WARWICK
ECONOMICS

Identification and Estimation of Group-Level Partial Effects

Kenichi Nagasawa

February 2020 No: 1243

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

Identification and Estimation of Group-Level Partial Effects*

Kenichi Nagasawa†

February 10, 2020

Abstract

This paper presents identification and estimation results for causal effects of group-level variables when agents select into groups. I specify a triangular system of equations to model outcome determination and group selection, accommodating general nonseparable models. Using conditional independence and completeness assumptions, I show that the group-level distribution of individual characteristics is a valid control function, conditional on which group-level variables of interest become exogenous. Building on this result, I identify average effects under a common support condition. The key identifying requirements are more plausible in settings where a rich array of individual characteristics are observed. For the identified parameter, I construct a kernel-based estimator and prove its consistency. Although the identification argument uses completeness, the estimation procedure does not involve solving for an ill-posed integral equation.

Keywords: Nonseparable models, control functions, bounded completeness, multi-dimensional unobserved heterogeneity, average structural functions.

*This is a revised version of my job market paper. 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, NC State, Pittsburgh, Rice, 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

Policy makers often design interventions to influence individual outcomes through group-level variables. For instance, a government may relocate disadvantaged children to higher quality schools to improve their academic performance. Given their potential impact, many studies in economics have sought to evaluate group-level policy interventions (see Durlauf, 2004; Durlauf and Ioannides, 2010; Graham, 2018, and references therein). Nevertheless, estimation of group-level treatment effects is challenging. The problem is that individuals select into groups in part based on their unobserved characteristics, and this sorting causes systematic dependence among group-level variables and those individual characteristics. Therefore, comparing outcomes across groups without accounting for differences in unobserved heterogeneity is subject to selection bias.

Specifically, the endogeneity issue arises because the group-level distribution of unobserved heterogeneity varies with group-level characteristics in a systematic way. As an example, consider a setting where students choose schools. Academically motivated students tend to prefer high quality schools, and therefore, high quality schools have a larger proportion of highly motivated students or a right-skewed distribution of student motivation. It is this correlation between school quality and the unobserved distribution of motivation that hinders consistently estimating effects of school quality. Thus, addressing the selection bias issue requires some way to control for the school-level distribution of unobserved heterogeneity.

In this paper, I develop a novel identification strategy motivated by the observation that unobserved motivation varies with other student characteristics observed by an econometrician, and thus observed covariates provide some information about the unobserved. To explain the idea, suppose that a researcher observes parents' education level and that parent's education is positively correlated with (unobserved) child's academic motivation. For ease of exposition, further assume that student's motivation and parents' education level take just two values, high and low. In this simple case, the school-level distribution of motivation reduces to the fraction of highly motivated students within a school. Then, with positive correlation between motivation and parents' education, a higher fraction of highly educated parents implies a larger proportion of highly motivated students within a school. That is, there exists a monotonic relationship between the fractions of highly educated parents and highly motivated students. Because strict monotonicity

implies one-to-one mapping, we conclude that conditioning on the observed distribution of parents' education holds constant the unobserved distribution of student motivation. Then, provided that school quality has sufficient variation remaining, controlling for the school-level fraction of highly educated parents enables identifying the ceteris paribus effect of interest.

I formalize the preceding argument by specifying a triangular system of equations, where the first-stage equation specifies group selection and the structural equation determines an outcome of interest. In my model, both selection and outcome equations are nonseparable in unobserved heterogeneity, and I do not impose monotonicity with respect to unobserved heterogeneity as group choice is a discrete object. To achieve identification, I rely on two main assumptions. One restriction is conditional independence, which is motivated by the structure of how group formulation and selection into groups take place. Another key identifying condition is that observed individual-level covariates have sufficient correlation with unobserved heterogeneity. To formalize the idea of sufficient correlation, I use the notion of bounded completeness, which has been applied in a wide range of nonparametric identification problems. With these conditions, I show that the group-level distribution of individual covariates plays a role of a control function, conditional on which group-level variables of interest become exogenous.

The identification result is somewhat non-standard in that the control variable is function-valued, which belongs to an infinite-dimensional space, and existing results for estimation do not apply directly. In this paper, I propose a kernel-based estimator for average effects of group-level variables and, leveraging results from nonparametric functional data analysis literature, I prove consistency of the estimator. Although I use a completeness assumption in the identification argument, the estimation procedure does not involve solving for an ill-posed integral equation. This feature may be desirable as ill-posedness adversely affects precision of estimates through slower convergence rates, although I do not have results on convergence rates in this paper.

This paper's identification strategy complements existing approaches in the program evaluation literature (c.f., Abadie and Cattaneo, 2018; Imbens and Wooldridge, 2009). As an example, consider a selection-on-observables type assumption, where selection becomes exogenous conditional on observed agent-level covariates W. This assumption means that conditional on W, the selection mechanism is like a random assignment for the purpose of particular analysis. In some observational studies, this type of assumption may be deemed too strong. This paper's identification

strategy allows for non-random selection even after conditioning on W and instead imposes that W has sufficient correlation with the random vector entering the selection equation. In other words, what I require is W be a good proxy for the individual characteristics determining the selection process. Therefore, this paper's identification strategy is more fruitful when a researcher observes a rich array of individual characteristics, as in large survey datasets. As another example of existing methods, consider instrumental variables (IV) method, whose exclusion restriction is often justified by natural/quasi experimental variation. Although a powerful source of identification, such setting is an exception rather than the rule. On the other hand, this paper's identification strategy can be reasonable in absence of experimental variation.

In the context of triangular simultaneous equations models, this paper presents a new set of identifying conditions for average effects of endogenous variables. In particular, this paper applies statistical completeness in a novel way to develop a control function approach. An advantage of my approach is that I obtain identification results in nonseparable models without imposing monotonicity, which can be important when choice variables are discrete or unobserved heterogeneity is multi-dimensional. To elaborate on these contributions, I now review the related literature.

1.1 Related Literature

This paper's model is a special case of nonseparable triangular simultaneous equations models (see Matzkin, 2007, for a review of identification results in these models). In a general triangular model,

$$\begin{aligned} Y &= m(X, \varepsilon) \\ X &= h(Z, \eta) \end{aligned} \tag{1}$$

where the interest lies in partial effects of X on Y and the identification issue is $X \not\perp\!\!\!\perp \varepsilon$. With the independence assumption $Z \perp\!\!\!\perp (\varepsilon, \eta)$, X becomes independent of ε after conditioning on η. As pointed out by Imbens (2007), some type of monotonicity condition on $h(\cdot)$ facilitates identification via restricting possible values of η. For instance, Chesher (2003) and Imbens and Newey (2009) impose strict monotonicity of $h(\cdot)$ with respect to η and thus fixing X and Z holds constant η. For related approaches, see also D'Haultfoeuille and Février (2015); Florens et al. (2008); Matzkin (2016); Torgovitsky (2015). In the case of discrete X with ordered values, Chesher (2005) imposes

weak monotonicity of $h(\cdot)$ that implies η lies in an intrval defined by some functions of X and Z, which, in combination with other restrictions, produces an identified set for the function $m(\cdot)$. In the case of binary X, Heckman and Vytlacil (1999) and Vytlacil and Yildiz (2007) use the threshold-crossing structure, which restricts the range of η via the propensity score. As evident from these existing results, monotonicity restrictions play a crucial role in solving the endogeneity problem in nonseparable models.

However, such monotonicity condition can fail to hold in empirically relevant situations. For instance, when X results from unordered multinomial choices, the conventional notion of monotonicity is not well-defined (see Heckman and Pinto, 2018, for a recent development in this setting). More broadly, when the unobserved heterogeneity η is multi-dimensional, monotonicity fails in general (see e.g., Florens et al., 2008; Imbens, 2007). In this paper, I present a novel identification result without relying on a monotonicity assumption. I instead impose existence of a proxy variable W for η as well as conditional independence assumptions. Importantly, I relax the joint independence assumption $Z \perp\!\!\!\perp (\varepsilon, \eta)$, and thus I do not rely on availability of IV.

There exist other studies that do not impose monotonicity in the first stage equation. Chernozhukov and Hansen (2005) impose a completeness assumption on Z in relation to X to point-identify the structural quantile function. In this paper, I impose completeness on a proxy variable W in relation to the unobserved heterogeneity η. Unlike Chernozhukov and Hansen, I do not require the outcome be continuously distributed. Altonji and Matzkin (2005) do not model the mechanism for X and impose exchangeability among group members to construct a control variable. Also, Chesher and Rosen (2019) present a unified framework of single-equation IV approach to address endogeneity in a wide class of models. Generally, this generalized IV method produces set identification of the structural function.

This paper's model can be viewed as a version of panel data models, although I focus on the group-individual structure rather than the individual-time setting as in many papers on panel data.[1] Recent papers studying models with group structures include Altonji and Mansfield (2018), Arkhangelsky and Imbens (2019), Chetverikov et al. (2016), and Graham et al. (2018). The work by Altonji and Mansfield is most closely related to this paper. Like their paper, I use the idea that the group-level distribution of observed covariates moves together with the distribution of unob-

[1] Here individual is broadly defined, including person, household, firm, and other economic agents.

served heterogeneity. Whereas Altonji and Mansfield exploit functional form restrictions to obtain identification in a linear model, I use a conceptually distinct approach to achieve identification in a nonseparable model. For other papers, they consider different models and assumptions. Yet, interestingly, Arkhangelsky and Imbens also show that the group-level distribution of individual covariates plays the role of a control variable despite differences in identifying assumptions.

In addition, this paper is related to the growing literature on nonparametric identification using completeness. Since the seminal work of Newey and Powell (2003), completeness has been applied to a wide range of econometric identification problems. Examples include nonparametric IV (e.g., Chernozhukov and Hansen, 2005; Darolles et al., 2011; Hall and Horowitz, 2005; Newey and Powell, 2003), errors-in-variables models (e.g., Hu and Schennach, 2008), nonparametric discrete choice models with unobserved product characteristics (Berry and Haile, 2014), nonseparable (dynamic) panel data models (e.g., Arellano et al., 2017; Cunha et al., 2010; Freyberger, 2018; Sasaki, 2015), and semiparametric random coefficients models (Hoderlein et al., 2017). My paper applies completeness in a novel way to construct a control function.

The rest of this paper proceeds as follows. In Section 2, I set up an econometric model and discuss my main identification result. Building on the identification result, in Section 3, I propose an estimator for average effects of group-level variables and prove its consistency. Section 4 concludes.

Notations For random elements a, b, $\text{supp}(a)$ is the support of a, $\text{supp}(a|b)$ indicates the support of the conditional distribution of a given b, $F_{a|b}$ denotes the conditional distribution function of a given b, and $f_{a|b}$ represents a conditional density function of a given b with respect to some measure. $|\cdot|$ is the Euclidean norm and $\|\cdot\|$ denotes the supremum norm on a function space. $\mathbb{1}\{\cdot\}$ is the indicator function, which takes 1 if the statement inside the bracket is true and takes 0 otherwise. For random vectors X, Z, I write d_x, d_z to denote the dimensions of X and Z, respectively.

2 Econometric Model and Identification Result

In this section, I describe the econometric model and discuss the main identification result. I show that the group-level distribution of individual covariates plays the role of a control function. As in the introduction, I use the school example throughout to keep the discussion on concrete terms.

In the model, there exist individuals and groups, indexed by i and g, respectively. Here, group corresponds to school and individual to student. The outcome of interest, denoted by Y_{ig}, can be student's test scores and it is determined by the following education production function:

$$Y_{ig} = m(X_g, \nu_i, u_g) \qquad i = 1, \ldots, N,\ g = 1, \ldots, G \tag{2}$$

where $m(\cdot)$ is an unknown function, X_g is observed school characteristics (e.g., school quality), ν_i is a student-level unobserved variable (e.g., academic motivation), and u_g represents unobserved school features. Also, we observe student-level covariates W_i. In principle, W_i can enter the outcome equation, but as I focus on the effect of X_g on Y_{ig}, I subsume W_i in ν_i as a subvector. In the sequel, I write $\varepsilon_{ig} = (\nu_i, u_g)$.

In (2), the outcome is defined for all groups, but a researcher only observes the outcome variable for the group to which an individual belongs. That is, with $J_i \in \{1, \ldots, G\}$ representing the school student i attends, we only observe $Y_{iJ_i} = \sum_{g=1}^{G} Y_{ig} \mathbb{1}\{J_i = g\}$. I model the group determination by

$$J_i = J(Z_1, \ldots, Z_G, \Theta_i, \zeta_{i1}, \ldots, \zeta_{iG}) \tag{3}$$

where $J(\cdot)$ is a nonparametric function, Z_g denotes observed school characteristics important for student's choice, Θ_i represents student's unobserved preference for different school features, and ζ_{ig} is an idiosyncratic term. The school-level covariates Z_g may contain X_g as a subvector and generally contain elements excluded from the outcome equation. In the school setting, for instance, the quality of athletic programs may matter for student's decision of school but presumably does not affect academic performance directly and thus is excluded from the outcome equation. In Appendix B.1, I consider an extension with no observed excluded elements under an alternative sampling scheme. I note that without affecting the results of this paper, I can include unobserved school characteristics in the selection equation but I omit it for ease of exposition.

The specification in (3) is quite general and encompasses many models of interest. For example, one canonical model takes the form

$$J_i = \underset{g \in \{1, \ldots, G\}}{\arg\max} \left\{ Z_g' \Theta_i + \zeta_{ig} \right\}, \tag{4}$$

which is a version of random utility discrete choice models and prevalent in empirical work. As seen from this example, the unobserved heterogeneity in the selection equation can be multi-dimensional and I do not impose monotonicity of $J(\cdot)$ with respect to Θ_i.

Before proceeding, I point out that the model (2)-(3) is a special case of general triangular models (1). To verify the claim, rewrite the model as

$$\begin{aligned} Y_i &= m(X_i, \varepsilon_i) \\ X_i &= h(Z_1, \cdots, Z_G, \eta_i), \qquad \eta_i = (\Theta_i, \zeta_{i1}, \cdots, \zeta_{iG}) \end{aligned} \tag{5}$$

where $Y_i \equiv Y_{iJ_i}$, $X_i \equiv X_{J_i}$, $\varepsilon_i \equiv \varepsilon_{iJ_i}$, $h(Z_1, \cdots, Z_G, \eta_i) = \sum_{g=1}^{G} X_g \mathbb{1}\{J(Z_1, \cdots, Z_G, \eta_i) = g\}$ is viewed as some nonparametric function, and Z_g includes X_g as a subvector. I use this formulation to compare the identifying assumptions in this paper with those in the existing work.

In the model (2)-(3), the parameter of interest is the partial effect of X_g on Y_{ig}, e.g., the ceteris paribus effect of school quality on student's test scores. One such measure is the average structural function (ASF) (Blundell and Powell, 2003):

$$\mu(x) = \int m(x, e) f_\varepsilon(e) de.$$

We can also consider other measures such as the quantile structural function (Imbens and Newey, 2009) and the local average response (Altonji and Matzkin, 2005). In the sequel, I focus on the ASF. Other measures of group-level partial effects are treated in Appendix B.2.

In observational studies, a non-trivial challenge to identify $\mu(x)$ is that due to selection, the group-level variables of interest X_{J_i} may be correlated with the unobserved heterogeneity ν_i. This type of endogeneity arises due to dependence among unobserved heterogeneity in outcome and selection equations (c.f., Heckman, 1976; Imbens, 2007). In the school example, student's preference Θ_i in the choice equation is correlated with their work ethics ν_i in the outcome equation because highly motivated students have strong preference for high-quality schools.

In light of the model, selection bias manifests as

$$\mathbb{E}[Y_{ig} | X_g = x, J_i = g] = \int m(x, e) f_{\varepsilon | X J}(e | x, g) de \neq \int m(x, e) f_\varepsilon(e) de = \mu(x) \tag{6}$$

because $f_{\varepsilon|XJ} \neq f_{\varepsilon}$ in general. In words, the school-level distribution of student motivation $f_{\varepsilon|XJ}$ varies across schools due to the variation in school quality X_g. One way to solve this identification problem is to find a control function such that conditional on this variable, the conditional distribution of ε_{iJ_i} becomes invariant with respect to X_{J_i}. In the literature, various structures of outcome and/or selection equations have been exploited to construct a control function. For example, Altonji and Matzkin (2005) exploit exchangeability among group members, Das et al. (2003) use the threshold-crossing structure of the selection equation as well as additive separability of the error term in the outcome equation, and Imbens and Newey (2009) leverage strict monotonicity of the endogenous variable with respect to the unobserved heterogeneity.

In this paper, I propose a novel approach via conditional independence and completeness assumptions to construct a control variable. To describe my approach, I first discuss the conditional independence assumption.

Assumption 1. *(i)* $(X_g, Z_g, u_g)_{g=1}^G \perp\!\!\!\perp (W_i, \nu_i)|\Theta_i$, *(ii)* $(X_g, Z_g)_{g=1}^G \perp\!\!\!\perp (u_g)_{g=1}^G|\Theta_i$, *and (iii)* $(\zeta_{ig})_{g=1}^G$ *is independent of everything else conditional on* Θ_i.

The first part of Assumption 1 roughly states that school characteristics and student variables are independent *before* selection into groups occurs. In particular, it does not impose $(X_{J_i}, Z_{J_i}, u_{J_i}) \perp\!\!\!\perp (W_i, \nu_i)$, where the index by J_i denotes that they are observed *after* selection. To rationalize such independence condition, consider the following scenario: first, school features and student characteristics are drawn independently from some distributions, and then students determine which school to attend. This may be a reasonable model given that schools are established first and then households make decisions to relocate near their desired schools. Note that this independence requirement only needs to hold after conditioning on Θ_i, which allows for more general settings than complete independence of school and student characteristics. For instance, it accommodates some dependence through multiple-stage selection (e.g., first select into a metropolitan area, and then select into a school district/neighborhood) as long as the first-stage selection only depends on Θ_i. In a sense, this conditional independence assumption is essential for analysis of selection bias because without such independence assumption, endogeneity would be present even if there were no selection.

The part (ii) of Assumption 1 requires that the unobserved school characteristics u_g be indepen-

dent of other school features. This requirement means that a researcher has good measurements of school features that enter the education production function. Admittedly, this independence condition can be stringent. Yet, without such restriction, endogeneity would be present even if it were not for selection into groups, via dependence between X_g and u_g. Since I focus on the issue of selection bias, I maintain this assumption. The third part of the assumption imposes that $(\zeta_{i1}, \ldots, \zeta_{iG})$ is a purely idiosyncratic term, independent of all the other variables.

To compare Assumption 1 with identifying restrictions used in the literature, consider the formulation in (5). From how it is used in the proof of Theorem 1 below, Assumption 1 essentially translates to $Z \perp\!\!\!\perp (W, \varepsilon)|\eta$, whereas a standard assumption in the literature is $Z \perp\!\!\!\perp (\varepsilon, \eta)$. By subsuming the extra term W in ε as a subvector, we see that the conditional independence assumption in this paper is implied by the standard assumption. As endogeneity arises due to dependence between ε and η, the conditional independence of Z and ε given η is considerably weaker than the full independence of Z from (ε, η). In particular, this paper does not rely on Z being an instrument in the sense that $Z \perp\!\!\!\perp (\varepsilon, \eta)$ need not hold. This weaker condition suffices in part because I impose another restriction somewhere else, namely completeness on the proxy variable W. I will elaborate on this point later.

Going back to the model (2)-(3), one implication of the independence assumption is that the issue of endogeneity (i.e., across-school variation in $f_{\varepsilon|XJ}$, see (6)) arises from the variation in the school-level distribution of the preference variable Θ_i. To see this point, subsume X_g in Z_g as a subvector and write $Z_{-g} = (Z_1, \ldots, Z_{g-1}, Z_{g+1}, \ldots, Z_G)$ for the collection of Z's excluding Z_g. Then,

$$\begin{aligned} f_{\varepsilon|JZ}(\cdot|g,z) &= \int f_{\varepsilon|JZ\Theta Z_{-g}}(\cdot|g,z,\theta,z_{-g}) f_{\Theta Z_{-g}|JZ}(\theta,z_{-g}|g,z) d(\theta,z_{-g}) \\ &= \int f_{\varepsilon|Z\Theta Z_{-g}}(\cdot|z,\theta,z_{-g}) f_{\Theta Z_{-g}|JZ}(\theta,z_{-g}|g,z) d(\theta,z_{-g}) \\ &= \int f_{\varepsilon|\Theta}(\cdot|\theta) f_{\Theta Z_{-g}|JZ}(\theta,z_{-g}|g,z) d(\theta,z_{-g}) \\ &= \int f_{\varepsilon|\Theta}(\cdot|\theta) f_{\Theta|JZ}(\theta|g,z) d\theta \end{aligned} \tag{7}$$

where the second equality follows from $J_i = J(Z_1, \ldots, Z_G, \Theta_i, \zeta_{i1}, \ldots, \zeta_{iG})$ and independence of $(\zeta_{i1}, \ldots, \zeta_{iG})$ from other variables, and the third equality uses $(Z_1, \ldots, Z_G) \perp\!\!\!\perp (\nu_i, u_g)|\Theta_i$. From

the last line in the display, we see that the variation in the school-level distribution of motivation comes from the variation in the school-level distribution of the preference variable. This observation suggests that accounting for the across-school variation in $f_{\Theta|JZ}(\cdot|g,z)$ would address the endogeneity issue. Put it differently, we have $\varepsilon_{iJ_i} \perp\!\!\!\perp X_{J_i}|f_{\Theta|ZJ}(\cdot|J_i,Z_{J_i})$.[2] Of course, this insight is not directly useful as we do not observe Θ_i. Yet, we can infer some information about this unobserved heterogeneity from the observed individual-level covariates W_i.

Now, I indicate that the school-level distribution of student covariates $f_{W|JZ}(\cdot|J_i,Z_{J_i})$ plays the role of a control function. Doing the same calculation as above, we obtain

$$f_{W|JZ}(\cdot|g,z) = \int f_{W|\Theta}(\cdot|\theta) f_{\Theta|JZ}(\theta|g,z)d\theta,$$

which implies that the school-level distribution of student characteristics varies as $f_{\Theta|JZ}(\cdot|g,z)$ changes. Then, if the above mapping from $f_{\Theta|JZ}(\cdot|g,z)$ to $f_{W|JZ}(\cdot|g,z)$ is one-to-one, conditioning on $f_{W|JZ}(\cdot|g,z)$ holds constant $f_{\Theta|JZ}(\cdot|g,z)$. That is, if the above mapping is indeed injective, there exists some function $\mathbf{\Psi}^\dagger$ satisfying $\mathbf{\Psi}^\dagger(f_{W|JZ}(\cdot|g,z))(\theta) = f_{\Theta|JZ}(\theta|g,z)$. And from (7), we obtain

$$f_{\varepsilon|JZ}(e|g,z) = \int f_{\varepsilon|\Theta}(e|\theta)\mathbf{\Psi}^\dagger(f_{W|JZ}(\cdot|g,z))(\theta)d\theta.$$

Therefore, by holding constant the school-level distribution of W_i, we can hold constant $f_{\varepsilon|JZ}(\cdot|g,z)$. Now, we can identify ceteris paribus effects of X_g on Y_{ig} by computing $\mathbb{E}[Y_{iJ_i}|X_{J_i}=x, f_{W|JZ}(\cdot|J_i,Z_{J_i})]$ and varying x while holding constant $f_{W|JZ}$.

Here, the crucial hypothesis of the identification argument is that the above integral transform is one-to-one. To discuss the plausibility of this condition, I first provide a formal statement of this injectivity requirement.

Assumption 2. *The following mapping, defined on the set of bounded and integrable functions,*

$$(\mathbf{\Psi}h)(w) = \int h(\theta) f_{W|\Theta}(w|\theta)d\theta$$

is injective.

[2] Note that $f_{\Theta|JZ}(\cdot|J_i,Z_{J_i})$ is a random element whose randomness comes from that of J_i and Z_{J_i}.

This assumption, usually referred to as bounded completeness, ensures that the integral transform is an injective mapping. This condition can be viewed as a generalization of the instrument relevance condition in linear IV models (Newey and Powell, 2003) and has been widely used in the recent econometric literature on nonparametric identification (see Section 1.1). Intuitively, injectivity requires the function $f_{\Theta|W}(\cdot|w)$ to sufficiently vary in the conditioning value w. For sufficient conditions for different types of completeness, see Andrews (2017); D'Haultfoeuille (2011); Hu et al. (2017); Hu and Shiu (2018); Mattner (1993) and references therein.

To understand this assumption, suppose for a moment that a researcher observed student's preference Θ_i. Under this hypothetical situation, we could solve the endogeneity problem by conditioning on Θ_i because we have $\varepsilon_{iJ_i} \perp\!\!\!\perp X_{J_i}|J_i, \Theta_i$. This identification strategy relies on a so called selection on observables assumption: conditional on the observables Θ_i, selection becomes exogenous. However, in many cases, observing Θ_i seems implausible as preference is difficult to measure. In this paper, I provide an alternative identifying restriction that replaces the selection on observables assumption. In particular, instead of full measurement of Θ_i, I only require partial information on Θ_i through observed covariates W_i. For instance, we might posit that Θ_i is a function of W_i and another random element ω_i. Even when the selection on observables condition fails (i.e., $\omega_i \not\perp\!\!\!\perp \varepsilon_{ig}|W_i$), Assumption 2 can be satisfied provided that W_i has non-trivial influence on Θ_i via the functional relationship.

An alternative, related way to interpret this completeness assumption is to view observed student characteristics W_i as proxies or noisy measurements for the unobserved preference Θ_i (see e.g., Schennach, 2016). In recent development of identification in nonparametric errors-in-variables models, injectivity of certain integral transformations plays an important role, where Θ_i corresponds to unobserved correctly measured variables and W_i to noisy measurements. From the viewpoint of W_i as coarse measurements of Θ_i, the richer array of student characteristics in a dataset, the more plausible Assumption 2 becomes. Thus, this paper's identification strategy may be more fruitful when a researcher analyzes datasets with detailed information on individual characteristics, such as large survey datasets.[3]

[3]A notable distinction from the measurement error literature is the literautre considers unobserved heterogeneity distribution (i.e., the distribution of correctly measured variables) to be one of the main objects of interest, which is a nuisance parameter in this paper. Due to this difference in target estimands, whereas methods on measurement errors usually require at least two measurements of the unobserved variable, I only need one "proxy" for Θ_i.

In the current setting, if a researcher is willing to impose more structures on the selection equation (3), we can find some primitive conditions for Assumption 2. For instance, Altonji and Mansfield (2018) posit that the group choice is determined by a random utility discrete choice model like (4) and in particular the preference variable Θ_i takes the form

$$\Theta_i = \Gamma W_i + \omega_i$$

where Γ is some conformable non-stochastic matrix and ω_i is an unobserved random vector. A classical result states that completeness holds if ω_i given W_i has a mean-zero normal distribution with a fixed non-singular covariance matrix and Γ is of full column rank. More generally, (i) $\omega_i \perp\!\!\!\perp W_i$, (ii) the characteristic function of ω_i is non-zero everywhere, and (iii) $\text{supp}(\Gamma W_i) = \mathbb{R}^d$ with $d = \dim(\Theta_i)$ are sufficient conditions for bounded completeness (Mattner, 1993). Another possibility of identifying restrictions is to model unobserved heterogeneity as a discrete variable. With discrete unobserved heterogeneity, we can interpret Θ_i to represent an agent type. This strategy has been employed in a number of empirical studies. For instance, recent papers of Abowd et al. (2019) and Bonhomme et al. (2019) study the consequence of worker-firm matching on labor market earnings, and they model worker and firm heterogeneity as discrete types. Under this additional structure, the completeness assumption reduces to full column rank of the matrix $[\Pr(W_i = w_\ell | \Theta_i = \theta_k)]_{\ell,k}$ where $\text{supp}(\Theta_i) = \{\theta_1, \ldots, \theta_K\}$ and $\text{supp}(W_i) = \{w_1, \ldots, w_L\}$.[4]

Now I discuss Assumption 2 in relation to the assumptions used in the literature.[5] Among the existing work discussed in Section 1.1, common restrictions are (i) some form of monotonicity on $h(\cdot)$ and (ii) $Z \perp\!\!\!\perp (\varepsilon, \eta)$. Relative to them, Assumption 2 is a new restriction this paper imposes. In particular, it requires that a researcher observe a proxy variable W for the unobserved heterogeneity causing endogeneity. There is some trade off between imposing Assumptions 1-2 and the standard conditions. The approach in this paper does not require monotonicity of $h(\cdot)$ and the full independence $Z \perp\!\!\!\perp (\varepsilon, \eta)$. On the other hand, it requires to specify what Θ represents in the selection equation and to observe a good proxy for the unobserved heterogeneity. One may argue that this is a reasonable trade off because, although availability of a good proxy is a crucial

[4] It is without loss of generality to assume discrete distribution of W_i because we can use binning to transform a continuous random variable to discrete one.

[5] For this paragraph, refer to a general nonseparable model (1) and (5) for the notation.

and potentially stringent assumption, so is availability of valid instruments. Since the two sets of conditions are plausible in different circumstances, I view the two approaches as complementary. Researchers may use one method over the other depending on the type of data they have.

To present the identification result, I additionally impose the following conditions.

Assumption 3. *With respect to some σ-finite product measure, the distribution of $W_i \times \varepsilon_{ig} \times Z_1 \times \cdots \times Z_G \times \Theta_i$ is absolutely continuous, and the conditional density of Θ_i given Z_{J_i} is bounded.*

Assumption 4. *The joint distribution of $(W_i, X_{J_i}, Z_{J_i}, \Theta_i, \varepsilon_{iJ_i})$ is identical across i, and the distribution of $(Y_{iJ_i}, W_i, X_{J_i}, Z_{J_i})$ is identifiable from the data. The conditional density of W_i given Z_{J_i}, denoted by $f_{W|Z_J}$, is continuous.*

Assumption 5. *For some non-empty set $\mathcal{X} \subset \text{supp}(X_g)$, the support of $f_{W|Z_J}(\cdot|Z_{J_i})$ conditional on $X_{J_i} = x$ equals the unconditional one for $x \in \mathcal{X}$. That is, $\text{supp}(f_{W|Z_J}(\cdot|Z_{J_i})|X_{J_i} = x) = \text{supp}(f_{W|Z_J}(\cdot|Z_{J_i}))$ for $x \in \mathcal{X}$.*

With these assumptions, I now state the main identification result of this paper, which formalizes the heuristic identification argument above. The proof is in the appendix.

Theorem 1. *If Assumptions 1-4 hold, then the conditional distribution of ε_{iJ_i} given $X_{J_i} = x, f_{W|Z_J}(\cdot|Z_{J_i})$ is invariant across x. In addition, suppose $\mathbb{E}[|m(x, \varepsilon_{ig})|] < \infty$ for all $x \in \mathcal{X}$ and Assumption 5 holds. Then $\mu(x)$ is identified for $x \in \mathcal{X}$.*

The theorem states that the school-level distribution of student characteristics $f_{W|Z_J}(\cdot|Z_{J_i})$ plays the role of a control function, and furthermore, under the support condition Assumption 5, the ASF is identified. One important distinction from the heuristic argument above is that I use $f_{W|Z_J}(\cdot|Z_{J_i})$ as a control function rather than $f_{W|JZ}(\cdot|J_i, Z_{J_i})$. The latter object denotes the group-level distribution for a specific school $J_i = g$ with some fixed g, which might not be identified from the data if datasets contain only a moderate number of students for each school. This point is reflected in the proof of Theorem 1.

Although I focus on the model of selection into groups, Theorem 1 can be adapted to the general nonseparable model (1). The key assumptions are $Z \perp\!\!\!\perp (W, \varepsilon)|\eta$ and bounded completeness of the distribution of (W, ε) with respect to W. As discussed above, this result complements the existing results by providing an alternative set of identifying conditions. In particular, while most of the

existing methods rely on IV, this paper does not and instead employs a proxy variable for the unobserved heterogeneity influencing the selection process. Thus, Theorem 1 opens a new avenue for identification when good candidates of IV are not available.

Besides non-reliance on IV, there is another distinct aspect of Theorem 1: the control variable is function-valued, living on an infinite-dimensional space. Since the identification argument involves computing a nonparametric conditional expectation given the infinite-dimensional random element, it is not immediately clear how to construct an estimator for $\mu(x)$. In the next section, I leverage results from nonparametric functional data literature to propose an estimator.

Now, I discuss Assumptions 3-5. Assumption 3 imposes mild regularity conditions on the joint distribution of random elements. Absolute continuity with respect to some σ-finite measure means that there exists a density for the distribution of these variables. Assumption 4 concerns identical distribution and identifiability of the joint distribution of observed variables. A sufficient condition for identifiability, in combination with Assumption 1, is that group-level variables (X_g, Z_g, u_g) are i.i.d. draws across g, individual-level variables (W_i, ν_i, Θ_i) are i.i.d. draws across i, the idiosyncratic shock ζ_{ig} is i.i.d. across (i, g), and $(X_g, Z_g, u_g) \perp\!\!\!\perp (W_i, \nu_i, \Theta_i)$. This random sampling assumption is an easy-to-interpret sufficient condition, but we can accommodate a wide class of data generating processes. For instance, we can allow for school features (X_g, Z_g, u_g) to exhibit spatial dependence via mixing conditions or cluster structures. Also, Assumption 4 imposes some regularity on the conditional density of W_i given Z_{J_i} to ensure nice behavior of conditional distributions given the random element.

Assumption 5 requires that the support of the random element $f_{W|Z_J}(\cdot|Z_{J_i})$ be invariant after conditioning on X_{J_i}. Imbens and Newey (2009) refer to this restriction as a common support condition, and although essential, it can be a stringent assumption. It generally requires a large support of Z_g conditional on X_g. For illustration, suppose X_g is a subvector of Z_g and write $Z_g = (X_g', Z_{2g}')'$. Further assume that the group-level distribution of W_i is affected via some index function $\phi(Z_g)$ i.e., $F_{W|Z_J}(\cdot|z) = H(\cdot|\phi(z))$ for some function H. Then in order to satisfy Assumption 5, we need Z_{2g} to have sufficient variation conditional on X_g such that the range of ϕ on $\text{supp}(Z_g)$ equals the range of $\phi(x, \cdot)$ on $\text{supp}(Z_{2g}|X_g = x)$. Also, note that in this example, Z_g affects the conditional distribution only through a scalar function ϕ. For the common support condition to hold, the way Z_g affects the conditional distribution needs some restrictions.

If the support condition is not satisfied, we can still identify some versions of the ASF as done in the literature. One possibility is to focus on a local version of the ASF:

$$\int m(x,e) f_{\varepsilon|V}(e|v)de \qquad x \in \tilde{\mathcal{X}}$$

where I condition on the control variable $V_i = f_{W|Z_J}(\cdot|Z_{J_i})$ and $\tilde{\mathcal{X}} = \text{supp}(X_{J_i}|V_i = v)$. This type of parameter has been considered in the literature (e.g., Fernández-Val et al., 2018) as a local measure of average effects. For instance, if ε_{ig} denotes unobserved academic motivation, holding constant $f_{W|Z_J}(\cdot|Z_{J_i})$ means that we fix the distribution of motivation at some value. Although this distribution of motivation is not identified, this version of the ASF represents average effects for some subpopulation of schools. Another approach for identification without the common support condition is to develop bounds on the ASF as in Imbens and Newey (2009) if the outcome of interest Y_{ig} has known upper and lower bounds. Also, Chernozhukov et al. (2019) consider semiparametric triangular models where the support condition can be relaxed using structures of the outcome and choice equations. A similar approach may be possible, although developing a formal theory is beyond the scope of this paper.

3 Estimation

In this section, I propose a kernel-based estimator for the ASF identified in the previous section and provide a set of conditions for consistency.

We observe $(Y_i, W_i, X_i, Z_i) \equiv (Y_{J_i}, W_i, X_{J_i}, Z_{J_i})$ and denote $F_{W|Z}(\cdot|Z_i)$ by V_i, which is a random function. We do not directly observe V_i, but we can estimate it from data. With estimated $\hat{F}_{W|Z}(\cdot|\cdot)$, we take $\hat{V}_i = \hat{F}_{W|Z}(\cdot|Z_i)$. Any nonparametric method can be used to estimate $F_{W|Z}$ provided that the estimator satisfies a mild rate restriction for uniform convergence on a slowly expanding set. Suppose that an estimator of the control variable is given. Then, we can construct an estimator for the ASF $\mu(x_0)$ by

$$\hat{\mu}(x_0) = \frac{1}{N}\sum_{i=1}^{N} \hat{m}(x_0, \hat{V}_i)\mathbb{1}\{|Z_i| \le \tau_N\}, \qquad \hat{m}(x,v) = \frac{\sum_{j=1}^{N} Y_j K([X_j - x]/h_N) L(\|\hat{V}_j - v\|/h_N)}{\sum_{j=1}^{N} K([X_j - x]/h_N) L(\|\hat{V}_j - v\|/h_N)}$$

where x_0 is some fixed value chosen by a researcher, K and L are kernel functions, h_N is a bandwidth

sequence, $\|\cdot\|$ denotes the supremum norm on function spaces, and τ_N is a sequence of positive numbers tending to infinity. In the sequel, all asymptotic statements are understood as $N \to \infty$.

This estimator $\hat{\mu}(x)$ is a partial means estimator studied by Newey (1994), which averages the estimated conditional expectation over the control variable $\hat{V}_i$. As we need to evaluate the conditional expectation for a range of values, I introduce the trimming function $\mathbb{1}\{|Z_i| \leq \tau_N\}$ to achieve some type of uniformity. The estimator of the conditional expectation, $\hat{m}(x, v)$, is the usual Nadaraya-Watson estimator for conditional expectations with one important difference: the conditioning variable V_i is function-valued. This setting of function-valued covariates naturally arises in many fields of natural and social sciences, and there has been some work on nonparametric estimation with function-valued covariates. For a textbook treatment, see Ferraty and Vieu (2006).

One complication in this setting is that V_i is infinite-dimensional and from our knowledge on the finite-dimensional case, we suspect that the curse of dimensionality adversely affects properties of the estimator $\hat{m}(x_0, v)$. In general, estimators like $\hat{m}(x, v)$ might converge at logarithmic rates due to sparseness of data points in an infinite-dimensional space (see Chapter 13 of Ferraty and Vieu, 2006, for discussion). Fortunately, we have enough structures in this problem that prevent the issue of logarithmic convergence rates. In particular, the variation in the object $V_i = F_{W|Z}(\cdot|Z_i)$ comes from Z_i, which is a finite-dimensional random element, and using continuity, we see that V_i concentrates around a given point v like finite-dimensional variables do (see Lemma 1 for a formal result).

Here I provide a set of conditions ensuring consistency of $\hat{\mu}(x_0)$.

Assumption 6. *The covariate X_i is a subvector of Z_i. The observation $\{Y_i, W_i, Z_i\}_{i=1}^N$ is a random sample. The random vector Z_i has a bounded and continuous Lebesgue density f_Z, which is positive on $\mathbb{R}^{d_z}$. The conditional distribution function $F_{W|Z}$ is continuously differentiable with respect to z and the derivative is bounded and $\|\partial F_{W|Z}(\cdot|z)/\partial z\| > 0$ for all z. The conditional expectation $m(x, v) = \mathbb{E}[Y_i|X_i = x, V_i = v]$ is uniformly continuous on $x_0 \times \text{supp}(V_i)$, for some $s \geq 2$, $\mathbb{E}[|Y_i|^s] < \infty$, and $\mathbb{E}[|Y_i|^2|X_i = x, V_i = v]$ is bounded on $\{x : |x - x_0| \leq \delta\} \times \text{supp}(V_i)$ with some $\delta > 0$.*

This assumption specifies restrictions on the data generating process, most of which are natural extensions from the finite-dimensional covariate case. The condition $X_i \subset Z_i$ is without essential loss of generality as including X_i in Z_i does not change the identification argument. The restriction

$\text{supp}(Z_i) = \mathbb{R}^{d_z}$ is partially motivated by Assumption 5, which generally requires a large support of Z_i. Although including discrete random variables in Z_g is a relatively straightforward extension, I omit it to simplify presentation. Difficulties may arise in accommodating compactly supported continuous random variables since special care needs to be taken for the boundaries of the support. The next condition specifies properties of kernel functions.

Assumption 7. *The kernel functions K and L are bounded, non-negative, and compactly supported. Additionally, for some constants $c_L > 0$, $L(u) \geq c_L \mathbb{1}\{|u| \leq c_L\}$ and there exists a function L^* such that $|L(v) - L(u)| \leq L^*(u)|v - u|$ for $|v - u| \leq \delta$ with some $\delta > 0$ and L^* is bounded and compactly supported and satisfies $L^*(u) \geq c_L \mathbb{1}\{|u| \leq c_L\}$.*

Most of the existing work on nonparametric estimation with function-valued variables assume simple kernel functions. Following the practice, I impose non-negativity and compact support. In addition, to handle estimated control variables and achieve uniformity, I assume Lipschitz continuity of L.

With these assumptions, I show consistency of $\hat{\mu}(x_0)$ to $\mu(x_0)$. The proof is in the appendix.

Theorem 2. *Assume Assumptions 1-7 hold with $x_0 \in \mathcal{X}$ and $\max_{1 \leq i \leq N} \|\hat{V}_i - V_i\| \mathbb{1}\{|Z_i| \leq \tau_N\} = o_{\mathbb{P}}(h_N)$. Let $q_N = \inf_{|z| \leq \tau_N} f_Z(z)$ and if $\tau_N \to \infty$, $\log \tau_N (\log N)^{-1} = O(1)$, $h_N q_N^{-1} = o(1)$, and $N^{1-1/s} h_N^{d_z} q_N (\log N)^{-2} \to \infty$, then $\hat{\mu}(x_0) \to_{\mathbb{P}} \mu(x_0)$.*

The hypothesis $\max_{1 \leq i \leq N} \|\hat{V}_i - V_i\| \mathbb{1}\{|Z_i| \leq \tau_N\} = o_{\mathbb{P}}(h_N)$ is a mild requirement on the first-stage estimator $\hat{V}_i = \hat{F}_{W|Z}(\cdot|Z_i)$. For concreteness, consider the Nadaraya-Watson estimator for $F_{W|Z}$. For nonparametric kernel estimators, Hansen (2008) and Lemma B.1 of Cattaneo et al. (2013) among others present results on uniform convergence rates. With slight modifications in their proofs, we can show that the estimation error, $\max_{1 \leq i \leq N} \|\hat{V}_i - V_i\| \mathbb{1}\{|Z_i| \leq \tau_N\}$, has a polynomial convergence rate provided that τ_N tends to infinity at an appropriate rate. See Appendix B.3 for details.

4 Conclusion

This paper presents new identification and estimation results for group-level causal effects in a setting where individuals select into groups and selection is in part based on individual unobserved heterogeneity. It specifies a triangular model of the outcome and choice equations and imposes

conditional independence and completeness assumptions, which leads to a novel construction of a control variable. Building on the identification result, I propose a kernel-based estimator for the average structural function and prove its consistency.

5 Bibliography

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

Abowd, J. M., K. L. McKinney, and I. M. Schmutte (2019): "Modeling Endogenous Mobility in Earnings Determination," *Journal of Business & Economic Statistics*, 37, 405–418.

Altonji, J. G. and R. K. Mansfield (2018): "Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects," *American Economic Review*, 108, 2902–2946.

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.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Blundell, R. W. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnsovsky, Cambridge University Press, vol. 2, chap. 8, 321–357.

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

Cattaneo, M. D., R. K. Crump, and M. Jansson (2013): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1256.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2019): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," Forthcoming in Quantitative Economics.

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.

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

Chesher, A. and A. Rosen (2019): "Generalized Instrumental Variable Models Methods and Applications," Working Paper.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade," *Econometrica*, 84, 809–833.

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

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

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

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.

Durlauf, S. N. (2004): "Neighborhood Effects," in *Handbook of Regional and Urban Economics*, ed. by J. V. Henderson and J.-F. Thisse, Elsevier, vol. 4, 2173–2242.

Durlauf, S. N. and Y. M. Ioannides (2010): "Social Interactions," *Annual Review of Economics*, 2, 451–478.

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.

Graham, B. S. (2018): "Identifying and Estimating Neighborhood Effects," *Journal of Economic Literature*, 56, 450–500.

Graham, B. S., G. W. Imbens, and G. Ridder (2018): "Identification and Efficiency Bounds for the Average Match Function Under Conditionally Exogenous Matching," Forthcoming in Journal of Business & Economic Statistics.

Hall, P. and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 1–27.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726–748.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35.

Heckman, J. J. and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734.

Hoderlein, S., L. Nesheim, and A. Simoni (2017): "Semiparametric Estimation of Random Coefficients in Structural Economic Models," *Econometric Theory*, 33, 1265–1305.

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.

Hu, Y. and J.-L. Shiu (2018): "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness," *Econometric Theory*, 34, 659–693.

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.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Mattner, L. (1993): "Some Incomplete But Boundedly Complete Location Families," *Annals of Statistics*, 21, 2158–2162.

Matzkin, R. L. (2007): "Nonparametric identification," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5307 – 5368.

——— (2016): "On Independence Conditions in Nonseparable Models: Observable and Unobservable Instruments," *Journal of Econometrics*, 191, 302–311.

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

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

Pollard, D. (1984): *Convergence of Stochastic Processes*, New York, NY: Springer.

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

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.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York, NY: Springer.

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

A Proofs

In the sequel, I use the Greek letters λ, π, ρ to denote generic σ-finite measures.

A.1 Proof of Theorem 1

First note that by Assumption 1 (i) and (ii), $(W_i, \varepsilon_{ig}) \perp\!\!\!\perp (X_g, Z_g)_{g=1}^G | \Theta_i$.

Denote the conditional density of ε_i given Z_{J_i} by $f_{\varepsilon|Z_J}$. We have

$$\begin{aligned} f_{\varepsilon|Z_J}(\cdot|z) &= \int f_{\varepsilon|Z_J\Theta JZ_{-J}}(\cdot|z,\theta,g,\tilde{z}) f_{\Theta JZ_{-J}|Z_J}(\theta,g,\tilde{z}|z) d(\lambda\times\pi\times\rho)(\theta,g,\tilde{z}) \\ &= \int f_{\varepsilon|Z_J\Theta Z_{-J}}(\cdot|z,\theta,\tilde{z}) f_{\Theta JZ_{-J}|Z_J}(\theta,g,\tilde{z}|z) d(\lambda\times\pi\times\rho)(\theta,g,\tilde{z}) \\ &= \int f_{\varepsilon|\Theta}(\cdot|\theta) f_{\Theta JZ_{-J}|Z_J}(\theta,g,\tilde{z}|z) d(\lambda\times\pi\times\rho)(\theta,g,\tilde{z}) \\ &= \int f_{\varepsilon|\Theta}(\cdot|\theta) f_{\Theta|Z_J}(\theta|z) d\lambda(\theta) \end{aligned}$$

where the definition of conditional densities (Assumption 3 guarantees the existence of the conditional densities), the second equality uses $J_i = J(Z_1, \ldots, Z_G, \Theta_i, \zeta_{i1}, \ldots, \zeta_{iG})$ and ζ_{ig}'s are independent of everything else given Θ_i, the third equality uses $\varepsilon_{ig} \perp\!\!\!\perp (Z_g)_{g=1}^G | \Theta_i$, and the fourth equality follows from integrating over (J, Z_{-J}). Using the same argument,

$$f_{W|Z_J}(\cdot|z) = \int f_{W|\Theta}(\cdot|\theta) f_{\Theta|Z_J}(\theta|z) d\lambda(\theta).$$

By Assumption 2, there exists a function $\mathbf{\Psi}^\dagger$ such that

$$\mathbf{\Psi}^\dagger\big(f_{W|Z_J}(\cdot|z)\big)(\theta) = f_{\Theta|Z_J}(\theta|z).$$

Then, conditioning on $f_{W|Z_J}(\cdot|Z_{J_i})$, the random element $f_{\varepsilon|Z_J}(\cdot|Z_{J_i})$ is non-stochastic. Letting

$V_i = f_{W|Z}(\cdot|Z_{J_i})$, we have

$$\begin{aligned}
\mathbb{E}[Y_{iJ_i}|X_{J_i} = x, V_i] &= \mathbb{E}[m(x, \varepsilon_{iJ_i})|X_{J_i} = x, V_i] \\
&= \mathbb{E}[\mathbb{E}[m(x, \varepsilon_{iJ_i})|X_{J_i} = x, Z_{J_i}]|X_{J_i} = x, V_i] \\
&= \mathbb{E}[\mathbb{E}[m(x, \varepsilon_{iJ_i})|Z_{J_i}]|X_{J_i} = x, V_i] \\
&= \mathbb{E}\Big[\int m(x, e) f_{\varepsilon|Z_J}(e|Z_{J_i}) d\rho(e) \Big| X_{J_i} = x, V_i\Big] \\
&= \int m(x, e) \mathbb{E}\big[f_{\varepsilon|Z_J}(e|Z_{J_i}) \big| X_{J_i} = x, V_i\big] d\rho(e) \\
&= \int m(x, e) \mathbb{E}\big[f_{\varepsilon|Z_J}(e|Z_{J_i}) \big| V_i\big] d\rho(e)
\end{aligned}$$

where the second equality follows from V_i is a function of Z_{J_i}, the third equality follows from $\varepsilon_{ig} \perp\!\!\!\perp X_g|\Theta_i, Z_g$, the second-to-last equality uses the Fubini theorem to interchange the order of integration, and the last equality uses $f_{\varepsilon|Z_J}(\cdot|Z_{J_i})$ is non-stochastic conditional on $V_i = f_{W|Z_J}(\cdot|Z_{J_i})$. Finally, letting σ be the distribution of $f_{W|Z_J}(\cdot|Z_{J_i})$, which is identifiable from the data,

$$\begin{aligned}
\int \mathbb{E}[Y_{iJ_i}|X_{J_i} = x, V_i = v] d\sigma(v) &= \int\int m(x, e) \mathbb{E}\big[f_{\varepsilon|Z_J}(e|Z_{J_i}) \big| V_i = v\big] d\rho(e) d\sigma(v) \\
&= \int\int m(x, e) \mathbb{E}\big[f_{\varepsilon|Z_J}(e|Z_{J_i}) \big| V_i = v\big] d\sigma(v) d\rho(e) \\
&= \int m(x, e) \mathbb{E}\big[f_{\varepsilon|Z_J}(e|Z_{J_i})\big] d\rho(e) \\
&= \int m(x, e) f_\varepsilon(e) d\rho(e)
\end{aligned}$$

where the second equality uses the Fubini theorem and the third equality requires Assumption 5.

A.2 Proof of Theorem 2

Let $\pi_N = (N)^{1/s} \log N$ and define

$$\breve{\mu}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{N} Y_{jN} K([X_j - x]/h_N) L(\|\hat{V}_j - \hat{V}_i\|/h_N)}{\sum_{j=1}^{N} K([X_j - x]/h_N) L(\|\hat{V}_j - \hat{V}_i\|/h_N)} \mathbb{1}\{|Z_i| \le \tau_N\}, \qquad Y_{jN} = Y_j \mathbb{1}\{|Y_j| \le \pi_N\}$$

where $\breve{\mu}(x)$ is different from $\hat{\mu}(x)$ in replacing Y_j's with the truncated version $Y_{jN} = Y_j \mathbb{1}\{|Y_j| \le \pi_N\}$. Then,

$$\mathbb{P}\left[\breve{\mu}(x_0) \ne \hat{\mu}(x_0)\right] \le \mathbb{P}\left[|Y_j| > \pi_N \text{ for some } j\right] \le N \pi_N^{-s} \mathbb{E}[|Y_j|^s] = o(1)$$

and it suffices to show $\breve{\mu}(x_0) \to_{\mathbb{P}} \mu(x_0)$. Define

$$\varphi(v) = \mathbb{E}\left[K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|V_i - v\|}{h_N}\right)\right], \qquad \varphi^*(v) = \mathbb{E}\left[K\left(\frac{X_i - x_0}{h_N}\right) L^*\left(\frac{\|V_i - v\|}{h_N}\right)\right]$$

$$\hat{g}(v) = \frac{1}{N\varphi(v)} \sum_{i=1}^N Y_{iN} K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|\hat{V}_i - v\|}{h_N}\right), \quad \tilde{g}(v) = \frac{1}{N\varphi(v)} \sum_{i=1}^N Y_{iN} K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|V_i - v\|}{h_N}\right)$$

$$\hat{f}(v) = \frac{1}{N\varphi(v)} \sum_{i=1}^N K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|\hat{V}_i - v\|}{h_N}\right), \quad \tilde{f}(v) = \frac{1}{N\varphi(v)} \sum_{i=1}^N K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|V_i - v\|}{h_N}\right).$$

Then, $\breve{\mu}(x_0) = \sum_{i=1}^N \mathbb{1}\{|Z_i \le \tau_N\} \hat{g}(\hat{V}_i)/\hat{f}(\hat{V}_i) N$ and

$$\begin{aligned}\breve{\mu}(x_0) = \frac{1}{N}\sum_{i=1}^N \tilde{g}(V_i)\mathbb{1}\{|Z_i| \le \tau_N\} + \frac{1}{N}\sum_{i=1}^N \tilde{g}(V_i)\left[\hat{f}(\hat{V}_i)^{-1} - 1\right]\mathbb{1}\{|Z_i| \le \tau_N\} \\ + \frac{1}{N}\sum_{i=1}^N [\hat{g}(\hat{V}_i) - \tilde{g}(V_i)]\hat{f}(\hat{V}_i)^{-1}\mathbb{1}\{|Z_i| \le \tau_N\}.\end{aligned}$$

Below I show that

$$\max_{1\le i\le N} |\tilde{g}(V_i) - m(x_0, V_i)|\mathbb{1}\{|Z_i| \le \tau_N\} = o_{\mathbb{P}}(1), \quad \max_{1\le i\le N} |\tilde{f}(V_i) - 1|\mathbb{1}\{|Z_i| \le \tau_N\} = o_{\mathbb{P}}(1) \tag{8}$$

$$\max_{1\le i\le N} |\hat{g}(\hat{V}_i) - \tilde{g}(V_i)|\mathbb{1}\{|Z_i| \le \tau_N\} = o_{\mathbb{P}}(1), \quad \max_{1\le i\le N} |\hat{f}(\hat{V}_i) - \tilde{f}(V_i)|\mathbb{1}\{|Z_i| \le \tau_N\} = o_{\mathbb{P}}(1) \tag{9}$$

from which $\breve{\mu}(x_0) \to_{\mathbb{P}} \mu(x_0)$ follows.

For the first statement of (8), it suffices to show $\sup_{v\in A_N} |\tilde{g}(v) - m(x_0, v)| = o_{\mathbb{P}}(1)$ where $A_N = \{v : \exists z \text{ s.t. } v = F_{W|Z}(\cdot|z) \;\&\; |z| \le \tau_N\}$. Using uniform continuity of $m(x, v)$ and compact support of K and L, $|\mathbb{E}[\tilde{g}(v)] - m(x_0, v)| = o(1)$ uniformly in v. Thus, it remains to show $\sup_{v\in A_N} |\tilde{g}(v) - \mathbb{E}[\tilde{g}(v)]| = o_{\mathbb{P}}(1)$, for which I build on the discretization argument in Cattaneo et al. (2013).

Pick points $\{v_1^*, \cdots, v_M^*\} \subset A_N$ for which $\min_{1\le m\le M} \|v - v_m^*\| \le h_N^2$ for all $v \in A_N$ with $M = O(\tau_N^{d_z} h_N^{-2d_z})$. Define $\tilde{g}^*(v) = \sum_{i=1}^N Y_{iN} K([X_i - x_0]/h_N) L^*(\|V_i - v\|/h_N)/N\varphi(v)$ and by Lemma 2 and (10),

$$\begin{aligned}\sup_{v\in A_N} |\tilde{g}(v) - \mathbb{E}[\tilde{g}(v)]| \le C \max_{1\le m\le M} |\tilde{g}(v_m^*) - \mathbb{E}[\tilde{g}(v_m^*)]| + Ch_N q_N^{-1} \max_{1\le m\le M} |\tilde{g}^*(v_m^*) - \mathbb{E}[\tilde{g}^*(v_m^*)]| \\ + Ch_N q_N^{-1}\Big(\max_{1\le m\le M} \mathbb{E}[\tilde{g}(v_m^*)] + \max_{1\le m\le M} \mathbb{E}[\tilde{g}^*(v_m^*)]\Big)\end{aligned}$$

where $q_N = \inf_{|z|\le\tau_N} f_Z(z)$. Then arguing as in the proof of Lemma B-1 of Cattaneo et al. (2013),

it remains to verify that Bernstein's inequalities apply. Note that

$$\left| Y_{iN} K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|V_i - v\|}{h_N}\right) \right| \leq C\pi_N$$

and

$$\mathbb{E}\left[\left| Y_{iN} K\left(\frac{X_i - x_0}{h_N}\right) L\left(\frac{\|V_i - v\|}{h_N}\right) \right|^2 \right] \leq \varphi(v) C \sup_{|x-x_0|\leq\delta, \|u-v\|\leq\delta} \mathbb{E}[|Y_i|^2 | X_i = x, V_i = u].$$

Then, Bernstein's inequalities imply

$$\mathbb{P}\left[|\tilde{g}(v) - \mathbb{E}[\tilde{g}(v)]| > \delta\right] \leq 2\exp\left(-\frac{\delta^2 N\varphi(v)/2}{C + \pi_N C\delta/3}\right)$$

and $Nq_N\pi_N^{-1}(\log N)^{-1} \to \infty$ implies the desired result. The second statement of (8) follows from a similar argument.

For (9),

$$|\hat{g}(\hat{V}_i) - \tilde{g}(V_i)| \leq \frac{\varphi(V_i)}{\varphi(\hat{V}_i)} \frac{1}{N\varphi(V_i)} \sum_{j\neq i} |Y_{jN}| K\left(\frac{X_j - x}{h_N}\right) L^*\left(\frac{V_j - V_i}{h_N}\right) \frac{\|\hat{V}_j - V_j\|}{h_N} + \frac{|\varphi(V_i) - \varphi(\hat{V}_i)|}{\varphi(\hat{V}_i)} |\tilde{g}(V_i)|$$

and the desired result follows from $\max_{1\leq i\leq N} \|\hat{V}_i - V_i\| \mathbb{1}\{|Z_i| \leq \tau_N\} = o_{\mathbb{P}}(h_N)$, Lemma 1, (10), $Nh_N^{d_z} q_N \to \infty$, (8), and $m(X_j, v)\mathbb{1}\{|X_j - x| \leq \delta\} < \infty$.

A.2.1 Lemmas

The following inequality is a simple application of Lipschitz continuity. As it arises multiple times in the proof, I state it here for reference.

$$|\varphi(v) - \varphi(u)| \leq \mathbb{E}\left[K\left(\frac{X_i - x}{h_N}\right) L^*\left(\frac{V_i - u}{h_N}\right)\right] \frac{\|v - u\|}{h_N} = \varphi^*(u) \frac{\|v - u\|}{h_N}. \tag{10}$$

The following lemma provides a lower bound on $\varphi(v), \varphi^*(v)$, which are a version of small ball probabilities (see Chapter 13 of Ferraty and Vieu, 2006, for discussion).

Lemma 1. *Assumptions 6 and 7 hold and for a given $v \in \text{supp}(V_i)$, pick $z \in \text{supp}(Z_i)$ such that $v = F_{W|Z}(\cdot|z)$. Then, for some $C > 0$,*

$$\min\{\varphi(v), \varphi^*(v)\} \geq h_N^{d_z} \inf_{|e|\leq 1} f_Z(z + e)C$$

for sufficiently large N.

Proof. Using $\min\{L(v), L^*(v)\} \geq c_L \mathbb{1}\{|v| \leq c_L\}$ and Lipschitz continuity of $F_{W|Z}$,

$$\begin{aligned}
\min\{\varphi(v), \varphi^*(v)\} &\geq c_L \mathbb{E}\left[K\left(\frac{X_i - x}{h_N}\right) \mathbb{1}\left\{\|V_i - v\| \leq c_L h_N\right\}\right] \\
&\geq c_L \mathbb{E}\left[K\left(\frac{X_i - x}{h_N}\right) \mathbb{1}\left\{C_F|Z_i - z| \leq c_L h_N\right\}\right] \\
&= c_L h_N^{d_z} \int_{\mathbb{R}^{d_z}} K(s_1) \mathbb{1}\{C_F|s| \leq c_L\} f_Z(z + s h_N) ds
\end{aligned}$$

where $s_1 \in \mathbb{R}^{d_x}$ is a subvector of $s \in \mathbb{R}^{d_z}$ and using compact support of K, we obtain the desired result. □

Lemma 2. *Recall $A_N = \{v : \exists z \text{ s.t. } v = F_{W|Z}(\cdot|z) \ \& \ |z| \leq \tau_N\}$. Under Assumption 7,*

$$\sup_{v \in A_N} \frac{\varphi^*(v)}{\varphi(v)} = O\Big(\inf_{|z| \leq \tau_N} f_Z(z)\Big)$$

Proof. For v, pick z such that $v = F_{W|Z}(\cdot|z)$ and we have $\|V_i - v\| = \|\partial F_{W|Z}(\cdot|\tilde{z})/\partial z'(Z_i - z)\|$ where $\tilde{z}$ is some value between Z_i and z. Because the derivative is non-zero everywhere, for some $c_z > 0$,

$$\varphi^*(v) \leq \mathbb{E}\left[K\left(\frac{X_i - x}{h_N}\right) L^*\left(\frac{c_z|Z_i - z|}{h_N}\right)\right] = h^{d_z} \int K(s_1) L^*(c_z|s|) f_Z(z + s h_N) ds$$

and by Lemma 1, we obtain the desired result. □

B Additional Discussion

B.1 Identification Without Excluded Group-Level Variables

Here, I consider an alternative model where a researcher does not observe excluded group-level variables in the selection equation. Using the notation from the main paper, set up the following model.

$$\begin{aligned}
Y_{ig} &= m(X_g, \nu_i, u_g) \\
J_i &= J(A_1, \cdots, A_G, \Theta_i, \zeta_{i1}, \cdots, \zeta_{iG})
\end{aligned}$$

where A_g represents unobserved group-level features affecting group determination and it generally includes X_g as a subvector. The difference from (2)-(3) lies in the inability to observe group-level variables entering the selection equation. Since we do not observe A_g, we cannot compute the conditional distribution $f_{W|A_J}$, which we would use as a control variable if A_g were observed.

To accommodate this setting, assume that data contain many individuals for each group. Then, we can compute the group-level distribution by using observations within each group. Specifically, for each group g, we can compute

$$\frac{1}{N_g}\sum_{i=1}^{N}\mathbb{1}\{W_i \leq w\}\mathbb{1}\{J_i = g\}, \qquad N_g = \sum_{i=1}^{N}\mathbb{1}\{J_i = g\}$$

and under appropriate conditions, as $N_g \to \infty$, this within-group sample average converges in probability to

$$F_{W|AJ}(w|A_1, \ldots, A_G, g).$$

The conditioning on A_g's follows from $J_i = J(A_1, \cdots, A_G, \Theta_i, \zeta_{i1}, \cdots, \zeta_{iG})$ and the fact that averaging over i with $J_i = g$ holds $(A_1, \cdots, A_G)$ fixed. This averaging over i holding constant $J_i = g$ is akin to averaging individual variables over time dimension in panel data models: i.e., if Y_{it} denotes an outcome for individual i at time t and α_i represents individual fixed effects, $\sum_{t=1}^{T} Y_{it}/T \to_{\mathbb{P}} \mathbb{E}[Y_{it}|\alpha_i]$ under appropriate conditions.

With the identified group-level distribution $F_{W|AJ}$, we can now use the same identification argument as in the main paper to identify averages effects of X_g on Y_{ig}.

B.2 Other Measures of Partial Effects

In the main paper, I only discuss the average structural function to focus on main ideas. Here, I consider other measures of partial effects. The quantile structural function (QSF) (Imbens and Newey, 2009) is the quantile of $m(x, \varepsilon_{iJ_i})$ for a fixed x. Note that the quantile is computed using the marginal distribution of ε_{iJ_i}. To consider the identification of this parameter, note

$$F_{Y|XV}(y|x, v) = \mathbb{E}[\mathbb{1}\{Y_{iJ_i} \leq y\}|X_{J_i} = x, V_i = v] = \mathbb{E}[\mathbb{1}\{m(x, \varepsilon_{iJ_i}) \leq y\}|V_i = v]$$

where $V_i = f_{W|Z_J}(\cdot|Z_{J_i})$ and the second equality follows under the hypothesis of Theorem 1. Then, provided that the common support condition holds, integrating this object with respect to the

marginal distribution of V_i yields

$$\mathbb{E}[\mathbb{1}\{m(x, \varepsilon_{iJ_i}) \leq y\}],$$

which is the distribution function for the random variable $m(x, \varepsilon_{iJ_i})$, and the left-inverse of this distribution function is the QSF. Imbens and Newey (2009) discuss bounds for the QSF when the common support condition fails.

Next, I consider the local average response (LAR) of Altonji and Matzkin (2005). Assume X_g is continuously distributed and $m(x, \varepsilon)$ is continuously differentiable with respect to x. Then, the LAR is defined as

$$\int m_x(x, e) f_{\varepsilon|X_J}(e|x) de, \qquad m_x(x, e) = \frac{\partial}{\partial x} m(x, e).$$

To identify this parameter, suppose that $X_g \subset Z_g$ with Z_{2g} denoting the excluded elements, Z_g has a Lebesgue density f_Z, and define $\mathcal{Z}(x, v) = \{z_2 \in \text{supp}(Z_{2g}) : v = F_{W|Z}(\cdot|x, z_2)\}$. For a given $x_0 \in \text{supp}(X_g)$, assume that if $f_Z(x_0, z_2) > 0$, then there exists a $\delta > 0$ such that for all $|x - x_0| \leq \delta$, we can find $z_2' \in \mathcal{Z}(x, F_{W|Z}(\cdot|x_0, z_2))$ with $f_Z(x, z_2') > 0$. Furthermore, if for some $\delta > 0$

$$\mathbb{E}\Big[\int \sup_{|x - x_0| \leq \delta} |m_x(x, e)| dF_{\varepsilon|V}(e|V_i)\Big] < \infty$$

holds, then the LAR is identified at x_0. The involved condition for positivity of f_Z translates Assumption 2.2 in Altonji and Matzkin (2005) to this setting.

B.3 Uniform Convergence Rates for the Conditional Distribution Function

Hansen (2008) and Cattaneo et al. (2013) present results on uniform convergence rates for the following nonparametric estimator.

$$\frac{1}{N} \sum_{i=1}^{N} Y_i \kappa_N(X_i - x)$$

where $\kappa_N(u) = b_N^{-d_x} \kappa(u/b_N)$, κ is a kernel function, and b_N is a bandwidth sequence. In this paper, I need to obtain a rate for

$$\sup_{w \in \mathbb{R}^{d_w}, |z| \leq \tau_N} \left|\hat{F}_{W|Z}(w|z) - F_{W|Z}(w|z)\right|$$

where $\hat{F}_{W|Z}(w|z) = \hat{\Psi}(w|z)/\hat{f}(z)$,

$$\hat{\Psi}(w|z) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{W_i \leq w\}\kappa_N(Z_i - z), \qquad \hat{f}(z) = \frac{1}{N}\sum_{i=1}^{N} \kappa_N(Z_i - z).$$

For $\hat{f}(z)$, I can directly apply the results from the aforementioned papers. For $\hat{\Psi}(w|z)$, I need to handle uniformity in evaluation points w. We have the decomposition

$$\hat{\Psi}(w|z) - F_{W|Z}(w|z)f_Z(z) = \hat{\Psi}(w|z) - \mathbb{E}[\hat{\Psi}(w|z)] + \mathbb{E}[\hat{\Psi}(w|z)] - F_{W|Z}(w|z)f_Z(z)$$

and by assuming appropriate differentiability and boundedness conditions,

$$\left|\mathbb{E}[\hat{\Psi}(w|z)] - F_{W|Z}(w|z)f_Z(z)\right| \leq Cb_N^P$$

where P is the order of kernel κ and C is independent of w and z. Then, it remains to bound

$$\mathbb{P}\left[\sup_{w\in\mathbb{R}^{d_w}, |z|\leq\tau_N} \left|\hat{\Psi}(w|z) - \mathbb{E}[\hat{\Psi}(w|z)]\right| > Cr_N\right]$$

for some rate r_N. Assume Lipschitz continuity of κ as I did in Assumption 7 for L. Also, discretize the space $\{|z| \leq \tau_N\}$ by $\{z_m^*\}_{m=1}^M$ as done in the proof of Theorem 2. Then,

$$\begin{aligned}\left|\hat{\Psi}(w|z) - \mathbb{E}[\hat{\Psi}(w|z)]\right| \leq \left|\hat{\Psi}(w|z^*) - \mathbb{E}[\hat{\Psi}(w|z^*)]\right| + \left|\hat{\Psi}^*(w|z^*) - \mathbb{E}[\hat{\Psi}^*(w|z^*)]\right|\left|\frac{z-z^*}{b_N}\right| \\ + 2\left|\mathbb{E}[\hat{\Psi}^*(w|z^*)]\right|\left|\frac{z-z^*}{b_N}\right|\end{aligned}$$

where $\hat{\Psi}^*(w|z^*) = \sum_{i=1}^N \mathbb{1}\{W_i \leq w\}\kappa_N^*(Z_i - z^*)/N$. Then, I need to bound probabilities like

$$\mathbb{P}\left[\sup_{w\in\mathbb{R}^{d_w}} \max_{1\leq m\leq M} \left|\hat{\Psi}(w|z_m^*) - \mathbb{E}[\hat{\Psi}(w|z_m^*)]\right| > Cr_N\right].$$

Applying symmetrization technique from empirical process theory (e.g., Lemma 2.3.1 of van der Vaart and Wellner, 1996), the above probability is bounded by

$$\mathbb{P}\left[2\sup_{w\in\mathbb{R}^{d_w}} \max_{1\leq m\leq M} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}\{W_i \leq w\}\kappa_N(Z_i - z_m^*)\right| > Cr_N\right]$$

where $\varepsilon_i \in \{-1, 1\}$ is a Rademacher variable independent of data. Now, conditioning on the data,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{w\in\mathbb{R}^{d_w}} \max_{1\le m\le M} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}\{W_i \le w\}\kappa_N(Z_i - z_m^*)\right| > Cr_N \middle| (W_i, Z_i)_{i=1}^N\right] \\
&= \mathbb{P}\left[\max_{1\le \ell\le L} \max_{1\le m\le M} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}\{W_i \le w_\ell\}\kappa_N(Z_i - z_m^*)\right| > Cr_N \middle| (W_i, Z_i)_{i=1}^N\right] \\
&\le \sum_{\ell,j} \mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}\{W_i \le w_\ell\}\kappa_N(Z_i - z_m^*)\right| > Cr_N \middle| (W_i, Z_i)_{i=1}^N\right]
\end{aligned}$$

where replacing the supremum with maximum over a finite number of points is possible as we condition on W_i's. Note that $L = O(N^{d_w})$. Then, applying Hoeffding's inequality,

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \mathbb{1}\{W_i \le w_\ell\}\kappa_N(Z_i - z_m^*)\right| > Cr_N \middle| (W_i, Z_i)_{i=1}^N\right] \le 2\exp\left(-\frac{C^2 N r_n^2/2}{\frac{1}{N}\sum_{i=1}^{N} |\kappa_N(Z_i - z_m^*)|^2}\right).$$

Now applying Lemma 33 in Chapter 2 of Pollard (1984), $\max_{1\le m\le M} \frac{1}{N}\sum_{i=1}^{N} |\kappa_N(Z_i - z_m^*)|^2$ is bounded by a constant multiple of $\max_{1\le m\le M} \mathbb{E}[|\kappa_N(Z_i - z_m^*)|^2] = O(b_N^{-d_z})$ with probability approaching one if $nb_N^{d_Z}/\log N \to \infty$. Then, we can take $r_N = (\log N/Nb_N^{d_z})^{1/2}$. Finally, assuming all the regularity conditions hold for the above arguments, we obtain

$$\sup_{w\in\mathbb{R}^{d_w}, |z|\le\tau_N} \left|\hat{\Psi}(w|z) - F_{W|Z}(w|z)f_Z(z)\right| = O_{\mathbb{P}}\left(b_N^P + \sqrt{\frac{\log N}{Nb_N^{d_z}}}\right),$$

which can be used to obtain the convergence rate for $\sup_{w\in\mathbb{R}^{d_w}, |z|\le\tau_N} \left|\hat{F}_{W|Z}(w|z) - F_{W|Z}(w|z)\right|$.