NONPARAMETRIC TESTING OF AN INDEX MODEL

By Peter J. Bickel, Ya'acov Ritov, and Tom Stoker

October 24, 2001

Abstract

Following a framework proposed in Bickel, Ritov and Stoker (2001) we propose and analyze the behavior of a broad family of tests for $H : E(Y \mid U, V) = E(Y \mid U)$ when we observe $(U_i, V_i, Y_i) \in R^{d_u + dv + 1}$ i.i.d., $i = 1, \ldots, n$.

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 judgment of whether an estimated model displays decent goodness-of-fit to the empirical data. The paradigmatic situation we consider is one where the investigator views some departures from the hypothesized model as being of primary importance with others of interest if sufficiently gross but otherwise secondary. For instance consider a signal hypothesized to be constant. Low frequency departures from a constant value might be considered of interest, even if of low amplitude; while high frequency departures are less important, unless they are of high amplitude.

Bickel, Ritov and Stoker (2001) follow this point of view by proposing a general approach to testing semiparametric hypotheses within a nonparametric model in the context of observing n i.i.d. observations. They proposed that tests should be tailored in such a way that on the $n^{-1/2}$ scale power can be concentrated in a few selected directions with some power reserved at the same scale in all other directions. In that paper this methodology was applied to two classical problems, testing goodness-of-fit to a parametric model and testing independence. In this paper we show how this approach can be applied rigorously to generate tests for one of the simplest classical econometric hypotheses, that the conditional expectation of a response given a number of explanatory variables is in fact dependent only on a known subset of these. Such index hypotheses have been widely discussed in the econometric literature. A recent review and a more standard type of test may be found in Ait Sahalia, Bickel and Stoker (2001).

Formally we consider the following problem. We observe X_i, i.i.d. $i = 1, \ldots, n$ where $X = (W, Y)$ where $W = (U, V)$, $U \in \mathbf{R}^{d_u}$, $V \in \mathbf{R}^{d_v}$ and $Y \in \mathbf{R}$. Assume that the joint probability density function (with respect to Lebesgue measure) of W and Y is given $p\,(w, y; f, \nu) = f\,(w, y - \nu\,(w))$. Let $\mathcal{P}$ be the collection of all distribution functions with such a density (i.e. for all possible f and ν satisfying the regularity assumption specified below). 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 $E\,(Y \mid W) = E\,(Y \mid U)$. These models contain the special case $E(Y \mid W) = 0$. The extension of this last model where $E(Y \mid W)$ follows a parametric model was treated by Härdle and Mammen (1993).

In the general framework of Bickel, Ritov and Stoker (2001), we test $\mathcal{P}_0$ a proper set of probability functions, against "everything", $\mathcal{P} = \mathcal{M} \equiv \{\text{All probabilities dominated by } \mu\}$ or at least $\mathcal{P}$ such that the tangent space is saturated,

$$\overset{\bullet}{\mathcal{P}}\,(P) = L_2^0(P) = \{h \in L_2(P) : P(h) = 0\}.$$

See Bickel, Klaassen, Ritov and Wellner (1993) for a general discussion of semiparametric models and tangent spaces.

If $\overset{\bullet}{\mathcal{P}}_0\,(P)$ is the tangent space at $P_0 \in \mathcal{P}_0$, we can write the efficient score function at P_0 in a direction $a(\cdot) \in L_2^0(P)$, corresponding to a submodel of $\mathcal{P}$ containing P_0 as

$$\begin{aligned} Z_n(a, P_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (a - P_0(a) - \Pi(a, P_0))(X_i) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Pi^{\perp}(a, P_0)(X_i) \end{aligned} \tag{1.1}$$

for a in the tangent space, or at least in a subset $\mathcal{A}$ spanning the tangent space. Here, $\Pi(a, P_0)$ is the projection operator from $L_2(P_0)$ to the subspace $\overset{\bullet}{\mathcal{P}}_0\ (P_0)$ of $L_2^0(P_0)$ and $\Pi^\perp$ is the projection to the orthocomplement of $\overset{\bullet}{\mathcal{P}}_0\ (P_0)$ within $L_2^0(P_0)$. The identity uses $\Pi^\perp(h, P_0) = \Pi^\perp(h + c, P_0)$ for all c.

Call $Z_n(\cdot, P_0)$, the *score process.* In general, $Z_n(a, P_0)$ is not computable given the data, but if $\hat{P} \in \mathcal{P}_0$ is an estimate of P_0 we can consider

$$\hat{Z}_n(a) \equiv Z_n(a, \hat{P}) \tag{1.2}$$

defined on $\mathcal{A}$.

Typically we consider a parametric sub-family $\{a_\gamma, \gamma \in \Gamma\} \subset \mathcal{A}$. Having the score process, we can construct tailor made tests by considering any functional $T(\hat{Z}_n)$. For example two standard methods for constructing tests are

1. Cramér–von Mises type (or χ^2 goodness-of-fit) tests: $\int \omega(\gamma)\hat{Z}_n^2(a_\gamma)\, d\mu(\gamma)$ for some weight function ω and measure μ.

2. Kolmogorov–Smirnov type (or union-intersection) tests: $sup_{\gamma\in\Gamma}\omega(\gamma)|\hat{Z}_n(a_\gamma)|$.

This paper discusses the construction of $\hat{Z}_n(\cdot)$ in section 2 and establishes the properties needed for its use, in section 3. The definition of the actual test is left to the user although section 4 discusses setting of critical values and gives the results of a small simulation on some natural candidate tests. A brief discussion in section 5 and an appendix complete the paper.

2. Preliminaries

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

Lemma 2.1: *We have*

$$\begin{aligned} \overset{\bullet}{\mathcal{P}} &= \left\{a\left(W,Y\right) : E_P\left[a^2\left(W,Y\right)\right] < \infty, E_P\left[a\left(W,Y\right)\right] = 0\right\} \\ \overset{\bullet}{\mathcal{P}}_0 &= \Big\{ a\left(W,Y\right) = h\left(W, Y-\nu\left(U\right)\right) + \ell'_{Y|W}\left(Y-\nu\left(U\right)\right)g\left(U\right) : \\ & \qquad a, h \in \overset{\bullet}{\mathcal{P}}, \ \textstyle\int yh\left(W,y\right)dy = 0, \ a.s.\Big\} \\ \overset{\bullet}{\mathcal{P}}_0^{\perp} &= \Big\{ a\left(W,Y\right) = \left[b\left(W\right) - E\left(b(W)\mid U\right)\right]\left(Y - E\left(Y\mid U\right)\right) : a,b \in \overset{\bullet}{\mathcal{P}}\Big\}. \end{aligned}$$

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

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

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

if and only if $a\left(W,Y\right) = b\left(W\right)\left(Y-\nu\left(U\right)\right)$, a.s. This latter object is orthogonal to all functions in $\overset{\bullet}{\mathcal{P}}$ of the form $\ell'_y\left(W, Y-\nu\left(U\right)\right)g\left(U\right)$ if and only if $E\left(b\left(W\right)\mid U\right) = 0$ *a.s.* which follows from the fact that for any p.d.f. q (with mean 0), we have $\int xq'\left(x\right)dx = -1$. *Q.E.D.*

Therefore, our *score process* is defined by

$$\hat{Z}_n\left(a\right) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[a\left(W_i\right) - E_{\hat{P}}\left(a\left(W\right)\mid U_i\right)\right]\left(Y_i - E_{\hat{P}}\left(Y\mid U_i\right)\right) \tag{2.1}$$

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

3. Main Result

We consider the case that $\mathcal{A}$ doesn't depend on P_0, the joint distribution of $(W, Y - E(Y\mid W) + E(Y\mid U))$. We will consider the standard Nadaraya–Watson estimates of $E_P(Y\mid U=u)$, $E_P(a(W)\mid U=u)$. Let K be a symmetric kernel with bounded support on R and α vanishing moments, that is,

$$K : R \to R$$

a) $K = 0$ outside $[-1, 1]$

b) $\int K(u)du = 1$

c) $\int u^j K(u)du = 0$, for $1 \leq j \leq \alpha$.

Let $\mathbb{K}_d : R^d \to R$ be the product kernel

$$\mathbb{K}_d(x_1, \ldots, x_d) = \prod_{j=1}^{d} K(x_j)$$

and $\mathbb{K}_d(\mathbf{x}; \sigma) \equiv \sigma^{-d}\mathbb{K}_d(\mathbf{x}/\sigma)$. We abuse notation writing $\hat{p}(\mathbf{w}, y)$ for the estimated joint density of $(\mathbf{W}, Y)$, $\hat{p}(\mathbf{u}, y)$ for the marginal estimated joint density of $(\mathbf{U}, Y)$ and dropping the subscript d in $\mathbb{K}_d$ when it is implicit. Then,

$$\begin{aligned} \hat{p}(\mathbf{w}, y) &\equiv \int \mathbb{K}(\mathbf{w} - \mathbf{w}', y - y', \sigma) dP_n(\mathbf{w}', y') \\ \hat{p}(\mathbf{u}, y) &= \int \mathbb{K}(\mathbf{u} - \mathbf{u}', y - y'; \sigma) dP_n(\mathbf{u}', y') \end{aligned}$$

where we also use the convention that $P_n(\mathbf{w}, y)$ refers to the joint empirical distribution of $(\mathbf{W}, Y)$, etc. Finally,

$$\begin{aligned} \hat{E}(Y \mid \mathbf{U} = \mathbf{u}) &\equiv \int y\hat{p}(\mathbf{u}, y)dy \Big/ \int \hat{p}(\mathbf{u}, y)dy \\ &= \int y\mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma)dP_n(\mathbf{u}', y) \Big/ \hat{p}(\mathbf{u}) \end{aligned}$$

where $\hat{p}(\mathbf{u}) \equiv \int \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma)dP_n(\mathbf{u}')$. Here we use

$$\int y\hat{p}(\mathbf{u}, y)dy = \int y \int \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma)K(y - y'; \sigma)dP_n(\mathbf{u}', y')dy \tag{3.1}$$

and

$$\int yK(y - y'; \sigma)dy = y'.$$

We define $\hat{E}(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u})$ similarly. We introduce the following assumptions.

I0: $\int yf(W, y)\, dy = 0$ a.s., $\int \left(|w|^2 + y^2\right) f(w, y)\, dy\, dw < \infty$, and $\int \nu^2(w) f(w, y)\, dy\, dw < \infty$.

I1: The support of the distribution of $\mathbf{U}$ is a fixed compact say $[-1,1]^{d_{\mathbf{u}}}$ for all $P \in \mathcal{P}$.

I2: All $P \in \mathcal{P}$ are absolutely continuous with respect to Lebesgue measure and

a) the density $p(\mathbf{u})$ has bounded derivatives of order greater than $\frac{3}{2}d_{\mathbf{u}}$.

b) $Y \in L_2(P)$ and $\mathbf{u} \to E(Y \mid \mathbf{U} = \mathbf{u})$ is continuous.

Moreover

I3: There exists $\epsilon(P) > 0$ such that $\epsilon \leq p(\mathbf{u}) \leq \frac{1}{\epsilon}$ for all $\mathbf{u} \in [-1,1]^{d_{\mathbf{u}}}$.

I4: $\sup\{\|a\|_\infty : a \in \mathcal{A}\} < \infty$ and $\mathcal{A}^* \equiv \{a(\mathbf{u}) - Ea(\mathbf{W} \mid \mathbf{U} = \mathbf{u})\}$ is a VC class of functions in the sense of the definition on p. 141 of van der Vaart and Wellner (1996).

Discussion of I1–I4

1. Conditions (I1) and (I3) are very restrictive. Our argument suggests they can be weakened to a tail condition on $p(\mathbf{u})$ but at the cost of a great deal of technical labor. Alternatively test statistics which pay no attention to regions where $\mathbf{U}$ has low density, i.e., such that $a(\mathbf{W}) = 0$ for such $\mathbf{U}$ can be used.

2. Condition (I2) unfortunately seems necessary. It becomes more and more stringent as the dimension of $\mathbf{U}$ increases.

3. Condition (I4) is somewhat more restrictive than, say, universal Donsker. But all the usual classes, indicators of rectangles, etc., satisfy it given the smoothness conditions on $p(\mathbf{w}, \mathbf{u})$.

Then:

Theorem 3.1: *Under* I1–I4, *if* $\sigma = \Omega\left(n^{-\frac{1}{2d+d_{\mathbf{u}}}}\right)$ *and* K *has* α *vanishing moments where* $\alpha > \frac{3}{2}d_{\mathbf{u}}$ *then,*

$$\sup_{\mathcal{A}}\{|\hat{Z}_n(a) - Z_n(a, P_0)|\} = o_p(n^{-1/2}).$$

Proof. Write

$$
\begin{aligned}
&\hat{Z}_n(a) - Z_n(a, P_0) \\
= \;& \int (\hat{E}(Y \mid \mathbf{U} = \mathbf{u}) - E(Y \mid \mathbf{U} = \mathbf{u})(a(\mathbf{w}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) dP_n(\mathbf{w}) \\
& + \int (\hat{E}(a(\mathbf{W}) \mid \mathbf{U} - \mathbf{u}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))(Y - E(Y \mid \mathbf{U} = \mathbf{u})) dP_n(\mathbf{w}) \\
& + \int (\hat{E}(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))(\hat{E}(Y \mid \mathbf{U} = \mathbf{u}) - E(Y \mid \mathbf{U} = \mathbf{u})) dP_n(\mathbf{u}) \\
= \;& I + II + III, \quad \text{say.}
\end{aligned}
$$

We argue now that under our conditions $\sup_{\mathcal{A}} |I|$, $\sup_{\mathcal{A}} |II|$, and $\sup_{\mathcal{A}} |III|$ are all $o_p(n^{-1/2})$.

To do so we require a lengthy argument some of which will be given in the appendix.

Let

$$
\bar{p}(\mathbf{u}, y) = \int \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma) p(\mathbf{u}', y) d\mathbf{u}'
$$

and

$$
a^*(\mathbf{w}) \equiv a(\mathbf{w}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}).
$$

Then define

$$
\begin{aligned}
\Delta_n^{(1)}(\mathbf{a}) &\equiv \int \left\{ \frac{\int y \hat{p}(\mathbf{u}, y) dy}{\bar{p}(\mathbf{u})} - \frac{\int y \bar{p}(\mathbf{u}, y) dy}{\bar{p}(\mathbf{u})} \right\} a^*(\mathbf{w}) dP_n(\mathbf{w}) \\
&= \int \bar{p}^{-1}(\mathbf{u}) \int y \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma) d(P_n(\mathbf{u}', y) - P(\mathbf{u}', y)) a^*(\mathbf{w}) dP_n(\mathbf{w}).
\end{aligned} \tag{3.2}
$$

Similarly define

$$
\begin{aligned}
\Delta_n^{(2)}(\mathbf{a}) &= - \int \frac{\int y \bar{p}(\mathbf{u}, y) dy}{\bar{p}^2(\mathbf{u})} (\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u})) a^*(\mathbf{w}) dP_n(\mathbf{w}) \\
\Delta_n^{(3)}(\mathbf{a}) &= - \int \frac{\int y (p(\mathbf{u}, y) - \bar{p}(\mathbf{u}, y)) dy}{\bar{p}(\mathbf{u})} a^*(\mathbf{w}) dP_n(\mathbf{w}) \\
\Delta_n^{(4)}(\mathbf{a}) &= - \int \frac{\int y p(\mathbf{u}, y) dy}{\bar{p}(\mathbf{u}) p(\mathbf{u})} (\bar{p}(\mathbf{u}) - p(\mathbf{u})) a^*(\mathbf{w}) dP_n(\mathbf{w})
\end{aligned} \tag{3.3}
$$

$$
\Delta_n^{(5)}(\mathbf{a}) = - \int \frac{\int y (\hat{p}(\mathbf{u}, y) - \bar{p}(\mathbf{u}, y))}{\hat{p} \bar{p}(\mathbf{u})} (\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u})) a^*(\mathbf{w}) dP_n(\mathbf{w}) \tag{3.4}
$$

$$\Delta_n^{(6)}(\mathbf{a}) = \int \frac{\left(\int y\bar{p}(\mathbf{u}, y)dy\right)}{\bar{p}^2(\mathbf{u})\hat{p}(\mathbf{u})}(\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u}))^2 a^*(\mathbf{w})dP_n(\mathbf{w}). \tag{3.5}$$

Some algebra shows

$$I = \sum_{j=1}^{6} \Delta_n^{(j)}(\cdot).$$

For $g : \mathcal{A} \to R$ let $\|g\|_{\mathcal{A}} = \sup_{\mathcal{A}} |g(a)|$. We shall show that $\|\Delta_n^{(j)}\|_{\mathcal{A}} = o_p(n^{-1/2})$ for $j = 1, \ldots, 6$ and hence $\|I\|_{\mathcal{A}} = o_p(n^{-1/2})$. We can similarly establish $\|II\|_{\mathcal{A}} = o_p(n^{-1/2})$ and then argue in detail that $\|III\|_{\mathcal{A}} = o_p(n^{-1/2})$ establishing the theorem.

We proceed with $\Delta_n^{(1)}$ and note that

$$\Delta_n^{(1)}(\mathbf{a}) = \int \bar{p}^{-1}(\mathbf{u})y\mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma)a^*(\mathbf{w})d(P_n - P)(\mathbf{u}', y)d(P_n - P)(\mathbf{w}) \tag{3.6}$$

since for all $\mathbf{u}$

$$\int a^*(\mathbf{u}, \mathbf{v})p(\mathbf{v} \mid \mathbf{u})d\mathbf{v} = 0. \tag{3.7}$$

In the appendix we show that

$$\|\Delta_n^{(1)}(\cdot) - \tilde{\Delta}_n^{(1)}(\cdot)\|_{\mathcal{A}} = o_p(n^{-1/2}) \tag{3.8}$$

where

$$\tilde{\Delta}_n^{(1)}(a) = \frac{2}{n^2}\sum_{i<j} C((\mathbf{W}_i, Y_i), (\mathbf{W}_j, Y_j), a^*; \sigma)$$

with

$$\begin{aligned} &C((\mathbf{w}, y), (\mathbf{w}', y'), a^*; \sigma) \\ &\equiv \frac{1}{2}\left\{\frac{p(\mathbf{u})}{\bar{p}(\mathbf{u})}(yK(\mathbf{u} - \mathbf{u}'; \sigma) - E(YK(\mathbf{u} - \mathbf{U}; \sigma)))a^*(\mathbf{w})\right. \\ &\qquad \left. + \frac{p(\mathbf{u}')}{\bar{p}(\mathbf{u})}(y'K(\mathbf{u} - \mathbf{u}'; \sigma) - EYK(\mathbf{u}' - \mathbf{U}; \sigma))a^*(\mathbf{w}')\right\} \end{aligned} \tag{3.9}$$

is a degenerate U statistic process and that by Theorem 2.5(b) of Arcones and Gine (1995), $\|\tilde{\Delta}_n^{(1)}\|_{\mathcal{A}} =$

$o_p(n^{-1/2})$ under our conditions and hence $\|\Delta_n^{(1)}\|_{\mathcal{A}} = o_p(n^{-1/2})$. We now turn to $\Delta_n^{(2)}$. Again, by (3.7),

$$\begin{aligned}(3.10)\qquad \Delta_n^{(2)}(\mathbf{a}) &= -\int \frac{\int y\bar{p}(\mathbf{u},y)dy}{\bar{p}^2(\mathbf{u})}(\hat{p}(\mathbf{u})-\bar{p}(\mathbf{u}))a^*(\mathbf{w})d(P_n-P)(\mathbf{w}) \\ &= -\int\int\left(\frac{\int y\bar{p}(u,y)dy}{\bar{p}^2(\mathbf{u})}\right)a^*(\mathbf{w})\mathbb{K}(\mathbf{u}-\mathbf{u}';\sigma)d(P_n-P)(\mathbf{u}')d(P_n-P)(\mathbf{w}).\end{aligned}$$

This has the same structure as $\Delta_n^{(1)}$ and it can be similarly shown that $\|\Delta_n^{(2)}\|_{\mathcal{A}} = o_p(n^{-1/2})$. On the other hand, $\Delta_n^{(3)}$ and $\Delta_n^{(4)}$ can both be written in the form $\int Q(a^*;\sigma)(\mathbf{w})d(P_n-P)(\mathbf{w})$ where $\{Q(a^*;\sigma) : a \in \mathcal{A},\ 0 \le \sigma \le 1\}$ (with $Q(a^*,0) \equiv 0$) is a universal Donsker class in view of $(I4)$. Since in both cases

$$\int Q^2(a^*;\sigma)(\mathbf{w})p(\mathbf{w})d\mathbf{w} \to 0$$

as $\sigma \to 0$ we can conclude by theorem of van der Vaart and Wellner (1996) that $\|\Delta_n^{(j)}\|_{\mathcal{A}} = o_p(n^{-1/2})$ for $j = 3, 4$. Next,

$$\begin{aligned}(3.11)\qquad |\Delta_n^{(5)}(\mathbf{a})| &\le \left\|\frac{a^*}{2}\right\|_\infty \left(\int \frac{\left(\int y(\hat{p}(\mathbf{u},y)-\bar{p}(\mathbf{u},y))dy\right)^2}{\bar{p}^2(\mathbf{u})}dP_n(\mathbf{u})\right. \\ &\qquad\qquad \left. + \int \frac{(\hat{p}(\mathbf{u})-\bar{p}(\mathbf{u}))^2}{\hat{p}^2(\mathbf{u})}dP_n(\mathbf{u})\right).\end{aligned}$$

By (I2) and (3.13) below, $\|\hat{p}(\mathbf{u})-\bar{p}(\mathbf{u})\|_\infty = o_p(1)$. Hence, by (I3) the denominators of both terms in (3.11) are bounded away from 0 with probability tending to 1. Write

$$\begin{aligned}(3.12)\qquad \Delta_{n1}^{(5)} &\equiv \int\left(\int y(\hat{p}(\mathbf{u},y)-\bar{p}(\mathbf{u},y))dy\right)^2 dP_n(\mathbf{u}) \\ &= \frac{1}{n^3}\sum_{i,j,k} A_{ij}A_{kj}\end{aligned}$$

where

$$A_{ij} \equiv (Y_i\mathbb{K}(\mathbf{U}_j-\mathbf{U}_i;\sigma) - E(Y_i\mathbb{K}(\mathbf{U}_j-\mathbf{U}_i;\sigma) \mid \mathbf{U}_j)).$$

Note that $EA_{ij}A_{kj} = 0$ unless $i = k$. Thus

$$\begin{aligned}E\Delta_n^{(5)} &\le n^{-2}K^2(0;\sigma)EY_1^2 + n^{-1}EY_1^2\mathbb{K}(\mathbf{U}_1-\mathbf{U}_2;\sigma) \\ &= O(n^{-2}\sigma^{-2d_{\mathbf{u}}}) + O(n^{-1}\sigma^{-d_{\mathbf{u}}}) = o(n^{-1/2})\end{aligned}$$

by the assumption $\sigma = \Omega(n^{-1/2\alpha + d_{\mathbf{u}}})$, $\int(\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u}))^2 dP_n(\mathbf{u})$ is bounded similarly and $\|\Delta_n^{(5)}\|_{\mathcal{A}} = o_p(n^{-1/2})$ follows. Similarly,

$$|\Delta_n^{(6)}(\mathbf{a})| \leq \|a^*\|_\infty \sup_{\mathbf{u}} (\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u}))^2 \sup_{\mathbf{u}} \bar{p}^{-2}(\mathbf{u}) \sup_{\mathbf{u}} \hat{p}^{-2}(\mathbf{u}) \frac{1}{n^2} \sum_{i,j} Y_i \mathbb{K}(\mathbf{u}_i - \mathbf{u}_j; \sigma).$$

Again by (I2) and (7.1) of Härdle and Mammen (1993),

$$\|\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u})\|_\infty = O_p(n^{-\frac{\alpha}{2\alpha + d_{\mathbf{u}}}} \log n). \tag{3.13}$$

By (I3) the second two sups are $O_p(1)$, the first sup is $O_p(n^{-2\alpha(2\alpha + d_{\mathbf{u}})^{-1}} y^2 n)$. Finally, the last term is $O_p(1)$. Thus we conclude since $\alpha > \frac{3}{2} d_{\mathbf{u}}$ that $\|\Delta_n^{(6)}\|_{\mathcal{A}} = o_p(n^{-1/2})$ and $\sup_{\mathcal{A}} I = o_p(n^{-1/2})$.

For II we proceed similarly. Here

$$\begin{aligned} II(\mathbf{a}) &= \sum_{j=1}^{6} \tilde{\Delta}_n^{(j)}(\mathbf{a}) \\ \tilde{\Delta}_n^{(1)}(a) &= \int \bar{p}^{-1}(\mathbf{u}) \int a(\mathbf{w}) \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma) d(P_n - P)(\mathbf{u}', \mathbf{v}) e(y, \mathbf{u}) dP_n(y, \mathbf{u}) \end{aligned} \tag{3.14}$$

where $e(y, \mathbf{u}) \equiv y - E(Y \mid \mathbf{U} = \mathbf{u})$ and this is dealt with just as $\Delta_n^{(1)}$ was.

The same kind of argument applies to the terms corresponding to $\tilde{\Delta}_n^{(2)} - \tilde{\Delta}_n^{(6)}$. We finally turn to III.

$$\begin{aligned} |III(a)| &\leq \frac{1}{2} \left(\int (\hat{E}(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) \right)^2 dP_n(\mathbf{u}) \\ &\qquad + \int (\hat{E}(Y \mid \mathbf{U} = \mathbf{u}) - E(Y \mid \mathbf{U} = \mathbf{u}))^2 dP_n(\mathbf{u}). \end{aligned}$$

Decompose as for I and II. For instance,

$$
\begin{aligned}
(3.15)\qquad & \int (\hat{E}(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))^2 dP_n(\mathbf{u}) \\
& \leq C \Bigg(\int \left(\int \frac{a(\mathbf{w})}{\bar{p}(\mathbf{u})} \mathbb{K}(\mathbf{u} - \mathbf{u}'; \sigma) d(P_n - P)(\mathbf{u}', \mathbf{v}) \right)^2 dP_n(\mathbf{u}) \\
& \quad + \int \left(\int a(\mathbf{w})(\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u})) \frac{\bar{p}(\mathbf{w})}{\bar{p}^2(\mathbf{u})} d\mathbf{v} \right)^2 dP_n(\mathbf{u}) \\
& \quad + \int \left(\int a(\mathbf{w}) \frac{(p - \bar{p})}{\bar{p}(\mathbf{u})} (\mathbf{w}) d\mathbf{v} \right)^2 dP_n(\mathbf{u}) \\
& \quad + \int \left(\int a(\mathbf{w}) \frac{p(\mathbf{w})}{\bar{p}p(\mathbf{u})} (\bar{p}(\mathbf{u}) - p(\mathbf{u})) d\mathbf{v} \right)^2 dP_n(\mathbf{u}) \\
& \quad + \int \left(\int a(\mathbf{w}) \frac{(\hat{p}(\mathbf{w}) - \bar{p}(\mathbf{w}))(\hat{p}(\mathbf{u}) - \bar{p}(\mathbf{u}))}{\hat{p}\bar{p}(\mathbf{u})} d\mathbf{v} \right)^2 dP_n(\mathbf{u}) \\
& \quad + \int \left(\int \frac{a(\mathbf{w})p(\mathbf{w})}{\hat{p}^2\bar{p}^2(\mathbf{u})} d\mathbf{v} (\hat{p}(\mathbf{u}) - p(\mathbf{u})^2 d\mathbf{v} \right)^2 dP_n(\mathbf{u}) \Bigg)
\end{aligned}
$$

In the appendix we show that

$$
(3.16)\qquad \sup_{\mathcal{A}, \mathbf{u}} \left(\int \frac{a(\mathbf{w})}{\bar{p}(\mathbf{u})} \mathbf{K}(\mathbf{u} - \mathbf{u}'; \sigma) d(P_n - P)(\mathbf{u}', \mathbf{v}) \right)^2 = o_p(n^{-1/2})
$$

by using large deviation bounds on the empirical process applied to $\{a(\mathbf{u}, \cdot)\mathbb{K}(\mathbf{u} - \cdot; \sigma) : a \in \mathcal{A}, \mathbf{u} \in K\}$.

The remaining terms are more straightforward. We can pull out the inf of $\hat{p}$ and $\bar{p}$ as well as the L_∞ norm of a and then argue as we did for $\Delta_n^{(5)}$. The argument for the term which involves $\hat{E}(Y \mid \cdot)$ is easy. The theorem follows. *Q.E.D.*

A problem we have not yet faced is how to set critical values for our tests. As the discussion in Bickel, Ritov and Stoker (2001) indicates two bootstraps are in principle possible. In the current model the "wild" bootstrap—see Härdle and Mammen (1993) is also possible. We chose to implement the version proposed by Bickel and Ren (2001), i.e., simulate the distribution of $\sqrt{n}(\hat{Z}_n^*(\cdot) - \hat{Z}_n(\cdot))$ where $\hat{Z}_n^*$ is the $\hat{Z}_n$ process defined for the bootstrap sample $X_1^*, \ldots, X_n^*$ from the empirical of $X_1, \ldots, X_n$ where $X_j = (\mathbf{V}_j, Y_j)$. Unfortunately the conditions of Theorems 1 and 2 of Bickel and Ren are not satisfied. We give a more special argument.

Note that

$$\hat{Z}_n^*(a) = \int (y - \hat{E}^*(Y \mid \mathbf{U} = \mathbf{u}))(a(\mathbf{w}) - \hat{E}^*(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))dP_n^*(\mathbf{w}, y).$$

Let

$$\tilde{Z}_n(a) = \int (y - \hat{E}^*(Y \mid \mathbf{U} = \mathbf{u}))(a(\mathbf{w}) - E^*(a(\mathbf{w}) \mid \mathbf{U} = \mathbf{u}))dP_n(\mathbf{w}, y).$$

Showing that

$$\tilde{Z}_n(a) - Z_n(a, P_0) = o_p(n^{-1/2})$$

can be done by essentially the same argument as that used for Theorem 3.1. For instance, define $\tilde{\Delta}_n^{(1)}$ corresponding to $\Delta_n^{(1)}$ by simply replacing P_n by P_n^* in the inner differential. We are left with showing that

$$\int (E(Y \mid \mathbf{U} = \mathbf{u}) - \hat{E}^*(Y \mid \mathbf{U} = \mathbf{u}))(a(\mathbf{w}) - E(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))d(P_n^* - P_n)(\mathbf{w}) = o_p(n^{-1/2})$$

$$\int (y - E(Y \mid \mathbf{U} = \mathbf{u}))(Ea(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}) - \hat{E}^*(a(\mathbf{W}) \mid \mathbf{U} = \mathbf{u}))d(P_n^* - P_n)(\mathbf{u}, y) = o_p(n^{-1/2})$$

and

$$\int (E(Y \mid \mathbf{U} = \mathbf{u}) - \hat{E}^*(Y \mid \mathbf{U} = \mathbf{u}))(E(a(\mathbf{w}) \mid \mathbf{U} = \mathbf{u}) - \hat{E}^*(a(\mathbf{w}) \mid \mathbf{U} = \mathbf{u}))d(P_n^* - P_n)(\mathbf{w}) = o_p(n^{-1/2}).$$

These terms can all be approximated by quantities of the form appearing on the right in $\Delta_n^{(1)} - \Delta_n^{(5)}$ and the validity of the bootstrap approximation established.

4. Critical Values and Simulations

We checked the behaviour of different estimators using a small Monte–Carlo experiment. We consider a sample of 500 independent observations from (U, V, Y) where $Y = \nu_\lambda(U, V) + \epsilon$, where U, V, and ϵ are independent, $U, V \sim U(0, 1)$, $\epsilon \sim N(0, 1)$, and $\nu_\lambda(u, v) = 0.8 \sin(\lambda u) \sin(\lambda v)$, where $\lambda = 0, \pi/2, \pi, 6\pi$. Of course, $\lambda = 0$ is the null assumption. The three regression surfaces are shown in Figure 1.

The three test statistics we examined were all based on partition of the unit square to 10 × 5 blocks. Where the support of U was divided to 10 blocks. The reason that the partition was asymmetrical in the two variables, was in the way the bias the partition introduced. The discretization of the range of U introduce a bias, since if it is not fine enough, a distribution in which Y and W are conditionally independent given U, may be not conditionally independent given the blocks. Condition (I2) is necessary to ensure that the test will be asymptotically unbiased. On the other hand 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 into blocks, one simple test is a standard ANOVA for only U effect (i.e., no V effect and no interaction). This is our first test statistic. The second is the Kolmogorov-Smirnov like test with the quadrates $\{\mathbf{1}(u \geq \gamma_1,\, v \geq \gamma_2)\}$. The third is another Kolmogorov–Smirnov statistic with rectangles: $\{\mathbf{1}(\gamma_< u \leq \gamma_2,\, \gamma_3 < v \leq \gamma_4)\}$.

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

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

Then the three test statistics are:

$$\begin{aligned}
F &= \frac{\sum_{kl} \bar{Y}_{kl\cdot}^2 . n_{kl} - \sum_k \bar{Y}_{k\cdot\cdot}^2 . n_{k\cdot}}{\sum_{klm} Y_{klm}^2 - \sum_k \bar{Y}_{k\cdot\cdot}^2 . n_{k\cdot}} \\
KS_1 &= \max_{kl} \left| \sum_{k'=k}^{K} \sum_{l'=l}^{L} \sum_{m=1}^{n_{k'l'}} (Y_{k'l'm} - \bar{Y}_{k\cdot\cdot}) \right| \\
KS_2 &= \max_{k_1 l_1 k_2 l_2} \left| \sum_{k'=l_1}^{l_2} \sum_{l'=k_1}^{k_2} \sum_{m=1}^{n_{k'l'}} (Y_{k'l'm} - \bar{Y}_{k\cdot\cdot}) \right|
\end{aligned}$$

The three deviations were supposed to check the strength and weakness of these tests. The first KS test was appropriate for deviations like the one with $\lambda = \pi/2$, in which the corners are different from the average. The second KS was supposed to show its strength against deviation which are concentrated in the center as the case of $\lambda = \pi$. Finally, the F test diffuse its strength among 40 degree of freedoms. 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 was however two modifications. Also, theoretically the number of obsevation in a cell should increase to ∞, in practice is finite, and may be quite small (in our simulation there were, on the 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 that, we multiplied each observation in the kl cell by $\sqrt{n_{kl}/(n_{kl}-1)}$. See Silverman (1981) for a similar correction. The KS type tests were not conservative without the inflation. Of course the F test is invariant for this correction. The second modification was that the bootstrap was only on for the Y values (hence we conducted a conditional test on the W's).

Rejection was defined if the test statistics was one of the $100(1-\alpha)\%$ larger values among 200 observations where α is the declared level. The randomization (both the sampling and the bootstraping) were common to the twenty four combinations of test statistics and values of λ and α.

The powers at level $\alpha = .1$ and $\alpha = .05$ of the various statistics are given in Table I.

5. Discussion

The simulation results show that we are able to tailor tests to set expected departures.

The minimax F test does indeed perform far better than the other two for the $\lambda = 6\pi$ case but the relevance of this least favorable departure is unclear. All we can hope for is good power in interesting directions when the signal to noise ratio is moderate and in uninteresting directions when the signal to noise is really high.

Technical though it is our discussion does not cover the more important case where the index is unknown, i.e., $\mathbf{U} = \mathbf{W}^T\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ unknown. At the scale we are working with the distribution of $\boldsymbol{\theta}$ will have an effect but again we expect to be able to tailor though formulating and checking regulatory conditions becomes even more tedious.

APPENDIX: Proof of Theorem 6.1 Details

Proof of (3.8) and $\|\Delta_n^{(1)}\|_{\mathcal{A}} = o_p(n^{-1/2})$.

$$\begin{aligned} \Delta_n^{(1)}(a) &= n^{-2}\textstyle\sum_{i,j} \frac{p(\mathbf{U}_i)}{\bar{p}(\mathbf{U}_i)}(Y_j\mathbb{K}(\mathbf{U}_i - \mathbf{U}_j;\sigma) \\ &- E(Y_j\mathbb{K}(\mathbf{U}_i - \mathbf{U}_j;\sigma) \mid \mathbf{U}_i)a^*(\mathbf{W}_i) \\ &= \tilde{\Delta}_n^{(1)}(a) - n^{-2}\mathbb{K}(0;\sigma)\textstyle\sum_{i=1}^n \frac{p(\mathbf{U}_i)}{\bar{p}(\mathbf{U}_i)}(Y_i - E(Y_i \mid \mathbf{U}_i))a^*(\mathbf{W}_i). \end{aligned} \tag{A.1}$$

The second term here is evidently $O_p(n^{-1}\sigma^{-d_{\mathbf{u}}}) = o_p(n^{-1/2})$ by (I3).

Note that

$$\begin{aligned} \|C(\mathbf{x}, \mathbf{x}', a^*;\sigma)\|_{\mathcal{A}} &\leq \tfrac{1}{2}\sup_{\mathcal{A}} \|a^*\|_\infty \Big\| \tfrac{p(\mathbf{u})}{\bar{p}(\mathbf{u})}(y\mathbb{K}(\mathbf{u} - \mathbf{u}';\sigma) \\ &- E(YK(\mathbf{u} - \mathbf{U};\sigma)) + \tfrac{p(\mathbf{u}')}{\bar{p}(\mathbf{u}')}(y'\mathbb{K}(\mathbf{u} - \mathbf{u}';\sigma - E(YK(\mathbf{u} - \mathbf{u}';\sigma))\Big\| \\ &\leq \sup_{\mathcal{A}} \|a^*\|_\infty \left(\tfrac{p(\mathbf{u})}{\bar{p}(\mathbf{u})} + \tfrac{p(\mathbf{u}')}{\bar{p}(\mathbf{u}')}\right)(|y| + E|Y|)\sigma^{-d_{\mathbf{u}}}. \end{aligned} \tag{A.2}$$

By Theorem 2.5(b) of Arcones and Gine (1995)

$$n(\log\log n)^2 E\|\tilde{\Delta}_n^{(1)}\|_{\mathcal{A}}^p \leq \sigma^{-pd_{\mathbf{u}}} 2E|Y|\epsilon^{-2}(P)\sup_{\mathcal{A}} \|a^*\|_\infty$$

for $0 < p < 2$ where $\epsilon(P)$ is the lower bound on $p(\mathbf{u})$. Hence,

$$\|\tilde{\Delta}_n^{(1)}\|_{\mathcal{A}} = O_p(n^{-1}\sigma^{-d_{\mathbf{u}}}(\log\log n)^2) = o_p(n^{-1/2}).$$

Proof of (3.16). Let $\mathcal{A}$ have metric entropy for Q given by

$$N(\mathcal{A}, L_2(Q)).$$

Let $\tilde{\mathcal{A}}_n = \left\{\frac{a(\mathbf{u},\cdot)}{\bar{p}(\mathbf{u})}\mathbb{K}(\mathbf{u} - \cdot;\sigma) : a \in \mathcal{A}, \mathbf{u} \in R^{d_{\mathbf{u}}}\right\}$. Given $\epsilon > 0$ by the smoothness of $\mathbb{K}$ we can find $\mathbf{u}_1^{(\epsilon)}, \ldots, \mathbf{u}_n^{(\epsilon)} \ni$ for some $j(\mathbf{u})$, $\|\mathbb{K}(\mathbf{u} - \cdot;\sigma) - \mathbb{K}(\mathbf{u}_j(\epsilon) - \cdot;\sigma)\|_\infty \leq \epsilon$ and $M = \Omega((\epsilon\sigma)^{-d_{\mathbf{u}}})$. Therefore

$$N(\tau, \tilde{\mathcal{A}}, L_2(Q)) = \Omega(N(\tau, \mathcal{A}, L_2(Q)) \cdot \Omega((\tau\sigma)^{-d_{\mathbf{u}}})$$

where $a_n = \Omega(b_n)$ iff $a_n = O(b_n)$, $b_n = O(a_n)$ and we can conclude from Theorem 2.14.9 of van der Vaart and Wellner (1996) that if $\mathbb{G}_n$ is the empirical process $\sqrt{n}(P_n - P)$ then,

$$\|\mathbb{G}_n\|_{\tilde{\mathcal{A}}} = O_p(\sigma^{-d_{\mathbf{u}}}). \tag{A.3}$$

Now (A.3) implies (3.16) since the left-hand side is $O_p(n^{-1}\sigma^{-2d_{\mathbf{u}}})$. *Q.E.D.*

Department of Statistics, University of California at Berkeley, Berkeley, California 94720; Research partially supported by NSF Grant FD01-04075.

and

Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel; Research partially supported by NSF Grant FD01-04075.

and

Department of Economics, Sloan School, MIT, Cambridge, Massachusetts 02139

REFERENCES

Ait-Sahalia, Y., P. J. Bickel, and T. M. Stoker (1998): "Goodness-of-Fit Tests for Regression Using Kernel Methods," *J. of Econometrics*, to appear.

Arcones, M., and E. Gine (1995): "On the Law of the Iterated Logarithm for Canonical U Processes," *Stochastic Proceses and Their Applications*, 58, 217–245.

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

Bickel, P. J., Y. Ritov, and T. M. Stoker (2001): "Tailor-Made Tests for Goodness-of-Fit to Semiparametric Hypotheses," Technical report.

Bierens, H., and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica*, 65, 1129–1151.

Härdle, W., and E. Mammen (1993): "Comparing Nonparametric Versus Parametric Regression Fits," *Annals of Statistics*, 21, 1926–1947.

Silverman, B. (1981): "Using Bootstrap Kernel Density Estimates to Investigate Unimodality," JRSSB-43, 97–99.

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

TABLE I

$\alpha = .1$

Test statistic	$\lambda = 0$	$\lambda = \pi/2$	$\lambda = \pi$	$\lambda = 6\pi$
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

$\alpha = 0.05$:

Test statistic	$\lambda = 0$	$\lambda = \pi/2$	$\lambda = \pi$	$\lambda = 6\pi$
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