Identifying Marginal Effects Using Covariates[*]

Carolina Caetano[†]
University of Rochester

Juan Carlos Escanciano[‡]
Indiana University

May 10th, 2016.

Abstract

This paper proposes a new strategy for the identification of all the marginal effects of an endogenous multi-valued variable (which can be continuous, or a vector) in a nonparametric model with a lower dimensional Instrumental Variable (IV), which may even be a single binary variable. Identification is achieved by exploiting heterogeneity of the "first stage" in covariates through a new rank condition that we term covariance completeness. The covariates themselves may be endogenous, but their endogeneity does not affect the identification of the marginal effects of interest. This paper also provides parametric and nonparametric Two-Stage Least Squares (TSLS) estimators which are simple to implement, discusses their asymptotic properties, and shows that the estimators have satisfactory performance in moderate samples sizes. Finally, we apply our methods to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005).

Keywords: Conditional instrumental variables; Covariance completeness; Endogeneity; Nonparametric identification; Two-stage least squares.

JEL classification: C13; C14; C21; D24

[*] We would like to thank Gregorio Caetano, Kenneth Chay, Matias Cattaneo and Pedro Sant'Anna for helpful discussions. We would also like to thank seminar participants at many institutions for useful comments. This paper subsumes part of the working paper "Identifying Multiple Marginal Effects with a Single Binary Instrument or by Regression Discontinuity".

[†] Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627-0156, USA. E-mail: carol.caetano@rochester.edu. Web Page: http://www.carolinacaetano.net/.

[‡] Department of Economics, Indiana University, 105 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.

1 Introduction

Instrumental Variables (IV) methods are well established as one of the most useful approaches to identify causal effects in econometric models. Consider first the nonparametric additively separable model

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

where Y is the dependent variable, g is an unknown measurable function of X, and U is an unobservable error term. The vector X is endogenous, in the sense that $\mathbb{E}[U|X] \neq 0$ with positive probability. In addition to (Y, X), we also observe an instrument Z that is conditionally exogenous given a vector of observed covariates W, i.e.

$$\mathbb{E}[U|Z, W] = \mathbb{E}[U|W] \text{ almost surely (a.s.).} \tag{2}$$

This paper studies nonparametric identification of the marginal effects of X on Y, i.e. identification of the function g (up to a constant), in the model defined by (1)-(2).

Depending on the nature of X and functional form assumptions on g, a traditional IV approach requires, among other things, that the instrument Z is sufficiently complex (see Newey and Powell (2003)). For instance, if X is continuous and we wish to identify g nonparametrically, then Z must be continuous. If X is discrete with q points of support and we wish to identify all of its marginal effects, then Z needs to have at least q points of support. If X is a vector of continuous variables, then Z must have at least as many components as X, even if we impose substantial restrictions on the shape of g, such as linearity. Thus, the traditional IV order condition imposes restrictive assumptions on the support of the instrument Z (relative to that of X), which may not hold in applications.

In this paper we propose a strategy for the identification of g in equation (1) (up to a constant) which applies to cases in which the support of X is larger than that of Z (and thus it is impossible to achieve identification with a traditional IV approach.) We focus first on the most difficult case in which Z is a binary variable, say $Z \in \{0, 1\}$, while X takes 3 or more values, and may even be continuous, or a vector. Our results for binary IV open up the possibility of the identification of all the marginal effects of a complex variable X in cases where the instrument may be an experiment or a natural experiment.

Furthermore, we show that with almost no modifications our methodology can be extended to the nonseparable model

$$Y = m(X, U), \tag{3}$$

where $m(x, u)$ is a strictly increasing function in u, for each x in the support of the distribution of X, $\mathcal{S}_X$ say. Model (3) allows for unobserved heterogeneous marginal effects, as in, e.g., wage equations where returns to education depend on unobserved individual's ability; see, e.g., Card (2001).

Identification holds in either (1) or (3) under a new rank condition that we term *covariance completeness*. This nonparametric/semiparametric identifying assumption imposes support restrictions on covariates (relative to X) rather than on instruments and requires a semiparametric structural separability condition of covariates. In our specifications W is not excluded from the structural equation and may be endogenous (W is part of U). Thus, the support requirement is not too restrictive, being certainly more plausible than a support requirement on the IV.

The following example illustrates and formalizes some of these ideas in a simple model where identification with a standard IV approach is not possible.

Example 1.1 *(Bivariate linear case) Suppose that $X = (X_1, X_2)$ and g is linear, so the model in (1) is*

$$Y = \beta_1 X_1 + \beta_2 X_2 + U$$

where $\mathbb{E}[U|X_1, X_2] \neq 0$. Here X_1 and X_2 can be two different variables, such as "education" and "IQ," or a relaxation of the linearity of a variable, for example X_1 is "education," and $X_2 = \mathbf{1}(X_1 \geq 17)$ captures the "sheepskin effect of graduating from high school." The Standard IV methods are unable to identify β_1 and β_2 with a single binary instrument Z, because the classical order condition fails.

For our approach, we require that the instrument Z be excluded from the structural equation conditional on W, i.e. $\mathbb{E}[U|W, Z] = \mathbb{E}[U|W]$, but W itself may be endogenous. Then

$$\begin{aligned}\mathbb{E}[Y|Z = 1, W] - \mathbb{E}[Y|Z = 0, W] = \beta_1 \left[\mathbb{E}[X_1|Z = 1, W] - \mathbb{E}[X_1|Z = 0, W]\right] \\ + \beta_2 \left[\mathbb{E}[X_2|Z = 1, W] - \mathbb{E}[X_2|Z = 0, W]\right].\end{aligned}$$

To identify β_1 and β_2 we need to invert this equation. The condition that guarantees that we can invert it is what we call covariance completeness for the class $\mathcal{G} = \{g(X_1, X_2) = \beta_1 X_1 + \beta_2 X_2; \beta_1, \beta_2 \in \mathbb{R}\}$. In this example covariance completeness holds if there exists no pair $(\lambda_1, \lambda_2) \neq (0, 0)$ such that

$$\lambda_1 \left(\mathbb{E}[X_1|Z = 1, W] - \mathbb{E}[X_1|Z = 0, W]\right) + \lambda_2 \left(\mathbb{E}[X_2|Z = 1, W] - \mathbb{E}[X_2|Z = 0, W]\right) = 0 \quad a.s.$$

This condition states that the "first stage" effects of Z on X_1 and X_2 vary (in a linearly independent manner) with W. In other words, our identification strategy exploits the heterogeneity in the "first stages" to separate the marginal effects of X_1 and X_2.

Note that W itself may be endogenous and not excluded from the structural equation, i.e. we can specify $U = h(W, \varepsilon)$ for an unknown h and vector of unobservables ε. The nuisance parameter h is not identified, but we are not concerned by this, as we are only interested in the marginal effects of X_1 and X_2 on Y.

To clarify the structural separability requirement in the general case, consider the following example. An individual's earnings Y depend on schooling X and unobserved ability U as in model (3). Ability is produced using parents' education W and other observable and unobservable factors ε as inputs, i.e $U = h(W, \varepsilon)$. The structural separability assumption is that parents' education is not a direct input in wage equations, although returns to education may depend on parents' education through its impact in shaping ability. We may be giving the impression that all covariates must be structurally separable, which in this example means that all covariates must be indirect inputs which affect wages only through ability. That is not the case, our approach allows the inclusion of exogenous covariates which are themselves direct inputs in the wage production function. The structural separability is imposed only on those covariates which are used to implement our identification method, and those covariates are allowed to be endogenous.

Covariance completeness does not impose restrictions on the support of the instrument, which may even be binary. For example, following Card (1995) we could use as Z a binary variable indicating living close to a college, which is more likely to be exogenous after conditioning on parents' education W. We provide conditions under which returns to education, which are allowed to depend here on ability, are nonparametrically identified with a single binary instrument and a standard conditional exogeneity restriction.

The identification strategy is based on the observation that if the population is categorized according to a covariate W, the discrete variation that Z induces on the distribution of X may vary with W. This may allow us to recover a rich (e.g continuous) set of marginal effects. For example, in the returns to education application living close to a college may have a differential impact on individual's education X depending on the level of parents' education W. For instance, distance to college is likely to have a relatively larger impact for individuals with less educated parents. By comparing the effects of different changes in schooling across different parental education levels we can recover the effect of each year of college on wages.

Based on our identification results, we propose parametric and nonparametric estimators of g up to a constant. Our estimation approach avoids completely the need of estimating the propensity score or the conditional variance, and can be immediately implemented using packaged software.[1] In models that are linear in parameters we show that our identification strategy can be straightforwardly implemented with a simple Two-Stage Least Squares (TSLS) estimator that treats the possibly endogenous covariate W as if it were exogenous and uses interaction terms between W and Z as instruments. Additionally, we propose nonparametric estimators that relax the functional form assumptions, and discuss the rates of consistency based on results by Blundell, Chen and Kristensen (2007). Our identification strategy in the nonparameric separable case can also be implemented as a TSLS regression with off-the-shelf econometric software.

Our results for separable models complement the classical nonparametric IV approach in, e.g., Newey and Powell (2003), Darolles, Fan, Florens and Renault (2011), Blundell, Chen and Kristensen (2007) and Horowitz (2011). In the general case, our paper contributes to the literature of nonparametric identification of heterogeneous (observable and unobservable) marginal effects; see Matzkin (2013) for a recent survey of this literature. Our results for nonseparable models complement alternative identification strategies for binary instruments and continuous endogenous variables in Chesher (2003), D'Haultfoeuille and Fevrier (2014), Torgovitsky (2014), D'Haultfoeuille, Hoderlein and Sasaki (2013) and Masten and Torgovitsky (2014); and for continuous instruments in Altonji and Matzkin (2005), Chernozhukov and Hansen (2005) and Florens, Heckman, Meghir and Vytlacil (2008), among others. The main contribution of this paper is the way we use covariates for identification. In particular, none of the papers mentioned above exploit the heterogeneity of the "first stage" conditional on possibly endogenous covariates and discuss covariance completeness conditions. More broadly, traditional methods treat covariates as exogenous variables. If covariates turn out to be endogenous (see e.g. experience or parents' education in wage equations), then estimates of marginal effects of interest may be incon-

[1] Stata code to implement the parametric and nonparametric estimators of this paper is available at the first author's website.

sistent; see, e.g., Frolich (2008). We provide explicit conditions under which endogeneity of covariates does not affect consistency of marginal effects of interest with our methods and also with traditional IV methods. Given the ubiquitous presence of endogenous covariates in applications this robustness is of certain practical relevance. Both our identification method and estimators have empirical advantages in that they can be adapted immediately to the idiosyncrasies of applied work, including incorporating functional form restrictions, large number of covariates and fixed effects all without the need to modify neither the method nor the estimator.

The rest of the paper is organized as follows. For simplicity, we introduce all the ideas on the separable model, which is done in Section 2. There we present the identification results (Section 2.1) and also show that the identification strategy can be implemented with a suitable TSLS estimator (Section 2.2). Section 3 extends the identification results of the separable case to the nonseparable model (see the implementation on the nonseparable case in Appendix A.2.3). Section 4 reports the results of Monte Carlo experiments. Section 5 contains an empirical application of our method to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005). Finally, we conclude in Section 6. Mathematical proofs of the main results are gathered in the Appendix.

2 The Separable Case with Binary IV

2.1 Identification

Throughout this section we assume that the observed random vector (Y, X, W, Z) satisfies the model

$$Y = g(X) + U, \tag{4}$$

where the following exclusion restriction holds

Assumption 1 *(validity)* $\mathbb{E}[U|W, Z] = \mathbb{E}[U|W]$ *a.s.*

We introduce the idea in the case where Z is binary, i.e. $\mathcal{S}_Z = \{0, 1\}$. Taking conditional means in (4) and subtracting terms we can thus write

$$\mathbb{E}[Y|W, Z = 1] - \mathbb{E}[Y|W, Z = 0] = \mathbb{E}[g(X)|W, Z = 1] - \mathbb{E}[g(X)|W, Z = 0] \text{ a.s.} \tag{5}$$

Identifying g (up to location) from this implicit equation depends on our ability to invert it. To better understand the conditions that guarantee the invertibility of equation (5) consider first the following example for the case where X and W are discrete, which extends naturally to the general case. This example is followed by an empirical example which clarifies how some of the requirements and ideas translate to an actual applied problem.

Example 2.1 *(X and W discrete) Denote by* $\mathcal{S}_X := \{x_1, ..., x_q\}$ *and* $\mathcal{S}_W := \{w_1, ..., w_l\}$ *the supports of the distributions of* X *and* W*, respectively, with* $q < \infty$ *and* $l < \infty$*.*

Our identification strategy consists of inverting equation (5), which in this context can be written as a linear system $\mathbf{m} = \mathbf{A}\mathbf{g}$, *where,* $\mathbf{m} := (m(w_1), \ldots, m(w_l))'$, *($a'$ denotes the transpose of a) with* $m(w) := \mathbb{E}[Y|Z = 1, W = w] - \mathbb{E}[Y|Z = 0, W = w]$, $w \in \mathcal{S}_W$, *the matrix* $\mathbf{A}$ *is given by* $\mathbf{P}_1 - \mathbf{P}_0$, *where* $\mathbf{P}_z = (p_{zij})$ *is the* $l \times q$ *matrix with entries* $p_{zij} = \mathbb{P}[X = x_j|Z = z, W = w_i]$, $i = 1, ..., l$, $j = 1, ..., q$ *and* $z = 0, 1$, *and* $\mathbf{g} : = (g(x_1), ..., g(x_q))'$. *Notice that since* $\mathbf{P}_0$ *and* $\mathbf{P}_1$ *are matrices of probabilities,* $\mathbf{A}\iota = 0$, *where* ι *denotes the* $q \times 1$ *vector of ones. Therefore,* $\mathbf{A}$ *is not full-rank, and thus* g *is not identified from (5). However, in this context we can identify linear functionals* $c'\mathbf{g}$ *with* c *in a space of dimension rank*$(\mathbf{A})$. *In particular, if rank*$(\mathbf{A}) = q - 1$, *then all linear functionals* $c'\mathbf{g}$ *with* $c'\iota = 0$ *are identified. In this case, all increment effects* $g(x_h) - g(x_j)$, $h \neq j$, *are identified. Of course, this is only possible if the order condition* $l \geq q - 1$ *holds, so* W *needs to assume at least* $q - 1$ *different values.*

In contrast, the classic nonparametric IV strategy is based on the equation

$$\mathbb{E}[Y|Z] = \mathbb{E}[g(X)|Z],$$

which translates into the system of equations $\mathbf{r} = \mathbf{P}\mathbf{g}$, *where* $\mathbf{r} := (\mathbb{E}[Y|Z = 0], \mathbb{E}[Y|Z = 1])'$, *and* $\mathbf{P} = (p_{Zj})$ *is the* $2 \times q$ *matrix with entries* $p_{Zj} := \mathbb{P}(X = x_j|Z = z)$, $z = 0, 1$, $j = 1, \ldots, q$. *In this classic setting the matrix* $\mathbf{P}$ *has a rank of at most 2, and so* g *is not identified if* $q > 2$ *(see Newey and Powell (2003)). In fact, we can only identify linear functionals* $c'\mathbf{g}$ *where* c *is spanned by the two rows of* $\mathbf{P}$ *(see Severini and Tripathi (2006, 2012)), which are not necessarily of interest.*

Example 2.2 *(Effects of maternal smoking on birth weight) Consider the problem of estimating the marginal effect of the amount a woman smokes during pregnancy (average daily number of cigarettes) on the baby's weight at birth (see Almond and Currie (2011) and Lumley et al. (2011) for discussions of the literature on this problem.) The following setup is entirely fictitious, but we believe that the association of our notation to a real problem can be helpful. Suppose that smoking can take 3 values* $X \in \{0, 1, 3\}$. *Later it will be immediate to see how the argument extends when* X *assumes more values. Suppose that women are randomly divided in two groups, indexed by* Z, *and let the classification covariate* W *be the mother's years of education. Table 1 shows an overview of the situation. Column (I) represents the years of education, and at first we are considering only 3 possibilities: 6, 10 and 17. Columns (II) and (III) show the average amount smoked by the women in the control and treatment groups, respectively, for the given number of years of education. Curiously, on average all groups reduced one cigarette*

Table 1: **Identification Idea**

(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)
$\mathbf{W}$	$\bar{\mathbf{X}}_{\mathbf{0},\mathbf{W}}$	$\bar{\mathbf{X}}_{\mathbf{1},\mathbf{W}}$	$\mathbf{\Delta}_{\mathbf{Y}}(\mathbf{W})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{0})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{1})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{3})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{0})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{1})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{3})$	row #
6	3	2	10	0	0	1	0	1/2	1/2	(1)
10	3	2	22	0	0	1	1/5	1/5	3/5	(2)
17	3	2	30	0	0	1	1/3	0	2/3	(3)

because of the intervention. This is not a requirement of our method, we just want to show what can be achieved even when the "first stages" do not vary at all across the different values of W. Column (IV) shows the average difference (in grams) in the birth weight between the $Z = 1$ and the $Z = 0$ groups for that level of education ($\Delta_Y(W) = \mathbb{E}[Y|Z = 1, W] - \mathbb{E}[Y|Z = 0, W]$). Columns (V) to (VII) show the smoking distribution when $Z = 0$. In this example everyone in the $Z = 0$ group smokes 3 cigarettes, which is just a simplification for explanation purposes. Columns (VIII) to (X) show the corresponding fractions in the $Z = 1$ group. As we can see, each row is different. It means that Z affects each of the education groups differently. It is this variation in the distributions that is at the heart of our approach. It does not matter that all the average effects are the same, it would not even matter if there were no first stage effects at all. As we show next, our ability to identify the marginal effects comes from the fact that the instrument affected the distribution of X differently across the different W.

Following the previous example, we build the system $\mathbf{m} = \mathbf{Ag}$. Here $\mathbf{m} = \Delta_Y(W)$, and each column of $\mathbf{A}$ denotes the differences in the probabilities, e.g. column 1 is $\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{0}) - \mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{0})$. The resulting system of equations is

$$10 = 0.5g(1) - 0.5g(3) \tag{6}$$

$$22 = 0.2g(0) + 0.2g(1) - 0.4g(3) \tag{7}$$

$$30 = 0.33g(0) - 0.33g(3). \tag{8}$$

Just as pointed out in the previous example, note that only two equations are linearly independent (since $0.4(6) + 0.6(8) = (7)$). In fact, if we had used more values of the variable W, we could have more equations, but it would not change the fact that at most two equations would be independent. This is caused by the fact that the coefficients of each of these equations always add up to zero, since they are the subtraction of probabilities, which themselves always add up to one.

Since we have 2 linearly independent equations, we cannot recover the values of $g(0)$, $g(1)$, and $g(3)$, but we can recover the value of any increment. It is straightforward to see in this example that, from equation (6), $g(3) - g(1) = -20$, from equation (8), $g(3) - g(0) = -90$, and combining both results, $g(1) - g(0) = -70$. In a situation where X assumes more values, say q, we can get all the increments provided we have $q - 1$ linearly independent equations (and thus W must assume at least $q - 1$ values).

The discussion in Example 2.1 extends to the general case as follows. With some abuse of notation, we write equation (5) also as

$$m = Ag, \tag{9}$$

where now $Ag := \mathbb{E}[g(X)|W, Z = 1] - \mathbb{E}[g(X)|W, Z = 0]$ is a continuous (i.e. bounded) linear operator, $A : L_2(X) \to L_2(W)$, where henceforth, for a generic random vector ζ, $L_2(\zeta)$ denotes the Hilbert space of square-integrable functions with respect to the distribution of ζ, with support $\mathcal{S}_\zeta$. We introduce our identification assumption as follows. Define $\mathcal{N}(A) = \{g \in L_2(X) : Ag = 0\}$, the null space of A. Our relevance condition requires that the null space of A is composed exclusively of the constant functions:

Assumption 2 *(relevance)* $\mathcal{N}(A) = \{f \equiv c \in \mathbb{R}\}$.

Notice that the identification condition in Example 2.1 that $rank(A) = q-1$ is equivalent to Assumption 2 in the discrete support case, since $\dim(\mathcal{N}(A)) + rank(A) = q$. The proof of the following identification result can be found in the Appendix A.1.

Theorem 2.1 *Under Assumptions 1 and 2, g is identified up to location.*

In the general case, Assumption 2 is a nonparametric rank condition which is the analogue in our setting of the L_2-completeness condition required in nonparametric IV (see Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Andrews (2011) and D'Haultfoeuille (2011) for discussions on completeness).

Note that we can also write our identification equation (5) as

$$\mathbb{C}(Y, Z|W) = \mathbb{C}(g(X), Z|W),$$

where $\mathbb{C}(V_1, V_2|W) = \mathbb{E}[(V_1 - \mathbb{E}[V_1|W])(V_2 - \mathbb{E}[V_2|W])|W]$ is the conditional covariance of V_1 and V_2 given W. This way of writing equation (5) inspires the introduction of a rank condition equivalent to Assumption 2, which we term "covariance completeness" and which naturally generalizes to cases where Z is not binary, or the structural equation is not separable, as we show later. Next we define covariance completeness in detail and compare it to the classical completeness used in nonparametric IV. Some readers may prefer to skip to Examples 2.4 to 2.6, which translate the meaning of covariance completeness to some important special cases.

Covariance Completeness. We introduce a general definition of covariance completeness and provide examples. To refine our identification result above and provide sufficient and necessary conditions for identification, while allowing for the possibility of prior information on the parameter space for g, we introduce the following class of functions. Let $\mathcal{G}$ be a subset of $L_2(X)$ with the properties: (i) if $g_1, g_2 \in \mathcal{G}$ then $g_1 - g_2 \in \mathcal{G}$; (ii) if $g_1, g_2 \in \mathcal{G}$ then $g_1 + g_2 \in \mathcal{G}$; and (iii) elements in $\mathcal{G}$ satisfy the normalization restriction $g(\bar{x}) = 0$ for a fixed $\bar{x} \in \mathcal{S}_X$. In examples where $\mathcal{G}$ is a subspace, conditions (i) and (ii) hold automatically. Condition (iii) is a location normalization.

Definition 2.1 *We say (X, Z) given W is $\mathcal{G}$-covariance complete if for each $g \in \mathcal{G}$*

$$\mathbb{C}(g(X), Z|W) = 0 \;\; a.s. \Longrightarrow g = 0 \;\; a.s.$$

When $\mathcal{G}$ is unrestricted (except for the location normalization) we simply say (X, Z) given W is L_2-covariance complete or simply covariance complete.

The following assumptions are sufficient for covariances to be well-defined and for the equivalence with Assumption 2.

Assumption 3 *(moments) The variable Y has bounded second moment.*

Assumption 4 *(overlapping) The function $p(w) = \mathbb{E}[Z|W = w]$ satisfies $0 < p < 1$ a.s.*

The following result follows from the definition of covariance completeness. See the proof in Appendix A.1.

Theorem 2.2 *Let Assumptions 1, 3 and 4 hold. Then, g is point-identified in $\mathcal{G}$ if and only if (X, Z) given W is $\mathcal{G}$-covariance complete.*

To compare covariance completeness with the classic concept of L_2-completeness, we define the latter formally; see Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Andrews (2011) and D'Haultfoeuille (2011) for further discussion on completeness.

Definition 2.2 *We say that the conditional distribution of R given S is $\mathcal{F}-$complete if for each $f \in \mathcal{F}$ the following holds*

$$\mathbb{E}[f(R)|S] = 0 \ a.s. \Longrightarrow f = 0 \ a.s.$$

When $\mathcal{F} = L_2(R)$ we simply say that the distribution of R given S is L_2-complete.

The following result provides a sufficient condition for covariance completeness in terms of traditional completeness. Define $q(x, w) = \mathbb{E}[Z|X = x, W = w]$ and $k(x, w) = q(x, w) - p(w)$. Define the class of measurable functions

$$\mathcal{F} = \{f(x, w) = g(x)k(x, w) : g \in \mathcal{G}\}.$$

The sufficient condition is a simple implication of the definition of covariance and the law of iterated expectations, and therefore its proof is omitted.

Proposition 2.3 *(X, Z) given W is $\mathcal{G}$-covariance complete if the distribution of (X, W) given W is $\mathcal{F}-$complete.*

From Proposition 2.3 a necessary nonparametric relevance condition for covariance completeness is that

$$\mathbb{E}[Z|X, W] \neq \mathbb{E}[Z|W] \text{ a.s.} \tag{10}$$

That is, covariance completeness can be understood as a weighted completeness between the endogenous variables X and the covariates W, and the necessary condition (10) requires that X has to be a nonparametrically significant predictor of Z conditional on W, so that the weights $k(x, w)$ are nonzero. These restrictions on k can be relaxed if $\mathcal{G}$ includes separametric or parametric assumptions.

Remark 2.1 *Covariance completeness imposes restrictions on the support of covariates relative to that of endogeneous variables, while L_2-completeness of traditional IV imposes restrictions on the support of the instrument relative to that of endogeneous variables. In general, a necessary condition for covariance completeness is that both X and W have the same level of complexity (e.g. both are continuous). The following example illustrates this point and compares the two (equivalent) rank conditions.*

Example 2.3 *(X and W discrete, Example 2.1 cont.) Under Assumption 4, the system* $\mathbf{m} = \mathbf{Ag}$*, is equivalent to* $\mathbf{c} = \mathbf{Cg}$*, where* $\mathbf{c} = \mathbf{Dm}$*,* $\mathbf{C} = \mathbf{DA}$*, and* $\mathbf{D}$ *is a full-rank diagonal matrix with i-th diagonal term* $p(w_i)(1 - p(w_i)) > 0$*,* $i = 1, ..., l$*. Here* $L_2(X)$ *can be identified with* $\mathbb{R}^q$*. Let* $\mathcal{G}$ *be the subspace of vectors in* $\mathbb{R}^q$ $\mathbf{g} := (g(x_1), ..., g(x_q))'$ *such that* $g(x_1) = 0$ *(here* $\bar{x} = x_1$ *without loss of generality). Then,* $\mathcal{G}$ *satisfies covariance completeness in this example iff the homogeneous system* $\mathbf{Cg} = \mathbf{0}$ *has a unique solution in* $\mathcal{G}$*, which boils down to rank*$(\mathbf{C}) = q - 1$ *or equivalently rank*$(\mathbf{A}) = q - 1$*. Again, covariance completeness requires the order condition* $l \geq q - 1$*, so it restricts the support of covariates* W *relative to that of endogenous variables* X*. As mentioned earlier, traditional IV requires that the suport of the instrument is larger than* q *for identification of* $\mathbf{g}$.

For certain distributions and classes of functions $\mathcal{G}$, simple conditions for covariance completeness can be established, as the following examples illustrate.

Example 2.4 *(Gaussian variables) Suppose that* (X, W) *is jointly normal conditionally on a binary IV* Z*, i.e.*

$$(X, W)|Z \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_Z \\ \rho_Z & 1 \end{pmatrix}\right).$$

Following Dunker, Florens, Hohage, Johannes and Mammen (2014), we can compute

$$\mathbb{E}[g(X)|W = w, Z = z] = (2\pi)^{-3/4} \exp\left(-\frac{w^2}{2}\right) \sum_{j=0}^{\infty} \mu^j(\rho_z) \mathbb{E}[g(X)p_j(X)] \frac{w^j}{\sqrt{j!}},$$

where p_j *are the Hermite functions,* $p_j(x) = (j!2\pi)^{-1/2} \exp\left(-0.5x^2\right) H_j(x)$*, with* H_j *the j-th Hermite polynomial, and* $\mu(\rho) = \rho/\sqrt{1 - \rho^2}$*. Therefore,*

$$Ag(w) = (2\pi)^{-3/4} \exp\left(-\frac{w^2}{2}\right) \sum_{j=0}^{\infty} \left\{\mu^j(\rho_1) - \mu^j(\rho_0)\right\} \mathbb{E}[g(X)p_j(X)] \frac{w^j}{\sqrt{j!}}.$$

By the completeness of the Hermite polynomials, L_2*-covariance completeness in this context translates into* $\rho_1 \neq \rho_0$*. Notice that if* $f_{X|Z,W}$ *denotes the density of* X *conditional on* Z *and* W*, then*

$$f_{X|Z=z,W=w}(x) = \frac{1}{\sqrt{2\pi(1 - \rho_z^2)}} \exp\left(-\frac{(x - \rho_z w)^2}{2(1 - \rho_z^2)}\right).$$

Therefore, the condition $\rho_1 \neq \rho_0$ *is equivalent to saying that the difference between the distribution of* X *when* $Z = 1$ *and when* $Z = 0$ *varies with* W*.*

Example 2.5 *(Exponential family) To illustrate how Proposition 2.3 can be applied consider the case where* $q(\cdot)$ *satisfies an index restriction. That is, let* d_w *denote the dimension of* W *and assume* $\mathbb{E}[Z|X, W] = \mathbb{E}[Z|X, \eta(W)]$ *for a measurable function* $\eta : \mathbb{R}^{d_w} \to \mathbb{R}^{d_\eta}$*. Set* $a(W) = (\eta(W), p(W))'$*, and assume that a.s. the distribution of* X *given* W *is absolutely continuous with density*

$$f_{X|W=w}(x) = s(x, a(w))t(w) \exp\left(\mu(w) \times \tau(x, a(w))\right),$$

with $s(x, a(w)) > 0$*,* $\tau(x, a(w))$ *is one-to-one in* x*, and the support of* $\mu(w)$ *given* $a(w)$ *is an open set, then if (10) holds, covariance completeness is satisfied. This follows from an application of Theorem 2.2 in Newey and Powell (2003) and Proposition 2.3 above.*

Example 2.6 *(Linear multivariate model) Suppose $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$, so that the model is*

$$Y = \beta' X + U, \quad \mathbb{E}[U|W, Z] = \mathbb{E}[U|W], \tag{11}$$

where X is a d-dimensional vector which does not include a constant and Z is binary. Note that U is not required to have zero mean, so it may include an intercept and/or functions of W. In this model, $\mathcal{G}-$covariance completeness is equivalent to a unique solution for β in the equation

$$\mathbb{E}[Y|W, Z = 1] - \mathbb{E}[Y|W, Z = 0] = \beta' \left(\mathbb{E}[X|W, Z = 1] - \mathbb{E}[X|W, Z = 0] \right),$$

or in short (using the generic notation $\Delta_\xi = \mathbb{E}[\xi|W, Z = 1] - \mathbb{E}[\xi|W, Z = 0]$)

$$\Delta_Y = \beta' \Delta_X. \tag{12}$$

Hence, $\mathcal{G}-$covariance completeness is equivalent to

$$\mathbb{E}[\Delta_X \Delta_X'] \text{ is positive definite.}$$

It is straightforward to see why β is identifiable under this condition, since from equation (12) $\beta = \left(\mathbb{E}[\Delta_X \Delta_X']\right)^{-1} \mathbb{E}[\Delta_X \Delta_Y]$. In practice, this condition requires that the "first stages" of the several elements in the vector X vary with W in a linearly independent manner. Notice that in linear models we can relax the conditions on the complexity of W. For example, even though X is multivariate, W may be univariate (though it must assume at least $d - 1$ different values).

Remark 2.2 *(Adding nonseparable covariates to the model) In this section's separable model, a necessary condition is that W must be additively separable from X. It is important to notice that not all covariates in the model need to be structurally separable from X, just those which will be used as W in our identification approach. To better understand how to make this choice, consider the extended model*

$$Y = g(X, W_c) + U,$$

where $\mathbb{E}[U|Z, W_c, W] = \mathbb{E}[U|W]$. In this model the researcher chose to separate the covariates into two groups. The variables W_c are included into the model as exogenous controls. The variables W are used as the classification variable.

The first question is why would the researcher include W_c in the first place. In some cases it may be important to include such controls either because of suspected nonseparable effects or because it is hard to argue the validity of the IV unless it is conditional on W_c as well.

The next question is which variables should be used as W and which should be used as W_c. There is a trade-off: on the one hand, W can be endogenous, while W_c must be exogenous. On the other hand, W must be separable from X, while W_c may interact with X in arbitrary ways.

Notice that W needs to be only as complex as X, and so, for example if X is one continuous variable, it suffices to find just one continuous covariate which can be argued to be separable from X. The application Section 5 provides an explicit example of such considerations in an empirical problem. Note that if we assume a semiparametric structure for g, for example if $g(X, W_c) = \beta(W_c)'X$, then we can often drop the exogeneity requirement for W_c (see the example below).

Example 2.7 *(Linear model with heterogeneous effects in W_c) Consider the varying coefficient model*

$$Y = \beta(W_c)'X + U, \quad \mathbb{E}[U|Z, W_c, W] = \mathbb{E}[U|W_c, W],$$

where X is a d-dimensional vector. In this model, covariance completeness holds if

$$\mathbb{E}[\Delta_X(w_c, W)\Delta_X'(w_c, W)] \textit{ is positive definite for a.s. } w_c.$$

Under this covariance completeness assumption, we can estimate $\beta(\cdot)$ nonparametrically from local least squares regressions. Alternatively, we could specify $\beta(W_c)$ as a linear function of W_c, say $\beta(W_c) = \beta_0 + \beta_1' W_c$, which results in a linear-in-parameters model with endogenous variables X and interactions between X and W_c, which can be dealt with as in Example 2.6 above. By comparing $\beta(\cdot)$ with the estimator obtained from $\beta = (\mathbb{E}[\Delta_X \Delta_X'])^{-1}\mathbb{E}[\Delta_X \Delta_Y]$ we can test for heterogenous marginal effects in W_c. In the linear specification, this can be done by simply testing if $\beta_1 = 0$. Section 3 below considers a generalization that allows for unobserved as well as observed heterogeneity of marginal effects of the form $\partial m(x, h(W, \varepsilon))/\partial x$ at $x = X$, for nonparametric functions m and h.

2.2 Estimation

In this section we discuss estimation when g and $\mathbb{E}[U|W]$ are linear in parameters. To see the discussion of the estimation in the nonparametric case, refer to Appendixes A.2.1 and A.2.2. As shown in the Appendix, the nonparametric sieve estimator is also linear in parameters, therefore the implementation of this section is also relevant for the nonparametric case.

Our identification strategy in the linear multivariate model is based on the identity $\Delta_Y = \beta'\Delta_X$ as derived in Example 2.6. This identity suggests a three-step estimator for β : Step 1, estimate Δ_X by $\hat{\Delta}_X$ using regression methods; Step 2, estimate Δ_Y by $\hat{\Delta}_Y$ using regression methods; and Step 3, run a regression of $\hat{\Delta}_Y$ on $\hat{\Delta}_X$ by ordinary least squares (OLS) to obtain an estimate $\widehat{\beta}_{3Step}$.

It turns out that this estimation strategy can be easily implemented as a TSLS. Specifically, given a random sample $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ of (Y, X, W, Z), we propose to estimate β with the coefficient of the X_i on a TSLS regression of Y_i onto X_i and W_i, using Z_i and $Z_i W_i$ as instruments for X_i, and treating W_i as exogenous. This estimator, which we denote by $\widehat{\beta}$, can be implemented with off-the-shelf econometric software. Additionally, the standard errors are correctly estimated as the standard errors of the TSLS regression proposed above, without the need for any correction.

To see why $\widehat{\beta}_{3Step} = \widehat{\beta}$, suppose that $g(X) = \beta'X$ and $\mathbb{E}[U|W] = \alpha + \gamma'W$. Then, we can write the model (1), by adding and subtracting $\mathbb{E}[U|W]$, as

$$Y = \alpha + \beta'X + \gamma'W + u, \tag{13}$$

where $u = U - \mathbb{E}[U|W]$. We can see from this representation that W is different from a classical "control" variable, since X is still correlated with u even after controlling for W. Note that our validity condition, $\mathbb{E}[U|W, Z] = \mathbb{E}[U|W]$, can be equivalently written as

$$\mathbb{E}[u|W, Z] = 0 \text{ a.s.}, \tag{14}$$

and thus β in (13) can be identified as in a standard instrumental variable model, and estimated with a TSLS regression as $\hat{\beta}$ described above, provided the relevance condition holds; see Theorem 2.4 below.

Explicitly, let the "first-stage" regression fitted value be $\widehat{\mathbb{E}}[X|W,Z] = \widehat{\alpha}_{0X} + \widehat{\alpha}_{1X}Z + \widehat{\alpha}'_{2X}W + \widehat{\alpha}'_{3X}WZ$ and the "reduced-form" fitted value be $\widehat{\mathbb{E}}[Y|W,Z] = \widehat{\alpha}_{0Y} + \widehat{\alpha}_{1Y}Z + \widehat{\alpha}'_{2Y}W + \widehat{\alpha}'_{3Y}WZ$, then it is well known that the TSLS estimator is related to the reduced form fits through the equation

$$\widehat{\mathbb{E}}[Y|W,Z] = \widehat{\alpha} + \widehat{\beta}'\widehat{\mathbb{E}}[X|W,Z] + \widehat{\gamma}'W.$$

If we evaluate this empirical equation at $Z = 1$ and $Z = 0$ and subtract, we arrive at

$$\hat{\Delta}_Y = \widehat{\beta}'\hat{\Delta}_X.$$

Thus, by definition of OLS, $\widehat{\beta}_{3Step}$ must be equal to $\widehat{\beta}$.

The next result shows the consistency of $\widehat{\beta}$ under our identification assumptions. In particular, the result shows that endogeneity of W does not affect the consistency of the TSLS estimator $\widehat{\beta}$. Its proof can be found in the Appendix A.1.

Theorem 2.4 *Let Assumptions 1, 3, 4 and (14) hold. If (X,Z) given W is $\mathcal{G}$-covariance complete with $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$ then*

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d N(0, \Sigma),$$

where Σ is the classical TSLS asymptotic variance (which is assumed to be finite).

Remark 2.3 *Note that the TSLS can be applied without any modification in the cases when Z is not binary.*

Remark 2.4 *The same arguments as those used in the proof of Theorem 2.4 show that, when X is univariate, the standard TSLS that treats W as exogenous but does not consider interactions is in fact robust to endogeneity of W under our conditions.*

Remark 2.5 *Our TSLS will be consistent for β even when $\mathbb{E}[X|Z,W]$ is non-linear, as long as $g(X)$ is a linear function of X and the conditions above hold. The proof of Theorem 2.4 shows formally this robust property of the TSLS, which is later confirmed in simulations.*

Remark 2.6 *In the parametric setting of this section, the relevance condition is a standard TSLS rank condition that can be tested by traditional methods. The order condition here is that dimension of W needs to be at least $d-1$ (the dimension of X minus one). Therefore, if there are several variables which satisfy the identification conditions, we recommend that W be chosen as the variable (or variables) for which Z and ZW make up the strongest instruments. Additionally, comparisons of results using different W are an informal test of the functional form assumption.*

3 Nonseparable Case

This section extends our previous identification strategy to the nonseparable model (3), repeated here for convenience:

$$Y = m(X, U), \tag{15}$$

where $m(x, u)$ is a strictly increasing function in u, for each x in $\mathcal{S}_X$. The following example motivates this model in an economic setting.

Example 3.1 *Consider the following extension of the model in Imbens and Newey (2009) (where W was absent). Let Y denote some outcome such as firm revenue or individual lifetime earnings. Let X be inputs chosen by the agent, and let $h(W, \varepsilon)$ represent other inputs that are at most partially observed by agents (W is observed but ε is not). The agent chooses X by maximizing expected outcome, minus the cost associated with choosing X given her information set. At the time of the decision on X, the agent already observes W, which in turn was used to produce the inputs $h(W, \varepsilon)$. The variable W can be endogenous in the sense of being dependent on ε (as is likely to be the case, since W and ε are inputs used to produce $h(W, \varepsilon)$). In addition, the agent also observes a cost shifter Z and a vector of shocks ε_x (proxies for ε observed by the agent, but not by the econometrician). The cost function is $C(x, w, z)$. The input X is the solution to the economic agent's problem*

$$X = s(W, Z, \varepsilon_x) \equiv \arg\max_{x^*} \left\{ \mathbb{E}[m(x^*, h(W, \varepsilon))|W, \varepsilon_x] - C(x^*, W, Z) \right\}.$$

The final outcome is given by $Y = m(X, h(W, \varepsilon))$. Here $U = h(W, \varepsilon)$ is the partially unobserved input.

What is special about this setup is that (i) production is monotonic in U; (ii) there are some observed factors W and some unobserved factors ε which only enter the production function as components of one of its inputs, h, and (iii) there exist another observable variable Z which influences the choice of X (a cost shifter), but is excluded from the production of Y, in the sense that it is neither a direct input in the production of Y, nor a direct or indirect input in the production of h ($Z \perp \varepsilon|W$). As we will show below, properties (i) to (iii) are fundamental to our identification strategy. Properties (i) and (ii) imply the structural separability in covariates that we need for our approach. The fact that prior imputs W do not enter final production in arbitrary ways helps to justify our structural separability assumption, and imply the type of exclusion restrictions we exploit. In (iii) we further require that first stages are heterogenous. Informally speaking, the cross derivative of $s(w, z, \varepsilon_x)$ in w and z is non-zero. Conditions for this are hard to find analytically, but this condition is likely to hold if the marginal cost has cross variation in w and z, i.e. the cross derivative $\partial^3 C(x, w, z)/\partial x \partial w \partial z$ is non-zero.

We now investigate identification in the nonseparable model (see how to implement our method in the nonseparable case in Appendix A.2.3). We show that a similar identification strategy as used for separable models allows for nonparametric identification in this more general setting. Suppose that Assumption 1 holds. Let m^{-1} denote the inverse of m with respect to the u argument, so that $m^{-1}(Y, X) = U$ a.s., then

$$\mathbb{C}\left(m^{-1}(Y, X), Z|W\right) = 0 \ a.s. \tag{16}$$

We note that these restrictions are valid for a general instrument Z, not necessarily binary. It turns out that a simple reparametrization transforms the nonseparable case into a problem with a similar mathematical structure as that of the separable case, but where (Y, X) replaces X. That is, defining

$$g(Y, X) := Y - m^{-1}(Y, X),$$

the homogeneous system in (16) can be written as

$$\mathbb{C}(Y, Z|W) = \mathbb{C}(g(Y, X), Z|W). \tag{17}$$

Then, identifying g from this equation is equivalent to identifying m in (16). In the nonseparable case, however, under the standard conditional independence assumption considered in the literature

$$Z \perp U|W, \tag{18}$$

there is an additional normalization assumption we need to impose. Following Matzkin (2003), we introduce the following normalization $m^{-1}(y, \bar{x}) = y$ for all $y \in \mathcal{S}_Y$ and some known $\bar{x} \in \mathcal{S}_X$, which after our reparametrization is equivalent to the convenient $g(y, \bar{x}) = 0$.

Thus, we introduce formally our identification conditions for the nonseparable case. Let $\mathcal{G}$ be a subset of $L_2(Y, X)$ with the properties: (i) if $g_1, g_2 \in \mathcal{G}$ then $g_1 - g_2 \in \mathcal{G}$; and (ii) elements in $\mathcal{G}$ satisfy the normalization restrictions, i.e. if $g \in \mathcal{G}$ then $g(y, \bar{x}) = 0$. Note that the normalization rules out the trivial solution $g(y, x) = y$ of (17).

Definition 3.1 *We say (Y, X, Z) given W is $\mathcal{G}$-covariance complete if for each $g \in \mathcal{G}$*

$$Cov\,(g(Y, X), Z|W) = 0 \Longrightarrow g = 0 \;\; a.s.$$

The proof of the next theorem is the same as that of Theorem 2.2 and therefore is omitted.

Theorem 3.1 *Suppose (15), (18) and Assumption 3 hold. Then, g is point-identified in $\mathcal{G}$ if (Y, X, Z) given W is $\mathcal{G}$-covariance complete.*

Remark 3.1 *Theorem 3.1 provides sufficient conditions for identification in the nonseparable case. These conditions are, however, not necessary for two reasons. First, these conditions do not exploit that $y - g(y, x)$ (i.e. $m^{-1}(y, x)$) is monotonic in y, for each $x \in \mathcal{S}_X$. Second, they do not exploit higher order implications from the conditional independence (18). A method to incorporate the latter is described in the Appendix A.3. Thus, identification of g may hold under more general conditions than those given in Theorem 3.1.*

We extend Example 2.1 to the nonseparable case. For simplicity, we consider the binary IV case, other cases can be treated similarly.

Example 3.2 *(W discrete) Suppose $Z \in \{0, 1\}$ is a binary instrument. To simplify notation denote $V = (Y, X)$ and its support by $\mathcal{S}_V := \{v_1, ..., v_q\}$, and let $\mathcal{S}_W := \{w_1, ..., w_l\}$ denote the support of*

W, with $q < \infty$ and $l < \infty$. Without loss of generally the normalization restrictions are $g(v_j) = 0$, $1 \leq j \leq q_y$, where q_y denotes the cardinality of the support of Y. Likewise, let q_x denote the cardinality of the support of X, so $q = q_y q_x$. Covariance completeness and the normalization restrictions imply the homogenous system of linear equations

$$\mathbf{B}g = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{g} = 0,$$

where with some abuse of notation, we denote by $\mathbf{g} = (g(v_1), ..., g(v_q))'$ the q-dimensional vector of parameters of interest, $\mathbf{B}_1$ is a $(l + q_y) \times q$ matrix with $ij - th$ element

$$b_{1ij} = \mathbb{P}[V = v_j | Z = 1, W = w_i] - \mathbb{P}[V = v_j | Z = 0, W = w_i],$$

and $\mathbf{B}_2 = \begin{bmatrix} I_{q_y} & 0 \end{bmatrix}$ is a $q_y \times q$ matrix, with I_{q_y} denoting the identity matrix of dimension q_y. In the discrete case a necessary and sufficient condition for covariance completeness when $\mathcal{G}$ is unrestricted (beyond the normalization constraints) is then

$$rank\,(\mathbf{B}) = q - 1.$$

The order condition is $l \geq q_y(q_x - 2) - 1$. If covariance completeness fails, still our method provides identification power depending on the value of $rank\,(\mathbf{B})$ in $q_y \leq rank\,(\mathbf{B}) < q$. With our method we can identify linear functionals $c'g$ with c in a space of dimension $rank(\mathbf{B})$.

In general, necessary conditions for covariance completeness in the nonparametric nonseparable case are that (Y, X) and W have the same level of complexity. For example, if Y is continuous but X is discrete, then W needs to be continuous, which is a different requirement than the one in separable models, where the support of Y did not play a role.

When the model is nonseparable but the assumption of monotonicity does not hold, our identification strategy still identifies a weighted marginal effect, as shown in the following example.

Example 3.3 *(Random coefficients). Consider the random coefficient model*

$$Y = \beta_0 + \beta_1 X,$$
$$X = \alpha_0 + \alpha_1 Z,$$

where now $\beta = (\beta_0, \beta_1)'$ and $\alpha = (\alpha_0, \alpha_1)'$ are random coefficients, satisfying with $\theta = (\beta', \alpha')'$,

$$Cov\,(\theta, Z|W) = 0 \ a.s.$$

Define the conditional variance $\sigma^2(W) = Var\,(Z|W)$, and for a generic random variable ζ,

$$\Delta_\zeta = \frac{\mathbb{C}\,(\zeta, Z|W)}{\sigma^2(W)}.$$

Then, in the random coefficients model above

$$\Delta_Y = \mathbb{E}[\alpha_1 \beta_1 | W]$$

and

$$\Delta_X = \mathbb{E}[\alpha_1 | W].$$

Therefore, under covariance completeness

$$\beta = \mathbb{E}[w(W, \alpha_1)\beta_1],$$

where

$$w(W, \alpha_1) = \frac{\mathbb{E}[\alpha_1|W]\alpha_1}{\mathbb{E}[\mathbb{E}[\alpha_1|W]\alpha_1]}$$

are weights that integrate up to one. Therefore, our estimand has an interpretation as a weighted marginal effect in this random coefficients model. Note that without further assumptions the weights can be negative, although they are conditionally positive in the sense that $\mathbb{E}[w(W, \alpha_1)|W] > 0$ *a.s. If* $\alpha_1 \perp \beta_1 | W$ *then* $\beta = \mathbb{E}[\beta_1]$.

4 Monte Carlo Simulations

This Section investigates the finite sample performance of our TSLS proposed in Section 2.2. In particular, we aim to investigate the sensitivity of our TSLS to the endogeneity of covariates and misspecification of the first stages.

We begin with a linear model with three endogenous variables and one binary instrument:

$$\begin{aligned} Y &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_W W + u, \\ W &= \alpha_W u + U_W, \\ X_1 &= \alpha_{01} + \alpha_{11} Z + \alpha_{21} W + \alpha_{31} Z \cdot W + U_1, \\ X_2 &= \alpha_{02} + \alpha_{12} Z + \alpha_{22} W + \alpha_{32} Z \cdot W + U_2, \end{aligned}$$

where (u, U_W, U_1, U_2) are independent standard normal random variables independent of Z, which is distributed as Bernoulli random variable with probability $p = 0.5$. The classical order condition of standard IV does not hold in this model, and hence, classical IV is unable to identify the marginal effects β_1 and β_2. For identification of the marginal effects, our method requires that Assumption 2 holds, which in this case is equivalent to $\mathbb{E}[\Delta_X \Delta_X']$ being positive definite (see Example 2.6) or explicitly

$$\det \begin{vmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{12} & \alpha_{32} \end{vmatrix} \neq 0.$$

The parameters in the structural equation are set at $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 2$ and $\gamma_W = 1$. We set $\alpha_{01} = \alpha_{02} = \alpha_{21} = \alpha_{22} = 0$, and $\alpha_W = 1$, so W is endogenous. Table 2 provides the average bias and Mean Squared Error (MSE) based on 10,000 Monte Carlo simulations, sample sizes $n = 100$, 300, 500 and 1000, and several values for $(\alpha_{11}, \alpha_{31}, \alpha_{12}, \alpha_{32})$. There are three variables in the structural equation and three in each reduced form equation, and therefore, the TSLS estimator is an IV estimator that treats Z and the interaction term $Z \cdot W$ as instruments. We note that Table 2 does not offer a comparison with existing methods such as IV because they are not applicable (e.g. classical IV's order condition fails).

We observe a satisfactory bias performance uniformly over all parameter values. For the first two cases ($\alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1$) and ($\alpha_{11} = 0, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 0$) the sample variance of the estimators is already small for small sample sizes as 100, and it decreases to zero with the sample size, in accordance with the consistency of the estimator. For the parameter values ($\alpha_{11} = 1.25, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 1.25$) identification is much weaker, and larger sample sizes are required for a good performance, as expected. Overall, these results provide supporting evidence of the robustness of our identification strategy to the endogeneity of W.

Table 2: **IV Case**

α_{11}	α_{31}	α_{12}	α_{32}	n	$Bias(\beta_1)$	$MSE(\beta_1)$	$Bias(\beta_2)$	$MSE(\beta_2)$
1	0	0	1	100	-0.0041	0.0252	0.0013	0.0123
				300	0.0015	0.0070	0.0000	0.0034
				500	-0.0014	0.0041	0.0000	0.0020
				1000	0.0001	0.0020	0.0003	0.0010
0	1	1	0	100	-0.0002	0.0121	0.0009	0.0243
				300	-0.0004	0.0035	0.0002	0.0070
				500	0.0002	0.0020	0.0000	0.0041
				1000	-0.0001	0.0010	0.0004	0.0020
1.25	1	1	1.25	100	0.1247	259.7900	-0.0897	193.4900
				300	0.0528	11.1740	-0.0434	7.2709
				500	-0.0065	0.2693	0.0053	0.2085
				1000	0.0007	0.0157	-0.0003	0.0136

10000 Monte Carlo Simulations.

In the second set of simulations we show how the estimator performs when the first stages are not linear. Consider now the DGP:

$$\begin{aligned} Y &= \alpha + \beta_1 D + \beta_2 D 1(D > 0) + \gamma_W W + u, \\ W &= \alpha_W u + U_W, \\ D &= \alpha_d W + \gamma_d Z \cdot W + U_d, \end{aligned}$$

where (u, U_W, U_d) are independent standard normal random variables, independent of Z, which is again distributed as Bernoulli with probability $p = 0.5$. This corresponds to a linear model

$$Y = \alpha + \beta_0' X + U,$$

where $\beta_0 = (\beta_1, \beta_2)'$, $X = (D, D1(D > 0))'$ and $U = \gamma_W W + u$. Here

$$\mathbb{E}[D|W, Z = 1] - \mathbb{E}[D|W, Z = 0] = \gamma_d W,$$

so γ_d controls the identification strength. Since there is only one binary IV, Z, standard IV methods cannot be applied in this example. Note also that under this DGP the difference of conditional means Δ_X is nonlinear in W in its second component. In Remark 2.5 we discuss that our estimator is still consistent, and we illustrate this in the present experiment. This shows the robustness of our estimator to the failure of the linearity assumption in the conditional mean $\mathbb{E}[X|W, Z]$.

Table 3 provides the average bias and MSE based on 10000 Monte Carlo simulations. In all cases $\alpha = 0$, $\gamma_W = 1$, $\beta_1 = 1$, $\beta_2 = 2$, $\alpha_W = 1$, $\alpha_d = 1$. We consider two levels of identification, "moderate" ($\gamma_d = 1$) and "high" ($\gamma_d = 2$).

Table 3: **IV Case - Misspecified Model**

γ_d	n	$Bias(\beta_1)$	$MSE(\beta_1)$	$Bias(\beta_2)$	$MSE(\beta_2)$
1	100	0.00107	0.48403	-0.01446	2.64655
	300	-0.00067	0.01124	0.00039	0.03006
	500	0.00001	0.00638	0.00042	0.01664
	1000	0.00003	0.00303	-0.00072	0.00802
2	100	-0.00160	0.00924	0.00320	0.02321
	300	-0.00065	0.00252	0.00089	0.00645
	500	0.00000	0.00148	0.00024	0.00385
	1000	0.00031	0.00074	-0.00019	0.00189

10000 Monte Carlo Simulations.

The reported results show that the estimator is still consistent even though the conditional means are not linear. Estimates of β_2 require larger sample sizes than those of β_1 to achieve the same level of precision and bias performance. There is an efficiency loss in estimating β_2 relative to β_1, probably due to the nonlinearities in $\mathbb{E}[D1(D > 0)|W, Z = j]$ for $j = 0, 1$. As expected, the results improve with the identification strength. In sum, these simulations provide finite sample evidence of a satisfactory performance of the TSLS estimator and its robustness to the endogeneity of the covariates and the nonlinearity of the first stages.

5 An Application to the Estimation of the Effects of Air Pollution on House Prices

We apply our identification strategy to the problem of estimating the effects of pollution on house prices, as in Chay and Greenstone (2005). The concern with endogeneity in this problem is warranted, since counties may differ from each other in many ways which may not be accounted by their observable characteristics and amenities. Chay and Greenstone base their identification strategy on an instrumental variable approach, which takes advantage of the quasi-experiment generated by the Clean Air Act around the time it was first implemented.

As explained in Remark 2.2, even when using a separable model, it is not necessary that all the covariates be separable from X in the model. Let Y denote the change between 1970 and 1980 in the logs of the county's median property values, X is the change between 1970 and 1980 in the geometric mean total suspended particulates (TSP) across all monitors in the county, Z is the county's attainment status in 1975 according to the Clean Air Act, and W_c and W denote vectors of further variables which are used by Chay and Greenstone as controls in their model specification (2) (Chay and Greenstone (2005), p. 411). We can estimate a model

$$Y = g(X, W_c) + U,$$

with the exclusion restriction $\mathbb{E}[U|Z, W_c, W] = \mathbb{E}[U|W]$, thus allowing a set of covariates W_c to be non-separable from X as long as they are exogenous. We can generalize Chay and Greenstone (2005)'s approach in two directions. First, even though the instrument Z is binary we are able to identify g when it is more general than simply linear. Second, since W in our approach may be endogenous, the covariates which we choose to use as W need no longer be assumed exogenous.

In choosing which covariates are part of W, we must be concerned with their separability from X. The issue is that counties' preference for pollution may differ as a function of some of the covariates. Because of this, it is unadvisable to use covariates such as, for example, the percent change in income, education levels, racial composition and unemployment in the county population, as these could be reasonably assumed to influence the population's taste in pollution. The covariates which we believe are most likely to be separable are the county's changes between 1970 and 1980 in the percent spending in highways, health and education. We considered estimators of our method with each of those variables separately and also together, as can be seen in Table 4.

The estimation is done using the same data set as in Chay and Greenstone (2005) as well as the exact same covariate specification. The first column in Table 4 shows the results of a standard IV estimation of the effects of pollution change, which is what is done in Chay and Greenstone (2005). The replicated results are, not surprisingly, identical to that paper. Columns A to D show the results of our estimation approach using different variables as the separable classification covariate W. Row I uses a specification without exogenous control variables. Although in specification I our results are of similar magnitude to the standard IV, they are slightly smaller and vary depending on the covariate. They are particularly smaller when all three covariates are used. We believe that this happens because when covariates W_c are not used, the classic IV operates under the assumption that the IV is unconditionally valid, while our estimator operates under the assumption that the IV is valid only conditional on the separable covariate ($\mathbb{E}[U|Z, W] = \mathbb{E}[U|W]$). That said, Chay and Greenstone never suppose that their IV is valid, but only that it is valid conditional on controls. Row II shows the results conditional on controls. There the identifying assumption of the classic IV approach is that $\mathbb{E}[U|Z, W_c, W] = 0$,

Table 4: **Estimation Results - Linear Case, Binary IV**

	IV	A	B	C	D
I	-.347	-.340	-.327	-.317	-.327
	(.140)	(.138)	(.136)	(.134)	(.135)
II	-.203	-.208	-.202	-.203	-.208
	(.093)	(.094)	(.093)	(.093)	(.093)

Table 4: Columns A to D use our approach with Z equal to the change from 1970 to 1980 in the % spending on highways (A), health (B), and education (C). In column (D), W is the vector of all three variables. Specification I has no exogenous covariates W_c. Specification II uses as exogenous covariates W_c the exact same specification as in Chay and Greenstone (2005) excluding the covariates which we are using as W.

while our identifying assumption is that $\mathbb{E}[U|Z, W_c, W] = \mathbb{E}[U|W]$. Nevertheless, our results generally confirm the estimates found by Chay and Greenstone.

As a robustness check of the separability of the variables we chose as W, we ran the same regressions using each of the other covariates in the model as W instead. The results are extremely similar and thus confirm the separability assumption. The 5 covariates which yielded the most different results are the number of houses built between 1970 and 1980, the rate of vacancies in 1980, change in income per-capita, the change in government revenue per-capita and the change in the fraction of the population with at least a college degree, but even in these cases the differences between our estimates and -.203 was always less than .1. In comparison, Chay and Greenston's RDD estimate is -.275, which they consider to be a confirmation of their results. We note that the observed robustness of our method to the specification of W in this application can be theoretically justified by the overidentification of our method even in cases where classical IV just-identifies.

The comparative advantages of our method are better showcased in the nonlinear case, where the classical instrumental variables methods cannot identify marginal effects with a single instrumental variable. In order to compare our results to Chay and Greenstone's we maintain their specification in all aspects, but allow X to have richer marginal effects on Y. The model is thus $Y = g(X) + W_c'\gamma_c + W'\gamma + U$, assuming that $\mathbb{E}[U|Z, W_c, W] = \mathbb{E}[U|W]$, where g is a connected piecewise linear function and W_c is the entire specification of controls in Chay and Greenstone (2005) except for the variables in W. We recognize that this model is also separable in W_c, but the separability of W_c is incidental, and not fundamental for the identification, as only the terms in W are used as classification covariates in our method. This model is also consistent with Chay and Greenstone's own specification, which provides a closer comparison with their results. Figure 1 has three linear pieces, which connect at the terciles of the distribution of X. Hence, we can write $g(x) = a_1\psi_1^3(x) + a_2\psi_2^3(x) + a_3\psi_3^3(x)$, where the $\psi_j^3(x)$ are the elements of a B-spline basis of degree 1 and smoothness 0 with knots at the terciles just described. In practice this is the same as if we had three endogenous variables $\psi_1^3(X)$, $\psi_2^3(X)$

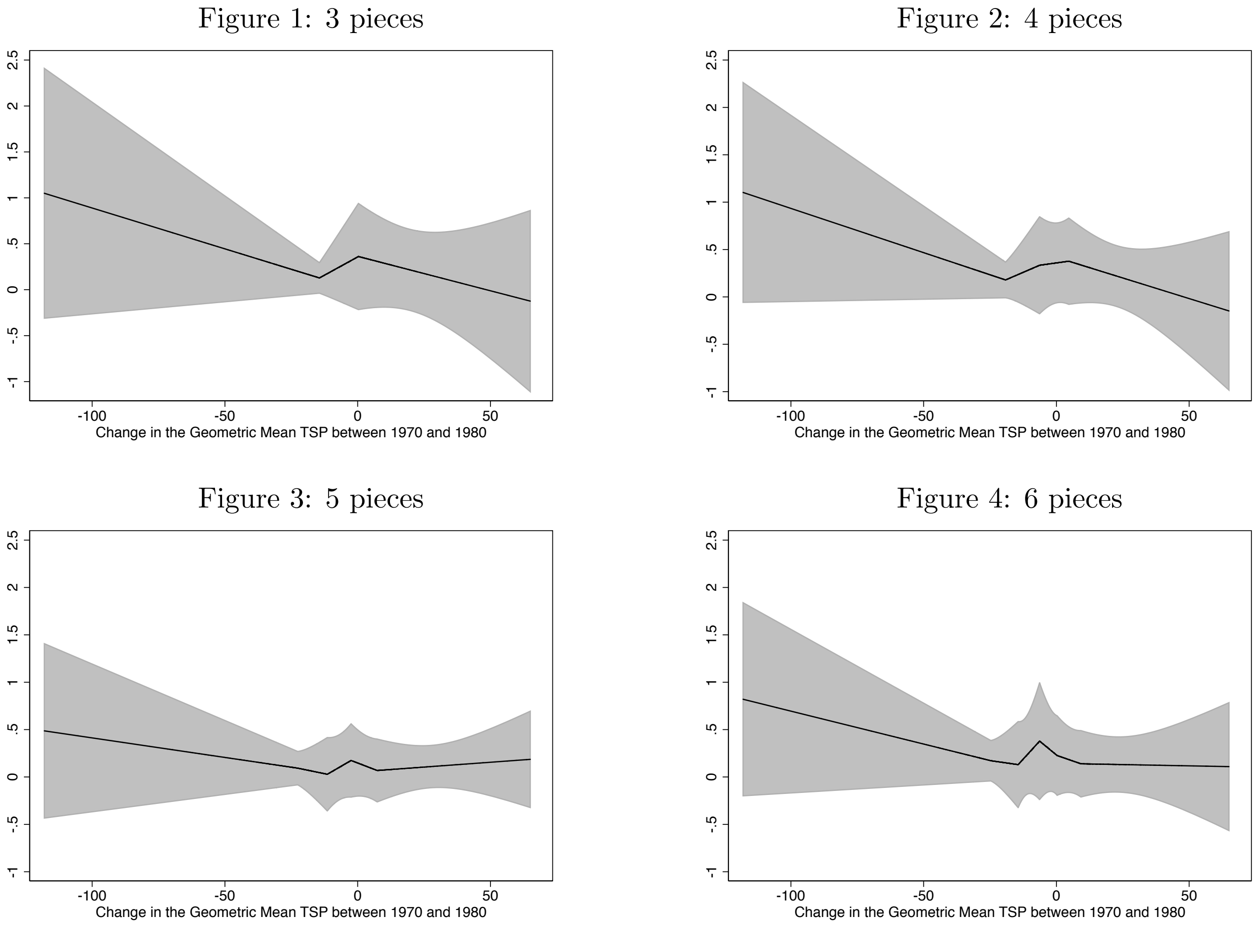

Figure 1: 3 pieces

Figure 2: 4 pieces

Figure 3: 5 pieces

Figure 4: 6 pieces

Figures 1 to 4: Nonlinear case – IV approach. Curves are the results of our approach with exogenous covariates, and Z as in column D in Table 4. The domain of each piece is determined by the quantiles. For instance, Figure 2's pieces connect at the 25th, 50t and 75th quantiles of the change in the geometric Mean TSP between 1970 and 1980.

and $\psi_3^3(X)$. For W we used the three variables in column D in Table 4 (call them $High$, $HealZh$ and $Educ$) expanded into the elements of the B-spline basis. So, with some abuse of notation $W = (\psi_1^3(High), \psi_2^3(High), \psi_3^3(High), \psi_1^3(HealZh), \psi_2^3(HealZh), \psi_3^3(HealZh), \psi_1^3(Educ), \psi_2^3(Educ), \psi_3^3(Educ))'$. Figures 2 to 4 are obtained analogously.

In Figure 1 the standard errors are calculated as the standard error of the predicted $\hat{g}(x) = \hat{a}_1\psi_1^3(x) + \hat{a}_2\psi_2^3(x) + \hat{a}_3\psi_3^3(x)$, so $SE(\widehat{g}(x)) = (\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))'\Omega(\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))$, where Ω is the estimated covariance matrix of $(\hat{a}_1, \hat{a}_2, \hat{a}_3)'$. In a pre-packaged software the standard errors can be obtained directly as the standard errors of the predicted g. Standard errors in Figures 2 to 4 are obtained analogously.

Our results qualitatively confirm the findings of the linear case, but it is important to notice that the effect may be even more negative than predicted in the linear case in the majority of the domain. Also, interestingly, for an important part of the domain the effect seems to go in the opposite direction. In fact, for small reductions in pollution, all regressions show derivatives which are not only positive, but rather high.

Figures 1 to 4 are representative of the patterns we found when we tried other strategies. For example, we also used as W each of the elements used in columns A to C in Table 4 separately (i.e. for g with three pieces we did $W = \left(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High)\right)'$, and we also expanded the elements of W in other ways, for example into two piece B-splines ($W = (\psi_1^2(High), \psi_2^2(High), \psi_1^2(HealZh), \psi_2^2(HealZh),$ $\psi_1^2(Educ), \psi_2^2(Educ))'$), four piece B-splines, etc., all with very similar results. The standard errors get substantially larger as we increase the number of pieces in g, but they are not affected (and seem in fact to decrease) as we increase the number of elements in W.

Our application on the effect of air pollution on house prices confirms the findings of Chay and Greenstone (2005) when a constant marginal effect model is considered, but also uncovers substantial heterogeneity in the effect of pollution on house prices when richer marginal effects are entertained. The impact of air quality on house prices is much larger for counties that significantly change their behaviour as a result of the Clean Air Act than for other counties that experience a minor decrease or an increase in pollution during the 1970-1980 period. Thus, our results are consistent with a nonparametric local average treatment effect interpretation where for the population of compliers the marginal effect is larger than the overall average marginal effect (averaged over the whole population) documented in Chay and Greenstone (2005).

6 Conclusions

In this paper we have proposed a strategy to identify marginal effects of "complex" variables using a lower dimension IV, in the presence of other possible endogenous covariates. The strategy hinges on the heterogeneity of the "first stages" and a structural separability of some covariates, and it can be extended to nonseparable models with unobserved marginal effects. It can be applied to parametric, semiparametric and nonparametric settings. In models that are linear in parameters (which also include nonparamatric models estimated by sieves), the identification strategy can be implemented with a simple TSLS estimator that treats the covariates as if they were exogenous, and runs a first stage with interactions between the Binary IV and the covariates. Thus, our identification strategy can be readily implemented with off-the-shelf econometric software. Monte Carlo simulations show that this TSLS estimator performs well in practice, and it is robust to endogeneity of covariates and misspecification of the first stage linear conditional expectation.

There are several extensions of our methods that deserve further investigation. First, as the first version of this paper shows, the proposed methods can be extended to the Regression Discontinuity Design (RDD) setting. In that setting a practically convenient semiparametric estimator is proposed that uses a varying coefficients specification of the first stages. Identification and semiparametric estimation in the RDD setting will be investigated in a companion paper.

A Appendix

A.1 Proofs of the Main Results

Proof of Theorem 2.1: Note that $A\widetilde{g} = m$, with $\widetilde{g}(X) = g(X) - \mathbb{E}[g(X)]$ and A is invertible on the orthocomplement of $\mathcal{N}(A)$, which by Assumption 2 is given by $\mathcal{N}^{\perp} = \{\lambda \in \mathcal{G} : \mathbb{E}[\lambda(X)] = 0\}$. Thus, since $\widetilde{g} \in \mathcal{N}^{\perp}$ it holds that $\widetilde{g} = A^{-1}m$ and therefore $\widetilde{g}$ is identified. ■

Proof of Theorem 2.2: Note that under Assumption 4, the equation $m = Ag$ is equivalent to

$$\mathbb{C}\left(Y, Z|W\right) = \mathbb{C}\left(g(X), Z|W\right). \tag{19}$$

Suppose (X, Z) given W is $\mathcal{G}$-covariance complete and let $g_1 \in \mathcal{G}$ such that

$$\mathbb{C}\left(Y, Z|W\right) = \mathbb{C}\left(g_1(X), Z|W\right) \text{ a.s.}$$

Thus, $\mathbb{C}\left(g_1(X) - g_0(X), Z|W\right) = 0$ a.s. Then, property (i) above and covariance completeness imply $g_0 = g_1$ a.s. This proves identification. The reciprocal also holds. Suppose that (X, Z) given W is not $\mathcal{G}$-covariance complete, then we can find $g_1 \in \mathcal{G}$, $g_1 \neq 0$, such that $\mathbb{C}\left(g_1(X), Z|W\right) = 0$. Then, identification of g_0 in $\mathcal{G}$ fails, because $g_2 \equiv g_0 + g_1 \in \mathcal{G}$ by property (ii) above and g_2 satisfies (19). ■

Proof of Theorem 2.4: We prove the consistency of the TSLS estimator. Its asymptotic distribution follows standard arguments, which are therefore omitted. TSLS identifies the coefficients of the regression of Y on a constant, X^* and W, where X^* is the population fitted value

$$X^* = \alpha_{0X} + \alpha_{1X}Z + \alpha_{2X}'W + \alpha_{3X}'WZ.$$

By the Frisch-Waugh-Lowell Theorem, the slope of X^*, say β^*, is given by

$$\beta^* = \left(\mathbb{E}[\Pi_W \Pi_W']\right)^{-1} \mathbb{E}[\Pi_W Y],$$

where $\Pi_W = X^* - \mathbb{E}[X^*|W] = \mathbb{E}[X|Z, W] - \mathbb{E}[X|W]$. We prove that is legitimate to take the inverse of $\mathbb{E}[\Pi_W \Pi_W']$ under our conditions, by showing that $\mathbb{E}[\Pi_W \Pi_W']$ is invertible if and only if $\mathbb{E}[\Delta_X \Delta_X']$ is invertible. To see that, suppose $\mathbb{E}[\Delta_X \Delta_X']$ is singular. Then, there exists a $\lambda \neq 0$ such that, a.s.

$$\mathbb{E}[\lambda' X|Z = 1, W] = \mathbb{E}[\lambda' X|Z = 0, W].$$

Then,

$$\mathbb{E}[\lambda' X|Z, W] = \mathbb{E}[\lambda' X|W],$$

and therefore, $\lambda'\Pi_W = 0$ a.s. (i.e. $\mathbb{E}[\Pi_W \Pi_W']$ is singular). The reciprocal follows the same arguments.

Then, using that $\mathbb{E}[\Pi_W] = \mathbb{E}[\Pi_W W] = \mathbb{E}[\Pi_W u] = 0$, and substituting (13) into β^*, yields

$$\begin{aligned}
\beta^* &= \left(\mathbb{E}[\Pi_W \Pi_W']\right)^{-1} \mathbb{E}[\Pi_W X']\beta \\
&= \left(\mathbb{E}[\Pi_W \Pi_W']\right)^{-1} \mathbb{E}[\Pi_W \mathbb{E}[X'|Z, W]]\beta \\
&= \beta,
\end{aligned}$$

thereby proving the consistency of the TSLS for β. We note that the arguments above do not depend on the linearity of $\mathbb{E}[X|Z,W]$, as we can replace the expectation operator above by the linear projection operator without affecting the conclusions. That is, with $\mathbb{L}[X|W]$ denoting the linear projection of X on W (and similarly for other variables), it follows that $X^* = \mathbb{L}[X|Z,W,ZW]$ and $\Pi_W = X^* - \mathbb{L}[X^*|W]$. Then, write

$$Y = \alpha + \beta' X + \gamma' W + u,$$

where $u = U - \mathbb{E}[U|W]$ and $\mathbb{E}[U|W] = \alpha + \gamma' W$. Then, using $\mathbb{E}[\Pi_W] = \mathbb{E}[\Pi_W W] = \mathbb{E}[\Pi_W u] = 0$, we obtain

$$\begin{aligned}\beta^* &= \left(\mathbb{E}[\Pi_W \Pi_W']\right)^{-1} \mathbb{E}[\Pi_W X']\beta \\ &= \left(\mathbb{E}[\Pi_W \Pi_W']\right)^{-1} \mathbb{E}[\Pi_W \mathbb{L}[X'|Z,W,ZW]]\beta \\ &= \beta,\end{aligned}$$

This shows the robustness of the TSLS to misspecification of the first stages. ■

A.2 Nonparametric Estimators

A.2.1 Separable and Binary IV Case

We first introduce some notation that will be used throughout this Section. Henceforth, A', $rank(A)$, A^-, $Tr(A)$ and $|A| := (Tr(A'A))^{1/2}$ denote the transpose, rank, Moore-Penrose generalized inverse, trace and the Euclidean norm of a matrix A, respectively. For generic random vectors ζ and ξ, let F_ζ and $F_{\zeta/\xi}$ be the cumulative distribution function (cdf) and conditional cdf of ζ and ζ given ξ, respectively. Denote the corresponding densities with respect to a σ-finite measure μ by f_ζ and $f_{\zeta/\xi}$. Unless otherwise stated, the underlying measure will be the Lebesgue measure. Let $\mathcal{S}_\zeta$ denote the support of ζ. Let $L_2(\zeta)$ denote the Hilbert space with inner product $\langle h, g\rangle := \int f(x)g(x)dF_\zeta(x)$ and the corresponding norm $\|g\|_2^2 := \langle g, g\rangle$. Henceforth, sometimes we drop the domain of integration for simplicity of notation. For a linear operator $K : L_2(X) \to L_2(Y)$, denote the subspaces $\mathcal{R}(K) := \{f \in L_2(Y) : \exists s \in L_2(X), Ks = f\}$ and $\mathcal{N}(K) := \{f \in L_2(X) : Kf = 0\}$. Let $\mathcal{D}(K)$ denote the domain of definition of K. Let K^* denote the adjoint operator of K. We will use some basic results from operator theory and Hilbert spaces. See Carrasco, Florens and Renault (2006) for an excellent review of these results.

Equation (9) provides an integral equation of the first kind that can be used for estimating g. Related estimators have been proposed before in Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Darolles, Fan, Florens and Renault (2011), Horowitz (2011), Chen and Pouzo (2012) and Santos (2012), to mention just a few. Here, we follow closely Blundell, Chen and Kristensen (2007). Although, strictly speaking, our model is not given by a conditional moment restriction on a unique set of covariates, we can easily adapt the existing results to make them applicable in our setting. For simplicity, we focus here on the univariate W and X case.

There are many nonparametric methods that can be used to estimate m and A. Here we follow Blundell, Chen and Kristensen (2007) and use a sieve OLS estimator (SLS), see also Ai and Chen

(2003) and Newey and Powell (2003). Optimally weighted estimators can be obtained applying ideas in Blundell, Chen and Kristensen (2007). We assume we have a random (i.e. independent and identically distributed, in short iid) sample $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ of size $n \geq 1$, with the same distribution as the fourth-dimensional vector (Y, X, W, Z). We assume g is in a suitable space of smooth functions. Suppose $\mathcal{S}_X$ is a bounded interval of $\mathbb{R}$, with non-empty interior. For any smooth function $h : \mathcal{S}_X \subset \mathbb{R} \to \mathbb{R}$ and some $r > 0$, let $[r]$ be the largest integer smaller than r, and

$$\|h\|_{\infty,r} := \max_{j \leq \underline{\eta}} \sup_{x \in \mathcal{S}_X} \left|\nabla^j h(x)\right| + \sup_{x \neq x'} \frac{\left|\nabla^{[r]} h(x) - \nabla^{[r]} h(x')\right|}{|x - x'|^{r - \underline{[r]}}}.$$

Further, let $C_c^r(\mathcal{S}_X)$ be the set of all continuous functions h with $\|h\|_{\infty,r} \leq c$. Since the constant c is irrelevant for our results, we drop the dependence on c and denote $C^r(\mathcal{S}_X)$. We shall assume that $g \in C^r(\mathcal{S}_X)$ for some r and approximate $C^r(\mathcal{S}_X)$ with a sieve space $\mathcal{G}_n$ satisfying some conditions below. Define $k_n = \dim(\mathcal{G}_n)$. Given an integer $s > 0$ define the Sobolev norm $\|h\|_s^2 := \sum_{l=0}^s \left\|h^{(s)}\right\|_2^2$, where $h^{(s)}(x) := \partial^s h(x)/\partial x^s$, with $h^{(0)} \equiv h$.

We approximate $m_z(w) \equiv m(w, z) := \mathbb{E}[Y|W = w, Z = z]$ by the function $\tilde{m}(w, z) := \sum_{j \in \mathcal{J}_n} m_{zj} p_{0j}(w, z)$, where p_{0j} are some known basis functions and $J_n := \#(\mathcal{J}_n) \to \infty$ as $n \to \infty$. We write $\tilde{m}(w, z) = p^{J_n}(w, z)' m^{J_n}(z)$, where $p^{J_n}(w, z) = (p_{01}(w, z), ..., p_{0J_n}(w, z))'$ and $m^{J_n}(z) = (m_{z1}, ..., m_{zJ_n})$. Define $P := (p^{J_n}(w_1, z_1), ..., p^{J_n}(w_n, z_n))'$. Then, the SLS is

$$\widehat{m}(w, z) = p^{J_n}(w, z)'(P'P)^- \sum_{i=1}^n p^{J_n}(W_i, Z_i) Y_i.$$

More precisely, we take $p^{J_n}(w, z) = (B^{J_{2n}}(w), z \cdot B^{J_{2n}}(w))$, where $B^{J_{2n}}(w)$ is a $J_{2n} \cdot 1$ vector of univariate B-splines or polynomial splines and $J_n = 2J_{2n}$. We define $\widehat{m}(w) := \widehat{m}(w, 1) - \widehat{m}(w, 0)$.

Similarly, for a fixed g, we consider the sieve estimator of Ag as $\widehat{A}g = \widehat{A}_1 g - \widehat{A}_0 g$, where

$$\widehat{A}_z g = p^{J_n}(w, z)'(P'P)^- \sum_{i=1}^n p^{J_n}(W_i, Z_i) g(X_i).$$

Finally, the SLS for g is given by the solution of

$$\widehat{g}_n = \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^n \left(\widehat{m}(W_i) - \widehat{A}g(W_i)\right)^2.$$

We assume the sieve space $\mathcal{G}_n$ is of the form

$$\begin{aligned} \mathcal{G}_n = \{g_n : \mathcal{S}_X \to \mathbb{R},\ & \sup_x |g_n(x)| < c, \sup_x \left|\nabla^{[r]} g_n(x)\right| < c \\ g_n(x) = \psi^{k_n}(x)'\Pi,\ & g_n(\bar{x}) = 0\}, \end{aligned}$$

where $\psi^{k_n}(\cdot)$ is a $k_n \times 1$ vector of known basis that are at least $\gamma = ([r] + 1)$ times differentiable and Π is a $k_n \times 1$ vector of coefficients to be estimated. In the application we use B-splines for ψ^{k_n}. Blundell, Chen and Kristensen (2007) discussed practical ways to incorporate the constraints into the computation of $\widehat{g}_n$. For large samples the unconstrained estimator performs well. Note that $g_n(\bar{x}) = 0$ is a normalization restriction (location), where $\bar{x}$ is an arbitrary point in $\mathcal{S}_X$.

The following sieve measure of ill-posedness plays a crucial role in the asymptotic theory of sieve estimators, see Blundell, Chen and Kristensen (2007),

$$\tau_n := \sup_{g \in \mathcal{G}_n} \frac{\|g\|}{\left\|(A^*A)^{1/2} g\right\|}.$$

Consider the following assumptions, which are the same as in Blundell, Chen and Kristensen (2007), and are discussed extensively there.

Assumption 5 *Suppose that*

1. *The data $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ are iid and Assumption 2 holds.*

2. *(i) $g \in C^r(\mathcal{S}_X)$ for $r > 1/2$ and $g(\bar{x}) = 0$; (ii) $\mathbb{E}[|X|^{2a}] < \infty$ for some $a > r$.*

3. *For $z = 0, 1$, $m_z \in C^{r_m}(\mathcal{S}_W)$ with $r_m > 1/2$ and $\mathbb{E}[g_n(X)|W = \cdot, Z = z] \in C^{r_m}(\mathcal{S}_W)$ for any $g_n \in \mathcal{G}_n$.*

4. *(i) The smallest and the largest eigenvalues of $\mathbb{E}[B^{J_{2n}}(W) \cdot B^{J_{2n}}(W)']$ are bounded and bounded away from zero for each J_{2n}; (ii) $B^{J_{2n}}(W)$ is a B-spline basis of order $\gamma > r_m > 1/2$; (iii) the density of W is continuous, bounded, and bounded away from zero over its support $\mathcal{S}_W$, which is a compact interval with non-empty interior.*

5. *(i) $k_n \to \infty$, $J_{2n}/n \to 0$; (ii) $\lim_{n\to\infty} (J_{2n}/k_n) = c_0 > 1$ and $\lim_{n\to\infty} \left(k_n^2/n\right) = 0$.*

6. *There is $g_n \in \mathcal{G}_n$ such that $\tau_n^2 \|A(g - g_n)\|^2 \leq C \|g - g_n\|^2$.*

The following Theorem establishes rates for $\|\widehat{g}_n - g\|$. Its proof is the same as that of Theorem 2 in Blundell, Chen and Kristensen (2007), hence it is omitted.

Theorem A.1 *Let Assumption 5 hold. Then,*

$$\|\widehat{g}_n - g\| = O_P\left(k_n^{-r} + \tau_n \cdot \sqrt{\frac{k_n}{n}}\right).$$

A.2.2 Separable and General IV Case: Implementation as TSLS

It turns out that the nonparametric estimator discussed earlier can be applied to a general instrument, not necessarily binary, and more importantly can be implemented as a TSLS, similar to that used in the parametric setting. For simplicty of exposition we consider the univariate case for X and W (the multivariate case is analogous and only introduces more notation). The structural equation is now

$$Y = \alpha + \beta' \psi^{k_n}(X) + \gamma' B^{J_n}(W) + \varepsilon_n, \tag{20}$$

where $\psi^{k_n}(\cdot)$ and $B^{J_n}(w)$ are a $k_n \times 1$ and a $J_n \times 1$ vector, respectively, of known basis (e.g. univariate B-splines or polynomial splines) satisfying some conditions below. Consider the first-stages and reduced form as

$$\widehat{\mathbb{E}}[\psi^{k_n}(X)|W, Z] = \hat{\alpha}_{0x} + \hat{\alpha}_{1x} Z + \hat{\alpha}_{2x}' B^{J_n}(W) + \hat{\alpha}_{3x}' B^{J_n}(W) Z$$

and

$$\widehat{\mathbb{E}}[Y|W,Z] = \hat{\alpha}_{0y} + \hat{\alpha}_{1y}Z + \hat{\alpha}_{2y}'B^{J_n}(W) + \hat{\alpha}_{3y}'B^{J_n}(W)Z.$$

These OLS fits are used to nonparametrically estimate Δ_Y and the linear operator

$$Ag(W) = \frac{\mathbb{C}\left(g(X), Z|W\right)}{\sigma^2(W)},$$

by

$$\hat{\Delta}_Y = \hat{\alpha}_{1y} + \hat{\alpha}_{3y}'B^{J_n}(\cdot)$$

and, for $g(X) = \beta'\psi^{k_n}(X)$,

$$\widehat{A}g(\cdot) = \beta'\left(\hat{\alpha}_{1x} + \hat{\alpha}_{3x}'B^{J_n}(W)\right).$$

Then, the three-step nonparametric estimator is the solution of

$$\widehat{g}_n = \arg\min_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_Y(W_i) - \widehat{A}g(W_i)\right)^2,$$

where $\mathcal{G}_n$ is a sieve space of the form

$$\begin{aligned}\mathcal{G}_n = \{g_n : \mathcal{S}_X \to \mathbb{R},\ & \sup_x |g_n(x)| < c, \sup_x \left|\nabla^{[r]}g_n(x)\right| < c \\ g_n(x) = \beta'\psi^{k_n}(x),\ & g_n(\bar{x}) = 0\},\end{aligned}$$

and the vector $\psi^{k_n}(\cdot)$ is at least $\gamma = ([r]+1)$ times differentiable and β is a $k_n \times 1$ vector of coefficients to be estimated. The estimator $\widehat{g}_n$ can be also computed as a simple TSLS where the endogenous variables $\psi^{k_n}(X)$ in (20) are instrumented with Z and $B^{J_n}(W)Z$ and $B^{J_n}(W)$ are treated as exogenous variables. Here, the order condition $J_n \geq k_n - 1$ needs to hold.

This TSLS nonparametric estimator is much simpler to compute that the somewhat more natural two-step least squares estimator based on equation (17), i.e.

$$\hat{g} = \arg\min_{g\in\mathcal{G}_n} \mathbb{E}[\left(\hat{C}_Y - (\hat{C}g)(W)\right)^2], \tag{21}$$

where $\hat{C}_Y$ is a consistent estimator of $\mathbb{C}\left(Y, Z|W\right)$ and $(\hat{C}g)(W)$ is a consistent estimator of

$$\left(Cg\right)(W) = \mathbb{C}\left(g(X), Z|W\right).$$

Estimators for $\hat{C}_Y$ and $\hat{C}g$ in turn would require estimating the conditional mean $p(\cdot)$ and the conditional variance in a first step.

A.2.3 Nonseparable Case

We note that the nonparametric separable estimator can be extended to the nonseparable case following the same arguments above but replacing $\psi^{k_n}(X)$ by $\psi^{k_n}(V)$ and incorporating the normalizations in the new sieve space $\mathcal{G}_n$

$$\begin{aligned}\mathcal{G}_n = \{g_n : \mathcal{S}_V \to \mathbb{R},\ & \sup_v |g_n(v)| < c, \sup_v \left|\nabla^{[r]}g_n(v)\right| < c \\ g_n(v) = \beta'\psi^{k_n}(v),\ & g_n(y,\bar{x}) = 0 \text{ for all } y\}.\end{aligned}$$

That is, consider the first-stages and reduced form as

$$\widehat{\mathbb{E}}[\psi^{k_n}(V)|W,Z] = \hat{\alpha}_{0v} + \hat{\alpha}_{1v}Z + \hat{\alpha}_{2v}'B^{J_n}(W) + \hat{\alpha}_{3v}'B^{J_n}(W)Z$$

and

$$\widehat{\mathbb{E}}[Y|W,Z] = \hat{\alpha}_{0y} + \hat{\alpha}_{1y}Z + \hat{\alpha}_{2y}'B^{J_n}(W) + \hat{\alpha}_{3y}'B^{J_n}(W)Z.$$

Then, we estimate the linear operator

$$Ag(W) = \frac{\mathbb{C}\left(g(V),Z|W\right)}{\sigma^2(W)},$$

for $g(v) = \beta'\psi^{k_n}(v)$ by

$$\widehat{A}g(\cdot) = \beta'\left(\hat{\alpha}_{1v} + \hat{\alpha}_{3v}'B^{J_n}(W)\right),$$

and Δ_Y by

$$\hat{\Delta}_Y = \hat{\alpha}_{1y} + \hat{\alpha}_{3y}'B^{J_n}(\cdot)$$

Then, the three-step nonparametric estimator is the solution of

$$\widehat{g}_n = \arg\min_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_Y(W_i) - \widehat{A}g(W_i)\right)^2,$$

where $\mathcal{G}_n$ is the sieve space given above. The asymptotic theory for the nonseparable case follows from the same steps as those of the separable case with X replaced by V. This has the same impact as increasing the number of endogenous variables in Blundell, Chen and Kristensen (2007), which leads to qualitatively the same method of proof, except that the exponent r in the bias term is replaced by $r/(d+1)$, where d is the dimension of X. This is a straightforward extension of Blundell, Chen and Kristensen (2007) and hence we omitt details. The implementation in the nonseparable case is different from the separable case due to the different normalizations. It is a TSLS with linear restrictions on parameters. Specifically, the normalizations $\beta'\psi^{k_n}(y,\bar{x}) = 0$ for all y can be implemented as a simple quadratic constraint on a least squares problem in the same way as in p.1635 in Blundell, Chen and Kristensen (2007), by adding to their equation (21) the term

$$\mu\beta'\left(\frac{1}{n}\sum_{i=1}^{n}\psi^{k_n}(Y_i,\bar{x})\psi^{k_n\prime}(Y_i,\bar{x})\right)\beta,$$

where μ is the corresponding Lagrange multiplier for the normalizations. We refer to Blundell, Chen and Kristensen (2007) for details.

A.3 Identification with Conditional Independence

We show in this section how our approach can be modified to accommodate all restrictions imposed by the conditional independence restriction in (18) when the instrument is binary. Here it is convenient to use an equivalent normalization to $g(y,\bar{x}) = 0$, which is discussed in Matzkin (2003). We assume that U follows a $U[0,1]$ distribution. Then, conditional independence is equivalent to

$$\mathbb{C}\left(1(U \le u), Z|W\right) = 0 \text{ a.s. for all } u \in [0,1]. \tag{22}$$

Let U^* be an auxiliary random variable distributed as $U[0,1]$ and independent of (Y, X, Z, W, U). Then, by independence (22) is equivalent to

$$\mathbb{C}\left(1(U \leq U^*), Z|W, U^*\right) = 0 \text{ a.s.}$$

Note that by monotonicity $1(U \leq U^*) = 1\left(Y \leq m(X, U^*)\right)$ a.s. Let $\mathcal{M}$ be a class of measurable functions for m, and define the class

$$\mathcal{G} = \{Y - 1\left(Y \leq m(X, U^*)\right) : m \in \mathcal{M}\}.$$

Then, identification of m holds if (Y, X, U^*, Z) given (W, U^*) is $\mathcal{G}$-covariance complete. Thus, by creating an artificial sample from U^* we transform the infinite number of moment restrictions in (22) to a covariance restriction similar to that used for the nonseparable case (but with a common component that is exogenous).

References

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

Almond, D., and J. Currie (2011): "Killing Me Softly: The Fetal Origins Hypothesis," *The Journal of Economic Perspectives*, 25(3), 153–172.

Altonji, J.G. and R.L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053-1102.

Andrews, D. (2011): "Examples of L2-Complete and Boundedly-Complete Distributions," Cowles Foundation for Research in Economics.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?, " *The Quarterly Journal of Economics*, 106(4), 979-1014.

Angrist, J. D., and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88, 450-477.

Angrist, J., Graddy, K. and G. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499-527.

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

Card, D. (1995): "The Wage Curve: A Review," *Journal of Economic Literature*, vol. 33(2), 285-299.

Card, D. (2001): "Estimating the return to schooling: progress on some persistent econometric problems," *Econometrica*, 69, 1127-1160.

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.

Chay, K., and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113(2), 376–424.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in Handbook of Econometrics (J. J. Heckman and E. E. Leamer, eds.) volume 6, 5549–5632. Elsevier, Amsterdam.

Chen, X., and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80(1), 277–321.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245-261.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405-1441.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 1, 1-12.

D'Haultfoeuille, X. and P. Fevrier (2014): "Identification of Nonseparable Models with Endogeneity and Discrete Instruments," *Econometrica*, forthcoming.

D'Haultfoeuille, X., Hoderlein, S. and Y. Sasaki (2013): "Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments", Boston College Working Paper wp839.

DiNardo, J., and D. S. Lee. (2011): "Program Evaluation and Research Designs," In Handbook of Labor Economics, ed. O. Ashenfelter and D. Card, vol. 4A, 463-536. Elsevier Science B.V.

Dunker, F., Florens, J-P., Hohage, T., Johannes, J. and Mammen, E. (2014): "Iterative Estimation of Solutions to Noisy Nonlinear Operator Equations in Nonparametric Instrumental Regression", *Journal of Econometrics*, 178, 444-455.

Fan, J., and Gijbels, I. (1996): *Local Polynomial Modelling and its Applications.* Chapman & Hall, London.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effect," *Econometrica*, 76, 1191-1207

Frolich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Frolich, M. (2008): "Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables," *International Statistical Review*, 76, 214–227.

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

Hoderlein, S., Holzmann, H. and Meister, A. (2015): "The Triangular Model with Random Coefficients," unpublished manuscript.

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

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 61, 2, 467-476.

Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–35.

Imbens, G. W. and W.K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Withou Additivity," *Econometrica*, 77, 1481-1512.

Kasy M. (2009): "Semiparametrically Efficient Estimation of Conditional Instrumental Variable Parameters," *International Journal of Biostatistics*, 5 (1).

Lumley, J., C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley, and L. Watson (2009): "Interventions for Promoting Smoking Cessation During Pregnancy (Cochrane Review)," *The Cochrane Library*, 8(3).

Masten, M. and Torgovitsky, A. (2014): "Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model," CEMMAP working paper CWP02/14.

Matzkin, R.L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-13785.

Matzkin, R.L. (2013) "Nonparametric Identification in Structural Economic Models," *Annual Review of Economics*, Vol. 5.

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

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

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

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

Torgovitsky, A. (2014): "Identification of Nonseparable Models using Instruments with Small Support," *Econometrica*, forthcoming.