Identification and Estimation of Partial Effects with Proxy Variables[*]

Kenichi Nagasawa[†]

September 30, 2020

Abstract

I develop a new identification approach for partial effects in nonseparable models with endogeneity. I use a proxy variable for the unobserved heterogeneity correlated with the endogenous variable to construct a valid control function, where the definition of a proxy variable is the same as in the measurement error literature. The identifying assumptions are distinct from existing methods, in particular instrumental variables and selection on observables approaches, and I provide an alternative identification strategy in settings where existing approaches are not applicable. Building on the identification result, I consider three estimation approaches, ranging from nonparametric to flexible parametric methods, and characterize asymptotic properties of the proposed estimators.

Keywords: Nonseparable models, bounded completeness, control function, the average structural function.

[*] This paper is based on a portion of my job market paper, which was circulated under the title "Identification and Estimation of Group-Level Partial Effects". I am grateful to my advisor Matias Cattaneo for advice and encouragement. I would like to thank Sebastian Calonico, Max Farrell, Yingjie Feng, Andreas Hagemann, Michael Jansson, Lutz Kilian, Xinwei Ma, Eric Renault, Rocío Titiunik, Gonzalo Vazquez-Bare, seminar participants at Brandeis, Bristol, Florida, LSE, NC State, Pittsburgh, Rice, UC Irvine, UPenn, and Warwick, and conference participants at Microeconometrics Class of 2019 and Causal Learning with Interactions for helpful comments.

[†] The University of Warwick, Department of Economics.

1 Introduction

I consider identification and estimation of partial effects of X on Y in the nonseparable model

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

where ε, ζ are unobserved heterogeneity, $X \perp\!\!\!\perp \zeta | \varepsilon$, and the term ε potentially causes the endogeneity issue (i.e., $X \not\perp\!\!\!\perp \varepsilon$). In addition, the model contains covariates Z, which is assumed to be independent of ζ given (X, ε) but may have arbitrary dependence with (X, ε). Here, the dimension of ζ is not restricted, the endogenous variable X can be continuous, discrete, or have a mixed distribution, and I do not explicitly model the determination of X. This model is important as it captures heterogeneous responses to treatments, and much effort has been devoted to develop various identification strategies (for recent review articles, see Abadie and Cattaneo, 2018; Matzkin, 2013). In this paper, I present a new identification result using a proxy variable W for ε. The definition of proxy is borrowed from the nonclassical measurement error literature (for a recent survey, see Schennach, 2016). Specifically, W satisfies the conditional independence restriction $W \perp\!\!\!\perp (X, Z) | \varepsilon$, and ε is assumed to be bounded complete for W.

To explain the main idea, recall that the identification challenge in (1) arises from the conditional distribution of ε varying with the value of X. Specifically, to identify partial effects, we need to compare groups of individuals with different levels of X while holding constant the distribution of ε, which is made non-trivial, if not impossible, by the presence of endogeneity. In this paper, I show that if two groups of individuals have the same proxy distribution (i.e., $f_{W|XZ}(\cdot|x, z) = f_{W|XZ}(\cdot|\tilde{x}, \tilde{z})$), then they also have the same unobserved heterogeneity distribution ($f_{\varepsilon|XZ}(\cdot|x, z) = f_{\varepsilon|XZ}(\cdot|\tilde{x}, \tilde{z})$).[1] Thus, conditional on $V = f_{W|XZ}(\cdot|X, Z)$, the conditional distribution of ε becomes invariant in changes in X. That is, V is a valid control function as it satisfies $X \perp\!\!\!\perp \varepsilon | V$ (Blundell and Powell, 2004). Identification of partial effects follows from this conditional independence result, provided that some support condition holds. Note that $f_{W|XZ}(\cdot|X, Z)$ is a random function as it is a density function whose randomness comes from (X, Z). Use of a random function as a control variable is non-standard, and yet it does not pose any difficulties in the identification argument.[2]

[1] Given random vectors A, B, I write $f_{A|B}$ for a conditional density of A given B with respect to some measure.

[2] Recently, Arkhangelsky and Imbens (2019) also used the conditional distribution of covariates as a control variable. Despite this similarity, their setting and identifying assumptions are quite distinct from those of this paper.

The main contribution of this paper is to provide a new identification strategy whose identifying assumptions are distinct from those of existing results. In particular, my approach neither requires instruments nor imposes selection on observables assumptions. To demonstrate, consider IV and selection on observables approaches using the above notation. For IV methods, we require Z to be an instrument i.e., independent of ε and related to X, whereas the proxy variable approach in this paper allows for arbitrary dependence between Z and ε. The cost of the proxy approach is to require an additional variable W that functions as a proxy for ε. For the selection on observables approach, it imposes $X \perp\!\!\!\perp \varepsilon | W, Z$, while the identifying assumption in this paper accommodates the case $X \not\perp\!\!\!\perp \varepsilon | W, Z$. Another way to compare the two approaches is that the selection on observables approach would assume that ε is observed and use it as a conditioning variable, whereas in this paper I use an imperfect measurement of ε, namely the proxy W.

This paper builds on results from nonclassical measurement error models, specifically, the insight from Hu and Schennach (2008) and the subsequent developments in the area. The identification approach developed by Hu and Schennach has been applied outside the context of measurement error models (e.g., Arellano et al., 2017; Cunha et al., 2010; Freyberger, 2018; Sasaki, 2015; Wilhelm, 2015), and this paper uses some of their insights. In particular, in this paper I use the definition of proxy that is taken from this strand of literature. One notable difference in this paper is that I treat distributional features of unobserved heterogeneity as nuisance parameters, whereas such objects are of main interest in the aforementioned papers. To be more specific, I use the completeness condition to ensure the existence of an injective mapping from $f_{\varepsilon|XZ}(\cdot|x,z)$ to $f_{W|XZ}(\cdot|x,z)$, but I do not need to recover the object $f_{\varepsilon|XZ}(\cdot|x,z)$ for my identification result. Consequently, I only require one proxy variable as opposed to the two required in the literature, and also the estimation procedures do not involve solving for integral equations, circumventing ill-posed inverse problems.

Many existing results employing IV jointly model the outcome determination and the first-stage equation using a triangular system (Chesher, 2003, 2005; D'Haultfoeuille and Février, 2015; Florens et al., 2008; Imbens and Newey, 2009; Torgovitsky, 2015; Vytlacil and Yildiz, 2007). These papers impose some form of monotonicity in the first-stage equation to achieve identification. In this paper, I do not explicitly model the first-stage equation, and consideration of special cases indicates that my method does not rely on monotonicity in the first stage. Therefore, the identification approach in this paper provides an interesting alternative in settings where the monotonicity assumption fails. In addition, my result treats continuous and discrete X in a unified way, which contrasts with the existing results in the triangular model literature as they develop specific approaches, depending on the nature of endogenous variables. As already mentioned, these advantages come at the cost of requiring a proxy variable for the unobserved heterogeneity, and it depends on specific applications which of the two methods is more appealing.

Other related papers include Altonji and Matzkin (2005), who showed that the equality of the unobserved heterogeneity distribution $f_{\varepsilon|XZ}(\cdot|x,z) = f_{\varepsilon|XZ}(\cdot|\tilde{x},\tilde{z})$ has identifying power in nonseparable models. They deduced this equality from a panel data structure and what they call the exchangeability condition, while the present paper uses the proxy condition. Also, whereas the leading example in Altonji and Matzkin is a panel data model, the focus in this paper is on cross-sectional settings.

Building on the identification result I propose several estimators, ranging from nonparametric to flexible parametric methods. To construct these estimators and characterize their asymptotic properties, I build on several strands of literature. For nonparametric estimation, I rely on the literature on nonparametric functional data analysis (see Ferraty and Vieu, 2006, for a textbook treatment). For semiparametric estimation, I rely on the extensive literature on estimation with generated covariates (e.g., Hahn and Ridder, 2019). For ease of implementation and relaxing some of the identifying assumptions, I also consider flexible parametric modelling, following the approaches of Chernozhukov et al. (2020) and Newey and Stouli (2019).

In the next section, I describe the econometric model and present the identification argument. In Section 3, I propose three estimators and discuss their asymptotic properties. Section 4 is the conclusion. The proofs of the theorems are in the supplemental appendix.

2 Identification

Given the model (1) described in the introduction, I focus on a conditional version of the average structural function (ASF). It is defined as

$$\mu(x) = \mathbb{E}[g(x,\varepsilon,\zeta)|(X,Z) \in \overline{\mathcal{XZ}}],$$

where $\overline{\mathcal{XZ}}$ is some subset of $\text{supp}(X, Z)$ with $\text{supp}(\cdot)$ denoting the support of a random variable, and $\mu(x)$ reduces to the ASF if we take $\overline{\mathcal{XZ}} = \text{supp}(X, Z)$. This object represents a ceteris paribus effect of X on the average outcome Y: as we vary x, the distribution of (ε, ζ) is held constant, so any change in the average outcome can be attributed to the variation in x. In the literature, some studies focused on conditional effects like $\mu(x)$ (Fernández-Val et al., 2018; Vytlacil and Yildiz, 2007) as doing so has the benefit of requiring a weaker support condition. Here I follow the same approach. There exist other parameters of interest beyond average effects, such as quantile effects (Imbens and Newey, 2009). For brevity, I consider these objects in the supplemental appendix.

2.1 Assumptions

Assumption ID Let $\mathcal{X}_0$ be some subset of $\text{supp}(X)$.

(i) (*Idiosyncratic term*) $\zeta \perp\!\!\!\perp (X, Z)|\varepsilon$.

(ii) (*Proxy*) $W \perp\!\!\!\perp (X, Z)|\varepsilon$ and for any bounded function $\tilde{f}$,

$$\mathbb{P}\big[\mathbb{E}[\tilde{f}(\varepsilon)|W] = 0\big] = 1 \quad \text{implies} \quad \mathbb{P}\big[\tilde{f}(\varepsilon) = 0\big] = 1.$$

(iii) (*Regularity*) The distribution of (W, X, Z, ε) is absolutely continuous with respect to some σ-finite product measure, $f_{\varepsilon|XZ}/f_\varepsilon$ is bounded, $f_{W|XZ}$ is continuous in W argument and bounded, and $\mathbb{E}[|g(x, \varepsilon, \zeta)||(X, Z) \in \overline{\mathcal{XZ}}] < \infty$ for $x \in \mathcal{X}_0$.

(iv) (*Support condition*) Let $V = \{f_{W|XZ}(w|X, Z) : w \in \text{supp}(W)\}$. For $x \in \mathcal{X}_0$, the conditional support of V given $X = x$ contains the support of V given $(X, Z) \in \overline{\mathcal{XZ}}$.

The condition $\zeta \perp\!\!\!\perp (X, Z)|\varepsilon$ formalizes the idea that ζ is an idiosyncratic term and it is ε that causes the endogeneity issue. One implication of this conditional independence assumption is

$$Y \perp\!\!\!\perp Z|X, \varepsilon.$$

This demands that the variable Z is excluded from the outcome equation, in the sense that conditional on (X, ε) Z has no influence on the outcome. This type of exclusion restriction is distinct from the instrument exclusion restriction. In particular, IV approaches require that $Z \perp\!\!\!\perp \varepsilon$, whereas this setting allows for arbitrary correlation between Z and ε. In fact, dependence between Z and ε may help satisfy the support condition (iv) as discussed below.

Condition (ii) describes the formal requirements for proxy. The conditional independence assumption states that once we condition on the "correctly measured variable" (ε), then its "noisy measurement" (W) is independent of other variables. This type of conditional independence is standard in errors-in-variables models (e.g., Schennach, 2016). As an example, consider wage regression, where the endogenous variable X is years of schooling and Z is some observed variable that is correlated with either educational attainment X or worker's ability ε (e.g., spouse's education). Here, W could be a test score that measures aspects of cognitive skills and it is a function of the ability ε. Then, I impose that conditional on the ability, the remaining variation in the test score is independent of worker's education X and spouse's education Z.

The second part of (ii) is bounded completeness and has been used extensively for nonparametric identification (see Carrasco et al., 2007, and references therein). For instance, Newey and Powell (2003) used a completeness condition to generalize the IV rank condition in linear models to nonparametric settings. In my model, the completeness assumption formalizes the idea that the proxy variable has a strong relationship with the unobserved heterogeneity. Since this is a rank condition, the dimension of W should be at least as large as that of ε, which is analogous to IV rank conditions. In applications, researchers should be explicit about what ε represents. Otherwise, it would be difficult to assess the plausibility of this assumption.

The definition of a proxy variable is borrowed from the literature on nonclassical measurement error models.[3] Condition (ii) follows Hu and Schennach, 2008, Assumptions 2 and 3. This type of proxy assumption has also been applied outside the context of measurement error problems, particularly in panel data models where past observations can be used as proxies for future observations (e.g., Arellano et al., 2017). Therefore, although my model applies to cross-sectional settings, panel datasets may provide natural candidates for proxy variables. One important distinction in this paper is that distributional features of unobserved heterogeneity are nuisance parameters, but they are of main interest in nonclassical measurement error models. Due to this difference in objects of

[3] In the measurement error literature, the term IV is sometimes used to refer to an additional mismeasured variable because it satisfies a version of exclusion restriction (conditional independence) and relevance condition (rank condition or completeness). In the current context, IV has a very specific meaning: it is related to the endogenous variable X and independent of ε. As W in Assumption ID violates these requirements, IV is not an appropriate term in this setting.

interest, I only require one proxy variable rather than two.

The completeness assumption has been studied extensively in the context of nonparametric identification, and there are several known sufficient conditions in the literature (e.g., Andrews, 2017; D'Haultfoeuille, 2011; Hu et al., 2017). These results can be used to assess the plausibility of condition (ii). In empirical applications, some studies model unobserved heterogeneity as finite discrete types, and this approach renders the full rank condition less demanding than in the continuous variable case. Recently, Bonhomme et al. (2019) used this approach to study earning distributions, explicitly accounting for worker and firm unobserved heterogeneity. Wilhelm (2015) also discussed justifications for completeness assumptions in a panel data setting. Often completeness conditions in identification arguments lead to solving for integral equations, but my identification strategy does not recover the distribution of ε, and consequently, I do not need to estimate the inverse of integral transforms. As a result, the identification result does not lead to ill-posed inverse estimation problems.

Condition (iii) collects mild regularity conditions. The dominating measure can be the Lebesgue measure, the counting measure, or the product measure of the Lebesgue and counting measures. Thus, the assumption accommodates continuous and discrete endogenous variables as well as mixtures of the two. Boundedness of $f_{\varepsilon|XZ}/f_\varepsilon$ can be modified by strengthening (ii) to a stronger completeness condition such as L^2-completeness.

Condition (iv) is known as a common support condition in the literature and it requires that the random element $V = f_{W|XZ}(\cdot|X,Z)$ has enough variation conditional on $X = x$. To explain, consider the support of $f_{\varepsilon|XZ}(\cdot|X,Z)$ instead, which is equivalent because condition (ii) implies the existence of a one-to-one mapping between $f_{\varepsilon|XZ}(\cdot|X,Z)$ and $f_{W|XZ}(\cdot|X,Z)$ (see the next subsection for details). One way for $f_{\varepsilon|XZ}(\cdot|X,Z)$ to have variation given $X = x$ is that ε and Z have a strong relationship. Another possibility is that Z affects X. To describe this mechanism, suppose the endogenous variable X is determined by the process

$$X = h(Z, \eta)$$

where h is some function and η is an unobserved random variable. This formulation resembles the first-stage equation of IV approaches. Endogeneity of X suggests that η and ε are dependent, and varying Z while holding constant X affects the distribution of ε through η. Note that this channel is exploited by control function methods in triangular models (e.g., Imbens and Newey, 2009). In contrast, the first mechanism (correlation between Z and ε) is specific to this setting, as IV approaches require independence of ε and Z. Also, for the support condition to hold, the functional form of $f_{\varepsilon|XZ}$ needs some restrictions. For instance, if X only affects the mean of ε while Z only influences its variance, the common support condition fails. Thus, we generally require that $f_{\varepsilon|XZ}(e|x,z) = p(e,\theta(x,z))$ where p, θ are some functions and θ is not injective.

To further examine the issue of common support, I consider the following special case: the endogenous variable is determined via the first-stage equation $X = Z_1\eta_1 + Z_2\eta_2$ where $X \in \mathbb{R}\backslash\{0\}$, $Z = (Z_1, Z_2) \in \mathbb{R}^2\backslash\{(0,0)\}$, and $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2\backslash\{(0,0)\}$. Also, impose $\varepsilon \perp\!\!\!\perp Z|\eta$, which states that endogeneity arises from dependence between η and ε. In this case, the conditional distribution of ε given (X,Z) is determined by the set $\{\eta \in \operatorname{supp}(\eta) : X = Z_1\eta_1 + Z_2\eta_2\}$. Then the common support condition $\operatorname{supp}(V|X=x) \supset \operatorname{supp}(V|(X,Z) \in \overline{\mathcal{XZ}})$ holds if for each $(\tilde{x}, \tilde{z}) \in \overline{\mathcal{XZ}}$, there exists $z \in \operatorname{supp}(Z|X=x)$ such that $z_1/\tilde{z}_1 = x/\tilde{x} = z_2/\tilde{z}_2$. Thus, in this special case, some positive identification result can be obtained even though the dimension of unobserved heterogeneity is greater than that of X. This observation contrasts with results in triangular models where low dimensionality of unobserved heterogeneity seems to be crucial (Imbens, 2007; Kasy, 2014). This example also illustrates that we cannot allow for arbitrary first-stage unobserved heterogeneity, as increasing the dimension of η to three (e.g., $X = Z_1\eta_1 + Z_2\eta_2 + \eta_3$) destroys the positive result.

Assumption ID does not distinguish a continuous/discrete endogenous variable X, and yet the nature of X has some implications for the support condition. Wooldridge (2015) suggests that explicit construction of a control function may be more challenging when the endogenous variable is discretely distributed. As an example, consider binary X which is determined by the threshold-crossing model $X = \mathbb{1}\{h(Z) \geq \eta\}$ where η follows the $(0,1)$ uniform distribution and I assume $Z \perp\!\!\!\perp (\varepsilon, \eta)$ for the sake of discussion. Then, the conditional distribution $F_{\varepsilon|XZ}(e|1,z)$ equals $\mathbb{P}[\varepsilon \leq e | h(z) \geq \eta]$. Thus, to identify average causal effects of changing $X = 0$ to $X = 1$, the common support condition requires the existence of values z and z' in the support of Z such that $h(z) = 1$ and $h(z') = 0$. Therefore, in this setting, the identification argument involves the identification-at-infinity approach (Chamberlain, 1986; Heckman, 1990). This feature is similar to the existing results. What is distinct is that the IV assumption $Z \perp\!\!\!\perp (\varepsilon, \eta)$ is not necessary as long as the proxy W satisfies condition (ii).

2.2 Identification result

Using the proxy condition (ii), we have

$$f_{W|XZ}(w|x,z)=\int f_{W|XZ\varepsilon}(w|x,z,e)f_{\varepsilon|XZ}(e|x,z)de=\int f_{W|\varepsilon}(w|e)f_{\varepsilon|XZ}(e|x,z)de$$

where the first equality follows from the law of total probability and the second equality from the conditional independence $W \perp\!\!\!\perp (X,Z)|\varepsilon$. Then, from simple calculations using $f_{W|\varepsilon}(w|e)=f_{W\varepsilon}(w,e)/f_\varepsilon(e)$, bounded completeness implies that the above integral transform is injective. Thus, there exists a mapping Ψ that takes a function of W to a function of ε such that

$$f_{\varepsilon|XZ}(e|X,Z)=\Psi(f_{W|XZ}(\cdot|X,Z))(e)\equiv\Psi(V)(e)$$

where $V=f_{W|XZ}(\cdot|X,Z)$. This equation implies that if $f_{W|XZ}(\cdot|x,z)=f_{W|XZ}(\cdot|\tilde{x},\tilde{z})$, then $f_{\varepsilon|XZ}(\cdot|x,z)=f_{\varepsilon|XZ}(\cdot|\tilde{x},\tilde{z})$; the latter equality was the basis of the identification results in Altonji and Matzkin (2005). For the current purpose, this observation indicates that conditional on V, the distribution of ε becomes invariant with respect to changes in X i.e., $X \perp\!\!\!\perp \varepsilon|V$.[4] Therefore, the random element $f_{W|XZ}(\cdot|X,Z)$ plays the role of a valid control function. Given the control function, identification of average partial effects follows provided that the common support condition (iv) holds.

The following statement formalizes this heuristic argument.

Theorem 1. *Suppose Assumption ID holds. Then, the conditional ASF $\mu(x)$ is identified from the joint distribution of (Y,X,Z,W) for $x\in\mathcal{X}_0$. In particular,*

$$\mu(x)=\int\mathbb{E}[Y|X=x,V=v]dF_{V|\overline{\mathcal{XZ}}}(v)$$

where $F_{V|\overline{\mathcal{XZ}}}$ denotes the distribution of V conditional on $(X,Z)\in\overline{\mathcal{XZ}}$.

[4] The identification argument hinges on this conditional independence result. Thus, researchers may use this identifying restriction to conduct sensitivity analysis (e.g., Masten and Poirier, 2018). The results of Masten and Poirier apply to binary X, and thus with continuous X, a researcher needs to transform it to a discrete one (e.g., binning) to conduct this exercise.

Assumption ID is distinct from the conditions of IV and selection on observables approaches, and thus Theorem 1 provides an identification result in settings where IV and conditional independence assumptions are difficult to justify. For example, a researcher might not have credible instruments but may have access to detailed information about individual characteristics e.g., large survey datasets. In such circumstances, the proxy assumption imposed in this paper may be reasonable and provide an alternative identification strategy. Specifically, this paper imposes a weaker exclusion restriction than IV methods. That is, I assume $Y \perp\!\!\!\perp Z|X, \varepsilon$, which allows for dependence between Z and ε and is weaker than the IV exclusion restriction $Z \perp\!\!\!\perp \varepsilon$. The cost of the weaker exclusion restriction is that a researcher needs to find some proxy variable satisfying Assumption ID (ii). The bounded completeness is analogous to the IV rank condition, and the conditional independence restriction $W \perp\!\!\!\perp (X, Z)|\varepsilon$ often holds if W represents a measurement specifically designed to capture an aspect of the unobserved heterogeneity (e.g., carefully designed tests to measure cognitive skills). In particular, Assumption ID (ii) can hold in observational datasets without quasi-experimental variation.

For the selection on observables assumption, such an approach imposes that $X \perp\!\!\!\perp \varepsilon|W, Z$. One way to rationalize this condition is to stipulate that $\varepsilon = k(W, \nu)$ and $X \perp\!\!\!\perp \nu|W, Z$ where k is some function and ν is unobserved heterogeneity. We can view this formulation as a Berkson-type measurement error problem (see e.g., Schennach, 2013) in the sense that the unobserved, error-free variable comes to the left-hand side and the mismeasured variable comes to the right-hand side. On the other hand, for my approach, a formulation following classical measurement errors is also valid. Specifically, observed variables come to the left-hand side and the unobserved variables to the right-hand side i.e., $W = k(\varepsilon, \nu)$ (e.g., Cunha et al., 2010). In this setting, conditioning on W does not necessarily eliminate the variation in ε causing the endogeneity issue. In particular, it is possible to have $X \not\!\perp\!\!\!\perp \varepsilon|W, Z$ under Assumption ID, so systematic selection into treatment is accommodated. Depending on specific applications, one assumption may be more reasonable than the other. For instance, when ε represents ability and W a test score, it is more natural to view test scores as some function of ability and other variables (i.e., $W = k(\varepsilon, \nu)$).

The underlying idea behind Theorem 1 may be of theoretical interest beyond the setting of this paper. As the discussion on the special case $X = Z_1\eta_1 + Z_2\eta_2$ highlights, the proxy variable approach in this paper does not require monotonicity in the first-stage equation. Thus, the identification approach can be applied to triangular models where monotonicity with respect to the first-stage unobserved heterogeneity fails. Such examples are considered in the supplemental appendix. Specifically, I analyze identification of treatment effects when the treatment is multivalued. In that setting, the values of the treatment variable may not have a natural ordering, and thus the conventional notion of monotonicity may fail to hold. As many papers on triangular models impose some form of monotonicity in the first-stage equation, the proxy variable method provides an interesting alternative.

3 Estimation

I consider three approaches for estimation of the average partial effects. Building on the identification result above, I estimate the object

$$\mu(x_0) = \int \mathbb{E}[Y|X = x_0, V = v] dF_{V|\overline{\mathcal{X}\mathcal{Z}}}(v)$$

where x_0 is some fixed value in $\operatorname{supp}(X)$ and $F_{V|\overline{\mathcal{X}\mathcal{Z}}}$ is the conditional distribution of V given $(X, Z) \in \overline{\mathcal{X}\mathcal{Z}}$. The estimation involves three steps: (1) estimation of the control function V, (2) estimation of the conditional expectation $\mathbb{E}[Y|X, V]$ given the first-stage estimate $\hat{V}$, and (3) estimation of $\mu(x_0)$ by integrating over the V argument. Different modelling strategies for each of these steps are possible, and I consider nonparametric and semiparametric estimators as well as flexible parametric modelling.

3.1 Nonparametric estimator

First, I consider a fully nonparametric estimation. The first-stage estimator can be any of nonparametric methods, provided that it converges at a reasonably fast rate. For convenience I use $V = F_{W|XZ}(\cdot|X, Z)$, which is equivalent to using the conditional density. Denote the estimator for $F_{W|XZ}$ by $\hat{F}_n$. For the second stage, the regressor V is function-valued, and I apply ideas from the nonparametric functional data analysis literature. Specifically, choose a (semi-)norm $\|\cdot\|$ on the space of distribution functions and the estimator for $\mathbb{E}[Y|X=x,V=v]$ is

$$\hat{m}_n(x,v)=\frac{\sum_{i=1}^n Y_i K_1\left(\frac{X_i-x}{b_n}\right)K_2\left(\frac{\|\hat{V}_i-v\|}{b_n}\right)}{\sum_{i=1}^n K_1\left(\frac{X_i-x}{b_n}\right)K_2\left(\frac{\|\hat{V}_i-v\|}{b_n}\right)}$$

where X is continuously distributed, K_1,K_2 are kernel functions, b_n is a sequence of bandwidth, and $\hat{V}_i=\hat{F}_n(\cdot|X_i,Z_i)$. When X has a discrete distribution, I modify the above estimator to be

$$\hat{m}_n(x,v)=\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i=x\}K\left(\frac{\|\hat{V}_i-v\|}{b_n}\right)}{\sum_{i=1}^n \mathbb{1}\{X_i=x\}K\left(\frac{\|\hat{V}_i-v\|}{b_n}\right)}.$$

This estimator is analogous to the Nadaraya-Watson estimator with changes to accommodate the function-valued regressor V. Given an estimate of the conditional expectation, we can form the estimator for the conditional ASF by

$$\hat{\mu}_n(x_0)=\frac{\sum_{i=1}^n \hat{m}_n(x_0,\hat{V}_i)\mathbb{1}\{(X_i,Z_i)\in\overline{\mathcal{XZ}}\}}{\sum_{i=1}^n \mathbb{1}\{(X_i,Z_i)\in\overline{\mathcal{XZ}}\}}.$$

The above estimator looks analogous to the counterpart in settings where V belongs to the Euclidean space, but one important distinction is that the dimension of V_i is not well-defined. Ferraty and Vieu (2006) explained that the notion of small ball probabilities can be used to define the effective dimension of function-valued regressors. Consider the expectation

$$\mathbb{E}\left[K\left(\frac{\|V_i-v\|}{b_n}\right)\right]$$

which roughly corresponds to the probability of V_i falling into the neighborhood of v with radius b_n when the kernel function is compactly supported and non-negative. In the standard case where $V\in\mathbb{R}^{d_v}$ for some positive integer d_v, this expectation is of order $b_n^{d_v}$ as b_n goes to zero if V has a Lebesgue density. Thus, if we can find some integer d_v such that $b_n^{-d_v}\mathbb{E}[K(\|V_i-v\|/b_n)]$ is bounded away from zero and bounded above, then such d_v represents the effective dimension. In this paper's setting, the random element V has the specific structure $F_{W|XZ}(\cdot|X,Z)$, and we might conjecture that the effective dimension is determined by how (X,Z) enters the conditional distribution function. For instance, if there is a single-index structure $F_{W|X,Z}(w|x,z)=F(w,x'\beta+z'\gamma)$ for some fixed function

F and finite-dimensional vectors β, γ, then the effective dimension will be one. This dimension is unknown to econometricians, but the sample analogue $\sum_i K(\|\hat{V}_i - v\|/b_n)/n$ will be of order $b_n^{d_v}$ under suitable conditions.

In the supplemental appendix, I provide a set of sufficient conditions for consistency of this nonparametric estimator. Although it is of theoretical interest to further characterize asymptotic properties of this estimator (e.g., asymptotic distribution), the practical usefulness of such results may be limited due to the potential curse of dimensionality and the restrictive common support condition. For this reason, I turn to semiparametric and flexible parametric modelling.

3.2 Semiparametric estimator

The nonparametric estimation considered in the previous subsection has some challenges due to the function-valued nature of the control function V. Here, I model the conditional distribution of the proxy W in a flexible parametric way. For example, the conditional distribution of W given (X, Z) is a normal distribution whose mean and variance are some flexibly specified functions of (X, Z). More generally, let $p(\cdot, \theta)$ be a density function with respect to some measure with a finite-dimensional parameter θ. In the above example, $p(\cdot)$ corresponds to the normal density and θ to the mean and variance parameters. I model the conditional distribution of the proxy variable by

$$f_{W|XZ}(w|x, z) = p(w, \theta(x, z)) \tag{2}$$

where $\theta(x, z)$ is some flexibly specified function of (x, z). To emphasize that $\theta(x, z)$ is parametrically specified, I write $\theta(x, z, \beta_0)$ with a finite-dimensional vector β_0 to be estimated. Given this specification, often some $\sqrt{n}$-consistent estimators for β_0 are available e.g., maximum likelihood estimator.

With the model (2), the equality $\theta(x, z, \beta_0) = \theta(\tilde{x}, \tilde{z}, \beta_0)$ implies $f_{W|XZ}(\cdot|x, z) = f_{W|XZ}(\cdot|\tilde{x}, \tilde{z})$. Since Theorem 1 uses this equality of the proxy distribution as the source of identification, we can set $V = \theta(X, Z, \beta_0)$ as a control function. This observation indicates that the estimation problem becomes one of three-step estimation with generated covariates, extensively studied in the literature. For completeness, I provide a set of primitive conditions to characterize the asymptotic distribution of the conditional ASF estimator.

To be specific about the estimator, let $\hat{\beta}_n$ be the first-stage estimator for β_0. Given the first-stage estimate, let $\hat{V}_i = \theta(X_i, Z_i, \hat{\beta}_n)$ and I estimate the conditional expectation $\mathbb{E}[Y|X, V]$ by the qth-order local polynomial regression, which is denoted by $\hat{m}_n(x, v)$. Finally, the estimator for the conditional ASF is the partial means estimator

$$\hat{\mu}_n(x_0) = \frac{\sum_{i=1}^n \hat{m}_n(x_0, \hat{V}_i)\mathbb{1}\{(X_i, Z_i) \in \overline{\mathcal{XZ}}\}}{\sum_{i=1}^n \mathbb{1}\{(X_i, Z_i) \in \overline{\mathcal{XZ}}\}}.$$

Regarding the asymptotic distribution of this estimator, we need to analyze how the first-stage estimation error $\hat{\beta}_n - \beta_0$ and the second-stage estimation of $\mathbb{E}[Y|X, V]$ contribute to the limit distribution. For this purpose, it is useful to look at the cases of continuous and discrete X separately. For the continuous case, the partial means estimator yields a slower-than-$\sqrt{n}$ convergence rate as it does not average over the X argument (Newey, 1994). Since the first-stage estimation error vanishes at the $\sqrt{n}$ rate, we conjecture that it does not affect the limit distribution. This intuition turns out to be true, and the asymptotic distribution equals that of the infeasible estimator treating V as observed. On the other hand, when X has a discrete distribution, the estimator $\hat{\mu}_n(x_0)$ becomes a full mean, giving rise to the $\sqrt{n}$ rate. Therefore, the first-stage estimation error has the first-order contribution. In the following I consider the continuous and discrete cases separately.

3.2.1 Continuous endogenous variables

I impose the following assumptions to analyze the asymptotic behavior of the estimator.

Assumption C Denote the regression function and regression error by $m_0(X, V) = \mathbb{E}[Y|X, V]$ and $\epsilon = Y - m_0(X, V)$, respectively. Let K and b_n be the kernel function and the bandwidth sequence used for the qth-order local polynomial estimation, respectively. $|\cdot|$ denotes the Euclidean norm, and for a set $\mathcal{A}$ and $\delta > 0$, let $\mathcal{A}^\delta = \{a : \inf_{\tilde{a} \in \mathcal{A}} |\tilde{a} - a| \leq \delta\}$. Define $\overline{\mathcal{V}} = \{v : v = F_{W|XZ}(\cdot|x, z), (x, z) \in \overline{\mathcal{XZ}}\}$.

(i) The observation $\{(Y_i, X_i, Z_i, W_i)\}_{i=1}^n$ is a random sample, the set $\overline{\mathcal{XZ}}$ is compact, $\tau_0 = \mathbb{P}[(X, Z) \in \overline{\mathcal{XZ}}] > 0$, V has a continuous distribution, and the conditional probability $\tau(v) = \mathbb{P}[(X, Z) \in \overline{\mathcal{XZ}}|V = v]$ is continuously differentiable.

(ii) The kernel function K is even, compactly supported, and twice continuously differentiable.

(iii) The function $\theta(x, z, \beta)$ is twice differentiable in β around β_0 with bounded derivatives. Also, the first-stage estimator satisfies $|\hat{\beta}_n - \beta_0| = O_{\mathbb{P}}(1/\sqrt{n})$.

(iv) The random vector (X, V) has a joint Lebesgue density, and there exists some $\delta > 0$ such that the density is continuous and positive on $(\{x_0\} \times \overline{\mathcal{V}})^\delta$. Also, $f_V(v)$ and $f_{XV}(x_0, v)$ are differentiable in v with bounded derivatives on $\overline{\mathcal{V}}^\delta$. The random vector (X, Z) has a joint Lebesgue density and the function $\mathbb{E}[\epsilon|X = x, Z = z]f_{XZ}(x, z)$ is differentiable in x around x_0 and satisfies

$$\int \sup_{|x-x_0|\leq\delta} |\mathbb{E}[\epsilon|X = x, Z = z]| f_{XZ}(x, z) + \sup_{|x-x_0|\leq\delta} \Big|\frac{\partial \mathbb{E}[\epsilon|X = x, Z = z] f_{XZ}(x, z)}{\partial x}\Big| dz < \infty.$$

(v) The conditional expectation $m_0(x, v)$ is $(q + 1)$-times differentiable with bounded derivatives on $(\{x_0\} \times \overline{\mathcal{V}})^\delta$ for some $q \geq 1$. The regression error ϵ satisfies $\mathbb{E}[|\epsilon|^s|X, V] \leq C$ on $(\{x_0\} \times \overline{\mathcal{V}})^\delta$ for some $s > 2$ and the conditional variance $\sigma^2(x, v) = \mathbb{E}[\epsilon^2|X = x, V = v]$ is continuous in x at x_0 for all $v \in \overline{\mathcal{V}}^\delta$.

Most of these restrictions are regularity conditions in that they impose sufficient differentiability, boundedness of functions, and finite moments of random variables. Substantive assumptions include positiveness of $\tau_0 = \mathbb{P}[(X, Z) \in \overline{\mathcal{XZ}}]$ and the density of V bounded away from zero on a certain set. These conditions prevent small denominator issues and are crucial for the asymptotic analysis. The existing results in the literature also require analogous assumptions to avoid small denominators. Here, I impose that the control function V has a continuous distribution. This assumption is not strictly necessary, and the results of this paper extend to the case where V has a discrete distribution.

Theorem 2. *Suppose Assumption C holds. Let d_x, d_v be the dimensions of X, V. If the bandwidth sequence satisfies $nb_n^{d_x+2(q+1)} = o(1)$ and $\log n/(nb_n^{d_x+2d_v})^{1/2} \max\{1, (\log n/n^{1-2/s}b_n^{d_x+d_v})^{1/2}\} = o(1)$, then*

$$\sqrt{nb_n^{d_x}}\big(\hat{\mu}_n(x_0) - \mu(x_0)\big) \rightsquigarrow N(0, \sigma_0^2)$$

where $\rightsquigarrow$ denotes convergence in distribution and

$$\sigma_0^2 = \mathbf{e}_1'\mathbf{S}_0^{-1}\mathbf{M}\mathbf{S}_0^{-1}\mathbf{e}_1 \int \frac{f_V^2(v)\sigma^2(x_0, v)\tau^2(v)}{f_{XV}(x_0, v)\tau_0^2} dv,$$

$\mathbf{e}_1$ *is the vector whose first element is unity and the remaining elements are zero,* $\mathbf{S}_0$ *and* $\mathbf{M}$ *are the matrices whose elements are integrals of* $K(u)$ *times polynomials of* u*: details are in the supplemental appendix.*

As hinted earlier, the asymptotic distribution of $\hat{\mu}_n(x_0)$ corresponds to that for the partial means estimator with observed V.

For inference, I propose plug-in variance estimators that replace unknown quantities with their estimates. Specifically,

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{nb_n^d}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(x_0,\hat{V}_j)\boldsymbol{\kappa}\Big(\frac{X_i - x_0}{b_n},\frac{\hat{V}_i - \hat{V}_j}{b_n}\Big)\frac{T_j\hat{\epsilon}_i}{\hat{\tau}_n}\right)^2$$

where $d = d_x + d_v$, $\hat{\epsilon}_i = Y_i - \hat{m}_n(X_i,\hat{V}_i)$, $\hat{m}_n$ is the qth order local polynomial estimator used for the original estimation, $T_i = \mathbb{1}\{(X_i,Z_i)\in\overline{\mathcal{XZ}}\}$, and $\hat{\tau}_n = \sum_{i=1}^{n} T_i/n$. The objects $\hat{\mathbf{S}}_n$ and $\boldsymbol{\kappa}$ involve the kernel function K and the bandwidth b_n and precise definitions are given in the supplemental appendix. Note that no additional estimates are needed for this variance estimator. This variance estimator is consistent under the slightly strengthened conditions of Theorem 2.

Theorem 3. *Assume that the hypothesis of Theorem 2 holds with* $s \geq 4$ *and* $nb_n^{3d_x}\to\infty$*. Then,* $\hat{\sigma}_n^2 \to_{\mathbb{P}} \sigma_0^2$*.*

3.2.2 Discrete endogenous variables

To analyze the estimator with discrete X, I impose the following conditions.

Assumption D I use the notation defined in Assumption C.

(i) The observation $\{(Y_i,X_i,Z_i,W_i)\}_{i=1}^n$ is a random sample, the set $\overline{\mathcal{XZ}}$ is compact, $\tau_0 = \mathbb{P}[(X,Z)\in\overline{\mathcal{XZ}}] > 0$, V has a continuous distribution, and the conditional probability $\tau(v) = \mathbb{P}[(X,Z)\in\overline{\mathcal{XZ}}|V=v]$ is continuously differentiable.

(ii) The kernel function K is even, compactly supported, and twice continuously differentiable.

(iii) The function $\theta(x,z,\beta)$ is twice differentiable in β around β_0 with bounded derivatives. The first-stage estimator admits an asymptotically linear representation:

$$\hat{\beta}_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n}\varphi(W_i, X_i, Z_i) + o_{\mathbb{P}}(1/\sqrt{n})$$

where $\mathbb{E}[\varphi(W,X,Z)] = 0$ and $\mathbb{E}[|\varphi(W,X,Z)|^2] < \infty$.

(iv) The random vector Z has a Lebesgue density conditional on $X = x_0$ and $\rho_0 = \mathbb{P}[X = x_0] > 0$. The random vector V has a Lebesgue density conditional on $X = x_0$, and there exists some $\delta > 0$ such that the conditional density is continuous and positive on $\overline{\mathcal{V}}^{\delta}$. Also, $f_{V|X}(v|x_0)$ is twice differentiable on $\overline{\mathcal{V}}^{\delta}$ with bounded derivatives, and $f_V(v)$ is continuously differentiable on $\overline{\mathcal{V}}^{\delta}$.

(v) The conditional expectation $m_0(x_0, v)$ is $(q+1)$-times differentiable in v with bounded derivatives on $\overline{\mathcal{V}}^{\delta}$ for some $q \geq 1$. The regression error ϵ satisfies $\mathbb{E}[|\epsilon|^s|X = x_0, V] \leq C$ on $\overline{\mathcal{V}}^{\delta}$ for some $s > 2$.

These restrictions are mostly counterparts to Assumption C in the previous subsection. The only substantive difference is that the first-stage estimator $\hat{\beta}_n$ has an asymptotically linear representation. This is a stronger requirement than in Assumption C, and this condition is necessary to characterize the limit distribution of the estimator with discrete X because the first-stage estimation error has the first-order contribution.

Theorem 4. *Suppose Assumption D holds. Let d_v be the dimension of V. If the bandwidth sequence satisfies $nb_n^{2(q+1)} \to 0$, $\log n/\sqrt{n}b_n^{\max\{d_v,3/2\}}\max\{1, (\log n/n^{1-2/s}b_n^{d_v})^{1/2}\}$, and $nb_n^4 \to \infty$, then*

$$\sqrt{n}\big(\hat{\mu}_n(x_0) - \mu(x_0)\big) \rightsquigarrow N(0, \sigma_0^2), \qquad \sigma_0^2 = \mathrm{Var}[\psi + \mathbf{\Gamma}\varphi(W,X,Z)]$$

where letting $T = \mathbb{1}\{(X,Z) \in \overline{\mathcal{X}\mathcal{Z}}\}$,

$$\psi = \frac{m_0(x_0,V)T}{\tau_0} - \mu(x_0) + \frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)\rho_0\tau_0}\mathbb{1}\{X = x_0\}\epsilon$$

$$\mathbf{\Gamma} = \tau_0^{-1}\bigg(- \mathbf{e}_1'\mathbf{S}_0^{-1}\int_{\mathbb{R}^{d_v}} \frac{\partial}{\partial u'}\boldsymbol{\kappa}(u)\mathbb{E}\Big[u'\frac{\partial}{\partial v}\Big\{\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\Big\}\frac{\partial\theta(x_0,Z,\beta_0)}{\partial\beta'}\epsilon\Big|X = x_0\Big]du$$
$$- \mathbb{E}\Big[\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\frac{\partial m(x_0,V)}{\partial v'}\frac{\partial\theta(x_0,Z,\beta_0)}{\partial\beta'}\big|X = x_0\Big] + \mathbb{E}\Big[\frac{\partial m_0(x_0,V)}{\partial v'}\frac{\partial\theta(X,Z,\beta_0)}{\partial\beta'}T\Big]\bigg),$$

$\mathbf{S}_0$ is analogous to the one in Theorem 2, and $\boldsymbol{\kappa}(u)$ is a vector of functions whose elements are products of $K(u)$ and polynomials of u: see the supplemental appendix for the precise definition.

From the representation of the asymptotic variance, we see that the effect of the first-stage estimation error appears as the term $\boldsymbol{\Gamma}\varphi(W, X, Z)$ where $\varphi(W, X, Z)$ is the influence function of the first-stage estimator $\hat{\beta}_n$.

For variance estimation, I assume the existence of some consistent estimator $\hat{\varphi}_n$ for the influence function of $\hat{\beta}_n$. The proposed estimator is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\psi}_{in} + \hat{\boldsymbol{\Gamma}}_n\hat{\varphi}_n(W_i, X_i, Z_i)\big)^2$$

where

$$\hat{\psi}_{in} = \frac{\widehat{m}_n(x_0, \hat{V}_i)T_i}{\hat{\tau}_n} - \hat{\mu}_n(x_0) + \hat{\tau}_n^{-1}\frac{1}{nb_n^{d_v}}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_j)\boldsymbol{\kappa}\Big(\frac{\hat{V}_i - \hat{V}_j}{b_n}\Big)T_j\mathbb{1}\{X_i = x_0\}\hat{\epsilon}_i,$$

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_n = \hat{\tau}_n^{-1}\Bigg(&- \frac{1}{n^2 b_n^{d_v+1}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\frac{\partial\boldsymbol{\kappa}}{\partial u'}\Big(\frac{\hat{V}_j - \hat{V}_i}{b_n}\Big)\frac{\partial\theta(x_0, Z_j, \hat{\beta}_n)}{\partial\beta'}\mathbb{1}\{X_j = x_0\}T_i\hat{\epsilon}_j \\ &- \frac{1}{n^2 b_n^{d_v}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\boldsymbol{\kappa}\Big(\frac{\hat{V}_j - \hat{V}_i}{b_n}\Big)\mathbb{1}\{X_j = x_0\}T_i\widehat{m}_n^{(1)}(x_0, \hat{V}_j)'\frac{\partial\theta(x_0, Z_j, \hat{\beta}_n)}{\partial\beta'} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_n^{(1)}(x_0, \hat{V}_i)'\frac{\partial\theta(X_i, Z_i, \hat{\beta}_n)}{\partial\beta'}T_i\Bigg),\end{aligned}$$

$\widehat{m}_n^{(1)}(x_0, \hat{V}_i)$ is the qth order local polynomial estimate of $\partial m_0(x_0, V_i)/\partial v$, and the definition of $\hat{\mathbf{S}}_n$ and $\boldsymbol{\kappa}$ is given in the supplemental appendix. The formula looks complicated, but estimation of the variance does not require additional nonparametric estimation. In particular, the estimates $\widehat{m}_n(x_0, v)$ and $\widehat{m}_n^{(1)}(x_0, v)$ are already obtained in the original estimation step, and the only additional estimate is $\hat{\varphi}_n$, which is often easy to construct using the structure (2). As in the continuous case, the variance estimator is consistent under a slightly stronger version of the sufficient conditions for asymptotic normality.

Theorem 5. *Assume that the hypothesis of Theorem 4 holds,* $\max_{1\leq i\leq n}|\hat{\varphi}_n(W_i, X_i, Z_i) - \varphi(W_i, X_i, Z_i)| = o_{\mathbb{P}}(1)$*,* $nb_n^{2q} = o(1)$*, and* $\sqrt{\log n/nb_n^{d_v+2}}\max\{1, \sqrt{\log n/n^{1-2/s}b_n^{d_v}}\} = o(1)$*. Then,* $\hat{\sigma}_n^2 \to_{\mathbb{P}} \sigma_0^2$*.*

3.3 Flexible parametric approach

In this section, I discuss how the approach of Chernozhukov et al. (2020) and Newey and Stouli (2019) can be used to estimate $\mu(x)$. This approach offers the benefits of easy implementation and dispensing with the common support condition.

In this approach, I maintain the model (2) for the conditional proxy distribution. Additionally, I model the conditional expectation of the outcome by

$$Y = \gamma_0' \big[p_1(X) \otimes p_2(V)\big] + \epsilon, \qquad \mathbb{E}[\epsilon|X, V] = 0 \tag{3}$$

where γ_0 is the parameter to be estimated, $p_1(X), p_2(V)$ are vectors of transformed X and V, respectively, and $\otimes$ represents the Kronecker product. This modelling approach builds on Newey and Stouli (2019), and identification of the average structural function follows from non-singularity of the matrix $\mathbb{E}[p_1(X) \otimes p_2(V)(p_1(X) \otimes p_2(V))'\mathbb{1}\{(X, Z) \in \overline{\mathcal{XZ}}\}]$. For various sufficient conditions for the non-singular second moment matrix, see the discussion in Newey and Stouli. To estimate the conditional ASF, we first estimate the control function V as in the semiparametric method and estimate γ_0 by the ordinary least squares. Then, the estimator for the conditional ASF can be constructed as

$$\hat{\mu}_n(x) = \frac{\sum_{i=1}^n \hat{\gamma}_n' \big[p_1(x) \otimes p_2(\hat{V}_i)\big] \mathbb{1}\{(X_i, Z_i) \in \overline{\mathcal{XZ}}\}}{\sum_{i=1}^n \mathbb{1}\{(X_i, Z_i) \in \overline{\mathcal{XZ}}\}}$$

where $\hat{\gamma}_n$ is the least squares estimator for γ_0. In this case, estimating the unconditional ASF is also feasible since the common support condition is not needed. In particular, the estimator for the unconditional ASF is

$$\hat{\mu}_n(x) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_n' \big[p_1(x) \otimes p_2(\hat{V}_i)\big].$$

Note that integrating over the V argument does not require any support condition for the conditional distribution of V given X. Dispensing with the common support condition is possible because we can extrapolate using the parametric model. This is in sharp contrast to the nonparametric and semiparametric methods, where I focused on the conditional ASF because the common support condition can be too restrictive to estimate the marginal expectation.

Chernozhukov et al. (2020) described baseline models that justify the model (3). For example, suppose that the outcome variable is determined by the following random coefficient model:

$$Y = \varepsilon_1 + \varepsilon_2 X$$

where $X \in \mathbb{R}$, $\mathbb{E}[\varepsilon_1|X, V] = \mathbb{E}[\varepsilon_1|V] = \gamma_{0,1}' p_2(V)$, and $\mathbb{E}[\varepsilon_2|X, V] = \mathbb{E}[\varepsilon_2|V] = \gamma_{0,2}' p_2(V)$ with $\gamma_0 = (\gamma_{0,1}', \gamma_{0,2}')'$. Then, (3) holds with $p_1(X) = (1, X)'$. This baseline model illustrates that while the flexible parametric model simplifies identification and estimation, it retains nonseparability between X and ε. This feature is crucial as it allows for heterogeneous treatment effects, which is salient in many empirical settings.

Given the identification, the estimation and inference problems become those of parametric three-stage estimation, which are well understood in the literature (e.g., Newey and McFadden, 1994). For completeness, I characterize the limit distribution of the unconditional ASF estimator. The asymptotic distribution of the conditional ASF estimator is analogous and can be obtained with obvious changes.

Theorem 6. *Let $r(x, v) = p_1(x) \otimes p_2(v)$. Assume that the function $r(x, v)$ is twice differentiable in v with bounded derivatives, Assumption D (iii) holds, $\mathbb{E}[|r(X, V)\epsilon|^2] + \mathbb{E}[\epsilon^2] + \mathbb{E}[|r(x_0, V)|^2] < \infty$, and the matrix $\mathbb{E}[r(X, V)r(X, V)']$ is non-singular. Then, under the random sampling assumption, the unconditional ASF estimator satisfies*

$$\sqrt{n}(\hat{\mu}_n(x_0) - \mu(x_0)) \rightsquigarrow N(0, \mathrm{Var}(\psi))$$

where

$$\begin{aligned}\psi = {} & \gamma_0' r(x_0, V) - \mu(x_0) + \mathbb{E}[r(x_0, V)]'\mathbb{E}[r(X, V)r(X, V)']^{-1} r(X, V)\epsilon \\ & + \left\{\mathbb{E}[r(x_0, V)]'\mathbb{E}[r(X, V)r(X, V)']^{-1}(\mathbb{E}[\epsilon D] - \mathbb{E}[r(X, V)\gamma_0' D]) + \gamma_0'\mathbb{E}[D]\right\}\varphi(W, X, Z)\end{aligned}$$

and $D = \partial r(x_0, V)/\partial v' \partial\theta(X, Z, \beta_0)/\partial\beta'$.

I omit the proof as the derivation is a straightforward exercise. For inference, the consistent variance estimator can be obtained using a plug-in estimator, replacing unknown objects with their estimates.

4 Conclusion

In this paper, I developed a new identification strategy for partial effects in nonseparable models with endogeneity. The identifying assumptions are distinct from those of the existing methods, so my result creates a new avenue for identification in settings where the current identification approaches are not adequate. Building on the identification result, I propose several estimators using different modelling assumptions and characterize their asymptotic properties.

As briefly discussed, the underlying idea behind Theorem 1 has wide applicability beyond the main econometric model considered in this paper. In future research, these results can be extended to other settings where nonseparability and endogeneity are key features. An advantage of the proxy method is that it does not rely on monotonicity assumptions, so it can apply to complex models where it is difficult to justify such shape restrictions.

5 Bibliography

Abadie, A. and M. D. Cattaneo (2018): "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

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. W. K. (2017): "Examples of L^2-Complete and Boundedly-Complete Distributions," *Journal of Econometrics*, 199, 213–220.

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734.

Arkhangelsky, D. and G. W. Imbens (2019): "The Role of the Propensity Score in Fixed Effect Models," Working Paper.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Reviews of Economic Studies*, 71, 655–679.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5633 – 5751.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2020): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," *Quantitative Economics*, 11.

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

——— (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73, 1525–1550.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

D'Haultfoeuille, X. and P. Février (2015): "Identification of Nonseparable Triangular Models with Discrete Instruments," *Econometrica*, 83, 1199–120.

Fernández-Val, I., A. van Vuuren, and F. Vella (2018): "Nonseparable Sample Selection Models with Censored Selection Rules," Working Paper.

Ferraty, F. and P. Vieu (2006): *Nonpparametric Functional Data Analysis*, New York, NY: Springer.

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 Effects," *Econometrica*, 76, 1191–1206.

Freyberger, J. (2018): "Nonparametric Panel Data Models with Interactive Fixed Effects," *Review of Economic Studies*, 85, 1824–1851.

Hahn, J. and G. Ridder (2019): "Three-Stage Semi-Parametric Inference: Control Variables and Differentiability," *Journal of Econometrics*, 211, 262–293.

Heckman, J. J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y., S. M. Schennach, and J.-L. Shiu (2017): "Injectivity of a Class of Integral Operators with Compactly Supported Kernels," *Journal of Econometrics*, 200, 48–58.

Imbens, G. W. (2007): "Non-Additive Models with Endogenous Regressors," in *Advances in Economics and Econometrics:Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. K. Newey, and T. Persson, Cambridge University Press, vol. 3, 17–46.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations models without Additivity," *Econometrica*, 77, 1481–1512.

Kasy, M. (2014): "Identification in Triangular Systems Using Control Functions," *Econometric Theory*, 27, 663–671.

Lee, S. and B. Salanié (2018): "Identifying Effects of Multivalued Treatments," *Econometrica*, 86, 1939–1963.

Masten, M. A. and A. Poirier (2018): "Identification of Treatment Effects under Conditional Partial Independence," *Econometrica*, 86, 317–351.

Matzkin, R. L. (2013): "Nonparametric Identification in Structural Economic Models," *Annual Review of Economics*, 5, 457–486.

Newey, W. and S. Stouli (2019): "Control Variables, Discrete Instruments, and Identification of Structural Functions," Forthcoming in Journal of Economietrics.

Newey, W. K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

Newey, W. K. and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111 – 2245.

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

Sasaki, Y. (2015): "Heterogeneity and Selection in Dynamic Panel Data," *Journal of Econometrics*, 188, 236–249.

Schennach, S. (2013): "Regressions with Berkson Errors in Covariates - A Nonparametric Approach," *Annals of Statistics*, 41, 1642–1668.

Schennach, S. M. (2016): "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8, 341–377.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments with Small Support," *Econometrica*, 83, 1185–1197.

Vytlacil, E. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.

Wilhelm, D. (2015): "Identification and Estimation of Nonparametric Panel Data Regression with Measurement Error," CEMMAP Working Paper.

Wooldridge, J. M. (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445.