Identifying Multiple Marginal Effects with a Single Instrument[*]

Carolina Caetano[†]
University of Rochester

Juan Carlos Escanciano[‡]
Indiana University

January 29th, 2018.

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 model with an Instrumental Variable (IV) of lower dimension and multiple endogenous controls. The IV may even be a single binary variable, like a natural experiment or intention to treat. Despite the failure of the classical order condition, we show that identification may be achieved by exploiting heterogeneity of the "first stage" in the controls through a new rank condition that we term covariance completeness. This paper also provides parametric and nonparametric Two-Stage Least Squares (TSLS) estimators, which are very simple to implement, discusses their asymptotic properties, and shows that the estimators have excellent 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, Pedro Sant'Anna and Robin Sickles for helpful discussions. We would also like to thank seminar participants at many institutions as well as guests at our 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", dated May 25, 2015.

[†] 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 the nonparametric 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. Depending on the nature of X and functional form assumptions on g, a traditional IV approach requires, among other things, that we observe an instrument Z that 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. Thus, the traditional IV order condition imposes rather restrictive assumptions on the support of the instrument Z (relative to that of X), which often do not hold in applications. As a result, the marginal effects of a complex variable X are not identified, for example, in common cases where the instrument may be an experiment or a natural experiment.

To address this problem, a researcher may consider adding more variables to the list of instruments, say W, and run an IV regression with instruments (Z, W). However, if some component of W is correlated with U, then the IV regression with instruments (Z, W) would yield inconsistent estimates of the marginal effects of X on Y. In this paper we propose an alternative identification strategy for marginal effects that allows for arbitrary correlation between W and U.

The starting point of our approach is the conditional exogeneity condition of Z given W, i.e.

$$\mathbb{C}(U, Z|W) = 0 \text{ almost surely (a.s.)}, \tag{2}$$

where $\mathbb{C}(V_1, V_2|W)$ is the conditional covariance of V_1 and V_2 given W. Substituting U in (1) in the exogeneity condition yields the conditional covariance restriction

$$\mathbb{C}(Y, Z|W) = \mathbb{C}(g(X), Z|W) \text{ a.s.} \tag{3}$$

The main contribution of this paper is the use of covariance restrictions such as (3) for nonparametric identification under endogeneity. Using the conditional covariance removes the arbitrary correlation of W and U, embodied by the nuisance function $\mathbb{E}[U|W]$, and permits focus on the parameter of interest g. This paper also introduces an identifying assumption, called *covariance completeness*, which provides conditions under which the right hand side of (3) can be uniquely solved (up to a constant) in g. For covariance completeness to hold W must be distinct from X (separability). In contrast to the related completeness assumption of nonparametric IV, covariance completeness, which is extensively discussed below, imposes restrictions on the support of W relative to X (rather than Z relative to X). As such, these necessary conditions are more likely to hold in applications.

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 is*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

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 "completing a training program," or a relaxation of the linearity of a variable, for example X_1 is "education," and $X_2 = \mathbf{1}(X_1 \geq 12)$ captures the "sheepskin effect of graduating from high school." 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 use W, which may not be excluded from the structural equation. The exogeneity condition is that $\mathbb{C}(U, Z|W) = 0$. This condition allows, for example, the structural model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 W + \varepsilon, \tag{4}$$

where $\beta_3 \neq 0$ and where $\mathbb{C}(\varepsilon, Z|W) = 0$ a.s.

Then,

$$\mathbb{C}(Y, Z|W) = \beta_1 \mathbb{C}(X_1, Z|W) + \beta_2 \mathbb{C}(X_2, Z|W). \tag{5}$$

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 $\mathbb{C}(X_1, Z|W)$ and $\mathbb{C}(X_2, Z|W)$ are linearly independent. For example, if first stages are of the form

$$X_j = \alpha_{0j} + \alpha_{1j} Z + \alpha_{2j} W + \alpha_{3j} Z \cdot W + U_j, \qquad j = 1, 2, \tag{6}$$

with errors U_j satisfying $\mathbb{E}[U_j|Z, W] = 0$ a.s., then covariance completeness holds if

$$rank \begin{bmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{12} & \alpha_{32} \end{bmatrix} = 2$$

and $\mathbb{V}(W) > 0$ (where $\mathbb{V}(W)$ denotes variance of W). Our identification strategy exploits the heterogeneity in the "first stages" to separate the marginal effects of X_1 and X_2.

The identification strategy has an intuitive interpretation. It is based on the observation that if the population is categorized according to W, the discrete variation on the distribution of X induced by Z may vary with W. This may allow us to recover a rich (e.g continuous) set of marginal effects even when Z is a binary variable, say $Z \in \{0, 1\}$, and X takes 3 or more values, and may even be continuous, or a vector. Thus, 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

$$Y = m(X, U), \tag{7}$$

where $m(x, u)$ is a strictly monotone function in the scalar u, for each x in the support of the distribution of X, $\mathcal{S}_X$ say. Model (7) 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). See also Example 4.1 for an economic example that satisfies the restrictions of the non-separable setting.

Based on our identification results, we propose simple parametric and nonparametric estimators of g up to a constant. Our estimation approach can be immediately implemented using standard 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 W as an exogenous control and uses interaction terms between W and Z as instruments. So, for example, in the bivariate linear model of Example 1.1 above, our estimator is a simple TSLS with first stages (6) and a structural equation given by (4). The implementation as TSLS also extends to the nonparametric model. Thus, this paper provides a new methodology for identification and estimation of multiple marginal effects, which is applicable in practical situations where instruments are discrete and which makes an effective use of potentially endogenous controls.

The rest of the paper is organized as follows. First, we provide a review of the literature in the next section. We then introduce the ideas on the additive model, which is done in Section 3. There, we present the identification results (Section 3.1), develop the concept of covariance completeness (Section 3.1.1), show how to include covariates and propose robustness checks of separability (Section 3.1.2). We also show that the identification strategy can be implemented with a suitable TSLS estimator (Section 3.2). Section 4 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 5 reports the results of Monte Carlo experiments. Section 6 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 7. Mathematical proofs of the main results, including rates of convergence for nonparametric models, are gathered in the Appendix.

2 Literature Review

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). Our asymptotic results for nonparametric models adapt to our setting previous results by Blundell, Chen and Kristensen (2007). In the more general nonseparable case, our paper contributes to the literature of nonparametric identification of heterogeneous (observable and unobservable) marginal effects; see Matzkin (2013) for a 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 (2015), Torgovitsky (2015), 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,

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

Meghir and Vytlacil (2008), among others. In independent and contemporaneous work, Huang, Khalil, and Yildiz (2015) propose an identification strategy for situations with insufficient instruments. Their identification strategy is based on a control function approach and does not use covariates or covariance restrictions.

The main contribution of our paper relative to the aforementioned literature 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, which are the main focus of this paper.

Covariance completeness is useful beyond the setting considered in this paper. In subsequent work to ours, Ben-Moshe, D'Haultfoeuille, and Lewbel (2016) and Kim and Song (2017) have used covariance completeness conditions to identify nonparametric models with mismeasured regressors and without instruments. The problem investigated by these authors is different from the one analyzed here. Nevertheless, these works illustrate the utility of covariance completeness in other applications of economic interest.

More broadly, our approach relates to traditional empirical methods in econometrics that treat controls as exogenous variables. If the controls turn out to be endogenous (see e.g. experience in wage equations), then estimates of marginal effects of interest may be inconsistent; see, e.g., Frolich (2008). We provide explicit conditions under which the consistency of the marginal effects of interest is not affected by the endogeneity of the controls both for our methods and also for traditional IV methods. Our results relax the strong linearity assumptions of traditional IV with controls in traditional econometrics, see e.g., Chapter 12 of Stock and Watson (2011), but more importantly, suggests additional "instruments" that can be used to aid identification, even when these instruments include endogenous controls. Given the ubiquitous presence of endogenous controls in applications, this robustness is of 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, very large number of controls and fixed effects all without the need to modify either the method or the estimator.

3 The Separable Case with Binary IV

3.1 Identification

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

$$Y = g(X) + U, \tag{8}$$

where Z is binary, with support $\mathcal{S}_Z = \{0, 1\}$, and the following exclusion restriction holds

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

Assumption 1 is equivalent to (2) when Z is binary and $\mathbb{E}[Z|W] > 0$ a.s. One interpretation of W is as invalid instruments, since they are potentially correlated with U, i.e. $\mathbb{E}[U|W] \neq 0$, but more broadly we can think of W as controls that are structurally separable from X. Without such controls the model is in general not identified, as discussed above. Thus, existence of such controls opens up the possibility of identification, as we now show.

The following assumptions are sufficient for conditional moments to be well-defined.

Assumption 2 *$\mathbb{E}[|Y|] < \infty$ and $p(w) = \mathbb{E}[Z|W = w]$ satisfies $0 < p < 1$ a.s.*

This assumption allows us to take conditional means in (8) and subtract terms, so we can thus write

$$\mathbb{E}[Y|W, Z = 1] - \mathbb{E}[Y|W, Z = 0] = \mathbb{E}[g(X)|W, Z = 1] - \mathbb{E}[g(X)|W, Z = 0] \text{ a.s.} \tag{9}$$

Identifying g (up to location) from this equation depends on our ability to invert it. To better understand the conditions that guarantee the invertibility of equation (9) consider first the following empirical example which clarifies how some of the requirements and ideas translate to an actual applied problem.

Example 3.1 *(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 into two groups, indexed by Z, and let the classification control W be the mother's years of education. Table 1 shows an overview of the situation.*

Table 1: **Identification Idea**

(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)
$\mathbf{W}$	$\bar{\mathbf{X}}_{\mathbf{0},\mathbf{W}}$	$\bar{\mathbf{X}}_{\mathbf{1},\mathbf{W}}$	$\boldsymbol{\Delta}_{\mathbf{Y}}(\mathbf{W})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{0})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{1})$	$\mathbb{P}_{\mathbf{0},\mathbf{W}}(\mathbf{3})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{0})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{1})$	$\mathbb{P}_{\mathbf{1},\mathbf{W}}(\mathbf{3})$	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)

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. In this example all groups reduced one cigarette on average because of the intervention. This is not a requirement of our method, we just want to show that identification 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.

If Assumption 1 holds, from Equation (9) we derive the following system of equations

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

$$22 = 0.2g(0) + 0.2g(1) - 0.4g(3) \tag{11}$$

$$30 = 0.33g(0) - 0.33g(3). \tag{12}$$

Note that only two equations are linearly independent (since $0.4(10)+0.6(12)=(11)$). 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 linearly 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 effect. It is straightforward to see in this example that, from equation (10), $g(3) - g(1) = -20$, from equation (12), $g(3) - g(0) = -90$, 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 3.1 extends to the general discrete case as follows.

Example 3.2 *(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 (9), which in this context can be written as a linear system $\mathbf{\Delta}_Y = \mathbf{Ag}$, where, $\mathbf{\Delta}_Y := (\Delta_Y(w_1), \dots, \Delta_Y(w_l))'$, ($a'$ denotes the transpose of a) with $\Delta_Y(w) := \mathbb{E}[Y|Z = 1, W = w] - \mathbb{E}[Y|Z = 0, W = w]$, $w \in \mathcal{S}_W$, 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 (9). 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. The identification condition rank($\mathbf{A}$) $= q - 1$ is the key identification assumption in this paper and formalizes the idea of identification by exploiting heterogeneity of first-stages in covariates, as the elements of $\mathbf{A}$ are these effects, i.e. $\mathbb{P}[X = x_j|Z = 1, W = w_i] - \mathbb{P}[X = x_j|Z = 0, W = w_i]$.*

The discussion and intuition of Example 3.2 further extends to the general continuous case as follows. With some abuse of notation, we write equation (9) also as

$$\Delta_Y = Ag, \tag{13}$$

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

Notice that the identification condition in Example 3.2 that $rank(A) = q-1$ is equivalent to Assumption 3 in the discrete support case, since $\dim(\mathcal{N}(A)) + rank(A) = q$ and $\mathcal{N}(A)$ of Assumption 3 has dimension one. The proof of the following identification result can be found in the Appendix A.1.

Theorem 3.1 *Under Assumptions 1-3, g is identified up to location.*

In the general case, Assumption 3 is the analogue in our setting of the L_2-completeness identification condition of nonparametric IV, which is assumed throughout that literature; see e.g. 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 (9) as

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

This way of writing equation (9) inspires the introduction of a rank condition equivalent to Assumption 3, 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 3.4 to 3.6, which translate the meaning of covariance completeness to some important special cases.

3.1.1 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 $g_0 \in \mathcal{G}$, and whose elements satisfy the location normalization restriction $g(\bar{x}) = 0$, for a fixed $\bar{x} \in \mathcal{S}_X$.

Definition 3.1 *We say (X, Z) given W is $\mathcal{G}$-covariance complete if for each $g \in \mathcal{G} - \{g_0\}$,*

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

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 result is proved in Appendix A.1.

Theorem 3.2 *Let Assumptions 1-2 hold. Then, g is point-identified in $\mathcal{G}$ iff (X, Z) given W is $\mathcal{G}$-covariance complete.*

Remark 3.1 *Theorem 3.2 is also valid for non binary IV provided Assumption 1 is replaced by the generally weaker exogeneity condition (2) and Assumption 2 by the condition that all conditional covariances are well-defined.*

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 3.2 *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.

It is important to stress that the nonparametric IV literature has used completeness of X given Z to identify marginal effects, which imposes conditions on the support of Z relative to that of X which may not hold in applications. The following result provides a sufficient condition for covariance completeness in terms of a weighted completeness for clarification. However, note that the completeness in this result is of X given W (rather than X given Z). Thus, our covariance completeness is fundamentally different from completeness as used in the nonparametric IV literature, as they involve different conditioning variables. To the extent that covariates have typically richer support than instruments, the support conditions required by our covariance completeness is more plausible than that of the standard nonparametric IV approach.

Define $q(x, w) = \mathbb{E}[Z|X = x, W = w]$ and $k(x, w) = q(x, w) - p(w)$, where $p(\cdot)$ is defined in Assumption 2. Define the class of measurable functions

$$\mathcal{F} = \{f(x, w) = g(x)k(x, w) : g \in \mathcal{G}\}. \tag{14}$$

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

Proposition 3.3 *(X, Z) given W is $\mathcal{G}$-covariance complete iff the distribution of (X, W) given W is $\mathcal{F}-$complete, where $\mathcal{F}$ is defined in (14).*

From Proposition 3.3 it follows that a necessary nonparametric relevance condition for covariance completeness is that $k(X, W)$ is non-zero with positive probability, i.e.

$$P(\Omega) > 0, \tag{15}$$

where $\Omega = \{(X, W) : \mathbb{E}[Z|X, W] \neq \mathbb{E}[Z|W]\}$. That is, covariance completeness can be understood as a weighted completeness between the endogenous variables X and the controls W, and the necessary condition (15) requires that the endogenous variables and the instrument are not independent conditional on W, so that the weights $k(x, w)$ are nonzero. This is a very intuitive necessary relevance condition: after controlling for W, X needs to be conditionally dependent of Z.

Remark 3.2 *Covariance completeness imposes restrictions on the support of W relative to that of the endogenous variables X, while L_2-completeness of traditional IV imposes restrictions on the support of the instrument Z relative to that of endogenous variables X. 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 3.3 *(X and W discrete, Example 3.2 cont.) Under Assumption 2, the system $\mathbf{\Delta}_Y = \mathbf{Ag}$, is equivalent to the covariance restrictions $\mathbf{c} = \mathbf{Cg}$, where $\mathbf{c} = \mathbf{D\Delta_Y}$, $\mathbf{C} = \mathbf{DA}$, 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 simply to rank$(\mathbf{C}) = q - 1$ or equivalently, the previous identification condition rank$(\mathbf{A}) = q - 1$. This is the precise sense of heterogenous first-stages needed in our method. Again, covariance completeness requires the order condition $l \geq q - 1$, so it restricts the support of the controls W relative to that of endogenous variables X. In contrast, traditional IV requires that the support of the instrument is larger than q for identification.*

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

Example 3.4 *(Gaussian variables) Suppose that (X, W) is jointly normal conditional 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, covariance completeness in this context translates into the intuitive condition $\rho_1 \neq \rho_0$, i.e. variation in the instrument changes the correlation of X and W.

Example 3.5 *(High-dimensional Regressions) The intuition from Example 1.1 extends to the high-dimensional case. Assume $g(X) = \alpha_g + \beta_g' \mathbf{X}$, where $\mathbf{X}$ is possibly a large dimensional vector of transformations of X. Then, taking as the location normalization $\alpha_g = 0$, covariance completeness holds if*

$$\mathbb{C}(Z, g(X)|W) = \mathbb{C}(Z, \mathbf{X}|W)\beta_g = 0 \Longrightarrow \beta_g = 0. \tag{16}$$

Thus, covariance completeness requires that first-stages are heterogenous in covariates. Specifically, in this example is needed that $\mathbb{C}(Z, \mathbf{X}|W)$ is not perfectly multicolinear, so that (16) holds. Of course, a neccesary condition for covariance completeness is that $\mathbb{C}(Z, \mathbf{X}|W)$ varies with W. This condition can be tested. For example, suppose $Z = \gamma + \delta'\mathbf{X} + V$, where $\mathbb{E}[V|W, X] = \mathbb{E}[V|W]$ a.s. Then, $\mathbb{C}(Z, \mathbf{X}|W) = \delta' Var(\mathbf{X}|W)$ and necessary conditions are that $\delta \neq 0$ and conditional heteroskedasticity, so that $Var(\mathbf{X}|W)$ is not constant. Both are intuitive and can be tested by traditional methods. The condition $\delta \neq 0$ is the relevance condition mentioned earlier ($\mathbb{E}[Z|X, W] \neq \mathbb{E}[Z|W]$). There is no need to check these conditions separately. Indeed, we recommend checking variability of $\mathbb{C}(Z, \mathbf{X}|W)$ in W by testing significance of interaction coefficients in first-stage regressions.

Example 3.6 *(Linear multivariate model) Suppose $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$, so that the model is*

$$Y = \beta' X + U, \quad \mathbb{E}[U|W, Z] = \mathbb{E}[U|W], \tag{17}$$

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 $\Delta_\xi = \mathbb{E}[\xi|W, Z=1] - \mathbb{E}[\xi|W, Z=0]$)

$$\Delta_Y = \beta' \Delta_X. \tag{18}$$

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 (18) $\beta = \left(\mathbb{E}[\Delta_X \Delta_X']\right)^{-1} \mathbb{E}[\Delta_X \Delta_Y]$. 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).

3.1.2 Relaxing and Testing Structural Separability

A necessary condition for covariance completeness is that W must be additively separable from X. It is important to notice that one can have non-separable covariates in the model. Suppose that Q is a vector of covariates, then the extended model

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

where $\mathbb{E}[U|Z, Q, W] = \mathbb{E}[U|W]$ can be treated identically to our procedure, but treating Q identically to Z. Here there are no rank conditions on Q, and the rank conditions on W remain unchanged. Note that if we assume a semiparametric structure for g, for example if $g(X, Q) = \beta(Q)'X$, with $\beta(\cdot)$ fully nonparametric, then we can even drop the exogeneity requirement for Q, as we show in the following example.

Example 3.7 *(Linear model with nonparametric heterogeneous effects in Q) Consider the varying coefficient model*

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

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

$$\mathbb{E}[\Delta_X(q, W)\Delta'_X(q, W)] \text{ is positive definite for a.s. } q,$$

where recall Δ_X is defined in (18). Under this covariance completeness assumption, we can estimate $\beta(\cdot)$ nonparametrically from local least squares regressions. Here, identification requires that Q is different from W, and W must assume at least $d-1$ different values.

Example 3.8 *(Linear model with parametric heterogeneous effects in Q) Alternatively, we could specify $\beta(Q)$ as a linear function of Q, say $\beta(Q) = \beta + \beta_1'Q$, which results in a linear-in-parameters model with endogenous variables X and endogenous interactions between X and Q, which can be dealt with as in Example 3.6 above. In this setting we could even have $Q = W$. Consider, for example, the case where X is univariate. The model with interactions in the structural equation is*

$$Y = \beta X + \beta_1' W X + U, \quad \mathbb{E}[U|Z, W] = \mathbb{E}[U|W],$$

and our covariance completeness condition becomes

$$\mathbb{E}\left[\begin{pmatrix} \Delta_X(W) \\ W\Delta_X(W) \end{pmatrix}\begin{pmatrix} \Delta_X(W) & W\Delta_X(W) \end{pmatrix}\right] \text{ is positive definite.}$$

This model is just identified under the rank condition.

In the previous example, note that since the parameter of the interaction term β_1 is identified, our method allows us to test for separability in the marginal effects, i.e. $H_0 : \beta_1 = 0$, with a simple Wald test. We implement tests for separability following this approach in our application Section 6. Since structural separability is an important component of our approach we recommend performing tests of separability when applying our procedure. If some of the coefficients in β_1 are zero the model becomes over-identified under the rank condition, which opens up the possibility of testing for the over-identifying restrictions and moreover, makes estimation more robust. Example 3.9 in the next section illustrates these points in the context of our estimator.

3.2 Estimation

Write the nonparametric model as

$$Y = g(X) + \mathbb{E}[U|W] + U - \mathbb{E}[U|W].$$

To propose an estimation method that is practical, we consider in this section the approximations $g(X) \approx \alpha + \beta'\mathbf{X}$, $\mathbb{E}[U|W] \approx \gamma'\mathbf{W}$, where $\mathbf{X}$ and $\mathbf{W}$ are possibly large dimensional vectors of transformations of X and W, respectively, and define $\varepsilon = U - \mathbb{E}[U|W]$. One can think of these linear approximations as sieve-approximations of nonparametric objects. This argument leads to a structural equation that is linear in parameters, i.e.,

$$Y = \alpha + \beta'\mathbf{X} + \gamma'\mathbf{W} + \varepsilon, \tag{19}$$

and where $\mathbf{X}$ and $\mathbf{W}$ are vectors of dimensions d and d_w, respectively. In this section we consider the case where d and d_w are finite, but Appendixes A.2.1 and A.2.2 consider the nonparametric case where these dimensions grow to infinite with the sample size, thus relaxing functional form assumptions.

With the reparametrization, $\mathbb{E}[\varepsilon|W, Z] = 0$, so any transformation of W and Z different from $\mathbf{W}$ could be used as an "instrument". This is the new insight of this paper, and should be compared with textbooks treatments of the topic which suggest just using Z as instruments and W as controls, see e.g. Appendix of Chapter 12 in Stock and Watson (2011). To incorporate heterogeneity in first-stages we suggest

$$\mathbf{X} = \alpha_{0X} + \alpha_{1X}Z + \alpha_{2X}\mathbf{W} + \alpha_{3X}\mathbf{W}Z + U_1,$$

where $\mathbb{E}[U_1|Z, W] = 0$ a.s. and Z is a single instrument. Of course, other transformations could be used, but since $\mathbf{W}$ is high dimensional, we believe this to be a sufficiently rich specification in practice.

With these specifications turns out that our identification strategy can be easily implemented as a TSLS estimation. Specifically, given an independent and identically distributed (iid) 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 $\mathbf{X}_i$ on a TSLS regression of Y_i onto $\mathbf{X}_i$ and $\mathbf{W}_i$, using Z_i and $Z_i\mathbf{W}_i$ as instruments for $\mathbf{X}_i$, and treating $\mathbf{W}_i$ as if they were exogenous controls. 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 how this estimation strategy relates to our identification strategy, let the fitted value of the "first-stage" be $\widehat{\mathbb{E}}[\mathbf{X}|W, Z] = \widehat{\alpha}_{0X} + \widehat{\alpha}_{1X}Z + \widehat{\alpha}_{2X}\mathbf{W} + \widehat{\alpha}_{3X}\mathbf{W}Z$ and the fitted value of the "reduced-form" be $\widehat{\mathbb{E}}[Y|W, Z] = \widehat{\alpha}_{0Y} + \widehat{\alpha}_{1Y}Z + \widehat{\alpha}_{2Y}\mathbf{W} + \widehat{\alpha}_{3Y}\mathbf{W}Z$. Then, it is well known that the TSLS estimator is related to the reduced form fits through the equation

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

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

$$\hat{\Delta}_Y = \widehat{\beta}'\hat{\Delta}_{\mathbf{X}},$$

where $\hat{\Delta}_Y$ and $\hat{\Delta}_{\mathbf{X}}$ are the empirical analogues of Δ_Y and $\Delta_{\mathbf{X}}$ in Example (3.6), respectively. Thus, indeed the TSLS estimator $\widehat{\beta}$ satisfies the sample analogue of the main equation we use for identification.

Note that $\hat{\Delta}_{\mathbf{X}} = \widehat{\alpha}_{1\mathbf{X}} + \widehat{\alpha}'_{3\mathbf{X}}\mathbf{W}$. Thus, adding the interaction term is critical for our completeness condition to hold in this example when $\mathbf{X}$ has dimension larger than one (i.e. when the classical IV order condition fails). A necessary and sufficient condition for identification is $\mathcal{G}$-covariance completeness of (X, Z) given W, for $\mathcal{G} = \{b'\mathbf{X} : b \in \mathbb{R}^d\}$, which in the specification of this section reduces to

$$\mathbb{V}(W) \text{ is positive definite and } rank\,(\mathbf{M}) = d, \tag{20}$$

where $\mathbf{M} = (\alpha_{1\mathbf{X}}, \alpha_{3\mathbf{X}})$ is a $d \times (d_w + 1)$ matrix of coefficients.

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 3.4 *Let Assumptions 1-2, (19) and (20) hold. Then,*

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

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

Remark 3.3 *Note that the TSLS can be applied without any modification when Z is not binary.*

Remark 3.4 *Our TSLS will be consistent for β even when $\mathbb{E}[\mathbf{X}|Z, W]$ is non-linear, as long as $g(X)$ is a linear function of $\mathbf{X}$ and the conditions above hold. This is formally established in the end of the proof of Theorem 3.4 (page 27), and we also confirm this in the simulations (see the second set of simulations, page 20).*

Remark 3.5 *The order condition here is that the dimension of $\mathbf{W}$, d_w, needs to be at least $d-1$ (the dimension of $\mathbf{X}$ minus one). Therefore, if there are several variables which satisfy the identification conditions, we recommend that $\mathbf{W}$ be chosen as the variable (or variables) for which Z and $Z\mathbf{W}$ make up the strongest instruments. Additionally, comparisons of results using different $\mathbf{W}$ are an informal test of the functional form assumption.*

The following example illustrates the estimation and provides practical recommendations for applying our results.

Example 3.9 *(Bivariate linear case, cont.) Consider the linear model of Example 1.1 in the Introduction, with two endogenous variables $X = (X_1, X_2)$, one control W and one binary IV, Z, our estimator is the TSLS with the structural model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 W + \varepsilon,$$

and the first stages

$$X_j = \alpha_{0j} + \alpha_{1j} Z + \alpha_{2j} W + \alpha_{3j} Z \cdot W + U_j, \qquad j = 1, 2.$$

Although the model is under-identified with classical IV, it is over-identified with our approach, and as result, an applied researcher can use our method to test for over-identifying restrictions by standard methods in TSLS, after estimating the model. Furthermore, if the researcher is concerned with the structural separability of W and the endogenous variables, as required in our method, we recommend, as robustness checks, running our estimator on the structural model with interactions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 W + \beta_4 W X_1 + \varepsilon,$$

and the same first stages as above. In this extended model the parameters are just-identified with our method, and a standard t-test for $\beta_4 = 0$ based on our TSLS provides empirical evidence of the separability (or lack thereof) used in the first specification. We can apply a similar argument with interactions with X_2. These recommendations, together with standard recommendations regarding weak instruments, provide specific guidance on how to apply and interpret our methods.

4 Nonseparable Case

This section extends our previous identification strategy to the nonseparable model (7), repeated here for convenience:

$$Y = m(X, U), \tag{21}$$

where $m(x, u)$ is a strictly monotone function in a scalar u, for each x in $\mathcal{S}_X$. The following example motivates our structural separability and monotonicity assumptions in an economic setting.

Example 4.1 *Consider the following example of production with an intermediate unobserved input. An economic agent produces some final outcome Y, such as firm revenue or individual lifetime earnings, following the steps:*

- *Step 1: W and ε are used to produce $U = h(W, \varepsilon)$.*
- *Step 2: Given W and ε_x , the agent chooses X*

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

  where $C(x, w, z)$ is a cost function and Z a cost shifter.
- *Step 3: The final outcome is given by $Y = m(X, h(W, \varepsilon))$.*

In this setting 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 input $U = h(W, \varepsilon)$. The variables in 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). Also, it is assumed that production is increasing in U (e.g. ability). 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. The fact that prior inputs W do not enter final production in arbitrary ways, because of the technological constraints embodied in h, justifies our structural separability assumption between X and W, 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.*

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 the exogeneity condition is

$$\mathbb{C}\left(m^{-1}(Y, X), Z|W\right) = 0 \; a.s. \tag{22}$$

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 (22) can be written as

$$\mathbb{C}\left(Y, Z|W\right) = \mathbb{C}\left(g(Y, X), Z|W\right). \tag{23}$$

Then, identifying g from this equation is equivalent to identifying m in (22). In the nonseparable case, however, under the standard conditional independence assumption considered in the literature

$$Z \perp U|W, \tag{24}$$

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 $g(y, \bar{x}) = 0$ for all $y \in \mathcal{S}_Y$.

Let $\mathcal{G}$ be a subset of $L_2(Y, X)$, whose elements g satisfy the normalization restrictions $g(y, \bar{x}) = 0$, for all $y \in \mathcal{S}_Y$. Note that the normalization rules out the trivial solution $g(y, x) = y$ of (23).

Definition 4.1 *We say (Y, X, Z) given W is $\mathcal{G}$-covariance complete if for each $g \in \mathcal{G} - \{g_0\}$,*

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

The proof of the next theorem is the same as that of Theorem 3.2 and therefore is omitted.

Theorem 4.1 *Suppose (21), (23) and Assumption 2 hold. Then, g is point-identified in $\mathcal{G}$ if (Y, X, Z) given W is $\mathcal{G}$-covariance complete.*

Remark 4.1 *Theorem 4.1 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 (24). 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 4.1.*

In the following example we extend Example 3.2 to the nonseparable case. For simplicity, we consider the binary IV case, as other cases can be treated similarly. Example 4.3 illustrates the theory in Example 4.2 in the maternal smoking case.

Example 4.2 *(Discrete data) 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 generality, we assume 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 $l \times q$ 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. Covariance completeness is equivalent to full column rank of $\mathbf{B}$,

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

Intuitively speaking, in the non-separable case covariance completeness means enough heterogeneity in both reduced forms (for X and for Y), and not just on X as in the separable case (see the definition of $\mathbf{B}_1$, which involves $V = (Y, X)$). The order condition is $l \geq q - q_y$.

Example 4.3 *(Effects of maternal smoking on birth weight) Suppose that we have information on whether the woman smoke during pregnancy or not ($X = 1$ if the woman smoked, $X = 0$ otherwise), and whether the baby had normal birth weight, which is defined as birthweight over 2,500 grams, or not: ($Y = 1$ if the child is of normal weight, $Y = 0$ otherwise). Suppose that women are randomly divided into two groups, indexed by Z, and let the classification control W be the mother's years of*

Table 2: **Probabilities in the nonseparable case**

	$\mathbf{Z=0}$				$\mathbf{Z=1}$				
(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)
W	$(\mathbf{0},\mathbf{0})$	$(\mathbf{0},\mathbf{1})$	$(\mathbf{1},\mathbf{0})$	$(\mathbf{1},\mathbf{1})$	$(\mathbf{0},\mathbf{0})$	$(\mathbf{0},\mathbf{1})$	$(\mathbf{1},\mathbf{0})$	$(\mathbf{1},\mathbf{1})$	row #
6	3/10	1/5	0	1/2	1/10	1/5	1/10	3/5	(1)
10	1/5	1/5	1/10	1/2	0	1/5	1/5	3/5	(2)
17	1/5	1/10	1/10	3/5	3/10	0	0	7/10	(3)

education, as in Example 3.1. Table 2 shows an overview of the probabilities. Columns (II) to (V) show the probabilities of the pairs (Y, X) in the control group. For example, the number assigned to column (V) row (1) is the probability of a person in the control group having had a baby with normal birth weight and being a smoker, which is 1/2 in this fictitious example. Columns (VI) to (IX) do the same in the treatment group. As we can see, Z affects each of the education groups differently.

Applying equation (23) to the first education group ($W = 6$) we arrive at the equation

$$1/5 = -1/5 \cdot g(0,0) + 0 \cdot g(0,1) + 1/10 \cdot g(1,0) + 1/10 \cdot g(1,1).$$

Analogously for the other education levels, the resulting system of equations is

$$\begin{aligned}
1/5 &= -1/5 \cdot g(0,0) + 0 \cdot g(0,1) + 1/10 \cdot g(1,0) + 1/10 \cdot g(1,1) && (25)\\
1/5 &= -1/5 \cdot g(0,0) + 0 \cdot g(0,1) + 1/10 \cdot g(1,0) + 1/10 \cdot g(1,1) && (26)\\
0 &= 1/10 \cdot g(0,0) - 1/10 \cdot g(0,1) - 1/10 \cdot g(1,0) + 1/10 \cdot g(1,1)\\
0 &= g(0,0)\\
0 &= g(1,0)
\end{aligned}$$

Noting that (25) and (26) are identical, and using the normalization restrictions, the solution to this system is

$$g(1,1) = 2 = g(0,1).$$

In terms of m^{-1}, the solution is given by $m^{-1}(0,0) = 0$, $m^{-1}(0,1) = -2$, $m^{-1}(1,0) = 1$, and $m^{-1}(1,1) = -1$.

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.

4.1 A nonseparable model without monotonicity

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

Example 4.4 *(Random coefficients). Consider the random coefficient model*

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X, \\
X &= \alpha_0 + \alpha_1 Z,
\end{aligned}
$$

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

$$\theta \perp Z | W.$$

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

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

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 $\beta = \mathbb{E}[\beta_1]$.

5 Monte Carlo Simulations

This Section investigates the finite sample performance of our TSLS proposed in Section 3.2. In particular, we aim to investigate the sensitivity of our TSLS to the endogeneity of the controls 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 3 holds, which in this case is equivalent to $\mathbb{E}[\Delta_X \Delta_X']$ being positive definite (see Example 3.6). 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 3 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 3 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 3: **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:

$$\begin{aligned} Y &= \alpha + \beta_1 D + \beta_2 D1(D > 0) + \gamma_W W + u, \\ W &= \alpha_W u + U_W, \\ D &= \alpha_d W + \gamma_d Z \cdot W + U_d, \end{aligned}$$

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 3.4 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 4 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 4: **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 controls and the nonlinearity of the first stages.

6 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.

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 and Z is the county's attainment status in 1975 according to the Clean Air Act. Since Chay and Greenstone use a rich set of controls (see p. 411 in that paper), we separate them into two vectors Q and W as explained in Section 3.1.2.

Our first concern when choosing W is the implications it has for the model in terms of the specific functional shapes as well as structural separability. To begin, we follow the guidance set out in Example 4.1. The county's changes between 1970 and 1980 in the percent spending in highways, health and education don't affect the valuation for air quality directly, though they can be influential through affecting U (for example when they cause a change in the composition of the county's population). Thus, the three variables above are prime candidates to serve as W (see considerations about the choice of W and robustness checks below). We began by considering 4 models with the structure $Y = \beta X + Q'\pi + U$, with $\mathbb{E}[U|Z, Q, W] = \mathbb{E}[U|W]$. The first three use a scalar W (each of the 3 variables above), and in the fourth W is a vector of all three variables at the same time (see Table 5). These models have a similar structure to that of Chay and Greenstone (2005), which allows us to do direct comparisons.

The estimation is done using the same data set as in Chay and Greenstone (2005), and Q is the exact same control list used in that paper. The first column in Table 4 shows the results of a standard IV estimation of β, 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 control W. Row (i) uses a specification without exogenous control variables ($\pi = 0$). Although in specification (i) our results are of similar magnitude to the standard IV, they are slightly smaller and vary depending on the chosen W. They are particularly smaller when W includes all three variables in column D. We believe that this happens because when controls Q are not used, the classic IV operates under the assumption that $\mathbb{E}[U|Z]$ is

Table 5: **Estimation Results - Linear Case, Binary IV**

	IV	A	B	C	D
(i)	-.347	-.340	-.327	-.317	-.327
β, no controls	(.140)	(.138)	(.136)	(.134)	(.135)
(ii)	-.203	-.208	-.202	-.203	-.208
β, with controls	(.093)	(.094)	(.093)	(.093)	(.093)
(iii)		-.226	-.220	-.194	
β in non-separable model		(.096)	(.098)	(.110)	
(iv)		-1.432	.854	.174	
β_1 (separability test)		(1.873)	(1.227)	(1.174)	

Table 5: Columns A to D use our approach with W 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 controls Q. Specification (ii) uses as exogenous controls Q the exact same specification as in Chay and Greenstone (2005) excluding the controls which we are using as W. Specification (iii) assumes a nonseparable model as described in Example 3.8. Row (iv) shows the results of the separability test described in Section 3.1.2.

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

As a robustness check of the separability of the three variables above we followed Example 3.8 to estimate the model $Y = \beta(W)X + Q'\pi + U$, with $\beta(W) = \beta + \beta_1' W$, under our assumption $\mathbb{E}[U|Z,Q,W] = \mathbb{E}[U|W]$. The results are very similar and can be seen in Row (iii). Also, as explained in Section 3.1.2, we can test the separability of W and X by testing whether $\beta_1 = 0$. Row (iv) shows the estimates $\hat{\beta}_1$, which are all insignificant, and thus we are confident that our choice of W satisfies the structural separability requirement of our method.

To further test the separability of W assumption, we also ran the same regressions in Row (ii) using each of the other controls in the model as W instead. The results are extremely similar. The 5 controls 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 Regression Discontinuity Design 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 over-identification 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. Given our test confirming the separability of W, we restrict ourselves to the model $Y = g(X) + Q'\pi + U$, with $\mathbb{E}[U|Z,Q,W] = \mathbb{E}[U|W]$, where g is a connected piecewise linear function, and thus X can have richer marginal effects on Y. This model allows us to compare our results directly with those of Chay and Greenstone. 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 $\psi_1^3(X)$, $\psi_2^3(X)$ and $\psi_3^3(X)$. For W we used the three variables in column D in Table 5 (call them $High$, $HealZh$ and $Educ$) expanded into the elements of the B-spline basis. So, with some abuse of notation $W =$

Figure 1: 3 pieces

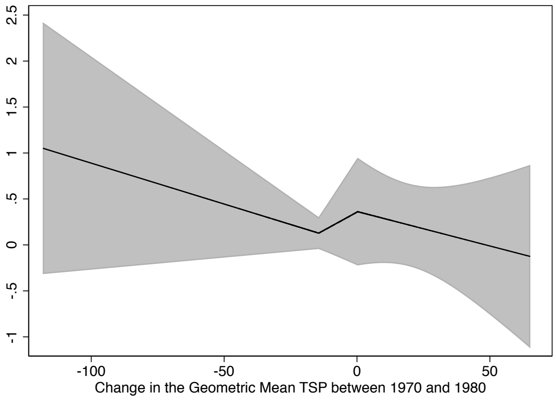


Figure 2: 4 pieces

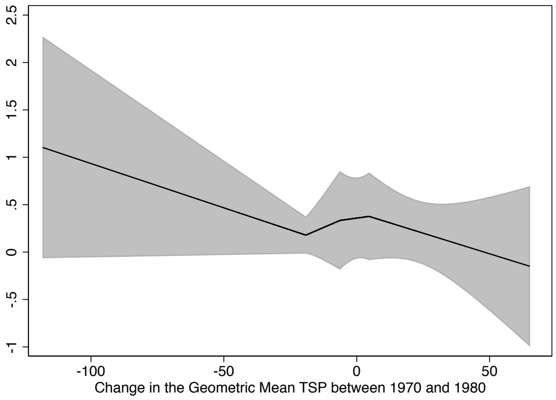


Figure 3: 5 pieces

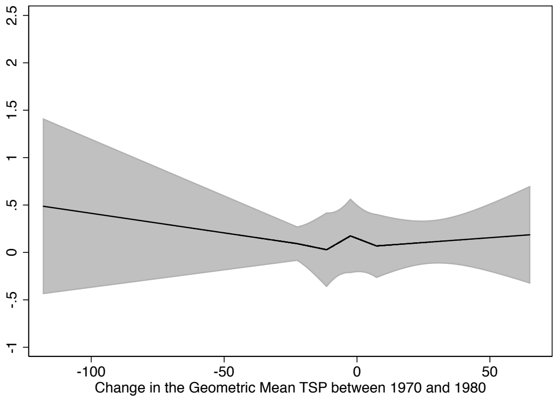
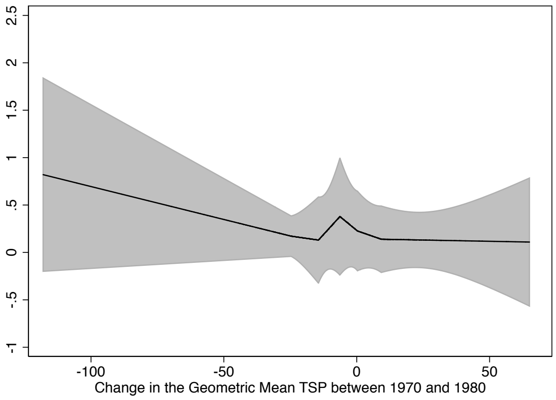


Figure 4: 6 pieces



Figures 1 to 4: Nonlinear case – IV approach. Curves are the results of our approach with exogenous controls, and Z as in column D in Table 5. 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.

$(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High), \psi_1^3(HealZh), \psi_2^3(HealZh), \psi_3^3(HealZh), \psi_1^3(Educ), \psi_2^3(Educ), \psi_3^3(Educ))'$. 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 note 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 5 separately (i.e. for g with three pieces we specified $W = \left(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High)\right)'$, and we also expanded the elements of W in other ways, e.g. 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).

7 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 controls. The strategy hinges on the heterogeneity of the "first stages" and a structural separability of some controls, 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 controls as if they were exogenous, and runs a first stage with interactions between the IV and the controls. 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 controls 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 we have devised a practically convenient semiparametric estimator 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 3.1: Note that $A\widetilde{g} = \Delta_Y$, with $\widetilde{g}(X) = g(X) - \mathbb{E}[g(X)]$ and A is invertible on the orthocomplement of $\mathcal{N}(A)$, which by Assumption 3 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 $\widetilde{g} = A^{-1}\Delta_Y$ and therefore $\widetilde{g}$ is identified. ■

Proof of Theorem 3.2: Note that under Assumption 2, the equation $m = Ag$ is equivalent to

$$\mathbb{C}\left(Y, Z|W\right) = \mathbb{C}\left(g(X), Z|W\right). \tag{27}$$

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, covariance completeness implies $g_0 = g_1$ 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 $g_1 \in \mathcal{G} - \{g_0\}$, $g_1 \neq 0$, such that $\mathbb{C}\left(g_1(X), Z|W\right) = 0$. Then, identification of g_0 in $\mathcal{G}$ fails, because $g_2 \equiv g_0 + g_1 \in \mathcal{G}$ and g_2 satisfies (27). ■

Proof of Theorem 3.4: 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, $\mathbf{X}^*$ and $\mathbf{W}$, where $\mathbf{X}^*$ is the population fitted value

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

By the Frisch-Waugh-Lowell Theorem, the slope of $\mathbf{X}^*$, say β^*, is given by

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

where $\Pi_W = \mathbf{X}^* - \mathbb{E}[\mathbf{X}^*|W] = \mathbb{E}[\mathbf{X}|Z, W] - \mathbb{E}[\mathbf{X}|W]$. 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 $\mathbb{E}[\Delta_{\mathbf{X}}\Delta_{\mathbf{X}}']$ is invertible. To see that, suppose $\mathbb{E}[\Delta_{\mathbf{X}}\Delta_{\mathbf{X}}']$ is singular. Then, there exists a $\lambda \neq 0$ such that, a.s.

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

Then,

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

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 (19) into β^*, yields

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

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

$$Y = \alpha + \beta'\mathbf{X} + \gamma'\mathbf{W} + \varepsilon,$$

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

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

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

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

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 $\hat{\Delta}_Y(w) := \hat{d}(w, 1) - \hat{d}(w, 0)$.

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

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

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

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

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 4 *Suppose that*

1. *The data $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ are iid and Assumption 3 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 $z = 0, 1$, $d_z \in C^{r_m}(\mathcal{S}_W)$ 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 4 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

$$Y = \alpha + \beta' \psi^{k_n}(X) + \gamma' B^{J_n}(W) + \varepsilon_n, \tag{28}$$

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

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

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 (28) 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 (23), i.e.

$$\hat{g} = \arg\min_{g \in \mathcal{G}_n} \mathbb{E}[\left(\hat{C}_Y - (\hat{C}g)(W) \right)^2], \tag{29}$$

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

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

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

$$Z \perp U|W, \tag{30}$$

which is traditionally imposed in the literature. Following Matzkin (2003), we assume that U follows a $U[0,1]$ distribution. Then, conditional independence is equivalent to

$$\mathbb{C}\left(1(U \leq u), Z|W\right) = 0 \text{ a.s. for all } u \in [0,1]. \tag{31}$$

Let U^* be an auxiliary random variable distributed as $U[0,1]$ and independent of (Y,X,Z,W,U). Then, by independence (31) 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 (31) 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 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.

Ben-Moshe, D., D’Haultfoeuille, X., and A. Lewbel (2016): “Identification of Additive and Polynomial Models of Mismeasured Regressors Without Instruments”, working paper.

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 (2015): "Identification of Nonseparable Models with Endogeneity and Discrete Instruments," *Econometrica*, 83, 1199-1210.

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.

Huang, L., Khalil, U. and N. Yildiz (2015): “Identification and Estimation of a Triangular model with Multiple Endogenous Varibles and Insufficiently Many Instrumental Variables,” working paper.

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.

Kim, K.I and S. Song (2017): “Estimation of Semiparametric Models with Mismeasured Endogenous Regressors Using Control Variables,” working paper.

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.

Stock, J. H., and M. W. Watson (2011): Introduction to econometrics (3rd ed.). Boston: Addison Wesley.

Torgovitsky, A. (2015): “Identification of Nonseparable Models using Instruments with Small Support,” *Econometrica*, 83, 1185-1197.