consistent specification testing approach. Indeed, Corollary 3.9 implies that the property that G is generically comprehensively revealing is itself "generic," that is, it is a property possessed by a dense set of functions G in $C(\mathbb{R})$. Let $\mathcal{G} \subset C(\mathbb{R})$ denote the class of continuous functions G such that $\Sigma(G,T)$ is dense in $C(\mathbb{R}^k)$ in the compact-open topology (uniform denseness in $C(B)$ for every compact B) for every set T having nonempty interior.

THEOREM 3.11. *The class* $\mathcal{G}$ *is dense in* $C(\mathbb{R})$ *in the compact-open topology.*

Note that the complement of $\mathcal{G}$ contains the polynomials and so is also dense.

4. IMPLICATIONS AND APPLICATIONS

We next consider the scope of the foregoing theory and show that it extends well beyond the standard regression framework and into probability and conditional probability models. Eubank and LaRiccia (1992) compare Cramér–von Mises and nonparametric tests for the equality of distributions. The Cramér–von Mises and Kolmogorov–Smirnov tests are examples of the uses of polar or other norm topologies to measure the distance between an estimated and a null hypothesis distribution. Eubank and LaRiccia cite a large body of literature on the problem of distinguishing between different one-dimensional distributions. Because the nuisance parameter approach does not need to directly estimate densities for comparison purposes, it is more easily applicable to multidimensional problems.

We begin with a study of likelihood models of a scalar random variable Y conditional on a random k-vector X distributed according to P_X.[5] Let the conditional likelihood function with respect to a σ-finite measure ν and parameterized by $\theta \in \Theta$, Θ an open subset of $\mathbb{R}^p$, be given by a measurable function, $f\colon \mathbb{R} \times \mathbb{R}^k \times \Theta \to \mathbb{R}^+$, with the property that for all $x \in \mathbb{R}^k$ and $\theta \in \Theta$, $\int_{\mathbb{R}} f(y|x,\theta)\, d\nu(y) = 1$.

Let $\mathcal{S}$ denote $\{ f(\cdot|\cdot,\theta) | \theta \in \Theta\}$. $\mathcal{S}$ is correctly specified for Y conditional on X when for P_X-almost all x, $f(\cdot|x,\theta_0)$ is a version of the true conditional density of Y given $X = x$ with respect to ν for some θ_0 in the interior of Θ. Under mild conditions on f, we can differentiate both sides of the identity $\int_{\mathbb{R}} f(y|x,\theta)\, d\nu(y) \equiv 1$ with respect to θ. Under correct specification this yields (e.g., White, 1994, Theorem 6.6) $E(s^*|X) = 0$ a.s. where $s^* = (s_1^*,\dots,s_p^*)' \equiv \nabla_\theta \log f(Y|X,\theta^*)$ is the $p \times 1$ "score" vector of the likelihood, with $\theta^* := \mathrm{argmax}_{\theta\in\Theta} E(\log f(Y|X,\theta))$. (Here $\theta^* = \theta_0$ under correct specification.) With $\xi = E(s^*|X)$, Theorem 3.8 ensures that for G in a dense set of choices and essentially every τ in $\mathbb{R}^{k+1}$

$$E(G(\widetilde{X}'\tau)s_j^*) \neq 0, \qquad j = 1,\dots,p, \tag{9}$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

for any misspecification of $\mathcal{S}$ leading to the failure of $E(s^*|X) = 0$ a.s. This forms the basis for statistically detecting any such failure as noted by Hansen (1990). Nevertheless, one may have $E(s^*|X) = 0$ in the presence of misspecification of $\mathcal{S}$, for example, one that leads to violation of the equation

$$E(\nabla' s^* + s^* s^{*\prime}|X) = 0 \quad \text{a.s.,} \quad \text{where } \nabla' s^* := \nabla^2 \log f(Y|X,\theta^*). \tag{10}$$

(To get this under correct specification, differentiate $\int_{\mathbb{R}} f(y|x,\theta)\, d\nu(y) \equiv 1$ twice w.r.t. θ.)

The question then arises whether the theory of Section 3 provides a way to detect arbitrary misspecification in $\mathcal{S}$.[6] To show it can, we use the fact that μ may be signed and consider testing whether two multivariate distributions, say P and Q, are the same. In the likelihood context, we can view P as the true distribution of $Z = (Y, X')'$ and Q as that implied by the likelihood at θ^*,

$$Q(A) = \int_A f(y|x,\theta^*)\, d\nu(y) dP_X(x), \qquad A \text{ a Borel subset of } \mathbb{R} \times \mathbb{R}^k. \tag{11}$$

Because P and Q each completely specify a joint distribution for Y and X, a test that will detect any deviation of P from Q is a test that will detect *any* misspecification in $\mathcal{S}$. Formally, we test

$$H_0 : \rho(P,Q) = 0 \quad \text{vs.} \quad H_A : \rho(P,Q) > 0, \tag{12}$$

where ρ is any metric on the space $\mathcal{M}$ of finite measures inducing the weak star topology on (variation) norm bounded subsets of $\mathcal{M}$, for example,

$$\rho(\mu_1,\mu_2) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{\left| \int f^n\, d\mu_1 - \int f^n\, d\mu_2 \right|}{1 + \left| \int f^n\, d\mu_1 - \int f^n\, d\mu_2 \right|}, \tag{13}$$

where $\{f^n\}_{n\in\mathbb{N}}$ is any uniformly dense subset of $C(B)$ (Dunford and Schwartz, 1958, p. 426).

Because ρ is a metric, H_0 is the hypothesis that $\mu := P - Q$ is the zero measure, that is, $\rho(\mu,0) = 0$. Let E_P and E_Q denote expectation with respect to the indicated measures. For any $h \in \text{sp}\,\mathcal{H}$,

$$E_P(h(X)) - E_Q(h(X)) = \int h(x)\, dP(x) - \int h(x)\, dQ(x) = \int \mathbf{1} h(x)\, d\mu(x). \tag{14}$$

We now set $\xi := \mathbf{1}$ and test whether $\mathbf{1} = 0$ a.e. $-\,\mu$. Now, $\mathbf{1} = 0$ a.e. $-\,\mu$ if and only if μ is the zero measure, that is, $P = Q$. Totality of sp $\mathcal{H}$ implies that $\mathbf{1} = 0$ a.e. $-\,\mu$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

if and only if $\int h(x)\, d\mu(x) = 0$ for all h in sp $\mathcal{H}$. Theorem 3.8 implies that under H_A for appropriate G and almost all $\tau \in \mathbb{R}^{k+2}$, $\tau = (\tau_1, \tau_2')'$,

$$E_P(G(\tilde{Z}'\tau)) - E_Q(G(\tilde{Z}'\tau)) \neq 0, \tag{15}$$

where $\tilde{Z} = (Y, \tilde{X}')'$. A statistic that estimates this difference is

$$n^{-1}\sum_{i=1}^{n} G(\tilde{Z}_i'\tau) - n^{-1}\sum_{i=1}^{n}\int G(y \times \tau_1 + \tilde{X}_i'\tau_2) f(y|X_i, \hat{\theta}_n)\, d\nu(y), \tag{16}$$

where $\hat{\theta}_n := \text{argmax}_{\theta\in\Theta}\, n^{-1}\sum_{i=1}^{n} \log f(Y_i|X_i, \theta)$. A test can be based on this difference, as described in the next section. We now see clearly the utility of treating signed measures, as this reasoning requires that μ be a signed measure.

In addition to testing the correctness of likelihood models, testing $\rho(P, Q) = 0$ against $\rho(P, Q) > 0$ permits testing whether the (unknown) joint distribution Q of a random vector X coincides with a specified joint distribution P—for example, that X is multivariate uniform—as well as testing that two independent random samples are both drawn from the same unknown distribution. In both cases, the integral $\int G(\tilde{x}'\tau)(dP(x) - dQ(x))$ can be approximated by the statistic

$$n_1^{-1}\sum_{i=1}^{n_1} G(\tilde{X}_{1i}'\tau) - n_2^{-1}\sum_{i=1}^{n_2} G(\tilde{X}_{2i}'\tau). \tag{17}$$

In the latter case, $\{X_{1i}, i = 1, \ldots, n_1\}$ and $\{X_{2i}, i = 1, \ldots, n_2\}$ are independent samples from the two populations. In the former, $\{X_{1i}, i = 1, \ldots, n_1\}$ is generated by nature according to Q, and $\{X_{2i}, i = 1, \ldots, n_2\}$ is generated pseudorandomly by the researcher according to P. (Alternatively, a numerical integration may be used.) When X is a scalar and G is chosen to be the indicator function $G(a) = 1[a > 0]$, using $G(\tilde{x}'\tau) = 1[x > \tau_0]$ with $\tau := (\tau_0, 1)$ leads to tests closely related to the familiar Kolmogorov–Smirnov test (e.g., Serfling, 1980, pp. 57–58) when the uniform norm is applied, as described in the next section; choosing the L^2-norm instead leads to tests related to the Cramér–von Mises statistic (e.g., Serfling, 1980, p. 58). Recent work of Andrews (1997) extends the Kolmogorov approach to testing parametric models with conditioning variables.

Applications of the sort discussed here approximate integrals $\int \phi(x, \tau)\, d\mu(x)$ for finite signed measures μ with statistics $\int \phi(x, \tau)\, d\mu_n(x)$, where μ_n converges to μ in the weak star topology. By definition, continuity and boundedness of each $\phi(\cdot, \tau)$ guarantees that $\int \phi(x, \tau)\, d\mu_n(x) \to \int \phi(x, \tau)\, d\mu(x)$ for each $\tau \in T$. The next result shows that if the class $\{\phi(\cdot, \tau) : \tau \in T\}$ (e.g., $\phi(x, \tau) = G(\tilde{x}'\tau)$) is comprehensively revealing, then μ_n converging to μ is equivalent to $\int \phi(x, \tau)\, d\mu_n(x) \to \int \phi(x, \tau)\, d\mu(x)$ for each $\tau \in T$. Further, if the class $\{\phi(\cdot, \tau) : \tau \in T\}$ has compact closure (in the sup norm), the convergence will be uniform over $\tau \in T$.

This uniformity is useful in two areas. It provides a uniform law of large numbers, and it guarantees that small deviations from correct specification give rise only to small values of our test statistics.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

THEOREM 4.1. *Let $\{\mu_n\}_{n=0}^{\infty}$ be a sequence of finite signed measures on B with $v\mu_n(B)$ uniformly bounded.*

(a) *If the class $\{\phi(\cdot,\tau):\tau \in T\} \subset C(B)$ is comprehensively revealing, then $\mu_n \to \mu_0$ in the weak star topology if and only if for every $\tau \in T$, $\int \phi(x,\tau)\, d\mu_n(x) \to \int \phi(x,\tau)\, d\mu_0(x)$.*

(b) *If $\{\phi(\cdot,\tau):\tau \in T\} \subset C(B)$ is comprehensively revealing and has compact closure, then $\mu_n \to \mu_0$ if and only if $\sup_{\tau\in T} |\int \phi(x,\tau)\, d\mu_n(x) - \int \phi(x,\tau)\, d\mu_0(x)| \to 0$.*

5. HYPOTHESIS TESTING WHEN A NUISANCE PARAMETER IS PRESENT ONLY UNDER THE ALTERNATIVE

Our preceding results establish an equivalence between the hypothesis of correct model specification and a family of moment conditions.[7] For appropriate moment function m, we write the null hypothesis as

$$H_0: E(m(Z,\theta_0,\tau)) = 0 \quad \text{for some } \theta_0 \in \Theta \quad \text{and all } \tau \in T, \tag{18}$$

where T is an appropriate compact set and Z is an $l \times 1$ random vector. For example, to test a nonlinear regression model, take m to be given by $m(Z,\theta,\tau) = G(\tilde{X}'\tau)[Y - f(X,\theta)]$. The alternative to H_0 is

$$H_A: E(m(Z,\theta,\tau_0)) \neq 0 \quad \text{for all } \theta \in \Theta \quad \text{and some } \tau_0 \in T. \tag{19}$$

Nevertheless, we have seen that for certain m the alternative is in fact

$$H_A^*: E(m(Z,\theta,\tau)) \neq 0 \quad \text{for all } \theta \in \Theta \quad \text{and essentially all } \tau \in T. \tag{20}$$

Now, τ is indeterminate under the null but not under H_A. Thus, τ is called a "nuisance parameter present only under the alternative." The phrase "identified only under the alternative" is also used. "Identification" here is not fully analogous to the usual concept arising in estimation of parametric models, because of H_A^*. Nevertheless, the terminology is now standard.

Hypothesis testing in such contexts presents challenges, as evidenced by our brief discussion of Bierens's (1990) approach in Section 2. The problem has been addressed by several authors: Davies (1977, 1987) gave bounds for certain statistics; Andrews (1993) and Hansen (1996) discussed a variety of examples in econometrics; Bierens (1990) and Hansen (1996) obtained the asymptotic distribution for $\max_{\tau\in T} \hat{W}_n(\tau)$; and Andrews and Ploberger (1994) proposed an optimality criterion and an optimal test. In this section we present a natural approach to solving this problem using Banach spaces and illustrate its application to testing for misspecification of a nonlinear regression model.

The key to our approach is to cast H_0 as a hypothesis about the expectation of a Banach space–valued random variable (a "Banach random element" or "Banach-r.e."). Recall that a Banach space, $\mathbb{B}$, is a complete normed linear space. We denote the norm $\|\cdot\|$. A Banach-r.e. is a measurable map $\mathcal{X}$ from a probability space $(\Omega, \mathcal{F}, P)$ into $\mathbb{B}$, equipped with the (Borel) σ-field generated by the open

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

sets of the norm topology, with the tightness property that for each $\epsilon > 0$ there exists a compact set K_ϵ in $\mathbb{B}$ such that $P[\mathcal{X} \subset K_\epsilon] \geq 1 - \epsilon$.

Because $\mathbb{B}$ is linear, the integral of a simple Banach-r.e., $\mathcal{X}^s$, is $E(\mathcal{X}^s) := \sum x\, P[\mathcal{X}^s = x] \in \mathbb{B}$, summing over the finitely many values taken by $\mathcal{X}^s$. Because by tightness the range of a general $\mathcal{X}$ is separable, it is the P-a.e. limit of some simple sequence $\{\mathcal{X}_j^s\}_{j\in\mathbb{N}}$ with $\|\mathcal{X}_j^s(\omega)\| \leq \|\mathcal{X}(\omega)\|$. If there is a unique $\|\cdot\|$-limit of $\{E\mathcal{X}_j^s\}$ for all such sequences, then the integral of $\mathcal{X}$, denoted $E\mathcal{X}$, is defined as this limit.

As previously discussed, $\hat{\theta}_n$ is consistent for θ^* in Θ, and $\theta^* = \theta_0$ under H_0. Denote by $\mathcal{M}^* = m(Z,\theta^*,\cdot)$ the mapping $\tau \to m(Z,\theta^*,\tau)$. As Z is random, $\mathcal{M}^*$ is a random function of τ, and under suitable conditions $\mathcal{M}^*$ is a Banach-r.e. We can then express H_0 as

$$H_0^{\mathbb{B}}: g(E(\mathcal{M}^*)) = 0, \tag{21}$$

where $g: \mathbb{B} \to \mathbb{R}^+$ is $\|\cdot\|$-continuous on $\mathbb{B}$ such that $g(x) = 0$ if and only if $\|x\| = 0$. We choose g, $\|\cdot\|$, and $\hat{\theta}_n$ jointly so that a convenient estimator of $g(E(\mathcal{M}^*))$ will have a tractable distribution under the null (with proper scaling) and have power under the alternative. For example, in the classical situation (with $\mathbb{B} = \mathbb{R}^p$) a leading case is $H_0: \theta^* = 0$. Let $\|\theta\| = [\theta'\theta]^{1/2}$, let $\hat{\theta}_n$ be the maximum likelihood estimator for θ^*, and let $g(\theta) = \theta' I^* \theta$, where I^* is the MLE information matrix. When $\sqrt{n}\hat{\theta}_n$ replaces θ^*, we obtain a statistic $g(\sqrt{n}\hat{\theta}_n) = n\hat{\theta}_n' I^* \hat{\theta}_n$ that has the convenient χ_p^2 distribution under the null asymptotically and has optimal asymptotic power properties under local alternatives.

A variety of useful norms is available for the present case. Often $m(Z,\theta^*,\cdot)$ is continuous on T, and so we are dealing with random elements of $C(T)$. This space can be endowed with the uniform norm $\|\cdot\|_\infty$, so that $\|E(\mathcal{M}^*)\|_\infty = \sup_{\tau\in T} |E(m(Z,\theta^*,\tau))|$. So long as the smallest weakly closed set supporting ν has weakly dense span, this procedure yields a separated topology. Taking $g(x) := \|x\|_\infty$ satisfies the conditions required of g and leads to statistics of the form

$$g\left(n^{-1/2}\sum_{i=1}^n \mathcal{M}_i^*\right) = \sup_{\tau\in T}\left|n^{-1/2}\sum_{i=1}^n m(Z_i,\theta^*,\tau)\right|, \tag{22}$$

where Z_i is a random sample on Z. These are the statistics considered by Bierens (1990), Hansen (1996), and Kolmogorov–Smirnov.

Alternatively, when $m(\cdot,\theta^*,\cdot)$ is properly integrable we can consider $L^p(\nu)$ norms of the form

$$\|E(\mathcal{M}^*)\|_{p,\nu} = \left[\int_T |E[m(Z,\theta^*,\tau)]|^p \, d\nu(\tau)\right]^{1/p}, \qquad 1 \leq p < \infty, \tag{23}$$

where ν is a given measure on T. Taking $g(x) = \|x\|_{p,\nu}^p$ leads to statistics of the form

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

$$g\left(n^{-1/2}\sum_{t=1}^{n}\mathcal{M}_i^*\right) = \int_T \left|n^{-1/2}\sum_{t=1}^{n} m(Z_i,\theta^*,\tau)\right|^p d\nu(\tau). \tag{24}$$

Bierens (1982) proposed a statistic of this form with $p = 2$ and $d\nu(\tau) = d\tau$. Hansen (1996) also examined a version of this statistic. For Cramér–von Mises type tests, take $p = 2$ and ν to be the distribution of Z. Andrews and Ploberger's (1994) optimal test arises with g of the form

$$g(x) = (1+c)^{-(l+1)} \int \exp[x(\tau)^2 c/2(1+c)]\, d\nu(\tau), \tag{25}$$

where $c > 0$ is a scalar constant determining whether the test has power against near or far local alternatives and ν is a given probability measure supported on T. For $x \in C(T)$ this choice is continuous with respect to $\|\cdot\|_\infty$.

Thus, we are led to consider the asymptotic behavior of statistics of the form $g(n^{-1/2}\mathcal{S}_n)$, where $\mathcal{S}_n = \sum_{i=1}^n \mathcal{M}_i^*$ is a sum of Banach-r.e.'s and the normalization by $n^{-1/2}$ stabilizes the distribution of the sum. The central limit theorem (CLT) and law of the iterated logarithm (LIL) for Banach-valued random sums provide the desired description of the asymptotic behavior of $n^{-1/2}\mathcal{S}_n$. Available results of Ossiander (1987) and Ledoux and Talagrand (1991) provide convenient sufficient conditions.

We therefore seek to test $g(E(\mathcal{M}^*)) = 0$ based on $g(n^{-1/2}\mathcal{S}_n)$, using the CLT and LIL for Banach-r.e.'s. For simplicity, we treat only the independent and identically distributed (i.i.d.) case. Generalizations may be taken up elsewhere. We first make the i.i.d. assumption formal.

Assumption 1. The term Z_i is a sequence of i.i.d. random variables on the probability space $(\Omega, \mathcal{F}, P)$ taking values in $\mathbb{R}^l$ and having the distribution of the random $l \times 1$ vector Z.

To permit $\hat{\theta}_n$ to be an m-estimator (Huber, 1967), we impose

Assumption 2. For each $n \in \mathbb{N}$, $\hat{\theta}_n : \Omega \to \Theta$ (with Θ a compact subset of $\mathbb{R}^p$) is measurable. Further, there exists a function $s : \mathbb{R}^l \times \Theta \to \mathbb{R}^p$ and a finite, nonsingular, nonstochastic $p \times p$ matrix A^* such that s is measurable on $\mathbb{R}^l$ for each θ in Θ and continuous in Θ for each z in $\mathbb{R}^l$;

$$n^{-1/2}\sum_{i=1}^{n} s(Z_i,\hat{\theta}_n) = o_P(1); \tag{26}$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = -A^{*-1}n^{-1/2}\sum_{i=1}^{n} s(Z_i,\theta^*) + o_P(1), \tag{27}$$

where θ^* is in the interior of Θ and $E(s(Z,\theta^*)) = 0$.

We next impose conditions on m sufficient to permit a Taylor series expansion around θ^*.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Assumption 3.

(a) Let T be a compact subset of $\mathbb{R}^q$, $q \in \mathbb{N}$. The function $m: \mathbb{R}^l \times \Theta \times T \to \mathbb{R}$ is measurable in $\mathbb{R}^l$ for each $(\theta, \tau) \in \Theta \times T$ and continuous on $\Theta \times T$ for each z in $\mathbb{R}^l$. Further, for each $(z, \tau) \in \mathbb{R}^l \times T$, $m(z, \cdot, \tau)$ is continuously differentiable on an open neighborhood of θ^*.
(b) The function $\sup_{(\theta,\tau)\in\Theta\times T} |\nabla m(Z, \theta, \tau)|$ is integrable.

For simplicity, we have m mapping into $\mathbb{R}$. Generalization to vector-valued m is straightforward. We write $M(\theta^*, \tau) := E(\nabla' m(Z, \theta^*, \tau))$. The present assumptions suffice for the following lemma.

LEMMA 5.1. *Suppose Assumptions* 1–3 *hold. Then*

$$\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^* \right\|_\infty = o_P(1), \tag{28}$$

where $\hat{\mathcal{M}}_i := m(Z_i, \hat{\theta}_n, \cdot)$ *and* $\mathcal{L}_i^* := m(Z_i, \theta^*, \cdot) - M(\theta^*, \cdot) A^{*-1} s(Z_i, \theta^*)$.

Consequently, the probabilistic behavior of $\|n^{-1/2} \sum_{i=1}^n \hat{\mathcal{M}}_i\|$ can be approximated asymptotically by that of $\|n^{-1/2} \sum_{i=1}^n \mathcal{L}_i^*\|$ for any norm weaker than $\|\cdot\|_\infty$, for example, $\|\cdot\|_{p,\nu}$, $1 \le p < \infty$, ν finite.

Inspecting $\mathcal{L}_i^*$, we see that its dispersion may vary over T. We standardize $\mathcal{L}_i^*$ using its standard deviation, say $\sigma_*: T \to \mathbb{R}_+$. The existence of σ_* is ensured by

Assumption 4.

(a) $\sup_{(\theta,\tau)\in\Theta\times T} m(Z, \theta, \tau)^2$ is integrable; and
(b) $\sup_{\theta\in\Theta} s(Z, \theta)' s(Z, \theta)$ is integrable.

This also ensures that $\tau \mapsto \sigma_*^2(\tau) := \mathrm{var}[m(Z, \theta^*, \tau) + M(\theta^*, \tau) A^{*-1} s(Z, \theta^*)]$ is continuous on T. To avoid division by zero, we impose

Assumption 5. T is chosen such that $\inf_{\tau\in T} \sigma_*^2(\tau) > 0$, that is, $\|\sigma_*^{-2}\|_\infty < \infty$.

From Lemma 5.1, the probabilistic behavior of $\|n^{-1/2} \sum_{i=1}^n \mathcal{L}_i^*/\sigma_*\|$ approximates that of $\|n^{-1/2} \sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_*\|$. Because σ_* is unknown, we replace it with a consistent estimator. We consider two specific estimators. The first is an uncentered estimator $\tilde{\sigma}_n^2$, given by

$$\tilde{\sigma}_n^2(\tau) = n^{-1} \sum_{i=1}^{n} [m(Z_i, \hat{\theta}_n, \tau) - \hat{M}_n(\hat{\theta}_n, \tau) \hat{A}_n^{-1} s(Z_i, \hat{\theta}_n)]^2, \tag{29}$$

where $\hat{M}_n(\hat{\theta}_n, \tau) := n^{-1} \sum_{i=1}^n \nabla' m(Z_i, \hat{\theta}_n, \tau)$. A centered estimator is $\hat{\sigma}_n^2$, given by

$$\hat{\sigma}_n^2(\tau) = \tilde{\sigma}_n^2(\tau) - \left(n^{-1} \sum_{i=1}^{n} m(Z_i, \hat{\theta}_n, \tau) \right)^2. \tag{30}$$

To ensure the consistency of either $\tilde{\sigma}_n^2$ or $\hat{\sigma}_n^2$ for σ_*^2 we impose the following assumption.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Assumption 6. $\hat{A}_n = A^* + o_{\text{a.s.}}(1)$.

The consistency of $\tilde{\sigma}_n^2$ for σ_*^2 holds under H_0. Under the alternative, $\tilde{\sigma}_n^2$ converges to a function not less than σ_*^2. This is all that is necessary. The estimator $\hat{\sigma}_n^2$ is consistent under both H_0 and H_A.

It remains to impose conditions that permit application of the Banach CLT and LIL. We use conditions of Ossiander (1987) that are reasonably broad and not difficult to verify.

First we define what it means for $n^{-1/2}S_n$ to obey the Banach CLT. The following definition is standard.

DEFINITION 5.2. *Let $\{\mathcal{Z}_n\}$ be a sequence of Banach-r.e.'s with corresponding distributions $\{\mu_n\}$. Then $\mathcal{Z}_n$ converges in distribution on $\mathbb{B}$ to $\mathcal{Z}$, written $\mathcal{Z}_n \Rightarrow_{\mathbb{B}} \mathcal{Z}$, if for every bounded continuous function $f: \mathbb{B} \to \mathbb{R}$, $\int f\,d\mu_n \to \int f\,d\mu$ as $n \to \infty$, where $\mathcal{Z}$ has distribution μ.*

We follow Ledoux and Talagrand in saying that $n^{-1/2}S_n$ obeys the Banach CLT if for some Banach-r.e. $\mathcal{Z}$, $n^{-1/2}S_n \Rightarrow_{\mathbb{B}} \mathcal{Z}$. We leave the distribution of $\mathcal{Z}$ unspecified.

For simplicity in stating our final assumption, define $\phi: \mathbb{R}^l \times T \to \mathbb{R}$ as

$$\phi(z,\tau) := [m(z,\theta^*,\tau) - M(\theta^*,\tau)A^{*-1}s(z,\theta^*)]/\sigma_*(\tau). \tag{31}$$

From Lemma 2.1 and Theorem 3.1 of Ossiander (1987), the Banach CLT holds for $n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_*$ under the following condition on ϕ, requiring essentially that ϕ not be too irregular locally.

Assumption 7. There exists a continuous strictly increasing function $\gamma: \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$E^*\left[\sup_{\tau \in N_\delta(\tau_0)} |\phi(Z,\tau) - \phi(Z,\tau_0)|^2\right] \leq \gamma(\delta)^2 \quad \text{for each } \delta \in \mathbb{R}_{++} \tag{32}$$

for each $\tau_0 \in T(\delta)$, where $T(\delta)$ is a δ-net for T and $\{N_\delta(\tau_0): \tau \in T(\delta)\}$ is a collection of δ neighborhoods of the members of $T(\delta)$ that covers T, and where E^* denotes outer expectation. Further, γ satisfies the "metric entropy" condition, that is, for some $\delta_0 > 0$,

$$\int_0^{\delta_0} (-\log[\gamma^{-1}(a/2)])^{1/2}\,da < \infty. \tag{33}$$

As an example, we give conditions on G that suffice for Assumption 7 in the case of Bierens-type specification testing for nonlinear regression. We put $m(Z,\theta,\tau) = G(\tilde{X}'\tau)(Y - f(X,\theta))$, $s(Z,\theta) = -\nabla f(X,\theta)(Y - f(X,\theta))$ with $\nabla' m(Z,\theta,\tau) = -G(\tilde{X}'\tau)\nabla' f(X,\theta)$ and $A^* = E[\nabla f(X,\theta^*)\nabla' f(X,\theta^*) - \nabla^2 f(X,\theta^*)\epsilon]$, where $\epsilon = Y - f(X,\theta^*)$ and θ^* solves

$$\min_{\theta \in \Theta} E([Y - f(X,\theta)]^2). \tag{34}$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Example 5.3.

Let Θ be compact and let $f: \mathbb{R}^k \times \Theta \to \mathbb{R}$ be measurable on $\mathbb{R}^k$ for each $\theta \in \Theta$ and twice continuously differentiable on an open neighborhood of Θ for each $x \in \mathbb{R}^k$. Suppose that Y, $f(X,\theta^*)$, and $\nabla f(X,\theta^*)$ have finite second moments, that Assumption 6 holds, and that A^{*-1} exists. Let T be a compact subset of $\mathbb{R}^{k+1}$. Then Assumption 7 holds if G is Lipschitz on compact intervals.

We can now state our first main result of this section.

THEOREM 5.4.

(a) *Let Assumptions* 1–7 *hold with* $\mathbb{B} = C(T)$ *and* sup norm $\|\cdot\|_\infty$. *Under* H_0

$$n^{-1/2}\sum_{i=1}^{n} \mathcal{L}_i^*/\sigma_* \Rightarrow_{\mathbb{B}} \mathcal{Z}, \qquad n^{-1/2}\sum_{i=1}^{n} \hat{\mathcal{M}}_i/\hat{\sigma}_n \Rightarrow_{\mathbb{B}} \mathcal{Z}, \qquad \text{and}$$

$$n^{-1/2}\sum_{i=1}^{n} \hat{\mathcal{M}}_i/\tilde{\sigma}_n \Rightarrow_{\mathbb{B}} \mathcal{Z}, \tag{35}$$

where $\{\mathcal{Z}(\tau): \tau \in T\}$ *is the zero mean Gaussian process with covariance given by*

$$\operatorname{cov}(\mathcal{Z}(\tau_1), \mathcal{Z}(\tau_2)) = \operatorname{cov}(\phi(Z,\tau_1), \phi(Z,\tau_2)). \tag{36}$$

(b) *Let* $g: \mathbb{B} \to \mathbb{R}^+$ *be* $\|\cdot\|_\infty$*-continuous on* $\mathbb{B}$ *such that* $g(x) = 0$ *if and only if* $x = 0$. *Then* $g(n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n) \Rightarrow_{\mathbb{R}} g(\mathcal{Z})$ *and* $g(n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\tilde{\sigma}_n) \Rightarrow_{\mathbb{R}} g(\mathcal{Z})$.

As a direct consequence of this result and Theorem 10.13 of Ledoux and Talagrand we also obtain convergence results for the norms $\|\cdot\|_{p,\nu}$, $1 < p < \infty$, ν finite.

COROLLARY 5.5. *Suppose Assumptions* 1–7 *hold and let* $\mathbb{B} = C(T)$. *Let* $g: \mathbb{B} \to \mathbb{R}^+$ *be* $\|\cdot\|_{p,\nu}$*-continuous on* $\mathbb{B}$ *such that* $g(x) = 0$ *if and only if* $x = 0$. *Then under* H_0

$$g\left(n^{-1/2}\sum_{i=1}^{n} \hat{\mathcal{M}}_i/\hat{\sigma}_n\right) \Rightarrow_{\mathbb{R}} g(\mathcal{Z}) \quad \text{and} \quad g\left(n^{-1/2}\sum_{i=1}^{n} \hat{\mathcal{M}}_i/\tilde{\sigma}_n\right) \Rightarrow_{\mathbb{R}} g(\mathcal{Z}) \tag{37}$$

for all $1 \le p < \infty$, ν *finite, where* $\mathcal{Z}$ *is the Gaussian process of Theorem* 5.4.

Obtaining asymptotic critical values for our statistics is computationally challenging but feasible, for example, by Monte Carlo (Hansen 1996). Andrews (1997) discussed a semiparametric bootstrap procedure in a related application. Nevertheless, a simple and sharp asymptotic bound follows from the Banach LIL (Ledoux and Talagrand, 1991, Theorem 8.2). We strengthen Assumption 2.

Assumption 2′. For each $n \in \mathbb{N}$, $\hat{\theta}_n: \Omega \to \Theta$ (with Θ a compact subset of $\mathbb{R}^p$) is measurable. For $\theta^* \in \operatorname{int}\Theta$, $\hat{\theta}_n = \theta^* + o_{\text{a.s.}}(1)$. There exists a function $s: \mathbb{R}^l \times \Theta \to \mathbb{R}^p$ measurable on $\mathbb{R}^l$ for each θ in Θ and continuously differentiable on int Θ for each z in $\mathbb{R}^l$ such that

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

$$n^{-1/2}\sum_{i=1}^{n} s(Z_i,\hat{\theta}_n) = o_{\text{a.s.}}(1), \tag{38}$$

with $E(s(Z_i,\theta^*)) = 0$ and $\sup_{\theta\in\Theta}|\nabla s(Z_i,\theta)|$ integrable. Define $A^* = E(\nabla s(Z_i,\theta^*))$ and assume that A^* is nonsingular.

THEOREM 5.6.

(a) *Suppose Assumptions* 1, 2′, *and* 3–7 *hold. Then with probability* 1 *under* H_0

$$\limsup_{n\to\infty}\left\|n^{-1/2}\sum_{i=1}^{n}\mathcal{L}_i^*/\sigma_*\right\|\Big/(2\log\log n)^{1/2}$$

$$= \limsup_{n\to\infty}\left\|n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\hat{\sigma}_n\right\|\Big/(2\log\log n)^{1/2} \tag{39}$$

$$= \limsup_{n\to\infty}\left\|n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\tilde{\sigma}_n\right\|\Big/(2\log\log n)^{1/2} = 1, \tag{40}$$

where $\|\cdot\| = \|\cdot\|_\infty$ *or where* $\|\cdot\| = \|\cdot\|_{p,\nu}$ *for* $1 \le p < \infty$ *and* ν *is any finite measure on* T.

(b) *Further, let* $g:\mathbb{B}\to\mathbb{R}^+$ *be such that there exists* $\bar{\gamma}:\mathbb{R}^+\to\mathbb{R}^+$ *monotone increasing for which* $g(x) \le \bar{\gamma}(\|x\|_\infty)$ *for all* $x\in\mathbb{B}$. *Then with probability one under* H_0 *we have*

$$\limsup_{n\to\infty} g\left(n^{-1/2}\sum_{i=1}^{n}\mathcal{L}_i^*/\sigma_*\right)\Big/\bar{\gamma}([2\log\log n]^{1/2}) \le 1, \tag{41}$$

$$\limsup_{n\to\infty} g\left(n^{-1/2}\sum_{t=1}^{n}\hat{\mathcal{M}}_i/\tilde{\sigma}_n\right)\Big/\bar{\gamma}([2\log\log n]^{1/2}) \le 1, \tag{42}$$

and

$$\limsup_{n\to\infty} g\left(n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\tilde{\sigma}_n\right)\Big/\bar{\gamma}([2\log\log n]^{1/2}) \le 1. \tag{43}$$

The statistic $g(n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\hat{\sigma}_n)$ can exceed $\bar{\gamma}([2\log\log n]^{1/2})$ only finitely many times, with probability 1 (w.p. 1) under H_0, so that rejecting H_0 if $g(n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\hat{\sigma}_n) > \bar{\gamma}([2\log\log n]^{1/2})$ delivers a test of asymptotic size *zero*. The LIL bound is independent of the norm. For moderate-sized samples, the LIL bound is likely to be rather conservative. Table 1 gives some sample values for $(2\log\log n)^{1/2}$. We therefore recommend basing a preliminary test on the LIL bound. If one fails to reject with the LIL bound, then the evidence is well in accord with H_0, and one can avoid further computation. If the statistic exceeds the LIL bound, we suggest using Monte Carlo or the bootstrap to compute an accurate *p*-value.

Global power of tests based on $n^{-1/2}\sum_{i=1}^{n}\hat{\mathcal{M}}_i/\hat{\sigma}_n$ is established by the next result.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

TABLE 1. Values for $(2 \log\log n)^{1/2}$

n	$(2 \log\log n)^{1/2}$
100	1.7477
500	1.9115
1,000	1.9660
5,000	2.0698
10,000	2.1073

THEOREM 5.7.

(a) *Suppose Assumptions* 1, 2′, *and* 3–6 *hold. Then under* H_A *and* H_A^*,

$$P\left[\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \hat{\sigma}_n \right\|_\infty > n^{1/2}\delta \quad a.a.n.\right]$$
$$= P\left[\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \tilde{\sigma}_n \right\|_\infty > n^{1/2}\delta \quad a.a.n.\right] = 1 \tag{44}$$

for some $\delta > 0$. *Further, under* H_A

$$P\left[\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \hat{\sigma}_n \right\|_{p,\nu} > n^{1/2}\delta \quad a.a.n.\right]$$
$$= P\left[\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \tilde{\sigma}_n \right\|_{p,\nu} > n^{1/2}\delta \quad a.a.n.\right] = 1 \tag{45}$$

for some $\delta > 0$, *for all* $1 \le p < \infty$ *and all* ν *placing positive measure in a sufficiently small neighborhood of* τ_0. *Under* H_A^*, *the preceding relations hold for all* $1 \le p < \infty$ *and all finite* ν.

(b)

(i) *Let* $g : \mathbb{B} \to \mathbb{R}^+$ *be such that there exists* $\underline{\gamma} : \mathbb{R}^+ \to \mathbb{R}^+$ *monotone increasing and a measure* ν *on* T *for which* $g(x) \ge \underline{\gamma}(\|x\|_{1,\nu})$ *for all* $x \in \mathbb{B}$. *Then under* H_A

$$P\left[g\left(n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \hat{\sigma}_n\right) > \underline{\gamma}(n^{1/2}\delta) \quad a.a.n.\right]$$
$$= P\left[g\left(n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \tilde{\sigma}_n\right) > \underline{\gamma}(n^{-1/2}\delta) \quad a.a.n.\right] = 1 \tag{46}$$

for some $\delta > 0$, *provided that* ν *places positive measure in a sufficiently small neighborhood of* τ_0. *Under* H_A^* *the restriction on* ν *is unnecessary.*

(ii) *Let* $g : \mathbb{B} \to \mathbb{R}^+$ *be such that there exists* $\underline{\gamma} : \mathbb{R}^+ \to \mathbb{R}^+$ *monotone increasing for which* $g(x) \ge \underline{\gamma}(\|x\|_\infty)$ *for all* $x \in \mathbb{B}$. *Then under* H_A *and* H_A^* *for some* $\delta > 0$,

$$P\left[g\left(n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \hat{\sigma}_n\right) > \underline{\gamma}(n^{1/2}\delta) \quad a.a.n.\right]$$
$$= P\left[g\left(n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i / \tilde{\sigma}_n\right) > \underline{\gamma}(n^{1/2}\delta) \quad a.a.n.\right] = 1. \tag{47}$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Thus, the asymptotic size zero test will be consistent when $g(\cdot)$ is properly chosen.

We complete the nonlinear regression example by specifying standard regularity conditions under which the results of this section can be invoked.

Regularity Condition 1. $\{Z_i = (Y_i, X_i')'\}$ is an i.i.d. sequence such that $E(Y_i^2) < \infty$ and X_i is bounded.

Regularity Condition 2. $f: \mathbb{R}^k \times \Theta \to \mathbb{R}$ satisfies (a) $\sup_{\theta\in\Theta} f(X_i,\theta)^2$ is integrable; (b) $\sup_{\theta\in\Theta} \nabla' f(X_i,\theta)\nabla f(X_i,\theta)$ is integrable; (c) $\sup_{\theta\in\Theta} |\nabla^2 f(X_i,\theta)(Y_i - f(X_i,\theta))|$ is integrable; (d) $\sup_{\theta\in\Theta} (Y_i - f(X_i,\theta))^2 \nabla' f(X_i,\theta)\nabla f(X_i,\theta)$ is integrable.

Regularity Condition 3. (a) $E([Y_i - f(X_i,\theta]^2)$ has a unique minimum at $\theta^* \in$ int Θ; (b) $\det A^* \neq 0$.

Regularity Condition 4. Assumption 5 holds.

Regularity Condition 5. G is Lipschitz on compacts and T is compact with nonempty interior.

Regularity Conditions 1 and 2 guarantee the existence of a measurable solution $\hat{\theta}_n$ to the problem

$$\min_{\theta\in\Theta} n^{-1} \sum_{i=1}^{n} [Y_i - f(X_i,\theta)]^2/2. \tag{48}$$

Set $\hat{A}_n = n^{-1}\sum_{i=1}^n \nabla\hat{f}_i\nabla\hat{f}_i' - \nabla^2\hat{f}_i\hat{\epsilon}_i$, where $\nabla\hat{f}_i = \nabla f(X_i,\hat{\theta}_n)$, $\nabla^2\hat{f}_i = \nabla^2 f(X_i,\hat{\theta}_n)$ and $\hat{\epsilon}_i = Y_i - f(X_i,\hat{\theta}_n)$. Regularity condition 4 imposes Assumption 5 directly for convenience. It suffices that the conditional variance of $\epsilon = Y - f(X,\theta^*)$ given X is bounded away from zero and that there is a less than perfect fit to $G(\tilde{X}_i'\tau)$ for all $\tau \in T$ from the "regression" $\nabla' f(X_i,\theta^*)A^{*-1}E(\nabla f(X_i,\theta^*)G(\tilde{X}_i'\tau))$.

COROLLARY 5.8. *The conclusions of Theorem* 5.4, *Corollary* 5.5, *Theorem* 5.6, *and Theorem* 5.7 *hold for* $\mathcal{L}_i^*$, $\hat{\mathcal{M}}_i$, $\hat{\sigma}_n$, *and* $\tilde{\sigma}_n$ *constructed under regularity conditions* 1–5. *In particular*, H_0 *occurs when* $P[E(Y|X) = f(X,\theta_0)] = 1$ *for some* θ_0 *interior to* Θ, *whereas* H_A^* *occurs when* $P[E(Y|X) = f(X,\theta)] < 1$ *for all* $\theta \in \Theta$ *and* G *is a nonpolynomial, real analytic function.*

6. MATHEMATICAL PROOFS

Proof of Theorem 2.3. Immediate from Corollary 3.9. ■

Proof of Theorem 3.1. The equivalence of (a), (b), and (d) is established in the text. The equivalence with (c) follows directly from the discussion of polar topologies (Robertson and Robertson, 1973, III.2). ■

Proof of Theorem 3.3. (a) If sp $\mathcal{H}$ is uniformly dense in $C(B)$, then Lusin's theorem (e.g., Ash, 1972, Corollary 4.3.17(b)) implies comprehensiveness. Suppose sp $\mathcal{H}$ is *not* uniformly dense in $C(B)$. The uniform closure of sp $\mathcal{H}$ is a closed linear subspace of $C(B)$. As the set of finite signed measures is the dual of $C(B)$, there is a nonzero finite signed measure μ such that for all η in the uniform

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

closure of sp $\mathcal{H}$, $\int_B \eta \, d\mu = \int_B \mathbf{1}\eta \, d\mu = 0$. Put $\xi = \mathbf{1}$; then $\mathcal{H}$ is not comprehensive because $\mathbf{1} \neq 0$ μ − a.e.

(b) By the dominated convergence theorem (DCT), uniform denseness in a comprehensive set implies comprehensiveness, hence condition (iii) suffices. Condition (i) thus suffices by Stone–Weierstrass. To prove condition (ii) suffices, we note that $M_b(B)$ is comprehensive. A variant of the monotone class theorem (e.g., Dellacherie and Meyer, 1978, Theorem 22.2, p. 15) implies that the smallest class of functions containing $\mathcal{A}$ closed under both uniform convergence and bounded monotone convergence is $M_b(B)$. Thus, for any μ, any $h \in M_b(B)$ is the μ a.e. limit of a uniformly bounded sequence of elements of $\mathcal{A}$. Therefore, the DCT implies comprehensiveness of $\mathcal{A}$. ■

Proof of Lemma 3.5. See Hornik (1991). ■

Proof of Lemma 3.7. If the uniform closure of sp $\mathcal{H}_G(T)$ contains $C(B)$, then so must the uniform closure of its superset, $\Sigma(G,T)$.

Now suppose the uniform closure of $\Sigma(G,T)$ contains $C(B)$ for every compact $B \subset \mathbb{R}$ and $T \subset \mathbb{R}^2$ has nonempty interior. By Stinchcombe and White (1990, Lemma 2.0), it suffices to treat this case. Suppose for purposes of contradiction that sp $\mathcal{H}_G(T)$ is not dense in $C(B)$ for some T and B. This happens if and only if there is a nonzero finite signed measure μ supported on B such that for all $\tau \in T$, $\int_{\mathbb{B}} G(\tilde{x}'\tau)\mu(dx) = 0$. Taking a subset of T if necessary, there is no loss in assuming T belongs entirely in one quadrant of $\mathbb{R}^2$.

Let $[a,b]$ be a closed interval containing an ϵ-neighborhood of B for some $\epsilon > 0$. Pick $\delta > 0$ and $\tau' \in T$ such that $S(\tau',2\delta)$, the ball of radius 2δ around τ', is contained in T. By assumption, $\Sigma(G,S(\tau',\delta))$ is uniformly dense in $C([a,b])$. In particular, for every $n \in \mathbb{N}$ and for every $a < a' \leq b' < b$, some element of $\Sigma(G,S(\tau',\delta))$ is uniformly within n^{-1} of the continuous function $f^n(x) := \max\{1 - nd(x,[a',b']),0\}$, where $d(x,[a',b'])$ is the distance from x to the set $[a',b']$.

The sequence f^n is uniformly bounded and converges pointwise to the indicator of the interval $[a',b']$. Therefore, as n goes to infinity, $\int_{[a,b]} f^n(x)\, d\mu(x)$ goes to $\mu([a',b'])$. Because each f^n is in the span of $\mathcal{H}_G(S(\tau',\delta))$ and $\mathbf{1}$, $f^n(x) = \beta_{0,n} + \sum_{j,n}^{J,n} \beta_{j,n} G(\tau_{j,n}x + \tau_{0,j,n})$, where each $(\tau_{j,n},\tau_{0,j,n}) \in S(\tau',\delta)$. The basic idea is that we can "horizontally stretch" the functions f^n without changing their integral against μ, and this cannot happen unless μ is equal to 0. Formally, as the integral of μ against any element of $\mathcal{H}_G(T)$ is 0, we can substitute any $(t_{j,n},t_{0,j,n}) \in T$ for each $(\tau_{j,n},\tau_{0,j,n})$ without changing the integral of f^n against μ. Let $c_{j,n}$ be that point in $\mathbb{R}$ such that $\tau_{j,n}c_{j,n} + \tau_{0,j,n} = a'$, and let $d_{j,n}$ be that point in $\mathbb{R}$ such that $\tau_{j,n}d_{j,n} + \tau_{0,j,n} = b'$. These exist because T belongs entirely to a single quadrant in $\mathbb{R}^2$.

Because $S(\tau',\delta) \subset S(\tau',2\delta) \subset T$, there exists some $\eta \in (0,\epsilon)$ such that for all (j,n)-pairs there exists $(t_{j,n},t_{0,j,n} \in T)$ such that $t_{j,n}c_{j,n} + t_{0,j,n} = a'$ and $t_{j,n}d_{j,n} + t_{0,j,n} = b' + \eta$. Because T belongs to a single quadrant in $\mathbb{R}^2$, the signs of $\tau_{j,n}$ and $t_{j,n}$ must agree.

Denote by $\{g^n\}$ the sequence of functions in $\Sigma(G,T)$ that are derived from $\{f^n\}$ by replacing each $(\tau_{j,n},\tau_{0,j,n})$ by the corresponding $(t_{j,n},t_{0,j,n})$. The sequence $\{g^n\}$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

converges pointwise to the indicator of the interval of $[a', b' + \eta]$. Therefore, for each $a < a' \leq b' < b$, $\mu([a', b']) = \mu([a', b' + \eta])$. Because the interval $[a, b]$ contains an ϵ-neighborhood of B and μ is supported on B, this implies that μ is the 0 measure, the contradiction that completes the proof. ■

Proof of Theorem 3.8. If G is not generically comprehensive, then there is a nonempty open set $T \subset \mathbb{R}^{k+1}$ and a compact set K such that $\text{sp}\{G(\cdot'\tau) : \tau \in T\}$ is not uniformly dense in $C(K)$. The Hahn–Banach theorem then implies existence of a nonzero finite signed measure μ supported on K such that for all $\tau \in T$, $m(\tau) := \int G(\tilde{x}'\tau)\mu(dx) = 0$. But m is real analytic because G is and because μ is compactly supported. As a real analytic function is equal to 0 on the open set T if and only if it is equal to 0 everywhere, $\mathcal{H}_G$ is not comprehensive. ■

Proof of Corollary 3.9. If G is a polynomial then it is clearly not comprehensive. If G is real analytic but is not a polynomial, then every interval is a very nice interval. By Lemma 3.5, this implies that G is comprehensive. By Theorem 3.8, this implies that G is generically comprehensive. ■

Proof of Theorem 3.10. See Stinchcombe and White (1992, Theorem 2.2). ■

Proof of Theorem 3.11. It is immediate that the real analytic functions that are not polynomials are dense in the compact-open topology; hence the result follows from Corollary 3.9. ■

Proof of Theorem 4.1. By Theorem 3.3, $\{\phi(\cdot, \tau) : \tau \in T\} \subset C(B)$ is comprehensive if and only if its span is uniformly dense in $C(B)$. (a) follows from Dunford and Schwartz (1958, Theorem V.5.1, p. 426). Given (a), (b) follows from the observation that the mapping $(\mu, f) \to \int f d\mu$ is jointly continuous on sets of measures with $v\mu$ uniformly bounded. ■

Proof of Lemma 5.1. Assumptions 1–3 permit a Taylor expansion around θ^* of the form

$$
\begin{aligned}
&n^{-1/2} \sum_{i=1}^{n} m(Z_i, \hat{\theta}_n, \tau) \\
&\quad = n^{-1/2} \sum_{i=1}^{n} m(Z_i, \theta^*, \tau) + \left[n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \bar{\theta}_n, \tau) \right] \sqrt{n}(\hat{\theta}_n - \theta^*) \\
&\quad = n^{-1/2} \sum_{i=1}^{n} m(Z_i, \theta^*, \tau) - \left[n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \bar{\theta}_n, \tau) \right] \\
&\qquad \times \left[A^{*-1} n^{-1/2} \sum_{i=1}^{n} s(Z_i, \theta^*) + o_P(1) \right] \\
&\quad = n^{-1/2} \sum_{i=1}^{n} m(Z_i, \theta^*, \tau) - M(\theta^*, \tau) \left[A^{*-1} n^{-1/2} \sum_{i=1}^{n} s(Z_i, \theta^*) + o_P(1) \right] \\
&\qquad - \left(n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \hat{\theta}_n, \tau) - M(\theta^*, \tau) \right) \left[A^{*-1} n^{-1/2} \sum_{i=1}^{n} s(Z_i, \theta^*) + o_P(1) \right] \\
&\quad = n^{-1/2} \sum_{i=1}^{n} [m(Z_i, \theta^*, \tau) - M(\theta^*, \tau) A^{*-1} s(Z_i, \theta^*)] + o_P(1) \qquad \textbf{(49)}
\end{aligned}
$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

uniformly on T. Uniformity follows as $M(\theta^*,\cdot)$ is continuous on the compact set T (implying that $M(\theta^*,\tau) = O(1)$ uniformly in τ) and from the WULLN for $\{\nabla' m(Z_i,\theta,\tau)\}$. Thus $(n^{-1}\sum_{i=1}^n \nabla' m(Z_i,\bar{\theta}_n,\tau) - M(\theta^*,\tau)) = o_P(1)$ uniformly in τ. The result now follows immediately. ■

Proof of Example 5.3. Because the integral (33) is finite when $\gamma(\delta) = \delta^a$, $a > 0$, it suffices that $|\phi|$ be bounded above by a square integrable random variable times a Lipschitz function. Given the moment conditions and that G is Lipschitz on the range of $\tilde{X}'T$, this is clearly true for the numerator of ϕ. Given that $\sigma^2(\tau)$ is bounded below, the denominator is also Lipschitz. Because T is compact, this suffices. ■

Proof of Theorem 5.4. (a) Assumptions 1–5 and 7 suffice to apply Ossiander's (1987) Banach CLT to $n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma^*$. The results for $n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n$ hold if $\|n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n - n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_*\|_\infty = o_P(1)$. The triangle inequality gives

$$\begin{aligned}
&\left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n - n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_* \right\|_\infty \\
&\quad \le \left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n - n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_* \right\|_\infty \\
&\qquad + \left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_* - n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_* \right\|_\infty \\
&\quad \le \|\sigma_*\hat{\sigma}_n^{-1} - 1\|_\infty \left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_* \right\|_\infty \\
&\qquad + \left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_* - n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_* \right\|_\infty . \qquad \textbf{(50)}
\end{aligned}$$

As $n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_*$ obeys the Banach CLT, Lemma 5.1 gives $\|n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\sigma_*\|_\infty = O_P(1)$. The desired approximation holds if $\|\sigma_*\hat{\sigma}_n^{-1} - 1\|_\infty = o_P(1)$. For this it suffices that $\|\hat{\sigma}_n^2 - \sigma_*^2\|_\infty = o_P(1)$ given Assumption 5, which follows easily given Assumptions 3, 4, and 6. The argument with $\tilde{\sigma}_n^2$ replacing $\hat{\sigma}_n^2$ is identical.

(b) Because $\{\mathcal{Z}(\tau):\tau \in T\}$ is a continuous process and g is continuous with respect to $\|\cdot\|_\infty$, it follows from the continuous mapping theorem (Billingsley, 1968, p. 30) that

$$g\left(n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat{\sigma}_n\right) \Rightarrow_{\mathbb{R}} g(\mathcal{Z}) \quad \text{and} \quad g\left(n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\tilde{\sigma}_n\right) \Rightarrow_{\mathbb{R}} g(\mathcal{Z}). \qquad \textbf{(51)}$$

■

Proof of Corollary 5.5. Apply the continuous mapping theorem. ■

Proof of Theorem 5.6. (a) Assumptions 1, 2′, 3–5, and 7 permit application of the Banach LIL (Ledoux and Talagrand, 1991, Theorem 8.2) to $n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_*$. For $\|\cdot\|$, it suffices that

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

$$\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\hat{\sigma}_n - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^*/\sigma^* \right\| \Big/ (2 \log\log n)^{1/2} = o_{\text{a.s.}}(1). \tag{52}$$

The argument with $\tilde{\sigma}_n$ replacing $\hat{\sigma}_n$ is identical. The triangle inequality gives

$$\begin{aligned}
&\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\hat{\sigma}_n - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^*/\sigma^* \right\| \Big/ (2 \log\log n)^{1/2} \\
&\quad \leq \|\sigma_* \hat{\sigma}_n^{-1} - 1\| \times \left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\sigma_* \right\| \Big/ (2 \log\log n)^{1/2} \\
&\qquad + \left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\sigma_* - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^*/\sigma_* \right\| \Big/ (2 \log\log n)^{1/2}.
\end{aligned} \tag{53}$$

Thus it suffices that $\|\sigma_* \hat{\sigma}_n^{-1} - 1\| = o_{\text{a.s.}}(1)$, $\|n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\sigma_*\|/(2 \log\log n)^{1/2} = O_{\text{a.s.}}(1)$ and that

$$\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\sigma_* - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^*/\sigma_* \right\| \Big/ (2 \log\log n)^{1/2} = o_{\text{a.s.}}(1). \tag{54}$$

The last equality and the Banach LIL for $n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^*/\sigma_*$ imply that $\|n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i/\sigma_*\|/(2 \log\log n)^{1/2} = O_{\text{a.s.}}(1)$. That $\|\sigma_* \hat{\sigma}_n^{-1} - 1\| = o_{\text{a.s.}}(1)$ follows under Assumptions 3, 4, and 6. The result thus follows by establishing (54). As $\sigma_*(\tau)$ is bounded below,

$$\left\| n^{-1/2} \sum_{i=1}^{n} \hat{\mathcal{M}}_i - n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_i^* \right\| \Big/ (2 \log\log n)^{1/2} = o_{\text{a.s.}}(1) \tag{55}$$

suffices. Taylor expansion around θ^* gives

$$\begin{aligned}
&n^{-1/2} \sum_{i=1}^{n} m(Z_i, \hat{\theta}_n, \tau)/(2 \log\log n)^{1/2} \\
&\quad = n^{-1/2} \sum_{i=1}^{n} m(Z_i, \theta^*, \tau)/(2 \log\log n)^{1/2} \\
&\qquad + \left[n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \bar{\theta}_n, \tau) \right] \sqrt{n}(\hat{\theta}_n - \theta^*)/(2 \log\log n)^{1/2} \\
&\quad = n^{-1/2} \sum_{i=1}^{n} [m(Z_i, \theta^*, \tau) - M(\theta^*, \tau) A^{*-1} s(Z_i, \theta^*)]/(2 \log\log n)^{1/2} \\
&\qquad - \left[n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \bar{\theta}_n, \tau) - M(\theta^*, \tau) \right] \\
&\qquad \times A^{*-1} n^{-1/2} \sum_{i=1}^{n} s(Z_i, \theta^*)/(2 \log\log n)^{1/2} \\
&\qquad + \left[n^{-1} \sum_{i=1}^{n} \nabla' m(Z_i, \bar{\theta}_n, \tau) \right] \Bigg\{ \sqrt{n}(\hat{\theta}_n - \theta^*)/(2 \log\log n)^{1/2} \\
&\qquad\qquad + A^{*-1} n^{-1/2} \sum_{i=1}^{n} s(Z_i, \theta^*)/(2 \log\log n)^{1/2} \Bigg\}.
\end{aligned} \tag{56}$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

The result follows if the second and third terms vanish as $n \to \infty$ a.s., uniformly in τ. Now Assumptions 1, 2′ (on s), and 4 ensure that the LIL applies to $n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)$, so that $n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2}$ is $O_{\text{a.s.}}(1)$. Here A^{*-1} is $O(1)$ by Assumption 2′ and $[n^{-1}\sum_{i=1}^n \nabla' m(Z_i,\bar\theta_n,\tau) - M(\theta^*,\tau)] = o_{\text{a.s.}}(1)$ uniformly in τ by the strong uniform law of large numbers, provided that $\bar\theta_n \to \theta^*$ a.s. For this it suffices that $\hat\theta_n \to \theta^*$ a.s., as imposed in Assumption 2′. The second term thus vanishes a.s. uniformly in τ as required.

Because $\sup_{(\theta,\tau)\in\Theta\times T}|M(\theta,\tau)|$ is finite given Assumption 3(b) and because $\sup_{(\theta,\tau)\in\Theta\times T}|[n^{-1}\sum_{i=1}^n \nabla' m(Z_i,\theta,\tau) - M(\theta,\tau)]| \to 0$ a.s., it follows that the term $[n^{-1}\sum_{i=1}^n \nabla' m(Z_i,\bar\theta_n,\tau)]$ appearing in the third term of (56) is $O_{\text{a.s.}}(1)$ uniformly in τ. The result now follows if

$$\sqrt{n}(\hat\theta_n - \theta^*)/(2\log\log n)^{1/2}$$

$$= -A^{*-1}n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2} + o_{\text{a.s.}}(1). \tag{57}$$

By the standard mean value expansion

$$\sqrt{n}(\hat\theta_n - \theta^*)/(2\log\log n)^{1/2}$$

$$= -\nabla\bar s_n^{-1}n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2}$$

$$= -A^{*-1}n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2}$$

$$+ (A^{*-1} - \nabla\bar s_n^{-1})n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2} \quad \text{a.s., a.a.n,} \tag{58}$$

where $\nabla\bar s_n$ represents the gradient matrix of $n^{-1}\sum_{i=1}^n s(Z_i,\theta)$ with each row evaluated at a (different) mean value lying between $\hat\theta_n$ and θ^*. The SULLN for $\nabla s(Z_i,\theta)$ ensured by Assumptions 1 and 2′ and the consistency of $\hat\theta_n$ for θ^* imply that $\nabla\bar s_n - E(\nabla s(Z_i,\theta^*)) = o_{\text{a.s.}}(1)$. But $n^{-1/2}\sum_{i=1}^n s(Z_i,\theta^*)/(2\log\log n)^{1/2}$ is $O_{\text{a.s.}}(1)$ by the LIL, implying that (57) holds, so we are done.

(b) It suffices to consider $n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*/\sigma_*$. Other cases are analogous. From (a), for any $\epsilon > 0$ and for each ω in a set with probability one there exists $N_\omega(\epsilon) < \infty$ such that for all $n > N_\omega(\epsilon)$ we have $\|n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*(\omega)/\sigma_*\|_\infty/(2\log\log n)^{1/2} < 1+\epsilon$. By the condition on g we have $g(n^{-1/2}\sum_{i=1}^n \mathcal{L}_i^*(\omega)/\sigma_*)/g[(2\log\log n)^{1/2}] < 1+\epsilon$ for all $n > N_\omega(\epsilon)$ also, and the result holds. ■

Proof of Theorem 5.7. (a) It suffices to treat $\|\cdot\|_{1,\nu}$ as $\|\cdot\|_{1,\nu} \le \|\cdot\|_{p,\nu} \le \|\cdot\|_\infty$, $1 \le p < \infty$. By definition

$$\left\| n^{-1/2}\sum_{i=1}^n \hat{\mathcal{M}}_i/\hat\sigma_n \right\|_{1,\nu} = \int \left(\left| n^{-1/2}\sum_{i=1}^n m(Z_i,\hat\theta_n,\tau) \right| \Big/ \hat\sigma_n(\tau) \right) \nu(d\tau). \tag{59}$$

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Assumptions 1, 2′, and 3–6 ensure $\|\hat{\sigma}_n^2 - \sigma_*^2\|_\infty = 0$ a.s. This and continuity of σ_* on T ensure that $\|\hat{\sigma}_n\|_\infty \leq \Delta < \infty$ a.a.n. a.s., so $\inf_{\tau\in T} \hat{\sigma}_n^{-1}(\tau) \geq \Delta^{-1} > 0$ a.a.n. a.s. The triangle inequality gives

$$\begin{aligned} |E(m(Z,\theta^*,\tau))| \leq & \left| n^{-1} \sum_{i=1}^n m(Z_i,\hat{\theta}_n,\tau) \right| \\ & + \left| n^{-1} \sum_{i=1}^n m(Z_i,\hat{\theta}_n,\tau) - E(m(Z,\hat{\theta}_n,\tau)) \right| \\ & + |E(m(Z,\hat{\theta}_n,\tau)) - E(m(Z,\theta^*,\tau))|. \end{aligned} \tag{60}$$

Given the continuity of $E[m(Z,\cdot,\cdot)]$ on $\Theta \times T$, if $\hat{\theta}_n \to \theta^*$ a.s., then we have that for any $\epsilon > 0$,

$$\sup_{\tau\in T} |E(m(Z,\hat{\theta}_n,\tau)) - E(m(Z,\theta^*,\tau))| < \epsilon \quad \text{a.a.n., a.s.} \tag{61}$$

Assumptions 4(a) and 1 deliver a SULLN for $\{m(Z_i,\theta,\tau)\}$. By Assumption 2′, $\hat{\theta}_n \to \theta^*$ a.s., which then implies

$$\sup_{\tau\in T} \left| n^{-1} \sum_{i=1}^n m(Z_i,\hat{\theta}_n,\tau) - E(m(Z_i,\hat{\theta}_n,\tau)) \right| < \epsilon \quad \text{a.a.n., a.s.} \tag{62}$$

Consequently,

$$|E(m(Z,\theta^*,\tau))| \leq \left| n^{-1} \sum_{i=1}^n m(Z_i,\hat{\theta}_n,\tau) \right| + 2\epsilon, \quad \text{a.a.n., a.s.,} \tag{63}$$

so that (with ν a probability measure for convenience)

$$n^{1/2} \int \left| n^{-1} \sum_{i=1}^n m(Z_i,\hat{\theta}_n,\tau) \right| d\nu(\tau) > n^{1/2} \left[\int |E(m(Z,\theta^*,\tau))|\, d\nu(\tau) - 2\epsilon \right]. \tag{64}$$

As we may take $\epsilon = \delta^*/4$, the test will be consistent whenever

$$\delta^* = \int |E(m(Z,\theta^*,\tau))|\, \nu(d\tau) > 0. \tag{65}$$

Under H_A, m is such that $|E(m(Z,\theta^*,\tau))| > 0$ for some $\tau_0 \in T$ under H_A. Because $E(m(Z,\theta^*,\tau))$ is continuous in τ, there will exist a neighborhood of τ_0 of positive Lebesgue measure for which $|E(m(Z,\theta^*,\tau))| > 0$. Because ν puts positive mass in this neighborhood and $\delta^* > 0$ we have a consistent test. Under H_A^*, ν is such that $|E(m(Z,\theta^*,\tau))| > 0$ for essentially all $\tau \in T$, and again $\delta^* > 0$.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

When $\|\cdot\| = \|\cdot\|_\infty$, the measure ν does not enter. The conditions given guarantee that

$$\sup_{\tau\in T}\left|n^{-1}\sum_{i=1}^{n} m(Z_i,\hat{\theta}_n,\tau)\right| \to \sup_{\tau\in T}|E(m(Z,\theta^*,\tau)| \quad \text{a.s.}, \tag{66}$$

so $\|n^{-1/2}\sum_{i=1}^{n}\hat{M}_i/\hat{\sigma}_n\|_\infty$ diverges almost surely under H_A, hence H_A^*.

(b) Immediate, given (a) and the conditions on g. ■

Proof of Corollary 5.8. It suffices to verify that Regularity Conditions 1–5 imply Assumptions 1–7. Clearly Regularity Condition 1 implies 1. Regularity Conditions 1–3 suffice for 2′ (see White, 1981). Regularity Condition 2 ensures 3(a). Boundedness of X_i and T combined with continuity of G imply $G(\tilde{X}_i'T)$ is bounded. Together with 1 and 2, these imply 3(b). Boundedness of $G(\tilde{X}_i'T)$ plus Regularity Conditions 1 and 3 imply 4(a) and 4(b). Assumption 6 is ensured for $\hat{A}_n$ given Regularity Conditions 1 and 2 (which deliver a ULLN) and Regularity Conditions 1, 2, and 3, which ensure $\hat{\theta}_n \to \theta^*$ w.p. 1. Assumption 7 holds given Regularity Conditions 2 and 5 by Example 5.3. That H_0 and H_A^* hold follows by choice of m and Corollary 3.9. ■

NOTES

1. An analytic function is one locally equal to its Taylor expansion at each point of its domain, such as $\exp(\cdot)$, the logistic, the hyperbolic tangent, the sine and cosine, polynomials, etc.

2. Note that $\mathcal{H}_G$ contains transformations of affine combinations of X, whereas $\mathcal{H}_{\exp}$ contains transformations of linear combination of X, so that $\mathcal{H}_G$ with $G = \exp$ contains all scalar multiples of the functions in $\mathcal{H}_{\exp}$. This has no substantive impact on the class of functions considered. Theorem 2.3 derives from the study of denseness properties of sp $\mathcal{H}_G$, a class of functions known in the study of artificial neural networks as the output functions of single hidden layer feedforward networks with activation function G.

3. See (12)–(16) et. seq.

4. Eubank and Hart (1993) note that several tests in the literature are "mostly of the form $T = H(\hat{\epsilon})/\hat{\sigma}^2$, where $H(\cdot)$ is a quadratic functional that vanishes when its argument is null," in other words, an L^2 norm.

5. Recall that we have in effect an assumption that all random variables have bounded support. We argued in the text following Theorem 3.1 that this loses no generality.

6. Zheng (1994) provided an information criteria based test for arbitrary misspecification.

7. In this section we will discuss "moments" and "expectations." To change to signed measures, replace "moments" with "integrals" and "expectation" with "integral under μ."

REFERENCES

Andrews, D.W.K. (1993) An introduction to econometric applications of functional limit theory for dependent random variables. *Econometric Reviews* 12, 183–216.

Andrews, D.W.K. (1997) A conditional Kolmogorov test. *Econometrica* 65, 1097–1128.

Andrews, D.W.K. & W. Ploberger (1994) Optimal tests when a nuisance parameter is identified only under the alternative. *Econometrica* 62, 1383–1414.

Ash, R.B. (1972) *Real Analysis and Probability.* New York: Academic Press.

Bierens, H.B. (1982) Consistent model specification tests. *Journal of Econometrics* 26, 323–353.

https://doi.org/10.1017/S0266466698143013 Published online by Cambridge University Press

Bierens, H.B. (1990) A consistent conditional moment test of functional form. *Econometrica* 58, 1443–1458.

Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.

Blum, J.R., J. Kiefer, & M. Rosenblatt (1961) Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics* 32, 485–498.

Cox, D. & E. Koh (1989) A smoothing spline based test of model adequacy in polynomial regression. *Annals of the Institute of Statistical Mathematics* 41, 383–400.

Cox, D., E. Koh, G. Wahba, & B. Yandell (1988) Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Annals of Statistics* 16, 113–119.

Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–254.

Davies, R.B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.

Dellacherie, C. & P.-A. Meyer (1978) *Probabilities and Potential.* Amsterdam: North-Holland.

Dunford, N. & J.T. Schwartz (1958) *Linear Operators. Part 1: General Theory.* New York: Wiley.

Eubank, R. & J. Hart (1992) Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* 20, 1412–1425.

Eubank, R. & J. Hart (1993) Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* 80, 89–98.

Eubank, R. & V. LaRiccia (1992) Asymptotic comparison of Cramér-Von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Annals of Statistics* 20, 2071–2086.

Eubank, R. & S. Speckman (1990) Curve fitting by polynomial-trigonometric regression. *Biometrika* 77, 1–9.

Eubank, R. & C. Spiegelman (1990) Testing the goodness of fit of linear models via nonparametric regression techniques. *Journal of the American Statistical Association* 85, 387–392.

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

Ghorai, J. (1980) Asymptotic normality of a quadratic measure of orthogonal series type density estimate. *Annals of the Institute of Statistical Mathematics* 19, 999–1009.

Gozalo, P. (1993) A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory* 9, 451–477.

Hansen, B. (1990) Lagrange Multiplier Tests for Parametric Instability in Nonlinear Models. Discussion paper, Department of Economics, University of Rochester.

Hansen, B. (1996) Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–430.

Härdle, W. & Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21, 1926–1947.

Hausman, J. (1978) Specification tests in econometrics. *Econometrica* 46, 1251–1272.

Holly, A. (1982) A remark on Hausman's specification test. *Econometrica* 50, 749–760.

Holst, L. & J.S. Rao (1980) Asymptotic theory for some families of two-sample nonparametric statistics. *Sankhyā, Series A*42, 19–52.

Hong, Y.-M. & H. White (1995) Consistent specification testing via nonparametric series regression. *Econometrica* 63, 1133–1159.

Hornik, K. (1991) Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4, 251–257.

Hornik, K., M. Stinchcombe, & H. White (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–368.

Huber, P. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In L.M. LeCam and J. Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, vol. 1, pp. 221–233. Berkeley: University of California Press.

Ledoux, M. & M. Talagrand (1991) *Probability in Banach Spaces: Isoperimetry and Processes.* New York: Springer-Verlag.

Leung, S.F. & S. Yu (1995) A New Regression Specification Error Test. Discussion paper, Department of Economics, University of Rochester.

Newey, W. (1985) Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–1070.

Ossiander, M. (1987) A central limit theorem under metric entropy with L_2 bracketing. *Annals of Probability* 15, 897–919.

Robertson, A.P. & W. Robertson (1973) *Topological Vector Spaces*, 2nd ed. Cambridge: Cambridge University Press.

Robinson, P.M. (1991) Consistent nonparametric entropy-based testing. *Review of Economic Studies* 58, 437–453.

Rumelhart, D.E., G. Hinton, & R. Williams (1986) Learning internal representations by error propagation. In D.E. Rumelhart & J. McClelland (eds.), *Parallel Distributed Processing*, vol. 1, pp. 318–362. Cambridge, MA: MIT Press.

Schweizer, B. & E.F. Wolff (1976) Sur une mesure de dépendance pour les variable aléatoires. *Comptes Rendus de l'Academie des Sciences de Paris* 283A, 659–661.

Schweizer, B. & E.F. Wolff (1981) On nonparametric measures of dependence for random variables. *Annals of Statistics* 9, 879–885.

Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Stinchcombe, M. & H. White (1990) Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 7–16. New York: IEEE Press.

Stinchcombe, M. & H. White (1992) Using feedforward networks to distinguish multivariate populations. *Proceedings of the International Joint Conference on Neural Networks*, vol. 1. pp. 788–793. New York: IEEE Press.

Tauchen, G. (1985) Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415–444.

White, H. (1981) Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76, 414–433.

White, H. (1987) Specification testing in dynamic models. In T. Bewley, (ed.), *Advances in Econometrics—Fifth World Congress*, vol. 1, pp. 1–58. New York: Cambridge University Press.

White, H. (1989a) Some asymptotic results for learning in single hidden layer feedforward networks. *Journal of the American Statistical Association* 84, 1003–1013.

White, H. (1989b) An additional hidden unit test for neglected nonlinearity. *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, pp. 451–455. New York: IEEE Press.

White, H. (1994) *Estimation, Inference and Specification Analysis.* New York: Cambridge University Press.

Wolff, E.F. (1981) *N*-dimensional measures of dependence. *Stochastica* 4, 175–188.

Zheng, J. (1994) A Consistent Test of Conditional Parametric Distributions. Working paper, Department of Economics, University of Texas at Austin.

Zheng, J. (1996) A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75, 263–289.