Optimal Linear Instrumental Variables Approximations*

Juan Carlos Escanciano†
Universidad Carlos III de Madrid

Wei Li ‡
Syracuse University

February 3th, 2020

Abstract

This paper studies the identification and estimation of the optimal linear approximation of a structural regression function. The parameter in the linear approximation is called the Optimal Linear Instrumental Variables Approximation (OLIVA). This paper shows that a necessary condition for standard inference on the OLIVA is also sufficient for the existence of an IV estimand in a linear model. The instrument in the IV estimand is unknown and may not be identified. A Two-Step IV (TSIV) estimator based on Tikhonov regularization is proposed, which can be implemented by standard regression routines. We establish the asymptotic normality of the TSIV estimator assuming neither completeness nor identification of the instrument. As an important application of our analysis, we robustify the classical Hausman test for exogeneity against misspecification of the linear structural model. We also discuss extensions to weighted least squares criteria. Monte Carlo simulations suggest an excellent finite sample performance for the proposed inferences. Finally, in an empirical application estimating the elasticity of intertemporal substitution (EIS) with US data, we obtain TSIV estimates that are much larger than their standard IV counterparts, with our robust Hausman test failing to reject the null hypothesis of exogeneity of real interest rates.

Keywords: Instrumental Variables; Nonparametric Identification; Hausman Test.

JEL classification: C26; C14; C21.

*We thank Andres Santos, Pedro Sant'Anna, Oliver Linton, an AE, two anonymous referees and seminar participants at Cambridge, LACEA-LAMES 2018, TSE, UCL and UC3M for useful comments.

†Department of Economics, Universidad Carlos III de Madrid, 15.2.19, Calle Madrid 126, Getafe, Madrid, 28903, Spain. E-mail: jescanci@econ.uc3m.es. Web Page: https://sites.google.com/view/juancarlosescanciano. Research funded by the Spanish Grant PGC2018-096732-B-I00.

‡Department of Mathematics, 215 Carnegie Building, Syracuse, New York 13244, USA. E-mail: wli169@syr.edu

1 Introduction

The Ordinary Least Squares (OLS) estimator has an appealing nonparametric interpretation—it provides the optimal linear approximation (in a mean-square error sense) to the true regression function. That is, the OLS estimand is a meaningful and easily interpretable parameter under misspecification of the linear model. Unfortunately, except in special circumstances (such as with random assignment), this parameter does not have a causal interpretation. Commonly used estimands based on Instrumental Variables (IV) do have a causal interpretation (see, e.g., Imbens and Angrist (1994)), but they do not share with OLS the appealing nonparametric interpretation (see, e.g., Imbens, Angrist and Graddy (2000)). The main goal of our paper is to fill this gap and to propose an IV estimand that has the same nonparametric interpretation as OLS, but under endogeneity.

The parameter of interest is thus the vector of slopes in the optimal linear approximation of the *structural* regression function. We call this parameter the Optimal Linear IV Approximation (OLIVA). We investigate regular identification of the OLIVA, i.e. identification with a finite efficiency bound, based on the results in Severini and Tripathi (2012). The main contribution of our paper is to show that a necessary condition for regular identification of the OLIVA is also sufficient for existence of an IV estimand in a linear structural regression. That is, we show that, under a minimal condition for standard inference on the OLIVA, it is possible to obtain an IV estimator for it.

The identification result is constructive and leads to a Two-Step IV (TSIV) estimation strategy. The necessary condition for regular identification is a conditional moment restriction that is used to estimate a suitable instrument in a first step. The second step is simply a standard linear IV estimator with the estimated instrument from the first step. The situation is somewhat analogous to optimal IV (see, e.g., Robinson (1976) and Newey (1990)), but more difficult due to the lack of identification of the first step and the first step problem being statistically harder than a nonparametric regression problem. To select an instrument among potentially many candidates, we use Tikhonov regularization, combined with a sieve approach to obtain a Penalized Sieve Minimum Distance (PSMD) first step estimator (cf. Chen and Pouzo (2012)). The instrument choice based on Tikhonov is statistically and empirically justified. Statistically, a Tikhonov instrument exhibits a certain sufficiency property explained below. Empirically, the resulting PSMD estimator can be computed with standard regression routines. The TSIV estimator is shown to be asymptotically normal and to perform favorably in simulations when compared to alternative estimators, being competitive with the oracle IV under linearity of the structural model, while robustifying it otherwise.

An important application of our approach is to a Hausman test for exogeneity that is robust to misspecification of the linear model. This robustness comes from our TSIV being nonparametrically comparable to OLS under exogeneity. The robust Hausman test is a standard t-test in an augmented regression that does not require any correction for standard errors for its validity, as we show below. Lochner and Moretti (2015) consider a different exogeneity test comparing the classical IV estimator with a weighted OLS estimator when the endogenous variable is discrete. In contrast, our test compares the standard OLS with our TSIV estimator–more in the spirit of the original Hausman (1978)'s exogeneity test–while allowing for general endogenous variables (continuous, discrete or mixed). Monte

Carlo simulations confirm the robustness of the proposed Hausman test, and the inability of the standard Hausman test to control the empirical size under misspecification of the linear model.

Our paper contributes to two different strands of the literature. The first strand is the nonparametric IV literature; see, e.g., Newey and Powell (2003), Ai and Chen (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Horowitz (2007), Horowitz (2011), Darolles, Fan, Florens and Renault (2011), Santos (2012), Chetverikov and Wilhem (2017), and Freyberger (2017), among others. Severini and Tripathi (2006, 2012) discuss identification and efficiency of linear functionals of the structural function without assuming completeness. Their results on regular identification are adapted to the OLIVA below. Santos (2011) establishes regular asymptotic normality for weighted integrals of the structural function in nonparametric IV, also allowing for lack of nonparametric identification of the structural function. Blundell and Horowitz (2007) develop a nonparametric test of exogeneity under the maintained assumption of nonparametric identification. The OLIVA functional was not considered in Severini and Tripathi (2006, 2012) or Santos (2011), and the semiparametric robust Hausman test complements the nonparametric test of Blundell and Horowitz (2007).

Our paper is also related to the Causal IV literature that interprets IV nonparametrically as a Local Average Treatment Effect (LATE); see Imbens and Angrist (1994). A forerunner of our paper is Abadie (2003). He defines the Complier Causal Response Function and its best linear approximation in the presence of covariates. He also develops two-step inference for the linear approximation coefficients when the endogenous variable is binary. Within this binary case, we show that the OLIVA's slope parameter is the IV estimand resulting from using the propensity score as instrument, a recommended IV estimator in the literature (see Imbens and Angrist (1994) and pg. 623 in Wooldridge (2002)). Our asymptotic results for the binary endogenous case can thus be viewed as extensions of existing methods (such as, e.g., Theorem 3 in Imbens and Angrist (1994)) to a nonparametrically estimated propensity score.

The main theoretical contributions of this paper are thus the interpretation of the regular identification of the OLIVA as existence of an IV estimand, the asymptotic normality of a TSIV estimator, and the robust Hausman test. The identification, estimation and exogeneity test of this paper are all robust to the lack of the identification of the structural function (i.e. lack of completeness) and lack of identification of the first step instrument. Furthermore, the proposed methods are also robust to misspecification of linear model, sharing the nonparametric robustness of OLS, but in a setting with endogenous regressors.

We illustrate the utility of our methods with an empirical application estimating the elasticity of intertemporal substitution (EIS) with quarterly US data, revisiting previous work by Yogo (2004). If the structural relationship between consumption growth and interest rates is linear, then the TSIV and standard IV estimands should be the same. In contrast, we obtain a TSIV estimate much larger than the standard IV estimate, with a similar level of precision, thereby suggesting that nonlinearities matter in this application. The TSIV and OLS estimates are rather close, and the robust Hausman test fails to reject the null hypothesis of exogeneity of real interest rates.

The rest of the paper is organized as follows. Section 2 defines formally the parameter of interest and its regular identification. Section 3 proposes a PSMD first step and establishes the asymptotic normality of the TSIV. Section 4 derives the asymptotic properties of the robust Hausman test for exogeneity. The finite sample performance of the TSIV and the robust Hausman test is investigated in Section 5. Finally, Section 6 reports the results of our empirical application to the EIS. Appendix A presents notation, assumptions and some preliminary results that are needed for the main proofs in Appendix B. A Supplemental Appendix contains further simulation results, including extensive sensitivity analysis.

2 Optimal Linear Instrumental Variables Approximations

2.1 Nonparametric Interpretation

Let the dependent variable Y be related to the $p-$dimensional vector X through the equation

$$Y = g(X) + \varepsilon, \tag{1}$$

where $E[\varepsilon|\, Z] = 0$ almost surely (a.s), for a $q-$dimensional vector of instruments Z.

The OLIVA parameter β solves, for g satisfying (1),

$$\beta = \arg\min_{\gamma \in \mathbb{R}^p} E[\left(g(X) - \gamma' X\right)^2], \tag{2}$$

where henceforth A' denotes the transpose of A. Note that X may (and in general, will) contain an intercept. For extensions to weighted least squares versions of (2) see Section 3.5.

If $E[XX']$ is positive definite, then

$$\beta \equiv \beta(g) = E[XX']^{-1} E[Xg(X)]. \tag{3}$$

When X is exogenous, i.e. $E[\varepsilon|\, X] = 0$ a.s., the function $g(\cdot)$ is the regression function $E[Y|\, X = \cdot]$ and β is identified and consistently estimated by OLS under mild conditions. In many economic applications, however, X is endogenous, i.e. $E[\varepsilon|\, X] \neq 0$, and identification and estimation of (3) becomes a more difficult issue than in the exogenous case, albeit less difficult than identification and estimation of the structural function g in (1). Of course, if g is linear $g(x) = \gamma_0' x$, then $\beta = \gamma_0$.

We first investigate regular identification of β in (1)-(2). The terminology of regular identification is proposed in Khan and Tamer (2010), and refers to identification with a finite efficiency bound. Regular identification of a parameter is desirable because it means possibility of standard inference (see Chamberlain (1986)). It will be shown below that a necessary condition for regular identification of β is

$$E[h(Z)|\, X] = X \text{ a.s}, \tag{4}$$

for an square integrable $h(\cdot)$; see Lemma 2.1, which builds on Severini and Tripathi (2012). We show that condition (4) is also sufficient for existence of an IV estimand identifying β. That is, we show that (4) implies that β is identified from a linear structural regression

$$Y = X'\beta + U, \qquad E[Uh(Z)] = 0. \tag{5}$$

The IV estimand uses the unknown, possibly not unique, transformation $h(\cdot)$ of Z as instruments. We propose below a Two-Step IV (TSIV) estimator that first estimates the instruments from (4) and then applies IV with the estimated instruments. The proposed IV estimator has the same nonparametric interpretation as OLS, but under endogeneity.

If the nonparametric structural function g is identified, then β is of course identified (from 3). Conditions for point identification and consistent estimation of g are given in the references above on the nonparametric IV literature. Likewise, asymptotic normality for continuous functionals of a point-identified g has been analyzed in Ai and Chen (2003), Ai and Chen (2007), Carrasco, Florens and Renault (2006), Carrasco, Florens and Renault (2014), Chen and Pouzo (2015) and Breunig and Johannes (2016), among others.

Nonparametric identification of g is, however, not necessary for identification of the OLIVA; see Severini and Tripathi (2006, 2012). It is indeed desirable to obtain identification of β without requiring completeness assumptions, which are known to be impossible to test (cf. Canay, Santos and Shaikh (2013)). In this paper we focus on regular identification of the OLIVA without assuming completeness, i.e. without assuming identification of g.

Section 2.2 below shows the necessity of the conditional moment restriction (4) for regular identification of the OLIVA. When regular identification of the OLIVA does not hold, but the OLIVA is identified, we expect our estimator to provide a good approximation to the OLIVA. This follows because (i) under irregular identification of the OLIVA, the first step instrument approximately solves the first step conditional moment, and (ii) small errors in the first step equation lead to small errors in the second step limit.[1] Inference under irregular identification is known to be less stable, see Chamberlain (1986), and it is beyond the scope of this paper. See Babii and Florens (2018) for recent advances in this direction, and Escanciano and Li (2013) for partial identification results.

2.2 Regular Identification of the OLIVA

We observe a random vector $W = (Y, X', Z')'$ satisfying (1), or equivalently,

$$r(z) := E[Y|\, Z = z] = E[g(X)|\, Z = z] := T^* g(z), \tag{6}$$

where T^* denotes the adjoint operator of the operator T, with $Th(x) = E[h(Z)|\, X = x]$ a.s. Let $\mathcal{G}$ denote the parameter space for g. Assume $g \in \mathcal{G} \subseteq L_2(X)$ and $r \in L_2(Z)$, where henceforth, for a generic random variable V, $L_2(V)$ denotes the space of (measurable) square integrable functions of V, i.e. $f \in L_2(V)$ if $\|f\|^2 := E\left[|f(V)|^2\right] < \infty$, and where $|A| = trace\,(A'A)^{1/2}$ is the Euclidean norm.[2]

The next result, which follows from an application of Lemma 4.1 in Severini and Tripathi (2012), provides a necessary condition for regular identification of the OLIVA. Define $g_0 := \arg\min_{g:r=T^*g} \|g\|$, and note that correct specification of the model guarantees that g_0 is uniquely defined; see Engl, Hanke and Neubauer (1996). Define $\xi = Y - g_0(X)$, $\Omega(z) = E[\xi^2 |\, Z = z]$, and let $\mathcal{S}_Z$ denote the support of Z. For future reference, define the range of the operator T as $\mathcal{R}(T) := \{f \in L_2(X) : \exists s \in L_2(Z), Ts = f\}$, and for a subspace V, let $V^\perp$ and $\overline{V}$ denote, respectively, its orthogonal complement and its closure.

[1] We thank Andres Santos for making this point to us.

[2] When f is vector-valued, by $f(V) \in L_2(V)$ we mean that its components are all in $L_2(V)$.

Assumption 1: (6) holds, $g \in \mathcal{G} \subseteq L_2(X)$, $r \in L_2(Z)$, and $E[XX']$ is finite and positive definite.

Assumption 2: $0 < \inf_{z \in \mathcal{S}_Z} \Omega(z) \leq \sup_{z \in \mathcal{S}_Z} \Omega(z) < \infty$ and T is compact.

Assumption 3: There exists $h(\cdot) \in L_2(Z)$ such that (4) holds.

Lemma 2.1 *Let Assumptions 1-2 hold. If β is regularly identified, then Assumption 3 must hold.*

The proof of Lemma 2.1 and other results in the text are gathered in Appendix B. Assumptions 1 and 2 are taken from Severini and Tripathi (2012) and are standard in the literature. Given the necessity of Assumption 3 and its importance for our results it is useful to provide some discussion on it. The first observation is that although sufficient conditions for Assumption 3 to hold can be obtained for parametric settings, such as those in the Monte Carlo section, it is hard to give primitive sufficient conditions in nonparametric settings. The second observation is that Assumption 3 may hold when L_2-completeness of X given Z fails and g is thus not identified (see Newey and Powell (2003) for discussion of L_2-completeness). To illustrate this point, we consider the empirically relevant case where X is continuous and Z is discrete. It is well known that in this case g is not identified. In contrast, Assumption 3 may hold and, importantly, it is testable. To see this, let $\{z_1, ..., z_J\}$ denote the support of Z, with $J < \infty$, and note that any function h can be identified with a $J \times p$ matrix through the representation

$$h(z) = \sum_{j=1}^{J} h(z_j) 1(z = z_j),$$

where $1(A)$ is the indicator function of the event A. Assumption 3 is then simply the conditional moment restriction with a finite number of parameters $\theta = (h(z_1), ..., h(z_J)) \in \mathbb{R}^{p \times J}$ given by

$$E[\theta \mathbf{1} - X | X] = 0 \text{ a.s.} \tag{7}$$

where $\mathbf{1} = (1(Z = z_1), ..., 1(Z = z_J))'$. To deal with the potential lack of identification of h (i.e. of θ) we use the minimum norm estimator described below, which is consistent for a population analog $h_0(Z) = \theta_0 \mathbf{1}$. Furthermore, the estimator of θ_0 can be shown to be asymptotically normal. Thus, standard tools from nonparametric regression testing can be used to test for (7); see, e.g., Bierens (1982) and Escanciano (2006). Whether the nonparametric conditional moment restriction in Assumption 3 is testable more generally (i.e. with continuous Z) is a delicate issue, see Chen and Santos (2018), and it will be investigated elsewhere.

When Assumption 3 does not hold two possibilities may arise: (i) β is identified, but it has infinite efficiency bound, and (ii) β is not identified. When β is identified and Assumption 3 fails, X belongs to the boundary of the range of T (i.e. $X \in \overline{\mathcal{R}(T)} \setminus \mathcal{R}(T)$, see Severini and Tripathi (2012)), and thus our IV estimand can be made arbitrarily close to β. As we explain below in Remark 3.1, even when Assumption 3 does not hold, our estimator has a well-defined IV estimand as its limit, provided a mild condition is satisfied.

The main observation of this paper is that the necessary condition for regular identification of β is also sufficient for existence of an IV estimand. This follows because by the law of iterated expectations, Assumption 3 and $E[\varepsilon|\, Z] = 0$ a.s.,

$$\begin{aligned}\beta &= E[XX']^{-1}E[Xg(X)] \\ &= E[E[h(Z)|\, X]X']^{-1}E[E[h(Z)|\, X]g(X)] \\ &= E[h(Z)X']^{-1}E[h(Z)Y], \end{aligned} \tag{8}$$

which is the IV estimand using $h(Z)$ as instruments for X. We note that to obtain this IV representation in (8) a weaker exogeneity than $E[\varepsilon|\, Z] = 0$ suffices, namely $E[\varepsilon h(Z)] = 0$. We maintain the "strict" exogeneity $E[\varepsilon|\, Z] = 0$ because it is often used in the literature and simplifies some of our subsequent asymptotic results, although see Remark 3.2. The following result summarizes this finding and shows that, although there are potentially many solutions to (4), the corresponding β is unique.

Proposition 2.2 *Let Assumptions 1-3 hold. Then, β is regularly identified as (8).*

Remark 2.1 *By (4), $E[h(Z)X'] = E[XX']$. Thus, non-singularity of $E[h(Z)X']$ follows from that of $E[XX']$. Thus, the strength of the instruments $h(Z)$ is measured by the level of multicollinearity in X.*

3 Two-Step Instrumental Variables Estimation

Proposition 2.2 suggests a TSIV estimation method where, first, an h is estimated from (4) and then, an IV estimator is considered using the estimated h as instrument. To describe the estimator, let $\{W_i \equiv (Y_i, X_i', Z_i')'\}_{i=1}^n$ be an independent and identically distributed (*iid*) sample of size n satisfying (1). The TSIV estimator follows the steps:

Step 1. Estimate an instrument $h(Z)$ satisfying $E[h(Z)|\, X] = X$ a.s., say $\hat{h}_n$, as defined in (13) below.

Step 2. Run linear IV using instruments $\hat{h}_n(Z)$ for X in $Y = X'\beta + U$, i.e.

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{h}_n(Z_i)X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{h}_n(Z_i)Y_i\right), \tag{9}$$

where $\hat{h}_n$ is the first step estimator given in Step 1.

For ease of exposition, we consider first the case where X and Z have no overlapping components (i.e. no included exogenous or controls) and both are continuous. We also analyze below the case of control variables and discrete variables.

3.1 First-Step Estimation

As argued in pg. 130 of Santos (2012) identification of h in (4) is problematic, as in most instances instruments posses a variation that is unrelated to the endogenous regressor (i.e., there exists a function

$\psi(z)$ such that $E[\psi(Z)|\, X] = 0$ a.s.). To deal with the problem of lack of uniqueness of h, we consider a Tikhonov-type estimator. This approach is commonly used in the literature estimating g, see Hall and Horowitz (2005), Carrasco, Florens and Renault (2006), Florens, Johannes and Van Bellegem (2011), Chen and Pouzo (2012) and Gagliardini and Scaillet (2012), among others. Chen and Pouzo (2012) propose a PSMD estimator of g and show the L_2-consistency of a solution identified via a strict convex penalty. These authors also obtain rates in Banach norms under point identification. Our first-step estimator $\hat{h}_n$ is a PSMD estimator of the form considered in Chen and Pouzo (2012) when identification is achieved with an L_2-penalty. As it turns out, the Tikhonov-type or L_2-penalty estimator is well motivated in our setting, as we explain below. It implies that our instrument satisfies a certain sufficiency property.

Defining $m(X;h) := E[h(Z) - X|X]$, we estimate the unique h_0 satisfying $h_0 = \lim_{\lambda\downarrow 0} h_0(\lambda)$, where

$$h_0(\lambda) = \arg\min\{||m(\cdot;h)||^2 + \lambda||h||^2 : h \in L_2(Z)\},$$

and $\lambda > 0$. Assumption 3 guarantees the existence and uniqueness of h_0, see Engl, Hanke and Neubauer (1996). The sufficiency property mentioned above is that for any distinct solution h_1 of (4), $h_1 \neq h_0$, it holds that in the first stage regression

$$X = c_0 + \alpha_0 h_0(Z) + \alpha_1 h_1(Z) + V, \qquad Cov(V, h_j(Z)) = 0,\ j = 0, 1, \tag{10}$$

α_1 must be zero, as shown in the next result. We note that V is simply a least squares (i.e. reduced form) error term.

Proposition 3.1 *Let $h_1 \neq h_0$ be another solution of (4). Then, $\alpha_1 = 0$ in (10).*

This result states that after controlling for $h_0(Z)$ in the first stage regression, any other distinct solution to (4) is irrelevant in the first stage. It is in this precise sense that we say $h_0(Z)$ is sufficient. We note, however, that this property does not imply that h_0 is better than any other solution to (4) in terms of leading to a more efficient estimation of β. For efficiency considerations see Severini and Tripathi (2012).

Remark 3.1 *The minimum norm h_0 is well-defined under a weaker condition than Assumption 3. From Engl, Hanke and Neubauer (1996), for existence of h_0 it suffices that X belongs to the dense set $\mathcal{R}(T) + \mathcal{R}(T)^{\perp}$. In particular, this assumption holds when X is a square integrable continuous variable and Z is discrete (since $\mathcal{R}(T) + \mathcal{R}(T)^{\perp} \equiv L_2(X)$ in this case). Thus, under mild conditions, $\hat{\beta}$ has a probabilistic limit satisfying (5).*

Having motivated the Tikhonov-type instrument, we introduce now its PSMD estimator. Let $E_n[g(W)]$ denote the sample mean operator, i.e. $E_n[g(W)] = n^{-1}\sum_i^n g(W_i)$, let $||g||_n = \left(E_n[|g(W)|^2]\right)^{1/2}$ be the empirical L_2 norm, and let $\hat{E}[h(Z)|\, X]$ be a series-based estimator for the conditional mean $E[h(Z)|\, X]$, which is given as follows. Consider a vector of approximating functions

$$p^{K_n}(x) = (p_1(x), ..., p_{K_n}(x))',$$

having the property that a linear combination can approximate $E[h(Z)|\, X = x]$ well. Then,

$$\hat{E}[h(Z)|\, X = x] = p^{K_n \prime}(x)(P'P)^{-1} \sum_{i=1}^{n} p^{K_n}(X_i)h(Z_i),$$

where $P = [p^{K_n}(X_1), ..., p^{K_n}(X_n)]'$ and $K_n \to \infty$ as $n \to \infty$.

Let $\mathcal{H} \subseteq L_2(Z)$ denote the parameter space for h. Then, define the estimator

$$\hat{h}_n := \arg\min\{||\hat{m}(X; h)||_n^2 + \lambda_n ||h||_n^2 : h \in \mathcal{H}_n\}, \tag{11}$$

where $\mathcal{H}_n \subset \mathcal{H} \subseteq L_2(Z)$ is a linear sieve parameter space whose complexity grows with sample size, $\hat{m}(X_i; h) = \hat{E}(h(Z) - X|X_i)$, and λ_n is a sequence of positive numbers satisfying that $\lambda_n \downarrow 0$ as $n \uparrow \infty$, and some further conditions given in the Appendix A. In our implementation $\mathcal{H}_n$ is the finite dimensional linear sieve given by

$$\mathcal{H}_n = \left\{ h : h = \sum_{j=1}^{J_n} a_j q_j(\cdot) \right\} \tag{12}$$

where $q^{J_n}(z) = (q_1(z), ..., q_{J_n}(z))'$ is a vector containing a linear sieve basis, with $J_n \to \infty$ as $n \to \infty$.

To better understand the first step estimator and how it can be computed by standard methods consider the approximation

$$X = E[h(Z)|\, X] \approx E[a' q^{J_n}(Z)\big|\, X] = a' E[q^{J_n}(Z)\big|\, X],$$

which suggests a two step procedure for obtaining $\hat{h}_n$: (i) first compute the fitted values $\hat{q}(X) = \hat{E}[q^{J_n}(Z)\big|\, X]$ by OLS of $q^{J_n}(Z)$ on $p^{K_n}(X)$; and then (ii) run Ridge regression X on $\hat{q}(X)$. Indeed, if we define $D_n = E_n[\hat{q}(X)X']$, $Q_{2n} = E_n[q^{J_n}(Z)q^{J_n}(Z)']$, and

$$\hat{A}_{\lambda_n} = E_n[\hat{q}(X)\hat{q}(X)'] + \lambda_n Q_{2n}.$$

Then, the closed form solution to (11) is given by

$$\hat{h}_n(\cdot) = D_n' \hat{A}_{\lambda_n}^{-1} q^{J_n}(\cdot). \tag{13}$$

This estimator can be easily implemented by an OLS and a *standard* Ridge regression steps: (i) standardize q^{J_n} so that Q_{2n} becomes the identity (simply multiply the original q^{J_n} by $Q_{2n}^{-1/2}$); (ii) run OLS $q^{J_n}(Z)$ on $p^{K_n}(X)$ and keep fitted values $\hat{q}(X)$; (iii) run standard Ridge regression of X on $\hat{q}(X)$; the slope coefficient in the last regression is $D_n' \hat{A}_{\lambda_n}^{-1}$. Section 3.4 further discusses implementation of the estimation of h_0 in a more general setting with additional exogenous variables.

An alternative minimum norm approach requires choosing two sequences of positive numbers a_n and b_n and solving the program

$$\tilde{h}_n := \arg\min\{||h||_n^2 : h \in \mathcal{H}_n, ||\hat{m}(X; h)||_n^2 \leq b_n/a_n\}.$$

This is the approach used in Santos (2011) for different functionals than the OLIVA. We prefer our implementation, since we only need one tuning parameter rather than two, and data driven methods for choosing λ_n are readily available; see Section 3.4.

3.2 Second-Step Estimation and Inference

This section establishes the consistency and asymptotic normality of $\hat{\beta}$, and the consistency of its asymptotic variance, which is useful for inference. Recall $W = (Y, X', Z')'$ and define

$$m(W, \beta, h, g) = (Y - X'\beta)h(Z) - (g(X) - X'\beta)(h(Z) - X), \tag{14}$$

with the short notation $m_0 = m(W, \beta, h_0, g_0)$. The second term in (14) accounts for the asymptotic impact of estimating the instrument h_0. When the minimum norm structural function g_0 is linear, like with a binary treatment, this second term is zero and there will be no impact from estimating h_0 on inference. Thus, we can interpret this second term in m as accounting for a "nonlinearity bias" in inference of the IV estimator.

To estimate the asymptotic variance of $\hat{\beta}$ is useful to estimate g_0. We introduce a Tikhonov-type estimator that is the dual of $\hat{h}_n$. Let $\hat{g}_n(\cdot)$ denote a PSMD estimator of g_0 given by

$$\hat{g}_n(\cdot) = G_n' B_{\lambda_n}^{-1} p^{K_n}(\cdot), \tag{15}$$

with $G_n = E_n[\hat{p}(Z)Y]$, $\hat{p}(Z) = \hat{E}[p^{K_n}(X) | Z]$, $\hat{E}[g(X)| Z = z] = q^{J_n\prime}(z)(Q'Q)^{-1} \sum_{i=1}^{n} q^{J_n}(Z_i)g(X_i)$, $Q = [q^{J_n}(Z_1), ..., q^{J_n}(Z_n)]'$, $P_{2n} = E_n[p^{K_n}(X)p^{K_n}(X)']$, and $\hat{B}_{\lambda_n} = E_n[\hat{p}(Z)\hat{p}(Z)'] + \lambda_n P_{2n}$. For ease of presentation, we use the same notation for the tuning parameters in $\hat{h}_n$ and $\hat{g}_n$, although of course we will use different tuning parameters K_n and J_n for estimating $\hat{h}_n$ or $\hat{g}_n$, see Section 3.4 for issues of implementation.

Theorem 3.2 *Let Assumptions 1-3 above and Assumptions A1-A5, A6(i-iii) in the Appendix A hold. Then, $\hat{\beta}$ is consistent and asymptotically normal, i.e.*

$$\sqrt{n}(\hat{\beta} - \beta) \longrightarrow_d N(0, \Sigma),$$

where $\Sigma = E[h_0(Z)X']^{-1} E[m_0 m_0'] E[X h_0(Z)']^{-1}$. Furthermore, Σ is consistently estimated by

$$\hat{\Sigma} = E_n[\hat{h}_n(Z_i)X_i']^{-1} E_n[\hat{m}_{ni}\hat{m}_{ni}'] E_n[X_i \hat{h}_n'(Z_i)]^{-1}, \tag{16}$$

where $\hat{m}_{ni} = m(W_i, \hat{\beta}, \hat{h}_n, \hat{g}_n)$.

Remark 3.2 *If $E[\varepsilon| Z] = 0$ is relaxed to only $E[\varepsilon h_0(Z)] = 0$, then the asymptotic normality of $\hat{\beta}$ goes through with $(g(X) - X'\beta)$ in m_0 replaced by v_n in (40) of the Appendix, provided v_n and the resulting m_0 have finite variances, see the proof of Theorem 3.2.*

The assumptions in Theorem 3.2 are standard in the literature of two-step semiparametric estimators. Theorem 3.2 can be then used to construct confidence regions for β and testing hypotheses about β following standard procedures. The proof of Theorem 3.2 relies on new L_2-rates of convergence for $\hat{h}_n$ and $\hat{g}_n$ under partial identification of h and g (note that the rates in Chen and Pouzo (2012) are given under point identification and Santos (2011) obtained related rates but for a weak norm).

3.3 Partial Effects Interpretation, Exogenous Controls and Discrete Variables

We provide now a partial effects interpretation for subvectors of the OLIVA parameter β that are analogous to OLS. Define $X = (X_1', X_2')'$ and partition β accordingly as $\beta = (\beta_1', \beta_2')'$. Suppose we are only interested in β_2. From standard OLS theory, we obtain

$$\beta_2 = E[V_2 V_2']^{-1} E[V_2 g(X)],$$

where V_2 is the OLS error from the regression of X_2 on X_1. This result could be used to obtain an estimator of β_2 that does not compute an estimator for β_1 and that reduces the dimensionality of the problem of estimating h (from the dimension of the original X to the dimension of X_2), since now we can use the condition

$$E[\, h(Z)|\, V_2] = V_2 \text{ a.s.}$$

This method might be particularly useful when the dimension of X_1 is large and g has a partly linear structure

$$g(X) = \gamma_1' X_1 + g_2(X_2), \tag{17}$$

since then $\beta_2 = E[V_2 V_2']^{-1} E[V_2 g_2(X_2)]$ can be interpreted as providing a best linear approximation to $g_2(X_2)$ with a linear function of V_2, i.e.

$$\beta_2 = \arg\min_{b_2} E[\big(g_2(X_2) - b_2' V_2\big)^2].$$

In this discussion, X_1 could be variables that are of secondary interest.

Suppose now that there are exogenous variables included in the structural equation g. This means X and Z have common components. Specifically, with some abuse of notation, define $X = (X_1', X_2')'$ and $Z = (Z_1', Z_2')'$ where $X_1 = Z_1$ denote the overlapping components of X and Z, with dimension $p_1 = q_1$. This is a very common situation in applications, where exogenous controls are often used. In this setting a solution of $E[\, h(Z)|\, X] = X$ a.s. has the form $h(Z) = (Z_1', h_2'(Z))'$, where

$$E[\, h_2(Z)|\, X] = X_2 \text{ a.s.} \tag{18}$$

Following the arguments of the general case, we could obtain an estimator given by $\hat{h}_n = (Z_1', \hat{h}_{2n}')'$, where

$$\hat{h}_{2n}(\cdot) = D_{2n}' \hat{A}_{\lambda_n}^{-1} q^{J_n}(\cdot), \tag{19}$$

and $D_{2n} := E_n[\hat{q}(X) X_2']$. This setting also covers the case of an intercept with no other common components, where $X_1 = Z_1 = 1$ and $q_1 = 1$. The asymptotic normality for $\hat{\beta}$ continues to hold, with no changes in the asymptotic distribution.

If the dimension of X_1 is high and the sample size is moderate, the method above may not perform well due to the curse of dimensionality. Equation (18) implies

$$E[\, h_2(Z)|\, X_2] = X_2 \text{ a.s.} \tag{20}$$

so that nonparametric estimation of h_{20} only involves functions $p^{K_n}(X_2)$ and $q^{J_n}(Z)$. Equation (20) is still necessary for regular identification. Summarizing, for implementing our methods with moderate or high dimensional controls X_1 we recommend our general algorithm above with bases $\{p^{K_n}(X_2)\}$ and $\left\{q^{J_n}(Z)\right\}$, which is consistent with the specification in (17). Further details on implementation are provided in Section 3.4.

Simplifications occur when some variables are discrete. When the endogenous variable X is discrete we do not need $K_n \to \infty$, and we can choose p^{K_n} as a saturated basis. Consider first the important case of a binary endogenous variable $X = (1, X_2)$ with $X_2 \in \{0,1\}$. Define the propensity score $\pi(z) := \Pr\left(X_2 = 1 | Z = z\right)$. We show below that under the mild assumption that $\pi(z)$ is not constant, Assumption 3 holds. Furthermore, the minimum norm solution h_0 is simply

$$h_0(z) = \alpha + \gamma\pi(z), \tag{21}$$

where $\alpha = \bar{\pi}\left(1-\gamma\right)$, $\gamma = \bar{\pi}(1-\bar{\pi})/var(\pi(Z))$ and $\bar{\pi} = \Pr\left(X_2 = 1\right)$. An implication of this representation is that the slope of the OLIVA is

$$\frac{Cov(Y, h_0(Z))}{Cov(X_2, h_0(Z))} = \frac{Cov(Y, \pi(Z))}{Cov(X_2, \pi(Z))} \equiv \alpha_\pi^{IV}, \tag{22}$$

i.e., the LATE estimand α_π^{IV} using the propensity score as instrument, which was suggested in Imbens and Angrist (1994). Thus, the OLIVA in the binary endogenous case coincides with an important IV estimand recommended in the literature. We summarize our findings in the following result. The proof can be found in the Appendix.

Proposition 3.3 *If $X = (1, X_2)$ with X_2 a binary endogenous variable, $0 < \bar{\pi} < 1$, and $var(\pi(Z)) > 0$, then Assumption 3 holds with a minimum norm solution h_0 given by (21). Furthermore, the OLIVA is $\beta = (c_\pi^{IV}, \alpha_\pi^{IV})'$, where $c_\pi^{IV} = E[Y] - \alpha_\pi^{IV}\bar{\pi}$ and α_π^{IV} is defined in (22).*

This result implies that for the binary endogenous case estimating h_0, and then β_0, simply requires estimating nonparametrically the propensity score.

More generally, if X has d points of support, say $\{x_1, ..., x_d\}$, then we can set $K_n = d$ and $p_k(x) = 1(x = x_k)$, $k = 1, .., K_n$, in our general algorithm. Define the unconditional probabilities $\Pr\left(X = x_j\right) = \pi_j$, $j = 1, ..., d$. Then, Assumption 3 boils down to existence of h satisfying the linear equalities, for $k = 1, ..., d$,

$$E[h(Z)p_k(X)] = \pi_k x_k. \tag{23}$$

Using Theorem 2, pg. 65, in Luenberger (1997), we can find a closed form solution for h_0 as follows. Define the generalized propensity scores $\pi_j(z) := \Pr\left(X = x_j | Z = z\right)$ and the random vector $\Pi \equiv \Pi(Z) = (\pi_1(Z), ..., \pi_d(Z))'$. If $E[\Pi\Pi']$ is positive definite, then the minimum norm solution to (4) is given by $h_0(z) = \gamma'\Pi(z)$ where $\gamma = \left(E[\Pi\Pi']\right)^{-1} S$ and $S = (\pi_1 x_1, ..., \pi_d x_d)'$. Thus, for discrete endogenous variables our nonparametric algorithm with $K_n = d$ and $p_k(x) = 1(x = x_k)$, $k = 1, .., K_n$, is a semiparametric method where $h_0(z) = \gamma'\Pi(z)$ is estimated by estimating the conditional probabilities $\Pi(z)$ by $\hat{\Pi}(z) = (\hat{\pi}_1(Z), ..., \hat{\pi}_d(Z))$, where $\hat{\pi}_k(z) = \hat{E}[p_k(X) | Z = z]$. In estimating γ, if the sample analog of $E[\Pi\Pi']$ is positive definite, then there is no need to choose λ for estimating h_0. If this matrix is not invertible, we can apply the Tikhonov-type estimator, as proposed above.

Similarly, when Z is discrete we do not need J_n diverging to infinity. As before, we can choose a linear sieve $\mathcal{H}_n$ that is saturated and $q^{J_n}(Z)$ could be a saturated basis for it. Specifically, if Z takes J discrete values, $\{z_1, ..., z_J\}$, we can take $q_j(z) = 1(z = z_j)$, $j = 1, ..., J_n \equiv J$.

Summarizing, all the different cases (with or without controls, nonparametric or semiparametric structural functions, discrete or continuous variables) can be implemented with the same algorithm but with different definitions of the approximation bases $\{p^{K_n}(X), q^{J_n}(Z)\}$. In all these cases, the formulas for the asymptotic variance of $\hat{\beta}$ remain the same. The following section provides further details on implementation.

3.4 Implementation

To enhance the practical applicability of our method we summarize its implementation in what we think is the most useful case in empirical applications: estimation in the presence of a vector of controls entering linearly in the models for g and h. Since the vector of controls can be high dimensional, we do not think of the linearity of the controls as a strong assumption. As before, we split $X = (X_1', X_2')'$ and $Z = (Z_1', Z_2')'$, where $X_1 = Z_1$ denote the vector of exogeneous controls (containing an intercept), with dimension $p_1 = q_1$. The endogenous variable of interest X_2 has dimension $p_2 = p - p_1$, and the instrument Z_2 has dimension $q_2 = q - q_1$. Following the discussion above, for implementation one has to choose bases $\{p^{K_n}(x), q^{J_n}(z)\}$ and the tuning parameters $\{J_n, K_n, \lambda_n\}$. Using these imputs, we estimate an instrument $\hat{h}_n = (Z_1', \hat{h}_{2n}')'$, where $\hat{h}_{2n}$ estimates a minimum norm solution h_{20} of

$$E[h_2(Z)|\, X_2] = X_2 \text{ a.s.}$$

An appealing feature of sieve estimation is that additional semiparametric restrictions can be imposed on h_2 simply by restricting the terms in the basis $\{q^{J_n}(z)\}$. These include additivity or exclusion restrictions, among others. For example, one restriction that we impose in this section is that h_2 is linear in Z_1, i.e. $h_2(Z) = a_1' Z_1 + h_{20}(Z_2)$. This is, of course, not necessary for regular identification, but it ameliorates the curse of dimensionality, specially when Z_1 is high dimensional, and it may lead to better finite sample performance (by reducing variance).

As explained in the previous section, the implementation varies according to the nature of the endogenous variable X_2 and the instrument Z_2 (whether continuous or discrete). In the continuous case we need to choose $\{J_n, K_n, \lambda_n\}$ for estimating h_0. We can make these choices simultaneously by Generalized Cross-validation (cf. Wahba (1990), GCV henceforth). To simplify the computations we implement GCV by setting first $J_n = q_1 + j_n$ for fixed value j_n in a small grid (e.g. $j_n \in \{4, 5, 6, 7\}$), then setting $K_n = p_1 + \lfloor c j_n \rfloor$, for a grid of values for c in $[1, 3]$, where $\lfloor x \rfloor$ is the floor function, and then minimizing in $\tau = \{j_n, c, \lambda\}$ the GCV criteria $GCV_n(\tau)$ given below in (25) over the grid values.

Details are given as follows. Let H_n denote the $n \times p$ matrix with rows $\hat{h}_n(Z_i)$ $i = 1, ..., n$. Let $\mathbf{X} = [X_1, ..., X_n]'$ and denote by $\mathbf{X}_1 \equiv \mathbf{Z}_1$ and $\mathbf{X}_2$, respectively, the corresponding $n \times p_1$ and $n \times p_2$ design matrices based on the partition $X = (X_1', X_2')'$. Construct the $n \times J_n$ matrix $Q = [\mathbf{Z}_1 \; Q_2]$, $J_n = q_1 + j_n$, $Q_2 = [q^{j_n}(Z_{21}), ..., q^{j_n}(Z_{2n})]'$ (Q_2 excludes an intercept), and similarly the $n \times K_n$ matrix $P = [\mathbf{X}_1 \; P_2]$, $K_n = p_1 + k_n$, $k_n = \lfloor c j_n \rfloor$, $P_2 = [p^{k_n}(X_{21}), ..., p^{k_n}(X_{2n})]'$ (P_2 excludes an intercept), and their corresponding projection matrices Π_P and Π_Q, where $\Pi_A = A(A'A)^{-1}A'$ for a generic matrix A. Denote also I_d as the $d \times d$ identity matrix. Then, to provide an expression for H_n we construct

$$H_{2n} = Q\hat{A}_{\lambda_n}^{-1}Q'\Pi_P\mathbf{X}_2,$$

where

$$\hat{A}_{\lambda_n} = Q'(\Pi_P + \lambda_n I_n)Q.$$

Finally,

$$H_n = [\mathbf{Z}_1 \ H_{2n}]$$

and

$$\hat{\beta} = \left(H_n'\mathbf{X}\right)^{-1} H_n'\mathbf{Y}, \tag{24}$$

where $\mathbf{Y} = [Y_1, ..., Y_n]'$. This provides a matrix formula implementation for our estimator.

To give the GCV criteria define $L_\tau = \mathbf{X}\left(H_n'\mathbf{X}\right)^{-1} H_n'$, $\hat{Y}_\tau = L_\tau\mathbf{Y} = (\hat{Y}_{\tau 1}, ..., \hat{Y}_{\tau n})'$ and $v_\tau = trace(L_\tau)$. Then, the GCV criteria for estimating $\hat{\beta}$ is

$$GCV_n(\tau) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_{\tau i}}{1 - (v_\tau/n)}\right)^2. \tag{25}$$

To estimate g_0 in the presence of a high dimensional vector of controls we follow the specification in (17). The $n \times 1$ vector G_n of fitted values $\hat{g}_n(X_i)$, $i = 1, ..., n$, is given by

$$G_n = P\hat{B}_{\lambda_n}^{-1}P'\Pi_Q\mathbf{Y}, \tag{26}$$

where

$$\hat{B}_{\lambda_n} = P'(\Pi_Q + \lambda_n I_n)P.$$

Since G_n is linear in $\mathbf{Y}$, we can easily set another GCV method for selecting $\{J_n, K_n, \lambda_n\}$ for $\hat{g}_n$ (simply replaced L_τ above by $L_\tau = P\hat{B}_{\lambda_n}^{-1}P'\Pi_Q$). See also Centorrino, Feve and Florens (2017).

The following algorithm summarizes the main steps for implementation[3]:

Step 1. Compute $\tau_n = \arg\min GCV_n(\tau)$, over a finite grid of values of $\tau = \{j, c, \lambda\}$.

Step 2. Compute $\hat{\beta}$ following (24).

Step 3. Compute $\hat{g}_n$ following (26).

Step 4. Compute $\hat{m}_{ni} = m(W_i, \hat{\beta}, \hat{h}_n, \hat{g}_n)$ and $\hat{\Sigma} = E_n[\hat{h}_n X_i']^{-1}E_n[\hat{m}_{ni}\hat{m}_{ni}']E_n[X_i\hat{h}_n']^{-1}$.

For continuous variables we recommend using B-splines as sieve basis. If Z_2 is discrete, with support $\{z_{21}, ..., z_{2j_2}\}$, we set $J_n = q_1 + j_2 - 1$ and $q_j(z_2) = 1(z_2 = z_{2j})$, $j = 2, ..., j_2$, in the algorithm above. Similarly, if X_2 is discrete, with support $\{x_{21}, ..., x_{2k_2}\}$, we set $K_n = p_1 + k_2 - 1$, and $p_k(x_2) = 1(x_2 = x_{2k})$, $k = 2, ..., k_2$. In this discussion, we exclude the first element in the indicators because the intercept is part of the exogenous controls.

[3] Matlab and R code to implement the TSIV and related inferences are available from the authors upon request.

3.5 Weighted least squares

Our previous discussion can be extended to weighted least squares criteria. That is, suppose that the OLIVA is now defined as

$$\beta_w = \arg\min_{\gamma\in\mathbb{R}^p} E[\left(g(X) - \gamma' X\right)^2 w(X)], \tag{27a}$$

where $w(X)$ is a positive weight function. This extension can be relevant in a number of applications. For example, if f is the density of X and f^* is a counterfactual density, by taking $w(x) = f^*(x)/f(x)$ the linear approximation is under a counterfactual density which might better summarized the interest of the researcher. Our theory can be extended to this setting as follows. The necessary condition for regular identification of β_w is now

$$E[h(Z)|\, X] = Xw(X) \text{ a.s}, \tag{28}$$

for an square integrable $h(\cdot)$; and under this condition and if $E[XX'w(X)]$ is positive definite, then it follows that

$$\begin{aligned}\beta_w &= E[XX'w(X)]^{-1}E[Xw(X)g(X)] \\ &= E[h(Z)X']^{-1}E[h(Z)Y].\end{aligned}$$

The estimation proceeds as in our basic case (where $w = 1$). If w is unknown, we can estimate w nonparametrically and use the plugging estimator with the estimated w to solve for h in (28). Our estimator will be consistent and asymptotically normal under regularity conditions, as in the basic case. It remains to study if estimation of w changes the asymptotic variance of the resulting estimator of β_w. This issue is, however, beyond the scope of this paper and is left for future research.

4 A Robust Hausman Test

Applied researchers are concerned about the presence of endogeneity, and they traditionally use tools such as the Hausman (1978)'s exogeneity test for its measurement. This test, however, is uninformative under misspecification; see Lochner and Moretti (2015). The reason for this lack of robustness is that in these cases OLS and IV estimate different objects under exogeneity, with the estimand of standard IV depending on the instrument itself. As an important by-product of our analysis, we robustify the classic Hausman test of exogeneity against nonparametric misspecification of the linear regression model.

The classical Hausman test of exogeneity (cf. Hausman (1978)) compares OLS with IV. If we use the TSIV as the IV estimator, we obtain a robust version of the classical Hausman test, robust to the misspecification of the linear model. For implementation purposes it is convenient to use a regression-based test (see Wooldridge (2015), pg. 481). We illustrate the idea in the case of one potentially endogenous variable X_2 and several exogenous variables X_1, with X_1 including an intercept.

In the model

$$Y = \beta_1' X_1 + \beta_2 X_2 + U, \qquad E[Uh(Z)] = 0,\ h(Z) = (X_1', h_2(Z))',$$

the variable X_2 is exogenous if $Cov(X_2, U) = 0$. If we write the first-stage as

$$X_2 = \alpha_1' X_1 + \alpha_2 h_2(Z) + V, \qquad E[Vh(Z)] = 0,$$

then weak exogeneity of X_2 is equivalent to $Cov(V, U) = 0$. This in turn is equivalent to $\rho = 0$ in the least squares regression

$$U = \rho V + \xi.$$

A simple way to run a test for $\rho = 0$ is to consider the augmented regression

$$Y = \beta' X + \rho V + \xi,$$

estimated by OLS and use a standard $t - test$ for $\rho = 0$.

Since V is unobservable, we first need to obtain residuals from a regression of the endogenous variable X_2 on X_1 and $\hat{h}_{2n}(Z)$, say $\hat{V}$. Then, run the regression of Y on X and $\hat{V}$. The new Hausman test is a standard two-sided t-test for the coefficient of $\hat{V}$, or its Wald version in the multivariate endogenous case. Denote the t-test statistic by t_n. The benefit of this regression approach is that under some regularity conditions given in Appendix A no correction is necessary in the OLS standard errors because $\hat{V}$ is estimated. Denote $S = (X, V)'$, and consider the following mild assumption.

Assumption 4: The matrix $E[SS']$ is finite and non-singular.

Theorem 4.1 *Let Assumptions 1-4 above and Assumptions A1-A6 in the Appendix A hold. Then, under the the null of exogeneity of* X_2, $t_n \longrightarrow_d N(0, 1)$.

The proof of Theorem 4.1 is involved and requires stronger conditions than that of Theorem 3.2. In particular, for obtaining the result that standard OLS theory applies under the null hypothesis we have used a conditional exogeneity assumption between U and Z, $E[U|\, Z] = 0$ a.s. Simulations below show that, at least for the models considered, this assumption leads to a robust Hausman test that is able to control the empirical size. We note that under the null of exogeneity we do not require the model to be linear in the sense of $E[U|\, X] = 0$ a.s.

5 Monte Carlo

This section studies the finite sample performance of the proposed methods. Consider the following Data Generating Process (DGP):

$$\begin{cases} Y = \sum_{j=1}^{p} H_j(X) + \varepsilon, \\ Z = s(D), \\ \varepsilon = \rho_\varepsilon V + \zeta, \end{cases} \qquad \begin{pmatrix} X \\ D \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix} \right),$$

where $H_j(x)$ is the $j - th$ Hermite polynomial, with the first four given by $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$ and $H_3(x) = x^3 - 3x$; $V = X - E[X|\, Z]$, ζ is a standard normal, drawn independently of X and D, and s is a monotone function given below. The DGP is indexed by p and the function s. To generate V note

$$E[X|\,Z] = E[E[X|\,D]|\,Z] = \gamma E[D|\,Z] = \gamma s^{-1}(Z),$$

where s^{-1} is the inverse of s. Thus, by construction Z is exogenous, $E[\varepsilon|\,Z] = 0$, while X is endogenous because $E[\varepsilon|\,X] = \rho X$, with $\rho = \rho_\varepsilon(1-\gamma^2)$, $\rho_\varepsilon > 0$ and $-1 < \gamma < 1$.

The structural function g is given by

$$g(x) = \sum_{j=1}^{p} H_j(X),$$

and is therefore linear for $p = 1$, but nonlinear for $p > 1$. It follows from the orthogonality of Hermite polynomials that the true value for OLIVA is $\beta = 1$ and that g is identified if $\gamma \neq 0$ (since $Var(E[g(X)|\,Z]) = \sum_{j=1}^{\infty} g_j^2 \gamma^{2j}$, where $g_j = E[g(X)H_j(X)]$ is the $j-th$ Hermite coefficient, and thus, $E[g(X)|\,Z] = 0 \Longrightarrow g = 0$).

Note also that the OLIVA is regularly identified, because $h(Z) = s^{-1}(Z)/\gamma$ solves

$$E[h(Z)|\,X] = X.$$

We consider three different DGPs, corresponding to different values of p and functional forms for s:

DGP1: $p = 1$ and $s(D) = D$ (linear; $s^{-1}(Z) = Z$);

DGP2: $p = 2$ and $s(D) = D^3$ (nonlinear; $s^{-1}(Z) = Z^{1/3}$);

DGP3: $p = 3$ and $s(D) = \exp(D)/(1 + \exp(D))$ (nonlinear; $s^{-1}(Z) = \log(Z) - \log(1 - Z)$);

Several values for the parameters (γ, ρ) will be considered: $\gamma \in \{0.4, 0.8\}$ and $\rho \in \{0, 0.3, 0.9\}$. We will compare the TSIV with OLS and standard IV (using instrument Z). For DGP1, $h(Z) = \gamma^{-1}Z$ and hence the standard IV estimator with instrument Z is a consistent estimator for the OLIVA. Indeed, the standard IV can be seen as an oracle (infeasible version of our TSIV) under DGP1, where h is known rather than estimated. This allows us to see the effect of estimating h_0 on inferences. For DGP2 and DGP3, IV is not consistent for the OLIVA. The number of Monte Carlo replications is 5000. The sample sizes considered are $n = 100$, 500 and 1000.

Tables 1-3 report the Bias and MSE for OLS, IV and the TSIV for DGP1-DGP3, respectively. Our estimator is implemented with B-splines, following the GCV described in Section 3.4, where to simplify the computations we set $J_n = 6$ and $K_n = 2J_n$, and optimize only in λ for each simulated data. A similar strategy was followed in Blundell, Chen and Kristensen (2007). Likewise, we have followed a simple rule for selecting $\{J_n, K_n, \lambda_n\}$ for $\hat{g}_n$: switch the values of J_n and K_n used for $\hat{h}_n$ to compute $\hat{g}_n$ (so now $J_n = 2K_n$), and use same value of λ_n for $\hat{g}_n$ as for estimating $\hat{h}_n$, which seems to work well in our simulations. Remarkably, for DGP1 in Table 1 our TSIV implemented with GCV performs comparably or even better than IV (which does not estimate h and uses the true h). Thus, our estimator seems to have an oracle property, performing as well as the method that uses the correct specification of the model. As expected, OLS is best under exogeneity, but it leads to large biases

Table 1: Bias and MSE for DGP 1.

ρ	γ	n	BIAS_OLS	BIAS_IV	BIAS_TSIV	MSE_OLS	MSE_IV	MSE_TSIV
0.0	0.4	100	-0.0021	-0.0019	0.0010	0.0109	0.0829	0.0554
		500	0.0017	0.0025	0.0020	0.0021	0.0127	0.0105
		1000	-0.0001	0.0018	0.0020	0.0010	0.0067	0.0054
	0.8	100	-0.0030	-0.0040	-0.0040	0.0102	0.0163	0.0159
		500	0.0001	-0.0004	-0.0004	0.0019	0.0030	0.0030
		1000	0.0019	0.0025	0.0026	0.0010	0.0016	0.0016
0.3	0.4	100	0.2950	-0.0101	0.0841	0.0968	0.0908	0.0729
		500	0.2993	0.0026	0.0347	0.0915	0.0145	0.0168
		1000	0.3006	-0.0003	0.0189	0.0914	0.0071	0.0080
	0.8	100	0.2956	-0.0107	0.0061	0.0987	0.0207	0.0216
		500	0.2991	0.0009	0.0038	0.0918	0.0039	0.0039
		1000	0.2987	-0.0023	-0.0012	0.0904	0.0019	0.0019
0.9	0.4	100	0.8993	-0.0827	0.1753	0.8213	0.1990	0.1569
		500	0.9028	-0.0145	0.0421	0.8173	0.0295	0.0296
		1000	0.8998	-0.0066	0.0231	0.8108	0.0130	0.0140
	0.8	100	0.8965	-0.0186	0.0287	0.8270	0.0573	0.0571
		500	0.8980	-0.0036	0.0030	0.8114	0.0108	0.0109
		1000	0.8993	0.0031	0.0058	0.8111	0.0049	0.0050

under endogeneity. For the nonlinear models DGP2 and DGP3, IV deteriorates because the linear model is misspecified. Our TSIV performs well, with a MSE that converges to zero as n increases. Increasing γ makes the instrument stronger, thereby reducing the MSE of IV estimates, while for a fixed γ, increasing the level of endogeneity increases the MSE.

We have done extensive sensitivity analysis on the performance of the TSIV estimator. Simulations in the Supplemental Appendix report the sensitivity of the estimator to different choices of tuning parameters, J_n, K_n and λ_n. From these results, we see that the TSIV estimator is not sensitive to the choice of these parameters, within the wide ranges for which we have experimented. This is consistent with the regular identification, which means that the estimator should be robust to local perturbations of the tuning parameters. Likewise, unreported simulations with other DGPs confirm the overall good performance of the proposed TSIV under different scenarios.

Table 4 provides the results for coverage of confidence intervals based on the asymptotic normality of the TSIV using the GCV-computed λ_n, along with that using $0.7\lambda_n$ and $0.9\lambda_n$. The coverage is very stable for the three choices of λ_n considered. The performance in DGP1 and DGP2 is fairly good, while in DGP3 it noticeably improves when the sample size increases.

We now turn to the Hausman test. Practitioners often use the Hausman test to empirically evaluate the presence of endogeneity. As mentioned above, the standard Hausman test is not robust to misspefication of the linear model, because in that case OLS and IV estimate different parameters (Lochner and Moretti (2015)). We confirm this by simulating data from DGP1-DGP3 and reporting rejection frequencies for the standard Hausman test for $\gamma \in \{0.4, 0.8\}$. Table 5 contains the results. For DGP1, the rejection frequencies for $\rho = 0$ are close to the nominal level of 5% across the different

Table 2: Bias and MSE for DGP 2.

ρ	γ	n	BIAS_OLS	BIAS_IV	BIAS_TSIV	MSE_OLS	MSE_IV	MSE_TSIV
0.0	0.4	100	0.0131	-0.0030	-0.0037	0.1009	0.6321	0.2226
		500	0.0083	0.0216	0.0126	0.0213	0.1319	0.0479
		1000	0.0021	0.0005	0.0034	0.0115	0.0764	0.0228
	0.8	100	-0.0012	0.0001	-0.0001	0.0990	0.4559	0.1286
		500	0.0015	0.0056	0.0032	0.0211	0.1261	0.0275
		1000	0.0019	0.0084	0.0030	0.0113	0.0689	0.0154
0.3	0.4	100	0.2932	-0.0472	0.0605	0.1859	0.6167	0.2342
		500	0.2874	-0.0325	0.0302	0.1023	0.1417	0.0594
		1000	0.3008	-0.0135	0.0402	0.1013	0.0778	0.0331
	0.8	100	0.3064	0.0083	0.0318	0.1987	0.4554	0.1400
		500	0.3020	0.0078	0.0208	0.1114	0.1226	0.0289
		1000	0.3046	0.0076	0.0248	0.1040	0.0647	0.0168
0.9	0.4	100	0.9053	-0.1359	0.2155	0.9270	1.0165	0.3615
		500	0.8968	-0.0093	0.0794	0.8260	0.1619	0.0914
		1000	0.8974	-0.0122	0.0493	0.8159	0.0817	0.0449
	0.8	100	0.9095	-0.0117	0.0491	0.9425	0.5482	0.1921
		500	0.8969	-0.0013	0.0226	0.8290	0.1405	0.0435
		1000	0.8981	-0.0021	0.0271	0.8185	0.0753	0.0220

Table 3: Bias and MSE for DGP 3.

ρ	γ	n	BIAS_OLS	BIAS_IV	BIAS_TSIV	MSE_OLS	MSE_IV	MSE_TSIV
0.0	0.4	100	-0.0570	-1.5268	-0.0717	0.5023	381.7332	0.6817
		500	-0.0021	-0.5039	-0.0346	0.1000	155.9296	0.1326
		1000	-0.0014	-0.0365	-0.0378	0.0550	0.6179	0.0681
	0.8	100	-0.0418	-0.4112	-0.1106	0.4795	2.6703	0.4935
		500	-0.0096	-0.2270	-0.0411	0.1072	0.4192	0.1084
		1000	-0.0113	-0.2150	-0.0330	0.0527	0.2452	0.0543
0.3	0.4	100	0.2899	-5.4825	0.0227	0.6475	28179.2626	0.8182
		500	0.2882	-0.1335	0.0060	0.1878	1.5707	0.1571
		1000	0.2887	-0.0822	0.0199	0.1351	0.6518	0.0926
	0.8	100	0.2693	-0.3815	-0.0857	0.5906	11.1463	0.5498
		500	0.3062	-0.1985	-0.0249	0.2061	0.4885	0.1221
		1000	0.2951	-0.2166	-0.0246	0.1395	0.2512	0.0570
0.9	0.4	100	0.8470	1.4445	0.1675	1.1993	1772.3946	0.8970
		500	0.8888	-0.3336	0.0449	0.9098	4.8599	0.2103
		1000	0.8914	-0.1313	0.0158	0.8473	0.8558	0.0982
	0.8	100	0.8341	-0.5724	-0.0917	1.1833	4.3735	0.6045
		500	0.8749	-0.2933	-0.0566	0.8668	0.6084	0.1301
		1000	0.8863	-0.2466	-0.0401	0.8380	0.2861	0.0681

Table 4: 95% coverage for TSIV.

			DGP1			DGP2			DGP3		
ρ	γ	n	0.7cv	0.9cv	1.0cv	0.7cv	0.9cv	1.0cv	0.7cv	0.9cv	1.0cv
0.0	0.4	100	0.973	0.976	0.976	0.950	0.954	0.955	0.899	0.901	0.903
		500	0.976	0.978	0.977	0.950	0.951	0.951	0.929	0.931	0.932
		1000	0.971	0.973	0.973	0.954	0.957	0.956	0.931	0.931	0.930
	0.8	100	0.964	0.965	0.966	0.929	0.929	0.931	0.837	0.837	0.838
		500	0.957	0.957	0.957	0.941	0.942	0.944	0.902	0.905	0.905
		1000	0.950	0.951	0.951	0.932	0.938	0.941	0.926	0.927	0.927
0.3	0.4	100	0.976	0.982	0.982	0.950	0.948	0.949	0.919	0.921	0.922
		500	0.957	0.957	0.959	0.949	0.952	0.950	0.931	0.933	0.932
		1000	0.964	0.965	0.965	0.938	0.939	0.938	0.936	0.936	0.934
	0.8	100	0.945	0.945	0.946	0.917	0.920	0.920	0.858	0.861	0.862
		500	0.944	0.941	0.941	0.946	0.946	0.946	0.917	0.920	0.921
		1000	0.961	0.960	0.960	0.940	0.941	0.941	0.917	0.923	0.923
0.9	0.4	100	0.903	0.901	0.902	0.938	0.943	0.943	0.955	0.957	0.956
		500	0.947	0.949	0.948	0.936	0.940	0.941	0.951	0.949	0.949
		1000	0.943	0.942	0.942	0.925	0.929	0.932	0.950	0.951	0.951
	0.8	100	0.931	0.930	0.930	0.920	0.921	0.921	0.899	0.898	0.898
		500	0.938	0.937	0.935	0.949	0.949	0.949	0.918	0.920	0.921
		1000	0.951	0.951	0.951	0.954	0.954	0.954	0.930	0.935	0.935

sample sizes, confirming the validity of the test under correct specification. However, for DGP2 and DGP3 we observe large size distortions for the standard Hausman test, as large as 85%. This shows that the standard Hausman test is unreliable under misspecification of the linear model. In contrast, the proposed robust tests is able to control type-I error uniformly across the three DGPs. We also report size-corrected empirical rejections under the alternative. For the linear model, the standard Hausman test has (slightly) larger power than the robust test, while for the nonlinear model DGP2, the robust test has much larger power. For DGP3, the robust Hausman test outperforms the standard test for low values of ρ, while for large values of ρ they have comparable powers. In all cases we observe an empirical power that increases with the sample size and the level endogeneity, suggesting consistency against these alternatives. Despite these simulation results and others in the Supplemental Appendix, we stress that standard and robust Hausman tests should be viewed as complements rather than substitutes, given that they work under different set of assumptions.

We also report in the Supplemental Appendix further simulation results for cases where Z is discrete and X is continuous. For these DGPs g is not identified, although Assumption 3 is satisfied. These additional simulation results provide further evidence of the excellent finite sample performance of the TSIV and the robust Hausman test relative to their standard IV counterparts.

Overall, these simulations confirm the robustness of the proposed methods to misspecification of the linear IV model and their adaptive behaviour when correct specification holds. Furthermore, the TSIV estimator does not seem to be too sensitive to the choice of tuning parameters. Finally, the proposed Hausman test is indeed robust to the misspecification of the linear model, which makes it a reliable tool for economic applications. These finite sample robustness results confirm the claims made

Table 5: Empirical size and size-corrected power for Standard (S) and Robust (R) Hausman tests.

			DGP1		DGP2		DGP3	
ρ	γ	n	S	R	S	R	S	R
0.0	0.4	100	0.068	0.051	0.142	0.038	0.054	0.016
		500	0.062	0.044	0.070	0.016	0.048	0.010
		1000	0.056	0.040	0.053	0.006	0.050	0.008
	0.8	100	0.071	0.063	0.221	0.015	0.090	0.001
		500	0.047	0.045	0.146	0.004	0.521	0.004
		1000	0.060	0.054	0.110	0.001	0.850	0.002
0.3	0.4	100	0.242	0.174	0.066	0.082	0.085	0.105
		500	0.802	0.718	0.174	0.292	0.243	0.284
		1000	0.985	0.950	0.296	0.549	0.414	0.492
	0.8	100	0.952	0.896	0.108	0.562	0.671	0.727
		500	1.000	0.999	0.216	0.924	0.982	1.000
		1000	1.000	1.000	0.340	0.942	0.992	1.000
0.9	0.4	100	0.958	0.754	0.281	0.386	0.412	0.418
		500	1.000	0.993	0.744	0.952	0.930	0.938
		1000	1.000	1.000	0.932	0.994	0.999	0.997
	0.8	100	1.000	0.992	0.370	0.956	1.000	0.998
		500	1.000	1.000	0.739	0.980	1.000	1.000
		1000	1.000	1.000	0.910	0.980	1.000	1.000

for the TSIV estimator as a nonparametric analog to OLS under endogeneity.

6 Estimating the Elasticity of Intertemporal Substitution

In its log-linearized version, the Consumption-based Capital Asset Pricing Model (CCAPM) leads to the equation

$$\Delta c_{t+1} = \alpha + \psi r_{t+1} + U_t, \qquad E[U_t \mid Z_t] = 0 \text{ a.s.}, \tag{29}$$

where ψ is the elasticity of intertemporal substitution (EIS), Δc_{t+1} is the growth rate of consumption (the first difference in log real consumption per capita), r_{t+1} is the real interest rate at time $t+1$, α is a constant and Z_t is a vector of variables in the agent's information set at time t. The parameters $\beta_0 = (\alpha, \psi)'$ can be estimated from (29) by several estimation strategies; see, e.g., Hansen and Singleton (1983). Yogo (2004), using data from Campbell (2003), applied Two-Step Least Squares (TSLS), among other methods, to obtain estimates of ψ across different countries, arguing that in most cases the TSLS is subject to weak identification. Here we focus on quarterly US interest rate data, for which there is empirical evidence suggesting identification (the first-stage F statistic is 15.5). The data set is available at Motohiro Yogo's web page. A full description of the data is given in Campbell (2003).[4]

Following Yogo (2004), we use as instruments $Z_t = (r_{t-1}, \pi_{t-1}, \Delta c_{t-1}, dp_{t-1})$, where r_t is the nominal interest rate, π_t is inflation, and dp_t is the log dividend-price ratio. The sample size is $n = 206$. The

[4]It should be possible to extend our asymptotic results above to strictly stationary and ergodic time series data, although doing so is beyond the scope of this paper. Following much of the literature, including Yogo (2004), we compute standard errors assuming that the influence functions of the reported estimators are uncorrelated.

Table 6: EIS for Quarterly US data

	OLS	TSLS		TSIV	
J_n			4	5	6
estimate	0.161	0.060	0.151	0.160	0.162
s.e.	(0.054)	(0.095)	(0.092)	(0.099)	(0.101)
Hausman p-value		0.154	0.897	0.995	0.984

TSLS point estimate of ψ is 0.06, with a standard error of 0.09. We compare the TSLS with the proposed TSIV. To deal with the curse of dimensionality, we estimate $h_0(Z)$ with an additive nonparametric model, $h_0(Z_t) = h_{01}(Z_{t1}) + \cdots + h_{04}(Z_{t4})$. Specifically, we follow the implementation in our Monte Carlo and use B-splines with $4, 5$ or 6 knots for each instrument, leading to $J_n = 12, 15, 18$, respectively, $K_n = 2J_n$ and GCV for choosing λ. The matrix $Q = [q^{J_n}(Z_1), ..., q^{J_n}(Z_n)]'$ simply concatenates the corresponding matrix for each instrument. For estimating g_0 for the TSIV's standard errors we choose the same J_n and K_n as before, and compute λ_n by GCV. Further details on implementation are given in Section 3.4 for the case with an intercept (so $X_1 = Z_1 \equiv 1$).

Not surprisingly, our TSLS coincides with that of Yogo (2004). The TSIV is relatively much larger than the TSLS, and closer to the OLS, while the standard errors of both IV methods are similar in magnitude. These results are robust to the choice of J_n and K_n (we have experimented with $K_n = cJ_n$ for c between 1 and 3 and obtain qualitatively the same conclusions). If we apply our robust Hausman test of exogeneity we obtain very large p-values. Again, this result is robust to the choice of J_n and K_n. In contrast, the standard Hausman test leads to a p-value of 0.154.

We reach several conclusions from these results. First, the difference between the TSLS and the TSIV suggests that nonlinearities might be important in this application (indeed, the plot of the estimated $\hat{g}_n$, which is not reported here for the sake of space, reveals a marked nonlinear estimate). Second, once one accounts for the misspecification uncertainty, the null hypothesis of exogeneity cannot be rejected, thereby suggesting that for the purpose of estimating a log-linearized version of the Euler equation, endogeneity bias may be a second-order concern.

7 Appendix A: Notation, Assumptions and Preliminary Results

7.1 Notation

Define the kernel subspace $\mathcal{N} \equiv \{f \in L_2(X) : T^* f = 0\}$ of the operator $T^* f(z) := E[f(X)|\, Z = z]$. Let $Ts(x) := E[s(Z)|\, X = x]$ denote the adjoint operator of T^* and let $\mathcal{R}(T) := \{f \in L_2(X) : \exists s \in L_2(Z), Ts = f\}$ its range. For a subspace V, $V^\perp$, $\overline{V}$ and $P_{\overline{V}}$ denote, respectively, its orthogonal complement, its closure and its orthogonal projection operator. Let $\otimes$ denote Kronecker product and let I_p denote the identity matrix of order p.

Define the Sobolev norm $\|\cdot\|_{\infty,\eta}$ as follows. Define for any vector a of p integers the differential operator $\partial_x^a := \partial^{|a|_1}/\partial x_1^{a_1} \ldots \partial x_p^{a_p}$, where $|a|_1 := \sum_{i=1}^p a_i$. Let $\mathcal{X}$ denote a finite union of convex, bounded subsets of $\mathbb{R}^p$, with non-empty interior. For any smooth function $h : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} := \max_{|a|_1 \leq \underline{\eta}} \sup_{x \in \mathcal{X}} |\partial_x^a h(x)| + \max_{|a|_1 = \underline{\eta}} \sup_{x \neq x'} \frac{|\partial_x^a h(x) - \partial_x^a h(x')|}{|x - x'|^{\eta - \underline{\eta}}}.$$

Let $\mathcal{H}$ denote the parameter space for h, and define the identified set $\mathcal{H}_0 = \{h \in \mathcal{H} : m(X, h) = 0 \text{ a.s.}\}$. The operator $Th(x) := E[h(Z)|\, X = x]$ is estimated by

$$\hat{T}h(x) := \hat{E}[h(Z)|\, X = x] = \sum_{i=1}^n \left(p^{K_n\prime}(x)(P'P)^{-1} p^{K_n}(X_i) \otimes h(Z_i) \right).$$

The operator $\hat{T}$ is considered as an operator from $\mathcal{H}_n$ to $\mathcal{G}_n \subseteq L_2(X)$, where $\mathcal{G}_n$ is the linear span of $\{p^{K_n}(\cdot)\}$. Let $E_n[g(W)]$ denote the sample mean operator, i.e. $E_{n,W}[g(W)] = n^{-1} \sum_i^n g(W_i)$, let $||g||_{n,W}^2 = E_n[|g(W)|^2]$, and let $\langle f, g \rangle_{n,W} = n^{-1} \sum_{i=1}^n f(W_i) g(W_i)$ be the empirical L_2 inner product. We drop the dependence on W for simplicity of notation. Denote by $\hat{T}^*$ the adjoint operator of $\hat{T}$ with respect to the empirical inner product. Simple algebra shows for $p = 1$,

$$\begin{aligned} \left\langle \hat{T}h, g \right\rangle_n &= n^{-1} \sum_{i=1}^n h(Z_i) p^{K_n\prime}(X_i)(P'P)^{-1} \sum_{j=1}^n p^{K_n}(X_j) g(X_j) \\ &= \left\langle h, \hat{T}^* g \right\rangle_n, \end{aligned}$$

so $\hat{T}^* g = P_{\mathcal{H}_n} \hat{E}[g(X)|\, X = \cdot] = P_{\mathcal{H}_n} \hat{T} g$. A similar expression holds for $p > 1$.

With this operator notation, the first-step has the expression (where I denotes the identity operator)

$$\hat{h}_n = \left(\hat{T}^* \hat{T} + \lambda_n I \right)^{-1} \hat{T}^* \hat{X}, \tag{30}$$

where $\hat{X} = \hat{E}[X|\, X = \cdot]$. Similarly, define the Tikhonov approximation of h_0

$$h_{\lambda_n} = A_{\lambda_n}^{-1} T^* X, \tag{31}$$

where $A_{\lambda_n} = T^*T + \lambda_n I$. Define also $B_{\lambda_n} = TT^* + \lambda_n I$. With some abuse of notation, denote the operator norm by

$$\|T\| = \sup_{h \in \mathcal{H}, \|h\| \leq 1} \|Th\|.$$

Let $\mathcal{G} \subseteq L_2(X)$ denote the parameter space for g. An envelop for $\mathcal{G}$ is a function G such that $|g(x)| \leq G(x)$ for all $g \in \mathcal{G}$. Given two functions l, u, a bracket $[l, u]$ is the set of functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. An ε-bracket with respect to $\|\cdot\|$ is a bracket $[l, u]$ with $\|l - u\| \leq \varepsilon$, $\|l\| < \infty$ and $\|u\| < \infty$ (note that u and l not need to be in $\mathcal{G}$). The *covering number with bracketing* $N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|)$ is the minimal number of ε-brackets with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$. Define the bracketing entropy

$$J_{[\cdot]}(\delta, \mathcal{G}, \|\cdot\|) = \int_0^\delta \sqrt{\log N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|)} d\varepsilon$$

Similarly, we define $J_{[\cdot]}(\delta, \mathcal{H}, \|\cdot\|)$. Finally, throughout C denotes a positive constant that may change from expression to expression.

Let $W = (Y, X', Z')'$ be a random vector defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$. For a measurable function f we denote $\mathbb{P}f := \int f d\mathbb{P}$,

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(W_i) \text{ and } \mathbb{G}_n f := \sqrt{n}\left(\mathbb{P}_n f - \mathbb{P} f\right).$$

7.2 Assumptions

The following assumptions are standard in the literature of sieve estimation; see, e.g., Newey (1997), Chen (2007), Santos (2011), and Chen and Pouzo (2012).

Assumption A1: (i) $\{Y_i, X_i, Z_i\}_{i=1}^n$ is an *iid* sample, satisfying (1) with $E[\varepsilon|\, Z] = 0$ a.s and $E[Y^2] < \infty$; (ii) X has a compact support with $E[|X|^2] < \infty$; (iii) Z has a compact support; (iv) the densities of X and Z are bounded and bounded away from zero.

Assumption A2: (i) The eigenvalues of $E[p^{K_n}(X)p^{K_n}(X)']$ are bounded above and away from zero; (ii) $\max_{1\leq k\leq K_n} \|p_k\| \leq C$ and $\xi_{n,p}^2 K_n = o(n)$, for $\xi_{n,p} = \sup_x \left|p^{K_n}(x)\right|$; (iii) there is $\pi_{n,p}(h)$ such that $\sup_{h\in\mathcal{H}} \left\|E[h(Z)|\, X = \cdot] - \pi_{n,p}'(h)p^{K_n}(\cdot)\right\| = O(K_n^{-\alpha_T})$; (iv) there is a finite constant C, such that $\sup_{h\in\mathcal{H},\|h\|\leq 1} |h(Z) - E[h(Z)|\, X]| \leq \rho_{n,p}(Z, X)$ with $E[|\rho_{n,p}(Z, X)|^2 \big|\, X] \leq C$.

Assumption A3: (i) The eigenvalues of $E[q^{J_n}(Z)q^{J_n}(Z)']$ are bounded above and away from zero; (ii) there is a sequence of closed subsets satisfying $\mathcal{H}_j \subseteq \mathcal{H}_{j+1} \subseteq \mathcal{H}$, $\mathcal{H}$ is closed, bounded and convex, $h_0 \in \mathcal{H}_0$, and there is a $\Pi_n(h_0) \in \mathcal{H}_n$ such that $\|\Pi_n(h_0) - h_0\| = o(1)$; (iii) $\sup_{h\in\mathcal{H}_n} \left|\|h\|_n^2 - \|h\|^2\right| = o_P(1)$; (iv) $\lambda_n \downarrow 0$ and $\max\{\|\Pi_n(h_0) - h_0\|^2, c_{n,T}^2\} = o(\lambda_n)$, where $c_{n,T} = \sqrt{K_n/n} + K_n^{-\alpha_T}$; (v) A_{λ_n} is non-singular.

Assumption A4: (i) $h_0 \in \mathcal{R}((T^*T)^{\alpha_h/2})$ and $g_0 \in \mathcal{R}((TT^*)^{\alpha_g/2})$, $\alpha_h, \alpha_g > 0$; (ii) $\max_{1\leq j\leq J_n} \|q_j\| \leq C$ and $\xi_{n,j}^2 J_n = o(n)$, for $\xi_{n,j} = \sup_z \left|q^{J_n}(z)\right|$; (iii) $\sup_{g\in\mathcal{G}} \left\|E[g(X)|\, Z = \cdot] - \pi_{n,q}'(g)q^{J_n}(\cdot)\right\| = O(J_n^{-\alpha_{T^*}})$ for some $\pi_{n,q}(g)$; (iv) $\sup_{g\in\mathcal{G},\|g\|\leq 1} |g(X) - E[g(X)|\, Z]| \leq \rho_{n,q}(Z, X)$ with $E[|\rho_{n,q}(Z, X)|^2 \big|\, Z] \leq C$; (v) $\lambda_n^{-1} c_n = o(1)$, where $c_n = c_{n,T} + c_{n,T^*}$ and $c_{n,T^*} = \sqrt{J_n/n} + J_n^{-\alpha_{T^*}}$; (vi) B_{λ_n} is non-singular.

Assumption A5: (i) $E[U^2 | Z] < C$ a.s.; (ii) $N_{[\cdot]}(\delta, \mathcal{G}, \|\cdot\|) < \infty$ and $J_{[\cdot]}(\delta, \mathcal{H}, \|\cdot\|) < \infty$ for some $\delta > 0$, and $\mathcal{G}$ and $\mathcal{H}$ have squared integrable envelopes.

Assumption A6: (i) $\lambda_n^{-1} c_n = o(n^{-1/4})$; (ii) $\sqrt{n}\lambda_n^{\min(\alpha_h,2)} = o(1)$ and $\sqrt{n}c_n\lambda_n^{\min(\alpha_h-1,1)} = o(1)$; (iii) $h_0 \in \mathcal{R}(T^*)$, $E\left[|X - h_0(Z)|^4 \middle| X\right]$ is bounded and $Var[h_0(Z)| X]$ is bounded and bounded away from zero; and (iv) $E[U| Z] = 0$ a.s.

For regression splines $\xi_{n,p}^2 = O(K_n)$, and hence A2(ii) requires $K_n^2/n \to 0$, see Newey (1997). Assumptions A2(iii-iv) are satisfied if $\sup_{h\in\mathcal{H}} \|Th\|_{\infty,\eta_h} < \infty$ with $\alpha_T = \eta_h/q$. Assumption A3(iii) holds under mild conditions if for example $\sup_{h\in\mathcal{H}} \|h\| < C$. Assumption A4(i) is a regularity condition that is well discussed in the literature, see e.g. Florens, Johannes and Van Bellegem (2011). A sufficient condition for Assumption A5(ii) is that for some $\eta_h > q/2$ and $\eta_g > p/2$ we have $\sup_{h\in\mathcal{H}} \|h\|_{\infty,\eta_h} < \infty$ and $\sup_{g\in\mathcal{G}} \|g\|_{\infty,\eta_g} < \infty$; see Theorems 2.7.11 and 2.7.1 in van der Vaart and Wellner (1996). Assumptions A6 is standard.

7.3 Preliminary Results

In all the preliminary results Assumptions 1-3 in the text are assumed to hold.

Lemma A1: Let Assumptions A1-A3 hold. Then, $\left\|\hat{h}_n - h_0\right\| = o_P(1)$.
Proof of Lemma A1: We proceed to verify the conditions of Theorem A.1 in Chen and Pouzo (2012). Recall $\mathcal{H}_0 = \{h \in \mathcal{H} : m(X, h) = 0 \text{ a.s.}\}$. By Assumption A3, $\mathcal{H}_0$ is non-empty. The penalty function $P(h) = ||h||^2$ is strictly convex and continuous and $||m(\cdot; h)||^2$ is convex and continuous. Their Assumption 3.1(i) trivially holds since $W = I_p$. Their Assumption 3.1(iii) is A3(i-ii). Their Assumption 3.1(iv) follows from A3(ii) since

$$||m(\cdot; \Pi_n(h_0))||^2 \leq \|\Pi_n(h_0) - h_0\|^2 = o(1).$$

To verify their Assumption 3.2(c) we need to check

$$\sup_{h\in\mathcal{H}_n} \left| \|h\|_n^2 - \|h\|^2 \right| = o_P(1) \tag{32}$$

and

$$\left| \|\Pi_n(h_0)\|^2 - \|h_0\|^2 \right| = o(1).$$

The last equality follows because $\left| \|\Pi_n(h_0)\|^2 - \|h_0\|^2 \right| \leq C \|\Pi_n(h_0) - h_0\| = o(1)$. Condition (32) is our Assumption A3(iii). Assumption 3.3 in Chen and Pouzo (2012) follows from their Lemma C.2 and our Assumption A2. Assumption 3.4 in Chen and Pouzo (2012) is satisfied for the L_2 norm. Finally, Assumption A3(iv) completes the conditions of Theorem A.1 in Chen and Pouzo (2012), and hence implies that $\left\|\hat{h}_n - h_0\right\| = o_P(1)$. ■

Lemma A2: Let Assumptions A1-A4 hold. Then, $\left\|\hat{h}_n - h_0\right\| = O_P(\lambda_n^{\min(\alpha_h,2)} + \lambda_n^{-1} c_n)$ and $\|\hat{g}_n - g_0\| = o_P(\lambda_n^{\min(\alpha_g,2)} + \lambda_n^{-1} c_n)$.

Proof of Lemma A2: For simplicity of exposition we consider the case $p = q = 1$. The proof for $p > 1$ or $q > 1$ follows the same steps. By the triangle inequality, with h_{λ_n} defined in (31),

$$\left\|\hat{h}_n - h_0\right\| \leq \left\|\hat{h}_n - h_{\lambda_n}\right\| + \left\|h_{\lambda_n} - h_0\right\|.$$

Under $h_0 \in \mathcal{R}((T^*T)^{\alpha_h/2})$, Lemma A1(1) in Florens, Johannes and Van Bellegem (2011) yields

$$\left\|h_{\lambda_n} - h_0\right\| = O(\lambda_n^{\min(\alpha_h,2)}). \tag{33}$$

With some abuse of notation, denote $\hat{A}_{\lambda_n} = \hat{T}^*\hat{T} + \lambda_n I$. Then, arguing as in Proposition 3.14 of Carrasco, Florens and Renault (2006), it is shown that

$$\hat{h}_n - h_{\lambda_n} = \hat{A}_{\lambda_n}^{-1}\hat{T}^*(\hat{X} - \hat{T}h_0) + \hat{A}_{\lambda_n}^{-1}(\hat{T}^*\hat{T} - T^*T)(h_{\lambda_n} - h_0), \tag{34}$$

and thus,

$$\left\|\hat{h}_n - h_{\lambda_n}\right\| \leq \left\|\hat{A}_{\lambda_n}^{-1}\right\| \left\|\hat{T}^*(\hat{X} - \hat{T}h_0)\right\| + \left\|\hat{A}_{\lambda_n}^{-1}\right\| \left\|\hat{T}^*\hat{T} - T^*T\right\| \left\|h_{\lambda_n} - h_0\right\|. \tag{35}$$

As in Carrasco, Florens and Renault (2006),

$$\left\|\hat{A}_{\lambda_n}^{-1}\right\| = O_P(\lambda_n^{-1}).$$

Since $\hat{T}^*$ is a bounded operator

$$\begin{aligned} \left\|\hat{T}^*(\hat{X} - \hat{T}h_0)\right\| &= O_P\left(\left\|(\hat{X} - \hat{T}h_0)\right\|\right) \\ &= O_P\left(c_{n,T}\right), \end{aligned}$$

where recall $c_{n,T} = K_n/n + K_n^{-2\alpha_T}$, and where the second equality follows from an application of Theorem 1 in Newey (1997) with $y = x - h_0(z)$ there. Note that Assumption 3 and Assumption A2(iv) imply that $Var[y|\, X]$ is bounded (which is required in Assumption 1 in Newey (1997)). Also note that the supremum bound in Assumption 3 in Newey (1997) can be replaced by our L_2-bound in Assumption A2(iii) when the goal is to obtain L_2-rates.

On the other hand,

$$\left\|\hat{T}^*\hat{T} - T^*T\right\| \leq O_P\left(\left\|\hat{T}^* - T^*\right\|\right) + O_P\left(\left\|\hat{T} - T\right\|\right) \tag{36}$$

and

$$\begin{aligned} \left\|\hat{T}^* - T^*\right\| &\leq \left\|P_{\mathcal{H}_n}\right\| \left\|\hat{T} - T\right\| + \left\|P_{\mathcal{H}_n} - T^*\right\| \\ &= O_P\left(\left\|\hat{T} - T\right\|\right) + O_P(c_{n,T^*}). \end{aligned} \tag{37}$$

We now proceed to establish rates for $\left\|\hat{T} - T\right\|$. As in Newey (1997), we can assume without loss of generality that $E[q^{J_n}(Z)q^{J_n}(Z)']$ is the identity matrix. Then, by the triangle inequality,

$$\begin{aligned} \left\|\hat{T} - T\right\| &= \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left\|\hat{T}h - Th\right\| \\ &\leq \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left\|\hat{T}h - \pi_{n,p}(h)p^{K_n}(\cdot)\right\| + \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left\|E[h(Z)|\, X = \cdot] - \pi_{n,p}(h)p^{K_n}(\cdot)\right\| \\ &\leq \sup_{h \in \mathcal{H}, \|h\| \leq 1} \left\|\hat{\pi}_{n,p}(h) - \pi_{n,p}(h)\right\| + O(K_n^{-\alpha_T}), \end{aligned}$$

where

$$\hat{\pi}_{n,p}(h) = (P'P)^{-1}\sum_{i=1}^{n} p^{K_n}(X_i)h(Z_i).$$

Write

$$\hat{\pi}_{n,p}(h) - \pi_{n,p}(h) = Q_{2n}^{-1}P'\varepsilon_h/n + Q_{2n}^{-1}P'(G_h - P\pi_{n,p}(h))/n,$$

where $\varepsilon_h = H - G_h$, $H = (h(Z_1), ..., h(Z_n))'$, and $G_h = (Th(X_1), ..., Th(X_n))'$. Similarly to the proof of Theorem 1 in Newey (1997), it is shown that

$$\sup_{h\in\mathcal{H},\|h\|\leq 1} \left\|Q_{2n}^{-1}P'\varepsilon_h/n\right\|^2 = O_P(K_n/n),$$

where we use Assumption A2(iv) to show that

$$\sup_{h\in\mathcal{H},\|h\|\leq 1} E[\varepsilon_h\varepsilon_h'|\, X] \leq CI_n.$$

That is,

$$\begin{aligned}
\sup_{h\in\mathcal{H},\|h\|\leq 1} E\left[\left|Q_{2n}^{-1/2}P'\varepsilon_h/n\right|^2 \middle|\, X\right] &= \sup_{h\in\mathcal{H},\|h\|\leq 1} E\left[\varepsilon_h P(P'P)^{-1}P'\varepsilon_h \middle|\, X\right]/n \\
&= \sup_{h\in\mathcal{H},\|h\|\leq 1} E\left[tr\{P(P'P)^{-1}P'\varepsilon_h\varepsilon_h'\} \middle|\, X\right]/n \\
&= \sup_{h\in\mathcal{H},\|h\|\leq 1} tr\{P(P'P)^{-1}P'E[\varepsilon_h\varepsilon_h'|\, X]\}/n \\
&\leq Ctr\{P(P'P)^{-1}P'\}/n \\
&\leq CK_n/n
\end{aligned}$$

Similarly, by A2(iii)

$$\sup_{h\in\mathcal{H},\|h\|\leq 1} \left\|Q_{2n}^{-1}P'(G_h - P\pi_{n,p}(h))/n\right\| = O_P(K_n^{-\alpha_T}).$$

Then, conclude $\left\|\hat{T} - T\right\| = O_P(c_{n,T})$, $\left\|\hat{T}^*\hat{T} - T^*T\right\| = O_P(c_n)$, where $c_n = c_{n,T} + c_{n,T^*}$, and by (35), (36) and (37)

$$\left\|\hat{h}_n - h_{\lambda_n}\right\| = O_P\left(\lambda_n^{-1}c_n\right).$$

The proof for $\hat{g}_n$ is the same and hence omitted. ■

Define the classes

$$\mathcal{F} = \{f(y,x,z) = h(z)(y - x'\beta_0) : h \in \mathcal{H}\}.$$

and

$$\mathcal{G} = \{g(y,x,z) = h(z)x : h \in \mathcal{H}\}.$$

Lemma A3:

(i) Assume $0 < E[|X|^2] < C$. Then, $N_{[\cdot]}(\epsilon, \mathcal{G}, \|\cdot\|_1) \leq N_{[\cdot]}(\epsilon/\left\|X\right\|_2, \mathcal{H}, \|\cdot\|_2)$.

(ii) Assume $Var[Y - X'\beta_0 | Z]$ is bounded. Then, $J_{[\cdot]}(\delta, \mathcal{F}, \|\cdot\|) < \infty$ if $J_{[\cdot]}(\delta, \mathcal{H}, \|\cdot\|) < \infty$ for some $\delta > 0$.

(iii) $N_{[\cdot]}(\epsilon, \mathcal{H}\cdot\mathcal{G}, \|\cdot\|_1) \leq N_{[\cdot]}(C\epsilon, \mathcal{H}, \|\cdot\|_2) \times N_{[\cdot]}(C\epsilon, \mathcal{G}, \|\cdot\|_2)$.

Proof of Lemma A3: (i) Let $[l_j(Z)X, u_j(Z)X]$ be an $\epsilon/E[|x|^2]$ bracket for $\mathcal{H}$. Then, by Cauchy-Schwartz inequality

$$\begin{aligned} \|l_j(Z)X - u_j(Z)X\|_1 &\leq \|l_j(Z) - u_j(Z)\| \, \|X\| \\ &\leq \epsilon. \end{aligned}$$

This shows (i). The proof of (ii) is analogous, and follows from

$$\|l_j(Z)U - u_j(Z)U\| \leq C \|l_j(Z) - u_j(Z)\| \leq C\epsilon,$$

where C is such that $Var[Y - X'\beta_0 | Z] < C$ a.s. The proof of (iii) is standard and hence omitted. ■

8 Appendix B: Proofs of Main Results

Proof of Lemma 2.1: The $n^{1/2}$-estimability of the OLIVA implies the $n^{1/2}$-estimability of the vector-valued functional

$$E[Xg(X)],$$

which in turn implies that of the functional

$$E[X_j g(X)],$$

for each component X_j of X (i.e. $X = (X_1, ..., X_p)'$). By Lemma 4.1 in Severini and Tripathi (2012), the latter implies existence of $h_j \in L_2(Z)$ such that

$$E[h_j(Z)| X] = X_j \text{ a.s.}$$

This implies Assumption 3 with $h(Z) = (h_1(Z), ..., h_p(Z))'$. ■

Proof of Proposition 2.2: We shall show that for any $h(Z) \in L_2(Z)$ such that

$$E[h(Z)| X] = X \text{ a.s.}$$

the parameter $\beta = E[h(Z)X']^{-1}E[h(Z)Y]$ is uniquely defined. First, it is straightforward to show that for any such h, $E[h(Z)X']^{-1} = E[XX']^{-1}$. Second, we can substitute $Y = g_0(X) + P_{\mathcal{N}}g(X) + \varepsilon$, where recall $\mathcal{N} \equiv \{f \in L_2(X) : T^*f = 0\}$ and $T^*f(z) := E[f(X)| Z = z]$. Note that for all h, $E[h(Z)P_{\mathcal{N}}g(X)] = 0$, so that

$$\begin{aligned} E[h(Z)Y] &= E[h(Z)g_0(X)] \\ &= E[Xg_0(X)], \end{aligned}$$

for all h satisfying $E[h(Z)|\, X] = X$ a.s. $\blacksquare$

Proof of Proposition 3.3: We shall show that under the conditions of the proposition there exists a $h(Z) \in L_2(Z)$ such that

$$E[h(Z)|\, X] = X \text{ a.s.}$$

Denote $\bar{\pi} = E[\pi(Z)]$. For a binary X, and since $0 < \bar{\pi} < 1$, the last display is equivalent to the system

$$E[Xh(Z)] = \bar{\pi} \text{ and } E[(1 - X)h(Z)] = 0,$$

or

$$E[h(Z)] = \bar{\pi} \text{ and } E[\pi(Z)h(Z)] = \bar{\pi}.$$

Each equation from the last display defines a hyperplane in h. Since $\pi(Z)$ is not constant, the normal vectors 1 and $\pi(Z)$ are linearly independent (not proportional). Hence, the two hyperplanes have an non-empty intersection, showing that there is at least one h satisfying $E[h(Z)|\, X] = X$ a.s.

Moreover, by Theorem 2, pg. 65, in Luenberger (1997) the minimum norm solution is the linear combination of 1 and $\pi(Z)$ that satisfies the linear constraints, that is, $h_0(Z) = \alpha + \gamma\pi(Z)$ such that α and γ satisfy the 2×2 system

$$\begin{cases} \alpha + \gamma\bar{\pi} = \bar{\pi} \\ \alpha\bar{\pi} + \gamma E[\pi^2(Z)] = \bar{\pi}. \end{cases}$$

Note that this system has a unique solution, since the determinant of the coefficient matrix is $Var(\pi(Z)) > 0$. Then, the unique solution is given by

$$\begin{aligned} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} &= \begin{bmatrix} 1 & \bar{\pi} \\ \bar{\pi} & E[\pi^2(Z)] \end{bmatrix}^{-1} \begin{bmatrix} \bar{\pi} \\ \bar{\pi} \end{bmatrix} \\ &= \begin{bmatrix} \bar{\pi}\left(1 - \frac{\bar{\pi}(1-\bar{\pi})}{var(\pi(Z))}\right) \\ \frac{\bar{\pi}(1-\bar{\pi})}{var(\pi(Z))} \end{bmatrix}. \end{aligned}$$

$\blacksquare$

Proof of Proposition 3.1: Assume without loss of generality that X is scalar and note that, by Engl, Hanke and Neubauer (1996), $h_1(Z) = h_0(Z) + h_\perp(Z)$, with $Cov(h_0(Z), h_\perp(Z)) = 0$ (note $E[h_0(Z)h_\perp(Z)] = 0$ and $E[h_\perp(Z)] = 0$). Thus, since $E[h_0(Z)|\, X] = X$ and $E[h_1(Z)|\, X] = X$, then $E[h_\perp(Z)|\, X] = 0$ a.s., and hence

$$0 = Cov(X, h_\perp(Z)) = \alpha_1 Var(h_\perp(Z)),$$

and hence, if $h_1 \neq h_0$ (i.e. $Var(h_\perp(Z)) > 0$) then $\alpha_1 = 0$. $\blacksquare$

Proof of Theorem 3.2: Write

$$\begin{aligned} \hat{\beta} &= \left(E_n\left[\hat{h}_n(Z_i)X_i'\right]\right)^{-1}\left(E_n\left[\hat{h}_n(Z_i)Y_i\right]\right) \\ &= \beta_0 + \left(E_n\left[\hat{h}_n(Z_i)X_i'\right]\right)^{-1}\left(E_n\left[\hat{h}_n(Z_i)U_i\right]\right). \end{aligned}$$

Note that

$$E_n\left[\hat{h}_n(Z_i)X_i'\right] = E_n\left[h_0(Z_i)X_i'\right] + o_P(1)$$
$$= E\left[h_0(Z_i)X_i'\right] + o_P(1), \tag{38}$$

where the first equality follows from Lemma A3(i), Lemma A1, Assumption A5 and $\hat{h}_n \in \mathcal{H}$ by an application of a Glivenko-Cantelli´s argument, and the second equality follows from the Law of Large Numbers. The same arguments show that $E_n\left[\hat{h}_n(Z_i)U_i\right] = o_P(1)$. Thus, $\hat{\beta}$ is consistent for β_0.

Likewise, Lemma A3(ii), Lemma A1, Assumption A5(ii) and $\hat{h}_n \in \mathcal{H}$, yields for $\hat{f} = \hat{h}_n(Z_i)U_i$ and $f_0 = h_0(Z_i)U_i$,

$$\mathbb{G}_n\hat{f} = \mathbb{G}_n f_0 + o_P(1),$$

since the class $\mathcal{F}$ is a Donsker class, see Theorem 2.5.6 in van der Vaart and Wellner (1996). Then,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \left(E\left[h_0(Z_i)X_i'\right] + o_P(1)\right)^{-1}\left(\sqrt{n}E_n\left[h_0(Z_i)U_i\right] + \sqrt{n}\mathbb{P}\left[\left\{\hat{h}_n(Z_i) - h_0(Z_i)\right\}U_i\right]\right). \tag{39}$$

We investigate the second term, which with the notation $\langle h_1, h_2\rangle = E[h_1(Z)h_2(Z)]$ can be written as

$$\sqrt{n}\mathbb{P}\left[\left\{\hat{h}_n(Z_i) - h_0(Z_i)\right\}U_i\right] = \sqrt{n}\left\langle \hat{h}_n - h_0, u\right\rangle$$

where $u(z) = E[U|\, Z = z]$ is in $L_2(Z)$ by A5(i).

From the proof of Lemma A2, and in particular (33) and (34), and Assumption A6(ii),

$$\begin{aligned}
\sqrt{n}\left\langle \hat{h}_n - h_0, u\right\rangle &= \sqrt{n}\left\langle \hat{h}_n - h_{\lambda_n}, u\right\rangle + \sqrt{n}\left\langle h_{\lambda_n} - h_0, u\right\rangle \\
&= \sqrt{n}\left\langle \hat{A}_{\lambda_n}^{-1}\hat{T}^*(\hat{X} - \hat{T}h_0), u\right\rangle + O_P\left(\sqrt{n}c_n\lambda_n^{\min(\alpha_h - 1, 1)}\right) + O\left(\sqrt{n}\lambda_n^{\min(\alpha_h, 2)}\right) \\
&= \sqrt{n}\left\langle \hat{A}_{\lambda_n}^{-1}\hat{T}^*(\hat{X} - \hat{T}h_0), u\right\rangle + o_P(1).
\end{aligned}$$

Next, we write

$$\begin{aligned}
\sqrt{n}\left\langle \hat{A}_{\lambda_n}^{-1}\hat{T}^*(\hat{X} - \hat{T}h_0), u\right\rangle &= \sqrt{n}\left\langle A_{\lambda_n}^{-1}T^*(\hat{X} - \hat{T}h_0), u\right\rangle \\
&+ \sqrt{n}\left\langle\left(\hat{A}_{\lambda_n}^{-1} - A_{\lambda_n}^{-1}\right)T^*(\hat{X} - \hat{T}h_0), u\right\rangle \\
&+ \sqrt{n}\left\langle A_{\lambda_n}^{-1}\left(\hat{T}^* - T^*\right)(\hat{T}X - \hat{T}h_0), u\right\rangle \\
&+ \sqrt{n}\left\langle\left(\hat{A}_{\lambda_n}^{-1} - A_{\lambda_n}^{-1}\right)\left(\hat{T}^* - T^*\right)(\hat{X} - \hat{T}h_0), u\right\rangle \\
&\equiv C_{1n} + C_{2n} + C_{3n} + C_{4n}.
\end{aligned}$$

From the simple equality $B^{-1} - C^{-1} = B^{-1}(C - B)C^{-1}$ we obtain $\hat{A}_{\lambda_n}^{-1} - A_{\lambda_n}^{-1} = \hat{A}_{\lambda_n}^{-1}\left(T^*T - \hat{T}^*\hat{T}\right)A_{\lambda_n}^{-1}$, and from this and Lemma A2,

$$\begin{aligned}
|C_{4n}| &= O_P(\sqrt{n}\lambda_n^{-2}c_n^3) = o_P(1), \text{ by A6(i);} \\
|C_{3n}| &= O_P(\sqrt{n}\lambda_n^{-1}c_n^2) = o_P(1), \text{ by A6(i);} \\
|C_{2n}| &= O_P(\sqrt{n}\lambda_n^{-2}c_n^2) = o_P(1), \text{ by A6(i).}
\end{aligned}$$

To analyze the term C_{1n} we use Theorem 3 in Newey (1997) after writing

$$C_{1n} = \sqrt{n}\left\langle \hat{T}\varphi, v_n \right\rangle,$$

where $\varphi = X - h_0$ and

$$v_n = TA_{\lambda_n}^{-1}u. \tag{40}$$

Note that

$$u = E[Y - \beta_0' X \big| Z] = E[g_0(X) - \beta_0' X \big| Z],$$

and hence $v_n = TA_{\lambda_n}^{-1}T^*\Delta$, for $\Delta(X) = (g_0(X) - \beta_0' X)$.

Assumption A6(iii) implies Assumptions 1 and 4 in Newey (1997). Assumption A2 implies Assumptions 2 and 3 in Newey (1997) (with $d = 0$ there). Note that by Lemma A1(A.4) in Florens, Johannes and Van Bellegem (2011)

$$\|v_n\| \leq \left\|TA_{\lambda_n}^{-1}T^*\right\| \|\Delta\| \leq \|\Delta\| < \infty.$$

Hence, Assumption 7 in Newey (1997) holds with $g_0 = T\varphi$ there. Hence, Theorem 4 in Newey (1997) applies to C_{1n} to conclude from its proof that

$$C_{1n} = -\frac{1}{\sqrt{n}}\sum_i^n v_n(X_i)(h_0(Z_i) - X_i) + o_P(1). \tag{41}$$

We will use that $g_0(X) - \beta_0' X$ is in $\mathcal{R}((TT^*)^{\alpha_g/2})$, $\alpha_g > 0$. Note the identities

$$T\left(T^*T + \lambda_n I\right)^{-1} T^* = \left(TT^* + \lambda_n I\right)^{-1} TT^*$$

and

$$I - \left(TT^* + \lambda_n I\right)^{-1} TT^* = \lambda_n \left(TT^* + \lambda_n I\right)^{-1}.$$

Then,

$$\frac{1}{\sqrt{n}}\sum_i^n v_n(X_i)(h_0(Z_i) - X_i) = \frac{1}{\sqrt{n}}\sum_i^n \left(g_0(X_i) - \beta_0' X_i\right)\left(h_0(Z_i) - X_i\right) + o_P(1), \tag{42}$$

since by Lemma A1(A1) in Florens, Johannes and Van Bellegem (2011), Assumption A4(i) and Assumption 3, it holds

$$\begin{aligned} Var\left(\frac{1}{\sqrt{n}}\sum_i^n \left[v_n(X_i) - \left(g_0(X_i) - \beta_0' X_i\right)\right]\left(h_0(Z_i) - X_i\right)\right) &\leq C\left\|v_n(X_i) - \left(g_0(X_i) - \beta_0' X_i\right)\right\|^2 \\ &= C\left\|\lambda_n \left(TT^* + \lambda_n I\right)^{-1}\left(g_0(X_i) - \beta_0' X_i\right)\right\|^2 \\ &\leq C\lambda_n^{\min(\alpha_g, 2)}. \end{aligned}$$

Thus, from (39), (41) and (42)

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \left(E\left[h_0(Z_i)X_i'\right]\right)^{-1}\sqrt{n}E_n\left[m(W_i, \beta_0, h_0, g_0)\right] + o_P(1).$$

The asymptotic normality then follows from the standard Central Limit Theorem.

We now show the consistency of $\hat{\Sigma} = E_n[\hat{h}_n(Z_i)X_i']^{-1}E_n[\hat{m}_{ni}\hat{m}_{ni}']E_n[\hat{h}_n(Z_i)X_i']^{-1}$. Write, with $m_{0i} = m(W_i, \beta, h_0, g_0)$,

$$E_n[\hat{m}_{ni}\hat{m}_{ni}'] - E_n[m_{0i}m_{0i}'] = E_n[m_{0i}(\hat{m}_{ni}' - m_{0i}')] + E_n[(\hat{m}_{ni} - m_{0i})m_{0i}'] + E_n[(\hat{m}_{ni} - m_{0i})(\hat{m}_{ni} - m_{0i})'] \quad (43)$$

and

$$\hat{m}_{ni} - m_{0i} = (Y - g_0(X_i))\left(\hat{h}_n(Z_i) - h_0(Z_i)\right) - (\hat{g}_n(X_i) - g_0(X_i))\left(\hat{h}_n(Z_i) - X_i\right).$$

By Cauchy-Schwartz inequality and Assumption 2

$$\left| E_n\left[m_{0i}(Y - g_0(X_i))\left(\hat{h}_n(Z_i) - h_0(Z_i)\right)'\right]\right|^2 \leq CE_n\left[\left|\hat{h}_n(Z_i) - h_0(Z_i)\right|^2\right].$$

The class of functions

$$\{|h(z) - h_0|^2 : h \in \mathcal{H}\}$$

is Glivenko-Cantelli under the conditions on $\mathcal{H}$, and thus $E_n\left[\left|\hat{h}_n(Z_i) - h_0(Z_i)\right|^2\right] = o_P(1)$ by Lemma A1. Likewise,

$$\begin{aligned}\left| E_n\left[m_{0i}'(\hat{g}_n(X_i) - g_0(X_i))\left(\hat{h}_n(Z_i) - X_i\right)'\right]\right|^2 &\leq CE_n\left[|\hat{g}_n(X_i) - g_0(X_i)|^2\right] \\ &= o_P(1),\end{aligned}$$

by Assumption A5(ii) and Lemma A1. Other terms in (43) are analyzed similarly, to conclude that they are $o_P(1)$. Together with (38), this implies the consistency of $\hat{\Sigma}$. ■

Proof of Theorem 4.1: We first show that the OLS first-stage estimator $\hat{\alpha} = (\hat{\alpha}_1', \hat{\alpha}_2)'$ of $\alpha_0 = (\alpha_1', \alpha_2)'$ in the regression

$$X_2 = \alpha_1'X_1 + \alpha_2\hat{h}_{2n}(Z) + e,$$

satisfies $\sqrt{n}(\hat{\alpha} - \alpha_0) = O_P(1)$. Note $e = V - \alpha_2(\hat{h}_{2n}(Z) - h_{20}(Z))$, and denote $\hat{h}_n(Z) = (X_1', \hat{h}_{2n}(Z))'$ and $h_0(Z) = (X_1', h_{20}(Z))'$. Then,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = \left(E_n\left[\hat{h}_n'\hat{h}_n'\right]\right)^{-1}\sqrt{n}E_n\left[\hat{h}_n e\right].$$

Lemma A2 and a Glivenko-Cantelli´s argument imply $E_n\left[\hat{h}_n\hat{h}_n'\right] = E_n\left[h_0(Z)h_0'(Z)\right] + o_P(1) = O_P(1)$.

By $\left\|\hat{h}_{2n} - h_{20}\right\| = o_P(n^{-1/4})$, it holds

$$\begin{aligned}\sqrt{n}E_n\left[\hat{h}_n(Z)e\right] &= \sqrt{n}E_n\left[\hat{h}_n(Z)V\right] - \alpha_2\sqrt{n}E_n\left[\hat{h}_n(Z)(\hat{h}_{2n}(Z) - h_{20}(Z))\right] \\ &= \sqrt{n}E_n\left[h_0(Z)V\right] - \alpha_2\sqrt{n}E_n\left[h_0(Z)(\hat{h}_{2n}(Z) - h_{20}(Z))\right] + \sqrt{n}E_n\left[(\hat{h}_n(Z) - h_0(Z))V\right] + o_P(1) \\ &\equiv A_1 - \alpha_2 A_2 + A_3 + o_P(1).\end{aligned}$$

The standard central limit theorem implies $A_1 = O_P(1)$.

An empirical processes argument shows

$$A_2 = \sqrt{n}E\left[h_0(Z)(\hat{h}_{2n}(Z) - h_{20}(Z))\right] + o_P(1).$$

By A6(ii),

$$\begin{aligned}\sqrt{n}E\left[h_0(Z)(\hat{h}_{2n}(Z) - h_{20}(Z))\right] &= \sqrt{n}E\left[h_0(Z)(\hat{h}_{2n}(Z) - h_{\lambda_n}(Z))\right] + \sqrt{n}E\left[h_0(Z)(h_{\lambda_n}(Z) - h_{20}(Z))\right] \\ &= \sqrt{n}E\left[h_0(Z)(\hat{h}_{2n}(Z) - h_{\lambda_n}(Z))\right] + o_P(1).\end{aligned}$$

While (34) and A6(ii) yield

$$\begin{aligned}A_2 &= \sqrt{n}E\left[h_0(Z)\hat{A}_{\lambda_n}^{-1}\hat{T}^*(\hat{X} - \hat{T}h_0)(Z)\right] + o_P(1) \\ &= \sqrt{n}E\left[h_0(Z)A_{\lambda_n}^{-1}T^*(\hat{X} - \hat{T}h_0)(Z)\right] + o_P(1) \\ &\equiv \sqrt{n}E\left[v(Z)(\hat{X} - \hat{T}h_0)(Z)\right] + o_P(1),\end{aligned}$$

where $v(Z) = TA_{\lambda_n}^{-1}h_0(Z)$. By $h_0 \in \mathcal{R}(T^*)$, $h_0 = T^*\psi$ for some ψ with $\|\psi\| < \infty$, then by Lemma A1(A.4) in Florens, Johannes and Van Bellegem (2011)

$$\begin{aligned}\|v\| &\leq \left\|TA_{\lambda_n}^{-1}T^*\right\| \|\psi\| \\ &\leq \|\psi\| < \infty.\end{aligned}$$

Then, by Theorem 3 in Newey (1997), $A_2 = O_P(1)$. A similar argument as for A_2 shows $A_3 = O_P(1)$, because $E[V|\, Z] \in \mathcal{R}(T^*)$. Thus, combining the previous bounds we obtain $\sqrt{n}(\hat{\alpha} - \alpha_0) = O_P(1)$.

We proceed now with second step estimator. Denote $\hat{S} = (X, \hat{V})'$ and $\theta = (\beta', \rho)'$. Let $\hat{\theta}$ denote the OLS of Y on $\hat{S}$. Since, since under the null $\rho = 0$, then

$$\begin{aligned}\hat{\theta} &= \left(E_n\left[\hat{S}\hat{S}'\right]\right)^{-1} E_n\left[\hat{S}Y\right] \\ &= \theta + \left(E_n\left[\hat{S}\hat{S}'\right]\right)^{-1} E_n\left[\hat{S}U\right] \\ &= \theta + \left(E\left[SS'\right]\right)^{-1} E_n\left[SU\right] + \left(E\left[SS'\right]\right)^{-1} E_n\left[(\hat{S} - S)U\right] + o_P(n^{-1/2}) \\ &= \theta + \left(E\left[SS'\right]\right)^{-1} E_n\left[SU\right] + o_P(n^{-1/2}),\end{aligned}$$

where the last equality follows because

$$\begin{aligned}\sqrt{n}E_n\left[(\hat{V} - V)U\right] &= \sqrt{n}(\hat{\alpha} - \alpha_0)'E_n\left[h_0(Z)U\right] + \hat{\alpha}_2\sqrt{n}E_n\left[U(\hat{h}_{2n}(Z) - h_{20}(Z))\right] \\ &= O_P(1) \times o_P(1) + O_P(1) \times o_P(1),\end{aligned}$$

with the term $\sqrt{n}E_n\left[U(\hat{h}_{2n}(Z) - h_{20}(Z))\right]$ being $o_P(1)$ because by A6(iv)

$$\begin{aligned}\sqrt{n}E_n\left[U(\hat{h}_{2n}(Z) - h_{20}(Z))\right] &= \sqrt{n}\mathbb{P}\left[U(\hat{h}_{2n}(Z) - h_{20}(Z))\right] + o_P(1) \\ &= o_P(1).\end{aligned}$$

Thus, the standard asymptotic normality for the OLS estimator applies. ■

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231-263.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Ai, C., and X. Chen (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables", *Journal of Econometrics,* 141, 5-43.

Imbens G., Angrist J., and K. Graddy (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 2000, 499-527.

Babii, A. and Florens, J.P. (2018): "Is Completeness Necessary? Estimation and Inference in Non-identified Linear Models," working paper.

Bierens, H. J. (1982): "Consistent model specification tests", *Journal of Econometrics,* 20**,** 105-134.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV Estimation of Shape-invariant Engel Curves, " *Econometrica*, 75, 1613-1670.

Blundell, R., and J.L. Horowitz (2007): "A Nonparametric Test for Exogeneity, " *Review of Economic Studies*, 74, 1035-1058.

Breunig, C. and Johannes, J. (2016): "Adaptive estimation of functionals in nonparametric instrumental regression," *Econometric Theory*, 32, 612-654.

Campbell, J.Y. (2003): "Consumption-Based Asset Pricing", Chapter 13 in George Constantinides, Milton Harris, and Rene Stulz eds. *Handbook of the Economics of Finance* Vol. IB, North-Holland, Amsterdam, 803-887.

Canay, I., Santos, A. and A. Shaikh (2013): "On the Testability of Identification in Some Non-parametric Models with Endogeneity," *Econometrica*, 81, 2535-2559.

Carrasco, M., J. P. Florens, and E. Renault (2006): "Linear Inverse Problem in Strucutral Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. 6, ed. by J. J. Heckman and E. E. Leamer. Amsterdam: North-Holland, 5633–5751.

Carrasco, M., J. P. Florens, and E. Renault (2014): "Asymptotic normal inference in linear inverse problems," in J. Racine, L. Su & A. Ullah, eds, 'The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics', Oxford University Press, pp. 65-96.

Centorrino, S., Feve, F., and Florens, J.-P. (2017): "AdditiveNonparametric Instrumental Regressions: A Guide to Implementation", *Journal of Econometric Methods*, 6(1), 1-25.

CHAMBERLAIN, G. (1986): "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics*, 34, 305-334.

CHEN (2007): "Large sample sieve estimation of semi-nonparametric models, " in *Handbook of Econometrics*, vol. 6, ed. by J. J. Heckman and E. E. Leamer. Amsterdam: North-Holland, 5549–5632.

CHEN, X., AND D. POUZO (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152, 46-60.

CHEN, X., AND D. POUZO (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Gneralized Residuals," *Econometrica*, 80(1), 277-321.

CHEN, X., AND D. POUZO (2015): "Sieve Quasi Likelihood Ratio Inference on Semi/Nonparametric Conditional Moment Models," *Econometrica* 83(3), 1013-1079.

CHEN, X., AND M. REISS (2011): "On Rate Optimallity for Ill-Posed Inverse Problems in Econometrics," *Econometric Theory*, 27, 497–521.

CHEN, X., AND A. SANTOS (2018): "Overidentification in Regular Models," *Econometrica*, 86, 1771-1817.

CHETVERIKOV, D. AND D. WILHEM (2017): "Nonparametric Instrumental Variable Estimation Under Monotonicity," *Econometrica*, 85(4), 1303–1320.

DAROLLES, S., Y. FAN, J. P. FLORENS, AND E. RENAULT (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

ENGL, H. W., M. HANKE, AND A. NUEBAUER (1996): *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic Publishers.

ESCANCIANO, J. C. (2006): "A Consistent Diagnostic Test for Regression Models Using Projections", *Econometric Theory*, 22, 1030-1051.

ESCANCIANO, J.C. AND W. LI (2013): "On the Identification of Structural Linear Functionals," CEMMAP working paper CWP48/13.

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. (2017): "On Completeness and Consistency in Nonparametric Instrumental Variable Models", *Econometrica,* 85(5), 1629-1644.

GAGLIARDINI, P., AND O. SCAILLET (2012): "Nonparametric Instrumental Variable Estimation of Structural Quantile Effects," *Econometrica*, 80: 1533–1562.

HALL, P., AND J. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

Hansen, L. P. and Singleton, K. J. (1983): "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, 91, 249-265.

Hausman, J.A (1978): "Specification Tests in Econometrics," *Econometrica* 6, 1251-1271.

Horowitz, J. (2007): "Asymptotic normality of a nonparametric instrumental variables estimator," *International Economic Review,* 48, 1329–1349.

Horowitz, J. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), 347–394.

Imbens, G.W. and J.D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

Khan, S. and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 6, 2021-2042.

Lochner, L., and E. Moretti (2015): "Estimating and Testing Models with Many Treatment Levels and Limited Instruments", *The Review of Economics and Statistics*, 97, 387-397.

Luenberger, D. G. (1997): *Optimization by Vector Space Methods*. New York: John Wiley & Sons. 1969 edition.

Newey, W. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica* 58, 809-837.

Newey, W. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

Newey, W. K., and J. Powell (2003): "Instrumental Variables Estimation for Nonparametric Models," *Econometrica*, 71, 1565–1578.

Robinson, P. M. (1976): "Instrumental Variables Estimation of Differential Equations," *Econometrica*, 44, 765-776.

Santos, A. (2011): "Instrumental Variable Methods for Recovering Continuous Linear Functionals," *Journal of Econometrics*, 161(2), 129–146.

Santos, A. (2012): "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, 80(1), 213–275.

Severini, T. A., and G. Tripathi (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theory*, 22(2), 258–278.

Severini, T. A., and G. Tripathi (2012): "Efficency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," *Journal of Econometrics*, 170(2), 491–498.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer Series in Statistics. Springer-Verlag, New York, 1 edn.

Wahba, G. (1990): *Spline Models for Observational Data*. Philadelphia: SIAM.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J.M. (2015): *Introductory Econometrics: A Modern Approach*, 6th Edition, Cengage Eds.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," *Review of Economics and Statistics*, 86, 797-810.