Optimal Linear Instrumental Variables Approximations

Juan Carlos Escanciano*
Indiana University

Wei Li †
North Carolina State University

May 7th, 2018

Abstract

Ordinary least squares provides the optimal linear approximation to the true regression function under misspecification. This paper investigates the Instrumental Variables (IV) version of this problem. The resulting population parameter is called the Optimal Linear IV Approximation (OLIVA). This paper shows that a necessary condition for regular identification of the OLIVA is also sufficient for existence of an IV estimand in a linear IV model. The necessary condition holds for the important case of a binary endogenous treatment, leading also to a LATE interpretation with positive weights. The instrument in the IV estimand is unknown and is estimated in a first step. A Two-Step IV (TSIV) estimator is proposed. We establish the asymptotic normality of a debiased TSIV estimator based on locally robust moments. The TSIV estimator does not require neither completeness nor identification of the instrument. As a by-product of our analysis, we robustify the classical Hausman test for exogeneity against misspecification of the linear model. Monte Carlo simulations suggest excellent finite sample performance for the proposed inferences.

Keywords: Instrumental Variables; Nonparametric Identification; Hausman Test.

JEL classification: C26; C14; C21.

*Department of Economics, Indiana University, 349 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number ECO2014-55858-P.

†Department of Statistics, North Carolina State University, 4217 SAS, 2311 Stinson Dr., Raleigh, NC 27695 – 8203, USA. E-mail: wli35@ncsu.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 even 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, but do not share with OLS the appealing nonparametric interpretation (see Imbens, Angrist and Graddy (2000)). The main goal of our paper is to fill this gap and propose an IV analog to OLS.

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 first 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 the necessary condition for regular identification of the OLIVA is also sufficient for existence of an IV estimand in a linear IV regression. That is, we show that under a minimal condition for regular estimation, it is possible to obtain an IV version of OLS.

The identification result is constructive and leads to a two-step 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 analogous to optimal IV (see, e.g., Robinson (1976) and Newey (1990)), but technically more difficult due to the possible lack of identification of the first step and the first step problem being statistically harder than a nonparametric regression problem. We overcome these difficulties combining two ingredients: a Penalized Sieve Minimum Distance (PSMD) first step estimator of the type discussed in Chen and Pouzo (2012) (to address the lack of identification) and the use of locally robust moments that have zero derivatives with respect to first steps (to obtain asymptotic normality under weak assumptions). The combination of these two ingredients for obtaining asymptotic normality appears to be new in the literature, and is of independent interest.

Locally robust moments in a general GMM setting have been discussed in Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018), including linear functionals of structural functions identified by conditional moment restrictions, such as the OLIVA . These authors provide a general asymptotic theory based on sample splitting. We complement their theory with an asymptotic theory that does not require neither sample splitting nor identification of the first steps. The proposed TSIV estimator has an excellent finite sample performance in simulations, being competitive with the oracle standard IV under linearity of the structural model, while robustifying it otherwise.

An important by-product of our approach is 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. We establish the asymptotic null distribution for the robust Hausman test. 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 IV estimator, allowing for general endogenous variables (continuous, discrete or mixed), more in the spirit of the original Hausman (1978)'s exogeneity test. 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) and Santos (2012), among others. Severini and Tripathi (2006, 2012) discuss regular and irregular identification of linear functionals of the structural function without completeness, and 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. The OLIVA functional was not discuss in neither Severini and Tripathi (2006, 2012) nor Santos (2011). The implementation and asymptotic normality proof for the OLIVA based local robustness can be also applied to the functionals considered in Santos (2011) and to other problems involving linear functionals of structural functions defined by conditional moment restrictions.

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 (2000). 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 resulting linear approximation coefficients. Like in much of this literature, the endogenous variable is binary and the instrument is also binary. In this case, we show that our IV estimator also has a LATE interpretation with non-negative weights; see Section 2.3.

The main 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 the instrument. Importantly, the proposed methods are also robust to misspecification of linear model, sharing the nonparametric interpretation of OLS, but in a setting with endogenous regressors.

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 establish 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. Appendix A presents notation, assumptions and some preliminary results that are needed for the main proofs in Appendix B. Appendix C discusses different implementations of first steps. Appendix D reports tables for simulation results on 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),

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

where henceforth A' denotes the transpose of A. 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 (2) becomes a more difficult issue than in the exogenous case, albeit less difficult than identification and estimation of the structural function g in (1).

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)). The necessary condition for regular identification of β is

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

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

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

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. Conditions for point identification and consistent estimation of g are given in the references on the nonparametric IV literature cited above. 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), and we could adapt these results to obtain asymptotic normality for the OLIVA when g is identified.

Nonparametric identification of g is, however, not necessary for identification of the OLIVA; see also 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. Inference under irregular identification is known to be less stable and non-standard, see Chamberlain (1986), and it is beyond the scope of this paper.

2.2 Regular Identification

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

$$r(z) := E[Y|\, Z = z] = E[g(X)|\, Z = z] := T^* g, \tag{5}$$

where T^* denotes the adjoint operator of T. Let $\mathcal{G}$ denote the parameter space for g, with $g \in \mathcal{G} \subseteq L_2(X)$. Assume $T^* : \mathcal{G} \to L_2(Z)$ 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.[1]

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 $g_0 := \arg\min_{g:r=T^*g} \|g\|$. Correct specification of the model guarantees that g_0 is uniquely defined; see Engl, Hanke and Neubauer (1996). Define $\xi = Y - g_0(X)$, $\Omega(z) = E[\xi^2 \big|\, Z = z]$, and let $\mathcal{S}_Z$ denote the support of Z. We consider the following assumptions.

Assumption 1: The model (1) holds with $E[XX']$ 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 $n^{1/2}$-regularly estimable, then Assumption 3 holds.*

The proof of Lemma 2.1 and other results in the text are gathered in Appendix B. The concept of $n^{1/2}$-regular estimation is defined in e.g. Chamberlain (1986). Another way to state Lemma 2.1 is that Assumption 3 is necessary for the OLIVA to be identified with a finite efficiency bound, i.e. to be regularly identified. Assumption 3 may hold when completeness fails (see Newey and Powell (2003) for discussion of completeness). If Z has discrete support, then Assumption 3 can be tested. We expect that this condition is also testable when Z and X are continuous and the distribution of X given Z is not complete (see Chen and Santos (2015)). When X is binary, Assumption 3 holds under a mild condition, as shown below. More generally, for X discrete, (4) becomes a finite system of equations, which makes the condition more likely to hold, provided the support of Z is large enough relative to that of X.

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

The main observation of this paper is that the necessary condition for $n^{1/2}$-estimability 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.,

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

which is the IV estimand using $h(Z)$ as instruments for X. The following Proposition 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 identified by the IV with instruments $h(Z)$.*

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.*

2.3 LATE Interpretation

As an important example, consider the case where the endogenous variable X is binary, like an endogenous treatment indicator. In this case Assumption 3 is satisfied under a mild condition. Furthermore, a unique minimum norm solution to (4) can be easily characterized (see the proof of Proposition 2.3). Such minimum norm solutions will also play an important role in our implementation of the continuous case as well.

Proposition 2.3 *If X is binary, and the propensity score $p(Z) = E[X|\, Z]$ is not constant, with $0 < E[p(Z)] < 1$, then Assumption 3 holds. Moreover, there exists a unique solution of (4) of the form $h_0(Z) = \alpha + \gamma p(Z)$, and this h_0 is the unique minimum norm solution among all solutions of (4).*

The last part of Proposition 2.3 is particularly important, as it implies that Condition 3 in Imbens and Angrist (1994) holds. This condition states that (i) for all z_1, z_2 in the support of Z, it follows that $p(z_1) \leq p(z_2)$ implies either $h_0(z_1) \leq h_0(z_1)$ or $h_0(z_1) \geq h_0(z_1)$; and (ii) $Cov(X, h_0(Z)) \neq 0$. Both conditions are satisfied by h_0 in Proposition 2.3 (note $Cov(X, h_0(Z)) = Var(X) > 0$). Hence, when other standard assumptions in Imbens and Angrist (1994) are satisfied (Conditions 1 and 2), their Theorem 2 implies that our IV estimator has a LATE interpretation as a weighted average of local average treatment effects with nonnegative weights. More generally, for continuous endogenous variables and continuous instruments our estimator, being an IV estimator, has a LATE interpretation as described in Imbens, Angrist and Graddy (2000).

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

$\{Y_i, X_i, Z_i\}_{i=1}^n$ be an independent and identically distributed (iid) sample of size n satisfying (1). The TSIV estimator follows the steps:

Step 1. Estimate a function h satisfying $E[h(Z)|\, X] = X$ a.s., say $\hat{h}_n$, as defined in (9) below.

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

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n \hat{h}_n(Z_i)X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{h}_n(Z_i)Y_i\right), \tag{6}$$

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 and are continuous. We also analyze below the case of overlapping components and discrete variables.

3.1 First-Step Estimation

To deal with the problem of lack of uniqueness of h, we consider a Tikhonov-type estimator. This approach is commonly used in the inverse problem literature, and is also popular in econometrics, 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. Their results are used below to establish consistency of our TSIV, but they are not applicable to establish asymptotic normality, for which rates are required, due to the possible lack of identification of h.

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. The function h_0 is the minimum norm solution of (4), as in Proposition 2.3. For the case where X is continuous, we propose to estimate h_0 by a 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 $E[h(Z)|\, X = x]$. Then,

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

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

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

$$\hat{h}_n := \arg\min\{||\hat{m}(X;h)||_n^2 + \lambda_n ||h||_n^2 : h \in \mathcal{H}_n\}, \tag{7}$$

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

$$\mathcal{H}_n = \left\{ h : h = \sum_{j=1}^{J_n} a_j q_j(\cdot) \right\} \tag{8}$$

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 $n \to \infty$.

The proposed TSIV estimator uses $\hat{h}_n$ in (7) with $\mathcal{H}_n$ as in (8), and has a simple closed-form expression given as follows. Define $\hat{q}(X) = \hat{E}[q^J(Z)\big|\, X]$, $D_n = E_n[\hat{q}(X)X']$, $Q_{2n} = E_n[q^J(Z)q^J(Z)']$, and

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

Then, the closed form solution to (7) is given by

$$\hat{h}_n(\cdot) = D_n' A_{\lambda_n}^{-1} q^J(\cdot). \tag{9}$$

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 his two-step setting. Appendix C shows the equivalence between Tikhonov-type estimators and minimum norm-type estimators, in the sense that there exists λ_n such that $\hat{h}_n = \tilde{h}_n$, and more importantly, we provide bounds for such λ_n in terms of b_n/a_n. This result is of independent interest. We prefer our implementation, since we only need one tunning parameter rather than two, and data driven methods to choose this parameter are available; see Section 3.3. We could combine this equivalence result with the uniform consistency result in Santos (2011) to show consistency of $\hat{\beta}$, but this would require compactness of the parameter space $\mathcal{H}$ with respect the supremum norm. Alternatively, the general L_2-consistency result for $\hat{h}_n$ in Chen and Pouzo (2012) can be used to establish the consistency of our TSIV estimator under more general conditions on the parameter space, as the following result shows.

Proposition 3.1 *Let Assumptions 1-3 and A1-A3 in Appendix A hold. Then, $\hat{\beta}$ is consistent for β.*

3.2 Second-Step Estimation and Inference

The moments that define the IV estimand are

$$E[(Y - X'\beta)h_0(Z)] = 0.$$

These moments are not locally robust in the sense of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018), meaning that the derivative of the moments with respect to h_0 is in general not zero, and hence the first step will have an impact in the asymptotic distribution of the TSIV. Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018) derive locally robust moments for linear functionals of structural functions defined by conditional moment restrictions, which include the OLIVA as special case. The locally robust moments are given by

$$m(W, \beta, h, g) = (Y - X'\beta)h(Z) - (g(X) - X'\beta)(h(Z) - X).$$

These moments are also doubly robust in the sense of Scharfstein, Rotnitzky and Robins (1999) and Robins, Rotnitzky and van der Laan (2000). However, the double robustness here holds even when first-steps are not identified, since for *any* $\bar{h}$ satisfying $E[\bar{h}(Z)\big|\, X] = X$ a.s. and *any* $\bar{g}$ satisfying $E[Y - \bar{g}(X)|\, Z] = 0$ a.s., and for all h and g,

$$E[m(W, \beta, h, \bar{g})] = 0 \text{ and } E[m(W, \beta, \bar{h}, g)] = 0. \tag{10}$$

That is, the moments continue to hold if one first step component is in the identified set and the identified set is not singleton. Estimators based on doubly robust moments have several advantages in terms of bias and mean squared error finite-sample performance, as illustrated in the context of treatment effects by Bang and Robins (2005) and Firpo and Rothe (2016).

Doubly robust moments can be also used to derive parametric inference for β that is robust to misspecification of g or h. That is, if g_θ and h_δ are parametric specifications of g and h, respectively, we only need either g_θ or h_δ to be correctly specified for consistent estimation of β with the doubly robust moments. For example, if Y and X are binary we could specify g_θ and the propensity score as parametric Probit models, and estimate β as the solution of the doubly robust moments with plugged in parametric estimates of g_θ and $h_\delta(z) = \alpha + \gamma p_\eta(z)$, $\delta = (\alpha, \gamma, \eta')'$. More generally, we can use standard GMM inference for any parametric estimates based on doubly robust moments. Since parametric inference is standard, we leave the details to the reader, and rather focus on the more complicated semiparametric case.

In the semiparametric two-step setting the locally robust or doubly robust moment leads to

$$\begin{aligned}\beta &= E[h(Z)X']^{-1}E[h(Z)Y] - E[h(Z)X']^{-1}E[(g(X) - X'\beta)(h(Z) - X)] \\ &\equiv \beta_{IV} - b,\end{aligned}$$

which suggests the debiased TSIV estimator

$$\tilde{\beta} = \hat{\beta} - \hat{b},$$

where $\hat{b} = E_n[\hat{h}_n(Z_i)X_i']^{-1}E_n[(\hat{g}_n(X_i) - X'\hat{\beta})(\hat{h}_n(Z_i) - X_i)]$ and $\hat{g}_n(\cdot)$ denotes a PSMD estimator of g_0 given by

$$\hat{g}_n(\cdot) = G_n' B_{\lambda_n}^{-1} p^K(\cdot), \tag{11}$$

with $G_n = E_n[\hat{p}(Z)Y]$, $\hat{p}(Z) = \hat{E}[p^K(X)\big|\, Z]$, $\hat{E}[g(X)|\, Z = z] = q^{J_n}{}'(z)(Q'Q)^{-1}\sum_{i=1}^n q^{J_n}(Z_i)g(X_i)$, $Q = [q^{J_n}(Z_1), ..., q^{J_n}(Z_n)]'$, $P_{2n} = E_n[p^K(X)p^K(X)']$, and $B_{\lambda_n} = E_n[\hat{p}(Z)\hat{p}(Z)'] + \lambda_n P_{2n}$. For ease of presentation, we use the same penalization parameter λ_n for $\hat{h}_n$ and $\hat{g}_n$, although it is possible to use two different parameters in the theory. Similarly, although we do not make it explicit in the notation, we will use different tuning parameters K_n and J_n for estimating $\hat{h}_n$ or $\hat{g}_n$, see Section 3.3 for issues of implementation.

The following result establishes the asymptotic normality of $\tilde{\beta}$. Its proof relies on new L_2-rates of convergence for $\hat{h}_n$ and $\hat{g}_n$ under partial identification of h and g. Santos (2011) obtained related rates but for a weak norm, which are not enough for our asymptotic normality. Although we focus on PSMD, the asymptotic normality proof applies to any first step estimators satisfying $\left\|\hat{h}_n - h_0\right\| = o_P(n^{-1/4})$ and $\|\hat{g}_n - g_0\| = o_P(n^{-1/4})$ under some entropy conditions for the parameter spaces. This genericity of the proof holds true by virtue of the double robustness of moments. These conditions can be further weakened to a simple rate condition (without entropy conditions) by means of sample splitting, as shown in Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018).

Theorem 3.2 *Let Assumptions 1-3 above and Assumptions A1-A5 in the Appendix A hold. Then,*

$$\sqrt{n}(\tilde{\beta} - \beta) \longrightarrow_d N(0, V),$$

where $V = E[XX']^{-1}E[ss']E[XX']^{-1}$ *and* $s = h_0(Z)U - g_0(X)\,(h_0(Z) - X)$. *Furthermore, a consistent estimator for* V *is given by*

$$\hat{V} = E_n[X_i X_i']^{-1}E_n[\hat{s}_{ni}\hat{s}_{ni}']E_n[X_i X_i']^{-1}, \tag{12}$$

where

$$\hat{s}_{ni} = \hat{h}_n(Z_i)\hat{U} - \hat{g}_n(X_i)(\hat{h}_n(Z_i) - X_i)$$

and $\hat{U} = Y - X'\tilde{\beta}$.

Remark 3.1 *When* h *is identified, and* λ_n *is set to zero, the TSIV becomes asymptotically first order equivalent to* $\tilde{\beta}$, *and hence asymptotically doubly robust. This follows from Escanciano and Song (2010), whose results imply that under these conditions* $\sqrt{n}\hat{b} = o_P(1)$ *(note* g *need not be identified).*

Theorem 3.2 can be then used to construct confidence regions for β and testing hypotheses about β following standard procedures.

3.3 Implementation

For implementation one has to choose the basis $\{p^{K_n}(X), q^{J_n}(Z)\}$ and the tuning parameters $\{K_n, J_n, \lambda_n\}$. The theory for estimating h_0 requires that $K_n \geq J_n$ (for A_{λ_n} to be invertible). In the simulations we

study rules of the form $K_n = cJ_n$ for several values of c such as 2 or 3. In practice, we recommend choosing first J_n, then set $K_n = 2J_n$ and choose λ_n by Generalized Cross-validation (cf. Wahba (1990)), $\lambda_n = \arg\min_{\lambda>0} GCV_n(\lambda)$, as follows. Note that

$$\hat{\beta} = \left(D_n' A_{\lambda_n}^{-1} Q' \mathbf{X}\right)^{-1} D_n' A_{\lambda_n}^{-1} Q' \mathbf{Y}, \tag{13}$$

where $\mathbf{X} = [X_1, ..., X_n]'$ and $\mathbf{Y} = [Y_1, ..., Y_n]'$. Similarly, define $L_\lambda = \mathbf{X}\left(D_n' A_\lambda^{-1} Q' \mathbf{X}\right)^{-1} D_n' A_\lambda^{-1} Q'$, $\hat{Y}_\lambda = L_\lambda Y = (\hat{Y}_{\lambda 1}, ..., \hat{Y}_{\lambda n})'$ and $v_\lambda = tr(L_\lambda)$. Then, the Generalized Cross-validation criteria for estimating $\hat{\beta}$ is

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

We then propose the following algorithm for implementation:

Step 1. Choose $\{p^{K_n}(X), q^{J_n}(Z)\}$ (e.g. B-splines or power series). Set J_n to small value (e.g. 4), set $K_n = 2J_n$ and compute $\lambda_n = \arg\min_{\lambda>0} GCV_n(\lambda)$. Compute $\hat{h}_n$ following (9).

Step 2. Switch the values of J_n and K_n (so now $J_n = 2K_n$) and compute $\hat{g}_n$ as in (11).

Step 3. Compute $\hat{\beta}$ as in (13) and $\hat{b} = E_n[\hat{h}_n(Z_i)X_i']^{-1} E_n[(\hat{g}_n(\cdot) - \hat{\beta}' X_i)(\hat{h}_n(Z_i) - X_i)]$.

Step 4. Compute $\tilde{\beta} = \hat{\beta} - \hat{b}$ and $\hat{V} = E_n[\hat{h}_n(Z_i)X_i']^{-1} E_n[\hat{s}_{ni}\hat{s}_{ni}'] E_n[X_i \hat{h}_n(Z_i)']^{-1}$.

In practice, we recommend to carry out sensitivity analysis with respect to $\{K_n, J_n, \lambda_n\}$ in the implementation above. Extensive simulations in Appendix D show that our methods are not sensitive to the tuning parameters $\{K_n, J_n, \lambda_n\}$.[2]

3.4 Overlapping components and Discrete Variables

Suppose now that there are exogenous variables included in the structural equation g. This means X and Z have common components. Specifically, 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

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

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

$$\hat{h}_{2n}(\cdot) = D_{2n}' A_{\lambda_n}^{-1} q^J(\cdot), \tag{14}$$

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 $\tilde{\beta}$ continues to hold, with no changes in the asymptotic distribution, due to the fact that the theory is the same with estimated h than for known h (thanks to the double robustness).

[2]Matlab and R code to implement our TSIV is available at the first author's website.

When Z is discrete, the theory above is applicable but we do not need J_n diverging to infinity. In that case, the linear sieve $\mathcal{H}_n$ is saturated and $q^J(Z)$ could be a saturated basis for it. For example, if Z takes J discrete values, $\{z_1, ..., z_J\}$, we can take $q_j(z) = 1(z = z_j)$. Similarly, if X is discrete we do not need $K_n \to \infty$, and we can choose p^K as a saturated basis. For example, if $X = (1, X_2)$ with X_2 binary (a treatment indicator), we can take $K_n = 2$, $p_1(x) = 1$, $p_2(x) = x_2$, $h(z) = \alpha + \gamma p(z)$, where the propensity score (and then α, γ) can be estimated by sieves, and $g_0(z) = \beta_0 + \beta_1 x_2 \equiv \beta' x$. The formulas for the asymptotic variance of $\tilde{\beta}$ are the same for discrete or continuous variables.

4 A Robust Hausman Test

Applied researchers are concerned about the presence of endogeneity, and have traditionally used 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 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

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

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

$$X_2 = \alpha_1' X_1 + \alpha_2 h_2(Z) + V,$$

then 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 $S = (X, V)'$.

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,

$$t_n \longrightarrow_d N(0,1).$$

5 Monte Carlo

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

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

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 m is a monotone function given below. The DGP is indexed by p and the function m. To generate V note

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

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

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 $\beta = 1$.

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

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

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

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

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

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

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. The standard IV then 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 expected not to be 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 (3.3) with $J_n = 6$ and $K_n = 2J_n$. 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 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. The level of endogeneity does not seem to have a strong impact on the performance of the TSIV estimator.

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

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

Unreported simulations with other DGPs confirm the overall good performance of the proposed TSIV under different scenarios, including cases where h and g are not identified. The sensitivity of the estimator to different choices of tuning parameters, J_n, K_n and λ is presented in Tables 6-8. In each

cell, the top element is for $n = 100$ and the bottom element is for $n = 1000$. 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.

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 misspecification 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 $\gamma \in \{0.4, 0.8\}$ under the null hypothesis of $\rho = 0$. Table 4 contains the results. For DGP1, the rejection frequencies are close to the nominal level of 5% across the different sample sizes, confirming the validity of the test under correct specification. However, for DGP2 and DGP3 we observe large size distortions, as large as 82.2%. This shows that the standard Hausman test is unreliable under misspecification of the linear model.

Table 4: Empirical Size of standard Hausman Test.

γ	n	DGP1	DGP2	DGP3
0.4	100	0.070	0.109	0.046
	500	0.046	0.064	0.053
	1000	0.064	0.072	0.059
0.8	100	0.067	0.223	0.094
	500	0.065	0.134	0.524
	1000	0.060	0.105	0.872

Table 5: Empirical Size and Power of robust Hausman Test.

ρ	γ	n	DGP1	DGP2	DGP3
0.0	0.4	100	0.055	0.037	0.013
		500	0.035	0.018	0.008
		1000	0.038	0.007	0.016
	0.8	100	0.059	0.015	0.013
		500	0.050	0.004	0.003
		1000	0.052	0.003	0.002
0.3	0.4	100	0.176	0.062	0.041
		500	0.649	0.153	0.107
		1000	0.915	0.290	0.222
	0.8	100	0.929	0.324	0.519
		500	1.000	0.710	0.993
		1000	1.000	0.793	1.000
0.9	0.4	100	0.785	0.336	0.249
		500	0.999	0.877	0.825
		1000	0.999	0.974	0.985
	0.8	100	0.993	0.923	0.991
		500	1.000	0.934	1.000
		1000	1.000	0.919	1.000

Table 5 reports rejection probabilities for the proposed robust Hausman test. In contrast to previous results based on the standard IV, we observe that the empirical size is now controlled, with a type-I error that is smaller for nonlinear models than for the linear model. We also report rejection probabilities under the alternative. We observe an empirical power that increases with the sample size and the level endogeneity, suggesting consistency for the proposed Hausman test.

Overall, these simulations confirm the robustness of the proposed methods to mispecification of the linear IV model and their adaptive behaviour when correct specification holds. Furthermore, the TSIV estimator seems to be not 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 for the TSIV estimator as a nonparametric analog to OLS under endogeneity.

6 Appendix A: Notation, Assumptions and Preliminary Results

6.1 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 $\partial_x^a := \partial^{|a|_1}/\partial x_1^{a_1} \dots \partial x_p^{a_p}$, 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

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

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 $p = 1$,

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

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)

$$\hat{h}_n = \left(\hat{T}^* \hat{T} + \lambda_n I \right)^{-1} \hat{T}^* \hat{X}, \tag{15}$$

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

$$h_{\lambda_n} = (T^* T + \lambda_n I)^{-1} T^* X. \tag{16}$$

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.

6.2 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; (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) $\lambda_n c_n = o(n^{-1/4})$, 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) $J_{[\cdot]}(\delta, \mathcal{G}, \|\cdot\|) < \infty$ and $J_{[\cdot]}(\delta, \mathcal{H}, \|\cdot\|) < \infty$ for some $\delta > 0$, and $\mathcal{G}$ and $\mathcal{H}$ have bounded envelopes; (ii) $P(\hat{h}_n \in \mathcal{H}) \to 1$ and $P(\hat{g}_n \in \mathcal{G}) \to 1$.

Assumption A6: (i) $E[U|\, Z] = 0$; (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) $h_0 \in \mathcal{R}(T^*)$, $E\left[|X - h_0(Z)|^4 \Big| X\right]$ is bounded and $Var[h_0(Z)|\, X]$ is bounded and bounded away from zero; and (iv) $E\left[(\hat{h}_{2n}(Z) - h_{20}(Z))V\right] = O_P(n^{-1/2})$.

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(i) 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). The bounded envelop assumption can be easily relaxed. Assumption A5(ii) is satisfied for sieve estimators. Assumptions A6(i-iii) are standard. Assumption A6(iv) is a high-level condition. If Z is independent of V this assumption trivially holds. More general, primitive conditions for Assumption A6(iv) to hold can be shown along the lines of the proof of $E\left[(\hat{h}_{2n}(Z) - h_{20}(Z))h_{20}(Z)\right] = O_P(n^{-1/2})$ in Theorem 4.1.

6.3 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

$$\sup_{h\in\mathcal{H}_n}\left|\|h\|_n^2-\|h\|^2\right|=o_P(1) \tag{17}$$

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 (17) 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, $\left\|\hat{h}_n-h_0\right\|=o_P(n^{-1/4})$ and $\|\hat{g}_n-g_0\|=o_P(n^{-1/4})$.
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 (16),

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

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

$$\|h_{\lambda_n}-h_0\|=O(\lambda_n^{\min(\alpha_h,2)}). \tag{18}$$

With some abuse of notation, denote $\hat{A}_{\lambda_n}=\left(\hat{T}^*\hat{T}+\lambda_n I\right)^{-1}$. Then, arguing as in Proposition 3.14 of Carrasco, Florens and Renault (2006), it is shown that

$$\hat{h}_n-h_{\lambda_n}=\hat{A}_{\lambda_n}\hat{T}^*(\hat{X}-\hat{T}h_0)+\hat{A}_{\lambda_n}(\hat{T}^*\hat{T}-T^*T)(h_{\lambda_n}-h_0), \tag{19}$$

and thus,

$$\left\|\hat{h}_n-h_{\lambda_n}\right\|\leq\left\|\hat{A}_{\lambda_n}\right\|\left\|\hat{T}^*(\hat{X}-\hat{T}h_0)\right\|+\left\|\hat{A}_{\lambda_n}\right\|\left\|\hat{T}^*\hat{T}-T^*T\right\|\|h_{\lambda_n}-h_0\|. \tag{20}$$

As in Carrasco, Florens and Renault (2006),

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

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,

$$\left\|\hat{T}^*\hat{T} - T^*T\right\| \leq O_P\left(\left\|\hat{T}^* - T^*\right\|\right) + O_P\left(\left\|\hat{T} - T\right\|\right) \tag{21}$$

and

$$\begin{aligned}\left\|\hat{T}^* - T^*\right\| &\leq \|P_{\mathcal{H}_n}\| \left\|\hat{T} - T\right\| + \|P_{\mathcal{H}_n} - T^*\| \\ &= O_P\left(\left\|\hat{T} - T\right\|\right) + O_P(c_{n,T^*}).\end{aligned} \tag{22}$$

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,

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

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

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

That is,

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

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 (20), (21) and (22)

$$\begin{aligned}\left\|\hat{h}_n - h_{\lambda_n}\right\| &= O_P\left(\lambda_n^{-1}c_n\right) \\ &= o_P\left(n^{-1/4}\right),\end{aligned}$$

where the last equality follows from A4(v).

The proof of $\|\hat{g}_n - g_0\| = o_P(n^{-1/4})$ is the same and hence omitted. ■

Lemma A3: *Let $\mathcal{H}$ and $\mathcal{G}$ be classes of functions with a bounded envelope F and G, respectively, and let ξ be a squared integrable random variable, then:*

(i) $N_{[\cdot]}(\epsilon, \mathcal{H}\cdot\xi, \|\cdot\|_2) \leq N_{[\cdot]}(C\epsilon, \mathcal{H}, \|\cdot\|_2)$.

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

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

Proof of Lemma A3: Follows from standard arguments in empirical processes theory, and hence is omitted. ■

7 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 $Y = g_0(X) + P_{\mathcal{N}}g(X) + \varepsilon$, and 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. ■

Proof of Proposition 2.3: 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 $\bar{p} = E[p(Z)]$. For a binary X, and since $0 < \bar{p} < 1$, the last display is equivalent to the system

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

or

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

Each equation from the last display defines a hyperplane in h. Since $p(Z)$ is not constant, the normal vectors 1 and $p(Z)$ 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 $p(Z)$ that satisfies the linear constraints, that is, $h_0(Z) = \alpha + \gamma p(Z)$ such that α and γ satisfy the 2×2 system

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

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

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

■

Proof of Proposition 3.1: By Markov's inequality and Lemma A1

$$\frac{1}{n}\sum_{i=1}^{n} \hat{h}_n(Z_i)Y_i = \frac{1}{n}\sum_{i=1}^{n} h_0(Z_i)Y_i + o_P(1),$$

and similarly,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{h}_n(Z_i)X_i' = \frac{1}{n}\sum_{i=1}^{n} h_0(Z_i)X_i' + o_P(1).$$

Then, conclude by the continuous mapping theorem. ■

Proof of Theorem 3.2: Define the class of functions

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

By several applications of Lemma A3 we conclude that

$$\log N_{[\cdot]}(\epsilon, \mathcal{F}, \|\cdot\|_2) \leq \log N_{[\cdot]}(C\epsilon, \mathcal{H}, \|\cdot\|_2) + \log N_{[\cdot]}(C\epsilon, \mathcal{G}, \|\cdot\|_2).$$

Assumption A5 and Theorem 2.5.6 in van der Vaart and Wellner (1996) then imply that $\mathcal{F}$ is a Donsker class. By Lemma A1 $\left\|\hat{h}_n - h_0\right\| = o_P(1)$, and similarly by Lemma A2 $\|\hat{g}_n - g_0\| = o_P(1)$. Then, by a standard empirical processes argument, since $P(\hat{h}_n \in \mathcal{H}) = 1$ and $P(\hat{g}_n \in \mathcal{G}) = 1$, and Assumption 3, it holds that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_i^n \hat{h}_n(Z_i)Y_i - \hat{g}_n(X_i)(\hat{h}_n(Z_i) - X_i) = \frac{1}{\sqrt{n}}\sum_i^n h_0(Z_i)Y_i - g_0(X_i)(h_0(Z_i) - X_i) \\
+ \sqrt{n}E[(\hat{h}_n(Z) - h_0(Z))(Y - g_0(X))] \\
- \sqrt{n}E[(\hat{g}_n(X) - g_0(X))(\hat{h}_n(Z) - h_0(Z))] \\
+ o_P(1) \\
\equiv I + II + III + o_P(1),
\end{aligned}$$

where the expectations in the right hand side are with respect to (Y, X, Z), which is a copy of the (Y_i, X_i, Z_i), independent of the original sample.

It is straightforward to prove that $II = o_P(1)$, since for all squared integrable h,

$$E[(Y - g_0(X))h(Z)] = 0.$$

The rate conditions $\left\|\hat{h}_n - h_0\right\| = o_P(n^{-1/4})$ and $\|\hat{g}_n - g_0\| = o_P(n^{-1/4})$ of Lemma A2 imply by Cauchy-Schwarz inequality

$$|III| \leq \sqrt{n}\left\|\hat{h}_n - h_0\right\| \|\hat{g}_n - g_0\| = o_P(1).$$

On the other hand, $\tilde{\beta}$ is the unique solution to the empirical equation $E_n[m(W, \tilde{\beta}, \hat{h}_n, \hat{g}_n)] = 0$, which yields

$$\tilde{\beta} = \left(\frac{1}{n}\sum_i^n X_i X_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_i^n \hat{h}_n(Z_i)Y_i - \hat{g}_n(X_i)(\hat{h}_n(Z_i) - X_i)\right).$$

Thus, by the invariance principle above

$$\begin{aligned}
\sqrt{n}\left(\tilde{\beta} - \beta\right) &= \left(\frac{1}{n}\sum_i^n X_i X_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_i^n \hat{h}_n(Z_i)Y_i - \hat{g}_n(X_i)(\hat{h}_n(Z_i) - X_i)\right) - \sqrt{n}\beta \\
&= E\left[X_i X_i'\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_i^n h_0(Z_i)Y_i - g_0(X_i)(h_0(Z_i) - X_i)\right) - \sqrt{n}\beta + o_P(1) \\
&= E\left[X_i X_i'\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_i^n s_i\right) + o_P(1),
\end{aligned}$$

The asymptotic normality then follows from the last equality and the central limit theorem.

The consistency of $\hat{V} = E_n[\hat{h}_n(Z_i)X_i']^{-1}E_n[\hat{s}_{ni}\hat{s}_{ni}']E_n[\hat{h}_n(Z_i)X_i']^{-1}$ follows from $\left\|\hat{h}_n - h_0\right\| = o_P(1)$, $\|\hat{g}_n - g_0\| = o_P(1)$ and the consistency of $\hat{\beta}$. ■

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 Markov's inequality 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 (19) and A6(ii) yield

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

where $v(Z) = TA_{\lambda_n}h_0(Z)$. 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)

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

Then, by Theorem 3 in Newey (1997), $A_2 = O_P(1)$. Similarly, an empirical processes argument and A6(iv) show $A_3 = O_P(1)$. 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

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

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

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

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

8 Appendix C: Penalized and Minimum Norm Solutions

The following result shows that our PSMD estimator is equivalent to a minimum norm estimator. This result is of independent interest. Define the optimization problem

$$\min\{||h||_n^2 : h \in \mathcal{H}_n, ||\hat{m}(X; h)||_n^2 \leq b_n/a_n\}, \tag{23}$$

for positive constants a_n and b_n. Define $\hat{X} = \hat{E}[X \,|\, X]$.

Lemma A1: *Assume that $b_n/a_n \downarrow 0$ and $||\hat{X}||_n^2 > 0$. Then, (7) is equivalent to (23), in the sense that for large n, we can find a solution for (7) that also solves (23) for a certain choice of λ_n. Moreover, $\lambda_n^2 = O\left(b_n/a_n\right)$.*

Proof of Lemma A1: Let $\hat{h}_\lambda(\cdot) = D_n' A_\lambda^{-1} q^J(\cdot)$ denote the solution to (7) corresponding to $\lambda > 0$, see (9). We shall show that there is λ_n such that

$$||\hat{m}(X; \hat{h}_{\lambda_n})||_n^2 = b_n/a_n$$

and that $\hat{h}_{\lambda_n}$ is a solution of (23).

Note that (23) is a convex optimization problem, whose necessary and sufficient condition for a solution $\hat{h}_{\lambda_n}$ is that

$$\left\langle \hat{h}_{\lambda_n}, \hat{h}_{\lambda_n} - h \right\rangle_n \leq 0,$$

for all $h \in \mathcal{H}_n$ with $||\hat{m}(X;h)||_n^2 \leq b_n/a_n$ (where $\langle \cdot, \cdot \rangle_n$ is the inner product corresponding to $|| \cdot ||_n$).

Define the linear operator $\hat{K}h = \hat{E}[h(Z)|\, X]$ on $\mathcal{H}_n$, and let $\hat{K}^*$ denote its adjoint (with respect to $\langle \cdot, \cdot \rangle_n$). The optimal solution satisfies the equation

$$\lambda_n \hat{h}_{\lambda_n} + \hat{K}^* \hat{K} \hat{h}_{\lambda_n} = \hat{K}^* \hat{X}.$$

Then, for all $h \in \mathcal{H}_n$ such that $||\hat{K}h - \hat{X}||_n \leq b_n/a_n$,

$$\begin{aligned}
\lambda_n \left\langle \hat{h}_{\lambda_n}, \hat{h}_{\lambda_n} - h \right\rangle_n &\leq \left\langle \hat{K}^*(\hat{K}\hat{h}_{\lambda_n} - \hat{X}), h - \hat{h}_{\lambda_n} \right\rangle_n \\
&\leq \left\langle \hat{K}\hat{h}_{\lambda_n} - \hat{X}, \hat{K}(h - \hat{h}_{\lambda_n}) \right\rangle_n \\
&\leq \left\langle \hat{K}\hat{h}_{\lambda_n} - \hat{X}, \hat{K}h - \hat{X} \right\rangle_n - \left\langle \hat{K}\hat{h}_{\lambda_n} - \hat{X}, \hat{K}\hat{h}_{\lambda_n} - \hat{X} \right\rangle_n \\
&\leq ||\hat{K}\hat{h}_{\lambda_n} - \hat{X}||_n ||\hat{K}h - \hat{X}||_n - ||\hat{K}\hat{h}_{\lambda_n} - \hat{X}||_n^2 \\
&\leq b_n/a_n - b_n/a_n \\
&= 0.
\end{aligned}$$

It remains to show that there exists a positive λ_n such that $||\hat{m}(X;\hat{h}_{\lambda_n})||_n^2 = b_n/a_n$ and $\lambda_n = O\left(b_n/a_n\right)$. Existence follows from Bolzano's Theorem, since $\lambda \to ||\hat{m}(X;\hat{h}_\lambda)||_n^2$ is continuous, $||\hat{m}(X;\hat{h}_\lambda)||_n^2 - b_n/a_n \to ||\hat{X}||_n^2 - b_n/a_n > 0$ as $\lambda \to \infty$ and $||\hat{m}(X;\hat{h}_\lambda)||_n^2 - b_n/a_n \to -b_n/a_n < 0$ as $\lambda \to 0$.

Define $r_{\lambda_n} = \hat{X} - \hat{K}\hat{h}_{\lambda_n}$, and note that $\hat{K}^* r_{\lambda_n} = \lambda_n \hat{h}_{\lambda_n}$ and $||r_{\lambda_n}||_n^2 = b_n/a_n$. Then,

$$\begin{aligned}
||\hat{X}||_n - \left(\frac{b_n}{a_n}\right)^{1/2} &= ||\hat{X}||_n - ||r_{\lambda_n}||_n \\
&\leq ||\hat{K}\hat{h}_{\lambda_n}||_n \\
&\leq \frac{1}{\lambda_n} ||\hat{K}\hat{K}^* r_{\lambda_n}||_n \\
&\leq \frac{1}{\lambda_n} ||\hat{K}||_n^2 \left(\frac{b_n}{a_n}\right)^{1/2},
\end{aligned}$$

or, equivalently

$$\lambda_n \leq \left(\frac{b_n}{a_n}\right)^{1/2} \frac{\lambda_n + ||\hat{K}||_n^2}{||\hat{X}||_n},$$

which shows $\lambda_n^2 = O\left(b_n/a_n\right)$. ■

9 Appendix D: Tables for Simulations

Table 6: Sensitivity analysis of MSE($\times 10^{-2}$) for DGP1.

			λ													
			$K_n = 2J_n$							$K_n = 3J_n$						
J_n	ρ	γ	0	0.001	0.01	0.1	0.2	0.3	0.6	0	0.001	0.01	0.1	0.2	0.3	0.6
4	0	0.4	10.58	9.84	8.37	7.05	6.38	6.62	6.54	8.93	8.67	7.65	6.98	6.42	6.61	6.59
			0.77	0.77	0.66	0.65	0.64	0.64	0.65	0.71	0.76	0.67	0.65	0.64	0.64	0.65
		0.8	1.89	1.62	1.60	1.67	1.56	1.65	1.60	1.87	1.62	1.60	1.67	1.55	1.65	1.60
			0.16	0.16	0.16	0.15	0.16	0.16	0.17	0.16	0.16	0.16	0.15	0.16	0.16	0.17
	0.3	0.4	11.25	10.95	9.82	7.35	7.32	8.24	6.65	8.85	8.73	8.67	7.45	7.22	8.30	6.63
			0.80	0.82	0.72	0.69	0.68	0.69	0.73	0.73	0.80	0.71	0.69	0.68	0.69	0.73
		0.8	2.07	2.17	2.09	2.01	2.00	1.88	2.03	2.05	2.14	2.10	2.02	2.00	1.89	2.03
			0.18	0.20	0.21	0.20	0.20	0.20	0.20	0.18	0.20	0.21	0.20	0.20	0.20	0.20
	0.9	0.4	17.70	19.46	15.45	13.49	12.37	12.04	12.33	15.17	16.57	14.92	13.47	12.57	12.04	12.37
			1.67	1.47	1.33	1.21	1.14	1.24	1.31	1.59	1.39	1.34	1.21	1.14	1.24	1.31
		0.8	5.84	5.72	5.34	5.35	5.52	5.18	5.13	5.53	5.62	5.35	5.39	5.52	5.18	5.13
			0.51	0.54	0.57	0.50	0.54	0.50	0.49	0.51	0.54	0.57	0.50	0.54	0.50	0.49
5	0	0.4	9.94	9.82	8.47	6.72	6.26	6.18	6.39	7.97	8.21	7.75	6.71	6.29	6.19	6.41
			0.86	0.84	0.66	0.65	0.63	0.64	0.64	0.76	0.80	0.67	0.65	0.63	0.64	0.64
		0.8	1.91	1.67	1.63	1.70	1.55	1.64	1.59	1.86	1.65	1.65	1.70	1.54	1.64	1.59
			0.16	0.16	0.16	0.15	0.15	0.16	0.17	0.16	0.16	0.16	0.15	0.16	0.16	0.17
	0.3	0.4	11.94	10.82	10.17	7.22	6.86	7.39	6.58	9.16	8.55	8.90	7.24	6.79	7.42	6.60
			0.89	0.87	0.71	0.69	0.69	0.68	0.73	0.78	0.83	0.72	0.69	0.69	0.68	0.73
		0.8	2.10	2.19	2.14	2.03	2.01	1.86	2.02	2.05	2.13	2.12	2.02	2.00	1.86	2.02
			0.19	0.20	0.21	0.20	0.20	0.20	0.20	0.18	0.20	0.21	0.20	0.20	0.20	0.20
	0.9	0.4	18.46	18.10	15.73	12.94	11.57	12.10	12.01	15.23	16.08	14.60	12.83	11.51	12.13	12.04
			1.77	1.55	1.35	1.21	1.13	1.24	1.30	1.59	1.47	1.35	1.22	1.13	1.23	1.30
		0.8	5.85	5.79	5.44	5.34	5.48	5.17	5.14	5.57	5.65	5.39	5.29	5.49	5.18	5.14
			0.53	0.55	0.57	0.50	0.54	0.50	0.49	0.52	0.55	0.57	0.50	0.54	0.50	0.49
6	0	0.4	9.69	10.05	8.21	6.27	6.20	5.67	6.02	7.84	7.94	7.26	6.32	6.22	5.65	6.04
			0.92	0.85	0.67	0.64	0.63	0.63	0.64	0.80	0.80	0.68	0.65	0.63	0.63	0.64
		0.8	1.96	1.78	1.70	1.69	1.55	1.62	1.58	1.91	1.66	1.63	1.68	1.54	1.62	1.58
			0.16	0.16	0.16	0.15	0.15	0.16	0.17	0.16	0.16	0.16	0.15	0.15	0.16	0.17
	0.3	0.4	11.08	10.10	9.65	7.02	6.80	7.22	6.51	8.80	8.23	8.77	7.14	6.91	7.19	6.50
			1.04	0.91	0.73	0.69	0.69	0.68	0.73	0.82	0.87	0.73	0.69	0.69	0.68	0.73
		0.8	2.23	2.22	2.19	2.03	2.01	1.85	2.02	2.04	2.11	2.17	2.02	2.00	1.84	2.01
			0.19	0.20	0.21	0.20	0.19	0.20	0.20	0.19	0.20	0.21	0.20	0.20	0.20	0.20
	0.9	0.4	19.37	18.72	15.26	12.61	11.74	12.03	12.69	14.26	14.86	13.95	12.51	11.56	11.93	12.61
			1.92	1.58	1.34	1.19	1.13	1.23	1.29	1.60	1.46	1.34	1.20	1.13	1.23	1.29
		0.8	5.92	5.90	5.55	5.29	5.45	5.10	5.13	5.55	5.70	5.48	5.28	5.47	5.07	5.13
			0.53	0.56	0.57	0.51	0.54	0.50	0.49	0.52	0.55	0.57	0.51	0.54	0.50	0.49
7	0	0.4	10.71	8.60	7.32	5.86	5.88	5.43	5.56	7.95	7.71	6.88	5.93	5.92	5.46	5.61
			0.95	0.85	0.68	0.65	0.63	0.63	0.63	0.82	0.80	0.69	0.65	0.63	0.63	0.63
		0.8	2.07	1.74	1.68	1.69	1.54	1.63	1.58	1.92	1.66	1.64	1.68	1.54	1.62	1.58
			0.16	0.16	0.16	0.15	0.15	0.16	0.17	0.16	0.16	0.16	0.15	0.15	0.16	0.17
	0.3	0.4	11.22	9.43	9.12	6.88	6.72	7.02	6.25	8.70	7.85	8.21	6.87	6.74	6.95	6.21
			1.03	0.96	0.74	0.68	0.68	0.68	0.72	0.83	0.87	0.75	0.68	0.68	0.68	0.72
		0.8	2.37	2.24	2.27	2.04	1.99	1.84	2.02	2.11	2.13	2.19	2.02	2.00	1.84	2.00
			0.19	0.20	0.21	0.20	0.19	0.20	0.20	0.19	0.20	0.21	0.20	0.20	0.20	0.20
	0.9	0.4	19.78	18.28	15.58	13.06	12.13	12.53	13.02	14.80	15.07	14.24	12.95	12.12	12.52	13.07
			1.98	1.66	1.31	1.21	1.12	1.23	1.31	1.62	1.51	1.34	1.21	1.12	1.23	1.30
		0.8	6.04	6.07	5.48	5.21	5.42	5.09	5.13	5.71	5.76	5.31	5.23	5.46	5.10	5.14
			0.53	0.56	0.57	0.51	0.54	0.50	0.49	0.53	0.56	0.57	0.50	0.54	0.50	0.49

Table 7: Sensitivity analysis of MSE($\times 10^{-2}$) for DGP2.

			λ													
			$K_n = 2J_n$							$K_n = 3J_n$						
J_n	ρ	γ	0	0.001	0.01	0.1	0.2	0.3	0.6	0	0.001	0.01	0.1	0.2	0.3	0.6
4	0	0.4	36.49	34.99	31.23	32.82	36.11	36.04	38.73	33.99	34.41	32.02	33.59	35.61	35.98	38.43
			3.49	3.13	3.32	4.86	5.26	5.30	6.15	3.66	3.38	3.55	4.86	5.22	5.24	6.11
		0.8	13.80	15.88	15.68	17.08	16.79	17.37	17.46	14.79	17.27	16.49	18.05	17.41	17.91	17.74
			2.25	2.22	2.45	2.58	2.68	2.95	2.87	2.42	2.37	2.63	2.78	2.81	3.06	2.93
	0.3	0.4	41.70	34.96	34.40	36.76	37.38	38.83	37.93	39.48	31.79	34.65	37.43	37.12	38.59	37.64
			3.64	3.36	3.14	4.72	5.43	5.42	6.02	3.88	3.58	3.30	4.69	5.35	5.36	5.93
		0.8	15.21	16.66	15.59	17.44	17.60	18.77	20.40	16.19	17.29	16.70	18.43	18.17	19.27	20.63
			2.50	2.41	2.33	2.57	2.68	2.93	3.16	2.62	2.58	2.50	2.77	2.83	3.06	3.22
	0.9	0.4	51.43	56.95	41.81	43.76	41.78	48.76	48.29	43.82	49.86	42.62	44.71	42.08	48.78	48.02
			4.30	4.56	4.44	5.28	6.07	6.09	6.29	4.05	4.62	4.67	5.26	6.05	6.02	6.23
		0.8	23.87	22.37	20.47	20.34	19.39	21.47	24.11	23.58	22.94	20.95	21.22	19.69	22.05	24.62
			3.28	2.91	2.74	3.09	3.56	3.28	3.48	3.21	2.96	2.90	3.27	3.71	3.40	3.54
5	0	0.4	32.80	36.47	29.03	31.08	32.71	32.21	34.81	30.60	32.29	29.12	31.74	32.92	32.35	34.69
			3.46	3.10	3.08	4.46	4.72	5.27	5.52	3.46	3.22	3.26	4.46	4.67	5.16	5.46
		0.8	12.60	14.56	13.88	15.41	15.28	15.86	15.66	13.05	15.26	14.59	16.41	15.77	16.34	15.91
			1.62	1.54	1.70	1.74	1.90	2.04	2.26	1.75	1.62	1.84	1.83	1.97	2.07	2.29
	0.3	0.4	46.68	32.80	32.50	32.72	32.27	36.03	35.73	43.05	32.18	33.01	33.73	32.99	35.84	35.78
			3.77	3.19	2.94	4.31	4.76	5.24	5.77	3.49	3.42	3.14	4.28	4.70	5.19	5.69
		0.8	13.90	15.09	14.25	16.12	15.98	16.98	18.18	14.54	15.85	14.93	16.86	16.60	17.28	18.46
			1.84	1.83	1.69	1.79	1.93	2.19	2.19	1.81	1.90	1.78	1.90	1.99	2.23	2.22
	0.9	0.4	49.09	54.26	38.87	38.62	38.08	44.49	42.72	41.66	42.61	38.63	39.38	38.44	45.13	42.81
			4.62	4.37	4.09	4.82	5.33	5.57	6.04	4.04	4.33	4.29	4.80	5.24	5.48	5.97
		0.8	21.56	20.61	18.54	18.11	18.29	20.63	22.32	21.22	20.74	18.80	18.96	18.62	21.05	22.91
			2.54	2.37	2.26	2.30	2.56	2.57	2.64	2.42	2.29	2.32	2.38	2.65	2.65	2.67
6	0	0.4	53.93	29.47	27.59	28.54	29.66	30.13	32.74	33.06	27.94	29.22	29.64	30.07	30.51	33.01
			3.34	2.99	2.92	4.19	4.52	4.77	5.21	3.01	3.24	3.17	4.14	4.47	4.69	5.12
		0.8	12.60	14.28	13.17	15.08	14.98	15.34	14.86	12.88	14.92	13.97	15.90	15.39	15.81	15.12
			1.71	1.48	1.62	1.74	1.86	2.06	2.10	1.62	1.55	1.74	1.82	1.89	2.11	2.12
	0.3	0.4	40.03	29.34	29.99	30.17	29.78	33.68	33.86	35.84	27.83	31.29	31.79	30.68	33.82	34.03
			3.62	3.14	2.67	4.05	4.60	4.70	5.21	3.47	3.11	2.83	4.00	4.57	4.67	5.14
		0.8	13.62	14.06	13.98	15.77	15.40	16.48	17.31	14.11	14.52	14.67	16.34	15.92	16.83	17.53
			1.83	1.64	1.54	1.78	1.85	2.10	2.27	1.73	1.71	1.64	1.85	1.89	2.13	2.30
	0.9	0.4	60.72	46.57	35.46	36.41	35.46	40.62	41.88	42.88	38.53	35.39	37.34	36.13	41.08	42.29
			4.39	4.33	3.84	4.62	5.21	5.24	5.61	3.87	4.20	4.05	4.64	5.14	5.19	5.53
		0.8	20.90	20.27	17.87	17.85	17.71	19.02	21.84	20.17	20.08	18.06	18.60	18.12	19.42	22.26
			2.41	2.22	1.94	2.19	2.59	2.45	2.50	2.27	2.12	1.98	2.24	2.64	2.49	2.54
7	0	0.4	117.41	29.85	26.96	27.86	28.58	28.32	31.52	33.51	28.19	27.72	29.50	29.26	28.74	31.74
			3.25	3.05	2.79	4.05	4.24	4.62	5.09	3.38	3.09	3.05	4.05	4.22	4.56	5.01
		0.8	12.54	14.01	12.92	14.70	14.49	14.82	14.59	12.75	14.55	13.54	15.36	14.91	15.23	14.85
			1.46	1.36	1.54	1.58	1.74	1.87	1.95	1.44	1.37	1.63	1.63	1.77	1.91	1.97
	0.3	0.4	43.41	29.13	30.90	29.18	29.03	32.56	33.17	31.83	28.27	31.45	30.96	29.74	33.54	33.57
			3.43	2.90	2.67	3.90	4.14	4.46	5.16	3.35	3.02	2.84	3.84	4.08	4.42	5.10
		0.8	14.24	14.29	13.88	15.25	15.23	15.98	16.62	14.43	14.31	14.37	15.98	15.67	16.34	16.84
			1.59	1.54	1.44	1.59	1.76	1.98	1.94	1.57	1.55	1.51	1.65	1.79	2.01	1.97
	0.9	0.4	77.30	44.87	34.52	34.60	34.92	38.92	40.54	53.12	37.77	34.27	35.77	35.30	39.87	40.83
			4.78	4.18	3.84	4.28	4.85	5.06	5.35	3.95	4.21	3.96	4.30	4.83	5.01	5.30
		0.8	20.53	19.65	16.80	17.38	16.84	18.61	21.06	19.57	19.96	17.00	18.32	17.28	18.87	21.40
			2.29	2.12	2.00	2.00	2.27	2.29	2.43	2.09	2.07	2.03	2.03	2.30	2.33	2.45

Table 8: Sensitivity analysis of MSE($\times 10^{-2}$) for DGP3.

			λ													
			$K_n = 2J_n$							$K_n = 3J_n$						
J_n	ρ	γ	0	0.001	0.01	0.1	0.2	0.3	0.6	0	0.001	0.01	0.1	0.2	0.3	0.6
4	0	0.4	89.50	79.24	86.87	85.35	89.05	93.34	107.17	90.67	82.94	89.85	87.03	89.82	94.38	107.13
			7.64	7.80	7.65	9.97	9.82	9.73	11.02	7.96	8.21	7.92	9.91	9.81	9.68	11.00
		0.8	53.60	47.34	51.40	48.01	53.19	54.75	49.33	53.39	47.18	50.88	48.33	52.86	54.43	49.25
			5.00	4.96	4.73	5.25	5.27	5.64	5.80	5.04	5.04	4.80	5.27	5.30	5.64	5.82
	0.3	0.4	82.01	81.80	77.76	87.58	89.50	106.04	90.46	82.71	83.17	81.33	89.08	89.96	105.42	90.14
			7.68	8.17	8.88	9.52	11.00	10.64	10.21	7.89	8.46	9.06	9.47	10.94	10.56	10.19
		0.8	55.98	51.81	52.21	52.12	56.89	55.73	49.28	55.34	51.92	52.10	52.34	56.79	55.43	48.96
			5.80	5.85	5.38	5.47	6.11	6.14	6.09	5.85	5.92	5.46	5.52	6.13	6.18	6.09
	0.9	0.4	97.77	96.58	101.87	102.88	106.35	122.48	126.38	102.13	98.56	104.74	104.85	107.05	122.36	124.53
			9.99	8.99	9.52	10.55	11.78	12.66	12.69	9.93	9.09	9.76	10.53	11.72	12.61	12.65
		0.8	64.62	62.55	66.26	61.41	63.56	63.49	60.79	63.63	62.49	65.04	60.96	63.21	62.92	60.36
			6.17	6.03	6.79	7.14	7.44	7.18	7.60	6.24	6.08	6.88	7.16	7.46	7.18	7.61
5	0	0.4	88.84	79.45	87.03	80.84	84.82	94.25	105.07	91.18	83.15	91.48	83.97	86.96	96.56	105.12
			7.58	7.96	7.72	10.00	9.97	9.85	11.45	8.16	8.20	7.94	9.91	9.94	9.79	11.41
		0.8	53.51	47.52	51.81	48.51	53.95	55.35	50.87	53.70	46.67	52.31	48.97	53.81	54.85	50.42
			5.07	5.03	4.79	5.32	5.36	5.76	5.91	5.12	5.13	4.86	5.39	5.37	5.76	5.91
	0.3	0.4	81.37	74.98	75.92	81.46	87.46	103.16	92.33	85.74	79.16	79.95	85.07	88.91	104.55	92.32
			7.46	7.77	8.69	9.43	11.13	10.78	10.40	7.99	8.13	8.98	9.37	11.06	10.71	10.36
		0.8	55.68	51.65	52.02	51.98	57.78	56.88	50.81	55.93	51.62	51.85	52.42	57.58	56.71	50.72
			5.86	5.95	5.44	5.58	6.24	6.21	6.26	5.91	6.03	5.53	5.62	6.25	6.24	6.26
	0.9	0.4	96.89	94.77	96.32	99.18	104.32	119.06	123.94	96.91	95.00	97.44	102.57	105.87	119.05	123.89
			9.59	8.78	9.24	10.47	11.90	12.90	12.97	9.66	9.28	9.60	10.43	11.85	12.83	12.91
		0.8	63.97	62.13	65.43	61.15	63.98	63.78	61.76	63.33	62.07	64.64	60.71	64.02	63.28	61.24
			6.29	6.15	6.86	7.28	7.57	7.31	7.75	6.36	6.21	6.99	7.29	7.58	7.31	7.77
6	0	0.4	86.06	77.41	71.12	79.02	81.58	90.28	102.63	86.40	80.97	81.00	81.98	84.73	92.78	103.54
			7.69	7.76	7.75	9.91	9.97	9.84	11.74	7.98	8.18	7.97	9.84	9.89	9.76	11.67
		0.8	53.87	46.98	51.86	48.59	54.17	55.23	51.42	54.67	46.77	52.37	49.05	54.67	54.88	51.14
			5.09	5.10	4.86	5.41	5.44	5.88	6.05	5.14	5.19	4.91	5.48	5.48	5.88	6.05
	0.3	0.4	76.92	74.25	75.31	80.21	86.00	99.48	87.68	83.40	79.15	80.89	83.99	88.55	103.19	87.78
			7.67	7.90	8.48	9.37	11.09	10.83	10.60	8.24	8.22	8.83	9.27	10.97	10.77	10.53
		0.8	55.46	51.27	51.85	51.97	57.86	57.62	51.55	56.03	50.96	51.81	52.06	58.13	57.56	51.55
			5.90	6.05	5.51	5.68	6.34	6.33	6.41	5.95	6.09	5.62	5.72	6.33	6.35	6.39
	0.9	0.4	95.32	94.26	92.09	98.61	99.25	115.32	122.08	95.61	92.97	95.75	100.82	100.28	115.61	123.20
			9.42	8.98	9.19	10.48	11.96	13.03	13.20	9.69	9.15	9.49	10.30	11.86	12.91	13.11
		0.8	63.90	61.49	65.11	60.39	63.88	63.52	61.84	63.14	61.26	64.76	60.31	63.73	63.42	61.50
			6.39	6.25	7.01	7.40	7.68	7.45	7.91	6.44	6.31	7.08	7.38	7.67	7.43	7.90
7	0	0.4	84.62	75.74	69.21	76.80	78.68	89.40	98.63	85.16	80.13	77.55	80.96	82.60	91.46	100.63
			7.72	7.71	7.62	9.82	9.94	9.85	11.80	8.13	8.05	7.97	9.80	9.85	9.73	11.68
		0.8	54.26	47.25	52.15	48.55	54.52	55.30	51.60	54.84	46.79	52.37	49.09	55.11	55.00	51.47
			5.08	5.14	4.87	5.47	5.46	5.91	6.10	5.19	5.22	4.98	5.53	5.50	5.90	6.10
	0.3	0.4	72.10	74.49	73.88	79.47	85.55	100.92	85.44	78.83	77.45	81.06	82.28	87.01	103.07	85.78
			7.86	7.70	8.40	9.33	11.04	10.74	10.62	8.03	8.09	8.78	9.26	10.93	10.66	10.54
		0.8	55.32	51.55	51.33	51.80	58.06	57.40	51.56	55.42	50.94	51.98	52.07	58.23	57.67	51.71
			5.88	6.06	5.51	5.70	6.37	6.36	6.47	5.98	6.11	5.63	5.75	6.36	6.39	6.45
	0.9	0.4	90.60	90.58	91.68	98.10	98.24	111.18	119.73	91.37	91.41	92.91	101.61	100.15	113.37	121.70
			9.62	9.17	8.97	10.35	11.92	13.01	13.11	9.86	9.26	9.24	10.18	11.76	12.85	12.99
		0.8	62.05	60.28	65.22	59.96	63.60	63.41	62.21	62.24	60.52	65.19	60.12	63.37	63.10	61.89
			6.43	6.23	7.02	7.37	7.70	7.51	7.98	6.43	6.28	7.10	7.36	7.69	7.48	7.95

References

Abadie, A. (2000): "Semiparametric Estimation of Instrumental Variable Models for Causal Effects," NBER Technical Working Paper No. 260.

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.

Bang, and J.M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics* 61, 962–972.

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

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

Canay, I., Santos, A. and A. Shaikh (2013): "On the Testability of Identification in Some Nonparametric 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.

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 (2015): "Overidentification in Regular Models," working paper.

Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W.K. and J. Robins (2018): "Locally Robust Semiparametric Estimation", working paper.

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. and K. Song (2010): "Testing Single-index Index Restrictions with a Focus on Average Derivatives," *Journal of Econometrics*, 156, 377-391.

Firpo, S. and C. Rothe (2016): "Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions," working paper.

Florens, J.-P., J. Johannes, and S. Van Bellegem (2011): "Identification and Estimation by Penalization in Nonparametric Instrumental Regression, " *Econometric Theory*, 27, 472–496.

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.

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.

Robins, J.M., A. Rotnitzky, and M. van der Laan (2000): "Comment on 'On Profile Likelihood' by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association* 95, 431-435.

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.

Scharfstein D.O., A. Rotnitzky, and J.M. Robins (1999): Rejoinder to "Adjusting For Nonignorable Drop-out Using Semiparametric Non-response Models," *Journal of the American Statistical Association* 94, 1135-1146.

Severini, T. A., and G. Tripathi (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theorey*, 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.

van der Vaart, A. W. (1991): "On differentiable functionals," *The Annals of Statistics*, 19(1), 178–204.

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

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