Matching Points: Supplementing Instruments with Covariates in Triangular Models *

Junlong Feng†

Job Market Paper

Current version: Click here.
This version: January 8, 2020

Abstract

We consider triangular models with a discrete endogenous variable and an instrumental variable (IV) taking on fewer values. Addressing the failure of the order condition, we develop the first approach to restore identification for both separable and nonseparable models in this case by supplementing the IV with covariates, allowed to enter the model in an arbitrary way. For the separable model, we show that it satisfies a system of linear equations, yielding a simple identification condition and a closed-form estimator. For the nonseparable model, we develop a new identification argument by exploiting its continuity and monotonicity, leading to weak sufficient conditions for global identification. Built on it, we propose a uniformly consistent and asymptotically normal sieve estimator. We apply our approach to an empirical application of the return to education with a binary IV. Though under-identified by the IV alone, we obtain results consistent with the literature using our approach. We also illustrate the applicability of our approach via an application of preschool program selection where the supplementation procedure fails.

Keywords: Nonparametric identification, triangular model, instrumental variable, endogeneity, generalized propensity scores.

*I thank my advisors Jushan Bai, Sokbae Lee and Bernard Salanié, who were gracious with their advice, support and feedback. I have also greatly benefited from comments and discussions with Karun Adusumili, Isaiah Andrews, Andres Aradillas-Lopez, Sandra Black, Ivan Canay, Xiaohong Chen, Leonard Goff, Florian Gunsilius, Han Hong, Jessie Li, José Luis Montiel Olea, Ulrich Müller, Whitney Newey, Serena Ng, Christoph Rothe, Jörg Stoye, Matt Taddy, Alexander Torgotivtsky, Quong Vuong, Yulong Wang, Kaspar Wuthrich and participants of the Columbia Econometrics Colloquium and Workshop as well as the participants of the seminar at the 2019 Econometric Society Asian Meeting in Xiamen. I also thank Research Connections for providing the data of the Head Start Impact Study. All errors are my own.

†Dept. of Economics, Columbia University, New York, NY 10027, U.S.A.; junlong.feng@columbia.edu.

1 Introduction

This paper considers identification and estimation of the outcome function $\boldsymbol{g}^* \equiv (g_d^*)_d$ in a triangular model:

$$Y = \sum_d \mathbb{1}(D = d) \cdot g_d^*(\boldsymbol{X}, U)$$
$$D = h(\boldsymbol{X}, Z, \boldsymbol{V})$$

where both the endogenous variable D and the instrumental variable (IV) Z are discrete, $\boldsymbol{X}$ is a vector of covariates, and the disturbances U and $\boldsymbol{V}$ are correlated (see also Newey, Powell and Vella (1999), Chesher (2003), Matzkin (2003), Newey and Powell (2003), Chernozhukov and Hansen (2005), Das (2005), Imbens and Newey (2009), etc.)

It is well-known that in general, $\boldsymbol{g}^*$ is not identified if Z takes on fewer values than D does. In many applications, however, IVs do have very small support while endogenous variables may take on more values.

Let us consider an example of the return to education. Suppose the log wage (Y) is determined by unobserved earning ability U and functions of covariates ($\boldsymbol{X}$) such as parents' education. These functions are heterogeneous in the level of education d: completing high school ($d = 1$), having some college education ($d = 2$), and at least completing college ($d = 3$). The available IV may be only binary. For instance, in Card (1995), Z indicates whether an individual lived near a 4-year college or not.

To see why identification may fail, suppose the unknown function is separable in ability: $g_d^*(\boldsymbol{X}, U) = m_d^*(\boldsymbol{X}) + U$. Then the model can be rewritten as $Y = \alpha(\boldsymbol{X}) + m_2^*(\boldsymbol{X})\mathbb{1}(D = 2) + m_3^*(\boldsymbol{X})\mathbb{1}(D = 3) + U$. The classical order condition thus does not hold: conditional on $\boldsymbol{X}$, there are two endogenous variables, $\mathbb{1}(D = 2)$ and $\mathbb{1}(D = 3)$, but only one binary IV.

Under the standard validity assumptions for the IV, it can be shown that the outcome function $\boldsymbol{m}^*$ satisfies the moment condition $\sum_{d=1}^3 p_d(\boldsymbol{x}_0, Z) m_d^*(\boldsymbol{x}_0) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}_0, Z)$ for some $\boldsymbol{x}_0$ where $p_d(\cdot, \cdot)$ is the generalized propensity score (e.g. Newey and Powell (2003)). With a binary Z, we obtain two equations by conditioning on each value that Z can take, but there are three unknowns. To the best of the author's knowledge, no existing method achieves point-identification in such a case.

This paper develops the first approach that obtains point-identification of $\boldsymbol{g}^*$ when the IV takes on fewer values than the discrete endogenous variable. This is achieved by supplementing the IV with variation in $\boldsymbol{X}$. We show that for a fixed $\boldsymbol{x}_0$, there may exist a *matching point* $\boldsymbol{x}_m$ such that the difference between $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ and $\boldsymbol{g}^*(\boldsymbol{x}_m, \cdot)$ is identified. Controlling for the difference, moment equations like the example above can be evaluated at $\boldsymbol{x}_m$ in addition to $\boldsymbol{x}_0$ without introducing new unknowns. In this way, the effective support set of the IV is

enlarged via the matching points, making identification possible.

To see why such outcome function differences may be identified before the outcome functions themselves are, note that in the triangular model, endogeneity is generally due to the dependency between U and $\boldsymbol{V}$. Suppose $\boldsymbol{X}$ and Z generate partitions in the space of $\boldsymbol{V}$ (for instance in an ordered choice model). Selecting into a value of D is determined by which partition $\boldsymbol{V}$ falls into. Hence, if for some $z \neq z'$, $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_m, z')$ generate exactly the same partitions, then the same selection choices would be made across the two schemes for any realization of the latent $\boldsymbol{V}$. The unknown selection biases at these two points would thus be equal and could be canceled out. The relationship between $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ and $\boldsymbol{g}^*(\boldsymbol{x}_m, \cdot)$ can thus be traced out from the distribution of the observed outcomes at $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_m, z')$.

Let us go back to the return to education example. Suppose we have a single covariate X, the average of parents' years of schooling. Let X and Z enter h via a linear single index $kZ + X$. It implies that k more years of the parents' schooling compensate for not living near a 4-year college ($Z = 0$) in terms of educational attainment choices. For any realization of $\boldsymbol{V}$, individuals with $(X, Z) = (x_0, 0)$ and with $(X, Z) = (x_0 - k, 1)$ would select into the same level of education. These two groups of individuals are equivalent in terms of selection, so their selection biases are presumably the same. Comparing their (average or distributions of) observed log wages, the biases may be differenced out.

To find the matching points of a given $\boldsymbol{x}_0$, we do not restrict the dimension of $\boldsymbol{V}$ and no notion of monotonicity is imposed. We propose a condition called *propensity score coherence*. Under it, the matching points can be found by matching the generalized propensity scores at different values of $\boldsymbol{X}$ and Z without specifying the selection model $\boldsymbol{h}$. We provide examples to illustrate that many widely used discrete choice models satisfy this condition.

Given the matching points, we derive the exact forms of the outcome function differences for two particular models of $\boldsymbol{g}^*$: additively separable in U, and nonseparable and strictly increasing in U. For each model, we provide sufficient conditions for identification and construct consistent and asymptotically normal estimators.

For the separable model, we show that the outcome function solves a system of linear equations, preserving a similar structure as in the standard IV approach. We thus obtain a closed-form estimator which is easy to implement in practice. We apply it to examine the return to education example using the same extract from 1979 National Longitudinal Surveys (NLS) as in Card (1995). We adopt the proximity-to-college IV and find that living near a four-year college and parents' years of schooling are indeed close substitutes in terms of children's educational attainment. Using the matching points generated from this covariate, we find that the return to education is (a) increasing in the level of schooling with slightly diminishing marginal return, and (b) heterogeneous in parents' education; individuals with

less educated parents enjoy higher potential returns. In contrast, the two-stage-least-squares (2SLS) estimates of parametric models using the interaction of parents' years of schooling and the proximity-to-college IV as an extra instrument leads to misleading results.

For the nonseparable model, we develop a new identification argument by exploiting continuity and monotonicity of $\boldsymbol{g}^*(\boldsymbol{X}, \cdot)$. We show that global identification of $\boldsymbol{g}^*(\boldsymbol{X}, \cdot)$ is achieved in the space of monotonic functions if $\boldsymbol{g}^*(\boldsymbol{X}, u)$ is only locally identified for each u. This new result also applies to the standard IV approach when the IV has large support. Based on our identification strategy, we construct a sieve estimator. We show that its large sample properties are guaranteed by simple low-level conditions, thanks to the nice properties of the monotonic function space.

It is worth noting that the success of our approach hinges on the covariates that are able to offset the impact of Z. For applications where the IV has the dominant effect, covariates may not have comparable effects on the selection so the matching points may not exist. This is testable in some cases. As an illustration, we consider another empirical application on the preschool program selection. We use the administrated Head Start Impact Study (HSIS) dataset following Kline and Walters (2016). The endogenous variable considered also takes on three values: participating in Head Start, in an alternative preschool program, and not participating in any programs. The binary IV indicates whether an individual won a lottery granting access to Head Start. The IV has very large effect on the choice of preschool programs. From the tests we develop, we find that no available covariate in the sample is able to generate a matching point.

We defer a detailed comparison of our method to the existing literature until Section 8. Here we highlight some major differences. Precursory methods that circumvent the problem of having an IV with small support include imposing homogeneity between adjacent levels of D when D is ordered, or specifying a parametric form for $\boldsymbol{g}^*$ and using interactions between Z and $\boldsymbol{X}$ as a second IV by assuming $\boldsymbol{X}$ is exogenous. Torgovitsky (2015, 2017) and D'Haultfœuille and Février (2015) show a binary IV is able to identify nonseparable models with a continuous endogenous variable. Continuity is crucial in their approach and they require the selection function strictly increasing in the scalar unobservable. Similar to this paper, Caetano and Escanciano (2018) also use covariates to identify models when the instruments do not have enough variation. Their approach does not rely on the first stage, but they need the covariates used for identification purpose to be "separable" in the model in a way that the model can be "inverted" and become free of them. In contrast, the covariates in our approach can enter the model in an arbitrary way. Huang, Khalil and Yıldız (2019) consider identification of separable models with multiple endogenous variable but a single instrument. They focus on a partial linear model and one of the endogenous

variable needs to be continuous to apply the control function technique. Ichimura and Taber (2000) and Vytlacil and Yıldız (2007) use shifts in some observables that compensate for a shift in a target variable to facilitate identification of different parameters than this paper. The shifting variables and the target variable are different from ours. Vuong and Xu (2017) and Feng, Vuong and Xu (2019) in the study of the individual treatment effect of a binary D develop a concept called the counterfactual mapping. It is also an identifiable function linking two outcome functions but at *different* values of D and the *same* value of X by exploiting the compliers' information.

The rest of the paper is organized as follows. In Section 2, we introduce the model, discuss the preliminary assumptions, and introduce the matching points. We also preview the basic idea of the new identification strategy. In Section 3, we discuss the existence of the matching points and provide sufficient conditions for identification of the matching points and the outcome functions. In Section 4, we propose estimators for them and discuss some implementation issues. Section 5 presents results of two empirical applications. Section 6 shows the estimators' asymptotic properties. Section 7 provides Monte Carlo simulations to illustrate the estimator's finite sample performance. Section 8 discusses the relation of our approach to the related work. Section 9 concludes. Appendix A discusses general cases including models with multiple discrete endogenous variables. Appendix B contains proofs of the results in Sections 2 and 3. Appendix C illustrates the propensity coherence condition via various discrete choice models for a single or multiple endogenous variables. In the supplementary appendices, Appendix D provides additional simulation results, and Appendix E collects proofs of the asymptotic results.

Notation

We use upper-case Latin letters for random variables and the corresponding lower-cases for their realizations. Bold Latin letters denote vectors or matrices. For two generic random variables A and B, denote the conditional expectation of A given $B = b$ by $\mathbb{E}_{A|B}(b)$, with similar notation for conditional distribution functions, densities and variances. Denote the support set of A by $S(A)$, and the support of A given $B = b$ by $S(A|B = b)$, or simply $S(A|b)$ when it does not cause confusion. For a finite set H, $|H|$ denotes the number of elements in it, while for a generic vector $\boldsymbol{c}$, $|\boldsymbol{c}|$ denotes its Euclidean norm. For two generic sets H_1 and H_2, $H_1 \backslash H_2$ denotes the set difference $H_1 \cap H_2^c$. Throughout, we assume all the random variables involved are in a common probability space with the measure function $\mathcal{P}$. Whenever we say almost surely (a.s.) and measurable, we refer to almost surely and measurable with respect to $\mathcal{P}$.

2 The Model

To highlight the key features of our approach, we focus on a simple case where the endogenous D takes on three values ($|S(D)| = 3$) and Z is binary ($|S(Z)| = 2$). We will also discuss the usefulness of the approach when $|S(D)| = |S(Z)| = 2$. The general cases for arbitrary $|S(D)| \geq |S(Z)|$ and multiple Ds will be discussed in Appendix A.1.

We study the separable model and the nonseparable model respectively:

$$Y = \sum_{d \in S(D)} \mathbb{1}(D = d) \cdot \left(m_d^*(\boldsymbol{X}) + U \right) \tag{SP}$$

and

$$Y = \sum_{d \in S(D)} \mathbb{1}(D = d) \cdot g_d^*(\boldsymbol{X}, U) \tag{NSP}$$

where $S(D) \equiv \{1, 2, 3\}$, $\boldsymbol{X}$ is a vector of covariates, and U is a scalar unobservable. The goal of this paper is to identify and estimate the outcome functions at a fixed value of $\boldsymbol{X}$: $\boldsymbol{m}^*(\boldsymbol{x}_0) \equiv (m_d^*(\boldsymbol{x}_0))_d$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot) \equiv (g_d^*(\boldsymbol{x}_0, \cdot))_d$. Note the choice of $S(D)$ is without loss of generality because any set of three element can be one-to-one mapped onto it.

We rewrite the selection model for D as follows:

$$D = d \text{ if and only if } h_d(\boldsymbol{X}, Z, \boldsymbol{V}) = 1 \tag{SL}$$

where for all $d \in S_D$, the selection function $h_d(\boldsymbol{X}, Z, \boldsymbol{V}) \in \{0, 1\}$, and $\sum_{d=1}^{3} h_d(\boldsymbol{X}, Z, \boldsymbol{V}) = 1$ a.s. $\boldsymbol{V}$ is a vector of unobservables that is correlated with U. We assume that for every $(\boldsymbol{x}, z) \in S(\boldsymbol{X}, Z)$, $h_d(\boldsymbol{x}, z, \cdot)$ is measurable on $S(\boldsymbol{V})$.

In the rest of this section, we introduce and discuss preliminary assumptions for each model. We also illustrate why a binary Z in general fails to identify the outcome functions. Finally, we introduce the key idea to restore identification.

2.1 The Separable Model

Let us begin with the assumption for the separable model-SP:

Assumption E-SP (Exogeneity)**.** $\mathbb{E}_{U|\boldsymbol{X}}(\boldsymbol{x}_0) = 0$, $\mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_0, Z) = \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{V}, \boldsymbol{x}_0)$ *a.s., and* $Z \perp\!\!\!\perp \boldsymbol{V} | \boldsymbol{X} = \boldsymbol{x}_0$.

The first condition in Assumption E-SP is a normalization without which $\boldsymbol{m}^*(\boldsymbol{x}_0)$ can only be identified up to an additive constant. The second and the third conditions are standard in the literature of triangular models (e.g. Newey, Powell and Vella (1999)).

Remark 2.1. *Note that the unobservable U is not d-dependent, so our model is more restrictive than many models in the treatment effects literature, for example Heckman and Vytlacil (2005) and Lee and Salanié (2018). However, in these works, the methods usually need much richer variation in the instruments (continuous and multidimensional). In our setup, we can allow U to be d-dependent by making extra assumptions. For example, we can show that our results still hold if $\mathbb{E}_{U_d|D\boldsymbol{X}Z}(D, \boldsymbol{x}_0, Z) = \mathbb{E}_{U_{d'}|D\boldsymbol{X}Z}(D, \boldsymbol{x}_0, Z)$ for any $d \neq d'$. This assumption allows U_d for each d to have different conditional distributions so long as they have the same mean dependence of the endogenous variable.*

Proposition 1 (Newey and Powell (2003) equation (2.2); Das (2005) equation (2.5))**.** *Under Assumptions E-SP, the following equation holds for all $z \in S(Z)$,*

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) \tag{2.1}$$

where $p_d(\boldsymbol{x}_0, z) \equiv \mathbb{P}(D = d|\boldsymbol{X} = \boldsymbol{x}_0, Z = z)$.

Since all the terms in equation (2.1) are directly identified from the population except for $\boldsymbol{m}^*(\boldsymbol{x}_0)$, we have two linear equations by letting $z = 0, 1$ but three knowns: $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is not identified without additional information.

2.2 The Nonseparable Model

Compared to the separable model, assumptions for the nonseparable model-NSP are more stringent.

Assumption E-NSP (Exogeneity)**.** $(U|\boldsymbol{X} = \boldsymbol{x}_0) \sim Unif[0, 1]$ *and* $(U, \boldsymbol{V}) \perp\!\!\!\perp Z|\boldsymbol{X} = \boldsymbol{x}_0$.

Assumption FS (Full Support)**.** $(U, \boldsymbol{V})|\boldsymbol{x}_0$ *is continuously distributed and* $S(U|\boldsymbol{V}, \boldsymbol{x}_0) = S(U|\boldsymbol{x}_0)$.

Assumption CM (Continuity and Monotonicity)**.** *For all* $(d, \boldsymbol{x}) \in S(D, \boldsymbol{X})$, $g_d^*(\boldsymbol{x}, \cdot)$ *is continuous and strictly increasing on* $[0, 1]$.

Assumption E-NSP is the counterpart of Assumption E-SP for the nonseparable outcome functions; the first part is a popular normalization for identification of nonseparable models, while the second part is the same as in Imbens and Newey (2009) which is standard for triangular models. Similar to Model-SP, U is invariant with respect to d. We can relax it by adopting the *rank similarity* condition in Chernozhukov and Hansen (2005).

Assumption FS guarantees that the range of $g_d^*(\boldsymbol{x}, z, \cdot)$ on $[0,1]$ is equal to the conditional support $S(Y|d,\boldsymbol{x})$[1]. The same assumption can be found in related work that also focuses on identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ on the entire domain as this paper, for instance D'Haultfœuille and Février (2015), Torgovitsky (2015) and Vuong and Xu (2017).

Assumption CM regulates the behavior of the NSP-outcome function $\boldsymbol{g}^*(\boldsymbol{x}, \cdot)$. Continuity on $[0,1]$ and Assumption FS (a) imply that $Y|d,\boldsymbol{x},z$ is continuously distributed and that $S(Y|d,\boldsymbol{x},z)$ is compact. Continuity and strict monotonicity are two standard requirements in the literature of nonseparable models when the unobservable is a scalar. In addition to using these properties to construct moment conditions as in the related literature, in this paper we show that they deliver nice results for identification and for deriving the large sample properties of the estimator we propose.

Under these assumptions, we have the following result:

Proposition 2 (Chernozhukov and Hansen (2005), Theorem 1)**.** *Under Assumptions E-NSP, FS and CM, the following equation holds for all* $z \in \{0,1\}$ *and* $u \in [0,1]$*,*

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) = u \tag{2.2}$$

Similar to Model-SP, again we have two equations but three unknowns for a fixed u. Global uniqueness of the solution is in general not guaranteed.

2.3 The Selection Model and the Matching Points

Now we show how to use covariates $\boldsymbol{X}$ to supplement the binary Z to restore identification when the order condition fails.

The major challenge for identification by varying the conditioning value of $\boldsymbol{X}$ is that it results in unknown changes in the outcome function: while more moment conditions are generated, even more unknowns are introduced into the new system of equations. For instance, consider equation (2.1) for Model-SP. Suppose Assumptions E-SP also holds for $\boldsymbol{x}' \neq \boldsymbol{x}_0$. Similar to (2.1), we have

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}', z) \cdot m_d^*(\boldsymbol{x}') = \sum_{d=1}^{3} p_d(\boldsymbol{x}', z) \cdot \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}', z)$$

for $z \in \{0,1\}$. We then have 4 equations in total: two conditional on $\boldsymbol{X} = \boldsymbol{x}_0$ and two on $\boldsymbol{x}'$, yet the number of the unknowns is increased to 6 at the same time. Therefore, an arbitrarily

[1]This is because $S(Y|d,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},Z,\boldsymbol{V}) = 1,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},z,\boldsymbol{V}) = 1,\boldsymbol{x},z) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},z,\boldsymbol{V}) = 1,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U))$, which is the range of $g_d^*(\boldsymbol{x},\cdot)$ on [0,1].

chosen $\boldsymbol{x}'$ does not help identification.

Instead, we look for a point, denoted by $\boldsymbol{x}_m$, such that the difference between $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{m}^*(\boldsymbol{x}_m)$ can be identified first. To see this is possible, let us take conditional expectation on both sides of equation (SP). For all $d, \boldsymbol{x}, z$ in their support,

$$m_d^*(\boldsymbol{x}) + \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z) = \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z).$$

The term $\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z)$ captures the selection bias due to endogeneity of D. Then for $\boldsymbol{x}_m \neq \boldsymbol{x}_0$, the difference between $m_d^*(\boldsymbol{x}_m)$ and $m_d^*(\boldsymbol{x}_0)$ satisfies:

$$\begin{aligned} m_d^*(\boldsymbol{x}_m) - m_d^*(\boldsymbol{x}_0) = & \underbrace{\left(\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\right)}_{\text{Difference in the Observed Outcomes}} \\ & - \underbrace{\left(\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\right)}_{\text{Difference in the Biases}} \end{aligned}$$

When the two unknown bias terms are equal, they cancel out and the change in the outcome function is identified.

Let us consider the following example for illustration.

Example OC (Ordered Choice)**.** *Suppose D is ordered and there is only one covariate. Let $h_1(X, Z, V) = \mathbb{1}(V < \kappa_1 + \beta X + \alpha Z)$, $h_3(X, Z, V) = \mathbb{1}(V \geq \kappa_2 + \beta X + \alpha Z)$, and $h_2 = 1 - h_1 - h_3$. Assume $\alpha \cdot \beta \neq 0$, $\kappa_1 < \kappa_2$, and $(X, Z) \perp\!\!\!\perp V$ where V is continuously distributed on $\mathbb{R}$. Fix x_0, it is straightforward to see that $(x_0, 0)$ and $(x_0 - \frac{\alpha}{\beta}, 1)$ generate exactly the same partitions on $\mathbb{R}$. Then taking $d = 1$ as an example, we have*

$$\begin{aligned} \mathbb{E}_{U|D\boldsymbol{X}Z}(1, x_0 - \frac{\alpha}{\beta}, 1) &= \mathbb{E}_{U|V\boldsymbol{X}Z}(V < \kappa_1 + \beta x_0, x_0 - \frac{\alpha}{\beta}, 1) \\ \mathbb{E}_{U|D\boldsymbol{X}Z}(1, x_0, 0) &= \mathbb{E}_{U|V\boldsymbol{X}Z}(V < \kappa_1 + \beta x_0, x_0, 0) \end{aligned}$$

When the dependency of (U, V) on $(X, Z) = (x_0, 0)$ and $(x_0 - \frac{\alpha}{\beta}, 1)$ are identical, the two bias terms are equal.

From the example, we can see that in order to difference out the bias, $(\boldsymbol{x}_m, z')$ and $(\boldsymbol{x}_0, z)$ should (a) generate the same partitions of $S(\boldsymbol{V})$, and (b) have the same level of dependency with respect to the unobservables. The following conditions formally characterize these ideas.

Definition MP (Matching Points and Matching Pairs)**.** *A point $\boldsymbol{x}_m \in S(\boldsymbol{X})$ is a matching point of $\boldsymbol{x}_0 \in S(\boldsymbol{X})$ if there exist $z \neq z' \in S(Z)$ such that for all $d \in S(D)$,*

$$h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) \text{ a.s.}, \tag{2.3}$$

and for Model-SP,

$$\mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_m, Z) = \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_0, Z) \text{ a.s. and } (\boldsymbol{V}|\boldsymbol{x}_m, Z) \sim (\boldsymbol{V}|\boldsymbol{x}_0, Z), \tag{2.4}$$

or for Model-NSP,

$$((U, \boldsymbol{V})|\boldsymbol{x}_m, Z) \sim ((U, \boldsymbol{V})|\boldsymbol{x}_0, Z). \tag{2.5}$$

$(\boldsymbol{x}_0, z)$ *and* $(\boldsymbol{x}_m, z')$ *are called a matching pair.*

Equation (2.3) guarantees that the matching pair generate exactly the same partitions on $S(\boldsymbol{V})$. Equation (2.4) and (2.5) imply that U and $\boldsymbol{V}$ have the same level of dependence given $\boldsymbol{X} = \boldsymbol{x}_0$ or $\boldsymbol{x}_m$. A sufficient condition for these two equations is that $(Z, \boldsymbol{X})$ are jointly exogenous. This assumption is actually commonly made in practice, for instance Carneiro, Heckman and Vytlacil (2011). Also, it only needs to be satisfied by the covariates that are used to generate the matching points. Finally, all these conditions are indirectly testable under over-identification, as will be seen in the next section.

From Definition MP, it can be verified that if Assumptions E-SP or Assumptions E-NSP and FS hold at $\boldsymbol{x}_0$, they would also hold at the matching points of $\boldsymbol{x}_0$. The following theorem thus shows that the changes in the outcome functions from $\boldsymbol{x}_0$ to a matching point are identified:

Theorem MEQ (Matching Equation)**.** *Suppose* $\boldsymbol{x}_m \in S(\boldsymbol{X})$ *is a matching point for* $\boldsymbol{x}_0 \in S(\boldsymbol{X})$*, then the following claims hold for all* $d \in S(D)$*:*

(a) Model-SP. Under Assumptions E-SP, $p_d(\boldsymbol{x}_m, z') = p_d(\boldsymbol{x}_0, z)$ *and*

$$m_d^*(\boldsymbol{x}_m) = m_d^*(\boldsymbol{x}_0) + \Big(\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\Big). \tag{2.6}$$

(b) Model-NSP. Under Assumptions E-NSP, FS and CM, $p_d(\boldsymbol{x}_m, z') = p_d(\boldsymbol{x}_0, z)$ *and*

$$F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_m, u)|d, \boldsymbol{x}_m, z') = F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z). \tag{2.7}$$

The matching equation (2.6) directly establishes an identified one-to-one mapping from $\boldsymbol{m}^*(\boldsymbol{x}_0)$ to $\boldsymbol{m}^*(\boldsymbol{x}_m)$.

For the matching equation (2.7), by strict motononicity of the conditional CDFs of Y (implied by Assumptions FS and CM), we have

$$g_d^*(\boldsymbol{x}_m, u) = Q_{Y|D\boldsymbol{X}Z}\Big(F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z)\Big|d, \boldsymbol{x}_m, z'\Big) \equiv \varphi_d(g_d^*(\boldsymbol{x}_0, u); \boldsymbol{x}_m, z') \tag{2.8}$$

where $\varphi_d(\cdot; \boldsymbol{x}_m, z') : S(Y|d, \boldsymbol{x}_0) \mapsto S(Y|d, \boldsymbol{x}_m)$ is continuous and strictly increasing. Later

we may use the shorthand notation $\varphi_d(\cdot)$ for brevity.

Theorem MEQ allows us to condition on $\boldsymbol{X} = \boldsymbol{x}_m$ to help identify $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$. We use Model-SP as an example for illustration. By Proposition 2,

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot \mathbb{E}(Y|d, \boldsymbol{x}_0, z) \tag{2.9}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot \mathbb{E}(Y|d, \boldsymbol{x}_0, z') \tag{2.10}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot m_d^*(\boldsymbol{x}_m) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot \mathbb{E}(Y|d, \boldsymbol{x}_m, z) \tag{2.11}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z') \cdot m_d^*(\boldsymbol{x}_m) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z') \cdot \mathbb{E}(Y|d, \boldsymbol{x}_m, z') \tag{2.12}$$

Substitute equation (2.6) into equations (2.11) and (2.12) for all d. Then (2.12) is redundant with (2.9) as they become identical. With the extra equation (2.11), we end up with three equations and three unknowns; identification becomes possible.

Further, the augmentation of the moment conditions does not necessarily end here. Given $\boldsymbol{x}_m$, one would expect that if it also has a matching point $\boldsymbol{x}'_m \neq \boldsymbol{x}_0$, then the mapping between the outcome functions at $\boldsymbol{x}'_m$ and $\boldsymbol{x}_m$ is identified. Consequently, the mapping between those at $\boldsymbol{x}'_m$ and $\boldsymbol{x}_0$ is identified, too. The following example illustrates this possibility.

Example OC Cont'd. *Under the setup in Example OC, for any fixed $x_0 \in S(X)$, it has the following two matching points by equation (2.3) if they are in $S(X)$:*

$$(z = 0, z' = 1) : \beta x_{m1} + \alpha \cdot 1 = \beta x_0 + \alpha \cdot 0 \implies x_{m1} = x_0 - \frac{\alpha}{\beta} \tag{2.13}$$

$$(z = 1, z' = 0) : \beta x_{m2} + \alpha \cdot 0 = \beta x_0 + \alpha \cdot 1 \implies x_{m2} = x_0 + \frac{\alpha}{\beta} \tag{2.14}$$

Similarly, for x_{m1} and x_{m2}, each of them also has two matching points: One is x_0, and the other is $x_0 - 2\frac{\alpha}{\beta}$ and $x_0 + 2\frac{\alpha}{\beta}$ respectively. This process can be continued until the boundaries of $S(X)$ are reached, illustrated by the following figure:

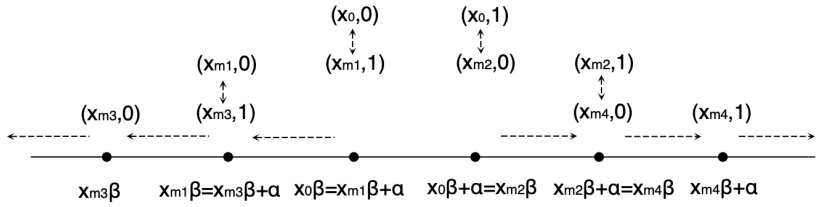

Figure 1: The Pyramid of Matching Points

The horizontal axis is the value of the single index $x\beta + z\alpha$. Starting from $(x_0, 0)$ and $(x_0, 1)$, we obtain x_{m1} and x_{m2} by solving the equations below the horizontal axis. Then we repeat this procedure to match $(x_{m1}, 0)$ with $(x_{m3}, 1)$ and match $(x_{m2}, 1)$ with $(x_{m4}, 0)$. Continuing the process, one can expect to see that the dotted points on this axis extend to both directions, until they reach the boundaries of $S(X)$.

These additional points in Example OC Cont'd are not the matching points of $\boldsymbol{x}_0$. But by recursively applying Theorem MEQ, the outcome functions at these points and at $\boldsymbol{x}_0$ are still linked by identified one-to-one mappings. To formalize this idea, we introduce the following concepts.

Definition MC (M-Connected Set)**.** *A set $\mathcal{X}_{MC}(\boldsymbol{x}_0) \subseteq S(\boldsymbol{X})$ is called the m-connected set of $\boldsymbol{x}_0$ if $\boldsymbol{x}_0 \in \mathcal{X}_{MC}(\boldsymbol{x}_0)$ and for any $\boldsymbol{x} \in \mathcal{X}_{MC}(\boldsymbol{x}_0)$, there exists $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{k(\boldsymbol{x})} \in \mathcal{X}_{MC}(\boldsymbol{x}_0)$ such that $\boldsymbol{x}_j$ is a matching point of $\boldsymbol{x}_{j-1}$, $j = 1, ..., k(\boldsymbol{x})$, and $\boldsymbol{x}$ is a matching point of $\boldsymbol{x}_{k(\boldsymbol{x})}$. Any two points in the m-connected set are said to be m-connected.*

By definition, the m-connected set is the largest subset of $S(\boldsymbol{X})$ such that the outcome functions' relationship at any two elements in it is identified by recursively applying Theorem MEQ. Coupled with $S(Z)$, the set $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ contains every possible value of $(\boldsymbol{X}, Z)$ that may be conditioned on to identify $\boldsymbol{m}^*(\boldsymbol{x}_0)$ or $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$.

3 Identification

In this section we first discuss the existence and identification of the matching points of a given $\boldsymbol{x}_0$. A similar argument holds for other points in $\boldsymbol{x}_0$'s m-connected set. Then we provide sufficient conditions under which the SP- and the NSP-outcome functions are identified.

3.1 The Existence and Identification of the Matching Points

The existence and identification of the matching points are closely related to the selection function $\boldsymbol{h}$ and features of $\boldsymbol{X}$, for example its dimensionality and support. Different ways to find them are available depending on how much we know about $\boldsymbol{h}$,

When the form of $\boldsymbol{h}$ or some of its structures are known, the matching points $\boldsymbol{x}_m$ may be obtained by directly applying the definition: $h_d(\boldsymbol{x}_m, z', \boldsymbol{v}) - h_d(\boldsymbol{x}_0, z, \boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in S(\boldsymbol{V})$. For instance, in Example OC Cont'd, we know D is determined by a single-index ordered choice model. Then the matching points are obtained via equations (2.13) and (2.14) when α and β are identified (up to a multiplicative constant).

When the model that determines D is unknown, as is common in many economic applications, solving equation (2.3) is infeasible. On the other hand, the generalized propensity scores are usually directly identified from the population. Under the exogeneity assumption for Z, a matching point $\boldsymbol{x}_m$ necessarily solves the following equation for $z' \neq z \in S(Z)$:

$$\Big(p_1(\boldsymbol{x}, z') - p_1(\boldsymbol{x}_0, z)\Big)^2 + \Big(p_2(\boldsymbol{x}, z') - p_2(\boldsymbol{x}_0, z)\Big)^2 = 0 \tag{3.1}$$

If the converse is also true, the solutions to equation (3.1) are then the matching points of $\boldsymbol{x}_0$.

Definition PSC (Propensity Score Coherence, PSC)**.** *Suppose* $\boldsymbol{p}(\boldsymbol{x}, z) = \boldsymbol{p}(\boldsymbol{x}', z')$. *The selection model is said to be propensity score coherent at* $(\boldsymbol{x}, z)$ *and* $(\boldsymbol{x}', z')$ *if* $\boldsymbol{h}(\boldsymbol{x}, z, \boldsymbol{V}) = \boldsymbol{h}(\boldsymbol{x}', z', \boldsymbol{V})$ *a.s.*

Note that if $\boldsymbol{h}$ is identified by propensity scores at $(\boldsymbol{x}_0, z)$, that is, there does not exist $(\boldsymbol{x}', z')$ such that $\boldsymbol{p}(\boldsymbol{x}', z') = \boldsymbol{p}(\boldsymbol{x}_0, z)$ but $\boldsymbol{h}(\boldsymbol{x}', z', \boldsymbol{V}) \neq \boldsymbol{h}(\boldsymbol{x}, z, \boldsymbol{V})$ with positive probability, then PSC holds at $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}, z')$ for any $\boldsymbol{x}$ that solves equation (3.1). Many familiar discrete choice models satisfy PSC. We present examples in Appendix C.

In principle, the existence of a solution to equation (3.1) depends on how much $\boldsymbol{X}$, within its support, can affect the propensity scores at $z' \in S(Z)$. For example if the propensity scores at z' have full support, i.e., $(p_1(\cdot, z'), p_2(\cdot, z')) : S(\boldsymbol{X}|z') \mapsto [0,1] \times [0,1]$ is surjective, then a solution always exists. In general, the higher the dimension of $\boldsymbol{X}$ is, the larger its effect on the propensity score, and the larger its support is, the more likely a solution is to exist. For instance, in Example OC, $x_0 \pm \frac{\alpha}{\beta} \in S(X)$ if $S(X)$ is large and/or α/β, i.e., the relative effect of Z with respect to X, is small.

As for the rest of the points in the m-connected set, since each of them is a matching point of some other point, a similar argument applies. In principle, the larger $S(\boldsymbol{X})$ is, the larger the m-connected set is. Recall Figure 1 in Example OC Cont'd, if $S(X) = \mathbb{R}$ and $(U, V) \perp\!\!\!\perp (X, Z)$, then the m-connected set of $\boldsymbol{x}_0$ is countably infinite.

3.2 Identification of the SP-Outcome Functions

Given the m-connected set of $\boldsymbol{x}_0$, we are ready to study the identification of the outcome functions. We begin with the separable model-SP.

For illustrative purposes, let us only consider one matching point $\boldsymbol{x}_m$ of $\boldsymbol{x}_0$ first. Substituting the matching equation (2.6) into (2.11) and (2.12) for each d and deleting the

redundant equation, we can see $\boldsymbol{m}^*(\boldsymbol{x}_0)$ satisfies the following system of equations:

$$\Pi_{SP} \cdot \boldsymbol{m}^*(\boldsymbol{x}_0) = \begin{pmatrix} \sum_{d=1}^3 \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) p_d(\boldsymbol{x}_0, z) \\ \sum_{d=1}^3 \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z') p_d(\boldsymbol{x}_0, z') \\ \underbrace{\sum_{d=1}^3 \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z) p_d(\boldsymbol{x}_m, z)}_{=\sum_{d=1}^3 m_d^*(\boldsymbol{x}_m) p_d(\boldsymbol{x}_m, z)} + \sum_{d=1}^3 \Big[\underbrace{\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z')}_{m_d^*(\boldsymbol{x}_0) - m_d^*(\boldsymbol{x}_m)} \Big] p_d(\boldsymbol{x}_m, z) \end{pmatrix} \tag{3.2}$$

where $\Pi_{SP} = \begin{pmatrix} p_1(\boldsymbol{x}_0, z) & p_2(\boldsymbol{x}_0, z) & p_3(\boldsymbol{x}_0, z) \\ p_1(\boldsymbol{x}_0, z') & p_2(\boldsymbol{x}_0, z') & p_3(\boldsymbol{x}_0, z') \\ p_1(\boldsymbol{x}_m, z) & p_2(\boldsymbol{x}_m, z) & p_3(\boldsymbol{x}_m, z) \end{pmatrix}$.

The first two equations in the system are directly from Proposition 1. In the third equation, we condition on $\boldsymbol{x}_m$ instead of $\boldsymbol{x}_0$; the first term on the right hand side again follows from Proposition 1. The second term, obtained from Theorem MEQ, then accounts for the difference sending $\boldsymbol{m}^*(\boldsymbol{x}_m)$ back to $\boldsymbol{m}^*(\boldsymbol{x}_0)$. Since the system of equations (3.2) is linear in $\boldsymbol{m}^*(\boldsymbol{x}_0)$, it is identified if Π_{SP} is full rank.

More generally, recall the augmented set of conditioning points $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ introduced in Section 2.3. The equation system (3.2) can be easily adapted for any point in $\mathcal{Z}(\boldsymbol{x}_0)$. Then $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is identified if Π_{SP} constructed by any three points in $\mathcal{Z}(\boldsymbol{x}_0)$ is full rank. Further, once $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is identified, $\boldsymbol{m}^*(\cdot)$ at any other points in $\mathcal{X}_{MC}(\boldsymbol{x}_0)$ is also identified.

Theorem ID-SP. *Under Assumptions E-SP, if there exists $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$ such that*

$$\Pi_{SP} = \begin{pmatrix} p_1(\tilde{\boldsymbol{z}}_1) & p_2(\tilde{\boldsymbol{z}}_1) & p_3(\tilde{\boldsymbol{z}}_1) \\ p_1(\tilde{\boldsymbol{z}}_2) & p_2(\tilde{\boldsymbol{z}}_2) & p_3(\tilde{\boldsymbol{z}}_2) \\ p_1(\tilde{\boldsymbol{z}}_3) & p_2(\tilde{\boldsymbol{z}}_3) & p_3(\tilde{\boldsymbol{z}}_3) \end{pmatrix} \text{ is full rank,}$$

then $\boldsymbol{m}^*(\boldsymbol{x})$ is identified for all $\boldsymbol{x} \in \mathcal{X}_{MC}(\boldsymbol{x}_0)$.

Remark 3.2. *Note that the conditioning values in the theorem does not necessarily include $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_0, z')$. For instance, in Example OC Cont'd, they can be any three of the dotted points in Figure 1.*

Since Π_{SP} does not contain any components of $\boldsymbol{m}^*$, whether the full-rank condition holds or not solely depends on the selection model. In what follows, we provide sufficient and necessary conditions for Π_{SP} to be full-rank.

The Full Rank Condition

For simplicity, we go back to the case in the beginning of this section and consider one matching point $\boldsymbol{x}_m$ such that $\boldsymbol{p}(\boldsymbol{x}_m, z') = \boldsymbol{p}(\boldsymbol{x}_0, z)$. Recall that in this case,

$$\Pi_{SP} = \begin{pmatrix} p_1(\boldsymbol{x}_0, z) & p_2(\boldsymbol{x}_0, z) & p_3(\boldsymbol{x}_0, z) \\ p_1(\boldsymbol{x}_0, z') & p_2(\boldsymbol{x}_0, z') & p_3(\boldsymbol{x}_0, z') \\ p_1(\boldsymbol{x}_m, z) & p_2(\boldsymbol{x}_m, z) & p_3(\boldsymbol{x}_m, z) \end{pmatrix}$$

Using the facts that the sum of the three columns in Π_{SP} is equal to vector $\mathbf{1}$, it can be shown that Π_{SP} is full rank if and only if

$$\Big(p_1(\boldsymbol{x}_m, z) - p_1(\boldsymbol{x}_0, z)\Big)\Big(p_3(\boldsymbol{x}_0, z) - p_3(\boldsymbol{x}_0, z')\Big) \neq \Big(p_1(\boldsymbol{x}_0, z) - p_1(\boldsymbol{x}_0, z')\Big)\Big(p_3(\boldsymbol{x}_m, z) - p_3(\boldsymbol{x}_0, z)\Big) \tag{3.3}$$

Inequality (3.3) does not hold if both sides are simultaneously zero. This is the case when Z has no effect on $\boldsymbol{p}$ at $\boldsymbol{X} = \boldsymbol{x}_0$ or $\boldsymbol{X}$ has no effect on $\boldsymbol{p}$ at $Z = z$. Both can be ruled out by a local relevance condition saying that $\boldsymbol{X}$ and Z have nonzero effects on the propensity scores at $(\boldsymbol{x}_0, z)$.

Now suppose neither side is 0. By $\boldsymbol{p}(\boldsymbol{x}_0, z) = \boldsymbol{p}(\boldsymbol{x}_m, z')$, inequality (3.3) can be rewritten as

$$\frac{p_1(\boldsymbol{x}_m, z) - p_1(\boldsymbol{x}_0, z)}{p_3(\boldsymbol{x}_m, z) - p_3(\boldsymbol{x}_0, z)} \neq \frac{p_1(\boldsymbol{x}_m, z') - p_1(\boldsymbol{x}_0, z')}{p_3(\boldsymbol{x}_m, z') - p_3(\boldsymbol{x}_0, z')} \tag{3.4}$$

The inequality generally holds unless the propensity score differences are locally uniform. For example, one can verify that the inequality is satisfied in the ordered choice model in Example OC for almost all x_0 in its support unless V is (locally) uniformly distributed. In particular, it holds for widely applied Logit and Probit models. The following example provides sufficient and necessary conditions for inequality (3.4) in an ordered choice model with multi-dimensional unobservables.

Example OC Cont'd 2. *Suppose now there are two unobservables in the ordered choice model:* $h_1(X, Z, \boldsymbol{V}) = \mathbb{1}(V_1 \leq \kappa_1 + \alpha Z + \beta X)$, $h_3(X, Z, \boldsymbol{V}) = \mathbb{1}(V_2 > \kappa_2 + \alpha Z + \beta X)$, *and* $h_2(X, Z, \boldsymbol{V}) = 1 - h_1(X, Z, \boldsymbol{V}) - h_3(X, Z, \boldsymbol{V})$. *To guarantee* $V_1 < V_2$ *a.s., we assume both are continuously distributed on* $S(V_1) \equiv (-\infty, c]$ *and* $S(V_2) \equiv [c, \infty)$ *respectively where* $c \in \mathbb{R}$. *Finally, assume* $\alpha \cdot \beta \neq 0$ *and we only consider the matching point* $x_m = x_0 - \frac{\alpha}{\beta}$.

Theorem ID-OC (Identification under Example OC Cont'd 2)**.** *Under the setup in Example OC Cont'd 2,* Π_{SP} *is full rank if and only if the single index* $X\beta + Z\alpha$ *evaluated at* $(x_0, 0)$, $(x_0, 1)$ *and* $(x_m, 0)$ *do not all fall into* $S(V_1)$ *or* $S(V_2)$ *at the same time.*

3.3 Identification of the NSP-Outcome Functions

As before, let us start from one matching point $\boldsymbol{x}_m$ such that, $\boldsymbol{p}(\boldsymbol{x}_m, z') = \boldsymbol{p}(\boldsymbol{x}_0, z)$. Similar to Section 3.2, we can substitute equation (2.8) into equation (2.2) for $(\boldsymbol{x}_m, z)$. Then $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ solves the following system of equations for every $u \in [0,1]$:

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) = u \tag{3.5}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z') = u \tag{3.6}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot F_{Y|D\boldsymbol{X}Z}(\varphi_d(g_d^*(\boldsymbol{x}_0, u); \boldsymbol{x}_m, z')|d, \boldsymbol{x}_m, z) = u \tag{3.7}$$

Unlike identification of nonseparable models with a continuous D (e.g. Chernozhukov, Imbens and Newey (2007), Chen et al. (2014)), here we do not face the ill-posed problem due to the discreteness of D.

As the system is nonlinear in finite dimensional unknowns for a fixed u, it is well-known that the Jacobian of the system being full-rank at $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ only implies local identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ (see Chernozhukov and Hansen (2005) and Chen et al. (2014) for examples). In what follows, we show that by continuity and monotonicity of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$, local identification at all $u \in [0,1]$ actually implies global identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ in the class of monotonic functions.

Let us first define a solution path, a concept widely adopted in differential equations.

Definition SolP (Solution Paths)**.** *For a system of equations $\boldsymbol{M}(\boldsymbol{y}, u) = \boldsymbol{0}$, where $\boldsymbol{y}$ is a real vector and $u \in \mathcal{U}$, a solution path $\boldsymbol{y}^*(\cdot)$ is a function on $\mathcal{U}$ such that $\boldsymbol{M}(\boldsymbol{y}^*(u), u) = \boldsymbol{0}$ for all $u \in \mathcal{U}$.*

Stack the left hand side of equations (3.5) to (3.7) into a vector denoted by $\Psi(\boldsymbol{g}^*(\boldsymbol{x}_0, u))$. Denote the vector $(u, u, u)'$ by $\boldsymbol{u}$. Then $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is one solution path to $\boldsymbol{M}(\boldsymbol{y}, \boldsymbol{u}) \equiv \Psi(\boldsymbol{y}) - \boldsymbol{u} = 0$. The pathwise approach we propose focuses on the uniqueness of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ in a class of admissible solution paths $\mathcal{G}$. By Assumption CM, $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is continuous and each component is strictly increasing on $[0,1]$. We thus set $\mathcal{G}$ to be the closure of the set of all such functions:

$$\mathcal{G} \equiv \{\boldsymbol{g} : [0,1] \mapsto \mathbb{R}^3 \text{ and is weakly increasing}\}.$$

Recall that $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ contains all the points that can be conditioned on to identify $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$. Let $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3)$ be the moment equations adapted from equations (3.5) to (3.7) by conditioning on $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$. For example, the k-th component in

Ψ is $\sum_{d=1}^{3} p_d(\tilde{\boldsymbol{z}}_d) \cdot F_{Y|D\boldsymbol{X}Z}(\varphi_d(\cdot)|d, \tilde{\boldsymbol{z}}_k)$. The following theorem provides sufficient conditions that guarantee global identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ on $\mathcal{G}$.

Theorem ID-NSP. *Under Assumptions E-NSP, FS,and CM, if there exist $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$ such that $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3)$ is continuously differentiable on $\prod_{d=1}^{3} S(Y|d, \boldsymbol{x}_0)$, and that its Jacobian matrix at $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$, $\Pi_{NSP}(\boldsymbol{g}^*(\boldsymbol{x}_0, u))$, is full-rank for all $u \in [0, 1]$, then $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is the unique solution path (up to $u = 0, 1$) to $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3) - \boldsymbol{u} = 0$ in $\mathcal{G}$.*

By "unique up to $u = 0, 1$", we mean that if there is another solution path $\tilde{\boldsymbol{g}} \in \mathcal{G}$, it has to equal $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ for all $u \in (0, 1)$. Indeterminacy in the end points arises naturally from the fact that at $u = 0, 1$, any point that lies outside the range of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$, i.e., $\prod_d S(Y|d, \boldsymbol{x}_0)$, trivially satisfies the moment equations. If we restrict the parameter space $\mathcal{G}$ to only contain functions whose ranges are contained in $\prod_d S(Y|d, \boldsymbol{x}_0)$, then uniqueness holds on $[0, 1]$.

Remark 3.3. *By definition, $\varphi_d(g_d^*(\boldsymbol{x}_0, \cdot); \boldsymbol{x}_0, z) = g_d^*(\boldsymbol{x}_0, \cdot)$. So Theorem ID-NSP also applies to the standard IV approach when D is discrete with $|S(Z)| \geq |S(D)|$, for example, Chernozhukov and Hansen (2005).*

Let us sketch the proof to see how monotonicity and continuity of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ convert local identification pointwise in u to global identification. For illustrative purpose, we focus on the subspace $\mathcal{G}^* \subseteq \mathcal{G}$: The range of every function in $\mathcal{G}^*$ is contained in $\prod_{d=1}^{3} S(Y|d, \boldsymbol{x}_0)$. The proof for the larger $\mathcal{G}$ is in Appendix B.

Note that the functions in $\mathcal{G}^*$ may be continuous or discontinuous. We rule out the existence of a hypothetical solution path $\tilde{\boldsymbol{g}}(\cdot) \neq \boldsymbol{g}(\boldsymbol{x}_0, \cdot)$ of each type respectively.

First, suppose there exists a continuous $\tilde{\boldsymbol{g}}(\cdot) \neq \boldsymbol{g}^*(\boldsymbol{x}_0, \cdot) \in \mathcal{G}_0$. By Assumption CM, $\prod_d S(Y|d, \boldsymbol{x}_0)$ is bounded. Thus, $\tilde{\boldsymbol{g}}(\cdot)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ must intersect at $u = 0$ and $u = 1$, equal to the lower or the upper bound of $\prod_d S(Y|d, \boldsymbol{x}_0)$. Hence, the equation $\Delta(u) \equiv \tilde{\boldsymbol{g}}(u) - \boldsymbol{g}^*(\boldsymbol{x}_0, u) = \boldsymbol{0}$ has at least one solution. Let $\bar{u} = \sup_{u'}\{u' : \Delta(u) = \boldsymbol{0}, \forall u \leq u'\}$. By $\tilde{\boldsymbol{g}}(\cdot) \neq \boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$, we have $\bar{u} < 1$. At $\bar{u}$, by the full-rank Jacobian, there exists a neighborhood around $\boldsymbol{g}^*(\boldsymbol{x}_0, \bar{u}) = \tilde{\boldsymbol{g}}(\bar{u})$ in which $\Psi(\cdot)$ is injective. However, by continuity, $\tilde{\boldsymbol{g}}(\cdot)$ must enter this neighborhood from the right of $\bar{u}$. Once it enters the neighborhood, for any $u'' > \bar{u}$, $\Psi(\cdot) = \boldsymbol{u}''$ has two different solutions: $\tilde{\boldsymbol{g}}(u'') \neq \boldsymbol{g}^*(\boldsymbol{x}_0, u'')$, contradicting injectivity.[2] See Figure 2 for illustration. The solid curve is $g_d^*(\boldsymbol{x}_0, \cdot)$ and the dashed curve is the hypothetical $\tilde{g}_d(\cdot)$. The shaded region is where Ψ is injective. The dashed line must enter the shaded region by continuity, and thus local uniqueness is violated.

[2] A similar argument can also be found in Ortega and Rheinboldt (1970), pp. 133-134, Ambrosetti and Prodi (1995), pp. 48-49, and De Marco, Gorni and Zampieri (2014), as an intermediate step to show variants of the Hadamard Theorem.

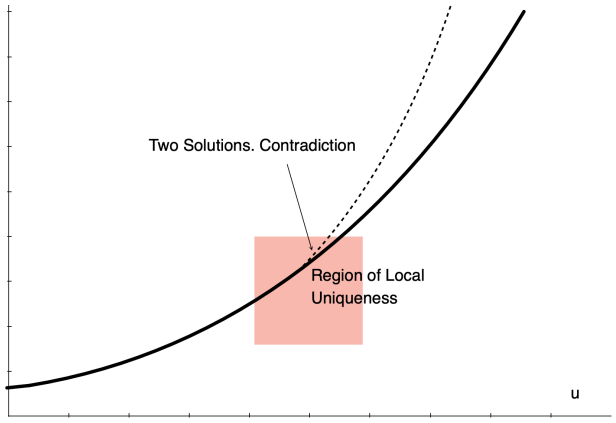

Figure 2: $\Delta(u)$ Continuous at $\bar{u}$

Second, suppose $\tilde{\boldsymbol{g}}(\cdot)$ is discontinuous at $\bar{u}$. As $\tilde{\boldsymbol{g}}(\cdot)$ is increasing, it has to jump up at $\bar{u}$. By construction, since the conditional CDFs and the φ_ds in $\Psi(\cdot)$ are all continuous and strictly increasing, Ψ is continuous and strictly increasing in each argument too. Therefore, Ψ also jumps up at $\bar{u}$. However, the right hand side of the moment conditions, $\boldsymbol{u}$, is continuous at $\bar{\boldsymbol{u}}$, so the equation does not hold. Alternatively, to make the equation hold, there must be a component in $\tilde{\boldsymbol{g}}$ jumping down at $\bar{u}$, but this violates monotonicity. See Figure 3. The two solid curves are $g^*_d(\boldsymbol{x}_0, \cdot)$ and $g^*_{d'}(\boldsymbol{x}_0, \cdot)$ for $d' \neq d$. The dashed curves are hypothetical $\tilde{g}_d$ and $\tilde{g}_{d'}$ after the jump.

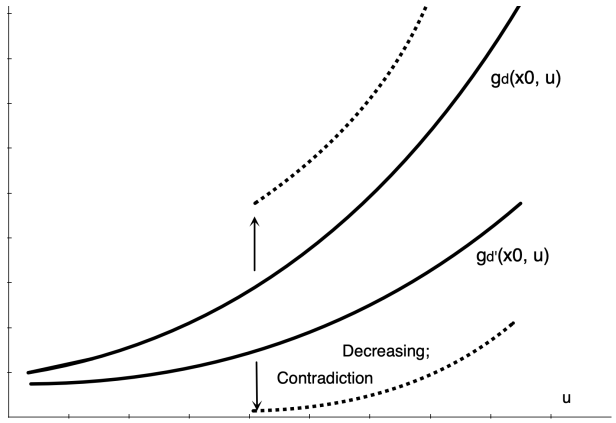

Figure 3: $\Delta(u)$ Not Continuous at $\bar{u}$

Before we close this section, let us emphasize that the identification notion in Theorem ID-NSP is in terms of the uniqueness of monotonic solution path. It does not rule out the possibility that at certain u, the solution to $\Psi(\cdot) = \boldsymbol{u}$ is not unique. This is expected because the conditions we require are much weaker than the sufficient conditions for global invertibility of $\Psi(\cdot)$ on its entire domain $\Pi_{d=1}^{3} S(Y|d, \boldsymbol{x}_0)$ (see variants of Hadamard's theorem,

e.g. Ambrosetti and Prodi (1995)). Under this weaker notion of identification, estimation cannot be conducted for a fixed u; as will be seen in the next section, we will estimate $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ at multiple nodes jointly by imposing monotonicity and assuming the number of the nodes grows to infinity with the sample size.

4 Estimation

In this section, we propose estimators for the matching points and the outcome functions given an independently and identically distributed sample $(Y_i, D_i, \boldsymbol{X}_i, Z_i)_{i=1}^n$. We also discuss some practical issues for implementation.

The estimation strategy follows our constructive identification. From Section 3, the matching points can be obtained by either matching the propensity scores or matching the selection functions, depending on the assumptions made on $\boldsymbol{h}$. Meanwhile, the moment conditions for $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ essentially can be constructed by conditioning on any three values in $\mathcal{Z}(\boldsymbol{x}_0)$. For illustrative purpose, we focus on the following benchmark case to highlight the key features of the estimation procedure.

1. $\boldsymbol{X}$ is one-dimensional, denoted by X.

2. Two matching pairs exist: $(x_0, 0)$, $(x_{m1}, 1)$ and $(x_0, 1)$, $(x_{m2}, 0)$. PSC holds at each pair.

The benchmark conditions setup the simplest scenario while both the matching points (due to Condition 1) and the outcome functions (due to Condition 2) are over-identified, allowing us to construct over-identification tests. Extending Condition 1 to multivariate $\boldsymbol{X}$ is straightforward. Condition 2 is testable by the over-identification test.

The Matching Points

Let $\hat{\boldsymbol{p}}(\cdot, z)$, $z = 0, 1$, be a consistent estimator of $\boldsymbol{p}(\cdot, z)$ uniformly on $S_0(X)$, a compact interior subset of $S(X)$. We assume both matching points are in it. Let $\Delta\hat{\boldsymbol{p}}(x_1, x_2) \equiv \big(\hat{p}_1(x_1, 1) - \hat{p}_1(x_0, 0), \hat{p}_2(x_1, 1) - \hat{p}_2(x_0, 0), \hat{p}_1(x_2, 0) - \hat{p}_1(x_0, 1), \hat{p}_2(x_2, 0) - \hat{p}_2(x_0, 1)\big)'$. Finally for some weighting matrix $\boldsymbol{W}_{xn}$ with positive definite probability limit, let $\hat{Q}_x(x_1, x_2) \equiv \Delta\hat{\boldsymbol{p}}(x_1, x_2)'\boldsymbol{W}_{xn}\Delta\hat{\boldsymbol{p}}(x_1, x_2)$. Under PSC, the estimator $(\hat{x}_{m1}, \hat{x}_{m2})$ we propose are points in $S_0^2(X)$ such that for some $a_n = o(1)$,

$$\hat{Q}_x(\hat{x}_{m1}, \hat{x}_{m2}) \leq \inf_{S_0^2(X)} \hat{Q}_x(x_1, x_2) + a_n \tag{4.1}$$

When $a_n = 0$, $(\hat{x}_{m1}, \hat{x}_{m2})$ is the minimizer of $\hat{Q}_x(x_1, x_2)$. In general, the minimizer of $\hat{Q}_x(x_1, x_2)$ is consistent of (x_{m1}, x_{m2}) only if the latter is the unique minimizer of the population objective function Q_x. When $Q_x(\cdot, \cdot)$ has multiple minima, which is allowed for the purpose of identification, the set of the minimizers of $\hat{Q}_x$ tends to be smaller than the true set, and its probability limit may not exist. For example, suppose Q_x has two global minima on $S_0^2(X)$ but $\hat{Q}_x$ may only have one. As $n \to \infty$, the minimum of $\hat{Q}_x$ may jump across the neighborhoods of the two minima of Q_x. The probability limit of being in any one particular neighborhood may thus be strictly smaller than one. To handle the general multiple minima case, we let $a_n > 0$ and converge to 0 at an appropriate rate similar to Chernozhukov, Hong and Tamer (2007). We discuss the general case in Appendix A.2. For simplicity, here we focus on the case where (x_{m1}, x_{m2}) is unique and let $a_n = 0$.

For concreteness, we consider the following kernel estimator for the propensity scores. We can use other nonparametric estimators for conditional probability too.

$$\hat{p}_d(x, z) = \frac{\sum_{i=1}^N \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_x}) \mathbb{1}(Z_i = z)}{\sum_{i=1}^N K(\frac{X_i - x}{h_x}) \mathbb{1}(Z_i = z)} \tag{4.2}$$

where $K(\cdot)$ is a kernel function and h_x is the bandwidth converging to 0. Regularity conditions for K and the convergence rate of h_x will be given in Section 6.

The SP-Outcome Functions

For the SP-outcome functions, by linearity of the moment conditions (3.2), we obtain the following closed-form estimator by inverting the estimated Π_{SP} matrix (weighted by $\boldsymbol{W}_{mn}$):

$$\hat{\boldsymbol{m}}(x_0) = (\widehat{\Pi}_{SP}' \boldsymbol{W}_{mn} \widehat{\Pi}_{SP})^{-1} \cdot \widehat{\Pi}_{SP}' \boldsymbol{W}_{mn} \widehat{\Phi}(\hat{x}_{m1}, \hat{x}_{m2}) \tag{4.3}$$

where in this case $\widehat{\Pi}_{SP} = \begin{pmatrix} \hat{p}_1(x_0, 0) & \hat{p}_2(x_0, 0) & \hat{p}_3(x_0, 0) \\ \hat{p}_1(x_0, 1) & \hat{p}_2(x_0, 1) & \hat{p}_3(x_0, 1) \\ \hat{p}_1(\hat{x}_{m1}, 0) & \hat{p}_2(\hat{x}_{m1}, 0) & \hat{p}_3(\hat{x}_{m1}, 0) \\ \hat{p}_1(\hat{x}_{m2}, 1) & \hat{p}_2(\hat{x}_{m2}, 1) & \hat{p}_3(\hat{x}_{m2}, 1) \end{pmatrix}$ and

$$\widehat{\Phi}(\hat{x}_{m1}, \hat{x}_{m2}) = \begin{pmatrix} \sum_{d=1}^3 \widehat{\mathbb{E}}_{Y|DXZ}(d, x_0, 0) \hat{p}_d(x_0, 0) \\ \sum_{d=1}^3 \widehat{\mathbb{E}}_{Y|DXZ}(d, x_0, 1) \hat{p}_d(x_0, 1) \\ \sum_{d=1}^3 \left[\widehat{\mathbb{E}}_{Y|DXZ}(d, \hat{x}_{m1}, 0) + \widehat{\mathbb{E}}_{Y|DXZ}(d, x_0, 0) - \widehat{\mathbb{E}}_{Y|DXZ}(d, \hat{x}_{m1}, 1) \right] \hat{p}_d(\hat{x}_{m1}, 0) \\ \sum_{d=1}^3 \left[\widehat{\mathbb{E}}_{Y|DXZ}(d, \hat{x}_{m2}, 1) + \widehat{\mathbb{E}}_{Y|DXZ}(d, x_0, 1) - \widehat{\mathbb{E}}_{Y|DXZ}(d, \hat{x}_{m2}, 0) \right] \hat{p}_d(\hat{x}_{m2}, 1) \end{pmatrix}.$$

The estimated propensity scores in $\widehat{\Pi}_{SP}$ and $\widehat{\Phi}$ follow equation (4.2). The conditional expectations are estimated by the standard Nadaraya-Waston estimator:

$$\widehat{\mathbb{E}}_{Y|DXZ}(d,x,z) = \frac{\sum_{i=1}^{N} Y_i \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_m}) \mathbb{1}(Z_i = z)}{\sum_{i=1}^{N} \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_m}) \mathbb{1}(Z_i = z)} \tag{4.4}$$

The NSP-Outcome Functions

In the benchmark case, we have the following moment functions:

$$\Psi(\boldsymbol{g}(u)) = \begin{pmatrix} \sum_{d=1}^{3} p_d(x_0, 0) \cdot F_{Y|DXZ}(g_d(u)|d, x_0, 0) \\ \sum_{d=1}^{3} p_d(x_0, 1) \cdot F_{Y|DXZ}(g_d(u)|d, x_0, 1) \\ \sum_{d=1}^{3} p_d(x_{m1}, 0) \cdot F_{Y|DXZ}(\varphi_d(g_d(u); x_{m1}, 1)|d, x_{m1}, 0) \\ \sum_{d=1}^{3} p_d(x_{m2}, 1) \cdot F_{Y|DXZ}(\varphi_d(\tilde{g}_d(u); x_{m2}, 0)|d, x_{m2}, 1) \end{pmatrix}.$$

Let $\boldsymbol{u} = (u, u, u)'$ and $Q_{NSP}(\boldsymbol{g}, u) \equiv \Big[\Big(\Psi(\boldsymbol{g}(u)) - \boldsymbol{u}\Big)' \boldsymbol{W}_g(u) \Big(\Psi(\boldsymbol{g}(u)) - \boldsymbol{u}\Big)\Big]$ for any positive definite matrix $\boldsymbol{W}_g(u)$. Our identification result for $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ implies that it is the unique minimizer (up to $u = 0, 1$) to the following minimization problem:

$$\min_{\boldsymbol{g} \in \mathcal{G}_0} \int_0^1 Q_{NSP}(\boldsymbol{g}(u), u) du \tag{4.5}$$

where $\mathcal{G}_0 \subseteq \mathcal{G}$ contains increasing functions on $[0,1]$ with ranges contained in $\prod_{d=1}^{3} S(Y|d)$.

Construct $\hat{Q}_{NSP}(\boldsymbol{g}, u)$ by plugging in estimators of Ψ and $\boldsymbol{W}_g(u)$. Let $u_j = \frac{j}{J}$ where $1 \leq j \leq J$ and $J \to \infty$. We estimate $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ by solving the following minimization problem:

$$\min_{\underline{\boldsymbol{y}} \leq \boldsymbol{g}(u_1) \leq \ldots \leq \boldsymbol{g}(u_J) \leq \bar{\boldsymbol{y}}} \frac{1}{J} \sum_{j=1}^{J} \hat{Q}_{NSP}(\boldsymbol{g}(u_j), u_j) + \lambda \sum_{j=2}^{J} \Big(\boldsymbol{g}(u_j) - \boldsymbol{g}(u_{j-1})\Big)' \Big(\boldsymbol{g}(u_j) - \boldsymbol{g}(u_{j-1})\Big) \tag{4.6}$$

Let us begin with the constraint. By connecting $\boldsymbol{g}$ at adjacent nodes with line segment, the constraint induces a finite dimensional sieve space $\widehat{\mathcal{G}}$ of piecewise affine increasing functions defined on $[0,1]$. By sending $J \to \infty$, elements in the sieve space are able to approximate any continuous and increasing functions that are bounded by $\underline{\boldsymbol{y}}$ and $\bar{\boldsymbol{y}}$, the lower and upper bounds of $\prod_d S(Y|d)$. As D is discrete, for each d the bounds can be estimated by $\underline{y}_d = \min(Y_i|D_i = d)$ and $\bar{y}_d = \max(Y_i|D_i = d)$. We treat the bounds as known parameters as these estimators converge faster than the nonparametric rate[3].

[3]Alternatively, we could shrink $\mathcal{G}_0$ so that only functions bounded within $\prod_d S(Y|d, x_0)$ are included and $\underline{\boldsymbol{y}}$ and $\bar{\boldsymbol{y}}$ are boundaries estimators for this smaller support set. In this space $\boldsymbol{g}^*(x_0, \cdot)$ is unqiue including at the end points. Also, as will be seen in Section 6.3, nicer boundary properties can be obtained. However,

The second term in equation (4.6) is a penalty making the estimator smoother in finite samples. We let $\lambda \to 0$ fast enough so the the penalty does not affect the estimator's asymptotic behavior.

As for $\widehat{\Psi}$, the conditional CDFs are estimated by the following smoothed kernel estimator (e.g. Hansen (2004) and Li and Racine (2008)):

$$\widehat{F}_{Y|DXZ}(y|d,x,z) = \frac{\sum_{i=1}^{N} L(\frac{y-Y_i}{h_0})\mathbb{1}(D_i=d)K(\frac{X_i-x}{h_g})\mathbb{1}(Z_i=z)}{\sum_{i=1}^{N} \mathbb{1}(D_i=d)K(\frac{X_i-x}{h_g})\mathbb{1}(Z_i=z)} \tag{4.7}$$

where $L(\cdot)$ is a smooth CDF supported on a bounded interval and $h_0 \mapsto 0$ faster than h_g. Another component in $\widehat{\Psi}$ is the function $\hat{\varphi}_d$:

$$\hat{\varphi}_d(y;x_m,z') = \arg\min_{y' \in [\underline{y}_d, \bar{y}_d]} \left(\widehat{F}_{Y|DXZ}(y'|d,\hat{x}_m,z') - \widehat{F}_{Y|DXZ}(y|d,x_0,z)\right)^2 \tag{4.8}$$

Remark 4.1. *Under pointwise identification, estimation can be simplified; one can minimize $\hat{Q}_{NSP}(\boldsymbol{g}(u),u)$ at each u of interest separately (e.g. Lewbel (2007)). The inequality constraints can be dropped. The dimension of each individual minimization problem is smaller. Computation is thus made easier. Under pathwise identification, joint estimation under the constraint of monotonicity is necessary because it is possible that $\boldsymbol{g}^*(\boldsymbol{x}_0,u)$ is not the unique solution to the moment equations for some u. The minimizers of $\hat{Q}_{NSP}(\boldsymbol{g}(u),u)$ at these u are then inconsistent.*

5 Empirical Applications

Before we move into the asymptotic theory of the estimators, we consider two applications to illustrate the usefulness and limitation of our approach. The first application is the return to education example we discussed earlier. The second studies preschool program selection using the administrated Head Start Impact Study (HSIS) dataset.

5.1 The Return to Schooling: A Binary D

We use the same extract from the 1979 NLS as in Card (1995). The outcome variable Y is the log wage. We adopt the same IV which indicates whether an individual grew up near an accredited four-year college. In this subsection, we assume that the latent selection mechanism yields two outcomes: $D = 1$ if the years of schooling is greater than 12 and

since X is continuous, boundary estimators of $\prod_d S(Y|d,x_0)$ (e.g. Guerre, Perrigne and Vuong (2000)) involve extra tuning parameters. For simplicity, we do not adopt this approach.

$D = 0$ otherwise. We will consider a three-valued D in the next subsection. We use the average across parents' years of schooling as the matching covariate X. Finally, we drop the observations who were still enrolled in a school at the time of the survey. The remaining sample size is 2000.

We assume the log wage is determined by Model-SP. As $\boldsymbol{m}^*$ is identified by the standard IV approach with the binary Z, we can compare the results using the standard IV method and our approach.

No Covariates

Let us first consider the following case assuming no covariates are in the outcome functions:

$$Y = \sum_{d=0}^{1} \mathbb{1}(D = d) m_d^* + U.$$

This model sets up a clean benchmark because m_0^* and m_1^* are identified by Z so no extra steps for propensity score matching are needed. In fact they can be estimated by the simple Wald estimator. The results provide us with references about the magnitudes of the outcome function and the effects. As shown in Table 1 (standard errors in parentheses), the return to education is increasing in D. The wage for individuals receiving post-high school education is on average 1.35% higher than those with at most high school education.

Table 1: IV Estimates

$\hat{m}_0$	5.58
	(0.18)
$\hat{m}_1$	6.93
	(0.16)

Covariates and Matching

Let us first illustrate the process of finding a matching point. Figure 4 depicts $\hat{p}_0(x, 0)$ and $\hat{p}_0(x, 1)$. The black dashed lines illustrate the how we find $\hat{x}_{m1}$: For the fixed x_0 in the left panel, we find the value of the propensity score $\hat{p}_0(x_0, 0)$, and find $\hat{x}_{m1}$ in the right panel such that $\hat{p}_0(\hat{x}_{m1}, 1) = \hat{p}_0(x_0, 0)$. Similarly, the blue dash-dot line starts from x_0 in the right panel for $Z = 1$, and the second matching point $\hat{x}_{m2}$ is found in the left panel. In Figure 5, the red solid curves in the left and right panels are $\hat{p}_0(x, 1) - \hat{p}_0(12, 0)$ and $\hat{p}_0(x, 0) - \hat{p}_0(12, 1)$ respectively. These propensity score differences clearly intersect with zero. The intersection points are the estimated matching points. The patterns for other values of x_0 we consider are similar and are thus omitted.

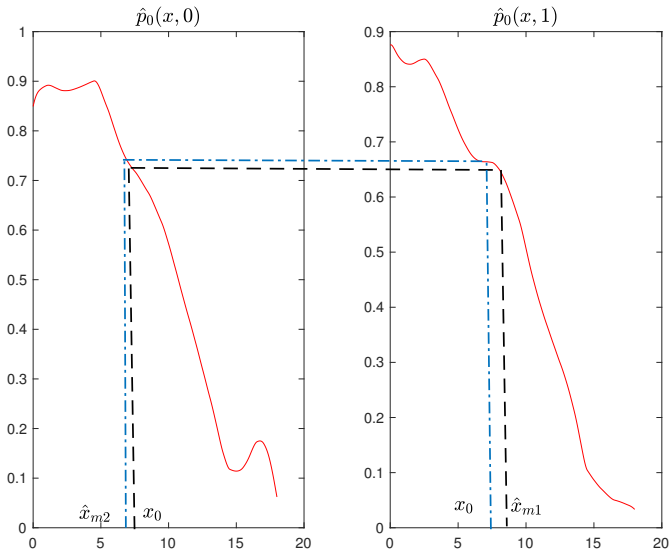

Figure 4: Propensity Scores: $|S(D)| = 2$

Figures 4 and 5 imply that individuals whose parents have more years of schooling are more likely to attain post-high school education. Also, from the values of the matching points, living close to a four-year college and parents education are close substitutes. At $X = 12$, an increase of about half a year in parents' education compensates for not living near a college.

Now let us turn to the outcome function estimates at $x_0 = 10, 11, 12$, shown in Table 2. The second row "Matching" indicates whether the matching points are estimated and used. When not using the matching points, we estimate $(m_0^*(x_0), m_0^*(x_1))$ by inverting the following moment conditions:

$$\begin{pmatrix} \hat{p}_0(x_0,0), \hat{p}_1(x_0,0) \\ \hat{p}_0(x_0,1), \hat{p}_1(x_0,1) \end{pmatrix} \begin{pmatrix} \hat{m}_0(x_0) \\ \hat{m}_1(x_0) \end{pmatrix} = \begin{pmatrix} \hat{p}_0(x_0,0)\hat{\mathbb{E}}_{Y|DXZ}(0,x_0,0) + \hat{p}_1(x_0,0)\hat{\mathbb{E}}_{Y|DXZ}(1,x_0,0) \\ \hat{p}_0(x_0,1)\hat{\mathbb{E}}_{Y|DXZ}(0,x_0,1) + \hat{p}_1(x_0,1)\hat{\mathbb{E}}_{Y|DXZ}(1,x_0,1) \end{pmatrix}.$$

When we use our approach, we first estimate the matching points by grid search over 500 nodes. The standard errors in parentheses are computed using the asymptotic variance estimators derived in the next section. Note that with the matching points, the model is over-identified because we have four moment conditions and two unknowns. Hence, we can perform the over-identification test with the null hypothesis that all the moment conditions are valid. The test is as in the standard GMM framework and we will provide details in Section 6.2. The p-values of the test results are in the last row.

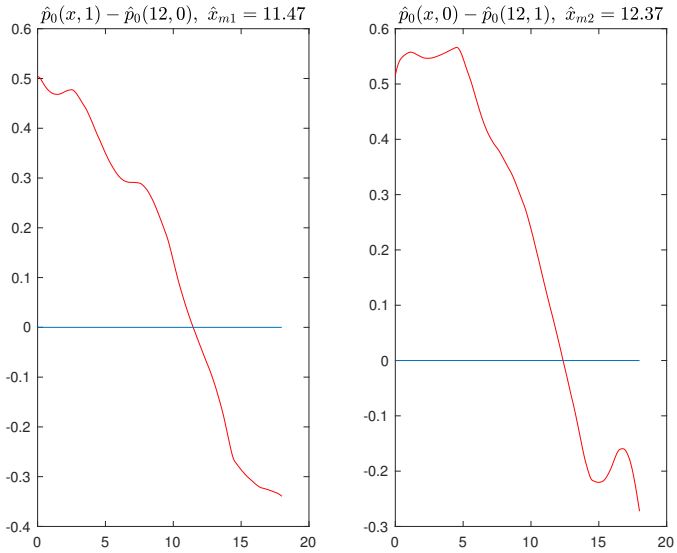

Figure 5: Propensity Score Differences: $x_0 = 12$, $|S(D)| = 2$

Table 2: Binary D with X

	$x_0 = 10$		$x_0 = 11$		$x_0 = 12$	
Matching:	✗	✓	✗	✓	✗	✓
$\hat{m}_0(x_0)$	5.63	5.64	5.59	5.56	5.35	5.37
	(0.28)	(0.17)	(0.33)	(0.20)	(0.60)	(0.33)
$\hat{m}_1(x_0)$	7.15	7.13	6.90	6.92	6.90	6.89
	(0.31)	(0.19)	(0.25)	(0.15)	(0.32)	(0.18)
Over-Id p value	N.A.	0.98	N.A.	0.99	N.A.	0.80

From Table 2, we can make three observations. First, the estimates using the two approaches are very close, but the variances are lower using our approach. Similar point estimates provide extra evidence in addition to the insignificant over-identification test statistics that the extra moment conditions brought in by the matching points are valid. Variance reduction is due to the use of more moment conditions. Consequently, the estimated effects are more significant. For instance, it can be computed that $\hat{m}_1(12) - \hat{m}_0(12)$ is significant at 10% level using the IV approach, but is significant at 1% level using our approach. Second, the outcome function is increasing in the level of education and heterogeneous in parents' education. Individuals with less educated parents have higher returns of education at both levels. Finally, for each level of education, the range of the heterogeneous estimates cover the results in Table 1, indicating the results in Table 2 are in a reasonable range.

5.2 The Return to Schooling: A Three-Valued D

In this subsection, we assume the underlying selection model yields three outcomes; we recode high school education by $D = 1$ and divide post-high school education into two groups: $D = 2$ if $12 <$ years of schooling ≤ 15 (some college), and $D = 3$ if years of schooling > 15 (college and above). In this case, no existing method can identify and estimate $\boldsymbol{m}^*(x_0)$ without imposing extra structures on it.

Again, let us first illustrate the finding of a matching point. Figure 6 depicts the propensity score functions at $Z = 0$ and $Z = 1$. Starting from x_0 on the left panel, we need to match both $\hat{p}_1(x_0, 0)$ and $\hat{p}_3(x_0, 0)$ with $\hat{p}_1(x, 1)$ and $\hat{p}_3(x, 1)$ at the same x. If such x exists, it is the estimated matching point $\hat{x}_{m1}$. Evidently, with a scalar X, x_{m1} is over-identified, so in the finite sample, it is very likely that we cannot exactly match both propensity scores, but we need the difference to be small enough. In Section 6.1, we propose an over-identification test to see whether all the propensity scores can be indeed matched with a single covariate.

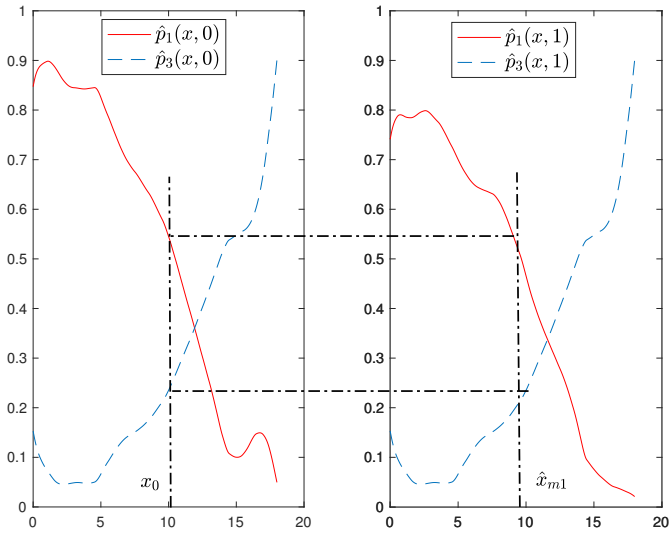

Figure 6: Propensity Scores: $|S(D)| = 3$

Figure 7 illustrates the matching points for $x_0 = 12$. Again we omit other values of x_0 as the patterns are similar. The solid red curves in the two panels in Figure 7 are $\hat{p}_1(x, 1) - \hat{p}_1(x_0, 0)$ and $\hat{p}_1(x, 0) - \hat{p}_1(x_0, 1)$, while the dashed blue curves are $\hat{p}_3(x, 1) - \hat{p}_3(x_0, 0)$ and $\hat{p}_3(x, 0) - \hat{p}_3(x_0, 1)$. Matching is successful if the solid curve and the dashed curve intersect with the horizontal line of zero at the same point. From the figure, the intersection points are indeed very close in both panels. This is also supported by the over-identification tests; $\mathcal{J}_{x1}$ and $\mathcal{J}_{x2}$ on top are insignificant in both cases. Finally, since the baseline level here (years of schooling ≤ 12) is the same as that in the previous case, the propensity scores at

the baseline level of D are equal. Hence, the estimated matching points in these two cases should be similar. Here the estimates are 11.54 and 12.34, indeed very close to those when D is binary (11.47 and 12.37).

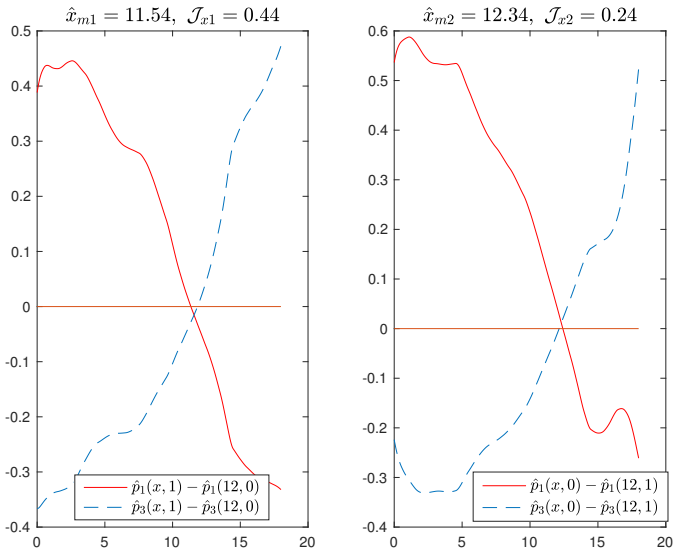

Figure 7: Propensity Score Differences: $x_0 = 12$, $|S(D)| = 3$

Next, let us turn to the estimates of the outcome function shown in Table 3. Since both the outcome function and the matching points are over-identified in this case, we present the p-values for each over-identification test in the bottom panel. First, we can see that none of the over-identification test statistics $\boldsymbol{m}^*(x_0)$ are significant at any reasonable level, similar to Table 2 in the binary case. Also, the joint over-identification tests for the matching points are also insignificant, confirming that all the propensity scores are matched. Second, the return to education is monotonic and the marginal return is slightly decreasing. Third, the returns are heterogeneous in parents education; similar to the binary case, individuals whose parents are less educated have slightly higher returns of education at almost all levels.

Table 3: Three-valued D

	$x_0 = 10$	$x_0 = 11$	$x_0 = 12$
$\hat{m}_1(x_0)$	5.62	5.56	5.33
	(0.23)	(0.22)	(0.38)
$\hat{m}_2(x_0)$	7.03	6.47	6.39
	(3.18)	(1.39)	(1.45)
$\hat{m}_3(x_0)$	7.28	7.32	7.31
	(2.72)	(1.09)	(1.00)
Over-Id p-value			
$\boldsymbol{m}^*(x_0)$	0.89	0.87	0.58
$\boldsymbol{x}_m$	0.54	0.36	0.41

5.3 Validity of the Exogeneity Assumption

In this subsection, we continue with this empirical example to show that in fact the covariate we choose may not be exogenous in the sense of $\mathbb{E}_{U|XZ}(X, Z) \neq 0$, so methods such as 2SLS using for example ZX as an extra instrument may not deliver correct estimates, even though they do not rely on the selection model. In contrast, our approach only imposes local exogeneity assumptions with respect to U and $\boldsymbol{V}$; though stronger than the standard nonparametric IV approach, we can still obtain informative results.

The conditional mean independence $\mathbb{E}_{U|XZ}(X, Z) = 0$ requires that $\mathbb{E}_{U|XZ}(x_0, Z) = 0$ for almost all $x_0 \in S(X)$. Since the latter equation is also required by our approach for fixed x_0, we can use our over-identification test to check if there are values of x_0 such that the condition may not hold. Specifically, we re-estimate the outcome function at $x_0 = 8$ and $x_0 = 14$, for both the binary D and the three-valued D. In each case, we conduct the over-identification test for the outcome function, and for three-valued D, we also conduct the test for the matching points as they are also over-identified then. The results are in Table 4.

Table 4: Values of x_0 Where Exogeneity May Fail

Over-Id	$x_0 = 8$		$x_0 = 14$	
p-value	$\lvert S(D)\rvert = 2$	$\lvert S(D)\rvert = 3$	$\lvert S(D)\rvert = 2$	$\lvert S(D)\rvert = 3$
$\boldsymbol{m}^*(x_0)$	0.05	0.01	0.01	0.00
$\boldsymbol{x}_m$	N.A.	0.23	N.A.	0.14

The results imply that not all the moment conditions are valid at these two values of X, although matching is still successful. It suggests that the invalidity of the moment conditions is more likely to be driven by violations of the exogeneity assumption at these choices of x_0.

For verification, now let us turn to the 2SLS estimates under the conditional mean independence assumption. We consider the following two specifications:

$$\text{Setup 1: } Y = \beta_0 + \mathbb{1}(D=2)\beta_1 + \mathbb{1}(D=3)\beta_2 + X\beta_3 + U$$
$$\text{Setup 2: } Y = \beta_0 + \mathbb{1}(D=2)\beta_1 + \mathbb{1}(D=3)\beta_2 + X\beta_3 + \mathbb{1}(D=2)X\beta_4 + \mathbb{1}(D=3)X\beta_5 + U$$

Under $\mathbb{E}_{U|XZ}(X, Z) = 0$, polynomials of X and their interactions with Z are all valid IVs by the law of iterated expectation. We present the results for $x_0 = 12$ in Table 5. The estimates in Column (1) are obtained using our approach. Note that we do not utilize the parametric form, so they are the same as in Table 5. Columns (2)-(5) contain the fitted values for $x_0 = 12$ using the 2SLS estimates. Estimates under Setup 1 are in Columns (2)-(4), using (Z, ZX), (Z, ZX, X^2) and (Z, ZX, X^2, ZX^2) as instruments respectively. Column (5) contains results under Setup 2, using (Z, ZX, X^2, ZX^2) as instruments. The last row reports the p-values of the over-identification tests for our approach and for the standard 2SLS.

Table 5: Comparison with 2SLS

	MP	2SLS			
	(1)	(2)	(3)	(4)	(5)
$\hat{m}_1(12)$	5.33	4.87	5.63	5.76	5.35
	(0.38)	(0.71)	(0.37)	(0.29)	(0.80)
$\hat{m}_2(12)$	6.39	7.24	7.56	7.34	5.20
	(1.45)	(0.73)	(0.54)	(0.41)	(2.10)
$\hat{m}_3(12)$	7.31	7.08	6.24	6.26	8.24
	(1.00)	(0.63)	(0.17)	(0.14)	(2.00)
Over-Id p-value	0.58	N.A.	0.07	0.07	N.A.

Table 5 shows that the 2SLS over-identification tests, when available, are indeed significant at 10% level, rejecting the null hypothesis that all these instruments are valid, which in turn rejects the conditional mean independence assumption $\mathbb{E}_{U|XZ}(X, Z) = 0$. From the estimates, the results from the 2SLS are misleading: they suggest that the return to education is not monotonic. In Columns (2)-(4), it first increases then decreases in the level of education. In Column (5), it decreases first and then increases to a level that is outside the range of Y in the sample ($\max(Y_i) = 7.78$). In contrast, the over-identification test is insignificant in our approach because we only need it to hold locally in x_0, and our results are consistent with the literature of return to education.

This example shows that although our approach relies on stronger assumptions than the standard IV approach, when the latter is not possible due to the failure of the order condition,

our approach may still obtain informative results. By contrast, alternative approaches that make stronger assumptions on exogeneity with respect to the outcome heterogeneity may not work well in some applications.

5.4 When Does Matching Fail?

In this section we illustrate two possibilities where a matching point does not exist. The first case is that the covariate only matches one propensity score at a time. We use the IQ score in place of parents' education for illustration. The second possibility is that the IV has dominant effects on the propensity scores such that the covariates cannot compensate for the shift in the IV. Using the HSIS dataset, we illustrate it by examining the impact of a randomly assigned lottery granting access to the Head Start preschool program compared to the impacts of other covariates.

Covariates Too Few

Recall that when D is three-valued, there are two propensity scores to be matched. One covariate may fail to match both even if it has large effects on each of them.

For illustration, we keep the setup in Section 5.2 but replace parents' education with the IQ score; IQ is a reasonable candidate for the matching point because it is likely to affect both an individual's educational attainment and her wage. However, as shown in the following figure, it is unable to generate a matching point that match all the propensity scores.

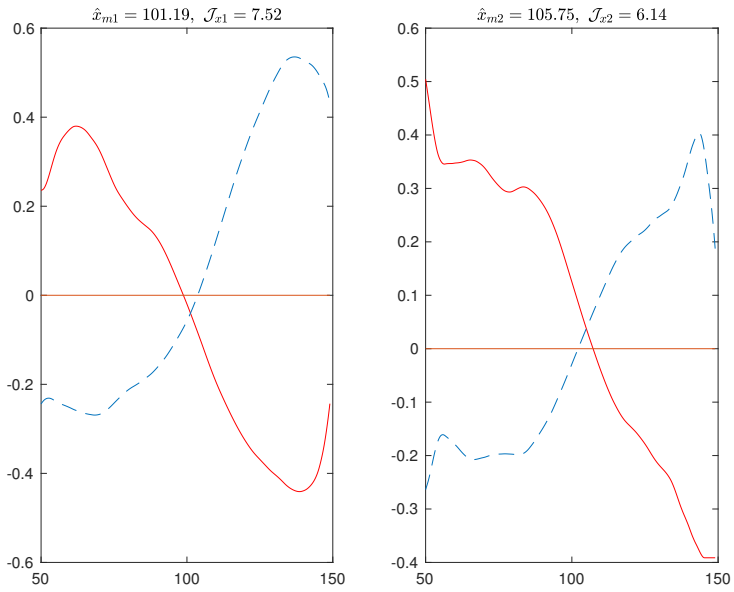

Figure 8: Propensity Score Differences: $X = \text{IQ}$, $x_0 = \text{med(IQ)}$

In Figure 8, the solid red curves are $\hat{p}_1(x,1) - \hat{p}_1(x_0,0)$ and $\hat{p}_1(x,0) - \hat{p}_1(x_0,1)$, and the dashed blue curves are $\hat{p}_3(x,1) - \hat{p}_3(x_0,0)$ and $\hat{p}_3(x,0) - \hat{p}_3(x_0,1)$. All the four curves indeed intersect with 0, so a solution does exist for each propensity score matching equation. The problem is that the intersection points do not coincide, so the two propensity score differences cannot be 0 at the same time. Indeed, the individual over-identification test statistics reject the null that both propensity scores are matched at 1% and 5% level. This type of matching failure is likely to be resolved by using more covariates that also have large effects on the propensity scores for matching.

Covariates Too Weak

Another reason for matching failure is that the effects of Z on the propensity scores dominate those of X, making it difficult for the covariate to compensate the change in Z within its support. The extreme of this scenario is that no covariates enter the selection model, and matching points obviously do not exist. For illustration, let us consider an application on preschool program selection, following Kline and Walters (2016) using the HSIS dataset. D takes on three values: participating in Head Start (h), participating in another competing preschool program (c), and not participating in any preschool programs (n). The binary instrument Z is a lottery granting access to Head Start. Available candidates for X are family income, baseline test scores and the centers' quality index.

Figure 9 shows that the propensity scores with $X =$ the baseline test score and $x_0 =$ the sample median. Patterns for other values of X and other covariates are similar. We see that

when an individual won the lottery, the probability attending the Head Start program is very high, and not much affected by the baseline scores. On the contrary, when not winning the lottery, she would most likely not participate in any programs, and in particular, the probability of attending the Head Start is lowest for almost any baseline scores. Apparently a matching point does not exist in this example because varying X never offsets the dominant effect of Z on the propensity scores.

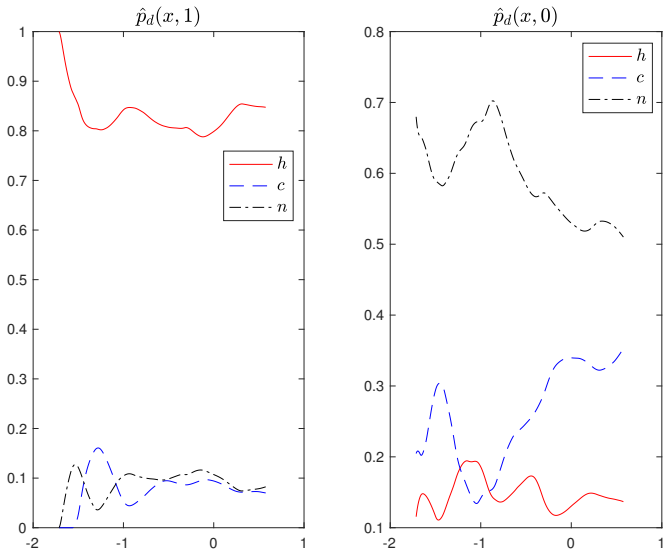

Figure 9: Propensity Scores of Different Preschool Program Choices

6 Asymptotic Properties

In this section we present the asymptotic properties of the estimators. We also discuss inference procedures and develop some specification tests.

6.1 The Matching Points

We start with consistency and asymptotic normality of $(\hat{x}_{m1}, \hat{x}_{m2})$. Recall that we focus on the simple benchmark case where (x_{m1}, x_{m2}) is unique and $a_n = 0$ in equation (4.1). The asymptotic property for the general case is presented in Appendix A.2. We make the following regularity conditions.

Assumption Reg-MP. *For every $d \in S(D)$ and $z \in \{0,1\}$, $p_d(\cdot, z)$ is twice continuously differentiable on $S_0(X)$ with bounded derivatives. The density of X exists and is bounded away from 0 on $S_0(X)$.*

Assumption Reg-K. *The kernel $K(\cdot)$ is symmetric at 0 with finite second moment and twice continuously bounded derivatives on $[-1,1]$.*

Under the regularity conditions, $\hat{Q}_x$ is uniformly consistent for Q_x. Consistency and asymptotic normality follow from the standard argument for GMM estimators.

Theorem Cons-MP. *Under Assumptions Reg-MP, Reg-K and the benchmark conditions, $|(\hat{x}_{m1}, \hat{x}_{m2}) - (x_{m1}, x_{m2})| = o_p(1)$.*

Denote the gradient of $\Delta\boldsymbol{p}(x_1, x_2)$ evaluated at x_{m1} and x_{m2} by $\partial_{x'}\Delta\boldsymbol{p}(x_{m1}, x_{m2})$. Let $\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_4$ be $(x_0, 0)$, $(x_0, 1)$, $(x_{m1}, 1)$ and $(x_{m2}, 0)$ respectively.

Theorem AsymDist-MP. *Under the conditions in Theorem Cons-MP, if (x_{m1}, x_{m2}) is in the interior of $S_0(X)$, $\Pi_x \equiv \partial_{x'}\Delta\boldsymbol{p}(x_{m1}, x_{m2})\boldsymbol{W}_x\partial_x\Delta\boldsymbol{p}(x_{m1}, x_{m2})$ is nonsingular, and $h_x^2 \cdot \sqrt{nh_x} = o(1)$, we have*

$$\sqrt{nh_x}\begin{pmatrix}\hat{x}_{m1} - x_{m1}\\ \hat{x}_{m2} - x_{m2}\end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Pi_x^{-1}\partial_{x'}\Delta\boldsymbol{p}(x_{m1}, x_{m2})\boldsymbol{W}_x\Sigma_x\boldsymbol{W}_x\partial_x\Delta\boldsymbol{p}(x_{m1}, x_{m2})\Pi_x^{-1}) \tag{6.1}$$

where $\Sigma_x = \kappa\begin{pmatrix}\Sigma_{x1} & \mathbf{0}\\ \mathbf{0} & \Sigma_{x2}\end{pmatrix}$, $\kappa \equiv \int v^2K(v)dv$, *and*

$$\Sigma_{x1} = \begin{pmatrix} \frac{p_1(\tilde{z}_3)(1-p_1(\tilde{z}_3))}{f_{XZ}(\tilde{z}_3)} + \frac{p_1(\tilde{z}_1)(1-p_1(\tilde{z}_1))}{f_{XZ}(\tilde{z}_1)} & -\frac{p_1(\tilde{z}_3)p_2(\tilde{z}_3)}{f_{XZ}(\tilde{z}_3)} - \frac{p_1(\tilde{z}_1)p_2(\tilde{z}_1)}{f_{XZ}(\tilde{z}_1)} \\ -\frac{p_1(\tilde{z}_3)p_2(\tilde{z}_3)}{f_{XZ}(\tilde{z}_3)} - \frac{p_1(\tilde{z}_1)p_2(\tilde{z}_1)}{f_{XZ}(\tilde{z}_1)} & \frac{p_2(\tilde{z}_3)(1-p_2(\tilde{z}_3))}{f_{XZ}(\tilde{z}_3)} + \frac{p_2(\tilde{z}_1)(1-p_2(\tilde{z}_1))}{f_{XZ}(\tilde{z}_1)} \end{pmatrix},$$

$$\Sigma_{x2} = \begin{pmatrix} \frac{p_1(\tilde{z}_4)(1-p_1(\tilde{z}_4))}{f_{XZ}(\tilde{z}_4)} + \frac{p_1(\tilde{z}_2)(1-p_1(\tilde{z}_2))}{f_{XZ}(\tilde{z}_2)} & -\frac{p_1(\tilde{z}_4)p_2(\tilde{z}_4)}{f_{XZ}(\tilde{z}_4)} - \frac{p_1(\tilde{z}_2)p_2(\tilde{z}_2)}{f_{XZ}(\tilde{z}_2)} \\ -\frac{p_1(\tilde{z}_4)p_2(\tilde{z}_4)}{f_{XZ}(\tilde{z}_4)} - \frac{p_1(\tilde{z}_2)p_2(\tilde{z}_2)}{f_{XZ}(\tilde{z}_2)} & \frac{p_2(\tilde{z}_4)(1-p_2(\tilde{z}_4))}{f_{XZ}(\tilde{z}_4)} + \frac{p_2(\tilde{z}_2)(1-p_2(\tilde{z}_2))}{f_{XZ}(\tilde{z}_2)} \end{pmatrix}.$$

It is easy to verify that the optimal weighting matrix that achieves the smallest asymptotic variance given (6.1) is $\boldsymbol{W}_x^* = \Sigma_x^{-1}$. It can be estimated by adopting the standard two-step or multiple-step GMM approach. Denote the estimator using the estimated optimal weighting matrix by $(\hat{x}_{m1}^*, \hat{x}_{m2}^*)$, it is straightforward that

$$\sqrt{nh_x}\begin{pmatrix}\hat{x}_{m1}^* - x_{m1}\\ \hat{x}_{m2}^* - x_{m2}\end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(0, (\partial_{x'}\Delta\boldsymbol{p}(x_{m1}, x_{m2})\Sigma_x^{-1}\partial_x\Delta\boldsymbol{p}(x_{m1}, x_{m2}))^{-1}\Big) \tag{6.2}$$

Note that in this benchmark case, only a single covariate is present, so (x_{m1}, x_{m2}) is over-identified. The null hypothesis $\mathbb{H}_0 : \Delta\boldsymbol{p}(x_{m1}, x_{m2}) = \mathbf{0}$ can be tested by an over-identification

test. For example, consider J-test $\mathcal{J}_x = nh_x \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}, \hat{x}^*_{m2})' \widehat{\Sigma}_x^{-1} \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}, \hat{x}^*_{m2})$. Under the null, it can be verified that $\mathcal{J}_x \xrightarrow{d} \chi^2_2$.

In addition to jointly testing whether (x_{m1}, x_{m2}) solves the propensity score matching equations, we can separately test either one of them if needed. By block-diagonality of the asymptotic variance in equation (6.2), $\hat{x}^*_{m1}$ and $\hat{x}^*_{m2}$ are asymptotically independent, and thus it is equivalent to estimate them separately. In each separate problem the matching point is still over-identified, so let

$$\mathcal{J}_{x1} = nh_x \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1})' \widehat{\Sigma}_{x1}^{-1} \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}), \tag{6.3}$$

and

$$\mathcal{J}_{x2} = nh_x \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m2})' \widehat{\Sigma}_{x2}^{-1} \Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m2}), \tag{6.4}$$

Under the null, each of two test statistics converges in distribution to χ^2_1.

6.2 The SP-Outcome Functions

From equation (4.3), consistency of $\hat{\boldsymbol{m}}(x_0)$ directly follows from consistency of each component in its formula, guaranteed by the following regularity conditions.

Assumption Reg-SP. *For every d, z, x, $Y|d, x, z$ has finite second moment. $\mathbb{E}_{Y|DXZ}(d, \cdot, z)$ is twice continuously differentiable on $S(X)$ with bounded derivatives.*

Theorem Cons-SP. *Under the conditions in Theorem ID-SP, Assumptions Reg-MP, Reg-SP, Reg-K, and the benchmark conditions, $\hat{\boldsymbol{m}}(x_0) - \boldsymbol{m}^*(x_0) = o_p(1)$.*

For the asymptotic distribution, we let $h_m/h_x \to 0$ so that the impacts of estimating (x_{m1}, x_{m2}) and the propensity scores are negligible. Let $\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_6$ be $(x_0, 0)$, $(x_{m1}, 0)$, $(x_{m1}, 1)$, $(x_0, 1)$, $(x_{m2}, 1)$ and $(x_{m2}, 0)$. We have

Theorem AsymDist-SP. *Under the conditions in Theorem Cons-SP, suppose $h_m/h_x \to 0$ where h_x satisfies the conditions in Theorem AsymDist-MP, then*

$$\sqrt{nh_m}(\hat{\boldsymbol{m}}(x_0) - \boldsymbol{m}(x_0)) \xrightarrow{d} \mathcal{N}(0, (\Pi'_{SP}\boldsymbol{W}_m\Pi_{SP})^{-1}\Pi'_{SP}\boldsymbol{W}_m\Sigma_{SP}\boldsymbol{W}'_m\Pi_{SP}(\Pi'_{SP}\boldsymbol{W}_m\Pi_{SP})^{-1}) \tag{6.5}$$

where $\Sigma_{SP} = \kappa(\Sigma_{SP,1} + \Sigma_{SP,2} + \Sigma_{SP,3})$ and $\Sigma_{SP,d}$ $(d = 1, 2, 3)$ equals

$$\begin{pmatrix} \frac{p_d(\tilde{z}_1)^2 \mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)^2 \mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} \\ \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \sum_{k=1}^{3} \frac{p_d(\tilde{z}_2)^2 \mathbb{V}_{Y|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \sum_{k=4}^{6} \frac{p_d(\tilde{z}_5)^2 \mathbb{V}_{Y|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} \end{pmatrix}$$

Again, the optimal weighting matrix is $\boldsymbol{W}_m^* = \Sigma_{SP}^{-1}$. Note that Σ_{SP} does not depend on $\boldsymbol{m}^*(x_0)$, so it can be directly estimated by plugging in the estimated matching points and consistent estimators for the conditional variances and densities. Denote the estimator under the estimated optimal weighting matrix by $\hat{\boldsymbol{m}}^*(x_0)$, then

$$\sqrt{nh_m}(\hat{\boldsymbol{m}}^*(x_0) - \boldsymbol{m}^*(x_0)) \xrightarrow{d} \mathcal{N}\Big(0, (\Pi_{SP}'\Sigma_{SP}^{-1}\Pi_{SP})^{-1}\Big) \tag{6.6}$$

A consistent estimator of the asymptotic variance in equation (6.6) is straightforward to compute. Alternatively, bootstrap inference can be implemented by fixing $\hat{x}_{m1}$, $\hat{x}_{m2}$ and $\hat{\boldsymbol{p}}$ and only re-estimate the conditional expectations in each bootstrap sample.

As with the matching points, the over-identifying restrictions can be tested; construct the test statistic:

$$\mathcal{J}_{SP} = nh_m\Big(\widehat{\Pi}_{SP}(\hat{x}_{m1}, \hat{x}_{m2})\hat{\boldsymbol{m}}^*(x_0) - \widehat{\Phi}_{SP}(\hat{x}_{m1}, \hat{x}_{m2})\Big)'\widehat{\Sigma}_m^{-1}\Big(\widehat{\Pi}_{SP}(\hat{x}_{m1}, \hat{x}_{m2})\hat{\boldsymbol{m}}^*(x_0) - \widehat{\Phi}_{SP}(\hat{x}_{m1}, \hat{x}_{m2})\Big)$$

Under the null that all the moment conditions hold, $\mathcal{J}_{SP} \xrightarrow{d} \chi_1^2$.

It is worth noting that the test in our approach not only examines exogeneity of Z, but also jointly tests whether PSC holds at the selected conditioning points and whether all the conditions in Definition MP are satisfied.

6.3 The NSP-Outcome Functions

In this subsection, we provide sufficient conditions that deliver uniform consistency and asymptotic normality of $\hat{\boldsymbol{g}}(\boldsymbol{x}_0, u)$. It turns out that monotonicity of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ and the structure of our estimator simplify the general theory of sieve estimators (e.g. Chen and Pouzo (2012, 2015)); simple low level conditions suffice.

Let us begin by establishing the following key condition for consistency: For any closed interval $\mathcal{U}_0$ in the interior of $[0, 1]$,

$$\inf_{\substack{\boldsymbol{g}\in\mathcal{G}_0: \\ \sup_{u\in\mathcal{U}_0}|\boldsymbol{g}(u)-\boldsymbol{g}^*(x_0,u)|\geq\delta}} \Big|\int_0^1 Q_{NSP}(\boldsymbol{g}(u), u)du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0, u), u)du\Big| > 0 \tag{6.7}$$

When $\boldsymbol{g}^*$ is the unique solution to $Q_{NSP}(\cdot) = 0$, inequality (6.7) holds if $\mathcal{G}$ is compact in sup-norm (see for instance Newey and McFadden (1994)). It is common that function spaces are not compact. Thus for sieve estimators, zero on the right hand side of equation (6.7) is usually replaced by a positive sequence that converges to zero. We show that for the space $\mathcal{G}_0$ of uniformly bounded monotonic functions on a compact domain, inequality (6.7) does hold under Theorem ID-NSP.

Theorem ID-Sup. *Under the conditions in Theorem ID-NSP, inequality (6.7) is true.*

Proof. Denote $\mathcal{G}_0^- = \mathcal{G}_0 \backslash \{\boldsymbol{g} : \sup_{u \in \mathcal{U}_0} |\boldsymbol{g}(u) - \boldsymbol{g}^*(x_0, u)| \geq \delta\}$. Suppose inequality (6.7) does not hold. Then there exists a sequence $\boldsymbol{g}_k \in \mathcal{G}_0^-$ such that

$$\lim_{k \to \infty} \Big(\int_0^1 Q_{NSP}(\boldsymbol{g}_k(u), u) du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0, u), u) du \Big) = 0$$

As the sequence $\{\boldsymbol{g}_k\}$ are uniformly bounded monotonic functions on a compact interval, by Helly's Selection Theorem there exists a pointwise convergent subsequence $\tilde{\boldsymbol{g}}_{k_l}$. Denote its limit by $\tilde{\boldsymbol{g}}$. Note the equation above also holds for this subsequence. Then by the Dominated Convergence Theorem, we can change the order of the limit and the integral operators:

$$\begin{aligned} &\int_0^1 \lim_{k_l \to \infty} Q_{NSP}(\boldsymbol{g}_{k_l}(u), u) du = \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0, u), u) du = 0 \\ \implies &\int_0^1 Q_{NSP}(\tilde{\boldsymbol{g}}(u), u) du = \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0, u), u) du = 0 \end{aligned}$$

where the last equation follows from continuity of Q_{NSP}.

As $\mathcal{G}_0$ is closed, $\tilde{g} \in \mathcal{G}_0$. By Theorem ID-NSP, $\tilde{\boldsymbol{g}}(\cdot) = \boldsymbol{g}^*(x_0, \cdot)$ on $\mathcal{U}_0$. Since pointwise convergence of a sequence of monotonic functions on a compact domain implies uniform convergence if the limiting function is continuous, by continuity of $\boldsymbol{g}^*(x_0, u)$,

$$\lim_{k_l \to \infty} \sup_{u \in \mathcal{U}_0} |\boldsymbol{g}^*(x_0, u) - \boldsymbol{g}_{k_l}(u)| \to 0,$$

contradicting $\boldsymbol{g}_{k_l} \in \mathcal{G}_0^-$. □

From the last step in the proof, $\mathcal{U}_0$ in the theorem can be replaced by $[0, 1]$ if uniqueness of $\boldsymbol{g}^*(x_0, \cdot)$ holds on the entire interval $[0, 1]$. As discussed in Section 3.3, this would be the case when the codomain in $\mathcal{G}_0$ was set to be $\prod_d S(Y|d, x_0)$ instead of $\prod_d S(Y|d)$.

Consistency of $\hat{\boldsymbol{g}}$ in sup-norm then follows from the uniform convergence of $\hat{Q}_{NSP}$ and the existence of an element in the sieve space $\hat{G}$ that converges to $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ in sup-norm (see Chen (2007), Chen and Pouzo (2012, 2015), etc). The latter is straightforward because any continuous increasing functions can be approximated by piecewise affine increasing functions arbitrarily well provided that the number of the nodes are sufficiently large. The former holds if the proposed CDF estimators are uniformly consistent, guaranteed by the following regularity assumptions.

Assumption Reg-NSP. *For every $d \in S(D)$ and $z \in \{0, 1\}$, $F_{Y|DXZ}(\cdot|d, \cdot, z)$ is twice continuously differentiable on the support with bounded derivatives. The conditional density*

$f_{Y|DXZ}(\cdot|d,\cdot,z)$ is continuous and uniformly bounded away from 0 over the support for all d and z.

Assumption Reg-L. *$L(\cdot)$ is a continuously differentiable CDF supported on $[-1,1]$ with bounded derivatives.*

Theorem Cons-NSP. *Under the conditions in Theorem ID-NSP, Assumptions Reg-L, Reg-K, Reg-MP, Reg-NSP and the benchmark conditions, if $J \to \infty$ and $\lambda \cdot J \to 0$,*

$$\sup_{u \in \mathcal{U}_0} \left| \hat{\boldsymbol{g}}(x_0, u) - \boldsymbol{g}^*(x_0, u) \right| = o_p(1). \tag{6.8}$$

In particular, if $S(Y|d,x) = S(Y|d)$ for $x = x_0, x_{m1}, x_{m2}$, equation (6.8) holds for $\mathcal{U}_0 = [0,1]$.

Now let us derive the asymptotic distribution of $\hat{\boldsymbol{g}}(x_0,u_0) - \boldsymbol{g}^*(x_0,u_0)$ for a fixed u_0. Recall that by construction, $\underline{\boldsymbol{y}} \leq \hat{\boldsymbol{g}}(x_0,u_1) \leq \hat{\boldsymbol{g}}(x_0,u_2) \leq ... \leq \hat{\boldsymbol{g}}(x_0,u_J) \leq \bar{\boldsymbol{y}}$. If all these inequalities are strict, the estimator at each node satisfies the first order condition due to the smoothness of the CDF estimators and the penalty function.

Theorem Cons-NSP implies that for large enough n, $\hat{\boldsymbol{g}}(x_0,u_j)$ for all $u_j \in \mathcal{U}_0$ are uniformly close to the true function values at the corresponding nodes. Meanwhile, the differences between $\boldsymbol{g}^*(x_0,\cdot)$ at adjacent nodes converge to 0 because $|\boldsymbol{g}^*(x_0,u) - \boldsymbol{g}^*(x_0,u \pm 1/J)| = O(1/J)$ if $\boldsymbol{g}^*(x_0,\cdot)$ is differentiable and its derivative is bounded away from 0. Therefore, for any node $u \in \mathcal{U}_0$, by the triangle inequality, if $\hat{\boldsymbol{g}}(x_0,u)$ converges to $\boldsymbol{g}(x_0,u)$ faster than $1/J$, for large enough n, $\hat{\boldsymbol{g}}(x_0,u)$ is strictly greater than $\hat{\boldsymbol{g}}(x_0,u-1/J)$ and strictly smaller than $\hat{\boldsymbol{g}}(x_0,u+1/J)$. The following theorem provides the uniform rate of convergence of $\hat{\boldsymbol{g}}(x_0,\cdot)$ at each node in $\mathcal{U}_0$.

Theorem RoC-NSP (Rate of Convergence)**.** *Let $r_n = \sqrt{\log(n)/nh_g} + h_g$. Suppose $h_g/h_x \to 0$, $h_0/h_g \to 0$, and $\lambda = o(r_n^2)$. Under all the conditions in Theorem Cons-NSP,*

$$\max_{u_j \in \mathcal{U}_0} |\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)| = O_p(\sqrt{J} r_n) \tag{6.9}$$

Corollary RoC-NSP. *Under the conditions in Theorem RoC-NSP, suppose $J \cdot \sqrt{J} r_n \to 0$. If the derivative of $g_d^*(x_0,\cdot)$ is bounded away from 0 on $\mathcal{U}_0$ for all d, then $\hat{\boldsymbol{g}}(x_0,\cdot)$ on the nodes in $\mathcal{U}_0$ are strictly increasing with probability approaching one.*

Remark 5.1. *The order in equation (6.9) the square root of the uniform convergence rate of $J\hat{Q}_{NSP}$. Note that the bias is of the order of h_g, instead of h_g^2 in the standard case. This is because of the nonsmoothness of $F_{Y|DXZ}(\cdot|d,x,z)$ at the boundaries; symmetry of the kernel function cannot be utilized because of such nonsmoothness, slowing down the uniform rate.*

Note that under Corollary RoC-NSP, none of the inequality constraints are binding. $\hat{\boldsymbol{g}}(x_0,\cdot)$ at the nodes are thus asymptotically equivalent as the unconstrained pointwise estimator described in Section 4 under global identification pointwise at each node.

Let $\tilde{\boldsymbol{z}}_1,...,\tilde{\boldsymbol{z}}_6$ be $(x_0,0)$, $(x_{m1},0)$, $(x_{m1},1)$, $(x_0,1)$, $(x_{m2},1)$ and $(x_{m2},0)$. Let $\delta = \mathbb{1}(Y \leq g_D^*(X,u_0))$, $\phi_{d1} = \frac{f_{Y|DXZ}(g_d^*(x_{m1},u_0)|d,x_{m1},0)}{f_{Y|DXZ}(g_d^*(x_{m1},u_0)|d,x_{m1},1)}$, and $\phi_{d2} = \frac{f_{Y|DXZ}(g_d^*(x_{m2},u_0)|d,x_{m2},1)}{f_{Y|DXZ}(g_d^*(x_{m2},u_0)|d,x_{m2},0)}$.

Theorem AsymDist-NSP. *Under all the conditions in Corollary RoC-NSP, if* $J^{3/2}r_n \to 0$, $h_g^2 = o(1/nh_g)$ *and* $h_0 = o(1/nh_g)$, *then for any node* $u_0 \in \mathcal{U}_0$,

$$\sqrt{nh_g}(\hat{\boldsymbol{g}}(u_0) - \boldsymbol{g}^*(u_0)) \xrightarrow{d}$$
$$\mathcal{N}\Big(0, (\Pi'_{NSP}\boldsymbol{W}_g(u_0)\Pi_{NSP})^{-1}\Pi'_{NSP}\boldsymbol{W}_g(u_0)\Sigma_{NSP}\boldsymbol{W}_g(u_0)'\Pi_{NSP}(\Pi'_{NSP}\boldsymbol{W}_g(u_0)\Pi_{NSP})^{-1}\Big) \tag{6.10}$$

where $\Sigma_{NSP} = \kappa(\Sigma_{NSP,1} + \Sigma_{NSP,2} + \Sigma_{NSP,3})$ *and* $\Sigma_{NSP,d}$ $(d = 1,2,3)$ *equals*

$$\begin{pmatrix} \frac{p_d(\tilde{z}_1)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} \\ \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \sum_{k=1}^{3}\frac{\phi_{d1}^2 p_d(\tilde{z}_2)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \sum_{k=4}^{6}\frac{\phi_{d2}^2 p_d(\tilde{z}_5)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} \end{pmatrix}$$

The asymptotic variance has similar form as in Theorem AsymDist-SP for $\hat{\boldsymbol{m}}(x_0)$. In particular, entries in $\Sigma_{NSP,d}$ are similar to those in $\Sigma_{SP,d}$ with only two distinctions: In $\Sigma_{NSP,d}$, the conditional variances are with respect to the indicator function δ instead of Y, and there are additional factors ϕ_{d1} and ϕ_{d2} in the last two diagonal elements. The first distinction is analogous to the comparison of the variance formulas for mean regression and quantile regression. The second one is also expected due to nonlinearity in φ_d.

Similar to previous sections, the optimal weighting matrix that achieves the smallest asymptotic variance under equation (6.10) is $\boldsymbol{W}_g(u_0) = \Sigma_{NSP}^{-1}$, and a consistent estimator can be obtained by plugging in $\hat{\boldsymbol{g}}(x_0,\cdot)$ into the CDF and density estimators.

Now let us discuss how to obtain the second-step estimator using the feasible optimal weighting matrix. Although we can plug the optimal weighting matrix at each node into $\hat{Q}_{NSP}$ and solve the joint minimization problem again, in terms of computation, joint minimization is less favorable than individual minimization due to its high dimensionality. We adopt it because only local identification holds pointwise at each u. However, this is no longer a problem once we obtain a uniformly consistent first-step estimator $\hat{\boldsymbol{g}}$ using any weighting matrix. Uniform consistency guarantees that at any interior u_0, $\hat{\boldsymbol{g}}(x_0,u_0)$ is arbitrarily close to $\boldsymbol{g}^*(x_0,u_0)$. Therefore, if we focus on a new parameter space that is shrinking towards

$\hat{\boldsymbol{g}}(u_0)$, identification at u_0 holds in that space for large enough n:

$$\hat{\boldsymbol{g}}^*(x_0,u_0) = \arg \min_{[\hat{\boldsymbol{g}}(x_0,u_0)-\frac{1}{J},\hat{\boldsymbol{g}}(x_0,u_0)+\frac{1}{J}]} \hat{Q}^*_{NSP}(\boldsymbol{g}(x_0,u_0),u_0) \tag{6.11}$$

where $\hat{Q}^*_{NSP}(\boldsymbol{g}(u_0),u_0) = \Big(\widehat{\Psi}(\boldsymbol{g}(u_0)) - \boldsymbol{u}_0\Big)' \widehat{\Sigma}^{-1}_{NSP} \Big(\widehat{\Psi}(\boldsymbol{g}(u_0)) - \boldsymbol{u}_0\Big)$.

Theorem AsymDist-Op-NSP. *Under all the conditions in Theorem AsymDist-NSP,*

$$\sqrt{nh_g}(\hat{\boldsymbol{g}}^*(x_0,u_0) - \boldsymbol{g}^*(x_0,u_0)) \xrightarrow{d} \mathcal{N}(0, (\Pi'_{NSP}\Sigma^{-1}_{NSP}\Pi_{NSP})^{-1}) \tag{6.12}$$

Like $\hat{\boldsymbol{m}}^*(x_0)$, the asymptotic variance of $\hat{\boldsymbol{g}}^*(x_0,u_0)$ is straightforward to estimate. Bootstrap inference would also work. It would be computationally intensive if we computed $\hat{\boldsymbol{g}}^*$ in every bootstrap sample. Instead, we only solve the minimization problem once, then estimate the linear expansion of $\hat{Q}^*_{NSP}(\hat{\boldsymbol{g}}^*(x_0,u_0),u_0)$ in each bootstrap sample. The resulting distribution approximates that of $\hat{\boldsymbol{g}}^*(x_0,u_0) - \boldsymbol{g}^*(x_0,u_0)$.

Finally, over-identification test can be constructed by $\mathcal{J}_{NSP} = nh_g \hat{Q}^*_{NSP}(\hat{\boldsymbol{g}}^*(u_0),u_0)$. Under the null that all the moment conditions hold jointly, $\mathcal{J}_{NSP} \xrightarrow{d} \chi^2_1$.

7 Monte Carlo Simulations

This section illustrates the finite sample performance of the estimator for two separable models similar to the return to education example. The first model we consider has a three-valued D and a binary Z. In the second model, D is binary too. It enables us to compare our approach and the standard IV approach. We find that the extra moment conditions do not increase finite sample bias by much but largely reduce variances.

7.1 A Three-Valued D

Let D follow the ordered choice model in Example OC. Let the outcome variable Y be determined by the following model:

$$Y = [\gamma_1 \mathbb{1}(D=1) + \gamma_2 \mathbb{1}(D=2) + \gamma_3 \mathbb{1}(D=3)] \cdot (X+1) + U$$

where $X \sim \text{Unif}[-3,3]$, $Z \sim Ber(0.5)$, $[U,V] \sim \mathcal{N}\Big(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big)$ and $X \perp\!\!\!\perp Z \perp\!\!\!\perp (U,V)$.

For the parameters, we fix $(\gamma_1,\gamma_2,\gamma_3,\kappa_1,\kappa_2) = (1.5,3,3.5,-0.7,0.1)$. The parameters (α,β) govern the strength of the instrument and the covariate. In this section we present

Table 6: $x_0 = 0$. $\boldsymbol{m}^*(0) = (1.5, 3, 3.5)$.

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0)$	1000	1.49	2e-4	0.12	0.12	90.2% (86.8%)	95.4% (92.6%)	99% (97.6%)
	2000	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
	3000	1.49	4e-5	0.04	0.04	88.4% (88%)	94.4% (93.8%)	99% (97.6%)
$\hat{m}_2(0)$	1000	2.89	0.01	0.78	0.79	93.2% (88.8%)	96.2% (92.6%)	99% (96.6%)
	2000	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
	3000	2.92	0.01	0.25	0.26	89.2% (88.6%)	95.6% (93.2%)	99.8% (98.4%)
$\hat{m}_3(0)$	1000	3.47	0.001	0.22	0.22	92.8% (88.2%)	97% (93.4%)	98.6% (96.8%)
	2000	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
	3000	3.49	1e-4	0.07	0.07	92.6% (89.6%)	97.2% (95%)	99.4% (98.6%)
$\mathcal{J}_x$	1000					90%	95%	99.2%
	2000					91.6%	95.8%	99.6%
	3000					92.6%	96.8%	99.2%
$\mathcal{J}_{SP}$	1000					91.6%	94.8%	98.2%
	2000					93.6%	96.4%	98.4%
	3000					91.6%	96.2%	98.8%

the results for the cases $(\alpha, \beta) = (0.8, 0.4)$. The value is selected for two reasons: (a) all the propensity scores are far away from 0 so that in the simulated sample, there are sufficient observations to estimate each conditional expectation and propensity score; (b) X and Z have large effects on the propensity scores. Finally, we set $\rho = 0.5$ and $x_0 = 0$. Additional simulation results for small (α, β), different ρ and different x_0 are provided in Appendix D.

Table 6 contains the results for samples of size 1000, 2000 and 3000. The number of simulation replications is set at 500. In each replication, we estimate x_{m1} and x_{m2} using grid search with 500 grid nodes. The propensity scores are estimated using the biweight kernel. The bandwidth is equal to 1.6 times the Silverman's rule of thumb. The conditional expectations are estimated using the same kernel with a smaller bandwidth. Finally, we compute the actual coverage probabilities of the confidence intervals for $\boldsymbol{m}^*(x_0)$ based on both the asymptotic variance estimator (the top value in each cell in the last three columns) and 500 bootstrap samples (the bottom value in parentheses). The coverage probabilities of

Table 7: Binary D

x_0		Matching	Bias^2	Variance	MSE	90%	95%	99%
0	$\hat{m}_0(x_0)$	No	8e-4	0.07	0.07	91%	96%	99.2%
		Yes	6e-4	0.03	0.03	88%	94.4%	99.2%
	$\hat{m}_1(x_0)$	No	0.001	0.03	0.04	90.8%	95.6%	99.8%
		Yes	0.002	0.02	0.02	90.4%	94.8%	99.4%
-0.3	$\hat{m}_0(x_0)$	No	0.001	0.09	0.09	92.6%	95.8%	99.6%
		Yes	7e-4	0.04	0.04	89.6%	95.2%	99%
	$\hat{m}_1(x_0)$	No	8e-4	0.03	0.03	90%	95.6%	99.6%
		Yes	0.001	0.02	0.02	90%	94.6%	99.2%
0.3	$\hat{m}_0(x_0)$	No	5e-4	0.05	0.05	91.4%	96%	99.4%
		Yes	2e-4	0.03	0.03	90.4%	94.6%	99.4%
	$\hat{m}_1(x_0)$	No	0.002	0.03	0.04	92%	97%	99.6%
		Yes	0.002	0.02	0.02	90.4%	94.8%	99.2%

over-identification tests for (x_{m1}, x_{m2}) and for $\boldsymbol{m}^*(x_0)$ are also reported.

As is shown in Table 6, the variance of the estimator dominates in mean squared error (MSE) due to undersmoothing. The actual coverage probabilities are overall close to the nominal values. Bootstrap confidence intervals tend to undercover the true parameters while the asymptotic confidence intervals tend to be conservative.

7.2 A Binary D

We modify the data generating process in Section 7.1 to make D binary:

$$\begin{aligned} Y =& [\gamma_1 \mathbb{1}(D=0) + \gamma_2 \mathbb{1}(D=1)] \cdot (X+1) + U \\ D =& \mathbb{1}(V \geq \kappa + \alpha Z + \beta X) \end{aligned}$$

The distribution of (X, Z, U, V) is the same as in Section 7.1. Similarly, we set $(\gamma_1, \gamma_2, \kappa, \alpha, \beta) = (1.5, 3, -0.7, 0.8, 0.4)$. The correlation coefficient of U and V is 0.5.

The results are presented in Table 7. The third column indicates whether the matching points are estimated and used. We can see that adding more moment conditions from the matching points does not have much impact on the bias, but reduces the variance of the estimator. Meanwhile, the coverage probabilities are close to the nominal ones in all cases.

8 Relation to the Existing Literature

We discuss the most relevant approaches to identify nonparametric models with endogeneity in this section. The discussion is based on whether a selection model is explicitly exploited for identification of the outcome function.

8.1 Triangular Models

Triangular models are widely employed in the control function approach (e.g. Newey, Powell and Vella (1999), Chesher (2003) and Imbens and Newey (2009), etc.). This approach allows the outcome heterogeneity $\boldsymbol{U}$ to be multidimensional. On the other hand, D has to be continuous; these papers assume that V in the selection function is a scalar and $h(\boldsymbol{X}, Z, \cdot)$ is strictly increasing. Inverting $h(\boldsymbol{X}, Z, \cdot)$, the distribution of V can be traced out by $F_{D|\boldsymbol{XZ}}$. A "control variable" can be constructed, conditional on which endogeneity is eliminated.

Like this paper, D'Haultfœuille and Février (2015) and Torgovitsky (2015) study nonparametric identification in triangular models with a binary Z. As with the control function approach, they require D to be continuous and $h(\boldsymbol{X}, Z, \cdot)$ are strictly increasing. On the other hand, due to the small variation of the IV, they do not allow the unobservable in the outcome function to be multidimensional. This is more restrictive than the typical control function approach and the same as our paper. Gunsilius (2018) extends the model to allow for multidimensional heterogeneity, but the endogenous variable still has to be continuous.

Huang, Khalil and Yıldız (2019) consider a special additively separable triangular model where there are endogenous variables, denoted by D_1 and D_2, and a single IV Z. The triangular structure is with respect to one of the endogenous variables, for example D_1, in the sense that D_1 has a first stage as a function of Z and a scalar unobservable V. Specifically, the outcome variable Y is determined by $Y = m^*(D_1, D_2) + U$, and D_1 follows the equation $D_1 = h(Z, V)$. The benchmark case they focus on is that $m^*(D_1, D_2) = m_0^*(D_1) + \gamma D_2$. They follow the control function approach to construct a control variable for V, so $h(Z, \cdot)$ is assumed to be strictly increasing and D_1 needs to be continuously distributed (at least on a subset of its support).

One of the main contributions in this paper is that we allow for discrete D and dispense with monotonicity in the selection model. Although we are unable to trace out the entire distribution of $\boldsymbol{V}$ to construct a "control variable", the propensity scores provide useful information on the partitions of $S(\boldsymbol{V})$. Based on this, we find the matching points $\boldsymbol{x}_m$ that have the same level of endogeneity as $\boldsymbol{x}_0$. Endogeneity is eliminated by comparing the distribution or the mean of their outcomes.

8.2 Single Equation Approaches

Single equation approaches refer to methods that achieve identification without relying on the structures of the first stage. Perhaps the most important example is the standard IV approach for nonparametric identification (e.g. Newey and Powell (2003), Chesher (2004), Das (2005), Matzkin (2003), Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2007), Chen et al. (2014), etc.). As in this paper, the outcome function is usually assumed to be strictly increasing in the scalar unobservable. Typically, this approach requires Z to have large support.

Caetano and Escanciano (2018) develops a new identification strategy that achieves identification using small-support Z when D is multivalued. Their method does not rely on selection models. Similar to this paper, they utilize the variation in $\boldsymbol{X}$ for identification. But the strategy is different from this paper. Taking the nonseparable model as an example, their model essentially has a single index structure: $Y = g_D^*(U)$ and $U = \phi(\boldsymbol{X}, U_0)$. ϕ is a real-valued function and $g_D^*(\cdot)$ is strictly increasing a.s. Note that differently from our approach, they restrict the way in which the covariates enter the model. By contrast, we allow all the covariates to enter the model in arbitrary ways, but we need a selection model. Hence, their approach and ours are complementary.

9 Concluding Remarks

In this paper, we develop a novel approach to use covariates to identify separable and nonseparable models in a triangular system when the discrete endogenous variable takes on more values than the IV. This paper illustrates that information on endogenous selection has large identifying power. By tracing out the selection patterns across different values of the covariates and the IV, individuals that differ in observables but have the same selection patters may then be identified. Extrapolations can thus be made across them as they have the same degree of endogeneity, supplementing the insufficient information from the IV.

Moving forward, it would be of interest to extend the approach in this paper to other topics. In practice, the outcome variable is often limited, for example censored or truncated. Generalizing the approach to allow for such outcome variables would have a wide application. Another direction is to generalize the outcome function by allowing multidimensional heterogeneity, especially for the nonseparable model. For estimation, it would be interesting to investigate the optimal selection of matching points when they are uncountably infinite. It would also be useful to derive the optimal bandwidth choice for the multi-step local GMM or sieve estimation procedures proposed in this paper.

Appendix A General Cases

A.1 $|S(D)| > |S(Z)|$ and Multiple Endogenous Variables

Let us first discuss the general case of $|S(D)| > |S(Z)|$. For a given $\boldsymbol{x}_0$, at least $|S(D)| - |S(Z)|$ m-connected points are needed for identification. The size difference is not as formidable as it appears: When $|S(Z)|$ increases, the size of the m-connected points may increase at a faster rate. Recall Figure 1, each value of $z \in S(Z)$ induces an arm to grow m-connected points. Then, for instance, if $|S(Z)| = 3$, two matching points may be obtained by solving the following two equations:

$$\boldsymbol{p}(\boldsymbol{x}, z') = \boldsymbol{p}(\boldsymbol{x}_0, z) \text{ and } \boldsymbol{p}(\boldsymbol{x}, z'') = \boldsymbol{p}(\boldsymbol{x}_0, z)$$

Since z takes on 3 values, up to 6 matching points may be obtained even if each propensity score matching equation has only one solution. By induction, the number of matching points can be as many as $|S(Z)| \cdot (|S(Z)| - 1)$. With the variation in Z itself, a discrete IV taking on $|S(Z)|$ values may be able to identify a nonparametric model with an endogenous D with $|S(D)| = |S(Z)|^2$, instead of $|S(Z)|$ using the standard IV approach. For the m-connected points, the number can be even larger.

Example A1. *Consider an ordered choice model where $Z \in \{0, 1, 2\}$. Equivalently, define $Z_1 \geq Z_2 \in \{0, 1\}$ such that $Z = 0$ if and only if $Z_1 = Z_2 = 0$, $Z = 1$ if and only if $Z_1 = 1$ and $Z_2 = 0$, and $Z = 2$ if and only if $Z_1 = Z_2 = 1$. Let the linear single index be $X\beta + Z_1\alpha_1 + Z_2\alpha_2$, then we have the following six matching points:*

$$\begin{aligned}
(z = 0): &\beta x_{m1} + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 \implies x_{m1} = x_0 - \frac{\alpha_1}{\beta} \\
&\beta x_{m2} + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 = x_0\beta + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 \implies x_{m2} = x_0 - \frac{\alpha_1 + \alpha_2}{\beta} \\
(z = 1): &\beta x_{m3} + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 \implies x_{m3} = x_0 + \frac{\alpha_1}{\beta} \\
&\beta x_{m4} + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 \implies x_{m4} = x_0 - \frac{\alpha_2}{\beta} \\
(z = 2): &\beta x_{m5} + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 \implies x_{m5} = x_0 + \frac{\alpha_1 + \alpha_2}{\beta} \\
&\beta x_{m6} + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 \implies x_{m6} = x_0 + \frac{\alpha_2}{\beta}
\end{aligned}$$

Remark A1. *Note that the nonlinearity of Z's effect is needed to generate six matching points. If $\alpha_1 = \alpha_2$, then $x_{m1} = x_{m4}$ and $x_{m3} = x_{m6}$; only four matching points are generated.*

We can further generalize our approach to the case of multiple discrete endogenous vari-

ables. Suppose there are M discrete endogenous variables $D_1, ..., D_M$ in a model. It is equivalent to recode them as one single endogenous variable D_0. For instance, if $S(D_1, ..., D_M) = S(D_1) \times \cdots \times S(D_M)$, then let $S(D_0) = \{1, 2, ..., \Pi_{m=1}^{M} |S(D)_m|\}$. By construction, there exists a one-to-one mapping from $(D1, ..., D_M)$ to D_0, so the two models are equivalent.

The matching points can still be found by propensity score matching. Yet it is worth noting that since each D_m may be determined by different mechanisms, they may be affected by different components in $\boldsymbol{X}$. Therefore, the dimension of $\boldsymbol{X}$ needed tends to be larger than the case of a single endogenous variable. An example illustrating PSC for multiple Ds can be found in Appendix C in the supplement.

A.2 Multiple Solutions to $\mathbf{p}(x, z') = \mathbf{p}(x_0, z)$

In this section we adapt the estimator $(\hat{x}_{m1}, \hat{x}_{m2})$ to the general case where the solution to $\boldsymbol{p}(x, z') = \boldsymbol{p}(x_0, z)$ is not unique.

Recall equation (4.1) in Section 4, we define the estimator to be any $\hat{\boldsymbol{x}}_m \equiv (\hat{x}_{m1}, \hat{x}_{m2})$ that satisfies the following inequality:

$$\hat{Q}_x(\hat{\boldsymbol{x}}_m) \leq \inf_{S_0^2(X)} \hat{Q}_x(\boldsymbol{x}) + a_n$$

Let $\mathcal{X}_m$ be the set of all solutions to $Q(\boldsymbol{x}) = 0$. The estimator is consistent when $a_n = 0$ if $\mathcal{X}_m$ is a singleton. In general, we need $a_n > 0$ to consistently estimate $\mathcal{X}_m$ in Hausdorff distance[4].

Theorem Cons-MP-Set. *Let* $a_n = C\frac{(\log(n))^2}{nh_x}$ *for* $C > 0$. *Under Assumptions Reg-K and Reg-MP,* $\rho_H(\hat{\mathcal{X}}_m, \mathcal{X}_m) = O_p(\frac{\log(n)}{\sqrt{nh_x}})$.

Remark A2. *The rate of convergence is slower than that in the case of unique solution* $(\frac{1}{\sqrt{nh_x}})$. *This is because of the bias introduced by* a_n*; the convergence of the boundaries of* $\hat{\mathcal{X}}_m$ *is determined by the rate of* a_n*, and as* a_n *converges to 0 slower than* $\hat{Q}_x$*, the overall rate is slowed down.*

Once $\hat{\mathcal{X}}_m$ is obtained, one can select an element in it as the estimator of a matching point and use it to estimate $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$. However, in order to conduct the overidentification test, $\boldsymbol{x}_m$ has to be locally unique, i.e., an isolated solution, so that the Jacobian of $Q_x(\boldsymbol{x}_m)$ is full rank and the asymptotic distribution in Theorem AsymDist-MP holds.

[4]The Hausdorff distance ρ_H between two generic subsets A and B of a metric space endowed with metric ρ is defined as $\rho_H(A, B) = \max\{\sup_{a\in A} \inf_{b\in B} \rho(a, b), \sup_{b\in B} \inf_{a\in A} \rho(a, b)\}$. Intuitively, if the Hausdorff distance between two sets are small, for any point in either of the set, there exists a point close to it from the closure of the other set.

Therefore, we need to (a) find the isolated $\boldsymbol{x}_m$ and (b) reestimate the isolated matching point to obtain the optimal rate and the asymptotic distribution. We impose the following assumption.

Assumption ISO. *Let $\mathcal{X}_{ISO} \subseteq \mathcal{X}_m$ be the set of all isolated solutions. Suppose the following hold:*

(a) All the isolated solutions are in the interior of $S_0^2(X)$.

(b) There exists $\nu > 0$ such that:

$$\inf_{\boldsymbol{x}' \in \mathcal{X}_m \setminus \mathcal{X}_{ISO}, , \boldsymbol{x} \in \mathcal{X}_m} |\boldsymbol{x} - \boldsymbol{x}'| > \nu > 0.$$

(c) The Jacobian Π_x defined in Theorem AsymDist-MP is full rank at every isolated matching point.

Assumption ISO guarantees that each isolated solution is well separated from any other solutions. Under it, we now propose the following post-estimation procedure to estimate the isolated matching points:

- Step 0. Obtain $\hat{\mathcal{X}}_m$ by equation (4.1) with $a_n = C\frac{(\log(n))^2}{nh_x}$.
- Step 1 (Isolated solution selection). Cover $\hat{\mathcal{X}}_m$ with squares $\{I_k\}$ of side $b_n = \log(n)\sqrt{a_n}$. Select I_k if it fully covers a cluster of solutions.
- Step 2 (Reestimation). Minimize $\hat{Q}_x(\boldsymbol{x})$ on each selected I_k.

The rationale behind the procedure is as follows. From Theorem Cons-MP-Set and Assumption ISO, $\hat{\mathcal{X}}_m$ consists of isolated clusters of solutions with probability approaching one, and the diameter of a cluster converging in probability to an isolated solutions shrinks to zero at the rate of $\sqrt{a_n}$. Therefore, each of these cluster can be contained in one square I_k with probability approaching one since $b_n > \sqrt{a_n}$. For the reestimation step, again as $b_n/\sqrt{a_n} \to \infty$, by the rate in Theorem Cons-MP-Set, the true matching point is in the interior of the square with probability approaching 1. Therefore, the minimizer of $\hat{Q}_x$ satisfies the first order condition. By Assumption ISO, Theorem AsymDist-MP thus holds at this estimated matching point.

Appendix B Proofs of Results in Sections 2 and 3

Proof of Theorem MEQ. We first show that $\boldsymbol{p}(\boldsymbol{x}_0, z) = \boldsymbol{p}(\boldsymbol{x}_m, z')$. For any $d \in S(D)$,

$$\begin{aligned}
p_d(\boldsymbol{x}_0, z) =& \int_{h_d(\boldsymbol{x}_0,z,\boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_0, Z = z) \\
=& \int_{h_d(\boldsymbol{x}_0,z,\boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_0) \\
=& \int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_m) \\
=& \int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_m, Z = z') \\
=& p_d(\boldsymbol{x}_m, z')
\end{aligned}$$

where the second inequality is from Assumption E-SP or E-NSP and the third inequality is from Definition MP

Next we prove equation (2.6). For all $d \in S(D)$,

$$\begin{aligned}
\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) - m_d^*(\boldsymbol{x}_0, z) =& \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) \\
=& \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = 1, \boldsymbol{x}_0, z) \\
=& \frac{\int_{h_d(\boldsymbol{x}_0,z,\boldsymbol{v})=1} \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{v}, \boldsymbol{x}_0, z) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_0, Z = z)}{p_d(\boldsymbol{x}_0, z)} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{v}, \boldsymbol{x}_0) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_0)}{p_d(\boldsymbol{x}_m, z')} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{v}, \boldsymbol{x}_m) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_m)}{p_d(\boldsymbol{x}_m, z')} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{v}, \boldsymbol{x}_m, z') d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_m, z')}{p_d(\boldsymbol{x}_m, z')} \\
=& \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1, \boldsymbol{x}_m, z') \\
=& \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') \\
=& \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - m_d^*(\boldsymbol{x}_m, z')
\end{aligned}$$

where the fourth inequality follows Assumption E-SP. The fifth inequality is from Definition MP. For the sixth equality, Definition MP implies the exogeneity assumption also holds at $\boldsymbol{x}_m$.

Finally we show equation (2.7). For all $d \in S(D)$,

$$\begin{aligned}
F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) =& F_{U|D\boldsymbol{X}Z}(u|d, \boldsymbol{x}_0, z) \\
=& \frac{\int_{h_d(\boldsymbol{x}_0,z,\boldsymbol{v})=1} F_{U|\boldsymbol{V}\boldsymbol{X}Z}(u|\boldsymbol{v}, \boldsymbol{x}_0, z) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_0, Z = z)}{p_d(\boldsymbol{x}_0, z)} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m,z',\boldsymbol{v})=1} F_{U|\boldsymbol{V}\boldsymbol{X}Z}(u|\boldsymbol{v}, \boldsymbol{x}_m, z') d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v}|\boldsymbol{X} = \boldsymbol{x}_m, Z = z')}{p_d(\boldsymbol{x}_m, z')} \\
=& F_{U|\boldsymbol{V}\boldsymbol{X}Z}(u|h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1, \boldsymbol{x}_m, z') \\
=& F_{U|D\boldsymbol{X}Z}(u|d, \boldsymbol{x}_m, z') \\
=& F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_m, u)|d, \boldsymbol{x}_m, z')
\end{aligned}$$

where the first equality follows from Assumptions FS and CM. The third and the fourth equalities are from Definitions MP and Assumption E-NSP. For the third and the last equality, by Definition of MP, the exogeneity and the full support assumptions also hold at $\boldsymbol{x}_m$. □

Proof of Theorem ID-OC. Under the model setup, inequality (3.3) holds if

$$\begin{aligned}
&\Big(F_{V_1}(x_0\beta) - F_{V_1}(x_0\beta - \alpha)\Big)\Big(F_{V_2}(x_0\beta) - F_{V_2}(x_0\beta + \alpha)\Big) \\
\neq&\Big(F_{V_1}(x_0\beta) - F_{V_1}(x_0\beta + \alpha)\Big)\Big(F_{V_2}(x_0\beta) - F_{V_2}(x_0\beta - \alpha)\Big)
\end{aligned} \tag{B.1}$$

Without loss of generality, assume $\alpha > 0$. Then there are four cases.

- $x_0\beta - \alpha < x_0\beta < x_0\beta + \alpha \leq c$. Then $F_{V_2}(x_0\beta - \alpha) = F_{V_2}(x_0\beta) = F_{V_2}(x_0\beta + \alpha) = 0$. Inequality (B.1) does not hold.
- $x_0\beta - \alpha < x_0\beta \leq c < x_0\beta + \alpha$. The left hand side is negative but the right hand side is zero because $F_{V_2}(x_0\beta - \alpha) = F_{V_2}(x_0\beta) = 0$. Inequality (B.1) holds.
- $x_0\beta - \alpha < c < x_0\beta < x_0\beta + \alpha$. The left hand side is still negative but the right hand side is again zero because $F_{V_1}(x_0\beta) = F_{V_1}(x_0\beta + \alpha) = 1$.
- $c \leq x_0\beta - \alpha < x_0\beta < x_0\beta + \alpha$. Both sides are zero because $F_{V_1}(x_0\beta) = F_{V_1}(x_0\beta - \alpha) = F_{V_1}(x_0\beta + \alpha) = 1$.

Therefore, inequality (B.1) holds if there are two elements from $\{(x_0, 0), (x_0, 1), (x_m, 0)\}$ making the single indices located left and right to c respectively. □

Proof of Theorem ID-NSP. We prove the theorem in two steps. In the first step, we formalize the sketch of the proof in Section 3.3, which only considers the uniqueness of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ in $\mathcal{G}^* \subseteq \mathcal{G}$. In the second step, we extend the result to $\mathcal{G}$.

Step 1. Recall that the ranges of functions in $\mathcal{G}^*$ are contained in $\prod_{d=1}^3 S(Y|d, \boldsymbol{x}_0)$. This set is compact under Assumptions CM and Assumption FS. For each $d \in S(D)$, denote the lower and the upper bounds of $S(Y|d, \boldsymbol{x}_0)$ by $\underline{y}_{d\boldsymbol{x}_0}$ and $\bar{y}_{d\boldsymbol{x}_0}$. By Assumption FS,

$$g_d^*(\boldsymbol{x}_0, 0) = \underline{y}_{d\boldsymbol{x}_0} \text{ and } g_d^*(\boldsymbol{x}_0, 1) = \bar{y}_{d\boldsymbol{x}_0}, \forall d \in S(D)$$

Suppose $\boldsymbol{g}^*$ is not the unique solution path in $\mathcal{G}_0$, then any other solution path $\tilde{\boldsymbol{g}} \equiv (\tilde{g}_1, \tilde{g}_2, \tilde{g}_3)$ must also satisfy

$$\tilde{g}_d(0) = \underline{y}_{d\boldsymbol{x}_0} \text{ and } \tilde{g}_d(1) = \bar{y}_{d\boldsymbol{x}_0}, \forall d \in S(D)$$

Therefore, the set $\{u' : \tilde{\boldsymbol{g}}(u) = \boldsymbol{g}^*(\boldsymbol{x}_0, u), 0 \leq u \leq u'\}$ is nonempty and its supremum, denoted by $\bar{u}$, is well-defined. By continuity of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$, $\bar{u}$ is in the set.

If $\bar{u} = 1$, we are done.

If $\bar{u} < 1$, by monotonicity, $\tilde{\boldsymbol{g}}(u)$ has at most countable discontinuities. Thus, there exists an interval $(\bar{u}, \bar{u}')$ where $\bar{u}' < 1$ such that on the interval, $\tilde{\boldsymbol{g}}(u)$ is continuous and $\tilde{\boldsymbol{g}}(u) \neq \boldsymbol{g}^*(\boldsymbol{x}_0, u)$ for $u \in (\bar{u}, \bar{u}')$. Then there are the following two cases depending on whether $\tilde{\boldsymbol{g}}$ is continuous at $\bar{u}$.

Case 1. $\tilde{\boldsymbol{g}}(\cdot)$ is continuous at $\bar{u}$. Since the Jacobian of $\Psi(\cdot)$ at $\boldsymbol{g}^*(\boldsymbol{x}_0, \bar{u})$ is full rank, there exists a neighborhood $\mathcal{N}$ around $\boldsymbol{g}^*(\boldsymbol{x}_0, \bar{u})$ on which $\Psi(\cdot)$ is one-to-one. However, by continuity, there exists u'' such that $u'' \in (\bar{u}, \bar{u}')$ and $\tilde{\boldsymbol{g}}(u'') \in \mathcal{N}$. Then $\tilde{\boldsymbol{g}}(u'') \neq \boldsymbol{g}^*(\boldsymbol{x}_0, u'')$ but $\Psi(\tilde{\boldsymbol{g}}(u'')) = \Psi(\boldsymbol{g}^*(\boldsymbol{x}_0, u'')) = \boldsymbol{u}''$, a contradiction. As in footnote 2, a similar argument can also be found in proofs of the Hadamard Theorem in Ortega and Rheinboldt (1970), Ambrosetti and Prodi (1995), De Marco, Gorni and Zampieri (2014).

Case 2. $\tilde{\boldsymbol{g}}(\cdot)$ is not continuous at $\bar{u}$. By monotonicity, there is at least one $d \in S(D)$ such that $\lim_{u \searrow \bar{u}} \tilde{g}_d(u) > \lim_{u \nearrow \bar{u}} \tilde{g}_d(u)$, i.e., $\tilde{g}_d(\cdot)$ jumps up at $\bar{u}$. However, as both $F_{Y|D\boldsymbol{X}Z}$ and φ_d are continuous and strictly increasing, to make $\lim_{u \searrow \bar{u}} \Psi(\tilde{g}_d(u)) = \lim_{u \nearrow \bar{u}} \Psi(\tilde{g}_d(u)) = \boldsymbol{u}$, there must exist $d' \neq d$ such that $\lim_{u \searrow \bar{u}} \tilde{g}_d(u) < \lim_{u \nearrow \bar{u}} \tilde{g}_d(u)$. A contradiction with that $\tilde{g}_{d'}(\cdot)$ is increasing.

Therefore, $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is the unique solution path in $\mathcal{G}_0$ to $\Psi(\boldsymbol{g}(u)) = \boldsymbol{u}$.

Step 2. Now we consider the general case: $\mathcal{G}$ contains functions whose ranges contain sets that are out of the support set $\prod_{d=1}^3 S(Y|d, \boldsymbol{x}_0)$.

Suppose there exists another solution path $\breve{\boldsymbol{g}} : [0, 1] \in \mathcal{G}$. Construct $\boldsymbol{g}^\dagger$ as follows: For each $d \in S(D)$, let

$$g_d^\dagger(u) = \begin{cases} \underline{y}_{d\boldsymbol{x}_0}, & \text{if } \breve{g}_d(u) < \underline{y}_{d\boldsymbol{x}_0} \\ \breve{g}_d(u), & \text{if } \breve{g}_d(u) \in S(Y|d, \boldsymbol{x}_0) \\ \bar{y}_{d\boldsymbol{x}_0}, & \text{if } \breve{g}_d(u) > \bar{y}_{d\boldsymbol{x}_0} \end{cases}$$

Clearly, $\boldsymbol{g}^\dagger \in \mathcal{G}^*$. From Step 1, $\boldsymbol{g}^\dagger = \boldsymbol{g}^*$, so $\breve{\boldsymbol{g}}(u) = \boldsymbol{g}(\boldsymbol{x}, u)$ for all $u \in (0,1)$. For $u = 0, 1$, $\breve{g}_d(u)$ can take any value smaller than $\underline{y}_{d\boldsymbol{x}_0}$ or $\bar{y}_{d\boldsymbol{x}_0}$ respectively for all d. □

Appendix C Examples for Propensity Score Coherence

Example OC Cont'd 3. *In this example we generalize the ordered choice model in Example OC Cont'd 2.*

Let $h_1(\boldsymbol{X}, Z, V) = \mathbb{1}(V_1 \leq \gamma_1(\boldsymbol{X}, Z))$, $h_3(\boldsymbol{X}, Z, V) = \mathbb{1}(V_2 > \gamma_2(\boldsymbol{X}, Z))$, and $h_2 = 1 - h_1 - h_3$ where $V_1 < V_2$ a.s. are two scalar random variables that are continuously distributed on $\mathbb{R}$. Also, assume $\gamma_1(\boldsymbol{X}, Z) < \gamma_2(\boldsymbol{X}, Z)$ a.s. This model nests parametric ordered choice models and also some nonparametric models with more complicated structures, for instance the general ordered choice model in Cunha, Heckman and Navarro (2007).

The matching points are identified by propensity score matching (3.1): By strict monotonicity of F_{V_1} and F_{V_2}, it is straightforward that for any $(\boldsymbol{x}, z)$ and $(\boldsymbol{x}', z')$, $p_1(\boldsymbol{x}, z) = p_1(\boldsymbol{x}', z')$ implies $\gamma_1(\boldsymbol{x}, z) = \gamma_1(\boldsymbol{x}', z')$, and $p_3(\boldsymbol{x}, z) = p_3(\boldsymbol{x}', z')$ implies $\gamma_2(\boldsymbol{x}, z) = \gamma_2(\boldsymbol{x}', z')$.

Example MC (Multinomial Choice)**.** *In this model we consider a nonparametric multinomial choice model. Variants of it are considered in Matzkin (1993), Heckman, Urzua and Vytlacil (2008), and Lee and Salanié (2018). Let $R_d(\boldsymbol{X}, Z) + \tilde{V}_d$ be the indirect utility of choosing treatment d where $\tilde{V}_d$ is an unobservable continuous random variable. Alternative d is selected if $R_d(\boldsymbol{X}, Z) + \tilde{V}_d > R_{-d}(\boldsymbol{X}, Z) + \tilde{V}_{-d}$ where the subscript $-d$ refers to any selection other than d. Reparameterize the model by letting $V_1 = \tilde{V}_2 - \tilde{V}_1$, $V_2 = \tilde{V}_3 - \tilde{V}_1$, $V_3 = \tilde{V}_3 - \tilde{V}_2$, $\gamma_1(\boldsymbol{X}, Z) = R_1(\boldsymbol{X}, Z) - R_2(\boldsymbol{X}, Z)$ and $\gamma_2(\boldsymbol{X}, Z) = R_1(\boldsymbol{X}, Z) - R_3(\boldsymbol{X}, Z)$. The model can be rewritten as*

$$
\begin{aligned}
D = 1 &\iff V_1 < \gamma_1(\boldsymbol{X}, Z), V_2 < \gamma_2(\boldsymbol{X}, Z) \\
D = 2 &\iff V_1 > \gamma_1(\boldsymbol{X}, Z), V_3 < \gamma_2(\boldsymbol{X}, Z) - \gamma_1(\boldsymbol{X}, Z) \\
D = 3 &\iff V_2 > \gamma_2(\boldsymbol{X}, Z), V_3 < \gamma_2(\boldsymbol{X}, Z) - \gamma_1(\boldsymbol{X}, Z)
\end{aligned}
$$

If $0 < p_d(\boldsymbol{x}_0, z) < 1$ for all d, it can be verified that the solutions to equation (3.1) are matching points of $\boldsymbol{x}_0$.

Example TWF (Two-Way Flow)**.** *This example is adapted from the two-way flow model in Lee and Salanié (2018).*

Let $h_1(\boldsymbol{X}, Z, V) = \mathbb{1}\Big(V_1 \leq \gamma_1(\boldsymbol{X}, Z), V_2 \leq \gamma_2(\boldsymbol{X}, Z)\Big)$, $h_2(\boldsymbol{X}, Z, V) = \mathbb{1}\Big(V_1 \geq \gamma_1(\boldsymbol{X}, Z), V_2 \geq \gamma_2(\boldsymbol{X}, Z)\Big)$, and $h_3 = 1 - h_1 - h_2$. V_1 and V_2 are two scalar random variables that are continuously distributed.

The benefit of this double-threshold structure is that it breaks the traditional notion of monotonicity (in selection), allowing for richer selection patterns. The cost is that stronger assumptions are needed to identify $\gamma_1(\boldsymbol{X}, Z)$ and $\gamma_2(\boldsymbol{X}, Z)$. The key assumption in Lee and Salanié (2018) is that there exist covariates that are only in γ_1 and covariates only in γ_2. Then they show that γ_1 and γ_2 are identified up to an additive constant (see Theorem 4.1 in Lee and Salanié (2018), pp. 1551-1552), that is, $\gamma_1(\boldsymbol{X}, Z) = \tilde{\gamma}_1(\boldsymbol{X}, Z) + c$ and $\gamma_2(\boldsymbol{X}, Z) = \tilde{\gamma}_2(\boldsymbol{X}, Z) - c$, where $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are known functions but c is an unknown constant. Therefore, the matching points of $\boldsymbol{x}_0$ can be directly identified by solving $\tilde{\gamma}_d(\boldsymbol{x}, z') - \tilde{\gamma}_d(\boldsymbol{x}_0, z)$ for all d.

It is straightforward to see that PSC does not hold for all $(\boldsymbol{x}, z)$. For example, it is possible that there exists $(\boldsymbol{x}', z')$ such that $\gamma_1(\boldsymbol{x}', z') < \gamma_1(\boldsymbol{x}_0, z)$, $\gamma_2(\boldsymbol{x}', z') > \gamma_2(\boldsymbol{x}_0, z)$ and $F_{V_1V_2}(\gamma_1(\boldsymbol{x}', z'), \gamma_2(\boldsymbol{x}', z')) = F_{V_1V_2}(\gamma_1(\boldsymbol{x}_0, z), \gamma_2(\boldsymbol{x}_0, z))$.

However, under the following conditions there are subsets of $S(\boldsymbol{X}, Z)$ where PSC holds. For simplicity, let $\boldsymbol{X} = (X_1, X_2)$. Let γ_1 depend on Z and X_1 only, and γ_2 depend on Z and X_2. Suppose the support of V_1 and V_2 are bounded, say from above by c. If there exist $\bar{x}_1$ and $\bar{x}_2$ such that $\gamma_1(\bar{x}_1) > c$ and $\gamma_2(\bar{x}_2) > c$, then fixing $\bar{x}_2$ and $\bar{x}_1$, the matching points of $(x_1, \bar{x}_2)$ and of $(\bar{x}_1, x_2)$ for any $x_1 \in S(X_1)$ and $x_2 \in S(X_2)$ can be identified by propensity score matching; at $\bar{x}_2$ or $\bar{x}_1$, only one threshold is effective and the model becomes equivalent to the ordered choice model in Example OC Cont'd 3.

Example Two Endogenous Variables. *In this example, we have two dummy endogenous variables. We show that PSC still holds in this model. Let $D_1, D_2 \in \{0, 1\}$ be two endogenous variables. Suppose they are determined by the following model:*

$$D_1 = \mathbb{1}(\gamma_1(X, Z) \leq V_1)$$
$$D_2 = \mathbb{1}(\gamma_2(X, Z) \leq V_2)$$

where V_1 and V_2 are unobservables continuously distributed on $\mathbb{R}^2$. Let $D_0 = 1, 2, 3, 4$, corresponding to $(D_1, D_2) = (0, 0), (0, 1), (1, 0), (1, 1)$ respectively. Then the model can be rewritten with our notation:

$$\begin{aligned}
h_1(X, Z, V_1, V_2) &= \Big(1 - \mathbb{1}(\gamma_1(X, Z) \leq V_1)\Big) \cdot \Big(1 - \mathbb{1}(\gamma_2(X, Z) \leq V_2)\Big) \\
h_2(X, Z, V_1, V_2) &= \Big(1 - \mathbb{1}(\gamma_1(X, Z) \leq V_1)\Big) \cdot \mathbb{1}(\gamma_2(X, Z) \leq V_2) \\
h_3(X, Z, V_1, V_2) &= \mathbb{1}(\gamma_1(X, Z) \leq V_1) \cdot \Big(1 - \mathbb{1}(\gamma_2(X, Z) \leq V_2)\Big) \\
h_4(X, Z, V_1, V_2) &= \mathbb{1}(\gamma_1(X, Z) \leq V_1) \cdot \mathbb{1}(\gamma_2(X, Z) \leq V_2)
\end{aligned}$$

It is clear that $\sum_d h_d(X, Z) = 1$. Let us now verify that PSC holds for this model. Suppose

$\boldsymbol{p}(x_0, z) = \boldsymbol{p}(x_m, z')$. *Then*

$$\mathbb{P}\Big(V_1 < \gamma_1(x_0, z), V_2 < \gamma_2(x_0, z)\Big) = \mathbb{P}\Big(V_1 < \gamma_1(x_m, z'), V_2 < \gamma_2(x_m, z')\Big)$$

Suppose $\gamma_1(x_0, z) \neq \gamma_1(x_m, z')$. *Without loss of generality, let* $\gamma_1(x_0, z) < \gamma_1(x_m, z')$. *Then* $\gamma_2(x_0, z) > \gamma_2(x_m, z')$. *Consequently,*

$$\mathbb{P}\Big(V_1 \geq \gamma_1(x_0, z), V_2 < \gamma_2(x_0, z)\Big) > \mathbb{P}\Big(V_1 \geq \gamma_1(x_m, z'), V_2 < \gamma_2(x_m, z')\Big)$$

But this implies $p_3(x_0, z) > p_3(x_m, z')$, *a contradiction.*

References

Ambrosetti, Antonio, and Giovanni Prodi. 1995. *A Primer of Nonlinear Analysis.* Vol. 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Caetano, Carolina, and Juan C. Escanciano. 2018. "Identifying multiple marginal effects with a single instrument." *Working Paper.*

Card, David. 1995. "Using geographic variation in college proximity to estimate the return to schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp.* ed. Louis N. Christofides, E. Kenneth Grant, and Roebert Swidinsky, 201-222. Toronto: University of Toronto Press.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review*, 101(6): 2754–81.

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." In *Handbook of Econometrics, Vol. 6B.* ed. James J. Heckman and Edward E. Leamer, 5549–5632. Amsterdam: Elsevier.

Chen, Xiaohong, and Demian Pouzo. 2012. "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals." *Econometrica*, 80(1): 277–321.

Chen, Xiaohong, and Demian Pouzo. 2015. "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models." *Econometrica*, 83(3): 1013–1079.

Chen, Xiaohong, Victor Chernozhukov, Sokbae Lee, and Whitney K. Newey. 2014. "Local identification of nonparametric and semiparametric models." *Econometrica*, 82(2): 785–809.

Chernozhukov, Victor, and Christian Hansen. 2005. "An IV model of quantile treatment effects." *Econometrica*, 73(1): 245–261.

Chernozhukov, Victor, Guido W. Imbens, and Whitney K. Newey. 2007. "Instrumental variable estimation of nonseparable models." *Journal of Econometrics*, 139(1): 4–14.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and confidence regions for parameter sets in econometric models." *Econometrica*, 75(5): 1243–1284.

Chesher, Andrew. 2003. "Identification in nonseparable models." *Econometrica*, 71(5): 1405–1441.

Chesher, Andrew. 2004. "Identification in additive error models with discrete endogenous variables." *cemmap Working Paper, CWP11/04*.

Cunha, Flavio, James J. Heckman, and Salvador Navarro. 2007. "The identification and economic content of ordered choice models with stochastic thresholds." *International Economic Review*, 48(4): 1273–1309.

Das, Mitali. 2005. "Instrumental variables estimators of nonparametric models with discrete endogenous regressors." *Journal of Econometrics*, 124(2): 335–361.

De Marco, Giuseppe, Gianluca Gorni, and Gaetano Zampieri. 2014. "Global inversion of functions: An introduction." *arXiv preprint arXiv:1410.7902*.

D'Haultfœuille, Xavier, and Philippe Février. 2015. "Identification of nonseparable triangular models with discrete instruments." *Econometrica*, 83(3): 1199–1210.

Feng, Qian, Quang Vuong, and Haiqing Xu. 2019. "Estimation of heterogeneous individual treatment effects with endogenous treatments." *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2018.1543121.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong. 2000. "Optimal nonparametric estimation of first-price auctions." *Econometrica*, 68(3): 525–574.

Gunsilius, Florian. 2018. "Point-identification in multivariate nonseparable triangular models." *arXiv preprint arXiv:1806.09680*.

Hansen, Bruce E. 2004. "Nonparametric estimation of smooth conditional distributions." *Working Paper, Department of Economics, University of Wisconsin, Madison.*

Heckman, James J., and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73(3): 669–738.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2008. "Instrumental variables in models with multiple outcomes: The general unordered case." *Annales d'Economie et de Statistique*, 91/92: 151–174.

Huang, Liquan, Umair Khalil, and Neşe Yıldız. 2019. "Identification and estimation of a triangular model with multiple endogenous variables and insufficiently many instrumental variables." *Journal of Econometrics*, 208(2): 346–366.

Ichimura, Hidehiko, and Christopher Taber. 2000. "Direct estimation of policy impacts." *NBER Technical Working Paper No. 254.*

Imbens, Guido W., and Whitney K. Newey. 2009. "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica*, 77(5): 1481–1512.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics*, 131(4): 1795–1848.

Lee, Sokbae, and Bernard Salanié. 2018. "Identifying effects of multivalued treatments." *Econometrica*, 86(6): 1939–1963.

Lewbel, Arthur. 2007. "A local generalized method of moments estimator." *Economics Letters*, 94(1): 124–128.

Li, Qi, and Jeffrey S. Racine. 2008. "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data." *Journal of Business & Economic Statistics*, 26(4): 423–434.

Matzkin, Rosa L. 1993. "Nonparametric identification and estimation of polychotomous choice models." *Journal of Econometrics*, 58(1-2): 137–168.

Matzkin, Rosa L. 2003. "Nonparametric estimation of nonadditive random functions." *Econometrica*, 71(5): 1339–1375.

Newey, Whitney K., and Daniel. McFadden. 1994. "Large sample estimation and hypothesis testing." In *Handbook of Econometrics, Vol. 4.* ed. Robert F. Engle and Daniel L. McFadden, 2111–2245. Amsterdam: Elsevier.

Newey, Whitney K., and James L. Powell. 2003. "Instrumental variable estimation of nonparametric models." *Econometrica*, 71(5): 1565–1578.

Newey, Whitney K., James L. Powell, and Francis Vella. 1999. "Nonparametric estimation of triangular simultaneous equations models." *Econometrica*, 67(3): 565–603.

Ortega, James M., and Werner C. Rheinboldt. 1970. *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press.

Torgovitsky, Alexander. 2015. "Identification of nonseparable models using instruments with small support." *Econometrica*, 83(3): 1185–1197.

Torgovitsky, Alexander. 2017. "Minimum distance from independence estimation of nonseparable instrumental variables models." *Journal of Econometrics*, 199(1): 35–48.

United States Department of Health and Human Services. Administration for Children and Families. Office of Planning, Research and Evaluation. 2018-02-08. "Head Start Impact Study (HSIS), 2002-2006 [United States]." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR29462.v7.

Vuong, Quang, and Haiqing Xu. 2017. "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity." *Quantitative Economics*, 8(2): 589–610.

Vytlacil, Edward, and Neşe Yıldız. 2007. "Dummy endogenous variables in weakly separable models." *Econometrica*, 75(3): 757–779.

Supplement to "Matching Points: Supplementing Instruments with Covariates in Triangular Models"

Junlong Feng*

November 15, 2019

Abstract

Appendix D contains additional simulation results. Appendix E collects proofs for some of the asymptotic properties in Section 6.

*Dept. of Economics, Columbia University, New York, NY 10027, U.S.A.; junlong.feng@columbia.edu.

Appendix D Additional Simulation Results

In this section, we present the simulation results under different parameters in the model in Section 7.1. Table D1 and Table D2 present the results for $x_0 = \pm 0.3$ with parameters the same as in Table 6. We can see the results are similar to those in Table 6. Table D3 shows the results for weaker IV: $(\alpha, \beta) = (0.16, 0.08)$. These parameters are one fifth of the benchmark one, while the smallest eigenvalue of the matrix $\Pi'_{SP}\Pi_{SP}$ is 1/200 of the original. The variances blow up, but the biases are still very small like in the benchmark case. Table D4 shows the results for different correlation coefficients $\rho = 0.3, 0.7$. The results are again very similar to the benchmark case $\rho = 0.5$.

Table D1: $x_0 = -0.3, \boldsymbol{m}^*(-0.3) = (1.05, 2.1, 2.45)$

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(-0.3)$	1000	1.02	0.001	0.14	0.14	93% (87.2%)	97% (92%)	99.2% (98%)
	2000	1.04	2e-4	0.07	0.07	91% (88.6%)	96.4% (94.8%)	99.2% (98.6%)
	3000	1.04	1e-4	0.05	0.05	89.8% (88.6%)	94.4% (93.2%)	99% (98.8%)
$\hat{m}_2(-0.3)$	1000	2.09	1e-4	0.85	0.85	93.8% (88%)	96.8% (94.8%)	99% (98.2%)
	2000	2.04	0.003	0.34	0.35	92.8% (90.2%)	96.8% (95.4%)	99.6% (99.4%)
	3000	2.04	0.004	0.25	0.26	90% (88.2%)	95.8% (93.2%)	99.4% (97.8%)
$\hat{m}_3(-0.3)$	1000	2.36	0.01	0.21	0.22	91.4% (88.6%)	96% (93.4%)	98.4% (97.4%)
	2000	2.41	0.002	0.10	0.10	93.2% (90.6%)	96.6% (95.4%)	98.6% (99.4%)
	3000	2.43	3e-4	0.07	0.07	89.2% (88.4%)	94.4% (94.4%)	99.2% (98.8%)
$\mathcal{J}_x$	1000					88.6%	94.6%	99%
	2000					90.4%	95.2%	99.2%
	3000					91.2%	97%	99.4%
$\mathcal{J}_{SP}$	1000					93.4%	96.6%	99.2%
	2000					92.2%	96.4%	99.2%
	3000					91%	94.2%	99.2%

Table D2: $x_0 = 0.3, \boldsymbol{m}^*(0.3) = (1.95, 3.9, 4.55)$

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0.3)$	1000	1.95	2e-5	0.11	0.11	89.4% (85%)	93.4% (90.4%)	97.2% (96.2%)
	2000	1.98	8e-4	0.05	0.05	90.2% (89.8%)	95% (94.2%)	98.6% (98.2%)
	3000	1.95	2e-6	0.04	0.04	89% (88.8%)	94.2% (92.8%)	99% (98.6%)
$\hat{m}_2(0.3)$	1000	3.73	0.03	0.87	0.90	90.2% (85%)	93.8% (90.2%)	97% (95.6%)
	2000	3.72	0.03	0.38	0.41	92% (86.8%)	95.2% (94%)	99% (98.8%)
	3000	3.81	0.01	0.27	0.28	89.6% (89.4%)	95.6% (95.6%)	99.6% (98.4%)
$\hat{m}_3(0.3)$	1000	4.55	1e-5	0.27	0.27	92.4% (86.2%)	95.4% (91%)	97.6% (95.4%)
	2000	4.56	2e-4	0.13	0.13	93.8% (89%)	97.4% (94.2%)	99.2% (97.8%)
	3000	4.53	3e-4	0.08	0.08	93% (90.6%)	97.8% (95.6%)	99.2% (98.8%)
$\mathcal{J}_x$	1000					88.6%	94.6%	99%
	2000					88.8%	95.2%	99.2%
	3000					92%	95.6%	99%
$\mathcal{J}_{SP}$	1000					90.2%	94.2%	97.4%
	2000					91.4%	96.8%	99%
	3000					90.8%	96.2%	99.2%

Table D3: Different Strengths of (Z, X)

(α, β)	min. eig.		Average	Bias^2	Variance	MSE	90%	95%	99%
(0.8,0.4)	0.02	$\hat{m}_1(0)$	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
		$\hat{m}_2(0)$	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
		$\hat{m}_3(0)$	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
(0.16,0.08)	1e-4	$\hat{m}_1(0)$	1.48	3e-4	13.10	13.10	93% (90.6%)	96.4% (95.6%)	99.2% (99%)
		$\hat{m}_2(0)$	3.11	0.01	50.82	50.83	93.2% (89.2%)	95.6% (94.6%)	99.4% (98%)
		$\hat{m}_3(0)$	3.44	0.004	11.04	11.05	91.8% (90.6%)	95.2% (96%)	99.2% (99%)

Table D4: Different Degree of Endogeneity

	ρ	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0)$	0.3	1.50	2e-5	0.06	0.06	92.4% (89.8%)	96.8% (95.4%)	99.6% (99%)
	0.5	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
	0.7	1.51	4e-5	0.05	0.05	91.2% (89.2%)	95.4% (93.8%)	99.6% (97.6%)
$\hat{m}_2(0)$	0.3	2.88	0.01	0.38	0.39	91.8% (86.4%)	96.4% (93.4%)	99.4% (97.4%)
	0.5	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
	0.7	2.88	0.01	0.35	0.37	91% (87.6%)	96.2% (93.6%)	99% (98%)
$\hat{m}_3(0)$	0.3	3.49	5e-5	0.12	0.12	93% (90.8%)	96.8% (95.2%)	99.2% (98.6%)
	0.5	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
	0.7	3.49	2e-4	0.11	0.11	92% (89.8%)	96.8% (95.6%)	99.4% (98.8%)

Appendix E Proofs of Results in Section 6

In Section E.1 we prove some of the asymptotic results in Section 6. As the results of the matching point (the unique solution case) and of the separable model are standard, we only include proofs for the multiple matching points case (Theorem Cons-MP-Set) and the results of the nonseparable model estimator. Section E.2 contains used in Section E.1.

E.1 Asymptotic Properties

Let us first introduce the following lemmas. It will be needed in deriving the asymptotic properties for all the estimators.

Lemma E1. *Suppose* $h_0/h_g \to 0$. *Let* $S_0(Y|d,x)$ *be an interior set of* $S(Y|d,x)$. *Under*

Assumptions Reg-K, Reg-L, Reg-MP, Reg-SP, and Reg-NSP,

$$\sup_{x\in S_0(X)} \left|\hat{p}_d(x,z) - p_d(x,z))\right| = O_p\Big(\sqrt{\frac{\log(n)}{nh_x}} + h_x^2\Big) \tag{E.1}$$

$$\sup_{x\in S_0(X)} \left|\hat{\mathbb{E}}_{Y|DXZ}(d,x,z) - \mathbb{E}_{Y|XZ}(d,x,z)\right| = O_p\Big(\sqrt{\frac{\log(n)}{nh_m}} + h_m^2\Big) \tag{E.2}$$

$$\sup_{\substack{y\in S(Y|d)\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| = O_p\Big(\sqrt{\frac{\log(n)}{nh_g}} + h_g\Big) \tag{E.3}$$

$$\sup_{x\in S_0(X)} \left|\partial_x\hat{p}_d(x,z) - \partial_x p_d(x,z)\right| = o_p(1) \tag{E.4}$$

$$\sup_{x\in S_0(X)} \left|\partial_x\hat{\mathbb{E}}_{Y|DXZ}(d,x,z) - \partial_x\mathbb{E}_{Y|XZ}(d,x,z)\right| = o_p(1) \tag{E.5}$$

$$\sup_{\substack{y\in S_0(Y|d,x)\\ x\in S_0(X)}} \left|\partial_y\hat{F}_{Y|DXZ}(y|d,x,z) - f_{Y|DXZ}(y|d,x,z)\right| = o_p(1) \tag{E.6}$$

$$\sup_{\substack{y\in S_0(Y|d,x)\\ x\in S_0(X)}} \left|\partial_x\hat{F}_{Y|DXZ}(y|d,x,z) - \partial_x F_{Y|DXZ}(y|d,x,z)\right| = o_p(1) \tag{E.7}$$

The first three results are needed for consistency and the rate of convergence for each estimator proposed in the paper. The last four are needed to derive the rate of convergence as well and the asymptotic distributions. Tthese results are standard except equation (E.3). The bias term is $O_p(h_g)$ instead of the standard $O(h_g^2)$. This is due to the nonsmoothness of $F_{Y|DXZ}(\cdot|d,x,z)$ at the boundaries. We prove it in Appendix E.2.

The Matching Points

Proof of Theorem Cons-MP-Set. As the estimator is identical to Chernozhukov, Hong and Tamer (2007), we prove the theorem by verifying the conditions needed for their Theorem 3.1. Specifically, let $\tilde{Q}_x(\boldsymbol{x}) \equiv \hat{Q}_x(\boldsymbol{x}) - \inf_{\boldsymbol{x}\in S_0^2(X)} \hat{Q}_x(\boldsymbol{x})$. We need to verify that (a) $\sup_{\boldsymbol{x}\in S_0^2(X)} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})| = O_p\big(\frac{\log(n)}{nh_x}\big)$ and (b) there exist positive (δ,κ) such that for any $\varepsilon\in(0,1)$ there are $(\kappa_\varepsilon, n_\varepsilon)$ such that for all $n > n_\varepsilon$, $\tilde{Q}_x(\boldsymbol{x}) \geq \kappa[\rho(\boldsymbol{x},\mathcal{X}_m)\wedge\delta]^2$ uniformly on $\Delta \equiv \{\boldsymbol{x}\in S_0^2(X) : \rho(\boldsymbol{x},\mathcal{X}_m) \geq \sqrt{\frac{\kappa_\varepsilon \log(n)}{nh_x}}\}$ with probability at least $1-\varepsilon$.

We first derive the uniform rate of convergence for $\tilde{Q}_x(\boldsymbol{x})$. Let $\boldsymbol{x}_{mn}$ be such that

$\hat{Q}_x(\boldsymbol{x}_{mn}) = \inf_{S_0^2(X)} \hat{Q}_x(\boldsymbol{x})$. Then

$$\begin{aligned}\sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - \hat{Q}_x(\boldsymbol{x}_{mn}) - Q_x(\boldsymbol{x})\right| \leq & \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| + \left|\hat{Q}_x(\boldsymbol{x}_{mn}) - Q_x(\boldsymbol{x}_m)\right| \\ \leq & \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| + \hat{Q}_x(\boldsymbol{x}_m) - Q_x(\boldsymbol{x}_m) \\ \leq & 2 \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| \\ = & O_p\Big(\frac{\log(n)}{nh_x}\Big)\end{aligned}$$

where we used the definition of $\boldsymbol{x}_{mn}$, nonnegativity of $\hat{Q}_x$ and $Q_x(\boldsymbol{x}_m) = 0$. The last inequality is a consequence of Lemma E1 and the choice of h_m.

For (b), note that by Assumption Reg-MP, there exists $C > 0$ such that $Q_x(\boldsymbol{x}) \geq C\kappa[\rho(\boldsymbol{x}, \mathcal{X}_m) \wedge \delta]^2$ uniformly on Δ by continuity of Q_x and compactness of $S_0^2(\boldsymbol{X}) \cap \Delta$.

$$\begin{aligned}\inf_{\boldsymbol{x}\in\Delta} \tilde{Q}_x(\boldsymbol{x}) = & \inf_{\boldsymbol{x}\in\Delta} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x}) + Q_x(\boldsymbol{x})| \\ \geq & \inf_{\boldsymbol{x}\in\Delta} |Q_x(\boldsymbol{x})| - \sup_{\boldsymbol{x}\in\Delta} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})| \\ \geq & C\kappa[\rho(\boldsymbol{x}, \mathcal{X}_m) \wedge \delta]^2 - O_p(\frac{\log(n)}{nh_x})\end{aligned}$$

Therefore, we can choose $(\kappa_\varepsilon, n_\varepsilon)$ large enough so that the desired inequality holds uniformly on Δ. □

The Nonseparable Model-NSP

Let us begin with Theorem Cons-NSP. We need the following lemmas. Let x_m be a generic matching point.

Lemma E2. *Let $r_n = O_p\Big(\sqrt{\frac{\log(n)}{nh_g}} + h_g\Big)$. Suppose $|\hat{x}_m - x_m| = O_p(a_n)$. Under the conditions in Theorem Cons-NSP and Lemma E1, we have the following for all $d \in S(D)$:*

$$\sup_{y\in[\underline{y}_d, \bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y; \hat{x}_m, z')|d, \hat{x}_m, z) - F_{Y|DXZ}(\varphi_d(y; x_m, z')|d, x_m, z)\right| = O_p(r_n + a_n)$$

The main challenge to show this lemma is that the supremum is taken on $S(Y|d) \equiv [\underline{y}_d, \bar{y}_d]$, which contains the support set $S(Y|d, x_0)$ and $S(Y|d, x_m)$ as two subsets. By definition, $\varphi_d(y; x_m, z')$ is not unique when y is outside $S(Y|d, x_0)$, so it is not valid to show uniform consistent of $\hat{\varphi}_d$. The key lies in the fact that when $\varphi_d(y; x_m, z')$ is not unique, $F_{Y|DXZ}(\varphi_d(y; x_m, z')|d, x_m, z)$ is still unique (equal to 0 or 1). So instead of showing it by

establishing uniform convergence for $\hat{F}_{Y|DXZ}$ and $\hat{\varphi}_d$ separately, we treat it as one object.

Under Lemma E2, it is straightforward that $\sup_{\boldsymbol{y}\in\prod_d S(Y|d)} |\hat{Q}_{NSP}(\boldsymbol{y},u) - Q_{NSP}(\boldsymbol{y},u)| = o_p(1)$. Then we have the following lemmas.

Lemma E3. *Under all the conditions in Lemma E2, the following holds:*

$$\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \Big| \frac{1}{J}\sum_{j=1}^{J} \hat{Q}_{NSP}(\boldsymbol{g}(u_j),u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du \Big| = o_p(1)$$

Lemma E4. *There exists $\boldsymbol{g}_0 \in \hat{\mathcal{G}}$ such that $|\boldsymbol{g}_0(u) - \boldsymbol{g}^*(x_0,u)| = o(1)$ for all $u \in (0,1)$.*

Proof of Theorem Cons-NSP. Suppose there exists $\delta > 0$ such that $\sup_{u\in\mathcal{U}_0} |\hat{\boldsymbol{g}}(x_0,u) - \boldsymbol{g}^*(x_0,u)| > \delta$. By construction, $\hat{\boldsymbol{g}} \in \mathcal{G}_0$. By Theorem ID-Sup, there exists $\varepsilon > 0$ such that

$$\Big(\int_0^1 Q_{NSP}(\hat{\boldsymbol{g}}(x_0,u),u)du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0,u),u)du \Big) > \varepsilon$$

For simplicity, denote the sample objective function by $\hat{\mathcal{L}}$. By Lemma E3 and the rate of λ, $\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(\boldsymbol{g}(u)) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du| = o_p(1)$. Then

$$\begin{aligned}
&\Big(\int_0^1 Q_{NSP}(\hat{\boldsymbol{g}}(x_0,u),u)du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0,u),u)du \Big) \\
\leq& \hat{\mathcal{L}}(\hat{\boldsymbol{g}}(x_0,u)) - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0,u),u)du \\
&+ \int_0^1 Q_{NSP}(\hat{\boldsymbol{g}}(x_0,u),u)du - \hat{\mathcal{L}}(\hat{\boldsymbol{g}}(x_0,u)) \\
\leq& \hat{\mathcal{L}}(\boldsymbol{g}_0(u)) - \int_0^1 Q_{NSP}(\boldsymbol{g}_0(u),u)du \\
&+ \int_0^1 Q_{NSP}(\boldsymbol{g}_0(u),u)du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0,u),u)du \\
&+ \sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(\boldsymbol{g}(u)) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du| \\
\leq& 2\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(\boldsymbol{g}(u)) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du| + \int_0^1 (Q_{NSP}(\boldsymbol{g}_0(u),u) - Q_{NSP}(\boldsymbol{g}^*(x_0,u),u))du \\
=& o_p(1)
\end{aligned}$$

where $\boldsymbol{g}_0$ is defined in Lemma E4. The last inequality follows from Lemma E4 and the dominated convergence theorem. □

Proof of Theorem RoC-NSP. It is straightforward that

$$\max_{u_j\in\mathcal{U}_0} |\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)| \leq \sqrt{\sum_{u_j\in\mathcal{U}_0} (|\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)|)'(|\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)|)}$$

We derive the rate of convergence for the right hand side.

By Theorem Cons-NSP, $\hat{\boldsymbol{g}}(\cdot)$ on $\mathcal{U}_0$ in the interior of $S(Y|d,x_0)$ with probability approaching one. Under this event, $\Psi(\cdot)$ is differentiable at $\hat{\boldsymbol{g}}(x_0,u_j)$ and thus

$$\Psi(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j)) = \tilde{\Pi}_{NSP}(u_j)\cdot(\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)) \tag{E.8}$$

where $\tilde{\Pi}_{NSP}(u_j)$ is the Jacobian evaluated at the mean value. Again, by uniform convergence and the full rank condition in Theorem ID-NSP, $\tilde{\Pi}_{NSP}(u_j)$ is full rank uniformly in $u_j \in \mathcal{U}_0$. Then,

$$\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j) = \Big(\tilde{\Pi}'_{NSP}(u_j)\tilde{\Pi}_{NSP}(u_j)\Big)^{-1}\tilde{\Pi}'_{NSP}(u_j)\cdot\Big(\Psi(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j)\Big)$$

Therefore, by boundedness of all the conditional densities, there exists a constant $C>0$ such that

$$\begin{aligned}
&\sum_{u_j\in\mathcal{U}_0}(\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j))'(\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)) \\
\leq& C\sum_{u_j\in\mathcal{U}_0}\Big(\Psi(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j)\Big)'W_g(u_j)\Big(\Psi(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j)\Big)
\end{aligned} \tag{E.9}$$

Add and subtract $\hat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))$, the right hands side of inequality (E.9) can be expanded to the sum of the following three terms for some $C',C'',C'''>0$:

$$\begin{aligned}
(A):\quad &\sum_{u_j\in\mathcal{U}_0}\Big(\hat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big)'W_g(u_j)\Big(\hat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big) \\
\leq& C'J\sup_{\boldsymbol{y}\in S(Y|d)^3}\Big(\hat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\Big)'\Big(\hat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\Big) \\
=& O(Jr_n^2)
\end{aligned} \tag{E.10}$$

where the last inequality is from Lemma E2 and the rate condition $h_g/h_x \to 0$.

$$
\begin{aligned}
(B): \quad & \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\Big)' W_g(u_j) \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\Big) \\
&= \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j)\Big)' W_g(u_j) \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\Big) \\
&\le \sum_{j=1}^{J} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j)\Big)' W_g(u_j) \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\Big) \\
&\quad + \lambda \sum_{j=2}^{J} \Big(\hat{\boldsymbol{g}}(x_0,u_j) - \hat{\boldsymbol{g}}(x_0,u_{j-1})\Big)' \Big(\hat{\boldsymbol{g}}(x_0,u_j) - \hat{\boldsymbol{g}}(x_0,u_{j-1})\Big) \\
&\le \sum_{j=1}^{J} \Big(\widehat{\Psi}(\boldsymbol{g}_0(u_j)) - \boldsymbol{u}_j)\Big)' W_g(u_j) \Big(\widehat{\Psi}(\boldsymbol{g}_0(u_j)) - \boldsymbol{u}_j\Big) \\
&\quad + \lambda \sum_{j=2}^{J} \Big(\boldsymbol{g}_0(u_j)) - \boldsymbol{g}_0(u_{j-1}))\Big)' \Big(\boldsymbol{g}_0(u_j)) - \boldsymbol{g}_0(u_{j-1}))\Big)
\end{aligned}
$$

where the first inequality is due to non-negativity of the penalty. The second inequality is by the definition of the estimator. $\boldsymbol{g}_0$ is the same as in Lemma E4. In particular, let $\boldsymbol{g}_0(u_j) = \boldsymbol{g}^*(x_0,u_j)$. Then the right hand side of the last inequality is

$$
C'' J \sup_{\boldsymbol{y} \in S(Y|d)^3} \Big(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\Big)' \Big(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\Big) + \lambda/J = O(J r_n^2) \tag{E.11}
$$

$$
\begin{aligned}
(C): \quad & 2 \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big)' W_g(u_j) \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\Big) \\
=& 2 \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big)' W_g(u_j) \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big) \\
& + 2 \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big)' W_g(u_j) \Big(\Psi(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\Big) \\
=& 2 \cdot (A) + 2 \sum_{u_j \in \mathcal{U}_0} \Big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0,u_j))\Big)' W_g(u_j) \tilde{\Pi}_{NSP} \cdot \Big(\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j)\Big) \\
\le& 2 \cdot (A) + 2 \sqrt{\sum_{j=1}^{J} C''' \cdot (A)} \cdot \sqrt{\sum_{u_j \in \mathcal{U}_0} (\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j))' (\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j))} \\
=& O_p(J r_n^2) + O_p(\sqrt{J} r_n) \sqrt{\sum_{u_j \in \mathcal{U}_0} (\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j))' (\hat{\boldsymbol{g}}(x_0,u_j) - \boldsymbol{g}^*(x_0,u_j))}
\end{aligned} \tag{E.12}
$$

where the inequality follows from the Cauchy-Schwartz inequality.

Combining equations (E.8), (E.9), (E.10), (E.11) and (E.12), we have the desired result.

□

Proof of Theorem AsymDist-NSP. Fix $u_0 \in \mathcal{U}_0$. By Corollary RoC-NSP, $\hat{\boldsymbol{g}}(x_0, u_0)$ is in the interior so satisfies the first order condition. Under the rates of the bandwidths, the estimated propensity scores, the matching points and the penalty converge faster than $1/\sqrt{nh_g}$. By Lemma E1, under some manipulation, we obtain the following expansion for $\hat{\boldsymbol{g}}(x_0, u_0) - \boldsymbol{g}^*(x_0, u_0)$:

$$\hat{\boldsymbol{g}}(x_0,u_0) - \boldsymbol{g}^*(x_0,u_0) = -\Big(\Pi'_{NSP}\boldsymbol{W}_g(u_0)\Pi_{NSP}\Big)^{-1} \cdot \Pi'_{NSP}\boldsymbol{W}_g(u_0) \cdot \Big[\Big(\hat{\Psi}_{NSP}\Big(\boldsymbol{g}^*(\boldsymbol{x}_0,u_0)\Big) - \boldsymbol{u}\Big) + \begin{pmatrix} 0 \\ 0 \\ \sum_{d=1}^3 \phi_{d1}\Big(\hat{F}_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,0) - \hat{F}_{Y|DXZ}(g_d^*(x_{m1},u_0)|d,x_{m1},1)\Big) \\ \sum_{d=1}^3 \phi_{d2}\Big(\hat{F}_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,1) - \hat{F}_{Y|DXZ}(g_d^*(x_{m2},u_0)|d,x_{m2},0)\Big) \end{pmatrix} + o_p\Big(\frac{1}{\sqrt{Nh_g}}\Big)\Big]$$

where Π_{NSP}, ϕ_{d1}, and ϕ_{d2} are as defined in Section 6.3. Recall that $\Psi_{NSP}\Big(\boldsymbol{g}^*(\boldsymbol{x}_0,u_0)\Big) = \boldsymbol{u}$, $F_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,0) = F_{Y|DXZ}(g_d^*(x_{m1},u_0)|d,x_{m1},1)$ and $F_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,1) = F_{Y|DXZ}(g_d^*(x_{m2},u_0)|d,x_{m2},0)$. Also, by Lemma E1, the denominator of $\hat{F}_{YDXZ}(y,d,x,z)$ converges in probability to $f_{Y|DXZ}(d,x,z)$. Let

$$\mathbb{G}_d(y,x,z) \equiv \hat{F}_{YDXZ}(y,d,x,z) - F_{Y|DXZ}(y|d,x,z)\hat{f}_{DXZ}(d,x,z).$$

Then the asymptotic distribution is determined by the following vector:

$$\begin{pmatrix} \sum_{d=1}^3 \frac{\mathbb{G}_d(g_d^*(x_0,u_0),x_0,0)}{f_{DXZ}(d,x_0,0)} \\ \sum_{d=1}^3 \frac{\mathbb{G}_d(g_d^*(x_0,u_0),x_0,1)}{f_{DXZ}(d,x_0,1)} \\ \sum_{d=1}^3 \Big[\frac{\mathbb{G}_d(g_d^*(x_{m1},u_0),x_{m1},0)}{f_{DXZ}(d,x_{m1},0)} + \phi_{d1}\frac{\mathbb{G}_d(g_d^*(x_0,u_0),x_0,0)}{f_{DXZ}(d,x_0,0)} - \phi_{d1}\frac{\mathbb{G}_d(g_d^*(x_{m1},u_0),x_{m1},1)}{f_{DXZ}(d,x_{m1},1)}\Big] \\ \sum_{d=1}^3 \Big[\frac{\mathbb{G}_d(g_d^*(x_{m2},u_0),x_{m2},1)}{f_{DXZ}(d,x_{m2},1)} + \phi_{d1}\frac{\mathbb{G}_d(g_d^*(x_0,u_0),x_0,1)}{f_{DXZ}(d,x_0,1)} - \phi_{d1}\frac{\mathbb{G}_d(g_d^*(x_{m2},u_0),x_{m2},0)}{f_{DXZ}(d,x_{m2},0)}\Big] \end{pmatrix}$$

The variance of each $\mathbb{G}_d$ follows Theorem 2.2 in Li and Racine (2008):

$$\mathbb{V}(\mathbb{G}_d(y,x,z)) = \frac{\kappa f_{DXZ}(d,x,z)F_{Y|DXZ}(y|d,x,z)\cdot(1-F_{Y|DXZ}(y|d,x,z))}{nh_g} + o(\frac{1}{nh_g})$$

Now let us derive the covariances: $\mathbb{C}\Big(\mathbb{G}_d(y,x,z), \mathbb{G}_{d'}(y',x',z')\Big)$. By i.i.d., the covariance is

equal to

$$\frac{1}{nh_g^2}\mathbb{E}\Big[L(\frac{y-Y_i}{h_0})L(\frac{y'-Y_i}{h_0})K(\frac{X_i-x}{h_g})K(\frac{X_i-x'}{h_g})\mathbb{1}(D_i=d)\mathbb{1}(D_i=d')\mathbb{1}(Z_i=z)\mathbb{1}(Z_i=z')\Big]+O_p(\frac{1}{n})$$

where the $O_p(\frac{1}{n})$ term arises because the bias is of the order h_g^2 by Lemma A.2 in Li and Racine (2008). It is clear that when $z' \neq z$ or $d' \neq d$, the leading term is 0. When $x \neq x'$, for large enough n, $|x'-x| > 2h_g$, and thus $\left|\left|\frac{X_i-x}{h_g}\right| - \left|\frac{X_i-x'}{h_g}\right|\right| > 2$. As $K(\cdot) = 0$ outside $[-1,1]$, for any X_i, one of the K functions must be 0. Therefore, all the covariances are of the order of $O_p(\frac{1}{n})$ which is $o_p(\frac{1}{nh_g})$.

By Lyapunov's Central Limit Theorem and the delta method, we obtain the asymptotic distribution in the theorem. □

E.2 Proofs of Lemmas

Lemma E1

We only show equation (E.3) as others are standard in the literature of kernel estimation (e.g. Mack and Silverman (1982), Silverman (1986), Härdle, Marron and Wand (1990), Masry (1996)), etc.). Let us first prove the following lemma which generalizes Lemma A.5 in Li and Racine (2008) to a bounded Y.

Lemma E5. *Under the conditions in Lemma E1, we have the following equations for all d and z:*

$$\sup_{\substack{y\in[\underline{y}_{dx},\bar{y}_{dx}]\\ x\in S_0(X)}}\Big|\mathbb{E}\Big(L(\frac{y-Y_i}{h_0})|D_i=d,X_i=x,Z_i=z\Big)-F_{Y|DXZ}(y|d,x,z)\Big| = O(h_0) \tag{E.13}$$

Proof of Lemma E5. By definition,

$$\mathbb{E}\Big(L(\frac{y-Y_i}{h_0})|D_i=d,X_i=x,Z_i=z\Big)=\int_{\underline{y}_{dx}}^{\bar{y}_{dx}}L(\frac{y-y'}{h_0})dF_{Y|DXZ}(y'|d,x,z)$$

Let $\nu = \frac{y-y'}{h_0}$, by the rule of change-of-variable and integration by part, the right hand side

can be rewritten as

$$
\begin{aligned}
&-\int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} L(\nu) dF_{Y|DXZ}(y-\nu h_0|d,x,z) \\
=&L(\frac{y-\bar{y}_{dx}}{h_0}) + \int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} L'(\nu) F_{Y|DXZ}(y-\nu h_0|d.x,z)d\nu \\
=&L(\frac{y-\bar{y}_{dx}}{h_0}) + F_{Y|DXZ}(y|d,x,z)\Big(L(\frac{y-\underline{y}_{dx}}{h_0}) - L(\frac{y-\bar{y}_{dx}}{h_0})\Big) \\
&+ h_0 \int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} \nu f_{Y|DXZ}(y-\tilde{\nu}h_0|d,x,z)L'(\nu)d\nu
\end{aligned}
$$

where $\tilde{\nu}$ is a mean value between 0 and ν. Rearrange the terms, we obtain

$$
\begin{aligned}
&\Big|\mathbb{E}\Big(L(\frac{y-Y_i}{h_0})|D_i=d,X_i=x,Z_i=z\Big) - F_{Y|DXZ}(y|d,x,z)\Big| \\
\leq&\Big|L(\frac{y-\bar{y}_{dx}}{h_0})\Big(1-F_{Y|DXZ}(y|d,x,z)\Big) - \Big(1-L(\frac{y-\underline{y}_{dx}}{h_0})\Big)F_{Y|DXZ}(y|d,x,z)\Big| \\
&+ h_0 \sup_{(y,x)\in S(Y,X|d)} f_{Y|DXZ}(y|d,x,z)\int_0^1 \nu L'(\nu)du
\end{aligned}
$$

where the last term is $O(h_0)$. For the first term, if $\underline{y}_{dx}+h_0 < y < \bar{y}_{dx}-h_0$, then $L(\frac{y-Y_i}{h_0}) = 0$ and $\Big(1-L(\frac{y-\underline{y}_{dx}}{h_0})\Big) = 0$. Now we consider the remaining cases. For $\underline{y}_{dx} \leq y \leq \underline{y}_{dx}+h_0$, $L(\frac{y-Y_i}{h_0}) = 0$. Then the first term is bounded by $F_{Y|DXZ}(\underline{y}_{dx}+h_0|dxz)$, which is $O(h_0)$ by the mean value theorem. Similarly, if $\bar{y}_{dx}-h_0 \leq y \leq \bar{y}_{dx}$, then the same term is bounded by $1-F_{Y|DXZ}(\bar{y}_{dx}-h_0|dxz)$, which is again $O(h_0)$. As Y's conditional density is uniformly bounded, these bounds are uniform on $S(Y,X|d)$ □

Remark E1. *The rate in Li and Racine (2008) is $O(h_0^2)$, faster than the rate here. Intuitively, this is because at the boundaries, L systematically overestimate (at the lower bound) or underestimate (at the upper bound) the CDF, thus introducing larger bias.*

Now we are ready to show equation (E.3).

Proof of Equation (E.3) in Lemma E1. By construction, $\hat{F}_{Y|DXZ}(\cdot|d,x,z) \in [0,1]$ and is increasing. Therefore,

$$
\begin{aligned}
&\sup_{\substack{y\in S(Y|d)\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &\sup_{x\in S_0(X)} \sup_{y\leq \underline{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{x\in S_0(X)} \sup_{y\geq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
= &\sup_{x\in S_0(X)} \sup_{y\leq \underline{y}_{dx}} \hat{F}_{Y|DXZ}(y|d,x,z) + \sup_{x\in S_0(X)} \sup_{y\geq \bar{y}_{dx}} \left(1 - \hat{F}_{Y|DXZ}(y|d,x,z)\right) \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &\sup_{x\in S_0(X)} \hat{F}_{Y|DXZ}(\underline{y}_{dx}|d,x,z) + \sup_{x\in S_0(X)} \left(1 - \hat{F}_{Y|DXZ}(\bar{y}_{dx}|d,x,z)\right) \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &3 \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
= &3 \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\frac{\hat{F}_{YDXZ}(y,d,x,z) - F_{Y|DXZ}(y|d,x,z)\hat{f}_{DXZ}(d,x,z)}{\hat{f}_{DXZ}(d,x,z)}\right|
\end{aligned}
$$

We now show that the right hand side of the last inequality has the desired rate. As the denominator is independent of y, and it's uniformly consistent for $f_{DXZ}(d,x,z)$ on the interior set $S_0(X)$ under the regularity conditions, it's infimum is bounded away from 0.

For the numerator, denote $\hat{q}(y,x) \equiv \hat{F}_{YDXZ}(y,d,x,z) - F_{Y|DXZ}(y|d,x,z)\hat{f}_{DXZ}(d,x,z)$, then

$$
\sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{q}(y,x)\right| \leq \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\mathbb{E}(\hat{q}(y,x))\right| + \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{q}(y,x) - \mathbb{E}(\hat{q}(y,x))\right|
$$

We derive the rate of convergence for each of the two terms.

By the law of iterated expectation,

$$
\left|\mathbb{E}(\hat{q}(y,x))\right| = \left|\frac{1}{h_g}\int \left[\mathbb{E}_{Y|DXZ}\Big(L(\frac{y-Y_i}{h_0})|d,x',z\Big) - F_{Y|DXZ}(y|d,x,z)\right] K(\frac{x'-x}{h_g}) f_X(x')dx'\right|
$$

By Lemma E5, $\mathbb{E}_{Y|DXZ}\Big(L(\frac{y-Y_i}{h_0})|d,x',z\Big) = F_{Y|DXZ}(y|d,x',z) + O(h_0)$ uniformly. Therefore,

$\Big|\mathbb{E}(\hat{q}(y,x))\Big|$ is uniformly bounded by

$$\Big|\frac{1}{h_g}\int\Big[F_{Y|DXZ}(y|d,x',z)-F_{Y|DXZ}(y|d,x,z)\Big]K(\frac{x'-x}{h_g})f_X(x')dx'\Big|+O(h_0)$$

Recall the supremum is taken on $[\underline{y}_{dx},\bar{y}_{dx}]$, note that it may be the case that y is outside the support for some x'' in between of x and x'. Then $F_{Y|DXZ}(y|d,\cdot,z)$ is nondifferentiable, making the second order Taylor expansion invalid at x. However, by Lipschitz continuity of $F_{Y|DXZ}(y|d,\cdot,z)$, we can still bound the CDF difference:

$$\begin{aligned}
&\sup_{x\in S_0(X)}\sup_{\underline{y}_{dx}\le y\le\bar{y}_{dx}}\Big|\frac{1}{h_g}\int\Big[F_{Y|DXZ}(y|d,x',z)-F_{Y|DXZ}(y|d,x,z)\Big]K(\frac{x'-x}{h_g})f_X(x')dx'\Big|\\
\le& C\sup_{x\in S_0(X)}\int\Big|\frac{x'-x}{h_g}\Big|\cdot K(\frac{x'-x}{h_g})f_X(x')dx'\\
\le& h_gC'\int_0^1\nu K(\nu)d\nu\\
=& O(h_g)
\end{aligned}$$

for some $C,C'>0$.

Next let us derive the rate of $\sup_{(x,y)\in S_0(X,Y|d)}\Big|\hat{q}(y,x)-\mathbb{E}(\hat{q}(y,x))\Big|$ where for the ease of notation, $S_0(X,Y|d)\equiv\{(x,y):x\in S_0(X),\underline{y}_{dx}\le y\le\bar{y}_{dx}\}$. As $S_0(X,Y|d)$ is compact, it can be covered by $T_n<\infty$ squares $I_1,I_2,...,I_{T_n}$ with length τ_n where $\tau_n\propto O(\sqrt{1/T_n})$. Let the center in each square be (y_k,x_k) $(k=1,2,...,T_n)$, then

$$\begin{aligned}
&\sup_{(x,y)\in S_0(X,Y|d)}\Big|\hat{q}(y,x)-\mathbb{E}(\hat{q}(y,x))\Big|\\
\le&\max_{1\le k\le T_n}\sup_{S_0(X,Y|d)\cap I_k}\Big|\hat{q}(y,x)-\hat{q}(y_k,x_k)\Big|\\
&+\max_{1\le k\le T_n}\sup_{S_0(X,Y|d)\cap I_k}\Big|\mathbb{E}(\hat{q}(y,x))-\mathbb{E}(\hat{q}(y_k,x_k))\Big|\\
&+\max_{1\le k\le T_n}\Big|\hat{q}(y_k,x_k)-\mathbb{E}(\hat{q}(y_k,x_k))\Big|
\end{aligned}$$

Consider the first and the second terms. For the first term,

$$\hat{q}(y,x)-\hat{q}(y_k,x_k)=\frac{1}{nh_g}\sum_{i=1}^n\mathbb{1}(D_i=d)\mathbb{1}(Z_i=d)\Big(L(\frac{y-Y_i}{h_0})K(\frac{X_i-x}{h_g})-L(\frac{y_k-Y_i}{h_0})K(\frac{X_i-x_k}{h_g})\Big)$$

By Lipschitz continuity of $L(\cdot)$ and $K(\cdot)$, the right hand side is bounded by $\frac{C\tau_n}{h_gh_0}$ given $h_0<h_g$ uniformly in k. Similarly the second term is also bounded by $\frac{C\tau_n}{h_gh_0}$.

For the third term, let

$$W_i(y,x) = \frac{1}{nh_g}\Big[\Big(L(\frac{y-Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\Big)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)K(\frac{x-X_i}{h_g}) \\ - \mathbb{E}\Big(\Big(L(\frac{y-Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\Big)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)K(\frac{x-X_i}{h_g})\Big)\Big]$$

We consider the probability $\mathbb{P}\Big(\max_{1\le k\le T_n}\Big|\sum_{i=1}^n W_i(y_k,x_k)\Big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big)$.

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le k\le T_n}\Big|\sum_{i=1}^n W_i(y_k,x_k)\Big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big) \\
\le&\sum_{k=1}^{T_n}\mathbb{P}\Big(\Big|\sum_{i=1}^n W_i(y_k,x_k)\Big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big) \\
\le& T_n \sup_{(y,x)\in S_0(Y,X|d)}\mathbb{P}\Big(\Big|\sum_{i=1}^n W_i(y,x)\Big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big) \\
\le& T_n \sup_{(y,x)\in S_0(Y,X|d)}\Big(\mathbb{P}\Big(\sum_{i=1}^n W_i(y,x) > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big) + \mathbb{P}\Big(\sum_{i=1}^n W_i(y,x) < C_1\sqrt{\frac{\log(n)}{nh_g}}\Big)\Big) \\
\le& T_n \sup_{(y,x)\in S_0(Y,X|d)}\frac{\mathbb{E}\Big(\exp(a_n\sum_i W_i(y,x))\Big) + \mathbb{E}\Big(\exp(-a_n\sum_i W_i(y,x))\Big)}{\exp\Big(a_n C_1\sqrt{\frac{\log(n)}{nh_g}}\Big)}
\end{aligned}$$

where the last inequality follows from the Markov inequality for some $a_n > 0$. Since K and L are bounded, $|W_i(y,x)|$ is bounded. Let $a_n = \sqrt{\log(n)nh_g}$, then for large enough n, $a_n|W_i(y,x)| < 1/2$. Therefore, by the inequality $\exp(c) \le 1 + c + c^2$ for any $c \in [-1/2, 1/2]$ and $1 + c \le \exp(c)$ for $c \ge 0$, we have

$$\begin{aligned}
\mathbb{E}\Big(\exp(\pm a_n W_i(y,x))\Big) \le& 1 \pm \mathbb{E}(a_n W_i(y,x)) + \mathbb{E}(a_n^2 W_i^2(y,x))) \\
\le& \exp(\mathbb{E}a_n^2 W_i^2(y,x)))
\end{aligned}$$

since $\mathbb{E}(W_i(y,x))$ by construction. Therefore,

$$\begin{aligned}
&T_n \sup_{(y,x)\in S_0(Y,X|d)}\frac{\mathbb{E}\Big(\exp(a_n\sum_i W_i(y,x))\Big) + \mathbb{E}\Big(\exp(-a_n\sum_i W_i(y,x))\Big)}{\exp\Big(a_n C_1\sqrt{\frac{\log(n)}{nh_g}}\Big)} \\
\le& \frac{2T_n}{n^{C_1}}\cdot\sup_{(y,x)\in S_0(Y,X|d)}\exp(\log(n)nh_g\sum_i \mathbb{E}W_i^2(y,x)))
\end{aligned}$$

For $\mathbb{E}W_i^2(y,x)$, since $L(\cdot) \in [0,1]$, we have

$$\begin{aligned}\mathbb{E}W_i^2(y,x) \leq & \frac{1}{n^2 h_g^2}\mathbb{E}\Big[\Big(L(\frac{y-Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\Big)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)K(\frac{x-X_i}{h_g})\Big]^2 \\ \leq & \frac{1}{n^2 h_g^2}\mathbb{E}\Big[K^2(\frac{x-X_i}{h_g})\Big] \\ \leq & C_2\frac{1}{n^2 h_g}\end{aligned}$$

for some $C_2 > 0$. Therefore,

$$\frac{2T_n}{n^{C_1}} \cdot \sup_{(y,x)\in S_0(Y,X|d)} \exp(\log(n) n h_g \sum \mathbb{E}W_i^2(y,x))) \leq \frac{2T_n}{n^{C_1 - C_2}}$$

Let $T_n = \frac{n}{\log(n)h_0^2 h_g}$. Then for large enough C_1, there exists $\alpha \geq 2$,

$$\sum_{i=1}^{n} \mathbb{P}\Big(\max_{1\leq k\leq T_n} \Big|\hat{q}(y_k,x_k) - \mathbb{E}(\hat{q}(y_k,x_k))\Big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\Big) \leq \sum_n \frac{1}{n^\alpha} < \infty$$

Therefore, by the Borel-Cantelli lemma, $\max_{1\leq k\leq T_n}\Big|\hat{q}(y_k,x_k) - \mathbb{E}(\hat{q}(y_k,x_k))\Big| = O_p(\sqrt{\frac{\log(n)}{nh_g}})$.

With this choice of T_n, $\tau_n = O(\frac{\sqrt{\log(n)h_g}h_0}{\sqrt{n}})$, so the first two terms are $O_p(\sqrt{\frac{\log(n)}{nh_g}})$ as well. □

Lemma E2

Proof. By the triangle inequality,

$$
\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{\substack{y\in[\underline{y}_d,\bar{y}_d]\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{\substack{y\in[\underline{y}_d,\bar{y}_d]\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| + O_p(a_n)
\end{aligned}
$$

where the last inequality holds because of Lipschitz continuity of $F_{Y|DXZ}(\cdot|d,\cdot,z)$. Note the first term on the right hand side is $O_p(r_n)$ by Lemma E1. Now we only need to show that

$$
\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| = O_p(r_n+a_n)
$$

This is straightforward if $\hat{\varphi}_d(y)$ is uniformly consistent at that rate. However, as $[\underline{y}_d,\bar{y}_d]$ is larger than $[\underline{y}_{dx_0},\bar{y}_{dx_0}]$, $\hat{\varphi}_d(y)$ is not consistent when y is outside $[\underline{y}_{dx_0},\bar{y}_{dx_0}]$ because then $\varphi_d(y)$ is not unique. Let us prove the following equation first:

$$
\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| = O_p(r_n+a_n)
$$

By adding and subtracting $\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')$,

$$\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|F_{Y|DXZ}(\hat{\varphi}_d(y;x_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z')\right| \\
&+\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z')-F_{Y|DXZ}(y|d,x_0,z)\right|+O_p(r_n+a_n)
\end{aligned}$$

where we use $F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')=F_{Y|DXZ}(y|d,x_0,z)$ for the last inequality by Theorem MEQ. By adding and subtracting $\hat{F}_{Y|DXZ}(y|d,x_0,z)$,

$$\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(y|d,x_0,z)\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')-\hat{F}_{Y|DXZ}(y|d,x_0,z)\right|+O_p(r_n) \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z')-\hat{F}_{Y|DXZ}(y|d,x_0,z)\right|+O_p(r_n) \\
= &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right|+O_p(r_n) \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]}\left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z')-F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z')\right|+O_p(r_n+a_n) \\
=&O_p(r_n+a_n)
\end{aligned}$$

where the third line follows from the definition of $\hat{\varphi}_d(y;\hat{x}_m,z')$. Finally, let us establish the relation between $F_{Y|DXZ}(y'|d,x,z)-F_{Y|DXZ}(y|d,x,z)$ and $F_{Y|DXZ}(y'|d,x,z')-F_{Y|DXZ}(y|d,x,z')$ for any $y,y'\in[\underline{y}_d,\bar{y}_d]$. By the exogeneity condition of Z in Assumption E-NSP, $S(Y|d,x,z)=S(Y|d,x,z')=S(Y|d)$. Therefore, we have the following two equations:

$$F_{Y|DXZ}(y'|d,x,z)-F_{Y|DXZ}(y|d,x,z)=f_{Y|DXZ}(\tilde{y}_1|d,x,z)\Big[\Big(\underline{y}_{dx}\vee(y'\wedge\bar{y}_{dx})\Big)-\Big(\underline{y}_{dx}\vee(y\wedge\bar{y}_{dx})\Big)\Big],$$

and

$$F_{Y|DXZ}(y'|d,x,z')-F_{Y|DXZ}(y|d,x,z')=f_{Y|DXZ}(\tilde{y}_2|d,x,z')\Big[\Big(\underline{y}_{dx}\vee(y'\wedge\bar{y}_{dx})\Big)-\Big(\underline{y}_{dx}\vee(y\wedge\bar{y}_{dx})\Big)\Big].$$

where $\tilde{y}_1$ and $\tilde{y}_2$ are the mean values between $\Big(\underline{y}_{dx}\vee(y'\wedge\bar{y}_{dx})\Big)$ and $\Big(\underline{y}_{dx}\vee(y\wedge\bar{y}_{dx})\Big)$.

By Assumption Reg-NSP, $\frac{f_{Y|DXZ}(\tilde{y}_1|d,x,z)}{f_{Y|DXZ}(\tilde{y}_2|d,x,z')}$ is uniformly bounded from above on $S(Y|d)\times$

$S(Y|d)$. Therefore, there exists a constant $C > 0$ such that

$$\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| \\
\leq & C \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right|
\end{aligned}$$

This completes the proof. □

Lemma E3

Proof of Lemma E3. By the triangle inequality,

$$\begin{aligned}
&\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \left|\frac{1}{J}\sum_{j=1}^{J} \hat{Q}_{NSP}(\boldsymbol{g}(u_j),u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du\right| \\
\leq & \sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \left|\frac{1}{J}\sum_{j=1}^{J} \hat{Q}_{NSP}(\boldsymbol{g}(u_j),u_j) - \frac{1}{J}\sum_{j=1}^{J} Q_{NSP}(\boldsymbol{g}(u_j),u_j)\right| \\
&+ \sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \left|\frac{1}{J}\sum_{j=1}^{J} Q_{NSP}(\boldsymbol{g}(u_j),u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du\right| \\
\leq & \sup_{\boldsymbol{y}\in\prod_d S(Y|d)} \left|\hat{Q}_{NSP}(\boldsymbol{y},u_j) - Q_{NSP}(\boldsymbol{y},u_j)\right| \\
&+ \sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \left|\frac{1}{J}\sum_{j=1}^{J} Q_{NSP}(\boldsymbol{g}(u_j),u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du\right| \\
= & \sup_{\boldsymbol{g}\in\hat{\mathcal{G}}} \left|\frac{1}{J}\sum_{j=1}^{J} Q_{NSP}(\boldsymbol{g}(u_j),u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du\right| + o_p(1)
\end{aligned}$$

where the $o_p(1)$ in the last inequality follows from Lemma E2. We now show the remaining term is also $o_p(1)$.

$$
\begin{aligned}
&\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(\boldsymbol{g}(u_j),u_j)-\int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du\Big| \\
=&\sup_{\boldsymbol{g}\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(\boldsymbol{g}(u_j),u_j)-\sum_{j=1}^{J}\int_{\frac{j-1}{J}}^{\frac{j}{J}} Q_{NSP}(\boldsymbol{g}(u),u)du\Big| \\
=&\sup_{\boldsymbol{g}\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\frac{1}{J}Q_{NSP}(\boldsymbol{g}(u_j),u_j)-\int_{\frac{j-1}{J}}^{\frac{j}{J}} Q_{NSP}(\boldsymbol{g}(u),u)du\Big)\Big| \\
=&\sup_{\boldsymbol{g}\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\int_{\frac{j-1}{J}}^{\frac{1}{J}} Q_{NSP}(\boldsymbol{g}(u_j),u_j)du-\int_{\frac{j-1}{J}}^{\frac{j}{J}} Q_{NSP}(\boldsymbol{g}(u),u)du\Big)\Big| \\
\leq&\sup_{\boldsymbol{g}\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\int_{\frac{j-1}{J}}^{\frac{1}{J}} C\cdot(\boldsymbol{g}(u_j)-\boldsymbol{g}(u_{j-1}))du\Big)\Big| \\
\leq&\frac{1}{J}C\cdot\sum_{j=1}^{J}\Big|\boldsymbol{g}(u_j)-\boldsymbol{g}(u_{j-1})\Big|
\end{aligned}
$$

where in the second equality, $\hat{\mathcal{G}}$ is changed to $\mathcal{G}^*$ because for any $\boldsymbol{g}(u)$ that is outside $\prod_d S(Y|d,x_0)$ gives the same value of Q_{NSP} as that at the boundaries of $\prod_d S(Y|d,x_0)$. The third equality is due to the fact that $Q_{NSP}(\boldsymbol{g}(u_j),u_j)$ is a constant so $\int_{\frac{j-1}{J}}^{\frac{j}{J}} Q_{NSP}(\boldsymbol{g}(u_j),u_j)du = \frac{1}{J}Q_{NSP}(\boldsymbol{g}(u_j),u_j)$. The first inequality holds because for all values in $\prod_d S(Y|d,x_0)$, Q_{NSP} is differentiable and the derivative is uniformly bounded. Finally, by monotonicity, $\boldsymbol{g}(u_j)-\boldsymbol{g}(u_{j-1})>\mathbf{0}$. Hence

$$
\frac{1}{J}C\cdot\sum_{j=1}^{J}\Big|\boldsymbol{g}(u_j)-\boldsymbol{g}(u_{j-1})\Big|\leq\frac{1}{J}C\sum_{d}(\bar{y}_{dx_0}-\underline{y}_{dx_0})=O(\frac{1}{J})=o(1)
$$

□

Lemma E4

Proof of Lemma E4. For each u_j, let $\boldsymbol{g}_0(u_j)=\boldsymbol{g}^*(x_0,u_j)$. Then for any u between nodes u_{j-1} and u_j, by monotonicity,

$$
|\boldsymbol{g}_0(u)-\boldsymbol{g}^*(x_0,u)|\leq\boldsymbol{g}^*(x_0,u_j)-\boldsymbol{g}^*(x_0,u_{j-1})=O(\frac{1}{J})=o(1).
$$

□

References

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and confidence regions for parameter sets in econometric models." *Econometrica*, 75(5): 1243–1284.

Härdle, Wolfgang, J. S. Marron, and Matt Wand. 1990. "Bandwidth choice for density derivatives." *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1): 223–232.

Li, Qi, and Jeffrey S. Racine. 2008. "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data." *Journal of Business & Economic Statistics*, 26(4): 423–434.

Mack, Yue-pok, and Bernard W. Silverman. 1982. "Weak and strong uniform consistency of kernel regression estimates." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3): 405–415.

Masry, Elias. 1996. "Multivariate local polynomial regression for time series: uniform strong consistency and rates." *Journal of Time Series Analysis*, 17(6): 571–599.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.