Testing and the Method of Sieves

Peter J. Bickel
University of California at Berkeley

Ya'acov Ritov
Hebrew University

Thomas M. Stoker
Massachusetts Institute of Technology

July 1998

Abstract

This paper develops test statistics based on scores for the specification of regression in nonparametric and semiparametric contexts. We study how different types of test statistics focus power on different directions of departure from the null hypothesis. We consider index models as basic examples, and utilize sieves for nonparametric approximation. We examine various goodness-of-fit statistics, including Cramer-von Mises and Kolmogorov-Smirnov forms. For a "box-style" sieve approximation, we establish limiting distributions of these statistics. We develop a bootstrap resampling method for estimating critical values for the test statistics, and illustrate their performance with a Monte Carlo simulation.

1. Introduction

The practice of statistical testing plays several roles in empirical research. These roles range from the careful assessment of the evidence against specific scientific hypotheses to the judgement of whether an estimated model displays decent goodness-of-fit to the empirical data. The careful analysis of size and power in statistical theory is often obscured by the presentation of a battery of test results in support of a basic model or a view about the behavioral processes that generate the empirical data. An essential part of the assessment of the results of statistical

testing is an understanding of what kinds of departures from the basic model have been checked.

In this paper we develop some general theory for score tests in semiparametric and nonparametric contexts. The approach of score testing is based on the specific delineation of alternative hypotheses, and permits a cohesive discussion of testing for departures in specific directions as well as the combination of tests in multiple directions. Our focus is on approximation of large spaces of alternatives by sieves, or the set analogue of series expansions of functions. This facilitates the orderly consideration of increasing ranges of alternative directions. While our main results encompass many familiar tests, our view is that the approach will provide structure to the understanding of exactly how 'nonparametric' approximation of alternatives is being achieved in a practical testing context.

The literature on testing with nonparametric alternatives is reasonably large and growing rapidly. Hart (1997) provides an excellent overview of the statistical literature in the context of analyzing (one-dimensional) regression structure. Work more closely in line with our focus on score testing include Choi, Hall and Schick (1996) and Fan (1996). Yatchew (1998) provides coverage of the econometric literature on testing with nonparametric alternatives; also see Aït Sahalia, Bickel and Stoker (1998) for references. Our focus on alternative directions is reasonably close in spirit to the testing of conditional moment restrictions in econometric models; see Bierens (1990), Lewbel (1995), Bierens and Ploberger (1997) and Chen and Fan (1998), among others.

We start with a short description of index models and the specific testing problem we consider. In Section 3 we describe a heuristic construction of score tests of a semiparametric hypothesis against a one-dimensional alternative, and then we present ways to put score tests together against composite alternatives. The sieve method and its importance to the testing problem is presented in Section 4. The index model is revisited in Section 5 and particular tests are suggested. Section 6 gives (for index models) the asymptotics of the score process. Section 7.1 contains the main results for inference in index models under the hypothesis and contiguous alternatives, while 7.2 shows how bootstrap critical values can be used to implement the tests. Section 8 deals with consistency under fixed alternatives. Section 9 gives simulations that show the importance of alternative directions and testing power. Section 10 contains some concluding remarks, and various proofs are given in the Appendix.

2. The Framework and Some Examples

We consider a situation where a data sample of size n is drawn on the random variable

$$X = (W, Y) \sim P \in \mathcal{P}$$

where $W = (U, V)$ is a $d_w = d_u + d_v$ vector ($U \in \mathbf{R}^{d_u}$, $V \in \mathbf{R}^{d_v}$) and Y is a scalar. Our primary focus is on regression structure, so we take

$$\mathcal{P} = \left\{ \text{all probability measures } P \text{ on } \mathbf{R}^{d_w+1} : E_P |X|^2 < \infty \right\}.$$

where the second absolute moment condition assures the existence of conditional expectations. We are interested in testing whether the population distribution P lies in a proper subset $\mathcal{P}_0 \subset \mathcal{P}$, namely

$$H : P \in \mathcal{P}_0.$$

While we retain this general framework for our analysis, some examples help to add some concreteness to the kinds of applications we discuss in detail. We begin with

Example 2.1 (Dimension Reduction in Regression). *For $P \in \mathcal{P}_0$, we have that*

$$E_P(Y|W) = E_P(Y|U)$$

Therefore, the test amounts to a test of whether the variables V have significant impacts on the mean of Y.

Example 2.2 (Dimension Reduction with Parameters). *For $P \in \mathcal{P}_0$, we have that*

$$E_P(Y|W) = E_P(Y|I(\theta))$$

where

$$\theta \in \Theta \subset \mathbf{R}^s,\ I(\theta) \equiv I(W, \theta) : \mathbf{R}^{d_w} \times \mathbf{R}^s \to \mathbf{R}^k$$

A familiar special case of this is the index model, *where $k = 1$ and $I(\theta)$ is a weighted linear combination of the predictors:*

$$I(\theta) = \sum_{j=1}^{d_w} \theta_j W_j$$

This type of model is a direct generalization of a linear regression model, with its practical appeal derived from its parsimonious representation of multivariate effects. This model also arises as the exploratory method of one-dimensional projection pursuit*, and as a generalization of GLIM models.*

A second category of examples focus on the distribution of Y around its conditional mean. These cases focus on the departure of Y from its conditional mean: write Y as its mean plus departure from mean as

$$Y = E(Y|W) + \varepsilon \cdot \sigma(W).$$

Some important practical examples are

Example 2.3 (Pure Heteroskedasticity). *ε is distributed independently of W.*

Example 2.4 (Pure Homoskedasticity). *ε is distributed independently of W and $\sigma(W) = 1$.*

Example 2.5 (Normal Heteroskedasticity). *$\varepsilon \sim \mathcal{N}(0, 1)$.*

These examples cover traditional issues of mean and variance modeling in regression analysis. The first category of examples (2.1 and 2.2) involve questions on the appropriate way to specify a regression function to capture empirical relationships, in particular, whether certain predictor variables can be omitted from the analysis or whether the impacts of several predictor variables can be adequately captured by an index. The second category of examples (2.3, 2.4 and 2.5) involve questions of how to specify the spread of the response around its conditional mean. This is the central issue in some empirical applications, such as the study of volatility of prices of financial securities, where variance estimates are a central ingredient for pricing options or other derivatives. Alternatively, variance estimates are needed for efficient weighting of data for the estimation of mean regression functions. These and many other examples are discussed in Bickel, Klaassen, Ritov and Wellner (1993) and Stoker (1991) among many others.

The question of interest to us is, given an observed sample $X_1, \ldots, X_n$, how can one test $H : P \in \mathcal{P}_0$?

3. Some Testing Heuristics

3.1. Simple Null Hypothesis

Suppose that $X_1, \ldots, X_n$ is a (i.i.d) random sample from the probability $P \in \mathcal{Q}$, where $P \ll \mu$, $p = \partial P / \partial \mu$. Suppose that $\mathcal{Q} = \{P_\theta : \theta \in \mathbf{R}\}$ is a regular (one-dimensional) submodel of probabilities. Consider testing the hypothesis

$$H : P = P_0$$

against

$$K : P = P_\theta \text{ where } \theta > 0.$$

Denote

$$\ell(\cdot, P) \equiv \ln p, \ \ell_1(\cdot, P_0) \equiv \frac{\partial \ell(\cdot, P_0)}{\partial \theta}$$

where $\ell_1 \in L_2(P_0)$, $E_0(\ell_1(\cdot, P_0)) = 0$. *Score test statistics* refer to measures of the size of the 'scores' $\ell_1(\cdot, P_0)$ computed from the sample. The familiar scoring test of Rao (1947) is the sample analogue of the mean of the scores, with

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_1(X_i, P_0) \to \mathcal{N}\left(0, \|\ell_1\|_0^2\right)$$

with an asymptotic test based on the critical value

$$\zeta_{1-\alpha} \|\ell_1\|_0,$$

where $\|h\|_0^2 = \int h^2 dP_0$. Note that this test is consistent if and only if

$$E_\theta(\ell_1(X, P_0)) \neq 0 \text{ for } \theta > 0;$$

namely if a nonzero θ implies a positive mean score.

3.2. Composite Null Hypothesis

The case of a simple null hypothesis is indeed the simplest type of testing situation; the model under the null is specified by setting $\theta = 0$ and the alternatives (within $\mathcal{Q}$) are associated with nonzero values $\theta > 0$. If the null set $\mathcal{P}_0$ is composite, then the question arises as to how to formulate and compute scores that depend on $P \in \mathcal{P}_0$, or $\ell_1(\cdot, P)$. In particular, one needs two ingredients:

(i) A "consistent" estimate $\hat{P}_n$ of P.

(ii) The *tangent space* $\dot{\mathcal{P}}_0(P_0)$, $P_0 \in \mathcal{P}_0$. $\dot{\mathcal{P}}_0(P_0)$ is the linear closure (in $L_2(P)$) of the set of all score functions of regular one-dimensional submodels in $\mathcal{P}_0$ through P_0.

The space $\dot{\mathcal{P}}_0(P_0)$ captures the directions of variation from P_0 that are consistent with the null hypothesis of interest. Therefore, the effective direction of interest for the alternative $\mathcal{Q}$ is given by *efficient score function*

$$\ell_1^*(\cdot, P_0) \equiv \ell_1(\cdot, P_0) - \Pi\left(\ell_1 | \dot{\mathcal{P}}_0(P_0)\right)$$

where Π denotes the projection operator in $L_2(P_0)$.

A natural score test statistic is then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_1^*\left(X_i, \hat{P}_n\right)$$

with asymptotic tests based on the critical value

$$\zeta_{1-\alpha} I\left(\hat{P}_n\right)^{1/2}$$

where

$$I(P) \equiv E_P\left(\ell_1^*\left(X, \hat{P}_n\right)^2\right)$$

is the variance of the efficient score. This statistic is analyzed by Choi, Hall and Schick (1996).

For concreteness, we examine how these concepts arise in traditional tests of parametric models. Suppose that the general family is defined as

$$\mathcal{P} = \left\{P_{(\eta,\theta)} : \eta \in \mathbf{R}^p,\ \theta \in \mathbf{R}\right\}$$

so that the general log-density takes the form $\ell \equiv \ell(\eta, \theta)$.

The restricted family $\mathcal{P}_0$ is associated with the null hypothesis

$$H : \theta = 0$$

so that

$$\mathcal{P}_0 \equiv \left\{P_0 : P_0 = P_{(\eta_0,0)}, \eta_0 \in \mathbf{R}^p\right\}.$$

The score vector of an alternative direction is

$$\ell_1 \equiv \frac{\partial \ell\left(\eta_0, 0\right)}{\partial \theta}.$$

and the tangent space associated with the null hypothesis is

$$\dot{\mathcal{P}}_0\left(P_0\right) = \text{ Linear Span } \left\{ \frac{\partial \ell\left(\eta_0, 0\right)}{\partial \eta_j} : 1 \leq j \leq p \right\}.$$

Therefore, the efficient score is

$$\ell_1^* = \frac{\partial \ell\left(\eta_0, 0\right)}{\partial \theta} - \sum_{j=1}^{p} \left[a_j\left(\eta_0, 0\right) \frac{\partial \ell\left(\eta_0, 0\right)}{\partial \eta_j} \right]$$

where the a_j's are projection (least squares) weights; namely $\{a_j\left(\eta_0, 0\right)\}$ minimize

$$\left\| \ell_1 - \sum_{j=1}^{p} \left[a_j\left(\eta_0, 0\right) \frac{\partial \ell\left(\eta_0, 0\right)}{\partial \eta_j} \right] \right\|_0^2.$$

If $\hat{\eta}$ is a consistent estimator of η under H, then the Neyman $C\left(\alpha\right)$ test statistic is formed

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_1^*\left(X_i, \hat{\eta}, 0\right)$$

For instance, $\hat{\eta}$ can be taken as $\hat{\eta}_H$, the maximum likelihood estimator (MLE) of η under H. In that case

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_1\left(X_i, \hat{\eta}_H, 0\right)$$

since $\hat{\eta}_H$ solves the likelihood equations:

$$0 = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \eta_j}\left(X_i, \hat{\eta}_H, 0\right).$$

In fact, under additional smoothness conditions on $\ell(\cdot, \eta, \theta)$, essentially any estimate $\hat{\eta}$ with $\hat{\eta} = \eta_0 + o_p\left(n^{-1/4}\right)$ (under $P_{(\eta_0,0)}$) may be substituted into ℓ_1^*. Essentially any efficient estimator $\hat{\eta}$ may be substituted into ℓ_1.

3.3. Composite Alternatives

We now discuss how to put score tests of one-dimensional alternatives together for testing composite alternatives. Consider the situation where the null hypothesis is simple:

$$\mathcal{P}_0 = \{P_0\}$$

and let

$$\mathcal{Q} = \{P_\theta : \theta \in \mathbf{R}\}$$

denote a regular model through P_0. Clearly, the score $\ell_1 \equiv h_{\mathcal{Q}}$ depends on the $\mathcal{Q}$ direction. A composite alternative $\mathcal{P}$ is the union of several such $\mathcal{Q}$'s. The relevant set of (alternative) scores is the tangent space $\dot{\mathcal{P}}(P_0)$ defined as the linear closure of all the associated scores $h_{\mathcal{Q}}$.

One way of combining the score tests associated with different one-dimensional submodels is classical. Consider the situation of a finite number of submodels (directions), with individual scores denoted $h_1, \ldots, h_p$. Begin by assuming that the scores are mutually orthogonal under P_0, i.e. $(h_a, h_b)_0 = \delta_{ab}$ for $a, b = 1, \ldots, p$. In this case it is easy to verify that

$$T = \sum_{j=1}^{k} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i) \right)^2 \to \chi^2_p$$

under H, as would be expected.

When the scores from different submodels are correlated, then their covariance structure can be diagonalized in the standard way to produce an analogous test statistic. That is, suppose that

$$\|(h_a, h_b)_0\|^2 = \Sigma_0$$

and Σ_0 is nonsingular, then the natural test statistic is

$$T = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_0^{-1/2} \tilde{h}(X_i) \right\|^2$$

$$= \frac{1}{n} \left[\sum_{i=1}^{n} \tilde{h}(X_i) \right]' \Sigma_0^{-1} \left[\sum_{i=1}^{n} \tilde{h}(X_i) \right] \to \chi^2_p$$

where $\tilde{h}' \equiv (h_1, \ldots, h_p)$.

These formulae arise naturally in the parametric case as follows. If

$$\mathcal{P} = \{P_\theta : \theta \in \mathbf{R}^p\}$$

then the tangent space $\dot{\mathcal{P}}(P_0)$ is the linear closure of $\{\partial\ell/\partial\theta_j : j = 1, \ldots, p\}$, and the test statistic is

$$T = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{\partial \ell(X_i)}{\partial \theta}\right]' \mathcal{I}_0^{-1} \left[\sum_{i=1}^{n} \frac{\partial \ell(X_i)}{\partial \theta}\right]$$

where $\mathcal{I}_0$ is the information matrix.

Score tests of this familiar form fall under the rubric of *Lagrange multiplier* tests in econometrics, which refer to tests that examine departures making use only of estimates of the statistical model under the null hypothesis. In contrast, Wald tests or likelihood ratio tests are based on comparing estimates of the model under alternatives with those of the model estimated under the null. In the parametric context, to first order, both Wald and likelihood ratio tests are equivalent to score based tests.

It is important to note that T is just one way of combining the different test directions. There is nothing magic in the Mahalanobis distance or the likelihood ratio test. The appropriate weighted average should be such that the directions of interest would get more weight than the other directions, so that the test will have power in those directions.

This approach applies equally well to the situation where the alternative hypothesis contains an infinite number of directions, or where $\dot{\mathcal{P}}(P_0)$ has infinite basis. In particular, suppose that the tangent space is the linear closure of the set of directions

$$\{h_\gamma : \gamma \in \mathbf{R}^q\}$$

and μ is a measure on $\mathbf{R}^q$ with dense support. Two forms of statistics arise in analogy to the finite dimensional case, namely the weighted squared average score

$$T_a = \int \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_\gamma(X_i)\right)^2 d\mu(\gamma) \tag{3.1}$$

and the maximum average score

$$T_b = \sup_\gamma \left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_\gamma(X_i)\right|. \tag{3.2}$$

Some familiar statistics from density estimation, as well as classical nonparametric goodness-of-fit tests, fall into this category of score tests.

Example 3.1 (Goodness-of-Fit Statistics). *Consider testing the null hypothesis that a distribution on* $\mathbf{R}$ *is* P_0 *against "all" alternatives, namely where*

$$\dot{\mathcal{P}}(P_0) = \{h \in L_2(P_0) : E_0[h(x)] = 0\}$$

If we consider the family of directions

$$h_\gamma(\cdot) = 1(\cdot \leq \gamma) - F_0(\gamma);\ \gamma \in \mathbf{R}$$

where F_0 *is the cumulative distribution function of* P_0*, then the following two statistics arise in association with those above. Associated with (3.1) is the familiar* Cramer-Von Mises *goodness-of-fit statistic*

$$T_a = n \int (F_n(\gamma) - F_0(\gamma))^2 \, dF_0(\gamma)$$

where F_n *is the empirical distribution function, and the weighting measure is* $\mu = F_0$*. Corresponding to (3.2) is the familiar* Kolmogorov-Smirnov *goodness-of-fit test statistic*

$$T_b = \sup_\gamma \left| \sqrt{n} (F_n(\gamma) - F_0(\gamma)) \right|.$$

This approach is also known as the union-intersection test (UIT).

It is easy to extend these ideas to situations where the alternative directions depend on the size of the sample and each alternative approximates departure in the density more and more finely as n increases. For instance, if the tangent space is the linear closure of a finite set of such directions $\{h_{jn} : 1 \leq j \leq k_n\}$, then $\chi^2_{k_n}$ statistics are derived as above. If the number of directions is unlimited as in $\{h_{\gamma_n} : \gamma_n \in \mathbf{R}^q\}$, then test statistics of the Bickel and Rosenblatt (1973) form arise.

4. Score Tests and the Method of Sieves

4.1. Testing Paradigm

The above ideas motivate our general testing paradigm. For a simple null hypothesis, we consider a basic score statistic of the form

$$Z_n(h) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i)$$

where h is a direction in the tangent space

$$h \in \dot{\mathcal{P}}\left(P_0\right)$$

Our overall approach is to build test statistics using $Z_n\left(h\right)$.

For composite hypotheses, consider the space of alternatives

$$\mathcal{P} = \left\{P_{(\alpha,\beta)} : \alpha \in A, \beta \in B\right\}$$

where A, B are function spaces. The null hypothesis of interest is

$$H : \beta = 0$$

Define the "full", "null" and "alternative" tangent spaces

$$\begin{aligned}
\dot{\mathcal{P}}\left(\alpha,\beta\right) &= \text{Tangent Space of the Model } \mathcal{P} \text{ at } P_{(\alpha,\beta)} \\
\dot{\mathcal{P}}_0\left(\alpha,0\right) &= \text{Tangent Space of } \left\{P_{(\alpha,0)} : \alpha \in A\right\} \text{ at } P_{(\alpha,0)} \\
\dot{\mathcal{P}}_0^{\perp}\left(\alpha,0\right) &= \text{Orthogonal Complement of } \dot{\mathcal{P}}_0\left(\alpha,0\right) \text{ in } \dot{\mathcal{P}}\left(\alpha,0\right)
\end{aligned}$$

That is, $\dot{\mathcal{P}}_0\left(\alpha,0\right) \perp \dot{\mathcal{P}}_0^{\perp}\left(\alpha,0\right)$ and $\dot{\mathcal{P}}\left(\alpha,0\right) = \dot{\mathcal{P}}_0\left(\alpha,0\right) \oplus \dot{\mathcal{P}}_0^{\perp}\left(\alpha,0\right)$. Suppose $\dot{\mathcal{P}}_0^{\perp}\left(\alpha,0\right)$ is representable as the image of a fixed Hilbert space $\mathcal{H}$ under $\Pi\left(\cdot,\alpha\right) : \mathcal{H} \rightarrow \dot{\mathcal{P}}_0^{\perp}\left(\alpha,0\right)$. Define the score process by

$$Z_n\left(h,\alpha\right) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Pi\left(h,\alpha\right)\left(X_i\right).$$

Now, if $P_{(\hat{\alpha},0)}$ is a "consistent" estimate of P_0 under H, then we form a proxy for $Z_n\left(h,\alpha\right)$ by

$$\hat{Z}_n\left(h\right) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^*\left(X_i,\hat{\alpha}\right)$$

and build test statistics using this proxy.

We think of $Z_n\left(\cdot\right)$ and $\hat{Z}_n\left(\cdot\right)$ as stochastic processes defined on $\mathcal{H}$ in analogy to empirical processes — see Pollard (1989), for instance.

4.2. Sieve Approximation

As with approximation of general functions by truncated power series, we use the concept of sieves to provide approximation to "large" spaces of alternatives for the purpose of testing. In particular, we define a sieve for $\mathcal{P}$ as a class of sets

$$\mathcal{Q}_k \equiv \{Q_\theta : \theta \in \Theta_k\}, \ k = 1, \ldots, \infty$$

where the Q_θ's are probability measures satisfying the following conditions:

$$(i) \quad \mathcal{Q}_k \subset \mathcal{Q}_{k+1}, \ k = 1, \ldots, \infty$$

$$(ii) \quad \mathcal{Q}_k \text{ is a regular parametric family}$$

$$(iii) \quad \lim_{k\to\infty} \mathcal{Q}_k = \mathcal{P}$$

where (iii) is to be understood in the sense of a projective weak limit, i.e. if $P \in \mathcal{P}$ there exist $Q_k \in \mathcal{Q}_k$ such that $Q_k \to P$ in law.

These conditions capture the notion that the $\mathcal{Q}_k$ sets provide an increasingly rich set of probability measures to approximate the set of alternatives. On technical grounds, (i) is not necessary for what follows, (ii) is applied only to assure that procedures such as maximum likelihood estimation are well behaved in $\mathcal{Q}_k$, and (iii) need only hold in a weak sense to ensure that these estimators converge. The typical framework, though not necessarily needed in our case, will have a sieve defined with an explicit nesting: $\Theta_k = \mathbf{R}^k$, and for elements $Q_{(\theta_1,\theta_2,\ldots,\theta_k)} \in \mathcal{Q}_k$ there exist corresponding elements $Q_{(\theta_1,\theta_2,\ldots,\theta_k,0)} \in \mathcal{Q}_{k+1}$; again as an analogue to series expansion of functions.

Our results apply to the practical methods of estimation and testing using a set $\mathcal{Q}_k$ in place of the true alternative set $\mathcal{P}$. Specifically, we employ the following

Sieve Estimation Principle: Given $X_1, \ldots, X_n$, choose $\hat{k}\,(X_1, \ldots, X_n)$ and act as if $\mathcal{Q}_{\hat{k}}$ were the true family of alternatives. For instance, if $\hat{\theta}^{\hat{k}}$ is the maximum likelihood estimator of θ in $\mathcal{Q}_{\hat{k}}$, then we estimate $P \in \mathcal{P}$ by $Q_{\hat{\theta}^{\hat{k}}} \in \mathcal{Q}_{\hat{k}}$.

5. Index Models

5.1. Introduction

Again assume that $X = (W, Y)$ where $W = (U, V)$, $U \in \mathbf{R}^{d_u}$, $V \in \mathbf{R}^{d_v}$ and $Y \in \mathbf{R}$. Suppose that $\int yf(W, y)\, dy = 0$ a.s., and $\int \left(|w|^2 + y^2\right) f(w, y)\, dy\, dw < \infty$. Let $p(w, y; f, \nu) = f(w, y - \nu(w))$ for some function $\nu(\cdot)$ such that $\int \nu^2(w) f(w, y)\, dy\, dw < \infty$. Let $\mathcal{P}$ be the collection of all distribution functions with such a density (i.e. for all possible f and ν). Finally, let H_0 be the hypothesis that $\nu(U, V) = \nu(U)$ almost surely where the ν on the left hand side maps $R^{d_u+d_v}$ to R while that on the right maps R^{d_u} to R. That is $E(Y|W) = E(Y|U)$.

The tangent spaces are easy to characterize as shown in Newey (1990) and Bierens and Ploberger(1997), among others. The following lemma is proved for completeness.

Lemma 5.1. *We have*

$$\dot{\mathcal{P}} = \left\{a(W, Y) : E_P\left[a^2(W, Y)\right] < \infty, E_P[a(W, Y)] = 0\right\}$$

$$\begin{aligned}
\dot{\mathcal{P}}_0 &= \left\{a(W, Y) = h(W, Y - \nu(U)) + \ell'_{Y|W}(Y - \nu(U))\, g(U) : \right. \\
&\quad \left. a, h \in \dot{\mathcal{P}},\ \int yh(W, y)\, dy = 0,\ a.s.\right\} \\
\dot{\mathcal{P}}_0^{\perp} &= \left\{a(W, Y) = [b(W) - E(b(W)|U)](Y - E(Y|U)) : a, b \in \dot{\mathcal{P}}\right\}.
\end{aligned}$$

where $\ell'_{Y|W}(y|w)$ is the derivative of the conditional log-likelihood of Y given W at (y, w).

Proof. Since the "large" space is unrestricted, $\dot{\mathcal{P}}$ is "everything," but with the moment conditions. The structure of $\dot{\mathcal{P}}_0$ is obtained by considering the derivative of the general one-dimensional submodel $p_t(w, y) = f_t(w, y - \nu(u) + tg(u))$, where $h = f'_t/f_t\big|_{t=0}$. Finally, $\dot{\mathcal{P}}_0^{\perp}$ is the orthocomplement of $\dot{\mathcal{P}}_0$ in $\dot{\mathcal{P}}$. But $a(W, Y)$ is orthonormal to

$$\left\{h(W, Y - \nu(U)), \int yh(W, y)\, dy = 0\ a.s.\right\}$$

if and only if $a(W, Y) = b(W)(Y - \nu(U))$, a.s. This latter object is orthogonal to all functions in $\dot{\mathcal{P}}$ of the form $\ell'_y(W, Y - \nu(U))\, g(U)$ if and only if $E(b(W)|U) =$

0 *a.s.* which follows from the fact that for any p.d.f. q (with mean 0), we have $\int xq'(x)\,dx = -1$. □

Therefore, our *score process* is defined by

$$\hat{Z}_n(a) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[a(W_i) - E_{\hat{P}}(a(W)|U_i)\right](Y_i - E_{\hat{P}}(Y|U_i)) \tag{5.1}$$

where the estimator $\hat{P}$ is yet to be defined.

Let $\mathcal{B} = \mathcal{B}_n = \{\mathcal{B}_{n1}, \ldots, \mathcal{B}_{nK_n}\}$ be a partition of the range of U. Our sieve estimator $\hat{P}$ is the nonparametric maximum likelihood estimator (NPMLE) of P, for the family of distributions such that the conditional distribution of $W|U = u$ is the same for all $u \in \mathcal{B}_{nj}$, $j = 1, \ldots, K_n$. Thus, $\nu(U) \equiv c_j$ on $\mathcal{B}_{nj}$. The "natural" NPMLE $\hat{P}$ is easily seen to concentrate on $\{(u, w, y) : u = U_i,\ (w, y) = (W_j, Y_j)$ for some $1 \leq i, j \leq n\}$ and be given by,

$$\hat{P}[U = U_i] = \frac{1}{n} \sum_{k=1}^{n} 1(U_k \in \mathcal{B}_{nj})$$

$$\hat{P}[(V, Y)|U = U_i] = \frac{\sum_{k=1}^{n} 1(U_k \in \mathcal{B}_{nj}) 1((V_k, Y_k) = (V, Y))}{\sum_{k=1}^{n} 1(U_k \in \mathcal{B}_{nj})}$$

if $U_i \in \mathcal{B}_{nj}$, $1 \leq j \leq K_n$. This leads to

$$E_{\hat{P}}(a(W)|U) = \sum_{j=1}^{K_n} 1(U \in \mathcal{B}_{nj}) \frac{\sum_{i=1}^{n} a(W_i) 1[U_i \in \mathcal{B}_{nj}]}{\sum_{i=1}^{n} 1[U_i \in \mathcal{B}_{nj}]} \tag{5.2}$$

$$E_{\hat{P}}(Y|U) = \sum_{j=1}^{K_n} 1(U \in \mathcal{B}_{nj}) \frac{\sum_{i=1}^{n} Y_i 1[U_i \in \mathcal{B}_{nj}]}{\sum_{i=1}^{n} 1[U_i \in \mathcal{B}_{nj}]} \tag{5.3}$$

Given the score process we can construct several different test statistics.

1. Cramer-von Mises type test statistics: Let

$$a_\gamma(w) = 1(w \leq \gamma),\ \gamma \in \mathbf{R}^{d_u + d_v}$$

Then,

$$T = \int \hat{Z}_n^2(a_\gamma)\, d\mu(\gamma), \tag{5.4}$$

where μ is Lebesgue measure, gives an analogue to traditional Cramer-von Mises statistics. If both U and V are real, it may make sense to consider $\mathcal{B}_{nj}$ to be intervals for $\hat{Z}_n$, and to consider discrete μ with atoms at the intervals' ends instead of Lebesgue measure.

2. Kolmogorov-Smirnov type test statistics: With the same set $\{a_\gamma\}$, consider the test statistics
$$T = \sup_\gamma \left| \hat{Z}_n(\gamma) \right|, \tag{5.5}$$
where the 'sup' can be unrestricted or limited to a finite grid.

3. χ^2 type test statistics: We can define $a_{\gamma,\delta}$ to be a function centered at γ concentrated as $\delta \searrow 0$. For example, $a_{\gamma,\delta}(\cdot) = (1/\delta)\, a_0((\cdot - \gamma)/\delta)$, where a_0 may be any indicator function, or any other window function. Then one can consider the test statistic:
$$T = \int \hat{Z}_n^2(a_{\gamma,\delta_n})\, d\gamma \tag{5.6}$$
where $\delta_n \searrow 0$ as is done in Aït-Sahalia, Bickel and Stoker (1998) and Hart (1997), Bickel and Ritov (1992) and Fan (1996).

4. Bickel-Rosenblatt type statistics: In analogy to the relation between the "Cramer-von Mises type" statistics and the "χ^2 type" statistics we can consider
$$T = \sup_K |\hat{Z}_n(a_{\gamma,\delta_n})| \tag{5.7}$$
where K is a compact set. In the goodness-of-fit problem these correspond to the statistics proposed by Bickel and Rosenblatt (1973).

The first two approaches are appropriate when one wants to focus on particular departures, such as departures from the null hypothesis of the type of a change point in the conditional distribution. The last two approaches, on the other hand, are relevant when comparable interest is attached to all equi-distant local departures from the assumptions. Further, the first two types of tests can be designed to have non-trivial power against any $\sqrt{n}$ alternative (i.e. if we consider a sequence of problems in which $E(Y|W=w) = \nu(u) + n^{-1/2} g(w)$). The price for this is having considerable power against only a few directions, and very little power against all other possible departures from the null assumption. In contrast,

the third (χ^2) type test and fourth Bickel-Rosenblatt test has asymptotically no power against $\sqrt{n}$ alternatives, but has comparable power in all directions. In short, the choice between these types of testing approaches is a choice between concentrating the testing power in a specific way versus diffusing the power. One further observation is that the Kolmogorov-Smirnov type or Bickel-Rosenblatt type tests have an additional benefit: when the null hypothesis is rejected, they can identify the direction of departure.

We can consider many different sets of directions that yield different tests. Thus we could consider a variation of the above family by scaling the functions: $a_\gamma(w) = a(\gamma)\,1(w \leq \gamma)$. This scaling gives more weight to some directions (where $a(\gamma)$ is large), thus making the tests have more power in those directions. On the other hand, such tests would behave poorly in directions where the weights $a(\gamma)$ are small. We can change the family to the larger family of $a_{\gamma,\kappa}(w) = 1(\gamma < w \leq \kappa)$. Compared with the original test, this test will have more power against some alternatives (i.e., those which are characterized by a different behavior in the middle of the range), but less power against other alternatives (in particular, if the alternative is in the direction of a_γ for γ in the middle of the range).

We highlight these aspects of testing direction and power in simulations presented in Section 9. First, however, we consider the basic asymptotic properties of the tests as well as bootstrap approximation for critical values.

6. Asymptotics for the Score Process

We turn to the basic asymptotics which will justify the inferential results for Kolmogorov-Smirnov and Cramer-von Mises type statistics claimed in the next section. The third and fourth types of statistics ("χ^2" and Bickel-Rosenblatt) involve new technical problems that we will deal with in a subsequent paper. The Aït-Sahalia, Bickel, Stoker (1998) paper deals with a special case of the χ^2 type.

We now give a general approach to proving convergence and to determining limiting distributions for these statistics as well as a wide range of others.

6.1. The Score Process Under H

We begin by establishing asymptotic normality and tightness. The required notation and assumptions are as follows. Let $\mathcal{B}_n = \{\mathcal{B}_{n1}, \ldots, \mathcal{B}_{nK_n}\}$ be a partition of

the support of U. Let $M_a(U) = E(a(U,V)|U)$ and $\bar{M}_a^{(n)}(U) = E(a(U,V)|\mathcal{B}_n)$, the conditional expectation given which $\mathcal{B}_{nj}$ corresponds to U. Thus

$$\bar{M}_a^{(n)}(U) = \sum_{j=1}^{K_n} 1(U \in \mathcal{B}_{nj}) E(a(W)|U \in \mathcal{B}_{nj}). \tag{6.1}$$

Let $\xi_k \equiv \xi_{nk}$ be a point in $\mathcal{B}_{nk}$; for example the center if $\mathcal{B}_{nk}$ is a rectangular block. Finally, let $\mathcal{A}$ be a family of bounded functions. We now make the following assumptions

[A1] Let $\varepsilon = Y - \nu(U)$. Then, $Var(\varepsilon|U,V) \leq \sigma^2$.

[A2] Suppose $K_n \to \infty$, $K_n/n \to 0$. Let $m_{nk} = \sum_{i=1}^n 1_{\mathcal{B}_{nk}}(U_i)$. Then $m_n^* = \min_k m_{nk} \xrightarrow{P} \infty$ and $\max_k m_{nk}/n \xrightarrow{P} 0$.

[A3] Let $\Lambda_n = \sup\{|\nu(u) - \nu(u')| : u, u' \in \mathcal{B}_{nj},\ 1 \leq j \leq K_n\}$. Then $\Lambda_n \to 0$. Let $\gamma_n = \sup\{|a(u) - a(u')| : a \in \mathcal{A},\ u, u' \in \mathcal{B}_{nj},\ 1 \leq j \leq K_n\}$. Then, $\gamma_n \to 0$. Let $\Delta_n = \sup\{|\nu(u) - \nu(u')||\bar{M}_a^{(n)}(u) - \bar{M}_a^{(n)}(u')| : u, u' \in \mathcal{B}_{nj},\ 1 \leq j \leq K_n,\ a \in \mathcal{A}\}$. Then, $\Delta_n n^{1/2} \to 0$.

[A4] Given a metric d on $\mathcal{A}$ let $\log N_{[\,]}(\varepsilon, \mathcal{A}, d)$ be the bracketing entropy with respect to d. That is, $N_{[\,]}(\varepsilon, \mathcal{A}, d)$ is the minimal cardinality of a set of brackets $[a_\ell, a_\ell']$, $1 \leq \ell \leq N_{[\,]}$ such that $d(a_\ell, a_\ell') \leq \varepsilon$ and for every $a \in \mathcal{A}$, there exists ℓ such that $d(a_\ell, a_\ell') \leq \varepsilon$. With $\|\cdot\|_2$ the $L_2(P)$ norm on $\mathcal{A}$, assume that

$$N_{[\,]}(\varepsilon, \mathcal{A}, \|\cdot\|_2) \leq C\varepsilon^{-\beta}$$

for all $\varepsilon > 0$ some $C, \beta > 0$. Note that then

$$\int_0^1 \sqrt{\log N_{[\,]}(\varepsilon^r, \mathcal{A}, \|\cdot\|_2)}\, d\varepsilon < \infty \text{ for all } 0 \leq r < \infty. \tag{6.2}$$

Remark. We say that a function f satisfies the Lipschitz condition of order α if for some ε, $\sup\{f(x+y) - f(x)/\|y\|^\alpha : x, 0 < \|y\| < \varepsilon\} < \infty$. Suppose that the support of U is compact, and U has density bounded away from 0. Assumption [A3] is natural if ν is Lipschitz of order β and M_a is Lipschitz of order β, where $\alpha + \beta > 1/2$. We can then take the blocks to have a width of order δ_n, $\delta_n \to 0$, $n\delta_n \to \infty$, $\delta_n^{\alpha+\beta} n^{1/2} \to 0$. Thus both block assumptions [A2] and [A3] are satisfied.

If $\mathcal{A}$ is such that $a(u,\cdot) = a\left(u',\cdot\right)$ for any u and u' in the same block, then the above condition is essentially a condition on how close the conditional density of U given V and $U \in \mathcal{B}_{nk}$ is to uniform:

$$\begin{aligned} M_a(u) - M_a\left(u'\right) &= \int \left(a(u,v) f(v|u) - a\left(u',v\right) f\left(v|u'\right)\right) dv \\ &= f(v) \int a(u,v) \left(\frac{f(u|v)}{f(u)} - \frac{f\left(u'|v\right)}{f\left(u'\right)}\right) dv. \end{aligned}$$

Similarly, if $\mathcal{A}$ is a class of indicators of quadrant, or hypercubes as we discussed [A4] is guaranteed by smoothness of the conditional density of V given U in both arguments.

Other metrics including random ones will prove important. In particular for $\mathcal{B}_n$ as above, let $D_n^2(a,a') \equiv (1/n)\sum_{i=1}^n E\left[(a-a')^2(W_i) \mid \mathcal{B}_n\right]$ where $W_1,\ldots,W_n$ are i.i.d. In the Appendix, we show that

Lemma 6.1. *If [A2] and [A4] hold then*

$$P\left[\int_0^1 \sqrt{\log N_{[\ ]}(\varepsilon,\mathcal{A},D_n)}d\varepsilon < \infty\right] \to 1. \tag{6.3}$$

This lemma permits us to demonstrate tightness, as in

Theorem 6.2. *Suppose assumptions [A1]–[A4] are satisfied and $\nu(U,V) = \nu(U)$. Then the process $\left\{\hat{Z}_n(a) : a \in \mathcal{A}\right\}$ is tight. Moreover, let*

$$\tilde{Z}_n(a) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(a(W_i) - \bar{M}_a(U_i)\right)\varepsilon_i \tag{6.4}$$

where, as above, $\varepsilon_i = Y_i - E(Y_i|U_i)$. Then

$$\sup_{a\in\mathcal{A}} \left|\hat{Z}_n(a) - \tilde{Z}_n(a)\right| \xrightarrow{P} 0.$$

Thus,

$$\hat{Z}_n(\cdot) \Rightarrow Z(\cdot) \tag{6.5}$$

where $Z(\cdot)$ is Gaussian mean 0 with

$$cov\left(Z(a), Z(a')\right) = E\varepsilon_1^2 (a(W) - M_a(U))(a'(W) - M_{a'}(U)).$$

The structure of the proof is seen as follows. Let $\hat{M}_a^{(n)}(u) = E_{\hat{P}}(a(W)|U=u)$, where the right-hand side is defined by (5.2). Since

$$\sum_{i=1}^{n} b(U_i)(a(W_i) - \hat{M}_a^{(n)}(U_i)) = 0$$

if b is constant on $\mathcal{B}_{nj}$, $1 \le j \le K_n$ we have the following decomposition

$$\begin{aligned}
\hat{Z}_n(a) &= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(W_i) - \hat{M}_a^{(n)}(U_i))(Y_i - E_{\hat{P}}(Y|U_i)) \\
&= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(W_i) - \hat{M}_a^{(n)}(U_i))Y_i \\
&= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(W_i) - M_a(U_i))\varepsilon_i \\
&+ \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (M_a(U_i) - \bar{M}_a^{(n)}(U_i))\varepsilon_i \\
&+ \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (\bar{M}_a^{(n)}(U_i) - \hat{M}_a^{(n)}(U_i))\varepsilon_i \\
&+ \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(W_i) - \hat{M}_a^{(n)}(U_i))\nu(U_i).
\end{aligned}$$

The proof of this theorem follows from Lemmas 6.3–6.6. In Lemma 6.3 we establish the Donsker property of $\{(a - M_a(\cdot))\varepsilon : a \in \mathcal{A}\}$ and hence by Theorem 2.11.9 of van der Vaart and Wellner (1996),[1] the central limit theorem claimed in (6.5) for $\tilde{Z}_n$. In Lemma 6.4–6.6 we show that all the remaining three terms are negligible. The proofs of the last four lemmas are given in the Appendix.

Lemma 6.3. *The class $\mathcal{D} \equiv \{(a - M_a(\cdot))\varepsilon : a \in \mathcal{A}\}$ satisfies the conditions of Theorem 2.11.9 of van der Vaart and Wellner (1996).*

Proof. We need only check that

$$E(a - M_a)(W)^2\varepsilon^2 \le \|a\|_2^2\sigma^2$$

which is obvious. The lemma then follows from [A4], $\|a\|_\infty \le M$ for all $a \in \mathcal{A}$, and [A1]. □

Lemma 6.4. *Under [A3], [A4]*

$$\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(M_a(U_i) - \bar{M}_a^{(n)}(U_i))\varepsilon_i\right| : a \in \mathcal{A}\right\} \xrightarrow{P} 0. \tag{6.6}$$

[1]Theorem 2.11.9 is stated in the Appendix.

Lemma 6.5. *Under [A1], [A2] and [A4],*

$$\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{M}_a^{(n)}(U_i)-\bar{M}_a^{(n)}(U_i))\varepsilon_i\right| : a\in\mathcal{A}\right\}\xrightarrow{P}0. \tag{6.7}$$

Lemma 6.6. *Under [A1]–[A4],*

$$\sup_{a\in\mathcal{A}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a(U_i,V_i)-\hat{M}_a^{(n)}(U_i))\nu(U_i)\xrightarrow{P}0.$$

6.2. The Score Process Generally

Suppose, $E(Y|W)\neq E(Y|U)$. Intuitively, $\hat{Z}_n(a)$ should diverge. More precisely let,

$$\tilde{Z}_{ng}(a)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a(W_i)-M_a(U_i))\varepsilon_{ig}$$

where $\varepsilon_{ig}=Y_i-E(Y_i|W_i)$ denotes the 'general' error. Evidently $\tilde{Z}_{ng}$ and $\tilde{Z}_n$ coincide under H. We need a trivial strengthening of assumptions

[A1g] Var $(\varepsilon_{1g}|W_1)\leq\sigma^2$,

and an assumption which as we shall see comes from a possibly unsatisfactory definition of $\hat{M}_a^{(n)}(U_i)$.

[A2g] In addition to [A2] suppose that

$$\frac{K_n}{\sqrt{n}}\to 0.$$

Then

Theorem 6.7. *Suppose [A1g], [A2]–[A4] hold and $\nu(U,V)$ is uniformly bounded and hence so is $\nu(U)$. Then,*

$$\sup\left\{\left|\hat{Z}_n(a)-\tilde{Z}_{ng}(a)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a(W_i)-M_a(U_i))(\nu(U_i)-\nu(U_i,V_i))\right| : a\in\mathcal{A}\right\}\xrightarrow{P}0. \tag{6.8}$$

In fact,

$$\hat{Z}_n(a) - \sqrt{n}E(a(W_1) - M_a(U_1))(\nu(U_1) - \nu(U_1, V_1))$$

is tight and converges weakly to Z_g, a Gaussian process with mean 0 and

$$cov\,(Z_g(a), Z_g(a') = E\left(\varepsilon_{1g}^2(a(W_1) - M_a(U_1))(a'(W_1) - M_{a'}(U_1))\right).$$

The proof of this rests on the identity.

$$\hat{Z}_n(a) = \hat{Z}_{ng}(a) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a(W_i) - \hat{M}_a^{(n)}(U_i))(\nu(U_i) - \nu(U_i, V_i))$$

where $\hat{Z}_{ng}(a)$ is obtained by replacing Y_i by $Y_i - \nu(U_i, V_i) + \nu(U_i)$ in $\hat{Z}_n(a)$.

The first part of Theorem 6.7 will follow from Theorem 6.2 (since $(U, V, Y - \nu(U, V))$ satisfies [A1]–[A4]) provided that we prove

Lemma 6.8. *If [A1g], [A2g], [A3], [A4] hold then*

$$\sup\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\bar{M}_a^{(n)}(U_i) - \hat{M}_a^{(n)}(U_i))(\nu(U_i) - \nu(U_i, V_i)) : a \in \mathcal{A}\right\} \xrightarrow{P} 0. \qquad (6.9)$$

Equation (6.9) is proved in the appendix. Showing the second part of Theorem 6.7 merely requires a check of the tightness of $(1/\sqrt{n})\sum_{i=1}^{n}\{(a(W_i) - \bar{M}_a^{(n)}(U_i))(\nu(U_i) - \nu(U_i, V_i)) - E(a(W) - M_a(U))(\mu(U) - \nu(U, V))\}$. The boundedness of $\nu(U, V)$ and [A4] guarantee the required tightness. □

As we shall see in the proof of Lemma 6.8 the unwanted [A2g] comes only because we have defined $\hat{M}_a^{(n)}$ as we have, rather than the intuitively at least as appealing

$$\hat{M}_a^{(n)-}(u) = \sum_{j=1}^{K_n} 1(u \in \mathcal{B}_{nj})\frac{1}{m_{nj} - 1}\sum\{a(W_k) : U_k \in \mathcal{B}_{nj},\ U_k \neq u\}. \qquad (6.10)$$

7. Inference Under the Null Hypothesis and Contiguous Alternatives

7.1. Theory

Let $S : \ell_\infty \to R$ be a function defined on bounded functions on $\mathcal{A}$, which is continuous with respect to the sup norm on ℓ_∞. Consider $S(\hat{Z}_n(\cdot))$. By Theorem 6.2, under H, if [A1]–[A4] hold we have

$$\mathcal{L}(S(\hat{Z}(\cdot))) \to \mathcal{L}(S(Z(\cdot))).$$

Examples of such functions S are $\sup_{\mathcal{A}} |z(a)|$, and $\int_\Gamma (z(a_\gamma))^2 d\mu(\gamma)$ for μ a finite measure over Γ (with $\gamma \to a_\gamma$ continuous); namely the tests that we have designated as of Kolmogorov-Smirnov or Cramer-von Mises type. Let $c_S(\alpha, F)$ be the α quantile of the distribution of $S(Z(\cdot))$ when F is the distribution of (U, V, Y) and F obeys H.

Suppose we are given consistent estimates $\hat{c}_n$ of $c_S(\alpha, F)$ for F satisfying H. Then we clearly have

$$\overline{\lim} \left\{ P_F[S(\hat{Z}_n(\cdot)) > \hat{c}_n] \leq P_F[S(Z(\cdot)) > c_S(\alpha, F)] \right\} \leq 0. \tag{7.1}$$

On the other hand suppose $\{F^{(n)}\}$ is a sequence of contiguous alternatives to F which are locally asymptotically normal (LAN) at the rate $\frac{1}{\sqrt{n}}$ in the sense of Le Cam

$$\sum_{i=1}^{n} (\log p_{F^{(n)}}(W_i, Y_i) - \log p_F(W_i, Y_i)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, Y_i) - \frac{1}{2} E\psi^2(W_1, X_1) + o_{P_F}(1) \tag{7.2}$$

where $E_F \psi(W_1, Y_1) = 0$, $E_F \psi^2(W_1, Y_1) < \infty$. Contiguity implies that $\hat{c}_n \overset{P_{F_n}}{\to} c_S(\alpha, F)$.

Moreover $\left(\tilde{Z}_n(\cdot), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, Y_i) \right)$ is tight and has a limiting mean 0 Gaussian distribution $(Z(\cdot), N)$, with covariances specified by those of $Z(\cdot)$; Var $\psi(W_1, Y_1)$ and cov $(Z(a), N) = \text{cov}_F((a(W_1) - M_a(U_1))\varepsilon_1, \psi(W_1, Y_1))$. Therefore, by Le Cam's third lemma (c.f. Bickel, Klaasen, Ritov and Wellner (1993)), we have

$$\tilde{Z}_n(\cdot) \overset{F_n}{\Rightarrow} \tilde{Z}(\cdot) \tag{7.3}$$

where $\tilde{Z}(\cdot)$ is a Gaussian process with

$$E\tilde{Z}(a) = \text{cov}_F(a(W_1) - M_a(U_1))\varepsilon_1, \psi(W_1, Y_1)) \tag{7.4}$$

and the same covariance structure as $Z(\cdot)$. Since, by contiguity and Theorem 6.2, (7.3) applies to $\hat{Z}_n$ as well we have proved.

Theorem 7.1. *Suppose F obeying H satisfies [A1]–[A4], F_n is LAN as in (7.2) and S is as specified. Then the test which rejects iff $S(\hat{Z}_n(\cdot)) > \hat{c}_n$ is asymptotically level α and, for all α such that $c_S(\alpha, F)$ is a continuity point of the distribution of $\mathcal{L}_F S(Z(\cdot))$, has asymptotic power against F_n given by*

$$P[S(\tilde{Z}(\cdot)) > c_S(\alpha, F)].$$

Theoretically we can see how the power of such tests behave by specializing to the homogeneous Gaussian case where under F, $Y_i - \nu(U_i)$ are i.i.d. $\mathcal{N}(0,1)$ while under F_n, $Y_i - \nu_n(U_i, V_i)$ are i.i.d. $\mathcal{N}(0,1)$ with $\nu_n(U,V) = E_{F_n}(Y|U,V)$. This leads to $\psi(W,Y) = b(W)\varepsilon$ for some $b(W)$. So for (7.4),

$$\begin{aligned} E_F\tilde{Z}(a) &= E\left((a(W) - M_a(U))b(W)\varepsilon^2\right) \\ &= E\left((a(W) - M_a(U))b(W)\right) = E\left(a(W)(b(W) - M_b(U)\right). \end{aligned} \tag{7.5}$$

If $\{a(W) : a \in \mathcal{A}\}$ is dense in the set of all L_2 functions of W then $E_F\tilde{Z}(\cdot)$ does not vanish identically unless $E(b(W) - M_b(U))^2 = 0$, that is b is a function of U only which means that the F_n are not a genuine sequence of alternatives. Thus a test of the Kolmogorov-Smirnov or Cramer-von Mises type (with $\mathcal{A} = \{a_\gamma : \gamma \in \Gamma\}$) with $\mathcal{A}$ dense will have power against all contiguous alternatives of this type.

7.2. Approximation of Critical Values

In general it is impossible to find an analytic approximation for the limiting distribution of the Kolmogorov-Smirnov type test statistics. Statistics of the Cramer-von Mises type have distributions depending on eigenvalues of integral equations which can not be solved explicitly. One possible approach is to sample from the limiting Gaussian process. A more natural approach is to sample from a distribution that is close to the empirical distribution of the data but obeys the null hypothesis.[2] However, since the distribution of Y can depend on V under the null hypothesis (only the mean of Y does not), this kind of sampling cannot be done without making some further smoothness assumptions.

[2]We could study the situation of taking samples of size m with $m \to \infty$, $m/n \to 0$, as in Bickel, Götze and Van Zwet (1997), however we leave this to future research.

We propose the following general prescription for sampling. Recall that $\mathcal{B}_n = \{\mathcal{B}_{n1}, \ldots, \mathcal{B}_{nK_n}\}$ is a partition of the support of U, and let $\mathcal{C}_n = \{\mathcal{C}_{n1}, \ldots, \mathcal{C}_{nL_n}\}$ be a partition of the support of V. Assume that $\min_{kl} m_{nkl} \xrightarrow{p} \infty$ and $\max_{kl} m_{nkl}/n \xrightarrow{p} 0$, where $m_{nkl} = \sum_{i=1}^n 1\left(W_i \in \mathcal{B}_{nk} \times \mathcal{C}_{nl}\right)$. Let

$$\hat{\nu}_k = \frac{\sum_{i=1}^n Y_i 1\left[U_i \in \mathcal{B}_{nj}\right]}{\sum_{i=1}^n 1\left[U_i \in \mathcal{B}_{nj}\right]},\ k = 1, \ldots, K_n$$

$$\hat{\nu}_{kl} = \frac{\sum_{i=1}^n Y_i 1\left[W_i \in \mathcal{B}_{nk} \times \mathcal{C}_{nl}\right]}{\sum_{i=1}^n 1\left[W_i \in \mathcal{B}_{nk} \times \mathcal{C}_{nl}\right]},\ k = 1, \ldots, K_n, l = 1, \ldots, L_n$$

The bootstrap sample can be taken as follows. Sample W from $W_1, \ldots, W_n$, and if the resulting W_i obeys $W_i \in \mathcal{B}_{nk} \times \mathcal{C}_{nl}$, then sample Y from $\{Y_i - \hat{\nu}_{kl} + \hat{\nu}_k : W_i \in \mathcal{B}_{nk} \times \mathcal{C}_{nl}\}$.

Now, the bootstrap sample is taken from a distribution under which

$$E^*\left(Y|W\right) = E^*\left(Y|U\right)$$

where * denotes an operation under the bootstrap distribution. When the null hypothesis is true, $\hat{\nu}_{kl} \approx \hat{\nu}_k$, and the bootstrap distribution is close to the empirical distribution, and hence to the true distribution.

If (W_i^*, Y_i^*) $1 \leq i \leq n$ is a bootstrap sample we can define $\hat{Z}_n^*(\cdot)$ as $\hat{Z}_n$ defined for the bootstrap sample. We can prove the asymptotic correctness of the bootstrap critical values. under a slight strengthening of [A1]–[A4].

[A1'] Var $(\varepsilon|U, V) \leq \sigma^2$ and $E\varepsilon^4 < \infty$.

[A2'] In the partition $\{\mathcal{B}_{nj} \times \mathcal{C}_{n\ell}\}$ if $m_{nj\ell}$ denotes the number of $W_i \in \mathcal{B}_{nj} \times \mathcal{C}_{n\ell}$ $(1 \leq j \leq K_n, 1 \leq \ell \leq L_n)$, then $P\left[c \leq \frac{\max m_{nj\ell}}{\min m_{nj\ell}} \leq \frac{1}{c}\right] \to 1$ for some $c \to 0$. This plus our previous assumptions on $m_{nj\ell}$ implies $K_n, L_n \to \infty$, $K_n/n, L_n/n \to 0$. But we also require $K_n^2 L_n^2 = O(n)$.

Theorem 7.2. *Under [A1]–[A4]*

$$P[\hat{Z}_n^*(\cdot) \Rightarrow Z(\cdot)] \to 1 \tag{7.6}$$

where $Z(\cdot)$ is given in (6.4).

The proof hinges on the following two lemmas whose proofs are given in the appendix. Let $\hat{M}_a^{(n)*}(u), \bar{M}_a^{(n)*}(u)$ denote the bootstrap versions of $\hat{M}_a^{(n)}, \bar{M}_a^{(n)}$. Thus,

$$\bar{M}_a^{(n)*}(u) = \sum_{j=1}^{K_n} 1(u \in \mathcal{B}_{nj}) \frac{\sum a(W_i) 1(U_i \in \mathcal{B}_{nj})}{\sum 1(U_i \in \mathcal{B}_{nj})}$$

and

$$\hat{M}_a^{(n)*}(u) = \sum_{j=1}^{K_n} 1(u \in \mathcal{B}_{nj}) \frac{\sum a(W_i^*) 1(U_i^* \in \mathcal{B}_{nj})}{\sum 1(U_i^* \in \mathcal{B}_{nj})}.$$

Let

$$\varepsilon_i^* = Y_i^* - \sum_k \hat{\nu}_k 1(U_i^* \in \mathcal{B}_{nk})$$

$$\tilde{Z}_n^*(a) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (a(W_i^*) - \bar{M}_a^{(n)*}(U_i^*)) \varepsilon_i^*.$$

Lemma 7.3. *Suppose [A1'], [A2'], [A3], [A4] hold. Then*

$$\sup \left\{ \left| \hat{Z}_n^*(a) - \tilde{Z}_n^*(a) \right| : a \in \mathcal{A} \right\} \xrightarrow{P} 0. \tag{7.7}$$

This convergence in probability is in terms of the joint probability of the (W_i, Y_i) $1 \le i \le n$ *and the* (W_i^*, Y_i^*) $1 \le i \le n$.

Lemma 7.4. *Under [A1'], [A2'], [A3], [A4],*

$$\tilde{Z}_n^*(\cdot) \Rightarrow Z(\cdot)$$

in probability. That is, say the Prohorov distance between $S(\tilde{Z}_n^*(\cdot))$ *and* $S(Z(\cdot))$ *tends to* 0 *in probability for all continuous* $S : \mathcal{C}^\infty \to R$.

As a consequence of Theorem 7.2 we can use the α quantile of the bootstrap distribution of $S(\hat{Z}_n^*(\cdot))$ as a consistent estimate of $c_S(\alpha, F)$.

8. Consistency Under Fixed Alternatives

If conditions [A1g], [A2], [A3], [A4] are satisfied, but F does not satisfy the null hypothesis, then (as noted in section 6.2) we expect that the process $\hat{Z}_n(\cdot)$ drifts off to ∞ in some direction a. Evidently then a test based on $S(\hat{Z}_n(\cdot))$ will be consistent against a fixed alternative if

(i) $S(\hat{Z}_n(\cdot)) \xrightarrow{P} \infty$, which is expected if S pays attention to a dense set of a's (as in the Kolmogorov-Smirnov or Cramer-von Mises case), and

(ii) $\hat{c}_n = O_p(1)$, where $\hat{c}_n$ is the critical value used.

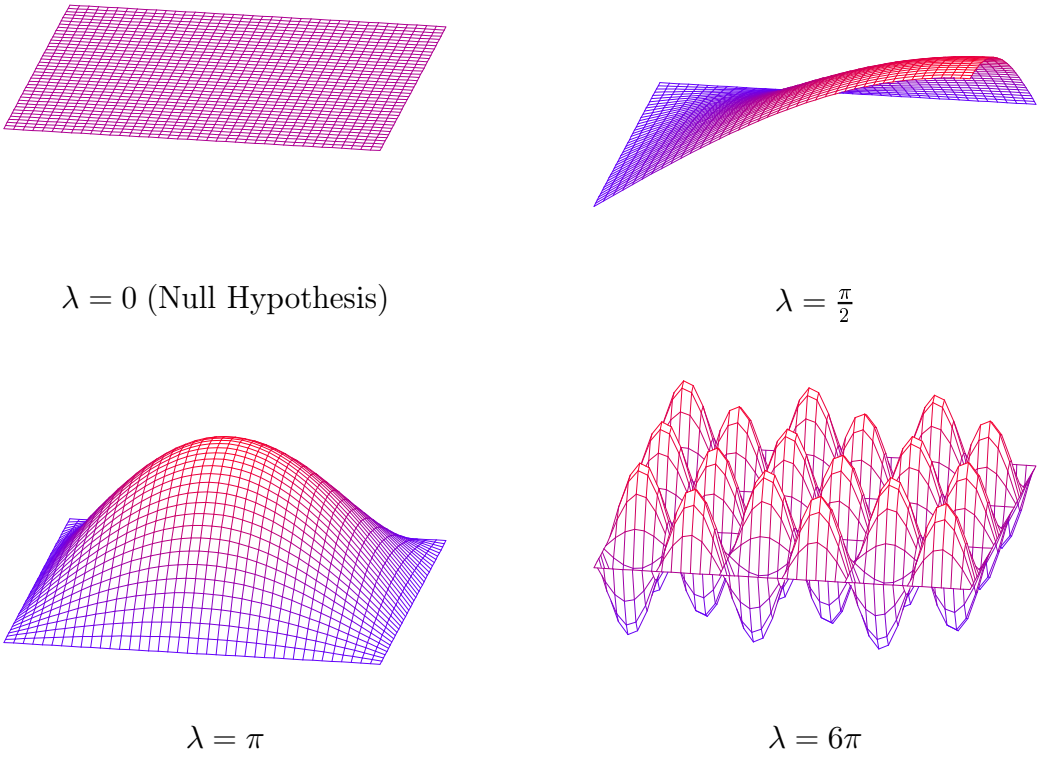
In particular if we obtain $\hat{c}_n$ from the bootstrap this will happen if [A1'g],[A2'], [A3], [A4] hold where [A1'g] is [A1g] and in addition $E\left(\varepsilon_{ig}^4\right) < \infty$. As in section 6.2 this requires proving that

$$\sup\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(M_a^*(U_i^*) - \hat{M}_a^{(n)(*)}(U_i^*))(\gamma^*(U_i^*) - \nu^*(U_i^*, V_i^*)) : a \in \mathcal{A}\right\} \xrightarrow{P} 0$$

and that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{(a(W_i^*) - M_a^*(U_i^*))(\nu^*(U_i^*) - \nu^*(U_i^*, V_i^*))$ is tight. The arguments are analogous to those given for Lemma 6.8 and we omit them.

9. Simulations

We checked the behavior of different test statistics using a small Monte Carlo experiment. We consider a sample of 500 independent observations from $U, V, Y = \nu_\lambda(U, V) + \varepsilon$, where U, V and ε are independent, $U, V \sim U(0, 1)$, $\varepsilon \sim N(0, 1)$. We take $\nu_\lambda(u, v) = 0.8\sin(\lambda u)\sin(\lambda v)$, where $\lambda = 0,\ \pi/2,\ \pi,\ 6\pi$. With $\lambda = 0$ taken as the null assumption, this design gives a wide range of different types of departures from the null hypothesis, as illustrated in Figure 9.1.



$\lambda = 0$ (Null Hypothesis) $\lambda = \frac{\pi}{2}$

$\lambda = \pi$ $\lambda = 6\pi$

Figure 9.1: Simulation Departures: $\nu_\lambda(u, v) = 0.8 \sin(\lambda u) \sin(\lambda v)$

We examine three different test statistics as described below. All are all based on a partition of the unit square into 10×5 blocks, where the support of U was divided to 10 blocks. We chose a partition that is asymmetrical in the U and V because of the way bias is introduced by the partitioning. In particular, the discretization of the range of U can introduce a bias — if it is not fine enough, a distribution in which Y and W are conditionally independent given U, may not adequately display such conditional independence given the blocks. Condition [A3] is necessary to ensure that the test will be asymptotically unbiased. In contrast, the wideness of the blocks on the V dimension is secondary and enters only through efficiency considerations, and the behavior of the bootstrap.

With the division to blocks, one simple test statistic is the standard ANOVA F statistic for testing for the presence of only an effect of U (i.e. no V effect and no

interaction). This is our first test statistic. The second is the Kolmogorov-Smirnov statistic with $\{a_\gamma\}$ taken as the quadrant indicators $\{1(u \geq \gamma_1, v \geq \gamma_2)\}$. The third is another Kolmogorov-Smirnov statistic with $\{a_\gamma\}$ taken as the rectangle indicators $\{1(\gamma_1 < u \leq \gamma_2, \gamma_3 < v \leq \gamma_4)\}$.

The tests are defined formally as follows. With some abuse of notation, let Y_{klm}, $k = 1, \ldots, K$, $l = 1, \ldots., L$, $m = 1, \ldots, m_{kl}$ be the Y-value of the m-th observation in the kl block. Denote as usual $\bar{Y}_{kl.} = m_{kl}^{-1} \sum_m Y_{klm}$ and $\bar{Y}_{k..} = m_{k.}^{-1} \sum_{lm} Y_{klm}$. Note that

$$\sum_{i=1}^{n} \left(a(W_i) - E_{\hat{P}}(W|U_i)\right)\left(Y_i - E_{\hat{P}}(Y|U_i)\right) = \sum_{i=1}^{n} a(W_i)\left(Y_i - E_{\hat{P}}(Y|U_i)\right)$$

The three test statistics are then

$$F = \frac{\sum_{kl} \bar{Y}_{kl.}^2 m_{kl} - \sum_k \bar{Y}_{k..}^2 m_{k.}}{\sum_{klm} Y_{klm}^2 - \sum_k \bar{Y}_{k..}^2 m_{k.}}$$

$$KS_1 = \max_{kl} \left| \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{m=1}^{m_{k'l'}} \left(Y_{k'l'm} - \bar{Y}_{k..}\right) \right|$$

$$KS_2 = \max_{k_1 l_1 k_2 l_2} \left| \sum_{k'=k_1}^{k_2} \sum_{l'=l_1}^{l_2} \sum_{m=1}^{m_{k'l'}} \left(Y_{k'l'm} - \bar{Y}_{k..}\right) \right|$$

The three simulated departures ($\lambda = \pi/2$, π, 6π of Figure 9.1) are intended to check the strength and weakness of these test statistics. The first Kolmogorov-Smirnov statistic, KS_1, is appropriate for deviations like the one with $\lambda = \pi/2$, in which the corners are different from the average. The second Kolmogorov-Smirnov statistic, KS_2, is supposed to show its strength against deviations which are concentrated in the center as the case of $\lambda = \pi$. Finally, the statistic F diffuses its strength among 40 degrees of freedom. Hence it will be weak against particular deviations, but unlike the two KS tests, it will be relatively strong against more complicated deviations like the one with $\lambda = 6\pi$. (This paragraph was written before any simulation was done).

The bootstrap was done essentially as described above. There were, however, two modifications. Theoretically the number of observations in a cell should increase to ∞, but in practice it is finite, and may be quite small (in our simulation, there were, on average, 10 observations in a cell). Since we center the observations in a cell (so that we sample under H), this decreases the variance of the

distribution from which the bootstrap samples are taken, and as a result, the spread of the test statistics is reduced. To correct this, our first modification was to multiply each observation in the kl cell by $\sqrt{m_{kl}/(m_{kl}-1)}$. See Silverman (1981) for a similar correction. The F test is invariant to this correction, but the Kolmogorov-Smirnov type tests were not conservative without the inflation. The second modification was to bootstrap sample only the Y values (hence our critical values apply to tests conditional on the W's).

Rejection was defined as occurring whenever the test statistic was one of the $100(1-\alpha)\%$ larger values among 200 bootstrap observations, where α is the declared value. The randomization (both the sampling and the bootstrapping) was common to the twenty four combinations of test statistics and values of λ and α.

The results are given in the following table

Test Statistics	$\lambda = 0$	$\lambda = \pi/2$	$\lambda = \pi$	$\lambda = 6\pi$
$\alpha = 0.1$				
F	0.072	0.492	0.443	0.453
KS_1	0.115	0.970	0.565	0.122
KS_2	0.095	0.838	0.887	0.113
$\alpha = 0.05$				
F	0.025	0.355	0.290	0.307
KS_1	0.052	0.922	0.395	0.072
KS_2	0.050	0.728	0.818	0.060

These results confirm the intuition behind the design of our simulation. In particular, the F statistic displays roughly the same rate of rejection among the different alternatives. The Kolmogorov-Smirnov statistics display higher power against the alternatives $\lambda = \pi/2$ and $\lambda = \pi$ and very low power against the alternative $\lambda = 6\pi$. Specifically, KS_1 displays relatively higher power against $\lambda = \pi/2$ than $\lambda = \pi$, and the opposite is true for KS_2. In any case, this simulation highlights how the different types of departures are captured by the three statistics, as well as how much that affects the performance of the different test statistics.

10. Conclusion

In this paper we have given a systematic coverage of score tests in semiparametric and nonparametric contexts. We have highlighted the way different approaches to testing give differential treatment to alternative departures from the null hypothesis. For the case of testing index restrictions on regression functions, we have examined several tests and illustrated how important the construction of the test is to performance under different alternatives.

Our development is clearly only the first step in a general coverage of test design for index models. The family of tests analyzed in the previous section is clearly not the only possible one. For example, we have focused on nonparametric approximation via blocks, and one may design other more modern or more elegant ways of nonparametric estimation of the required conditional expectations. In addition, we have only considered the situation in which the index model does not depend on an unknown parameter. Certainly this should be generalized to meet the needs of testing in applied research. However, we have limited our attention to this simplest case because the formal analysis above was complicated enough, and analyzing many other methods would obscure much of our general conceptual ('soft') discussion.

A. Appendix

Our results are based on Theorem 2.11.9 of van der Vaart and Wellner (1996), which we state here for ease in following our arguments below. As above (c.f. [A3]), for every n, define the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_2^n)$ as the minimal number of sets N_ε in a partition $\mathcal{F} = \cup_{j=1}^{N_\varepsilon} \mathcal{F}_{\varepsilon j}^n$ of the index sets into sets $\mathcal{F}_{\varepsilon j}^n$ such that, for every partitioning set $\mathcal{F}_{\varepsilon j}^n$

$$\sum_{i=1}^{m_n} E^* \left| Z_{ni}(f) - Z_{ni}(g) \right|^2 \leq \varepsilon^2,$$

where the partitions are clearly allowed to depend on n. If we denote the outer integral as E^*, then Theorem 2.11.9 of van der Vaart and Wellner (1996) is stated as:

Bracketing Central Limit Theorem: For each n, let $Z_{n1}, ..., Z_{n,m_n}$ be independent stochastic processes with finite secomd moments indexed by a to-

tally bounded semimetric space $(\mathcal{F}, \rho)$. Suppose

$$\sum_{i=1}^{m_n} E^* \left\| Z_{ni} \right\|_{\mathcal{F}} \left\{ \left\| Z_{ni} \right\|_{\mathcal{F}} > \eta \right\} \to 0, \quad \text{for every } \eta > 0.$$

$$\sup_{\rho(f,g)<\delta_n} \sum_{i=1}^{m_n} E\left(Z_{ni}\left(f\right) - Z_{ni}\left(g\right)\right)^2 \to 0, \quad \text{for every } \delta_n \searrow 0$$

$$\int_0^{\delta_n} \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{F}, L_2^n\right)} d\varepsilon \to 0, \quad \text{for every } \delta_n \searrow 0$$

Then the sequence $\sum_{i=1}^{m_n} \left(Z_{ni} - E\left(Z_{ni}\right)\right)$ is asymptotically tight in $\ell^\infty\left(\mathcal{F}\right)$ and converges in distribution provided it converges marginally. If the partitions can be chosen independent of n, then the middle of the displayed conditions is unnecessary.

Suppose that the family $\mathcal{A}$ of functions is bounded by a constant L.

Proof of Lemma 6.1. Write

$$D_n^2(a, a') = \sum_{j=1}^{K_n} \frac{m_{nj}}{n} E[(a - a')^2(W) | U \in \mathcal{B}_{nj}]. \tag{A.1}$$

Then

$$\left| D_n^2(a, a') - \|a - a'\|_2 \right| = \left[\sum_{j=1}^{K_n} E[(a - a')^2(W) | U \in \mathcal{B}_{nj}] \left(\frac{m_{nj}}{n} - P[U \in \mathcal{B}_{nj}] \right) \right]$$

Apply Cauchy-Schwarz to obtain

$$|D_n^2(a, a') - \|a - a'\|_2^2| \leq \|a - a'\|_2 R_n(a, a') \tag{A.2}$$

where

$$R_n(a, a') = \left[\sum_{j=1}^{K_n} E((a - a')^2(W) | U \in \mathcal{B}_{nj}) \frac{\left(\frac{m_{nj}}{n} - P[U \in \mathcal{B}_{nj}] \right)^2}{P[U \in \mathcal{B}_{nj}]} \right]^{1/2}.$$

But

$$R_n(a, a') \leq 2L \sum_{j=1}^{K_n} \frac{\left(\frac{m_{nj}}{n} - P[U \in \mathcal{B}_{nj}] \right)^2}{P[U \in \mathcal{B}_{nj}]} \tag{A.3}$$

and

$$E\left[\sum_{j=1}^{K_n}\left(\frac{m_{nj}}{n}-P[U\in\mathcal{B}_{nj}]\right)^2/P[U\in\mathcal{B}_{nj}]\right]\leq\frac{K_n}{n}\to 0. \tag{A.4}$$

Now, $|D_n^2(a,a')-\|a-a'\|_2^2|\leq\|a-a'\|_2$ for all a,a' implies

$$N_{[\ ]}(\varepsilon,\mathcal{A},D_n)\leq N_{[\ ]}\left(\frac{\varepsilon^2}{2},\mathcal{A},\|\cdot\|_2\right)$$

for $\varepsilon\leq 1$. Thus (A.2), (A.3), and (A.4) yield the lemma. □

Proof of Lemma 6.4. Evidently,

$$E\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(M_a(U_i)-\bar{M}_a^{(n)}(U_i))\varepsilon_i\right)^2\leq\sigma^2E(M_a(U_1)-\bar{M}_a^{(n)}(U_1))^2\leq\sigma^2\gamma_n\to 0$$

by [A3]. We apply the Bracketing Central Limit Theorem and check that,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[\sup\{|M_a(U_i)-\bar{M}_a^{(n)}(U_i)|:a\in\mathcal{A}\}|\varepsilon_i|1(|\varepsilon_i|\geq\eta\sqrt{n}\left|\left(M_a-\bar{M}_a^{(n)}\right)(U_i)\right|^{-1})\right]\to 0.$$

This is true by [A3] as above by decomposing according to which $\mathcal{B}_{nj}$ that U_i belongs to and bounding the expected absolute value of each term by

$$\frac{\gamma_n}{\sqrt{n}}E|\varepsilon_1|1\left[|\varepsilon_1|\geq\frac{\eta\sqrt{n}}{\gamma_n}\right]\leq\frac{\sigma^2\gamma_n^2}{n\eta}.$$

so that the sum is bounded by $\sigma^2\gamma_n^2\to 0$, and the first condition of Theorem 2.11.9 applies. Further,

$$\sup\{E((M_a-M_{a'})(U_1)-(M_a^{(n)}-M_{a'}^{(n)})(U_1))^2\varepsilon_1^2:d(a,a')\leq\delta_n\}\leq\sigma^2\delta_n^2\to 0 \tag{A.5}$$

and the second condition of Theorem 2.11.9 follows.

Similarly apply the vector form of Jensen's inequality to check that for any $\mathcal{A}^*\subset A$

$$\begin{aligned}&E\sup\left\{\left[(M_a-M_{a'})(U_1)-(\bar{M}_a^{(n)}-\bar{M}_{a'}^{(n)})(U_1)\right]^2\varepsilon_1^2:a,a'\in\mathcal{A}^*\right\}\\&\leq 2\sigma^2E\sup\{(a-a')^2(U_1):a,a'\in\mathcal{A}^*\}.\end{aligned} \tag{A.6}$$

It is then easy to see that $|a|_\infty \le M$ and [A4] guarantee the third condition of Theorem 2.11.9 and tightness of $\frac{1}{\sqrt{n}}\sum_{i=1}^n (M_a(U_i) - \bar{M}_a^{(n)}(U_i))\varepsilon_i$ and the lemma follows. □

Proof of Lemma 6.5. We can write, since both $\hat{M}_a^{(n)}$ and $\bar{M}_a^{(n)}$ are constant on $\mathcal{B}_{nj}$, $1 \le j \le K_n$,

$$\begin{aligned} R_n^{(a)} &\equiv \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n (\hat{M}_a^{(n)}(U_i) - \bar{M}_a^{(n)}(U_i))\varepsilon_i \\ &= \tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{K_n} (\hat{M}_a^{(n)}(\xi_j) - \bar{M}_a^{(n)}(\xi_j))S_{nj} \end{aligned} \tag{A.7}$$

where ξ_j is a representative value in $\mathcal{B}_{nj}$ and $S_{nj} = \sum\{\varepsilon_i : U_i \in \mathcal{B}_{nj}\}$.

Condition $\Delta_n(a)$ on $W_1, \ldots, W_n$ and take second moments noting that $E(S_{nj}|W_1, \ldots, W_n) = 0$. We obtain

$$E(R_n^2(a)|W_1, \ldots, W_n) = \frac{1}{n}\sum_{j=1}^{K_n} (\hat{M}_a^{(n)}(\xi_j) - \bar{M}_a(\xi_j))^2 \; m_{nj} \cdot \text{Var}\,(\varepsilon|U \in \mathcal{B}_{nj}).$$

Hence

$$\begin{aligned} E(R_n^2(a)|\mathcal{B}_n) &\le \sigma^2 \tfrac{1}{n}\textstyle\sum_{j=1}^{K_n} E(a^2(W)|U \in \mathcal{B}_{nj}) \\ &\le \sigma^2 \tfrac{K_n}{n} Ea^2(W) \to 0. \end{aligned} \tag{A.8}$$

In (A.8) we use [A1] and $E(a^2(W)|U \in \mathcal{B}_{nj}) = Ea^2(W)1(U \in \mathcal{B}_{nj})/P(U \in \mathcal{B}_{nj}) \le Ea^2(W)$. To prove tightness of $\{\Delta_n(a)\}$ we use the Bracketing Central Limit Theorem, conditional on $\mathcal{B}_n$.

Since the second condition of Theorem 2.11.9 is trivial, we need to check

(a)

$$\tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{K_n} E[\sup_{a\in\mathcal{A}}\{|\hat{M}_a^{(n)}(\xi_j) - \bar{M}_a^{(n)}(\xi_j)||S_{nj}|\}$$

$$1\{\sup_{a\in\mathcal{A}}\{|\hat{M}_a^{(n)}(\xi_j) - \bar{M}_a^{(n)}(\xi_j)||S_{nj}| > \eta\sqrt{n}\}|\mathcal{B}_n\} \xrightarrow{P} 0$$

for every $\eta > 0$.

(b) If $N_{[\,]}(\varepsilon, \mathcal{A}, L_2^n)$ = Minimal number of sets in a partition of $\mathcal{A}$ such that for each member $\mathcal{A}_\ell$, $1 \le \ell \le N_{[\,]}$

$$\tfrac{1}{n}\textstyle\sum_{j=1}^{K_n} E[\sup\{(\hat{M}_{a_1}^{(n)}(\xi_j) - \bar{M}_{a_1}(\xi_j) - \hat{M}_{a_2}^{(n)}(\xi_j)$$

$$+\bar{M}_{a_2}^{(n)}(\xi_j))^2 S_{nj}^2 : a_1, a_2 \in \mathcal{A}_\ell\}|\mathcal{B}_n] \le \varepsilon^2$$

with probability 1, then

$$\int_0^{\delta_n} \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{A}, L_2^n)}d\varepsilon \xrightarrow{P} 0 \text{ if } \delta_n \to 0.$$

We begin by proving (a). Apply the Markov inequality given $W_1, \ldots, W_n$ to obtain as a bound for the left-hand side of (a)

$$\begin{aligned} &\tfrac{1}{\sqrt{n}} \textstyle\sum_{j=1}^{K_n} \sup\{|\hat{M}^{(n)}(\xi_j) - \bar{M}_a^{(n)}(\xi_j)|^2 : a \in \mathcal{A}\} E(S_{nj}^2 | W_1, \ldots, W_n) \tfrac{1}{\eta\sqrt{n}} \\ &= \tfrac{1}{\eta} E \sup\{|\hat{M}_a^{(n)}(U_1) - \bar{M}_a^{(n)}(U_1)|^2 : a \in \mathcal{A}\}. \end{aligned} \tag{A.9}$$

Now,

$$\hat{M}_a^{(n)}(U_1) - \bar{M}_a^{(n)}(U_1) \xrightarrow{P} 0$$

by an elementary argument conditioning on $\mathcal{B}_n$. Moreover condition [A4] is easily seen to imply that a condition of Bickel and Millar (1992) trivially applies to the process $\hat{M}_a^{(n)}(U_1) - \bar{M}_a^{(n)}(U_1)$ so that

$$\sup\{|\hat{M}_a^{(n)}(U_1) - \bar{M}_a^{(n)}(U_1)| : a \in \mathcal{A}\} \xrightarrow{P} 0. \tag{A.10}$$

Then, (A.10) and the boundedness of $\mathcal{A}$ imply that the right-hand-side of (A.9) $\to 0$ and (a).

To prove (b) we note that,

$$\begin{aligned} &\tfrac{1}{n} \textstyle\sum_{j=1}^{K_n} E[\sup\{(\hat{M}_{a_1}^{(n)}(\xi_j) - \bar{M}_{a_1}^{(n)}(\xi_j) - \hat{M}_{a_2}^{(n)}(\xi_j) \\ &+ \bar{M}_{a_2}^{(n)}(\xi_j))^2 S_{nj}^2 : a_1, a_2 \in \mathcal{A}_\ell\} | \mathcal{B}_n] \\ &\leq \tfrac{\sigma^2}{n} \textstyle\sum_{i=1}^n E[\sup\{(a_1 - a_2)^2(W_i) : a_1, a_2 \in \mathcal{A}_\ell\} | \mathcal{B}_n] \end{aligned}$$

and hence

$$N_{[\,]}^*(\varepsilon, \mathcal{A}, L_2^n) \leq N_{[\,]}\left(\frac{\varepsilon}{\sigma^2}, \mathcal{A}, D_n\right). \tag{A.11}$$

Thus, (b) follows from [A4] and Lemma 6.1. □

Proof of Lemma 6.6. For $U_i \in \mathcal{B}_{nj}$, let $\bar{\nu}(U_i) = m_{nj}^{-1} \sum \{\nu(U_i), U_i \in \mathcal{B}_{nj}\}$, then

$$\begin{aligned} &\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(U_i, V_i) - \hat{M}_a^{(n)}(U_i))\nu(U_i) \\ &= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(U_i, V_i) - \tilde{M}_a^{(n)}(U_i))(\nu(U_i) - \bar{\nu}(U_i)) \\ &= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (a(U_i, V_i) - M_a(U_i))(\nu(U_i) - \bar{\nu}(U_i)) \\ &+ \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n (M_a(U_i) - \bar{M}_a^{(n)}(U_i))(\nu(U_i) - \bar{\nu}(U_i)). \end{aligned} \tag{A.12}$$

We show that the first term in (A.12) is tight as a process in a, and tends in probability to 0 by applying the Bracketing Central Limit Theorem. The first condition (a) is easy since

$$\sup\{|a(W) - M_a(U)||\nu(U) - \bar{\nu}(U)| : a \in \mathcal{A}\} \leq 2L|\nu(U) - \bar{\nu}(U)| \leq 2L\Lambda_n \to 0.$$

Define $N^{**}_{[\ ]}(\varepsilon, \mathcal{A}, L_2^n)$ to be the smallest cardinality for a partition $\mathcal{A}_1, \ldots, \mathcal{A}_{N^{**}_{[\ ]}}$ such that,

$$\frac{1}{n}\sum_{i=1}^{n} E\sup\{(\nu(U_i) - \bar{\nu}(U_i))^2[(a_1 - a_2)(W_i) - (M_{a_1} - M_{a_2})(U_i)]^2 : a_1, a_2 \in \mathcal{A}_\ell\} \leq \varepsilon^2$$

for $1 \leq \ell \leq N^{**}_{[\ ]}$. But

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i=1}^{n} E[\sup\{(\nu(U_i) - \bar{\nu}(U_i))^2(a(W_i) - M_a(U_i))^2 : a \in \mathcal{A}_\ell - \mathcal{A}_\ell\} \\ &\leq \tfrac{\Lambda_n^2}{n}\textstyle\sum_{i=1}^{n} E[\sup\{(a(W_i) - M_a(U_i))^2 : a \in \mathcal{A}_\ell - \mathcal{A}_\ell\}|\mathcal{B}_n] \end{aligned}$$

and since $E((a-b)(W) - (M_a(U) - M_b(U)))^2 \leq E(a-b)^2(W)$,

$$N^{**}(\varepsilon, \mathcal{A}, L_2^n) \leq N_{[\ ]}\left(\frac{\varepsilon}{\Lambda_n}, \mathcal{A}, D_n\right).$$

So

$$\begin{aligned} &\textstyle\int_0^{\delta_n} \sqrt{\log N^{**}_{[\ ]}(\varepsilon, \mathcal{A}, L_2^n)}d\varepsilon \leq \Lambda_n \int_0^{\delta_n/\Lambda_n} \sqrt{\log N_{[\ ]}(\varepsilon, \mathcal{A}, D_n)}d\varepsilon \\ &\leq \Lambda_n \textstyle\int_0^{2L} \sqrt{\log N_{[\ ]}(\varepsilon, \mathcal{A}, D_n)}d\varepsilon \xrightarrow{P} 0 \end{aligned} \tag{A.13}$$

by [A4] and Lemma 6.1 since $N_{[\ ]}(2L, \mathcal{A}, D_n) = 1$.

Thus the first term in (A.12) is a tight process. It is elementary to check that

$$E\left(n^{-1/2}\sum_{i=1}^{n}(a(W_i) - M_a(U_i))(\nu(U_i) - \bar{\nu}(U_i))|U_1, \ldots, U_n\right) = 0$$

and

$$\begin{aligned} &E\left(n^{-1/2}\textstyle\sum_{i=1}^{n}(a(W_i) - M_a(U_i))(\nu(U_i) - \bar{\nu}(U_i))\right)^2 \\ &\leq \tfrac{4L^2}{n}\textstyle\sum_{i=1}^{n} E(\nu(U_i) - \bar{\nu}(U_i))^2 \\ &= 4L^2E(\nu(U_1) - \bar{\nu}(U_1))^2 \leq 4L^2\Lambda_n^2 \to 0. \end{aligned}$$

Thus,

$$\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a(W_i)-M_a(U_i))(\nu(U_i)-\bar{\nu}(U_i))\right| : a\in\mathcal{A}\right\}\xrightarrow{P}0. \tag{A.14}$$

Consider the second term in (A.12). Note that

$$\begin{aligned}
&\left|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}(M_a(U_i)-\bar{M}_a^{(n)}(U_i))(\nu(U_i)-\bar{\nu}(U_i))\right| \\
&=\left|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}(M_a(U_i)-\bar{M}_a(U_i))(\nu(U_i)-\bar{\nu}(U_i))\right| \\
&\le \tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{K_n}\sum_{U_i\in\mathcal{B}_{nj}}\{|M_a(U_i)-\bar{M}_a(U_i))(\nu(U_i)-\bar{\nu}(U_i))\} \\
&\le \textstyle\sum_{j=1}^{K_n}\frac{m_{nj}}{n}\left(n^{1/2}\Delta_n\right)\le n^{1/2}\Delta_n\xrightarrow{P}0.
\end{aligned} \tag{A.15}$$

Hence, combining (A.14) and (A.15) the lemma follows. □

Proof of Lemma 6.8. We begin by establishing the lemma with $\hat{M}_a^{(n)}(U_i)$ replaced by $\hat{M}_a^{(n)-}$ given by (6.10) a familiar alternative estimate of the conditional expectation. We can write

$$\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}(\hat{M}_a^{(n)-}(U_i)-\bar{M}_a^{(n)}(U_{j\ell}))(\nu(U_i)-\nu(U_i,V_i)) \\
&=\tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{K_n}\left(\frac{1}{m_{nj}-1}\sum_{k\neq i}\{(a(W_k)-E(a(W_k)|U_k\in\mathcal{B}_{nj}))(\nu(U_i)-\nu(U_i,V_i))\} : U_k,U_i\in\mathcal{B}_{nj}\}\right)
\end{aligned} \tag{A.16}$$

Given $\mathcal{B}_n$ the K_n summands are independent and each summand is a multiple of a degenerate U statistic since

$$\begin{aligned}
E\{(a(W_n)-E(a(W_k))|U_k\in\mathcal{B}_{nj},W_i,U_i\in\mathcal{B}_{nj}\}=0 \\
E\{(\nu(U_i)-\nu(U_i,V_i))|U_i\in\mathcal{B}_{nj},W_k,U_k\in\mathcal{B}_{nj}\}=0.
\end{aligned}$$

Thus,

$$E\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{M}_a^{(n)-}(U_i)-\bar{M}_a^{(n)}(U_i))(\nu(U_i,V_i)-\nu(U_i))\right)^2=O\left(\frac{K_n}{n}\right). \tag{A.17}$$

To prove tightness we need to bound

$$\frac{1}{n}\sum_{j=1}^{K_n}E(\sum\{(\hat{M}_a^{(n)-}-M_{a'}^{(n)-}-\bar{M}_a^{(n)}+\bar{M}_{a'}^{(n)})(U_i))(\nu(U_i,V_i)-\nu(U_i))^2\}. \tag{A.18}$$

But again by appealing to the boundedness of ν and the variance calculation for degenerate U statistics, (A.18) is,

$$\leq L^2 \frac{1}{n} \sum_{i=1}^{n} E((a-a')^2(W_i)|\mathcal{B}_n) \tag{A.19}$$

and tightness follows in the same fashion as we have argued in Lemma 6.1. To go from $\hat{M}_a^{(n)-}(U_i)$ which is in fact a natural and attractive alternative to $\hat{M}_a^{(n)}(U_i)$ (see Aït Sahalia, Bickel and Stoker (1998) for instance) we need consider

$$\begin{array}{l} \frac{1}{\sqrt{n}} \sum (\hat{M}_a^{(n)-}(U_i) - \hat{M}_a^{(n)}(U_i))(\nu(U_i) - \nu(U_i, V_i)) \\ = \frac{-1}{\sqrt{n}} \sum_{j=1}^{K_n} \frac{1}{m_{nj}} \sum \{(a(W_i) - \bar{M}_a^{(n)}(U_i))(\nu(U_i) - \nu(U_i, V_i)) : U_i \in \mathcal{B}_{nj}\} \\ + \frac{1}{\sqrt{n}} \sum_{j=1}^{K_n} \\ \frac{1}{m_n(m_{n-1})} \sum_{i \neq k} \{(a(W_i) - Ea(W_i)|U_i \in \mathcal{B}_{nj})(\nu(U_k) - \nu(U_k, V_k)) : U_i, U_k \in \mathcal{B}_{nj}\}. \end{array} \tag{A.20}$$

The second term is familiar but of small order because of the additional m_{nj}. The second is an uncentered sum of independent variables with expectation bounded by $O\left(K_n/\sqrt{n}\right) = o(1)$ by [A2g]. Its variance is evidently $o(1)$. Tightness here can be argued as for the $\hat{M}_a^{(n)-}$ case. □

Proof of Lemma 7.3. The proof of Lemma 7.3 involves paralleling the treatment in Lemma 6.5 with one extra level of randomness. Thus, if

$$\begin{array}{l} R_n^*(a) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{M}_a^{(n)*}(U_i^*) - \bar{M}_a^{(n)*}(U_i^*))\varepsilon_i^* \\ E^*([R_n^*]^2(a)|W_1^*, W_n^*) = \frac{1}{n} \sum_{j=1}^{K_n} (\hat{M}_a^{(n)*}(\xi_j) \\ -\bar{M}_a^{(n)*}(\xi_j))^2 \sum \{\mathrm{Var}^*(\varepsilon_i^*|W_i^*, U_i^* \in \mathcal{B}_{jn}) : U_i^* \in \mathcal{B}_{jn}\}. \end{array} \tag{A.21}$$

Now

$$\begin{array}{l} \mathrm{Var}^*(\varepsilon_i^*|W_i^*, U_i^* \in \mathcal{B}_{jn}) = \\ \frac{1}{m_{nj\ell}} \sum_{\ell} \{(Y_i - \hat{Y}_{j\ell})^2 1(W_i \in \mathcal{B}_{nj} \times \mathcal{C}_{n\ell})\} \end{array}$$

if $W_i \in \mathcal{B}_{nj} \times \mathcal{C}_{n\ell}$. Further,

$$\begin{array}{l} P[\max_j \mathrm{Var}^*(\varepsilon_i^*|W_i^*, U_i^* \in \mathcal{B}_{jn}) \geq M] \\ \leq K_n L_n \max_{j,\ell} P\left[\frac{1}{m_{nj\ell}} \sum_i (Y_i - E(Y_i|W_i \in \mathcal{B}_{nj} \times \mathcal{C}_{n\ell}))^2 \geq L\right] \\ \leq E(Y - E(Y|W \in \mathcal{B}_{nj} \times \mathcal{C}_{n\ell})^4 \frac{C(K_n L_n)^2}{L^2 n} + o(1)) \end{array}$$

since by assumption [A2'] $\min_{j,\ell} m_{nj\ell} \geq \frac{n}{CK_nL_n}$ with probability tending to 1. Hence,

$$\max_j \ \mathrm{Var}^*(\varepsilon_i^*|W_i^*) = O_p(1) \tag{A.22}$$

so that

$$E^*([\Delta_n^*]^2(a)|\mathcal{B}_n^*) \leq \frac{O_p(1)}{n} \sum_j m_{nj} \frac{\mathrm{Var}^*(a(W^*)|U^* \in \mathcal{B}_{nj})}{m_{nj}}. \tag{A.23}$$

Now we argue as for (A.4) that this last expression is $O_p\left(K_n/n\right) = o_p(1)$. The proof of tightness proceeds in the same way. The only new difficulty is that we now have to deal with the metric

$$D_n^{2*}(a,a') = \frac{1}{n}\sum_{i=1}^n E^*\left((a-a')^2(W_i^*)|\mathcal{B}_n^*\right).$$

But we can argue as for Lemma 6.1 that

$$P[|D_n^{2*}(a,a') - D_n^2(a,a')| \leq \varepsilon D_n(a,a')] \to 1$$

for every $\varepsilon > 0$ and the lemma follows. □

Proof of Lemma 7.4. Lemma 7.4 is proved in the same way as Lemma 6.1, using tightness according to D_n^* and the double array Bracketing Central Limit Theorem. □

References

[1] Aït-Sahalia, Y. , P. J. Bickel and T.M. Stoker (1998), "Goodness-of-Fit Tests for Regression Using Kernel Methods," MIT Sloan School of Management Working Paper, SWP #3970, November 1994, revised June 1998.

[2] Bickel, P.J., F. Götze, and W.R. Van Zwet (1997), "Resampling Fewer Than n Observations: Gains, Losses and Remedies for Losses," *Statistica Sinica (ck)*, 7, 1-31.

[3] Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J. A.Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, London.

[4] Bickel, P.J. and Millar, P.W. (1992), "Uniform Convergence of Probability Measure on Classes of Functions," *Statistica Sinica*, 2, 1-15.

[5] Bickel, P.J. and Y. Ritov (1991), "Large Sample Theory of Estimation in Biased Sample Regression Models," *Annals of Statistics*, 19, 797-816,

[6] Bickel, P.J. and M. Rosenblatt (1973), "On Some Global Measures of the Deviations of Density Function Estimates," *Annals of Statistics* , 1, 1071-1096.

[7] Bierens, H.J. (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1143-1458.

[8] Bierens, H.J. and W. Ploberger (1997), "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica*, 65, 1129-1151.

[9] Chen, X. and Y. Fan (1998), "Consistent and Directional Tests via Funcational Principal Components Analysis," draft, Department of Economics, University of Chicago.

[10] Choi, S., W.J. Hall and A. Schick (1996), "Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models," *Annals of Statistics*, 24, 841-861.

[11] Fan, J. W. (1996), "Tests of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association,* , 91, 674-688.

[12] Hart (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer Verlag, New York.

[13] Lewbel, A. (1995), "Consistent Nonparametric Hypothesis Tests with an Application to Slutsky Symmetry," *Journal of Econometrics*, 67, 379-401.

[14] Newey, W.K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

[15] Pollard, D. (1990), *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics 2, IMS and ASA, Hayward.

[16] Rao, C.R. (1947). "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Application to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50-57.

[17] Silverman (1981), "Using Bootstrap Kernel Density Estimates to Investigate Unimodality," *Journal of the Royal Statistical Society* B, 43, 97-99.

[18] Stoker, T.M. (1992), *Lectures on Semiparametric Econometrics* , CORE Foundation, Louvain-la-Neuvre.

[19] van der Vaart, A.J. and J. Wellner (1996), *Weak Convergence and Empirical Processes*, Springer Verlag, New York.

[20] Yatchew, A. (1998), "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36, 669-721.