On completeness and consistency in nonparametric instrumental variable models*

Joachim Freyberger‡

March 6, 2015

Abstract

This paper provides a first test for the identification condition in a nonparametric instrumental variable model, known as completeness, by linking the outcome of the test to consistency of an estimator. In particular, I show that uniformly over all distributions for which the test rejects with probability bounded away from 0, an estimator of the structural function is consistent. This is the case for a large class of complete distributions as well as certain sequences of incomplete distributions. As a byproduct of this result, the paper makes two additional contributions. First, I present a definition of weak instruments in the nonparametric instrumental variable model, which is equivalent to the failure of a restricted version of completeness. Second, I show that the null hypothesis of weak instruments, and thus failure of a restricted version of completeness, is testable and I provide a test statistic and a bootstrap procedure to obtain the critical values. Finally, I demonstrate the finite sample properties of the tests and the estimator in Monte Carlo simulations.

Keywords: Completeness, consistency, instrumental variables, nonparametric estimation, weak instruments.

JEL Classification: C14, C21

*Thanks to Ivan Canay, Bruce Hansen, Joel Horowitz, Jack Porter, Azeem Shaikh, Xiaoxia Shi, participants at the Cowles Conference 2014, and seminar participants at Penn State, Cornell, Ohio State, and NYU for helpful comments and Jangsu Yoon for research assistance.

‡Department of Economics, University of Wisconsin - Madison. Email: jfreyberger@ssc.wisc.edu.

1 Introduction

There has been much recent work on nonparametric models with endogeneity, which relies on a nonparametric analog of the rank condition, known as completeness. Specifically, consider the nonparametric instrumental variable (IV) model

$$Y = g_0(X) + U, \qquad E(U \mid Z) = 0, \tag{1}$$

where Y, X, and Z are observed scalar random variables, U is an unobserved random variable, and g_0 is a structural function of interest. It is well known that identification in this model is equivalent to the completeness condition (Newey and Powell, 2003), which says that $E(g(X) \mid Z) = 0$ almost surely implies that $g(X) = 0$ almost surely for all g in a certain class of functions $\mathcal{G}$.[1] Next to this nonparametric IV model, completeness has also been used in various other settings including measurement error models (Hu and Schennach, 2008), panel data models (Freyberger, 2012), and nonadditive models with endogeneity (Chen, Chernozhukov, Lee, and Newey, 2014). Although completeness has been employed extensively, existing results so far have only established that the null hypothesis that completeness fails is not testable. In particular, Canay, Santos, and Shaikh (2013) show that any test that controls size uniformly over a large class of incomplete distributions, has power no greater than size against any alternative. Intuitively, the null hypothesis that completeness fails cannot be tested because for every complete distribution, there exists an incomplete distribution which is arbitrarily close to it. They conclude that "it is therefore not possible to provide empirical evidence in favor of the completeness condition by means of such a test".

In an application researchers most likely do not just want to test completeness by itself, but are instead interested in estimating the structural function g_0. If completeness holds, then standard estimators of g_0 have good properties, such as being consistent, and hence, a test which provides evidence in favor of completeness also implies evidence in favor of consistency. Contrarily, without completeness standard estimators are not consistent for g_0. Using a test that controls size uniformly over all incomplete distributions would then be crucial if a nonparametric estimator of g_0 had poor properties for any sequence of incomplete distributions (which may depend on the sample size). For example, in the linear IV model, where $Y = \alpha_0 + \beta_0 X + U$ and $cov(U, Z) = 0$, β_0 is point identified if and only if $cov(X, Z) \neq 0$. Moreover, for any sequence of distributions for which point identification fails, the two stage least squares (TSLS) estimator is not consistent for β_0. Thus, any test of the null hypothesis

[1] The class of functions typically depends on the restrictions imposed on g_0, such as being square integrable ("L^2 completeness") or bounded ("bounded completeness").

$cov(X, Z) = 0$, which can provide evidence for consistency of the TSLS estimator, needs to control size uniformly over all distributions for which point identification fails.

This paper explores a test for completeness, which does not control size uniformly over all incomplete distributions, and it links the outcome of the test to consistency of an estimator of g_0. The main motivation for this approach is that there are sequences of incomplete distributions which imply great instruments and under which a standard nonparametric estimator of g_0 is consistent. Therefore, if the main goal is to estimate the function g_0 consistently, uniform size control over all incomplete distributions is not necessary. In particular, I provide a test statistic $\hat{T}$, a critical value c_n, and an estimator $\hat{g}$, such that uniformly over a large class of distributions

$$P\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n\right) \to 0,$$

where $\|\cdot\|_c$ is a consistency norm, n is the sample size, and $\varepsilon_n \to 0$ as $n \to \infty$. An important implication of this result is that for any sequence of distributions for which $n\hat{T} \geq c_n$ with probability bounded away from 0, $\hat{g}$ will be consistent for g_0. I show that this is the case for a large class of complete distributions and certain sequences of incomplete distributions. Consequently, $n\hat{T} \geq c_n$ provides empirical evidence for consistency.

As a byproduct of the test, the paper also provides a framework of how one could think about weak instruments in the nonparametric IV model. Specifically, I define instruments to be weak if the asymptotic bias of certain estimators of g_0 can be larger than a prespecified ε. Various estimators can be used. The only requirement is that in case of partial identification, the estimator is close to some function in the identified set as the sample size increases. It follows that if the instruments are strong, $\|\hat{g} - g_0\|_c \leq \varepsilon + o_p(1)$ for various estimators $\hat{g}$. This definition of weak instruments is similar to that of Stock and Yogo (2005) in a linear IV model, but it has some distinct features as discussed in Section 4. It turns out that a version of the test described above, is a test for the null hypothesis that the instruments are weak. Furthermore, although we can have strong instruments with an incomplete distribution, I show that weak instruments are equivalent to the failure of a restricted version of completeness, where the class of functions $\mathcal{G}$ contains particular constraints. Thus, failure of this restricted version of completeness, or equivalently weak instruments, is testable and I provide a test statistic and a bootstrap procedure to obtain the critical values. The test has an intuitive interpretation because it is essentially a constrained version of a rank test of the covariance matrix of vectors of transformations of X and Z.

Note that for a fixed distribution of the data an estimator of g_0 is not consistent if completeness fails. The results in this paper show that the asymptotic bias of the estimator is

small for many incomplete distributions, in particular for those that are arbitrarily close to complete distributions. Furthermore, certain sequences of incomplete distributions lead to a consistent estimator, which explains the testability result. These results seem intuitive because if an incomplete distribution is arbitrary close to a complete distribution, we would expect that the properties of a nonparametric estimator of g_0 are almost identical under both distributions. Also note that this paper does not address the question of conducting uniformly valid inference following the test. Santos (2012) and Tao (2014) provide inference methods, which are robust to a failure of point identification, but they do not discuss properties of estimators of g_0 under partial identification. It is also not clear if their inference methods are uniformly valid over a large class of distributions. I complement their work because a consistent estimator of g_0 could be of interest next to confidence intervals for certain functionals of g_0, which is delivered by the results in this paper if the instruments are strong enough. Moreover, since the test provides evidence for the strength of the instruments, it can help to distinguish how much of an estimated confidence set is due to sampling noise and how much is due to the size of the identified set.

Several recent papers (including Mattner (1993), Newey and Powell (2003), Andrews (2011), D'Haultfoeuille (2011), and Hu and Shiu (2012)) have provided sufficient conditions for different versions of completeness, such as bounded completeness or L^2 completeness. The results of Canay et al. (2013) imply that these versions of completeness are not testable, while the sufficient conditions might be testable if they are strong enough. My restrictions on the class of functions $\mathcal{G}$, which imply testability, are different than the versions of completeness considered in those papers, but I am concerned with consistent estimation, testing, and the relationship to weak instruments, instead of providing sufficient conditions.

This paper contributes to the nonparametric IV literature by showing that the data can provide empirical evidence for consistency. Most theoretical work relies on the completeness assumption, such as Newey and Powell (2003), Hall and Horowitz (2005), Blundell et al. (2007), Darolles et al. (2011), Horowitz (2011), Horowitz and Lee (2012), Chen and Christensen (2013), and Horowitz (2014). Santos (2012) provides an inference method for linear functionals of g_0 without completeness. Other settings and applications which use completeness include Hu and Schennach (2008), Berry and Haile (2014), Chen et al. (2014), and Sasaki (2014). There is also a growing literature on general models with conditional moment restrictions which include instrumental variable models as special cases. Several settings assume point identification (for example Ai and Chen (2003), Chen and Pouzo (2009, 2012, 2014)) while others allow for partial identification (Tao, 2014).

Additional related papers are those on weak instruments in other models, in particular the linear model. Some papers (for example Staiger and Stock (1997)) consider estimation under weak instruments while others develop tests for weak instruments (among many others Cragg and Donald (1993), Stock and Yogo (2005), Montiel Olea and Pflueger (2013)). Han (2014) deals with weak instruments in the nonparametric triangular model, where identification follows from a finite dimensional rank condition. He then considers estimation under certain sequences of distributions, whereas I consider a different model with a different identification condition and I am mainly concerned with testing and uniform consistency.

The following section provides important definitions and a derivation of the population test statistic. Section 3 presents the sample test statistic and the formal result which links the outcome of the test to consistency of an estimator. Section 4 defines weak instruments and shows the equivalence between weak instruments and a restricted version of completeness. It also discusses the test statistic for the null hypothesis of weak instruments, characterizes its asymptotic distribution, and provides a bootstrap procedure to obtain the critical values. The remaining sections contain a discussion of choices of norms, extensions to functions on $\mathbb{R}$, Monte Carlo simulation results, and a conclusion.

2 Definitions and population test statistic

This section starts by introducing function spaces and norms that are used throughout the paper. I then explain why, uniformly over a class of distributions, completeness is not equivalent to consistency and how this idea can be used as a basis for the test.

2.1 Notation

Let $\|\cdot\|$ be the Euclidean norm for vectors in $\mathbb{R}^J$. Let $\mathcal{X}$ be the support of X and let $\|\cdot\|_c$ and $\|\cdot\|_s$ be norms for functions from $\mathcal{X}$ to $\mathbb{R}$. Define the parameter space

$$\tilde{\mathcal{G}} = \{g : \|g\|_s \leq C\}$$

where C is a positive constant. Properties of the norms $\|\cdot\|_c$ and $\|\cdot\|_s$ will be discussed in detail below but useful examples to think of are:

$$\|g\|_c^2 = \int_{\mathcal{X}} g(x)^2 dx \quad \text{and} \quad \|g\|_s^2 = \int_{\mathcal{X}} \left(g(x)^2 + g'(x)^2\right) dx$$

or

$$\|g\|_c = \sup_{x\in\mathcal{X}} |g(x)| \quad \text{and} \quad \|g\|_s = \sup_{x\in\mathcal{X}} |g(x)| + \sup_{x_1,x_2\in\mathcal{X}} |g(x_1) - g(x_2)|/|x_1 - x_2|.$$

A standard smoothness assumption in many nonparametric models, which I also impose in this paper, is that $g_0 \in \tilde{\mathcal{G}}$ (see e.g. Newey and Powell (2003), Santos (2012), or Horowitz (2014) in nonparametric IV models). This assumption typically restricts function values and derivatives of g_0. Consistency is then usually proved in the weaker norm $\|\cdot\|_c$.

2.2 Derivation of population test statistic

If g_0 is not point identified relative to all functions in $\tilde{\mathcal{G}}$, then there exists $g_1 \neq g_0$ with $\|g_1\|_s \leq C$ and $E(g_1(X) \mid Z) = E(Y \mid Z)$. Let $g \equiv g_0 - g_1$. Then g satisfies $E(g(X) \mid Z) = 0$ and $\|g\|_s \leq 2C$. For a fixed distribution of the data an estimator of g_0 is not consistent if g_0 is not point identified. That is, it is not consistent if $\|g\|_c > 0$. However, for a sequence of distributions consistency typically follows as long as $\|g\|_c \leq \varepsilon_n \to 0$ for all functions g such that $E(g(X) \mid Z) = 0$ and $\|g\|_s \leq 2C$. To show this, let $\hat{g}$ be an estimator such that

$$\inf_{g_1 : \|g_1\|_s \leq C, E(g_1(X)|Z) = E(Y|Z)} \|\hat{g} - g_1\|_c = o_p(1).$$

That is, $\hat{g}$ is close to some function in the identified set $I = \{g_1 : \|g_1\|_s \leq C, E(g_1(X) \mid Z) = E(Y \mid Z)\}$ as the sample size increases. Many estimators, such as series or Tikhonov type estimators satisfy this property, even if g_0 is not point identified. Then

$$\begin{aligned}
\|\hat{g} - g_0\|_c &= \inf_{g_1 : \|g_1\|_s \leq C, E(g_1(X)|Z) = E(Y|Z)} \|\hat{g} - g_1 + g_1 - g_0\|_c \\
&\leq \inf_{g_1 : \|g_1\|_s \leq C, E(g_1(X)|Z) = E(Y|Z)} \|\hat{g} - g_1\|_c + \sup_{g_1 : \|g_1\|_s \leq C, E(g_1(X)|Z) = E(Y|Z)} \|g_1 - g_0\|_c \\
&\leq \inf_{g_1 : \|g_1\|_s \leq C, E(g_1(X)|Z) = E(Y|Z)} \|\hat{g} - g_1\|_c + \sup_{g : \|g\|_s \leq 2C, E(g(X)|Z) = 0} \|g\|_c \\
&\leq o_p(1) + \varepsilon_n
\end{aligned}$$

It turns out that under certain assumptions we can test the null hypothesis

$$H_0 : \text{There is } g \text{ such that } E(g(X) \mid Z) = 0,\ \|g\|_s \leq 2C,\ \text{and } \|g\|_c \geq \varepsilon_n$$

and we can link the outcome of the test to consistency of an estimator of g_0. Intuitively, as demonstrated above, if H_0 is false, then the estimator will be consistent. Contrarily if H_0 is true and if the critical value diverges, then we will not reject the null hypothesis. In other words a test of the null hypothesis then has the feature that

$$P\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ \text{ reject } H_0\right) \to 0.$$

I provide a test statistic, a critical value, and an estimator $\hat{g}$ such that this is true uniformly over all distributions satisfying Assumption 1 below. As a consequence, for any sequence of

distributions for which we reject with probability bounded away from 0, $\hat{g}$ will be consistent for g_0. I show that this is the case for a large class of complete distributions and certain sequences of incomplete distributions.[2]

To see how this null hypothesis can be tested, first rewrite

$$\begin{aligned} E(g(X) \mid Z = z) = 0 \text{ a.s.} &\Leftrightarrow E(g(X) \mid Z = z) f_Z(z) = 0 \text{ a.s.} \\ &\Leftrightarrow \int \left(E(g(X) \mid Z = z) f_Z(z)\right)^2 dz = 0 \\ &\Leftrightarrow \int \left(\int g(x) f_{XZ}(x, z) dx \right)^2 dz = 0. \end{aligned}$$

Next define

$$S_0(g) \equiv \int \left(\int g(x) f_{XZ}(x, z) dx \right)^2 dz$$

and

$$T \equiv \inf_{g : \|g\|_s \leq 2C, \|g\|_c \geq \varepsilon_n} S_0(g).$$

If the null hypothesis above is true, then $T = 0$. Contrarily, for any $\varepsilon_n > 0$ and under certain restrictions on the norms $\|\cdot\|_s$ and $\|\cdot\|_c$, it holds that $T > 0$ under any alternative. For instance, suppose that $\tilde{\mathcal{G}}$ is compact under $\|\cdot\|_c$ and that $S_0(g)$ is continuous under $\|\cdot\|_c$.[3] Then $\{g : \|g\|_s \leq 2C, \|g\|_c \geq \varepsilon_n\}$ is a closed subset of a compact set,

$$\inf_{g : \|g\|_s \leq 2C, \|g\|_c \geq \varepsilon_n} S_0(g) = \min_{g : \|g\|_s \leq 2C, \|g\|_c \geq \varepsilon_n} S_0(g),$$

and $T > 0$ under any alternative. Notice that without the constraint $\|g\|_s \leq 2C$, or without smoothness restrictions on g_0, it holds that $T = 0$ even for complete distributions. This is one reason for the non-testability result of Canay et al. (2013).

The remaining analysis will be easier if we notice that the infimum will be attained at a function where $\|g\|_c = \varepsilon_n$, because otherwise we could simply scale down g. Moreover,

$$\inf_{g : \|g\|_s \leq 2C, \|g\|_c = \varepsilon_n} S_0(g) = \inf_{g : \|g/\varepsilon_n\|_s \leq 2C/\varepsilon_n, \|g/\varepsilon_n\|_c = 1} \varepsilon_n^2 S_0(g/\varepsilon_n) = \inf_{g \in \bar{\mathcal{G}}(\varepsilon_n) : \|g\|_c = 1} \varepsilon_n^2 S_0(g),$$

where

$$\bar{\mathcal{G}}(\varepsilon_n) = \{g : \|g\|_s \leq 2C/\varepsilon_n\}.$$

If ε_n changes with the sample size, then the function space changes with the sample size as well. Neglecting ε_n^2 in front of the objective does not change the minimizer, so I will consider a test statistic based on a scaled sample analog of $\inf_{g \in \bar{\mathcal{G}}(\varepsilon_n) : \|g\|_c = 1} S_0(g)$.

[2] Theorem 2 and the following discussion in Section 3.4 describe the class of distributions and provide examples of sequences of incomplete distributions.

[3] Compactness and continuity are implied by Assumption 2.

3 Testing identification and consistent estimation

In this section I present the sample version of the test statistic, the estimator of g_0, and the result which links the outcome of the test to consistency of the estimator. Throughout the paper I will assume that X and Z are scalar and that they are continuously distributed with compact support and joint density f_{XZ} with $0 < f_{XZ}(x,z) \leq C_d < \infty$ almost everywhere. We can then assume without loss of generality that $X, Z \in [0,1]$.[4] In particular, I make the following assumption about the distribution of the data:

Assumption 1. The data $\{Y_i, X_i, Z_i\}_{i=1}^n$ is an iid sample from the distribution of (Y,X,Z), where (Y,X,Z) are continuously distributed, $(X,Z) \in [0,1]^2$, $0 < f_{XZ}(x,z) \leq C_d < \infty$ almost everywhere, and $E(Y^2 \mid Z) \leq \sigma_Y^2$ for some $\sigma_Y > 0$. The data is generated by model (1) and $\|g_0\|_s \leq C$ for some constant $C > 0$.

3.1 Sample analog of test statistic

To construct the test statistic, let ϕ_j be an orthonormal basis for functions in $L^2[0,1]$. Let

$$f_J(x,z) = \sum_{j=1}^{J}\sum_{k=1}^{J} a_{jk}\phi_j(x)\phi_k(z),$$

denote the series approximation of f_{XZ} where

$$a_{jk} = \int\int \phi_j(x)\phi_k(z) f_{XZ}(x,z)dxdz.$$

We can estimate f_{XZ} by

$$\hat{f}_{XZ}(x,z) = \sum_{j=1}^{J}\sum_{k=1}^{J} \hat{a}_{jk}\phi_j(x)\phi_k(z)$$

where

$$\hat{a}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\phi_k(Z_i).$$

The assumptions imply that $J \to \infty$ as $n \to \infty$. Denote the series approximation of a function g by

$$g_J(x) = \sum_{j=1}^{J} h_j \phi_j(x)$$

where $h_j = \int g(x)\phi_j(x)dx \in \mathbb{R}$ for all j. Define the sieve space

$$\bar{\mathcal{G}}_J(\varepsilon_n) = \left\{ g \in \bar{\mathcal{G}}(\varepsilon_n) : g(x) = \sum_{j=1}^{J} h_j\phi_j(x) \text{ for some } h_j \in \mathbb{R} \right\}.$$

[4] Section 6 outlines the extension to functions on $\mathbb{R}$.

We can now define the test statistic which is

$$\hat{T} = \inf_{g \in \bar{\mathcal{G}}_J(\varepsilon_n): \|g\|_c = 1} \int \left(\int g(x) \hat{f}_{XZ}(x,z) dx \right)^2 dz.$$

To obtain a simpler representation of the test statistic notice that for any $g \in \bar{\mathcal{G}}_J(\varepsilon_n)$,

$$\begin{aligned} \int \left(\int g(x) \hat{f}_{XZ}(x,z) dx \right)^2 dz &= \int \left(\int \sum_{l=1}^{J} h_l \phi_l(x) \sum_{j=1}^{J} \sum_{k=1}^{J} \hat{a}_{jk} \phi_k(x) \phi_j(z) dx \right)^2 dz \\ &= \int \left(\sum_{j=1}^{J} \sum_{k=1}^{J} \hat{a}_{jk} h_k \phi_j(z) \right)^2 dz \\ &= \sum_{j=1}^{J} \left(\sum_{k=1}^{J} \hat{a}_{jk} h_k \right)^2. \end{aligned}$$

Let $\hat{A}$ be the $J \times J$ matrix with elements $\hat{a}_{jk}$ and let A be the population analog. Let h be the $J \times 1$ vector containing h_j. Then

$$\sum_{j=1}^{J} \left(\sum_{k=1}^{J} \hat{a}_{jk} h_k \right)^2 = \left\| \hat{A} h \right\|^2 = h'(\hat{A}'\hat{A})h,$$

where $\| \cdot \|$ denotes the Euclidean norm. Hence

$$\hat{T} = \inf_{g \in \bar{\mathcal{G}}_J(\varepsilon_n): \|g\|_c = 1} h'(\hat{A}'\hat{A})h.$$

The test statistic depends on the norms $\| \cdot \|_c$ and $\| \cdot \|_s$ (through $\bar{\mathcal{G}}_J(\varepsilon_n)$), but as described in the next section using a specific choice of norms, it has the intuitive interpretation of being a constraint version of a rank test of $A'A$.

3.2 Interpretation of test statistic with Sobolev spaces

The function space and constraints can be simplified if we are dealing with Sobolev spaces. In particular, let

$$\|g\|_c^2 = \int_0^1 g(x)^2 dx \quad \text{and} \quad \|g\|_s^2 = \int_0^1 \left(g(x)^2 + g'(x)^2 \right) dx.$$

Moreover, define $b_{jk} = \int \phi_j'(x) \phi_k'(x) dx$ and B as the $J \times J$ matrix with element (j,k) equal to b_{jk}. Then

$$\begin{aligned} \left\{ g \in \bar{\mathcal{G}}_J(\varepsilon_n) : \|g\|_c = 1 \right\} &= \left\{ g_J : \int_0^1 g_J'(x)^2 dx \leq (2C/\varepsilon_n)^2 - 1, \ \int_0^1 g_J(x)^2 = 1 \right\} \\ &= \left\{ g_J : \sum_{j=1}^{J} \sum_{k=1}^{J} b_{jk} h_j h_k \leq (2C/\varepsilon_n)^2 - 1, \ \sum_{j=1}^{J} h_j^2 = 1 \right\} \\ &= \left\{ g_J : h'Bh \leq (2C/\varepsilon_n)^2 - 1, \ h'h = 1 \right\}. \end{aligned}$$

It follows that the test statistic is the solution to

$$\begin{aligned} \min_{h\in\mathbb{R}^J} \quad & h'(\hat{A}'\hat{A})h \\ \text{subject to} \quad & h'Bh \leq (2C/\varepsilon_n)^2 - 1 \\ & h'h = 1. \end{aligned}$$

Without the first constraint, the solution to the optimization problem is the smallest eigenvalue of $\hat{A}'\hat{A}$, which could be used to test the rank of $A'A$ if J was fixed (see for example Robin and Smith, 2000). Thus, the test in this paper can be interpreted as a constrained version of a rank test, where the dimension of the matrix increases with the sample size.

3.3 Estimator

The estimator, which I will use to prove the consistency result, is a series estimator from Horowitz (2012). To describe the estimator, let $\hat{m}$ be a $J \times 1$ vector with

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_k(Z_i).$$

Let

$$\hat{h} = \underset{h\in\mathbb{R}^J:\|g_J\|_s\leq C}{\arg\min} \left\|\hat{A}h - \hat{m}\right\|^2 \quad \text{and} \quad \hat{g}(x) = \sum_{j=1}^{J} \hat{h}_j \phi_j(x).$$

Notice that the test statistic above is based on a scaled version of

$$\min_{h\in\mathbb{R}^J:\|g_J\|_s\leq 2C, \|g_J\|_c=\varepsilon_n} \|\hat{A}h\|^2,$$

which is why this estimator relates nicely to the test.

3.4 Assumptions and main results

I will next state and discuss the remaining assumptions and the main results.

Assumption 2. $\tilde{\mathcal{G}}$ is compact under $\|\cdot\|_c$ and $C_o\|g\|_c^2 \geq \int g(x)^2 dx$ for some $C_o > 0$.

Assumption 3. The basis functions form an orthornomal basis of $L^2[0,1]$.

Assumption 4. For all $g \in \bar{\mathcal{G}}(\varepsilon_n)$, $\|g - g_J\|_c \leq C_g J^{-\bar{s}}$ with $\bar{s} \geq 2$.

Assumption 5. For all $g \in \bar{\mathcal{G}}(\varepsilon_n)$ with $\|g\|_c = 1$, $g_J/\|g_J\|_c \in \bar{\mathcal{G}}(\varepsilon_n)$ and $g_J \in \bar{\mathcal{G}}(\varepsilon_n)$.

Compactness is implied by many standard choices of norms, as discussed in more detail in Section 5. In particular, it holds with the norms used in Section 3.2. The assumption simplifies the analysis because it guarantees, for example, that the infimum in the definition of the test statistic is actually attained. It could possibly be relaxed using similar tools as those in Chen and Pouzo (2012). Also notice that to implement the test statistic, the constant C in the definition of the function space has to be chosen. Section 5 discusses how this can done in particular applications. The second part of Assumption 2 implies that $S_0(g)$ is continuous in g under the norm $\|\cdot\|_c$. The assumption allows $\|\cdot\|_c$ to be the L^2-norm, the sup-norm, and many other norms.

Assumption 3 is standard. Assumption 4 would also be standard if ε_n was fixed (see Chen (2007) for function spaces which satisfy this assumption). It also holds if $\varepsilon_n \to 0$ slow enough and $J \to \infty$ as $n \to \infty$, and if the restrictions on the function space are strong enough. For example, suppose that $\|g_J - g\|_c \leq C_g J^{-\tilde{s}}$ for all g such that $\|g\|_s \leq 2C$. Equivalently, it holds that $\|g_J/\varepsilon_n - g/\varepsilon_n\|_c \leq C_g J^{-\tilde{s}}/\varepsilon_n$ for all g such that $\|g/\varepsilon_n\|_s \leq (2C/\varepsilon_n)$ or simply $\|\tilde{g}_J - \tilde{g}\|_c \leq C_g J^{-\tilde{s}}/\varepsilon_n$ for all $\tilde{g} \in \bar{\mathcal{G}}(\varepsilon_n)$. Hence, if for example $\varepsilon_n = \frac{1}{J}$ and $\tilde{s} \geq 3$, Assumption 4 holds.

Assumption 5 implies that the series approximations of functions in $\bar{\mathcal{G}}(\varepsilon_n)$ are in $\bar{\mathcal{G}}_J(\varepsilon_n)$ and are therefore contained in the set that is minimized over in the definition of the test statistic. It is stronger than necessary and it can be relaxed at the expense of additional notation, but it appears to be reasonable as shown in Appendix B.2.

We now get the following result. All proofs are in Appendix A.

Theorem 1. *Suppose Assumptions 1 - 5 hold. Let $c_n \to \infty$ and $\varepsilon_n \to 0$ such that*

$$\frac{nJ^{-2\bar{s}}}{\varepsilon_n^2 c_n} \to 0 \quad \text{and} \quad \frac{J}{\varepsilon_n^2 c_n} \to 0.$$

Let $\mathcal{P}$ be the class of distributions P satisfying Assumption 1. Then

$$\sup_{P \in \mathcal{P}} P\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n\right) \to 0.$$

The theorem implies that for any sequence of distributions P_n for which

$$P_n\left(n\hat{T} \geq c_n\right) \geq \delta > 0,$$

it holds that

$$P_n\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n \mid n\hat{T} \geq c_n\right) \to 0$$

or equivalently

$$P_n\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n \mid \text{ reject } H_0\right) \to 0.$$

Since $\varepsilon_n \to 0$, the results imply that for any sequence of distributions under which we reject with probability larger than δ, $\hat{g}$ will be consistent for g_0. Interestingly this will not only hold for sequences of complete distributions, but also for certain sequences of incomplete distributions if c_n converges slow enough (see Theorem 2 below). In other words, under certain sequences of incomplete distributions, we get a consistent estimator of g_0. Therefore, if the main goal is to estimate g_0 consistently, controlling size uniformly over all incomplete distributions is not necessary. These findings contrast results in the linear IV model, where any sequence of distributions under which identification fails (that is $cov(X, Z) = 0$), leads to an inconsistent estimator of the slope coefficient.

Instead, another implication of Theorem 1 is that for any $\alpha > 0$, $\varepsilon > 0$, and $\delta > 0$,

$$P\left(\|\hat{g} - g_0\|_c \geq \varepsilon\right) \geq \delta$$

implies that

$$P\left(n\hat{T} \geq c_n \mid \|\hat{g} - g_0\|_c \geq \varepsilon\right) \leq \alpha$$

for all n large enough and any $P \in \mathcal{P}$. Thus, the test controls size uniformly over certain distributions for which $\hat{g}$ is not a consistent estimator. Finally, since fixed incomplete distributions do not yield a consistent estimator, it also follows that $P(n\hat{T} \geq c_n) \to 0$ for every fixed incomplete distribution.

Clearly, the rate conditions in Theorem 1 are satisfied if c_n diverges very fast. In this case, regardless of whether or not the estimator is consistent, $P(n\hat{T} \geq c_n) \to 0$. The results are most interesting if c_n diverges slowly and if ε_n converges to 0 slowly. In this case, $P(n\hat{T} \geq c_n) > \delta > 0$ and thus $\|\hat{g} - g_0\|_c \xrightarrow{p} 0$ for a large class of complete distributions. To state this result formally, let

$$\kappa_n = \inf_{g \in \bar{\mathcal{G}}_J(\varepsilon_n): \|g\|_c = 1} \int \left(\int g(x) f_J(x, z) dx \right)^2 dz,$$

where f_J is the series approximation of f_{XZ}. With these definitions, we get the following result.

Theorem 2. *Suppose Assumptions 1 - 5 hold. For all distributions for which*

$$\frac{n\kappa_n}{J^2 \ln(\ln(n))} \to \infty \quad \text{and} \quad \frac{n\kappa_n}{c_n} \to \infty$$

we have

$$P\left(n\hat{T} \geq c_n\right) \to 1.$$

To better understand the rate conditions, first suppose that ε_n is fixed and that $c_n = J\ln(n)$. Next notice that,[5]

$$\begin{aligned}
\kappa_n &= \inf_{g\in\bar{\mathcal{G}}_J(\varepsilon_n):\|g\|_c=1} \int \left(\int f_J(x,z)g_J(x)\right)^2 dz \\
&= \inf_{g\in\bar{\mathcal{G}}_J(\varepsilon_n):\|g\|_c=1} \int \left(\int \left((f_J(x,z)-f_{XZ}(x,z))g_J(x) + f_{XZ}(x,z)g_J(x)\right)dx\right)^2 dz \\
&\geq \inf_{g\in\bar{\mathcal{G}}(\varepsilon_n):\|g\|_c=1} \frac{3}{4}\int \left(\int f_{XZ}(x,z)g(x)dx\right)^2 dz \\
&\quad - \sup_{g\in\bar{\mathcal{G}}(\varepsilon_n):\|g\|_c=1} 3\int \left(\int (f_J(x,z)-f_{XZ}(x,z))g(x)dx\right)^2 dz.
\end{aligned}$$

Since $\tilde{\mathcal{G}}$ is compact with respect to $\|\cdot\|_c$, the first term on the right hand side is a fixed positive constant for any complete distribution, while the second term converges to 0. Hence, κ_n is bounded below by a positive constant. Then all rate conditions of Theorems 1 and 2 are satisfied as long as

$$\frac{n}{J^2\ln(n)} \to \infty \quad \text{and} \quad \frac{n}{J\ln(n)}J^{-2\bar{s}} \to 0. \tag{2}$$

Hence, J has to go to ∞ but it cannot diverge too fast relative to n. Since $\bar{s} \geq 2$, feasible choices would be $J = n^a$, where $a \in (1/5, 1/2)$. In this case $P(n\hat{T} \geq c_n) \to 1$ for any complete distribution. It now follows that as long as ε_n converges to 0 slow enough, the rate conditions in (2) together with Assumptions 1 - 5 imply that

$$\sup_{P\in\mathcal{P}} P\left(\|\hat{g}-g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n\right) \to 0$$

and

$$P\left(n\hat{T} \geq c_n\right) \to 1$$

for a large class of complete distributions. The only complete distributions for which the test then does not reject with probability approaching 1 are the ones for which $a_{jk} \to 0$ as $j,k \to \infty$ extremely rapidly. For those distributions $\|\hat{g}-g_0\|_c \xrightarrow{p} 0$ very slowly and if the rate of convergence is slower than ε_n, the test rejects with probability approaching 0. Also notice that we get $P(n\hat{T} \geq c_n) \to 1$ and a consistent estimator for sequences of incomplete distributions. One simple example is a sequence where the density is a series approximation f_J of a density f_{XZ} corresponding to a complete distribution. For such a sequence of distributions the previous arguments still apply because κ_n is still bounded below by a positive constant.

[5]Using that $(1/2a+2b)^2 = 1/4a^2+4b^2+2ab \geq 0$ implies $(a+b)^2 = a^2+b^2+2ab \geq 3/4a^2-3b^2$.

The previous discussion highlight that the slower c_n diverges and $\varepsilon_n \to 0$, the larger the rejection probabilities for distributions for which $\hat{g}$ is consistent for g_0. However, a faster rate of c_n allows ε_n to go to 0 faster, which strengthens the conclusion in case of rejection. Since the focus of this paper is on consistency rather than the rate of convergence, the results are most interesting if c_n diverges slowly (at rate $J \ln(n)$) and ε_n goes to 0 at a logarithmic rate of n. If the main goal was to find evidence in favor of a certain rate of convergence, ε_n and c_n can be adjusted accordingly.

An alternative to using a converging ε_n and a diverging c_n is to fix ε_n and to use a bootstrap procedure to obtain the critical value. While this reduces the ambiguity of how to exactly choose ε_n and c_n in finite samples, the test will have a different interpretation as explained in the next section.

4 Weak instruments and restricted completeness

For a fixed $\varepsilon_n = \varepsilon$ the test statistic above tests the null hypothesis

$$H_0 : \text{There is } g \text{ such that } E(g(X) \mid Z) = 0,\ \|g\|_s \leq 2C, \text{ and } \|g\|_c \geq \varepsilon.$$

I also showed that if H_0 is false, then for any estimator $\hat{g}$ of g_0 in a large class

$$\|\hat{g} - g_0\|_c \leq o_p(1) + \varepsilon.$$

In other words, the asymptotic bias of the estimator is smaller than ε. Since, the bias is guaranteed to be small, this situation could be interpreted as strong instruments. Contrarily, instruments would then be weak if the bias can be larger than ε. The value of ε depends on the particular application and on how much bias a researcher considers acceptable. Formally, we get the following definition.

Definition 1. Let $\varepsilon > 0$ and $C > 0$ and suppose that $\|g_0\|_s \leq C$. Instruments are *weak* if there is g with $\|g\|_s \leq 2C$, $\|g\|_c \geq \varepsilon$, and $E(g(X) \mid Z) = 0$. Instruments are *strong* if no such function exists.

This definition of weak instruments is similar to the definition of Stock and Yogo (2005) in the linear model who also think of weak instruments in terms of properties of estimators.[6]

[6] An alternative and commonly used way to define weak instruments in the parametric model is in terms of sequences of distributions as in Staiger and Stock (1997). This would be difficult in the nonparametric model, because the singular values of the operator $E(\cdot \mid Z)$ converge to 0 for both complete and incomplete distributions. Furthermore, we get a consistent estimator for certain sequences of incomplete distributions.

They define instruments to be weak if the bias of the TSLS estimator, relative to the bias of the OLS estimator, could exceed a certain threshold b, e.g. 10%. Both my definition and that of Stock and Yogo (2005) are based on the bias of estimators, but the parametric definition is about a relative bias, while I use an absolute bias. An advantage of using the relative bias in the parametric case is that it helps to separate relevance of the instrument and endogeneity of the regressor. The reason is that if U is mean independent of X and Z, then the TSLS estimator is unbiased no matter whether the instruments are strong or weak. Such issues do not arise with the definition of weak instruments in the nonparametric framework, because it is based on the identified set rather than a specific estimator. For instance, if X and Z were independent, instruments would always be weak. Hence, considering the absolute bias seems more reasonable in this setting.

An alternative, yet equivalent, way of interpreting this definition of weak instruments is in terms of functions in the identified set. Suppose that both g_0 and an alternative function g_1 satisfy the moment conditions of the model and the smoothness restrictions $\|g_0\|_s \leq C$ and $\|g_1\|_s \leq C$. Then $g \equiv g_0 - g_1$ satisfies $E(g(X) \mid Z) = 0$ and $\|g\|_s \leq \|g_0\|_s + \|g_1\|_s \leq 2C$. Hence, if the instruments are strong, then $\|g\|_c \leq \varepsilon$ and thus, the consistency norm of the difference between any two functions in the identified set is smaller or equal to ε.

The following proposition now provides a relationship between weak instruments and the failure of a restricted version of completeness where all functions g for which $E(g(X) \mid Z) = 0$ are restricted to be in

$$\mathcal{G} = \left\{ g : \|g\|_s \leq 2C, \quad \|g\|_s \leq \frac{2C}{\varepsilon} \|g\|_c \right\}.$$

Proposition 1. The null hypothesis

$$H_0 : \text{There is } g \text{ such that } E(g(X) \mid Z) = 0, \ \|g\|_s \leq 2C, \text{ and } \|g\|_c \geq \varepsilon$$

is equivalent to

$$H_0 : \text{There is } g \in \mathcal{G} \text{ such that } E(g(X) \mid Z) = 0 \text{ and } \|g\|_c > 0.$$

Proof. Suppose that there is g such that $E(g(X) \mid Z) = 0$, $\|g\|_s \leq 2C$, and $\|g\|_c \geq \varepsilon$. Then $\|g\|_s \leq 2C \leq \frac{2C}{\varepsilon}\|g\|_c$. Hence, $g \in \mathcal{G}$, $E(g(X) \mid Z) = 0$, and $\|g\|_c \geq \varepsilon > 0$.

Next let $g \in \mathcal{G}$ such that $E(g(X) \mid Z) = 0$ and $\|g\|_c > 0$. Define $\tilde{g} = g \cdot (\varepsilon / \|g\|_c)$. Then $\|\tilde{g}\|_s = \|g\|_s(\varepsilon/\|g\|_c) \leq 2C$. Hence, $E(\tilde{g}(X) \mid Z) = 0$, $\|\tilde{g}\|_s \leq 2C$, and $\|\tilde{g}\|_c = \varepsilon$. □

The null hypothesis that this restricted version of completeness fails, or equivalently that the instruments are weak, is testable using the test statistic outlined above with a fixed

value of ε_n and a bootstrap critical value instead of a diverging c_n.[7] In the next subsection, I describe additional assumptions, the asymptotic properties of $\hat{T}$, and the bootstrap.

4.1 Asymptotic properties of $\hat{T}$

The null hypothesis we would like to test is

$$H_0 : \text{There is } g \text{ such that } E(g(X) \mid Z) = 0,\ \|g\|_s \leq 2C,\ \text{and } \|g\|_c \geq \varepsilon.$$

In this section, since ε is fixed, I define $\bar{\mathcal{G}} = \{g : \|g\|_s \leq (2C/\varepsilon)\}$ and the corresponding sieve space $\bar{\mathcal{G}}_J$. Just as before, for any $g \in \bar{\mathcal{G}}$ we can write $g = \sum_{j=1}^{\infty} h_j \phi_j(x)$ and $g_J(x) = \sum_{j=1}^{J} h_j \phi_j(x)$ denotes the projection of g on the sieve space. Recall the test statistic

$$\hat{T} = \inf_{g \in \bar{\mathcal{G}}_J : \|g\|_c = 1} \int \left(\int g(x) \hat{f}_{XZ}(x,z) dx \right)^2 dz$$

and the operator

$$S_0(g) = \int \left(\int g(x) f_{XZ}(x,z) dx \right)^2 dz.$$

I will focus on the case where the space of functions $\{g : \|g\|_s < \infty\}$ is a Hilbert space, with inner product $\langle \cdot, \cdot \rangle_s$, so that I can make use of concepts such as orthogonality. The assumption holds for example if the strong norm is a Sobolev norm.

Assumption 6. Let $\mathcal{S} = \{g : \|g\|_s < \infty\}$. The function space $(\mathcal{S}, \|\cdot\|_s)$ is a Hilbert space.

Define the null space of S_0 by $\mathcal{N} = \{g \in \mathcal{S} : S_0(g) = 0\}$ and its orthogonal space by $\mathcal{N}^{\perp} = \{g \in \mathcal{S} : \langle g, \bar{g} \rangle_s = 0 \text{ for all } \bar{g} \in \mathcal{N}\}$. I now state and discuss the remaining assumptions.

Assumption 7. $E(|\phi_k(X_i)|^3 |\phi_j(Z_i)|^3) \leq C_m$ for all j and k.

Assumption 8. $n\delta_n \to 0$ where

$$\delta_n = \sup_{g \in \bar{\mathcal{G}} : \|g\|_c \leq 1} \int \left(\int g(x)(f_J(x,z) - f_{XZ}(x,z)) dx \right)^2 dz.$$

Assumption 9. For any $\bar{C} > 0$, there exists some constant C_t such that

$$\inf_{g_J \in \mathcal{S} : g \in \mathcal{N}^{\perp}, \|g_J\|_c \geq \frac{\bar{C}}{J\sqrt{\ln(n)}}, \|g_J\|_s \leq (2C/\varepsilon)} \int \left(\int f_J(x,z) g_J(x) dx \right)^2 dz \geq \frac{C_t J^2 \ln(n)}{\sqrt{n}}.$$

[7] See Appendix B.1 for an example illustrating testability with these smoothness restrictions.

Assumption 10. Define

$$\mathcal{H}_J = \{g_J \in \bar{\mathcal{G}}_J : g_J = g_J^1 + g_J^2, \|g_J^1\|_s + \|g_J^2\|_s \leq (2C/\varepsilon), \|g_J^1\|_c = 1, g^1 \in \mathcal{N}, g^2 \in \mathcal{N}^\perp\}.$$

Let h^1 and h^2 be the coefficients of the series expansions of g^1 and g^2, respectively. Let W be a random matrix distributed as the (normal) limiting distribution of $\sqrt{n}(\hat{A} - A)$. Assume

$$\sup_{t \geq 0} \left| P\left(\min_{g_J \in \mathcal{H}_J} \|Wh^1 + \sqrt{n}Ah^2\| \leq t + d_n \right) - P\left(\min_{g_J \in \mathcal{H}_J} \|Wh^1 + \sqrt{n}Ah^2\| \leq t \right) \right| = o(1)$$

for all d_n such that $d_n \to 0$.

Assumption 7 is a moment condition, which holds for example if the basis functions are bounded or if $\int |\phi_j(x)|^3 dx$ is bounded uniformly over j. Assumption 8 implies that $J \to \infty$ as $n \to \infty$ fast enough relative to n. Assumption 9 says that J cannot converge too fast. It directly implies for example that $\frac{J^2 \ln(n)}{\sqrt{n}} \to 0$ as $n \to \infty$. The assumptions hold as long as $\bar{s}$ is large enough relative to the smoothness of the density, which is similar to other assumptions in the nonparametric IV literature. These assumptions allow for both severely and mildly ill-posed problems. Appendix B.2 contains a detailed explanation of these assumptions. Assumption 10 is a continuity condition on the asymptotic distribution. It holds as long as the restrictions in $\bar{\mathcal{G}}$ are strong enough, in particular as long as h_j converges to 0 fast enough.

We now get the following result, which characterizes the asymptotic distribution of the test statistic.

Theorem 3. *Suppose Assumptions 1 - 10 hold and $J/n \to 0$. Then under H_0*

$$\sup_{t \geq 0} \left| P\left(n\hat{T} \leq t \right) - P\left(\min_{g_J \in \mathcal{H}_J} \|Wh^1 + \sqrt{n}Ah^2\|^2 \leq t \right) \right| = o(1).$$

4.2 Bootstrap critical value

Simply resampling the data, using a bootstrap analog of the test statistic, and using the resulting quantiles as critical values does not control size, even for a fixed distribution of the data. The reason is that the population version of the test statistic might have multiple solutions and the asymptotic distribution changes discontinuously with the number of solutions. Instead, the critical value I use is the $1 - \alpha$ quantile of nT^*, denoted by c_α^*, where

$$T^* = \min_{g_J \in \mathcal{S}: g_J = g_J^1 + g_J^2, \|g_J^1\|_s + \|g_J^2\|_s \leq (2C/\varepsilon), \|g_J^1\|_c = 1, \|g_J^2\|_c \leq \lambda_n} \|(A^* - \hat{A})h^1 + \hat{A}h^2\|^2 + \frac{\mu_n}{n}\|\hat{A}h^1\|^2,$$

A^* is the bootstrap analog of $\hat{A}$,

$$\frac{\mu_n}{n} \to 0 \quad \text{and} \quad \frac{\sqrt{n}}{\mu_n} \to 0,$$

and $\lambda_n = \bar{C}/(J\sqrt{\ln(n)})$ for some constant $\bar{C} > 0$. The penalty function guarantees that minimizing the penalized objective is asymptotically equivalent to minimizing over g_J^1 such that $g^1 \in \mathcal{N}$ (in the sense that the error is $o_p(1)$). Moreover, since μ_n converges slow enough, the penalty will then not affect the asymptotic distribution under H_0, but we get consistency against a fixed alternative. The constraint $\|g_J^2\|_c \leq \bar{C}/(J\sqrt{\ln(n)})$ is needed because $\hat{A}$ has full rank even if A does not, which is the reason why a standard nonparametric bootstrap procedure does not control size. It ensures that

$$\|(A^* - \hat{A})h^1 + \hat{A}h^2\| = \|(A^* - \hat{A})h^1 + Ah^2 + (A - \hat{A})h^2\| = \|(A^* - \hat{A})h^1 + Ah^2\| + o_p(1/\sqrt{n}),$$

which then also allows us to restrict $g^2 \in \mathcal{N}^\perp$ (and thus $g_J \in \mathcal{H}_J$) without affecting the asymptotic distribution. The formal result is now as follows.

Theorem 4. *Suppose Assumptions 1 - 10 hold. Then under H_0*

$$\sup_{t \geq 0} \left| P^*\left(nT^* \leq t\right) - P^*\left(\min_{g_J \in \mathcal{H}_J} \|W^*h^1 + \sqrt{n}Ah^2\|^2 \leq t\right)\right| = o(1),$$

where W^* is a random matrix distributed as the limit distribution of $\sqrt{n}(A^* - \hat{A})$ under P^*. It follows that under H_0

$$P(n\hat{T} \geq c_\alpha^*) \to \alpha.$$

Moreover under any fixed alternative

$$P(n\hat{T} \geq c_\alpha^*) \to 1.$$

Remark 1. The previous result is pointwise for each distribution P and (most likely) does not hold uniformly over $\mathcal{P}$ (all distributions satisfying Assumption 1). To be precise, I expect that the test does not control size for some sequences of distributions for which

$$\inf_{g_J \in \mathcal{S}: g \in \mathcal{N}^\perp, \|g_J\|_c \geq \frac{1}{J\sqrt{\ln(n)}}, \|g_J\|_s \leq (2C/\varepsilon)} \int \left(\int f_J(x,z) g_J(x) dx\right)^2 dz \geq \frac{C_t J^2 \ln(n)}{\sqrt{n}}$$

is violated. The finite dimensional analog of the test statistic is the minimum eigenvalue of a matrix, where the minimum population eigenvalue is 0. A violation of the previous

inequality in this case would occur if one or more population eigenvalues were close to 0. This issue is unrelated to the nontestability result of Canay et al. (2013), who say that, without restrictions on $\mathcal{G}$, having both uniform size control and consistency against a fixed alternative is not possible. In fact, it can be seen from the proofs of Theorems 1 and 2 that under H_0,

$$\sup_{P \in \mathcal{P}} n\hat{T} \leq J^2 \ln(n)$$

almost surely as $n \to \infty$. Contrarily, for any fixed complete distribution,

$$n\hat{T} \geq c\left(n - J^2 \ln(\ln(n))\right)$$

for some constant c. Hence, a critical value of $J^2 \ln(n)$ and $\frac{n}{J^2 \ln(n)} \to \infty$ would control size uniformly and yield consistency against fixed alternatives. The lack of uniformity here is due to the difficulty of obtaining critical values for a test statistic which is the minimum of a random function and where the population minimum might not be unique. This problem could potentially be solved using an approach along the lines of recent work by Bugni, Canay, and Shi (2014), which is left for future research.

5 Choice of norms

In this section I discuss classes of functions which satisfy Assumption 2. I also show how C can be chosen in particular applications and that it can be implied by economic theory.

5.1 Compact function spaces

The most commonly used consistency norms are the L^2-norm, $\|g\|_c = \left(\int g(x)^2 dx\right)^{1/2}$, and the sup-norm, $\|g\|_c = \sup_x |g(x)|$. Suppose that the consistency norm is the L^2-norm. Then, as shown in Section 3.2, a convenient choice for the strong norm $\|\cdot\|_s$ is the Sobolev norm

$$\|g\|_s = \sqrt{\sum_{0 \leq \lambda \leq m} \int \left(D^\lambda g(x)\right)^2 dx},$$

where $m \geq 1$ and D^λ denotes the λ weak derivative of the function $g(x)$. If instead the consistency norm is the sup-norm, one could either use the Sobolev norm above or the Hölder norm

$$\|g\|_s = \max_{0 \leq |\lambda| \leq m} \sup_{x \in (0,1)} |\nabla^\lambda g(x)| + \sup_{x_1, x_2 \in (0,1)} \frac{|\nabla^m g(x_1) - \nabla^m g(x_2)|}{|x_1 - x_2|^\nu},$$

where $\nabla^\lambda g(x)$ denotes the λ derivative of the function $g(x)$, and $0 < \nu \leq 1$. In the first case, $\tilde{\mathcal{G}}$ is a Sobolev space while in the second case, $\tilde{\mathcal{G}}$ is a Hölder space. Similar as the strong norm, the consistency norm could also be defined using derivatives of higher order.

In all these cases it can be shown that $\tilde{\mathcal{G}}$ is compact under $\|\cdot\|_c$. See Freyberger and Masten (2014) for an overview of the compactness results. Moreover, it is easy to see that with these choices $\|g\|_c^2 \geq \int g(x)^2 dx$. Note, however, that while a Hölder space could be used in Theorems 1 and 2, it does not satisfy Assumption 6.

5.2 Example of norm bound

One assumption which is maintained throughout the paper is that $\|g_0\|_s \leq C$. As a consequence, the test involves a constraint on the strong norm and C needs be chosen by the researcher. I now explain how this can be done in two popular examples, namely estimation of Engel curves and demand functions.

Let X^* be total household expenditure and let $X = \log(X^*)$. Let Y^* be the total expenditure on a certain good, such as food, and define $Y = \frac{Y^*}{X^*}$, which is the expenditure share. Let Z be the gross earnings of the head of the household. This setup is studied by Blundell et al. (2007) and Santos (2012) among others. A reasonable assumption is that if a household increases total expenditure by $\$\delta$, the total expenditure on food does not increase by more than $\$\delta$ and it does not decrease.

If $X^* = \bar{X}^*$ and if we want to increase $\log(\bar{X}^*)$ to $\log(\bar{X}^*) + \delta$, then we need to increase X^* by $\bar{X}^*(\exp(\delta) - 1)$. Then the total expenditure is $\bar{X}^* \exp(\delta)$ and expenditure on food is not more that $Y^* + \bar{X}^*(\exp(\delta) - 1)$ and not less than Y^* In other words, for any X^*, the derivative of the Engel curve is bounded above by

$$\lim_{\delta \to 0} \frac{\frac{Y^* + X^*(\exp(\delta)-1)}{X^* \exp(\delta)} - \frac{Y^*}{X^*}}{\delta} = \lim_{\delta \to 0} \frac{\frac{Y^*(1-\exp(\delta)) + X^*(\exp(\delta)-1)}{X^* \exp(\delta)}}{\delta} = \lim_{\delta \to 0} \frac{(\exp(\delta) - 1)}{\delta \exp(\delta)} \frac{X^* - Y^*}{X^*} \leq 1$$

since $\left|\frac{X^* - Y^*}{X^*}\right| \leq 1$ and $\lim_{\delta \to 0} \frac{(\exp(\delta)-1)}{\delta \exp(\delta)} = 1$. Similarly, the derivative is bounded below by

$$\lim_{\delta \to 0} \frac{\frac{Y^*}{X^* \exp(\delta)} - \frac{Y^*}{X^*}}{\delta} = \lim_{\delta \to 0} \frac{\frac{Y^*(1-\exp(\delta))}{X^* \exp(\delta)}}{\delta} = \lim_{\delta \to 0} -\frac{(\exp(\delta) - 1)}{\delta \exp(\delta)} \frac{Y^*}{X^*} \geq -1.$$

If $X \in [a, b]$ we can use the regressor $(X - a)/(b - a) \in [0, 1]$. Then $\sup |g_0'(x)| \leq b - a$ and clearly $\sup |g_0(x)| \leq 1$. Let

$$\|g_0\|_s = \sup_x |g_0(x)| + \sup_{x_1, x_2} \frac{|g_0(x_1) - g_0(x_2)|}{|x_1 - x_2|}.$$

Then $\|g_0\|_s \leq 1 + b - a$ and we can choose $C = 1 + b - a$. If instead

$$\|g_0\|_s = \left(\int \left(g_0(x)^2 + g_0'(x)^2 \right) dx \right)^{1/2},$$

we get $\|g_0\|_s \leq \sqrt{1 + (b-a)^2}$.

If $g_0(x)$ is a demand function, then one can use bounds on price elasticities, and bounds on the support of quantity and price. In this way, one can obtain bounds on the derivatives and function values of the demand function using similar arguments as above.

6 Extension to functions on $\mathbb{R}$

The analysis could be extended to functions on $\mathbb{R}$ by using weighted norms. In this section, I provide the main ideas including specific examples of norms which satisfy compactness, and the test statistic. Let $w(x) = e^{-x^2}$ and let $\phi_j(x)$ be Hermite polynomials (see for example Chen 2007) so that

$$\int \phi_j(x)^2 w(x) dx = 1$$

and for $j \neq k$

$$\int \phi_k(x)\phi_j(x)w(x)dx = 0.$$

Let the consistency norm be the weighted L^2-norm

$$\|g\|_c = \sqrt{\int g(x)^2 w(x) dx}.$$

Then for every function g for which $\int g(x)^2 w(x)dx < \infty$, we can write

$$g(x) = \sum_{j=1}^{\infty} h_j \phi_j(x),$$

where $h_j \equiv \int g(x)\phi_j(x)w(x)dx$. Moreover, if $f_{XZ}(x,z)$ is square integrable, we can write

$$f_{XZ}(x,z) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}\phi_j(x)\phi_k(z)$$

where

$$a_{jk} = \int f_{XZ}(x,z)\phi_j(x)\phi_k(z)w(x)w(z)dx = E(\phi_j(X)\phi_k(Z)w(X)w(Z)).$$

Hence, we can estimate a_{jk} by

$$\hat{a}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\phi_k(Z_i)w(X_i)w(Z_i)$$

and f_{XZ} by

$$\hat{f}_{XZ}(x,z) = \sum_{j=1}^{J}\sum_{k=1}^{J} \hat{a}_{jk}\phi_j(x)\phi_k(z).$$

Now let

$$\begin{aligned} S_0(g) &= \int \left(\int g(x) f_{XZ}(x,z) w(x) dx \right)^2 w(z) dz \\ &= \int \left(\int g(x) w(x)^{1/2} f_{XZ}(x,z) w(x)^{1/2} w(z)^{1/2} dx \right)^2 dz. \end{aligned}$$

It can be shown that $S_0(g)$ is continuous in g under $\|\cdot\|_c$ as long as $|f_{XZ}(x,z)| \leq C_d$. With this choice of the consistency norm, we get a compact parameter space for example if

$$\|g\|_s = \sqrt{\int \sum_{0 \leq \lambda \leq m} \left(D^\lambda g(x)\right)^2 \tilde{w}(x) dx},$$

where $\tilde{w}(x) = (1+x^2)^{-\delta}$ for any $\delta > 0$ and $D^\lambda g(x)$ denotes the λ weak derivative of g. See Freyberger and Masten (2014) for the formal compactness result, which builds on results of Gallant and Nychka (1987). With these norms we can define the parameter spaces $\tilde{\mathcal{G}}$ and $\mathcal{G}$ just as before. Notice that in this case, we would assume that $\|g_0\|_s \leq C$. Hence, g_0 could be unbounded and it could have unbounded derivatives.

Interestingly, with these choices of norms, the constraints and the objective function are again easy to implement because

$$\|g\|_c^2 = \int \left(\sum_{j=1}^{\infty} h_j \phi_j(x) \right)^2 w(x) dx = \sum_{j=1}^{\infty} h_j^2$$

as before. Moreover, when we approximate g with $g_J = \sum_{j=1}^{J} h_j \phi_j(x)$ we get

$$\hat{S}(g_J) = \int \left(\int \sum_{j=1}^{J} h_j \phi_j(x) \sum_{j=1}^{J}\sum_{k=1}^{J} \hat{a}_{jk}\phi_j(x)\phi_k(z) w(x) dx \right)^2 w(z) dz = h' \hat{A}' \hat{A} h.$$

Due to continuity of $S_0(g)$ and compactness of the parameter space, it again holds that the infimum of $S_0(g)$, over all functions in $\tilde{\mathcal{G}}$ with a consistency norm of ε_n, is nonzero if and only if

$$H_0 : \text{There is } g \text{ such that } E(g(X) \mid Z) = 0,\ \|g\|_s \leq 2C, \text{ and } \|g\|_c \geq \varepsilon_n$$

holds. This result, combined with similar assumptions as those in this paper, can be used to link the outcome of the test to consistency of the estimator. Notice that the consistency

result, analogous to that of Theorem 1, will hold with respect to the weighted L^2-norm. Consistency in this norm would for example imply consistency in the L^2-norm over any compact subset of the support. Finally, weak instruments can be defined (and tested for) just as in Section 4. The definition of weak instruments would imply that instruments are strong if the bias in the weighted L^2-norm is smaller than some ε.

7 Monte Carlo simulations

In this section, I illustrate the finite sample properties of the test and the estimator using two types of distributions of X and Z. First, I consider a sequence of incomplete distributions which converges to a complete distribution. Second, I consider a sequence of complete distributions which converges to an incomplete distribution. In both cases I use the norms

$$\|g\|_c^2 = \int_0^1 g(x)^2 dx \quad \text{and} \quad \|g\|_s^2 = \int_0^1 \left(g(x)^2 + g'(x)^2\right) dx.$$

As basis functions I use Legendre polynomials normalized such that they are orthonormal on $[0,1]$. All results are based on 1000 Monte Carlo simulations.

For the first example let

$$f_{XZ}(x,z) = \sum_{j=1}^{\infty} d_{jj}\varphi_j(x)\varphi_j(z),$$

where $d_{11} = 1$, $\varphi_1(x) = 1$, $d_{jj} = \sqrt{0.2}(j-1)^{-2}$, and $\varphi_j(x) = \sqrt{2}\cos((j-1)\pi x)$ for $j \geq 2$. This example is taken from Horowitz (2011) and from the arguments in Appendix B.2 it follows that the distribution is complete. Let $f_{XZ}^k(x,z)$ be defined as $f_{XZ}(x,z)$ but with $d_{kk} = 0$ for some $k \geq 2$. Notice that $f_{XZ}^k(x,z)$ does not satisfy completeness without restrictions on $\mathcal{G}$. Also notice that with all these distributions the marginal distributions of X and Z are uniform on $[0,1]$. Figure 1 shows plots of $F_{X|Z}(x;z)$ and $F_{X|Z}^k(x;z)$ as functions of x for different values of k. In the first panel $z = 0.1$, in the second $z = 0.5$, and in the third $z = 0.9$. The different shapes indicate that X and Z are not independent, and the dependence is least obvious when $d_{22} = 0$. The figure also illustrates that $F_{X|Z}(x;z)$, $F_{X|Z}^4(x;z)$, and $F_{X|Z}^5(x;z)$ are almost identical. Although the differences are very small, they do exist. For example $F_{X|Z}(0.6;0.9) = 0.3308$, $F_{X|Z}^5(0.6;0.9) = 0.3294$, and $F_{X|Z}^4(0.6;0.9) = 0.3271$. Under these three distributions the correlation between X and Z is approximately 0.44. In case $d_{22} = 0$, the correlation is only -0.0135.

I simulate the data by taking draws Z_i and U_i^1 from the uniform distribution on $[0,1]$ (the marginal distribution of Z). I then choose X_i which satisfies $F_{X|Z}^k(X_i;Z_i) = U_i^1$. Moreover,

$$g_0(x) = \frac{1}{2}\exp(x) - x^3$$

and I set

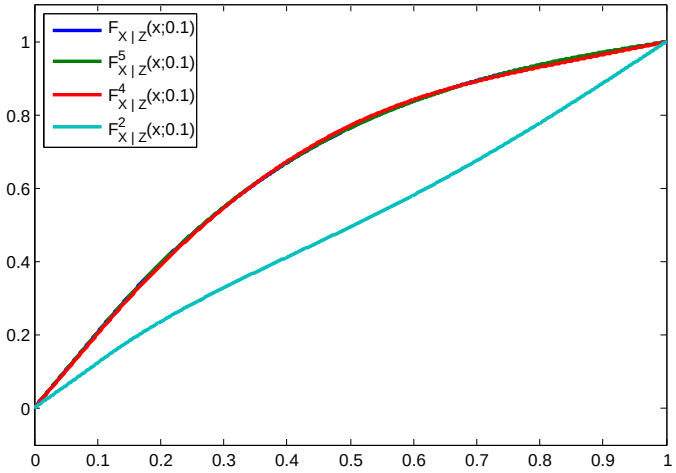
$$Y_i = g_0(X_i) + U_i,$$

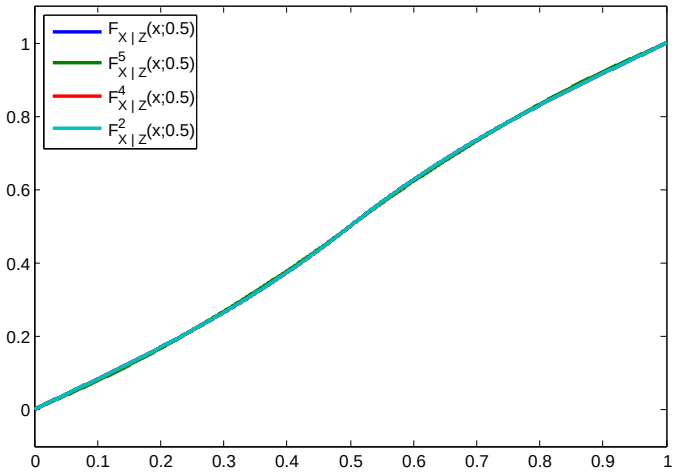
where $U_i = 0.01U_i^1 - U_i^2$ and $U_i^2 \sim N(1/200, 0.1^2)$. Then U_i and X_i are not independent and $var(U_i) \approx 0.01$, which is similar as in Horowitz (2011). With this choice $\|g_0\|_s = 0.9065$ and I set $C = 2$. I choose $\varepsilon_n = 1/(2\ln(n)^{1/3})$ and I set $c_n = J\ln(n)/10$. I divide by 10 because κ_n tends to be small and thus, without the division, the test tends to be too conservative.

Table 1 shows $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n)$, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$, and $P(n\hat{T} \geq c_n)$ for $(n, J) = (1000, 3)$ and $(n, J) = (5000, 4)$. When $n = 1000$ and $J = 3$, $P(n\hat{T} \geq c_n)$ is close to 95% in the first three cases and close to 0 in the last case. Hence, when J is small, the test cannot distinguish between the complete and the (very close to complete) incomplete distributions, but it can distinguish the last distribution. In the first three cases the estimator has similar properties as well, in particular $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ is quite small, while it is large in the last case. In all four cases $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n)$ is below 0.16 and it decreases further as n and J increase. Notice that the basis functions used to construct the densities and the ones used for the test are different. Thus, in all four cases the $J \times J$ matrix A has full rank for any fixed J and as a consequence, to control size, it is crucial that $J \to \infty$ as $n \to \infty$. When $n = 5000$ and $J = 4$, the rejection probability decreases when $d_{22} = 0$. More interestingly, now the rejection probability with $d_{44} = 0$ is only 5.7%, while the rejection probability for the complete distribution and the one with $d_{55} = 0$ is around 85%. That is, although the distributions are extremely close, the test is able to distinguish them if n and J are large enough. Nonetheless, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ is similar in the first three cases and the difference will show up more noticeably when n and J are even larger. Again, the test cannot distinguish between the complete distribution and the one with $d_{55} = 0$, but the properties of the estimator are also similar and thus $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\, n\hat{T} \geq c_n)$ is very small in both cases. The test can pick up this difference as well when n and J are larger. More generally, for every n and J there exists a k such that the test cannot distinguish between f_{XZ}^k and f_{XZ}, which illustrates the non-testability result of Canay et al. (2013), but the properties of the estimator will then also be almost identical under both distributions.

For the second example suppose that $\tilde{X}$ and $\tilde{Z}$ are jointly normally distributed with means 0, variances 1, and correlation ρ. Let $X = \Phi(\tilde{X})$ and $Z = \Phi(\tilde{Z})$. Since $f_{\tilde{X}\tilde{Z}}$ is

Figure 1: Distribution functions conditional on $Z = 0.1$, $Z = 0.5$, and $Z = 0.9$, respectively





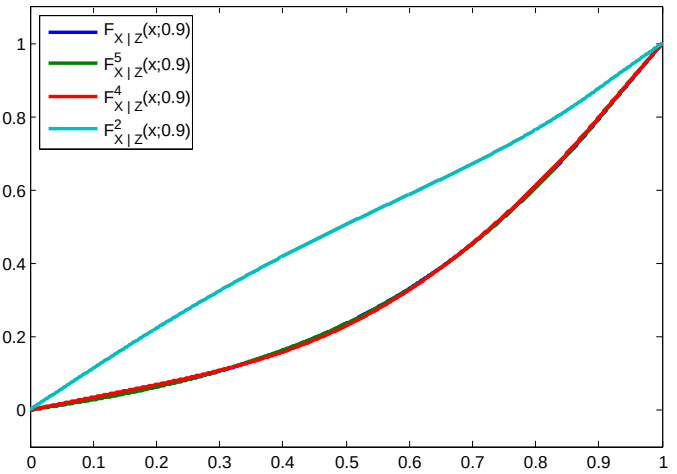


Table 1: Probabilities with sequence of incomplete distributions

		$n = 1000,\ J = 3$	$n = 5000,\ J = 4$
Complete	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.152	0.024
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.166	0.025
	$P(n\hat{T} \geq c_n)$	0.964	0.848
$d_{55} = 0$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.155	0.025
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.171	0.028
	$P(n\hat{T} \geq c_n)$	0.962	0.846
$d_{44} = 0$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.148	0.000
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.165	0.034
	$P(n\hat{T} \geq c_n)$	0.961	0.057
$d_{22} = 0$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.083	0.027
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.752	0.643
	$P(n\hat{T} \geq c_n)$	0.135	0.055

Table 2: Probabilities with sequence of complete distributions

		$n = 1000,\ J = 3$	$n = 5000,\ J = 4$
$\rho = 0.5$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.029	0.000
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.029	0.000
	$P(n\hat{T} \geq c_n)$	1.000	1.000
$\rho = 0.3$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.186	0.009
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.287	0.133
	$P(n\hat{T} \geq c_n)$	0.737	0.144
$\rho = 0.1$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.067	0.000
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.653	0.417
	$P(n\hat{T} \geq c_n)$	0.129	0.005
$\rho = 0$	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n,\ n\hat{T} \geq c_n)$	0.017	0.005
	$P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$	0.816	0.779
	$P(n\hat{T} \geq c_n)$	0.024	0.006

complete if and only if $\rho > 0$ and since completeness is preserved under strictly monotone transformations, also f_{XZ} is complete if and only if $\rho > 0$. In case $\rho = 0$, X and Z are independent. The function g_0 is defined just as before, and I also make use of the same parameter choices. Table 2 shows the outcomes for different choices of ρ, n, and J. As expected, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ decreases and $P(n\hat{T} \geq c_n)$ increases as ρ increases. When $\rho = 0.5$, we reject with probability close to 1 and the estimator is close to g_0 with large probability. Contrarily, when $\rho = 0$, we reject with probability close to 0 and $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ is large. Furthermore, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n, n\hat{T} \geq c_n)$ is very small when $n = 5000$ and $J = 4$ for all ρ. While $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ decreases as n increases for all $\rho > 0$, $P(n\hat{T} \geq c_n)$ decreases as n increases for $\rho = 0.1$ and $\rho = 0.3$. If $\rho = 0.1$, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ is quite large, so having a low rejection probability is desirable. If $\rho = 0.3$, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n)$ is small, so ideally we would like to reject in more cases. One reason for this feature is that the dimension of A increases as J increases and consequently, the minimum eigenvalue gets smaller. Thus, for a fixed n, the power decreases as J increases and the finite sample power might decrease when both n and J increase (see also the first example). Here J might be too large relative to the sample size, which is sensible because the data is based on a transformation of the normal distribution which is a very smooth. The properties improve when $n = 20,000$ and $J = 4$. Then with $\rho = 0.3$ we get $P(n\hat{T} \geq c_n) = 0.429$, $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n) = 0.019$ and $P(\|\hat{g} - g_0\|_c \geq \varepsilon_n, n\hat{T} \geq c_n) = 0.007$.

8 Conclusion

This paper provides the first positive testability result for the identification condition in a nonparametric IV model by linking the properties of the test to consistency of an estimator $\hat{g}$ for the structural function g_0. Specifically, I present a test statistic and an estimator such that uniformly over a large class of distributions, the joint probability that $\|\hat{g} - g_0\|_c \geq \varepsilon$ and that the test rejects goes to 0, for any $\varepsilon > 0$. It follows that for any sequence of distributions for which the test rejects with probability bounded away from 0, $\hat{g}$ will be consistent for g_0. Interestingly, not only is this the case for complete distributions but also for certain sequences of incomplete ones. These findings contrast results in the linear IV model, where any sequence of distributions under which identification fails, leads to an inconsistent estimator of the slope coefficient. However, in case of incomplete distributions, the resulting estimator of g_0 might converge at a slower than optimal rate, which suggests that uniform inference following the test will be problematic, just as in parametric models.

An alternative would be to use a partial identification robust inference approach, such as the one proposed by Santos (2012). Even in this case, the results in this paper will be useful because they provide information about the strength of the instruments and they yield a consistent estimator of g_0, if the instruments are strong enough.

As additional contributions I also provide a definition of weak instruments in the nonparametric IV model and I show that weak instruments are equivalent to the failure of a restricted version of completeness. This version of incompleteness, or equivalently weak instruments, is testable and I provide a test statistic and a bootstrap procedure to obtain the critical value. Rejecting this test leads to the conclusion of strong instruments which means that $\|\hat{g} - g_0\|_c \leq \varepsilon + o_p(1)$ and $\|g_1 - g_0\|_c \leq \varepsilon$ for all g_1 satisfying the moment and smoothness assumptions. Certain estimators of g_0 converge in probability even if g_0 is not point identified. For example, an estimator based on Tikhonov regularization converges to the function in the identified set, say g^*, with the minimal norm (Florens, Johannes, and Van Bellegem, 2011). This result leads to a potential alternative approach for obtaining confidence intervals for $g_0(x)$. If the instruments are strong and $\|\cdot\|_c$ is the sup-norm, a 95% confidence interval for g^* can easily be modified to a 95% confidence interval for g_0. Specifically, if $[\hat{L}(x), \hat{U}(x)]$ is a 95% confidence interval for $g^*(x)$, then with strong instruments $[\hat{L}(x) - \varepsilon, \hat{U}(x) + \varepsilon]$ is a 95% confidence interval for $g_0(x)$.

A Proofs

A.1 Proof of Theorem 1

Let $h \in \mathbb{R}^J$ contain the first J coefficients of the series expansion of g_0. By Assumption 5, the definition of $\hat{h}$ and the triangle inequality

$$\left\|\hat{A}h - \hat{m}\right\| \geq \left\|\hat{A}\hat{h} - \hat{m}\right\| = \left\|\hat{A}h - \hat{m} + \hat{A}(\hat{h} - h)\right\| \geq \left\|\hat{A}(\hat{h} - h)\right\| - \left\|\hat{A}h - \hat{m}\right\|$$

and thus

$$4\left\|\hat{A}h - \hat{m}\right\|^2 \geq \left\|\hat{A}(\hat{h} - h)\right\|^2.$$

Now suppose that $n\hat{T} \geq c_n$ and $\|\hat{g} - g_0\|_c \geq \varepsilon_n$ and notice that $\|\hat{g} - g_0\|_s \leq 2C$. From Assumption 5 it follows that

$$\frac{n}{\varepsilon_n^2}\left\|\hat{A}(\hat{h} - h)\right\|^2 \geq n\hat{T} \geq c_n$$

and thus

$$4\frac{n}{\varepsilon_n^2}\left\|\hat{A}h - \hat{m}\right\|^2 \geq c_n.$$

In other words,

$$\sup_{P\in\mathcal{P}} P\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n, n\hat{T} > c_n\right) \;\leq\; \sup_{P\in\mathcal{P}} P\left(4n\left\|\hat{A}h - \hat{m}\right\|^2 \geq c_n\varepsilon_n^2\right).$$

Next let m be a $J \times 1$ vector with $m_k = E(\hat{m}_k)$ and notice that since

$$\left\|\hat{A}h - \hat{m}\right\| \leq \|(\hat{A} - A)h\| + \|Ah - m\| + \|m - \hat{m}\|$$

we have

$$\left\|\hat{A}h - \hat{m}\right\|^2 \leq 4\|(\hat{A} - A)h\|^2 + 4\|Ah - m\|^2 + 4\|m - \hat{m}\|^2$$

and thus

$$\begin{aligned} P\left(4n\left\|\hat{A}h - \hat{m}\right\|^2 \geq c_n\varepsilon_n^2\right) &\leq P\left(16n\|(\hat{A} - A)h\|^2 + 16n\|Ah - m\|^2 + 16n\|m - \hat{m}\|^2 \geq c_n\varepsilon_n^2\right) \\ &\leq P\left(48n\|(\hat{A} - A)h\|^2 \geq c_n\varepsilon_n^2\right) + P\left(48n\|Ah - m\|^2 \geq c_n\varepsilon_n^2\right) \\ &\quad + P\left(48n\|m - \hat{m}\|^2 \geq c_n\varepsilon_n^2\right). \end{aligned}$$

It now suffices to prove that all three terms on the right hand side converge to 0 uniformly over $P \in \mathcal{P}$.

To show that the first term converges to 0 uniformly over $P \in \mathcal{P}$, write

$$n\|(\hat{A} - A)h\|^2 = \sum_{j=1}^{J} \left(\sum_{k=1}^{J} \sqrt{n}(\hat{a}_{jk} - a_{jk})h_k \right)^2$$

and notice that

$$\sum_{k=1}^{J} \sqrt{n}(\hat{a}_{jk} - a_{jk})h_k = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} (\phi_k(X_i)\phi_j(Z_i) - E(\phi_k(X_i))\phi_j(Z_i))h_k$$

and

$$Var\left(\sum_{k=1}^{J} (\phi_k(X_i)\phi_j(Z_i) - E(\phi_k(X_i))\phi_j(Z_i))h_k \right) \leq \max_{k=1,\ldots,J} E(\phi_k(X_i)^2\phi_j(Z_i)^2) \left(\sum_{k=1}^{J} |h_k| \right)^2.$$

Assumptions 2 and 4 imply that $|h_k| \leq C_g k^{-2}$ (see Appendix B.2 for a derivation). Moreover, by Assumptions 1 and 3

$$E(\phi_k(X_i)^2\phi_j(Z_i)^2) \leq C_d \int\int \phi_k(x)^2\phi_j(z)^2 dxdz = C_d$$

It follows that

$$Var\left(\sum_{k=1}^{J} (\phi_k(X_i)\phi_j(Z_i) - E(\phi_k(X_i))\phi_j(Z_i))h_k \right) \leq \sigma^2$$

where

$$\sigma^2 = C_d C_g^2 \left(\sum_{k=1}^{\infty} k^{-2} \right)^2 < \infty.$$

By Markov's inequality

$$\sup_{P \in \mathcal{P}} P\left(48n\|(\hat{A} - A)h\|^2 \geq c_n \varepsilon_n^2 \right) \leq \frac{48J\sigma^2}{c_n \varepsilon_n^2} \to 0.$$

Similarly,

$$\|\sqrt{n}(m - \hat{m})\|^2 = \sum_{k=1}^{J} \left(\sqrt{n}(\hat{m}_k - m_k) \right)^2 = \sum_{k=1}^{J} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i\phi_k(Z_i) - E(Y_i\phi_k(Z_i))) \right)^2$$

and by Assumptions 1 and 3

$$E\left(\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i\phi_k(Z_i) - E(Y_i\phi_k(Z_i))) \right)^2 \right) \leq \frac{1}{n} \sum_{i=1}^{n} E\left(Y_i^2 \phi_k(Z_i)^2 \right) \leq \sigma_Y^2 C_d.$$

It follows from Markov's inequality that

$$\sup_{P \in \mathcal{P}} P\left(48n\|m - \hat{m}\|^2 \geq c_n \varepsilon_n^2 \right) \leq \frac{48J\sigma_Y^2 C_d}{c_n \varepsilon_n^2} \to 0.$$

Next write

$$\|Ah - m\|^2 = \sum_{j=1}^{J} \left(\sum_{k=1}^{J} a_{jk} h_k - m_j \right)^2 .$$

Since

$$\sum_{j=1}^{\infty} \left(\sum_{k=1}^{\infty} a_{jk} h_k - m_j \right)^2 = 0$$

it holds that

$$m_j = \sum_{k=1}^{\infty} a_{jk} h_k .$$

Therefore

$$\begin{aligned}
\|Ah - m\|^2 &= \sum_{j=1}^{J} \left(\sum_{k=1}^{J} a_{jk} h_k - m_j \right)^2 \\
&= \sum_{j=1}^{J} \left(\sum_{k=J+1}^{\infty} a_{jk} h_k \right)^2 \\
&\leq \sum_{j=1}^{\infty} \left(\sum_{k=J+1}^{\infty} a_{jk} h_k \right)^2 \\
&= \int \left(\int f_{XZ}(x,z) \left(g(x) - g_J(x) \right) dx \right)^2 dz \\
&\leq \int \int f_{XZ}(x,z)^2 dx dz \int \left(g(x) - g_J(x) \right)^2 dx \\
&\leq C_d^2 C_g^2 C_o^2 J^{-2\bar{s}}
\end{aligned}$$

where the last inequality follows from Assumptions 2 and 4. Thus,

$$\sup_{P \in \mathcal{P}} P\left(48n \|Ah - m\|^2 \geq c_n \varepsilon_n^2 \right) \leq \sup_{P \in \mathcal{P}} P\left(C_d^2 C_g^2 C_o^2 J^{-2\bar{s}} n \geq c_n \varepsilon_n^2 \right) \to 0.$$

We can conclude that

$$\sup_{P \in \mathcal{P}} P\left(\|\hat{g} - g_0\|_c \geq \varepsilon_n, n\hat{T} > c_n \right) \to 0.$$

A.2 Proof of Theorem 2

First notice that

$$\|\hat{A}h\|^2 \geq \frac{3}{4} \|Ah\|^2 - 3\|(\hat{A} - A)h\|^2.$$

For any $g_J \in \bar{\mathcal{G}}_J(\varepsilon_n)$ with $\|g_J\|_c = 1$, let $h \in \mathbb{R}^J$ be the coefficients of the series expansion. Then

$$\|Ah\|^2 = \left(\int f_J(x,z) g_J(x) \right)^2 dz \geq \kappa_n .$$

Also notice that

$$\|(\hat{A}-A)h\|^2 = \sum_{j=1}^{J}\left(\sum_{k=1}^{J}(\hat{a}_{jk}-a_{jk})h_k\right)^2 \leq C_o\sum_{j=1}^{J}\sum_{k=1}^{J}(\hat{a}_{jk}-a_{jk})^2$$

for all h with $\|h\|^2 \leq C_o$ It follows that

$$\begin{aligned} n\hat{T} &\geq \inf_{g\in\bar{\mathcal{G}}(\varepsilon_n):\|g\|_c=1}\frac{3}{4}n\|Ah\|^2 - \sup_{g\in\bar{\mathcal{G}}(\varepsilon_n):\|g\|_c=1}3n\|(\hat{A}-A)h\|^2 \\ &\geq \frac{3}{4}n\kappa_n - 3C_o\sum_{j=1}^{J}\sum_{k=1}^{J}\left(\sqrt{n}(\hat{a}_{jk}-a_{jk})\right)^2. \end{aligned}$$

Since

$$\sqrt{n}(\hat{a}_{jk}-a_{jk}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\phi_k(X_i)\phi_j(Z_i) - E(\phi_k(X_i))\phi_j(Z_i))$$

and

$$Var\left(\phi_k(X_i)\phi_j(Z_i) - E(\phi_k(X_i))\phi_j(Z_i)\right) \leq C_d,$$

the law of iterated logarithm implies that

$$3C_o\sum_{j=1}^{J}\sum_{k=1}^{J}\left(\sqrt{n}(\hat{a}_{jk}-a_{jk})\right)^2 \leq 3C_oC_dJ^2\ln(\ln(n))$$

almost surely and thus

$$n\hat{T} \geq \frac{3}{4}n\kappa_n - 3C_oC_dJ^2\ln(\ln(n)).$$

Therefore for n and J large enough

$$\begin{aligned} P(n\hat{T}\geq c_n) &\geq P\left((3/4)n\kappa_n - 3C_oC_dJ^2\ln(\ln(n)) \geq c_n\right) \\ &\to 1. \end{aligned}$$

A.3 Proof of Theorem 3

Let $\bar{C}$ denote a generic constant that may differ in different uses. Let g_J be a series approximation of a function g such that $S_0(g) = 0$, $\|g\|_c = 1$ and $\|g\|_s \leq (2C/\varepsilon)$. Such a function exists under the null hypothesis and the series approximation g_J, divided by its consistency norm, is feasible by Assumption 5. Let $h \in \mathbb{R}^J$ be the vector containing the coefficients of this normalized series approximation. Then

$$n\|\hat{A}h\|^2 = n\|(\hat{A}-A)h + Ah\|^2 \leq 2\|\sqrt{n}(\hat{A}-A)h\|^2 + 2n\|Ah\|^2.$$

Notice that

$$\begin{aligned}\|Ah\|^2 &= \int \left(\int f_J(x,z) g_J(x) dx \right)^2 dz \\ &= \int \left(\int f_J(x,z) g(x) dx \right)^2 dz \\ &= \int \left(\int (f_J(x,z) - f_{XZ}(x,z)) g(x) dx \right)^2 dz \\ &\leq \delta_n.\end{aligned}$$

The third line follows because $S_0(g) = 0$. Assumption 8 implies that $n\|Ah\|^2 \to 0$. Moreover, notice that $\sqrt{n}(\hat{a}_{jk} - a_{jk})$ converges to a normally distributed random variable for each j and k by Assumptions 1 and 7. Also notice that the proof of Theorem 2 implies that for some constant $\bar{C}$ it holds almost surely that $2\|\sqrt{n}(\hat{A} - A)h\|^2 \leq \bar{C} J^2 \ln(\ln(n))$. It follows that

$$n\hat{T} \leq 2\bar{C} J^2 \ln(\ln(n)).$$

With this result, we can restrict the class of functions we need to minimize over. In particular recall the sets $\mathcal{S} = \{g : \|g\|_s < \infty\}$, the null space of S_0, namely $\mathcal{N} = \{g \in \mathcal{S} : S_0(g) = 0\}$, and the orthogonal space $\mathcal{N}^\perp = \{g \in \mathcal{S} : \langle g, \bar{g} \rangle_s = 0 \text{ for all } \bar{g} \in \mathcal{N}\}$. Since $(\mathcal{S}, \|\cdot\|_s)$ is Hilbert space by Assumption 6, and since $\mathcal{N}$ is a closed linear subspace of $\mathcal{S}$, it follows that we can decompose any $g \in \mathcal{S}$ as

$$g = g^1 + g^2,$$

where $g^1 \in \mathcal{N}$ and $g^2 \in \mathcal{N}^\perp$. If $\|g\|_s \leq (2C/\varepsilon)$, we get $\|g^1\|_s + \|g^2\|_s \leq (2C/\varepsilon)$, and thus both $\|g^1\|_s \leq (2C/\varepsilon)$ and $\|g^2\|_s \leq (2C/\varepsilon)$. Now for each $g \in \bar{\mathcal{G}}$ with $\|g\|_c = 1$, the series approximation can be written as $g_J = g_J^1 + g_J^2$ with coefficients $h = h^1 + h^2 \in \mathbb{R}^J$. Also notice that by the previous arguments

$$\begin{aligned}\sqrt{n}\|\hat{A}h\| &= \sqrt{n}\|\hat{A}h^1 + \hat{A}h^2\| \\ &\geq \sqrt{n}\|\hat{A}h^2\| - \sqrt{n}\|\hat{A}h^1\| \\ &\geq \sqrt{n}\|Ah^2\| - \sqrt{n}\|(\hat{A} - A)h^2\| - \sqrt{3\bar{C}J^2 \ln(\ln(n))} \\ &\geq \sqrt{n}\|Ah^2\| - 3\sqrt{\bar{C}J^2 \ln(\ln(n))}.\end{aligned}$$

Since any optimal g_J needs to satisfy $\sqrt{n}\|\hat{A}h\| \leq \sqrt{2\bar{C}J^2 \ln(\ln(n))}$ it follows that

$$5\sqrt{\bar{C}J^2 \ln(\ln(n))} \geq \sqrt{n}\|Ah^2\|.$$

Next notice that

$$\|Ah^2\|^2 = \int \left(\int f_J(x,z) g_J^2(x) dx \right)^2 dz.$$

Now suppose $\|g_J^2\|_c \geq \frac{\bar{C}}{J\sqrt{\ln(n)}}$ for some $\bar{C} > 0$. Then since $\|g_J^2\|_s \leq (2C/\varepsilon)$, Assumption 9 implies that

$$\|Ah^2\|^2 \geq \frac{C_t J^2 \ln(n)}{\sqrt{n}}.$$

But we also have

$$\|Ah^2\|^2 \leq \frac{25\bar{C}J^2 \ln(\ln(n))}{n},$$

which is a contradiction for n large enough. Define $\lambda_n = \frac{\bar{C}}{J\sqrt{\ln(n)}}$. Then

$$\|g_J^2\|_c \leq \lambda_n \to 0.$$

Let

$$\tilde{\mathcal{H}}_J = \{g_J \in \bar{\mathcal{G}}_J : g_J = g_J^1 + g_J^2, \|g_J^1\|_s + \|g_J^2\|_s \leq (2C/\varepsilon), \|g_J\|_c = 1, S_0(g^1) = 0, g^2 \in \mathcal{N}^\perp\}.$$

It follows that almost surely

$$\begin{aligned}
\sqrt{n\hat{T}} &= \min_{g_J \in \tilde{\mathcal{H}}_J} \|\sqrt{n}(\hat{A} - A)(h^1 + h^2) + \sqrt{n}A(h^1 + h^2)\| \\
&= \min_{g_J \in \tilde{\mathcal{H}}_J} \|\sqrt{n}(\hat{A} - A)h^1 + \sqrt{n}Ah^2 + \sqrt{n}Ah^1 + \sqrt{n}(\hat{A} - A)h^2\| \\
&= \min_{g_J \in \tilde{\mathcal{H}}_J} \|\sqrt{n}(\hat{A} - A)h^1 + \sqrt{n}Ah^2\| + o(1)
\end{aligned}$$

because

$$\|\sqrt{n}Ah^2\| \leq \sqrt{n\delta_n} \to 0$$

and

$$\|\sqrt{n}(\hat{A} - A)h^2\| \leq \bar{C}\sqrt{\ln(\ln(n))}J\lambda_n \to 0$$

uniformly over all $g_J \in \tilde{\mathcal{H}}_J$. Moreover, the optimal solution will satisfy $\|g_J^2\|_c \leq \lambda_n$.

By Skorokhod's representation theorem there exists a matrix W_n which has the same distribution as $\sqrt{n}(\hat{A} - A)$ and W_n converges to W almost surely. Moreover, since W_n is a sample average and we assume that the third moment is bounded, the Berry Esseen theorem implies that $W_n = W + R_n$, where each element of W_n is $O(1/\sqrt{n})$ and the upper bound is the same for each element. Also notice that each element of $R_n h$ is bounded in absolute value by a $O(1/\sqrt{n})$ term uniformly over all h for which $\|g\|_c = 1$ and $\|g\|_s \leq (2C/\varepsilon)$ because in this case $\sum_{j=1}^{\infty} |h_j| < \infty$. Therefore,

$$\begin{aligned}
P\left(\min_{g_J \in \tilde{\mathcal{H}}_J} \|\sqrt{n}(\hat{A} - A)h^1 + \sqrt{n}Ah^2\| \leq t\right) &= P\left(\min_{g_J \in \tilde{\mathcal{H}}_J} \|W_n h^1 + \sqrt{n}Ah^2\| \leq t\right) \\
&= P\left(\min_{g_J \in \tilde{\mathcal{H}}_J} \|Wh^1 + \sqrt{n}Ah^2 + O(1/\sqrt{n})\| \leq t\right)
\end{aligned}$$

where the $O(1/\sqrt{n})$ term does not depend on g_J. Also notice that

$$\|Wh^1+\sqrt{n}Ah^2\|-O(\sqrt{J/n}) \leq \|Wh^1+\sqrt{n}Ah^2+O(1/\sqrt{n})\| \leq \|Wh^1+\sqrt{n}Ah^2\|+O(\sqrt{J/n}).$$

Finally recall that $\|g_J^2\|_c \leq \lambda_n$ and since $\|g_J^1\|_c = \|g_J - g_J^2\|_c$ we get

$$1-\lambda_n \leq \|g_J^1\|_c = \|g_J - g_J^2\|_c \leq 1+\lambda_n,$$

which implies that $|\|g_J^1\|_c - 1| \leq \lambda_n$. Putting these results together implies that for some sequence $d_n \to 0$

$$\begin{aligned} P\left(\min_{g_J\in\tilde{\mathcal{H}}_J} \|\sqrt{n}(\hat{A}-A)h^1+\sqrt{n}Ah^2\| \leq t\right) &= P\left(\min_{g_J\in\mathcal{H}_J} \|Wh^1+\sqrt{n}Ah^2\| \leq t+d_n\right) \\ &= P\left(\min_{g_J\in\mathcal{H}_J} \|Wh^1+\sqrt{n}Ah^2\| \leq t\right)+o(1) \end{aligned}$$

where the last equality follows from Assumption 10. Hence, we can conclude that

$$\sup_{t\geq 0}\left|P\left(n\hat{T}\leq t\right) - P\left(\min_{g_J\in\mathcal{H}_J}\|Wh^1+\sqrt{n}Ah^2\|\leq t\right)\right| = o(1).$$

A.4 Proof of Theorem 4

Recall that

$$T^* = \min_{g_J\in\mathcal{S}: g_J=g_J^1+g_J^2, \|g_J^1\|_c=1, \|g_J^2\|_c\leq\lambda_n, \|g_J^1\|_s+\|g_J^2\|_s\leq(2C/\varepsilon)} \|(A^*-\hat{A})h^1+\hat{A}h^2\|^2 + \frac{\mu_n}{n}\|\hat{A}h^1\|^2,$$

where

$$\frac{\mu_n}{n}\to 0 \quad \text{and} \quad \frac{\sqrt{n}}{\mu_n}\to 0.$$

The constraint $\|g_2\|_c \leq \lambda_n$ is critical here because it guarantees that almost surely

$$\begin{aligned} \|\sqrt{n}(A^*-\hat{A})h^1+\sqrt{n}\hat{A}h^2\| &= \|\sqrt{n}(A^*-\hat{A})h^1+\sqrt{n}(\hat{A}-A)h^2+\sqrt{n}Ah^2\| \\ &= \|\sqrt{n}(A^*-\hat{A})h^1+\sqrt{n}Ah^2\|+o(1) \end{aligned}$$

uniformly over all relevant g_J because as shown in the proof of Theorem 3,

$$\|\sqrt{n}(\hat{A}-A)h^2\|^2 \leq \bar{C}\ln(\ln(n))J^2\lambda_J^2 \to 0.$$

It follows that nT^* is almost surely equal to

$$\min_{g_J\in\mathcal{S}: g_J=g_J^1+g_J^2, \|g_J^1\|_c=1, \|g_J^2\|_c\leq\lambda_n, \|g_J^1\|_s+\|g_J^2\|_s\leq(2C/\varepsilon)} \|\sqrt{n}(A^*-\hat{A})h^1+\sqrt{n}Ah^2\|^2 + \mu_n\|\hat{A}h^1\|^2 + o(1).$$

Now take any feasible g_J^1 such that $S_0(g^1) = 0$ and let $g_J^2 = 0$. Then for some constant $\bar{C}$

$$\begin{aligned}\|\sqrt{n}(A^* - \hat{A})h^1 + \sqrt{n}Ah^2\|^2 + \mu_n\|\hat{A}h^1\|^2 &\leq \bar{C}J^2\ln(\ln(n)) + \frac{\mu_n}{n}(\bar{C}J^2\ln(\ln(n))) + \frac{\mu_n}{n}n\delta_n \\ &\leq 3\bar{C}J^2\ln(\ln(n)).\end{aligned}$$

Moreover, for g_J^1 with $\|g_J^1\|_s \leq (2C/\varepsilon)$,

$$\sqrt{\mu_n}\|\hat{A}h^1\| = \sqrt{\frac{\mu_n}{n}}\|\sqrt{n}(\hat{A} - A)h^1 + \sqrt{n}Ah^1\| \geq \sqrt{\mu_n}\|Ah^1\| - \sqrt{\frac{\mu_n}{n}}\sqrt{\bar{C}J^2\ln(\ln(n))}.$$

Since any solution has to satisfy

$$\sqrt{\mu_n}\|\hat{A}h^1\| \leq \sqrt{3\bar{C}J^2\ln(\ln(n))}$$

it has to hold that

$$\|Ah^1\| \leq \sqrt{\frac{4\bar{C}J^2\ln(\ln(n))}{\mu_n}}.$$

Now let $g_J^1 = \bar{g}^1 + \bar{g}^2 = \bar{g}_J^1 + \bar{g}_J^2$, where $S_0(\bar{g}^1) = 0$ and $\bar{g}^2 \in \mathcal{N}^\perp$. It follows that $\|g_J^1\|_s = \|\bar{g}^1\|_s + \|\bar{g}^2\|_s$. Hence if $\|\bar{g}^1\|_s + \|\bar{g}^2\|_s \leq (2C/\varepsilon)$, then by Assumption 5 also $\|\bar{g}_J^1\|_s + \|\bar{g}_J^2\|_s \leq (2C/\varepsilon)$. Then

$$\sqrt{\frac{4\bar{C}J^2\ln(\ln(n))}{\mu_n}} \geq \|Ah^1\| \geq \|A\bar{h}^2\| - \|A\bar{h}^1\|.$$

By Assumption 8, $\sqrt{n}\|A\bar{h}^1\| \to 0$ and thus for n large enough

$$\|A\bar{h}^1\| \leq \sqrt{\frac{1}{\mu_n}}\sqrt{\frac{\mu_n}{n}}\sqrt{n}\|A\bar{h}^1\| \leq \sqrt{\frac{2}{\mu_n}}.$$

It follows that

$$\sqrt{\frac{5\bar{C}J^2\ln(\ln(n))}{\mu_n}} \geq \|A\bar{h}^2\|$$

and since $\sqrt{n}/\mu_n \to 0$,

$$\frac{5\bar{C}J^2\ln(\ln(n))}{\sqrt{n}} \geq \|A\bar{h}^2\|^2.$$

But then Assumption 9 implies that $\|\bar{g}_J^2\|_c \leq \lambda_n$ where $\lambda_n = \frac{\bar{C}}{J\sqrt{\ln(n)}}$ and the solution will be the minimizer of

$$\|\sqrt{n}(A^* - \hat{A})\left(\bar{h}^1 + \bar{h}^2\right) + \sqrt{n}Ah^2\|^2 + \mu_n\|\hat{A}(\bar{h}^1 + \bar{h}^2)\|^2$$

over the set $g_J \in \mathcal{S}$ such that $g_J = \bar{g}_J^1 + \bar{g}_J^2 + g_J^2$ with $\bar{g}^1 \in \mathcal{N}$ and $\bar{g}^2 \in \mathcal{N}^\perp$ and the constraints

$$\|\bar{g}_J^1\|_c = 1 + c_n, \|\bar{g}_J^2\|_c \leq \lambda_n, \|g_J^2\|_c \leq \lambda_n, \|\bar{g}_J^1\|_s + \|\bar{g}_J^2\|_s + \|g_J^2\|_s \leq (2C/\varepsilon),$$

where $|c_n| \leq \lambda_n$. But then $\bar{g}_J^2$ affects the objective function asymptotically only through the penalty because almost surely

$$\|\sqrt{n}(A^* - \hat{A})\bar{h}^2\|^2 \to 0.$$

Furthermore, since almost surely

$$\mu_n \|\hat{A}\bar{h}^1\|^2 \to 0$$

if $\bar{g}_J^2$ affects the penalty asymptotically it will increase the objective function. Moreover, a nonzero $\bar{g}_J^2$ reduces the constraint set. Thus, it will be optimal to set $\bar{g}_J^2 = 0$. It follows that $\sqrt{nT^*}$ is asymptotically equivalent to

$$\min_{g_J \in \mathcal{S}: g_J = \bar{g}_J^1 + g_J^2, \bar{g}^1 \in \mathcal{N}, \|\bar{g}_J^1\|_c = 1 + c_n, \|g_J^2\|_c \leq \lambda_n, \|\bar{g}_J^1\|_s + \|g_J^2\|_s \leq (2C/\varepsilon)} \|\sqrt{n}(A^* - \hat{A})\bar{h}^1 + \sqrt{n}Ah^2\| + o(1).$$

Next write $g_J^2 = \tilde{g}^1 + \tilde{g}^2 = \tilde{g}_J^1 + \tilde{g}_J^2$, where $\tilde{g}^1 \in \mathcal{N}$ and $\tilde{g}^2 \in \mathcal{N}^\perp$. Also here $\tilde{g}^1$ does not affect the objective but does reduce the constraint set if $\|\tilde{g}_J^1\|_s > 0$. Therefore, setting $\tilde{g}_J^1 = 0$ will be optimal. Hence $\sqrt{nT^*}$ is asymptotically equivalent to

$$\min_{g_J \in \mathcal{S}: g_J = \bar{g}_J^1 + \tilde{g}_J^2, \bar{g}^1 \in \mathcal{N}, \tilde{g}^2 \in \mathcal{N}^\perp, \|\bar{g}_J^1\|_c = 1 + c_n, \|\bar{g}_J^1\|_s + \|\tilde{g}_J^2\|_s \leq (2C/\varepsilon)} \|\sqrt{n}(A^* - \hat{A})\bar{h}^1 + \sqrt{n}A\tilde{h}^2\|$$

or, just as in the proof of Theorem 3, almost surely

$$\sqrt{n\hat{T}} = \min_{g_J \in \mathcal{H}_J} \|\sqrt{n}(A^* - \hat{A})h^1 + \sqrt{n}Ah^2\| + o(1).$$

Now the constraint $\|g_J^2\|_c \leq \lambda_n$ will also hold even without imposing it.

Then the same arguments as before show that

$$\sup_{t \geq 0} \left| P^*\left(nT^* \leq t\right) - P^*\left(\min_{g_J \in \mathcal{H}_J} \|W^* g_J^1 + \sqrt{n}Ag_J^2\|^2 \leq t\right) \right| = o(1),$$

where W^* is normally distributed. It follows that uniformly over $t \geq 0$

$$\begin{aligned} &\left| P^*\left(nT^* \leq t\right) - P\left(n\hat{T} \leq t\right) \right| \\ &\quad = \left| P^*\left(\min_{g_J \in \mathcal{H}_J} \|W^* h^1 + \sqrt{n}Ah^2\|^2 \leq t\right) - P\left(\min_{g_J \in \mathcal{H}_J} \|Wh^1 + \sqrt{n}Ah^2\|^2 \leq t\right) \right| + o(1). \end{aligned}$$

Finally notice that W^*h^1 and Wh^1 only differ by their covariance which converges at rate $J/\sqrt{n}$. More precisely denote by $\Sigma(h^1)$ the covariance matrix of Wh^1 and let $\Sigma^{1/2}(h^1)$ be such that $\Sigma^{1/2}(h^1)\Sigma^{1/2}(h^1) = \Sigma(h^1)$. Denote by $\hat{\Sigma}^{1/2}(h^1)$ the corresponding matrix of W^*h^1. Notice that

$$\left| \hat{\Sigma}^{1/2}(h^1) - \Sigma^{1/2}(h^1) \right| \leq \bar{C}\sqrt{\frac{\ln(\ln(n))}{n}}$$

where the inequality is understood element by element. Let V be a normally distributed $J \times 1$ vector with identity covariance matrix. It now follows that

$$\begin{aligned}
&P^*\left(\min_{g_J \in \mathcal{H}_J} \|W^* h^1 + \sqrt{n} A h^2\| \leq t\right) \\
&\quad = P^*\left(\min_{g_J \in \mathcal{H}_J} \|\hat{\Sigma}^{1/2}(h^1) V + \sqrt{n} A h^2\| \leq t\right) \\
&\quad = P^*\left(\min_{g_J \in \mathcal{H}_J} \|\Sigma^{1/2}(h^1) V + (\hat{\Sigma}^{1/2}(h^1) - \Sigma^{1/2}(h^1)) V + \sqrt{n} A h^2\| \leq t\right) \\
&\quad = P^*\left(\min_{g_J \in \mathcal{H}_J} \|\Sigma^{1/2}(h^1) V + \sqrt{n} A h^2\| \leq t + O(\sqrt{J \ln(\ln(n))/n})\right).
\end{aligned}$$

Now by Assumption 10 we get

$$P^*\left(\min_{g_J \in \mathcal{H}_J} \|W^* h^1 + \sqrt{n} A h^2\| \leq t\right) \;=\; P^*\left(\min_{g_J \in \mathcal{H}_J} \|\Sigma^{1/2}(h^1) V + \sqrt{n} A h^2\| \leq t\right) + o(1)$$

uniformly over $t \geq 0$. Therefore

$$\left|\alpha - P(n\hat{T} \geq c_\alpha^*)\right| \;=\; \left|P^*(nT^* \geq c_\alpha^*) - P(n\hat{T} \geq c_\alpha^*)\right| = o(1).$$

The remaining part is to show that the test is consistent against a fixed alternative. Under any alternative

$$\|\hat{A} h^1\| \geq \|A h^1\| - \|(\hat{A} - A) h^1\| = \|A h^1\| + o(1).$$

Moreover,

$$\begin{aligned}
\|Ah\|^2 &= \int \left(\int f_J(x,z) g_J(x,z)\right)^2 dz \\
&= \int \left(\int ((f_J(x,z) - f_{XZ}(x,z)) g_J(x) + f_{XZ}(x,z) g_J(x))\, dx\right)^2 dz \\
&\geq \frac{3}{4} \int \left(\int f_{XZ}(x,z) g_J(x) dx\right)^2 dz - 3 \int \left(\int (f_J(x,z) - f(x,z)) g_J(x) dx\right)^2 dz \\
&\geq \frac{3}{4} \inf_{g \in \bar{\mathcal{G}}: \|g\|_c \geq 1} \int \left(\int f(x,z) g(x) dx\right)^2 dz - 3\delta_n.
\end{aligned}$$

But under any alternative

$$\frac{3}{4} \inf_{g \in \bar{\mathcal{G}}: \|g\|_c \geq 1} \int \left(\int f(x,z) g(x)\right)^2 dz > 0$$

and hence $n\hat{T}$ diverges at rate n under any alternative.

Similarly, if $h^2 = 0$, we get

$$\|(A^* - \hat{A})h^1 + \hat{A}h^2\|^2 + \frac{\mu_n}{n}\|\hat{A}h^1\|^2 \geq \bar{C}\left(\frac{J^2\ln(\ln(n))}{n} + \frac{\mu_n}{n}\right)$$

and thus nT^* diverges at most at rate $\max\{J^2\ln(\ln(n)), \mu_n\}$ under any alternative. But since $\frac{n}{\mu_n} \to \infty$ and $\frac{n}{J^2\ln(\ln(n))} \to \infty$ it follows that under any fixed alternative

$$P(n\hat{T} \geq c_\alpha^*) \to 1.$$

B Examples satisfying testability and assumptions

I first illustrate why the failure of a restricted version of completeness, or equivalently weak instruments, is testable using a class of densities. In then use these densities to illustrate the technical assumptions.

B.1 Example of density functions

Assumption 1 implies that the density is square integrable. Therefore, we can write

$$f_{XZ}(x,z) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}\phi_j(x)\phi_k(z),$$

where $a_{jk} = \int\int f_{XZ}(x,z)\phi_j(x)\phi_k(z)dxdz$. Moreover, for any $g \in L^2[0,1]$, we can write

$$g(x) = \sum_{k=1}^{\infty} h_k\phi_k(x),$$

where $h_k = \int g(x)\phi_k(x)dx$. To obtain a tractable example, suppose that $a_{jk} = 0$ if $j \neq k$. Then

$$f_{XZ}(x,z) = \sum_{j=1}^{\infty} a_{jj}\phi_j(x)\phi_j(z)$$

and it can be shown that

$$S_0(g) \;=\; \sum_{j=1}^{\infty} h_j^2 a_{jj}^2.$$

Hence, if $a_{jj} = 0$ for some j, then $S_0(\phi_j) = 0$ and L^2-completeness fails. Contrarily, L^2-completeness holds if $a_{jj} \neq 0$ for all j. But since f_{XZ} is bounded

$$\int\int f_{XZ}(x,z)^2 dxdz = \sum_{j=1}^{\infty} a_{jj}^2 \leq C_d^2$$

and therefore, $a_{jj}^2 \to 0$ as $j \to \infty$. Now let $g = \phi_K$. Then $S_0(g) = a_{KK}^2 \to 0$ as $K \to \infty$. Hence, without the smoothness assumptions, taking the infimum of $S_0(g)$ over all functions in $L^2[0,1]$ with $\|g\|_c = 1$ yields 0, even if the distribution is complete.

Now for simplicity suppose that $\|g\|_s^2 = \int (g(x)^2 + g'(x)^2)\, dx$ and further suppose that $\int \phi_j'(x)\phi_k'(x)dx = 0$ if $j \neq k$, and $b_j \equiv \int \phi_j'(x)\phi_j'(x)dx \to \infty$ as $j \to \infty$. An example of such a basis would be trigonometric polynomials. While these are simplifying assumptions, used to obtain a tractable example, all commonly used basis function have the property that the derivatives get increasingly wiggly as j increases and therefore that $b_j \to \infty$. The constraint

$$\|g\|_s^2 = \int \left(g(x)^2 + g'(x)^2\right) dx \leq (2C/\varepsilon)^2$$

can now be written as

$$\sum_{j=1}^{\infty} h_j^2(1 + b_j) \leq (2C/\varepsilon)^2.$$

But since $b_j \to \infty$, it is not possible to have $h_K = 1$ for large K. In other words, we now cannot put all weight on very wiggly basis functions. As a consequence of this smoothness constraint, taking the infimum of $S_0(g)$ over all functions in $\bar{\mathcal{G}}$ with $\|g\|_c = 1$ does not yield 0 if the distribution is complete.

B.2 Explanation of assumptions

Assumption 1 is easy to interpret, while Assumptions 2 and 6 can be verified with popular parameter spaces, including Sobolev spaces, as discussed in Section 5.1. Assumptions 3 and 4 are discussed in Chen (2007) and hold for many popular basis functions as long as sufficient smoothness is imposed. Also notice that Assumptions 2 and 4 imply that for all functions g satisfying the smoothness restrictions it holds that

$$\sum_{j=J+1}^{\infty} h_j^2 = \int (g(x) - g_J(x))^2 dx \leq \frac{C_g^2}{C_0} J^{-2\bar{s}}.$$

It follows that

$$h_{J+1}^2 \leq \frac{C_g^2}{C_0} J^{-2\bar{s}} \leq \frac{C_g^2 2^{2\bar{s}}}{C_0} (J+1)^{-2\bar{s}}.$$

In other words, since the approximation error converges to 0 fast, the coefficients of the series approximation have to converge to 0 fast. We then also get

$$\sum_{j=1}^{\infty} |h_j| j^k \leq \frac{C_g 2^{\bar{s}}}{\sqrt{C_0}} \sum_{j=1}^{\infty} j^{-\bar{s}} j^k$$

Hence for any $k < \bar{s} - 1$, the sum on the right converges. In particular, it follows that $\sum_{j=1}^{\infty} |h_j| < \infty$ if $\bar{s} \geq 2$. This result is used in the proof of Theorems 3 and 4. Assumption 7 is a moment condition discussed in Section 4.1.

To obtain more intuition for the remaining assumptions suppose that

$$f_{XZ}(x,z) = \sum_{j=1}^{\infty} a_{jj}\phi_j(x)\phi_j(z).$$

Also suppose, similar as in Section B.1, that $\|g_J\|_c^2 = \sum_{j=1}^{J} h_j^2$, $\|g_J\|_s^2 = \sum_{j=1}^{J} h_j^2(1+b_j)$, where $b_j > 0$ and $b_j \to \infty$ as $j \to \infty$. For example, when $\|\cdot\|_s$ is the Sobolev norm, orthonormal trigonometric polynomials have this structure. Assumption 5, which could be relaxed at the expense of additional notation, says that if $\|g\|_s^2 = \sum_{j=1}^{\infty} h_j^2(1+b_j) \leq (2C/\varepsilon_n)^2$,

$$\sum_{j=1}^{J} h_j^2(1+b_j) \leq (2C/\varepsilon_n)^2 \quad \text{and} \quad \frac{\sum_{j=1}^{J} h_j^2(1+b_j)}{\sum_{j=1}^{J} h_j^2} \leq (2C/\varepsilon_n)^2.$$

The first inequality clearly holds because $b_j > 0$. Intuitively, the series truncation leaves out the very wiggly part of g and thus, the truncation has a smaller strong norm. The second inequality says that this is true even after normalizing by the consistency norm. To see why this is true rewrite

$$\frac{\sum_{j=1}^{J} h_j^2(1+b_j)}{\sum_{j=1}^{J} h_j^2} \leq \frac{\sum_{j=1}^{\infty} h_j^2(1+b_j) - \sum_{j=J+1}^{\infty} h_j^2(1+b_j)}{1 - \sum_{j=J+1}^{\infty} h_j^2} \leq \frac{(2C/\varepsilon_n)^2 - \sum_{j=J+1}^{\infty} h_j^2(1+b_j)}{1 - \sum_{j=J+1}^{\infty} h_j^2}.$$

Now notice that since $b_j \to \infty$ as $j \to \infty$, it holds that

$$\sum_{j=J+1}^{\infty} h_j^2(1+b_j) \geq \sum_{j=J+1}^{\infty} h_j^2(2C/\varepsilon_n)^2$$

if ε_n goes not 0 slow enough relative to J. Then

$$\frac{\sum_{j=1}^{J} h_j^2(1+b_j)}{\sum_{j=1}^{J} h_j^2} \leq \frac{(2C/\varepsilon_n)^2 - \sum_{j=J+1}^{\infty} h_j^2(1+b_j)}{1 - \sum_{j=J+1}^{\infty} h_j^2} \leq \frac{(2C/\varepsilon_n)^2 - (2C/\varepsilon_n)^2 \sum_{j=J+1}^{\infty} h_j^2}{1 - \sum_{j=J+1}^{\infty} h_j^2} = (2C/\varepsilon_n)^2,$$

which implies that Assumption 5 holds.

For the remaining assumptions suppose that ε_n is fixed. With the additional structure on the density and norms we get

$$\begin{aligned}
\int \left(\int g(x)(f_J(x,z) - f_{XZ}(x,z))dx \right)^2 dz &= \int \left(\int \sum_{j=1}^{\infty} h_j\phi_j(x) \sum_{j=J+1}^{\infty} a_{jj}\phi_j(x)\phi_j(z)dx \right)^2 dz \\
&= \int \left(\sum_{j=J+1}^{\infty} h_j a_{jj}\phi_j(z) \right)^2 dz \\
&= \sum_{j=J+1}^{\infty} h_j^2 a_{jj}^2
\end{aligned}$$

and thus

$$\delta_n = \max_{h_j^2, j=1,\ldots,\infty} \sum_{j=J+1}^{\infty} h_j^2 a_{jj}^2$$

Subject to

$$\sum_{j=1}^{\infty} h_j^2 \leq 1 \quad \text{and} \quad \sum_{j=1}^{\infty} h_j^2(1+b_j) \leq (2C/\varepsilon)^2.$$

It will be optimal to set $h_j = 0$ for $j \leq J$. An upper bound for δ_n can be obtained by ignoring the first constraint. In these examples, it will be the case that the more smoothness restrictions we impose on g_0, the larger $\bar{s}$ in Assumption 4, and the faster $b_j \to \infty$. Then for a given $\bar{s}$, the smoother the density, the smaller δ_n. If for example $b_j = j^{2s^*}$ and $a_{jj} = j^{-r}$, then h_j^2 has to converge faster than $j^{-2s^*-1-\eta}$ for some $\eta > 0$ and thus up to multiplicative constant

$$\delta_n \leq \sum_{j=J+1}^{\infty} j^{-2(s^*+r)-1-\eta} \leq (J+1)^{-2(s^*+r)} \sum_{j=J+1}^{\infty} j^{-1-\eta} \leq \bar{C} J^{-2(s^*+r)}$$

for some constant $\bar{C}$. Since δ_n goes to 0 at rate $J^{-2(s^*+r)}$, it holds that $n\delta_n \to 0$ as long as J is a polynomial of n and goes to ∞ fast enough. Similarly, if a_{jj} goes to 0 at an exponential rate, then δ_n goes to 0 at an exponential rate as well, and $n\delta_n \to 0$ holds even if J is a logarithmic function of n. Hence, Assumption 8 mainly says that J has to diverge fast enough relative to n for a given smoothness of g_0 and f_{XZ}.

Next notice that

$$\int \left(\int g_J(x) f_J(x,z) dx \right)^2 dz = \sum_{j=1}^{J} h_j^2 a_{jj}^2$$

To make more sense of Assumption 9, suppose that $a_{jj} = 0$ if and only if $j > \bar{J}$. Then $\mathcal{N}^{\perp}$ consists of all function g such that $g = \sum_{j=1}^{\bar{J}} h_j \phi_j$. For those functions and J large enough $\|g_J\|_c^2 = \sum_{j=1}^{\bar{J}} h_j^2$ and $\|g_J\|_s^2 = \sum_{j=1}^{\bar{J}} h_j^2(1+b_j)$. Suppose that $(1+b_{\bar{J}+1}) \leq (2C/\varepsilon)^2$. In this example, the density consists of low dimensional basis functions and some of the higher dimensional basis functions, which are in the null space of $S_0(\cdot)$, satisfy the smoothness restrictions. In this case, we want to solve

$$\min_{h_j^2, j=1,\ldots,\bar{J}} \sum_{j=1}^{\bar{J}} h_j^2 a_{jj}^2$$

subject to $\sum_{j=1}^{\bar{J}} h_j^2 \geq \frac{1}{J^2 \ln(n)}$ and $\sum_{j=1}^{\bar{J}} h_j^2(1+b_j) \leq (2C/\varepsilon)^2$. Assuming that a_{jj} is a decreasing sequence, the optimal solution is $h_{\bar{J}}^2 = \frac{1}{J^2 \ln(n)}$ in which case the minimum is $\frac{a_{\bar{J}\bar{J}}^2}{J^2 \ln(n)}$ and we need that

$$\frac{a_{\bar{J}\bar{J}}^2}{J^2 \ln(n)} \geq \frac{C_t J^2 \ln(n)}{\sqrt{n}}$$

or

$$\frac{a_{\bar{J}\bar{J}}^2\sqrt{n}}{C_t \ln(n)^2} \geq J^4.$$

This example highlights that the assumption mainly implies that J cannot diverge too fast relative to n.

Next suppose that $a_{jj} = 0$ if and only if $j = J^*$. Ignoring the smoothness restrictions leads to an optimal solution of $\frac{a_{JJ}}{J^2 \ln(n)}$ and we now require

$$\frac{a_{JJ}^2\sqrt{n}}{C_t \ln(n)^2} \geq J^4.$$

If $a_{JJ}^2 = J^{-2r}$ as before, then

$$\frac{\sqrt{n}}{C_t \ln(n)^2} \geq J^{4+2r}$$

which again says that J cannot increase too fast relative to n. Also recall that in this case

$$n\delta_n \leq \bar{C} n J^{-2(s^*+r)}.$$

Thus, for both Assumption 8 and 9 to be satisfied, we need that s^* is large enough relative to r. In that case, we can pick J such that both assumptions hold. Similar as other papers that deal with nonparametric IV estimation, these two assumptions link the smoothness of the density to the smoothness of g_0.

C Monte Carlo simulations for weak instruments

I use the same setup as in Section 7, but now I fix $\varepsilon = 0.05$ and I choose the critical values using the bootstrap procedure described in Section 4.2. I need to select two tuning parameters and I set $\lambda_n = 1/(J\sqrt{\ln(n)})$ and $\mu_n = n^{3/4}$. I use 1000 bootstrap samples. The nominal size of the test is $\alpha = 0.05$.

Table 3 shows the actual rejection probabilities when $n = 1000$ and $J = 3$ and when $n = 5000$ and $J = 4$ for the first setup. It can be shown that with this choice of ε, the instruments are weak with either $d_{22} = 0$, $d_{44} = 0$, or $d_{55} = 0$. The table also shows $\|\hat{g} - g_0\|_c$, the root mean squared error, averaged over the 1000 simulated data sets. When $n = 1000$ and $J = 3$, the rejection probability is close to 90% in the first three cases. Hence, the test cannot distinguish between the complete and the (very close to complete) incomplete distributions, but the RMSEs are also identical. When $n = 5000$ and $J = 4$, the rejection probability decreases when $d_{22} = 0$, it is 5.9% when $d_{44} = 0$, and above 80% with $d_{55} = 0$. So just as in Section 7, although the distributions are extremely close, the test is able to

distinguish them if n and J are large enough. Furthermore, the RMSE is now larger in case $d_{44} = 0$. Again, the test cannot distinguish between the complete distribution and the one with $d_{55} = 0$, but the RMSEs are also identical. The test can pick up this difference as well when n and J are larger. In particular, if $n = 10,000$ and $J = 5$, the rejection probability is 7.2% when $d_{55} = 0$.

Table 3: Rejection probabilities and RMSE with sequence of incomplete distributions

	Complete		$d_{55} = 0$		$d_{44} = 0$		$d_{22} = 0$	
	Reject	RMSE	Reject	RMSE	Reject	RMSE	Reject	RMSE
$n = 1000$, $J = 3$	0.906	0.165	0.901	0.166	0.903	0.165	0.100	0.370
$n = 5000$, $J = 4$	0.822	0.145	0.826	0.145	0.059	0.159	0.092	0.335

Table 4 shows the rejection probabilities and the RMSE for different choices of ρ, n, and J. As expected, the rejection probability increases with ρ and the RMSE decreases with ρ. Notice that the matrix A does not have full if $\rho = 0$. Therefore, we would expect the best finite sample properties of the test if J is small, because then we get great power properties while we can still control size. However, as seen in the first example, in general J has to increase as n increases in order to control size. Similar as in Section 7 J seems too large relative to the sample size because the power decreases significantly as n and J increase and even the RMSE increases when $\rho = 0.5$. The properties improve when $n = 20,000$ and $J = 4$. Then the RMSE with $\rho = 0.5$ is only 0.076 (and the rejection probability is 1) and if $\rho = 0.3$ the rejection probability goes up to 0.409 with a RMSE of 0.147. These results illustrate that both the properties of the estimator and the test can be sensitive to the choice of J.

Table 4: Rejection probabilities and RMSE with sequence of complete distributions

	$\rho = 0.5$		$\rho = 0.3$		$\rho = 0.1$		$\rho = 0$	
	Reject	RMSE	Reject	RMSE	Reject	RMSE	Reject	RMSE
$n = 1000$, $J = 3$	1.000	0.110	0.532	0.205	0.106	0.314	0.039	0.388
$n = 5000$, $J = 4$	0.993	0.119	0.178	0.180	0.031	0.236	0.041	0.362

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Andrews, D. W. K. (2011). Examples of L2-complete and boundedly-complete distributions. Working paper.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5).

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica 75*(6), 1613–1669.

Bugni, F. A., I. A. Canay, and X. Shi (2014). Inference for functions of partially identified parameters in moment inequality models. Working paper.

Canay, I. A., A. Santos, and A. M. Shaikh (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica 81*(6), 2535–2559.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 76, pp. 5550–5623. Elsevier.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014). Local identification of nonparametric and semiparametric models. *Econometrica 82*(2), 785–809.

Chen, X. and T. M. Christensen (2013). Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. Working paper.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*(1), 46–60.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica 80*(1), 277–321.

Chen, X. and D. Pouzo (2014). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica, forthcoming*.

Cragg, J. G. and S. G. Donald (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory 9*(2), 222–240.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

D'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory 27*(3), 460–471.

Florens, J.-P., J. Johannes, and S. Van Bellegem (2011). Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory 27*, 472–496.

Freyberger, J. (2012). Nonparametric panel data models with interactive fixed effects. Working paper.

Freyberger, J. and M. A. Masten (2014). Compactness of infinite dimensional parameter spaces. Working paper.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–90.

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics 33*(6), 1–27.

Han, S. (2014). Nonparametric estimation of triangular simultaneous equations models under weak identfication. Working paper.

Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica 79*(2), 347–394.

Horowitz, J. L. (2012). Specification testing in nonparametric instrumental variable estimation. *Journal of Econometrics 167*(2), 383–396.

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics 180*(2), 158–173.

Horowitz, J. L. and S. Lee (2012). Uniform confidence bands for functions estimated nonparametrically with instrumental variables. *Journal of Econometrics 168*(2), 175–188.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Hu, Y. and J.-L. Shiu (2012). Nonparametric identification using instrumental variables: sufficient conditions for completeness. Working paper.

Mattner, L. (1993). Some incomplete but boundedly complete location families. *The Annals of Statistics 21*(4), 2158–2162.

Montiel Olea, J. L. and C. Pflueger (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics 31*(3), 358–369.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Robin, J.-M. and R. J. Smith (2000). Tests of rank. *Econometric Theory 16*(2), 151–175.

Santos, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica 80*(1), 213–275.

Sasaki, Y. (2014). Heterogeneity and selection in dynamic panel data. Working paper.

Staiger, D. and J. H. Stock (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica 65*(3), 557–586.

Stock, J. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In D. Andrews and J. Stock (Eds.), *Identification and Inference for Econometric Models*. New York: Cambridge University Press.

Tao, J. (2014). Inference for point and partially identified semi-nonparametric conditional moment models. Working paper.