Supplemental to "Identification and Estimation of Partial Effects with Proxy Variables"

Kenichi Nagasawa

September 30, 2020

A Proofs

A.1 Proof of Theorem 1

It suffices to show $\mathbb{E}[g(x,\varepsilon,\zeta)|V] = \mathbb{E}[Y|X=x,V]$. Let ρ denote a σ-finite measure dominating the distribution of ε given X, Z. Using the law of total probability and conditional independence $W \perp\!\!\!\perp (X,Z)|\varepsilon$,

$$\begin{aligned} f_{W|XZ}(w|x,z) &= \int f_{W|\varepsilon XZ}(w|e,x,z) f_{\varepsilon|XZ}(e|x,z) d\rho(e) \\ &= \int f_{W|\varepsilon}(w|e) f_{\varepsilon|XZ}(e|x,z) d\rho(e) \\ &\equiv \tilde{\Psi}\big(f_{\varepsilon|XZ}(\cdot|x,z)\big)(w). \end{aligned}$$

To see injectivity of this mapping, suppose $\tilde{\Psi}(g_1) = \tilde{\Psi}(g_2)$ for some functions g_1, g_2 such that g_1/f_ε and g_2/f_ε are bounded. Since $\tilde{\Psi}(g_\ell)(w) = f_W(w)\mathbb{E}[g_\ell(\varepsilon)/f_\varepsilon(\varepsilon)|W=w]$ $\ell = 1,2$, for w with $f_W(w) > 0$, $\mathbb{E}[g_1(\varepsilon)/f_\varepsilon(\varepsilon)|W] = \mathbb{E}[g_2(\varepsilon)/f_\varepsilon(\varepsilon)|W]$ a.s., and Assumption ID implies that $g_1(\varepsilon)/f_\varepsilon(\varepsilon) = g_2(\varepsilon)/f_\varepsilon(\varepsilon)$ a.s., and thus $g_1 = g_2$ a.s. Thus, $\tilde{\Psi}$ is injective and there exists a mapping Ψ such that

$$f_{\varepsilon|XZ}(e|x,z) = \Psi\big(f_{W|XZ}(\cdot|x,z)\big)(e). \tag{1}$$

Now,

$$\begin{aligned}\mathbb{E}[g(x,\varepsilon,\zeta)|V] &= \mathbb{E}[\mathbb{E}[g(x,\varepsilon,\zeta)|X,Z]|V] \\ &= \mathbb{E}\bigg[\int\int g(x,e,s)dF_{\zeta|\varepsilon}(s|e)f_{\varepsilon|XZ}(e|X,Z)d\rho(e)\bigg|V\bigg].\end{aligned}$$

and by (1),

$$\begin{aligned}\mathbb{E}[Y|X=x,V] &= \mathbb{E}\bigg[\int\int g(x,e,s)dF_{\zeta|\varepsilon}(s|e)f_{\varepsilon|XZ}(e|X,Z)d\rho(e)\bigg|X=x,V\bigg] \\ &= \mathbb{E}\bigg[\int\int g(x,e,s)dF_{\zeta|\varepsilon}(s|e)\Psi(V)(e)d\rho(e)\bigg|V\bigg] \\ &= \mathbb{E}\bigg[\int\int g(x,e,s)dF_{\zeta|\varepsilon}(s|e)f_{\varepsilon|XZ}(e|X,Z)d\rho(e)\bigg|V\bigg]\end{aligned}$$

and thus, $\mathbb{E}[g(x,\varepsilon,\zeta)|V] = \mathbb{E}[Y|X=x,V]$. □

A.2 Estimation results

In this section, I provide proofs for estimation results from the main paper as well as some additional results. In the sequel, I use C to denote a generic positive constant independent of the sample size that has different values from place to place.

A.2.1 Nonparametric estimation

Assumption Nonparametric Let $\bar{\mathcal{V}} = \{v := F_{W|XZ}(\cdot|x,z), (x,z) \in \overline{\mathcal{X}\mathcal{Z}}\}$ and $\|v\|_\infty = \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} \|v(\cdot|x,z)\|$.

(i) The observation $\{(Y_i, X_i, Z_i, W_i)\}_{i=1}^n$ is a random sample. The random variable X and at least one element of Z have joint Lebesgue density. The norm $\|\cdot\|$ is the L^2 norm with some finite measure that dominates the conditional distribution of W given (X,Z).

(ii) The set $\overline{\mathcal{X}\mathcal{Z}}$ is compact and the conditional moment satisfies $\mathbb{E}[|Y|^s|X,Z] \le C < \infty$ for some $s > 2$. Also, for any vanishing sequence $c_n \to 0$,

$$\lim_{n\to\infty} \sup_{|\tilde{x}-x_0|\vee\|\tilde{v}-v\|\le c_n} \big|\mathbb{E}[Y|X=\tilde{x},V=\tilde{v}] - \mathbb{E}[Y|X=x_0,V=v]\big| = 0$$

where $v, \tilde{v} \in \bar{\mathcal{V}}$ and for some $\alpha \in (0, 1]$,

$$\|F_{W|XZ}(\cdot|\tilde{x}, \tilde{z}) - F_{W|XZ}(\cdot|x, z)\| \leq C(|\tilde{x} - x|^\alpha + |\tilde{z} - z|^\alpha)$$

where $|\tilde{x} - x| + |\tilde{z} - z|$ is sufficiently small and C is independent of evaluation points.

(iii) The kernel function K is non-negative, compactly supported, and continuously differentiable. For some $d > d_x$,

$$0 < c \leq b_n^{-d}\mathbb{E}\Big[K\Big(\frac{X - x_0}{b_n}\Big)K\Big(\frac{\|V - v\|}{b_n}\Big)\Big] + \frac{\mathbb{E}\Big[K\Big(\frac{X-x_0}{b_n}\Big)K\Big(\frac{\|V-v\|}{b_n}\Big)\Big]}{\mathbb{E}\Big[K\Big(\frac{X-x_0}{b_n}\Big)\bar{K}\Big(\frac{\|V-v\|}{b_n}\Big)\Big]} \leq C < \infty$$

for all $v \in \bar{\mathcal{V}}$ and sufficiently large n.

(iv) The first-stage estimator satisfies $\|\hat{F}_n - F_0\|_\infty = O_{\mathbb{P}}(\delta_n)$ with $\delta_n = o(b_n/n^{1/s})$.

The first condition posits that the observations are from i.i.d. sampling and at least one element of Z has a continuous distribution. Also, it specifies the norm $\|\cdot\|$ to be some L^2 norm. The proof goes through with little changes if we strengthen the norm to be the supremum norm. The second assumption imposes sufficient continuity on the regression function and the conditional proxy distribution. The third condition is specific to settings with function-valued regressors. It controls the behavior of small ball probabilities (Ferraty and Vieu, 2006). In some sense, we can regard $\mathbb{E}[K([X - x_0]/b_n)K(\|V - v\|/b_n)]/b_n^d$ as the joint "density" of (X, V) as $b_n \to 0$. In the standard case where V belongs to the Euclidean space, the limit is indeed the joint Lebesgue density. However, in this setting, there is no "standard" measure on the space of distribution functions and the idea of density is ambiguous. The last condition requires that the first-stage estimator converges at a reasonably fast rate.

Theorem : Nonparametric Estimation. *Suppose Assumption Nonparametric holds. If* $b_n \to 0$ *and* $\log n/n^{1-1/s}b_n^d \to 0$*, then* $\hat{\mu}_n(x_0) \to_{\mathbb{P}} \mu(x_0)$.

Proof. Define

$$\hat{m}_{1n}(x,v) = \frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{b_n}\Big)K\Big(\frac{\|\hat{V}_i - v\|}{b_n}\Big)$$
$$\hat{m}_{2n}(x,v) = \frac{1}{n}\sum_{i=1}^{n} Y_i K\Big(\frac{X_i - x}{b_n}\Big)K\Big(\frac{\|\hat{V}_i - v\|}{b_n}\Big)$$
$$\bar{\hat{m}}_{1n}(x,v) = \mathbb{E}\Big[K\Big(\frac{X - x}{b_n}\Big)K\Big(\frac{\|F(\cdot|X,Z) - v\|}{b_n}\Big)\Big]_{F=\hat{F}_n}$$
$$\bar{\hat{m}}_{2n}(x,v) = \mathbb{E}\Big[YK\Big(\frac{X - x}{b_n}\Big)K\Big(\frac{\|F(\cdot|X,Z) - v\|}{b_n}\Big)\Big]_{F=\hat{F}_n}.$$

I have the expansion

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}_n(x,\hat{V}_i)T_i = \frac{1}{n}\sum_{i=1}^{n}\frac{\bar{\hat{m}}_{2n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)}T_i + \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{m}_{2n}(x,\hat{V}_i) - \bar{\hat{m}}_{2n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)}T_i$$
$$- \frac{1}{n}\sum_{i=1}^{n}\hat{m}_n(x,\hat{V}_i)\frac{\hat{m}_{1n}(x,\hat{V}_i) - \bar{\hat{m}}_{1n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)}T_i.$$

By Lemma 1 and

$$|\bar{\hat{m}}_{1n}(x,v) - \bar{m}_{1n}(x,v)| \le 2\mathbb{E}\Big[K\Big(\frac{X - x}{b_n}\Big)\bar{K}\Big(\frac{\|V - v\|}{b_n}\Big)\Big]\frac{\|\hat{F} - F_0\|_\infty}{b_n}$$

where $\bar{m}_{1n}(x,v) = \mathbb{E}[K([X-x]/b_n)K(\|V-v\|/b_n)]$, we have

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{m}_n(x,\hat{V}_i)|T_i = O_{\mathbb{P}}\bigg(\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{\hat{m}}_{2n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)}T_i\bigg).$$

Then, it suffices to show $\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{\hat{m}}_{2n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)}T_i \to_{\mathbb{P}} \mu(x)$. This follows from $\frac{1}{n}\sum_{i=1}^{n}[\frac{\bar{\hat{m}}_{2n}(x,\hat{V}_i)}{\bar{\hat{m}}_{1n}(x,\hat{V}_i)} - \frac{\bar{m}_{2n}(x,V_i)}{\bar{m}_{1n}(x,V_i)}]T_i = o_{\mathbb{P}}(1)$ and continuity of $\mathbb{E}[Y|X,V]$ where $\bar{m}_{2n}(x,v) = \mathbb{E}[YK([X-x]/b_n)K(\|V - v\|/b_n)]$.

Lemma 1. *Suppose Assumption Nonparametric holds, $b_n \to 0$, and $\log n/n^{1-1/s}b_n^d \to 0$. Let $g_{in}(x,z,F) = Y_iK([X_i - x_0]/b_n)K(\|F(\cdot|X_i,Z_i) - F(\cdot|x,z)\|/b_n)$. Then,*

$$\sup_{F\in\mathcal{F}_n,(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\bigg|\frac{1}{n}\sum_{i=1}^{n}\big(g_{in}(x,z,F) - \mathbb{E}[g_{in}(x,z,F)]\big)\bigg| = O_{\mathbb{P}}(\rho_n)$$

where $\rho_n = \sqrt{b_n^d \log n/n}\max\{1, \sqrt{\log n/n^{1-2/s}b_n^d}\}$.

Proof. The proof follows the standard strategy: truncation, approximation by a suitable finite set, and bounding of tail probabilities. See, for example, Cattaneo et al. (2013); Hansen (2008); Pollard (1995).

Let $\mathcal{F}_n$ be a subset of functions classes such that the first-stage estimator belongs to $\mathcal{F}_n$ with probability approaching one and $\|F - F_0\| \leq C\delta_n$ for all $F \in \mathcal{F}_n$ with sufficiently large C. For the truncation argument, I replace Y_i with $Y_{in} := Y_i \mathbb{1}\{|Y_i| \leq Cn^{1/s}\}$. Note

$$\begin{aligned}
&\mathbb{P}\Big[\bigcup_{F\in\mathcal{F}_n,(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} \Big\{\frac{1}{n}\sum_{i=1}^n (Y_i - Y_{in})K\Big(\frac{X_i - x_0}{b_n}\Big)K\Big(\frac{\|F(\cdot|X_i,Z_i) - F(\cdot|x,z)\|}{b_n}\Big) \neq 0\Big\}\Big] \\
&\leq \mathbb{P}\big[\max_{1\leq i\leq n} |Y_i| > Cn^{1/s}\big]
\end{aligned}$$

and the last probability can be made arbitrary small by choosing sufficiently large C. For the expectation term, letting $\bar{K}(u) = \sup_{|v|\leq\eta} |\partial K(u+v)/\partial u|$ for some small $\eta > 0$,

$$\begin{aligned}
&\mathbb{E}\Big[Y\mathbb{1}\{|Y| > Cn^{1/s}\}K\Big(\frac{X - x_0}{b_n}\Big)K\Big(\frac{\|F(\cdot|X,Z) - F(\cdot|x,z)\|}{b_n}\Big)\Big] \\
&\leq C^{1-s}n^{-1+1/s}\mathbb{E}\Big[\mathbb{E}[|Y|^s|X,Z]K\Big(\frac{X - x_0}{b_n}\Big)K\Big(\frac{\|F(\cdot|X,Z) - F(\cdot|x,z)\|}{b_n}\Big)\Big] \\
&\leq n^{-1+1/s}C\mathbb{E}\Big[K\Big(\frac{X - x_0}{b_n}\Big)\Big\{K\Big(\frac{\|V - F_0(\cdot|x,z)\|}{b_n}\Big) + \bar{K}\Big(\frac{\|V - F_0(\cdot|x,z)\|}{b_n}\Big)\Big\}\Big]
\end{aligned}$$

where I use $\|F - F_0\|_\infty = o(b_n)$. Thus,

$$\begin{aligned}
\mathbb{P}\Big[\sup_{F\in\mathcal{F}_n(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} \Big|\frac{1}{n}\sum_{i=1}^n (Y_i - Y_{in})K\Big(\frac{X_i - x_0}{b_n}\Big)K\Big(\frac{\|F(\cdot|X_i,Z_i) - F(\cdot|x,z)\|}{b_n}\Big) \\
- \mathbb{E}\Big[(Y_i - Y_{in})K\Big(\frac{X - x_0}{b_n}\Big)K\Big(\frac{\|F(\cdot|X,Z) - F(\cdot|x,z)\|}{b_n}\Big)\Big]\Big| > C\rho_n\Big] = o(1).
\end{aligned}$$

Then, let $g_{in}(x,z,F) = Y_{in}K([X_i - x_0]/b_n)K(\|F(\cdot|X_i,Z_i) - F(\cdot|x,z)\|/b_n)$ in the sequel.

Now I construct an approximation set. Since $\overline{\mathcal{X}\mathcal{Z}}$ is compact, we can come up with points $(x_1,z_1),\ldots,(x_L,z_L)$ whose ϵ-balls cover $\overline{\mathcal{X}\mathcal{Z}}$ with $L \leq C\epsilon^{-(d_x+d_z)}$. Using

$$\begin{aligned}
&\Big|K\Big(\frac{\|F(\cdot|X_i,Z_i) - F(\cdot|\tilde{x},\tilde{z})\|}{b_n}\Big) - K\Big(\frac{\|V_i - F_0(\cdot|x,z)\|}{b_n}\Big)\Big| \\
&\leq \bar{K}\Big(\frac{\|V_i - F_0(\cdot|x,z)\|}{b_n}\Big)(2\|F - F_0\|_\infty + \|F_0(\cdot|\tilde{x},\tilde{z}) - F_0(\cdot|x,z)\|)b_n^{-1},
\end{aligned}$$

we have, for any $(x,z) \in \overline{\mathcal{X}\mathcal{Z}}$ by choosing appropriate (x_l, z_l),

$$|g_{in}(x,z,F) - g_{in}(x_l,z_l,F_0)| \leq Cn^{1/s}\frac{|x-x_l|^\alpha + |z-z_l|^\alpha + \delta_n}{b_n}\bar{K}\Big(\frac{\|V_i - v_l\|}{b_n}\Big)$$

where $v_l = F_0(\cdot|x_l, z_l)$. By letting $\epsilon = (\eta b_n / n^{1/s})^{1/\alpha}$,

$$\frac{1}{n}\sum_{i=1}^n \big|g_{in}(x,z,F) - g_{in}(x_l,z_l,F_0)\big| \leq \eta C$$

for any realization of the data. Note that $L = \eta^{-(d_x+d_z)/\alpha}O(n^a)$ for some positive a.

We have $\mathbb{E}[g_{in}^2(x,z,F)] \leq Cb_n^d$ for $F \in \mathcal{F}_n$, $(x,z) \in \overline{\mathcal{X}\mathcal{Z}}$ and by the hypothesis $nb_n^d \to \infty$. Then, arguing as in Section II.3 of Pollard (1984), for sufficiently large n,

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{F\in\mathcal{F}_n,(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\Big|\frac{1}{n}\sum_{i=1}^n g_{in}(x,z,F) - \mathbb{E}[g_{in}(x,z,F)]\Big| > C\rho_n\Big] \\
&\leq 4\mathbb{P}\Big[\sup_{F\in\mathcal{F}_n,(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i g_{in}(x,z,F)\Big| > C\rho_n/4\Big]
\end{aligned}$$

where $\mathbb{P}[\sigma_i = 1] = 1/2 = \mathbb{P}[\sigma_i = -1]$ and $\{\sigma_i\}_{i=1}^n$ is an i.i.d. sequence independent of the data. Now consider the above probability conditional on the data, and approximate the supremum by the finite set constructed above, which yields the bound

$$\begin{aligned}
&Cn^a \max_{1\leq l\leq L}\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i g_{in}(x_l,z_l,F_0)\Big| > C\rho_n/8\Big|(X_i,Z_i)_{i=1}^n\Big] \\
&\leq Cn^a \max_{1\leq l\leq L}\exp\Big(\frac{-Cn\rho_n^2}{\frac{1}{n}\sum_{i=1}^n g_{in}^2(x_l,z_l,F_0) + n^{1/s}\rho_n}\Big)
\end{aligned}$$

where the inequality follows from Bernstein's inequality. The theorem of Pollard (1995) implies $\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\frac{1}{n}\sum_{i=1}^n g_{in}^2(x,z,F_0) = O_{\mathbb{P}}(b_n^d)$ and arguing as in Cattaneo et al. (2013), the desired result follows. □

Remark 1. *The condition $\delta_n = o(b_n/n^{1/s})$ is a stronger condition than usually used in the literature. This feature seems to come from not imposing more smoothness on the first-stage estimator $\hat{F}_n$. The approach taken in Mammen et al. (2016) imposes that with probability approaching one, the first-stage estimator belongs to a function class with sufficient smoothness. This allows for con-*

struction of the approximation set utilizing a good bound on the covering number of such function class. This relaxes the condition $\delta_n = o(b_n/n^{1/s})$ *but inflates the uniform covering number from the order of* ϵ^{-a} *to* $\exp(\epsilon^{-b})$ *for some* $b \in (0,2)$.

□

A.2.2 Local polynomial regression

Here I review the setup of local polynomial regression, loosely following the exposition in Masry (1996). Given a vector of non-negative integers $\pi = (\pi_1 \dots \pi_d) \in \mathbb{N}^d$ and another vector $x = (x_1 \dots x_d) \in \mathbb{R}^d$, I write $x^\pi = (x_1^{\pi_1} \dots x_d^{\pi_d})$, $|\pi| = \pi_1 + \cdots + \pi_d$, and $\pi! = \pi_1! \dots \pi_d!$. Also, given a differentiable function f, I write $\partial^\pi f(x) = \partial^{|\pi|} f(x)/\partial^{\pi_1} x_1 \dots \partial^{\pi_d} x_d$. Denote the number of unique d-tuples with $|\pi| = k$ by $Q_k = \binom{d+k-1}{k-1}$ where $d = d_x + d_v$. The qth-order local polynomial estimator for $\mathbb{E}[Y|X = x, V = v]$ is characterized by minimization of

$$\sum_{i=1}^{n} \Big(Y_i - \sum_{|\pi|=0}^{q} \beta_\pi \begin{bmatrix} X_i - x \\ V_i - v \end{bmatrix}^\pi \Big)^2 K_n(X_i - x_0) K_n(V_i - v)$$

with respect to β's where $\sum_{|\pi|=0}^{q}$ is summation over all d-tuples π with $|\pi|$ ranging over $\{0, 1, \dots, q\}$. Here, the dimensions of X and V can be different, but with some abuse of notation, I denote the kernel functions by the same letter K. Denote the solution to the above minimization by $\hat{\beta}_n$ and arrange its elements using the following lexicographic order on the index π: 1) the smaller $|\pi|$ gets a priority and 2) among the tuples with the same $|\pi|$, the tuple having a larger integer in elements close to the dth position gets a priority. Let $Q = \sum_{k=0}^{p} Q_k$ and $\pi(i)$ be the mapping that takes integers $\{1, \dots, Q\}$ to the ith d-tuple according to the above lexicographic order. Then, $\hat{\beta}_n$ can be written as $(\hat{\beta}_{n\pi(1)}, \hat{\beta}_{n\pi(2)}, \dots, \hat{\beta}_{n\pi(Q)})'$ and the estimator $\hat{m}_n(x, v)$ is $\mathbf{e}_1' \hat{\beta}_n$ with $\mathbf{e}_1 = (1\ 0 \dots 0)'$.

The first-order condition of the above least squares problem is the collection of

$$\sum_{i=1}^{n} Y_i \begin{bmatrix} X_i - x \\ V_i - v \end{bmatrix}^s K_n(X_i - x_0) K_n(V_i - v) = \sum_{i=1}^{n} \sum_{|\pi|=0}^{q} \beta_\pi \begin{bmatrix} X_i - x \\ V_i - v \end{bmatrix}^{\pi+s} K_n(X_i - x_0) K_n(V_i - v)$$

where the d-tuple s ranges over $|s| \in \{0, 1, \dots, q\}$. Then, we can express $\hat{\beta}_n$ by

$$\hat{\beta}_n = \operatorname{diag}\left(b_n^{|\pi(1)|}\ b_n^{|\pi(2)|} \dots b_n^{|\pi(Q)|} \right)^{-1} [\mathbf{X}_n(x,v)' \mathbf{W}_n(x,v) \mathbf{X}_n(x,v)]^{-1} \mathbf{X}_n(x,v)' \mathbf{W}_n(x,v) \mathbf{y}_n$$

where the ith row of $\mathbf{X}_n(x,v)$ is $\left(\begin{bmatrix}(X_i-x)/b_n\\(V_i-v)/b_n\end{bmatrix}^{\pi(1)} \dots \begin{bmatrix}(X_i-x)/b_n\\(V_i-v)/b_n\end{bmatrix}^{\pi(Q)}\right)$, $\mathbf{W}_n(x,v) = \text{diag}(K_n(X_1-x)K_n(V_1-v)\dots K_n(X_n-x)K_n(V_n-v))$, and $\mathbf{y}_n = [Y_1 \dots Y_n]'$. Writing $\mathbf{S}_n(x,v)$ for $\mathbf{X}_n(x,v)'\mathbf{W}_n(x,v)\mathbf{X}_n(x,v)/n$, we have

$$\tilde{m}_n(x,v) = \mathbf{e}_1'\mathbf{S}_n^{-1}(x,v)\mathbf{X}_n(x,v)'\mathbf{W}_n(x,v)\mathbf{y}_n/n.$$

For the estimator with generated covariates, define analogous objects by replacing V_i with $\hat{V}_i$ and write $\hat{X}_n(x,v)$, $\hat{\mathbf{W}}_n(x,v)$, and $\hat{\mathbf{S}}_n(x,v)$. Then,

$$\hat{m}_n(x,v) = \mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(x,v)\hat{\mathbf{X}}_n(x,v)'\hat{\mathbf{W}}_n(x,v)\mathbf{y}_n/n.$$

For reference, let $\mathbf{S}_0$ be the $Q \times Q$ matrix

$$\mathbf{S}_0 = \begin{bmatrix} \int u^{\pi(1)+\pi(1)}K(u_1)K(u_2)du & \dots & \int u^{\pi(1)+\pi(Q)}K(u_1)K(u_2)du \\ \vdots & \ddots & \\ \int u^{\pi(Q)+\pi(1)}K(u_1)K(u_2)du & & \int u^{\pi(Q)+\pi(Q)}K(u_1)K(u_2)du \end{bmatrix},$$

where $u = (u_1'u_2')'$,

$$\boldsymbol{\kappa}(u_1,u_2) = \begin{bmatrix} u^{\pi(1)} \\ \vdots \\ u^{\pi(Q)} \end{bmatrix} K(u_1)K(u_2), \qquad \boldsymbol{\kappa}_n(u_1,u_2) = b^{-d}\boldsymbol{\kappa}(u_1/b_n, u_2/b_n).$$

A.2.3 Proof of Theorem 2

By $\mathbb{P}[(X,Z) \in \overline{\mathcal{XZ}}] > 0$, it suffices to consider the numerator of the estimator. Let $T_i = \mathbb{1}\{(X_i,Z_i) \in \overline{\mathcal{XZ}}\}$ and we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{m}_n(x_0,\hat{V}_i)T_i - \mu(x_0)\tau_0 = &\frac{1}{n}\sum_{i=1}^n m_0(x_0,V_i)T_i - \mu(x_0)\tau_0 \\ &+ \frac{1}{n}\sum_{i=1}^n [\tilde{m}_n(x_0,V_i) - m_0(x_0,V_i)]T_i \\ &+ \frac{1}{n}\sum_{i=1}^n [\hat{m}_n(x_0,\hat{V}_i) - \tilde{m}_n(x_0,V_i)]T_i.\end{aligned}$$

The first sum after the equality is $O_{\mathbb{P}}(1/\sqrt{n})$. In the sequel, I show that the second sum converges in distribution to $N(0, \sigma_0^2\tau_0^2)$ and the third sum is $o_{\mathbb{P}}(1/\sqrt{nb_n^{d_x}})$, from which the desired result follows.

Asymptotic normality I show that the term

$$\sqrt{nb_n^{d_x}}\frac{1}{n}\sum_{i=1}^{n}[\tilde{m}_n(x_0, \hat{V}_i) - m_0(x_0, V_i)]T_i$$

converges in distribution to the normal distribution with mean zero and variance $\sigma_0^2\tau_0^2$. By the standard argument,

$$\tilde{m}_n(x, v) - m_0(x, v) = \mathbf{e}_1'\mathbf{S}_n^{-1}(x, v)\mathbf{X}_n(x, v)'\mathbf{W}_n(x, v)\boldsymbol{\epsilon}/n + O_{\mathbb{P}}(b_n^{q+1})$$

where $\boldsymbol{\epsilon} = (\epsilon_1 \dots \epsilon_n)'$ and I use bondedeness of $(q+1)$th-order derivatives of m_0. Lemma B-1 of Cattaneo et al. (2013) implies

$$\sup_{v\in\mathcal{V}_0}|\mathbf{S}_n(x_0, v) - \bar{\mathbf{S}}_n(x_0, v)]| = O_{\mathbb{P}}\Big(\sqrt{\log n/nb_n^d}\Big) \tag{2}$$

with $\bar{\mathbf{S}}_n(x, v) = \mathbb{E}[\mathbf{S}_n(x_0, v)]$, $d = d_x + d_v$, and

$$\sup_{v\in\mathcal{V}_0}|\mathbf{X}_n(x_0, v)'\mathbf{W}_n(x_0, v)\boldsymbol{\epsilon}/n| = O_{\mathbb{P}}\Big(\sqrt{\log n/nb_n^d}\max\Big\{1, \sqrt{\log n/n^{1-2/s}b_n^d}\Big\}\Big). \tag{3}$$

Then, it suffices to look at $\sum_i T_i\mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(x_0, V_i)\mathbf{X}_n(x_0, V_i)'\mathbf{W}_n(x_0, V_i)\boldsymbol{\epsilon}/n^2$. Using the V-statistic structure and Hoeffding decomposition,

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{e}_1'\bar{\mathbf{S}}_0^{-1}(x_0, V_i)\mathbf{X}_n(x_0, V_i)'\mathbf{W}_n(x_0, V_i)\boldsymbol{\epsilon}T_i \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\mathbf{e}_1'\bar{\mathbf{S}}_n(x_0, V_i)\boldsymbol{\kappa}_n(X_j - x_0, V_j - V_i)\tau(V_i)|X_j, V_j]\epsilon_j + O_{\mathbb{P}}\big((nb_n^d)^{-1}\big).
\end{aligned}$$

Note that $\bar{\mathbf{S}}_n(x_0,v) = \mathbf{S}_0 f_{XV}(x_0,v) + O(b_n)$ uniformly over $v \in \mathcal{V}_0^{\delta/2}$. Then,

$$\begin{aligned}\lim_{n\to\infty} b_n^{d_x}\mathbb{E}\big[\mathbb{E}[\mathbf{e}_1'\bar{\mathbf{S}}_n(x_0,V_i)\boldsymbol{\kappa}_n(X_j-x_0,V_j-V_i)\tau(V_i)|X_j,V_j]^2\epsilon_j^2\big] \\ = \mathbf{e}_1'\mathbf{S}_0^{-1}\mathbf{M}\mathbf{S}_0^{-1}\mathbf{e}_1\int\frac{f_V^2(v)\sigma^2(x_0,v)\tau^2(v)}{f_{XV}(x_0,v)}dv\end{aligned}$$

where $\mathbf{M}$ is the $Q\times Q$ matrix whose (i,j)th element is $\int \left[\begin{smallmatrix}u_1\\u_2\end{smallmatrix}\right]^{\pi(i)}\left[\begin{smallmatrix}u_1\\\tilde{u}_2\end{smallmatrix}\right]^{\pi(j)}K^2(u_1)K(u_2)K(\tilde{u}_2)du_1du_2d\tilde{u}_2$.

First-stage estimation error We have the decomposition

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n[\hat{m}_n(x_0,\hat{V}_i)-\tilde{m}_n(x_0,V_i)]T_i = &\frac{1}{n}\sum_{i=1}^n[\hat{m}_n(x_0,\hat{V}_i)-m_0(x_0,\hat{V}_i)]T_i \\ &-\frac{1}{n}\sum_{i=1}^n[\tilde{m}_n(x_0,\hat{V}_i)-m_0(x_0,V_i)]T_i \\ &+\frac{1}{n}\sum_{i=1}^n[m_0(x_0,\hat{V}_i)-m_0(x_0,V_i)]T_i\end{aligned}$$

where the last sum is $O_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(1/\sqrt{nb_n^{d_x}})$. Let $\boldsymbol{m}_n = (m_0(X_1,V_1),\dots m_0(X_n,V_n))'$, $\boldsymbol{m}_n(\hat{\cdot}) = (m_0(X_1,\hat{V}_1),\dots m_0(X_n,\hat{V}_n))'$, and $\mathbf{1} = (1\dots 1)' \in \mathbb{R}^n$. Note that

$$\hat{m}_n(x,v)-m_0(x,v) = \mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(x,v)\hat{\mathbf{X}}_n(x,v)'\hat{\mathbf{W}}_n(x,v)[\boldsymbol{\epsilon}+\boldsymbol{m}_n(\hat{\cdot})-\mathbf{1}m_0(x,v)+\boldsymbol{m}_n-\boldsymbol{m}_n(\hat{\cdot})]/n.$$

and

$$\begin{aligned}&\max_{1\le i\le n}|\hat{\mathbf{S}}_n(x_0,\hat{V}_i)-\mathbf{S}_n(x_0,V_i)| = O_{\mathbb{P}}(n^{-1/2}b_n^{-1}) \\ &\max_{1\le i\le n}|[\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)-\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)]\boldsymbol{\epsilon}|/n = O_{\mathbb{P}}(n^{-1/2}b_n^{-1}) \qquad (4)\end{aligned}$$

which follow from continuous differentiability and compact support of K and $\max_i|\hat{V}_i-V_i| = O_{\mathbb{P}}(n^{-1/2})$. The standard smoothing argument implies that

$$\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(x_0,v)\hat{\mathbf{X}}_n(x_0,v)'\hat{\mathbf{W}}_n(x_0,v)[\boldsymbol{m}_n(\hat{\cdot})-\mathbf{1}m_0(x_0,v)]/n = O_{\mathbb{P}}(b_n^{p+1})$$

uniformly on $\mathcal{V}_0^\delta$. A similar bound applies to the bias component of $\tilde{m}_n(x_0,v)-m_0(x_0,v)$. Also,

$$\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(x_0,\hat{V}_i)\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)[\boldsymbol{m}_n-\boldsymbol{m}_n(\hat{\cdot})]T_i = O_{\mathbb{P}}(n^{-1/2})$$

follows from continuous differentiability of m_0 and $\max_i|\hat{V}_i-V_i| = O_{\mathbb{P}}(n^{-1/2})$. Thus it remains to analyze

$$\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\big(\hat{\mathbf{S}}_n^{-1}(x_0,\hat{V}_i)\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)-\mathbf{S}_n^{-1}(x_0,V_i)\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)\big)\boldsymbol{\epsilon}T_i$$

This term equals

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\big(\hat{\mathbf{S}}_n^{-1}(x_0,\hat{V}_i)-\mathbf{S}_n^{-1}(x_0,V_i)\big)\big(\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)-\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)\big)\boldsymbol{\epsilon}T_i\\
&+\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\big(\hat{\mathbf{S}}_n^{-1}(x_0,\hat{V}_i)-\mathbf{S}_n^{-1}(x_0,V_i)\big)\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)\boldsymbol{\epsilon}T_i\\
&+\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\big(\mathbf{S}_n^{-1}(x_0,V_i)-\bar{\mathbf{S}}_n^{-1}(x_0,V_i)\big)\big(\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)-\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)\big)\boldsymbol{\epsilon}T_i\\
&+\frac{1}{n^2}\sum_{i=1}^n \mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(x_0,V_i)\big(\hat{\mathbf{X}}_n(x_0,\hat{V}_i)'\hat{\mathbf{W}}_n(x_0,\hat{V}_i)-\mathbf{X}_n(x_0,V_i)'\mathbf{W}_n(x_0,V_i)\big)\boldsymbol{\epsilon}T_i.
\end{aligned}$$

Using (2), (3), and (4), the first three sums are $o_{\mathbb{P}}(1/\sqrt{nb_n^{d_x}})$. The last sum equals

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^n\sum_{j\neq i} \mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(x_0,V_i)\big(\boldsymbol{\kappa}_n(X_j-x_0,\hat{V}_j-\hat{V}_i)-\boldsymbol{\kappa}_n(X_j-x_0,V_j-V_i)\big)\epsilon_j T_i\\
&=\mathbf{e}_1'\mathbb{E}[\bar{\mathbf{S}}_n^{-1}(x_0,V_i)\big(\boldsymbol{\kappa}_n(X_j-x_0,\theta(X_j,Z_j,\beta)-\theta(X_i,Z_i,\beta))-\boldsymbol{\kappa}_n(X_j-x_0,V_j-V_i)\big)\epsilon_j T_i]_{\beta=\hat{\beta}_n}\\
&\qquad\times(1+n^{-1})+O_{\mathbb{P}}\big(n^{-1}b_n^{-(d_x/2+1)}+n^{-3/2}b_n^{-(d+1)}\big)
\end{aligned}$$

where I use Hoeffding decomposition and the maximal inequality of Chen and Kato (2020). To justify this claim, note continuous differentiability of K and θ can be used to show the class of functions

$$\Big\{K\Big(\tfrac{\cdot-x_0}{b_n}\Big)b_n^{-|\pi|}\Big(\Big[\substack{\cdot-x_0\\ \theta(\cdot,\beta)-\theta(\cdot,\beta)}\Big]^\pi K\Big(\tfrac{\theta(\cdot,\beta)-\theta(\cdot,\beta)}{b_n}\Big)-\Big[\substack{\cdot-x_0\\ \theta(\cdot,\beta_0)-\theta(\cdot,\beta_0)}\Big]^\pi K\Big(\tfrac{\theta(\cdot,\beta_0)-\theta(\cdot,\beta_0)}{b_n}\Big)\Big):\tfrac{|\beta-\beta_0|}{\sqrt{n}}\leq C\Big\}$$

is Euclidean for the natural envelope as defined in Pakes and Pollard (1989) where $\pi \in \{0, \ldots, p\}^d$. Note that the envelope function is bounded by

$$C\Big|K\Big(\frac{X_j - x_0}{b_n}\Big)\Big|\bar{K}\Big(\frac{V_j - V_i}{b_n}\Big)n^{-1/2}b_n^{-1}$$

where $\bar{K}(u) = |K(u)| + \sup_{|v|\leq\delta}|K^{(1)}(u+v)|$ for some small δ. Then, using the bound in Equation (21) of Chen and Kato, the desired result follows. For the expectation term, by Taylor expansion,

$$\begin{aligned}
&\mathbb{E}[\bar{\mathbf{S}}_n^{-1}(x_0, V_i)\big(\boldsymbol{\kappa}_n(X_j - x_0, \theta(X_j, Z_j, \beta) - \theta(X_i, Z_i, \beta)) - \boldsymbol{\kappa}_n(X_j - x_0, V_j - V_i)\big)\epsilon_j T_i] \\
&= \mathbb{E}[\bar{\mathbf{S}}_n^{-1}(x_0, V_i)\boldsymbol{\kappa}_n^{(1)}(X_j - x_0, V_j - V_i)\epsilon_j\theta^{(1)}(X_j, Z_j, \beta_0)T_i]\frac{\beta - \beta_0}{b_n} + O\Big(\frac{|\beta - \beta_0|^2}{b_n^2}\Big) \\
&= \mathbf{S}_0^{-1}\mathbb{E}[f_{XV}^{-1}(x_0, V_i)\boldsymbol{\kappa}_n^{(1)}(X_j - x_0, V_j - V_i)\epsilon_j\theta^{(1)}(X_j, Z_j, \beta_0)T_i]\frac{\beta - \beta_0}{b_n} + O\Big(|\beta - \beta_0| + \frac{|\beta - \beta_0|^2}{b_n^2}\Big)
\end{aligned}$$

where $\boldsymbol{\kappa}_n^{(1)}(\cdot) = b_n^{-d}\boldsymbol{\kappa}^{(1)}(\cdot/b_n)$, $\boldsymbol{\kappa}^{(1)}$ is the derivative of $\boldsymbol{\kappa}$ with respect to the argument v, and $\theta^{(1)}$ is the derivative of θ with respect to β. The first equality uses $\mathbb{E}[\epsilon_j|X_j, V_j] = 0$ and the random sampling assumption, and the second equality uses $\bar{\mathbf{S}}_n(x_0, v) = \mathbf{S}_0 f_{XV}(x_0, v) + O(b_n)$ uniformly over $v \in \overline{\mathcal{V}}$.

By the change of variables,

$$\begin{aligned}
&\mathbb{E}[f_{XV}^{-1}(x_0, V_i)\boldsymbol{\kappa}_n^{(1)}(X_j - x_0, V_j - V_i)\epsilon_j\theta^{(1)}(X_j, Z_j, \beta_0)T_i] \\
&= \int \mathbb{E}[\epsilon|X = x_0 + u_1 b_n, Z = z]\Big(\frac{\tau(v)f_V(v)}{f_{XV}(x_0, v)}\Big)_{v=\theta(x_0+u_1b_n, z, \beta_0)-u_2b_n} \\
&\qquad \times \boldsymbol{\kappa}^{(1)}(u_1, u_2)f_{XZ}(x_0 + u_1 b_n, z)du_1 dz du_2
\end{aligned}$$

Using continuity and boundedness of the integrand, the above expectation equals

$$\int \frac{\tau(v)f_V(v)}{f_{XV}(x_0, v)}\Big|_{v=\theta(x_0,z,\beta_0)}\mathbb{E}[\epsilon|X = x_0, Z = z]f_{XZ}(x_0, z)dz \int \boldsymbol{\kappa}^{(1)}(u_1, u_2)du_1 du_2 + O(b_n).$$

By evenness of K, $\mathbf{S}_0^{-1}\int \boldsymbol{\kappa}^{(1)}(u_1, u_2)du_1 du_2 = 0$, and the desired result follows.

A.2.4 Proof of Theorem 4

In this setting, the kernel only takes V as its argument. Thus, the ith row of the matrix $\mathbf{X}_n(v)$ is $[(V_i - v/b_n)^{\pi(1)} \dots (V_i - v/b_n)^{\pi(Q)}]$, and $\mathbf{W}_n(v) = \text{diag}(\mathbb{1}\{X_1 = x_0\}K_n(V_1 - v) \dots \mathbb{1}\{X_n = x_0\}K_n(V_n - v))$. Also, $\boldsymbol{\kappa}(u) = [u^{\pi(1)} \dots u^{\pi(Q)}]'K(u)$ for $u \in \mathbb{R}^{d_v}$. I use the same notation as above but objects involving the kernel function reflect this change. Let $D_i = \mathbb{1}\{X_i = x_0\}$. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{m}_n(x_0, \hat{V}_i)T_i - \mu(x_0)\tau_0 = \frac{1}{n}\sum_{i=1}^{n} \big(m_0(x_0, V_i)T_i - \mu(x_0)\tau_0\big) \\ + \frac{1}{n}\sum_{i=1}^{n} \big[\tilde{m}_n(x_0, V_i) - m_0(x_0, V_i)\big]T_i \\ + \frac{1}{n}\sum_{i=1}^{n} \big[\hat{m}_n(x_0, \hat{V}_i) - \tilde{m}_n(x_0, V_i)\big]T_i.\end{aligned}$$

The first sum after the equality constitutes part of the ψ term defined in Theorem 3. For the second and third terms, I show

$$\frac{1}{n}\sum_{i=1}^{n} \big[\tilde{m}_n(x_0, V_i) - m_0(x_0, V_i)\big]T_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\tau(V_i)f_V(V_i)}{f_{V|X}(V_i|x_0)\rho_0} D_i\epsilon_i + o_{\mathbb{P}}(1/\sqrt{n}) \tag{5}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \big[\hat{m}_n(x_0, \hat{V}_i) - \tilde{m}_n(x_0, V_i)\big]T_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Gamma}\varphi(W_i, X_i, Z_i) + o_{\mathbb{P}}(1/\sqrt{n}). \tag{6}$$

where $\boldsymbol{\Gamma}$ is defined in the statement of Theorem 3. From this representation, the desired result follows.

For (5), as in the continuous X case,

$$\frac{1}{n}\sum_{i=1}^{n} \big[\tilde{m}_n(x_0, V_i) - m_0(x_0, V_i)\big]T_i = \frac{1}{n}\sum_{j=1}^{n} \mathbf{e}_1'\mathbb{E}[\bar{\mathbf{S}}_n^{-1}(V_i)\boldsymbol{\kappa}_n(V_j - V_i)T_i|V_j]D_j\epsilon_j + o_{\mathbb{P}}(1/\sqrt{n})$$

and we have $\bar{\mathbf{S}}_n(v) \to \mathbf{S}_0 f_{V|X}(v|x_0)\rho_0$ uniformly over $\bar{\mathcal{V}}^{\delta/2}$. Then, (5) follows from the change-of-variables and $\mathbf{e}_1'\mathbf{S}_0^{-1}\int \boldsymbol{\kappa}(u)du = 1$.

First-stage estimation error We have the decomposition

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{m}_n(x_0,\hat{V}_i)-\tilde{m}_n(x_0,V_i)]T_i = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_n(x_0,\hat{V}_i)-m_0(x_0,\hat{V}_i)-\tilde{m}_n(x_0,\hat{V}_i)+m_0(x_0,V_i)]T_i + \frac{1}{n}\sum_{i=1}^{n}[m_0(x_0,\hat{V}_i)-m_0(x_0,V_i)]T_i.$$

I show that the first sum equals

$$\frac{1}{n}\sum_{i=1}^{n} -\Big\{\mathbf{e}_1'\mathbf{S}_0^{-1}\int\frac{\partial\boldsymbol{\kappa}(u)}{\partial u'}u'du\mathbb{E}\Big[\frac{\partial}{\partial v}\Big[\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\Big]\frac{\partial\theta(x_0,Z,\beta_0)}{\partial\beta'}\epsilon\Big|X=x_0\Big] + \mathbb{E}\Big[\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\frac{\partial m(x_0,V)}{\partial v'}\frac{\partial\theta(x_0,Z,\beta_0)}{\partial\beta'}\Big|X=x_0\Big]\Big\}\varphi(W_i,X_i,Z_i)+o_{\mathbb{P}}(1/\sqrt{n}) \tag{7}$$

and the second term equals

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\frac{\partial m_0(x_0,V)}{\partial v'}\frac{\partial\theta(X,Z,\beta_0)}{\partial\beta'}T\Big]\varphi(W_i,X_i,Z_i)+o_{\mathbb{P}}(1/\sqrt{n}),$$

from which (6) follows. For the second sum, using continuous differentiability of m_0 and θ,

$$\frac{1}{n}\sum_{i=1}^{n}[m_0(x_0,\hat{V}_i)-m_0(x_0,V_i)]T_i=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m_0(x_0,V_i)}{\partial v'}\frac{\partial\theta(X_i,Z_i,\beta_0)}{\partial\beta'}(\hat{\beta}_n-\beta_0)T_i+O_{\mathbb{P}}(|\hat{\beta}_n-\beta_0|^2)$$

and the desired result follows. It remains to show (7). Arguing as in the continuous case,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[\hat{m}_n(x_0,\hat{V}_i)-m_0(x_0,\hat{V}_i)-\tilde{m}_n(x_0,\hat{V}_i)+m_0(x_0,V_i)]T_i\\ &=\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{e}_1'[\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\hat{\mathbf{X}}_n(\hat{V}_i)'\hat{\mathbf{W}}_n(\hat{V}_i)-\mathbf{S}_n^{-1}(V_i)\mathbf{X}_n(V_i)'\mathbf{W}_n(V_i)]\boldsymbol{\epsilon}_nT_i \qquad (8)\\ &\quad-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\boldsymbol{\kappa}_n(\hat{V}_j-\hat{V}_i)D_j[m_0(x_0,\hat{V}_j)-m_0(x_0,V_j)]T_i+O_{\mathbb{P}}(b_n^{q+1}). \qquad (9)\end{aligned}$$

For (9), it equals

$$-\frac{1}{n^2}\sum_{i\neq j}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\boldsymbol{\kappa}_n(\hat{V}_j-\hat{V}_i)D_j\frac{\partial m(x_0,V_j)}{\partial v'}\frac{\partial\theta(x_0,Z_j,\beta_0)}{\partial\beta'}(\hat{\beta}_n-\beta_0)+O_{\mathbb{P}}\big(|\hat{\beta}_n-\beta_0|^2+1/nb_n^d\big)$$

and

$$\frac{1}{n^2}\sum_{i\neq j}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\boldsymbol{\kappa}_n(\hat{V}_j-\hat{V}_i)D_j\frac{\partial m(x_0,V_j)}{\partial v'}\frac{\partial\theta(x_0,Z_j,\beta_0)}{\partial\beta'}T_i$$
$$\to_{\mathbb{P}}\mathbf{e}_1'\mathbf{S}_0^{-1}\int\boldsymbol{\kappa}(u)du\int\frac{\tau(v)f_V(v)}{f_{V|X}(v|x_0)}\frac{\partial m(x_0,v)}{\partial v}\Big|_{v=\theta(x_0,z,\beta_0)}\frac{\partial\theta(x_0,z,\beta_0)}{\partial\beta'}f_{Z|X}(z|x_0)dz.$$

Then, we have

$$(9)=-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\frac{\partial m(x_0,V)}{\partial v'}\frac{\partial\theta(x_0,Z,\beta_0)}{\partial\beta'}\Big|X=x_0\Big]\varphi(W_i,X_i,Z_i)+o_{\mathbb{P}}(1/\sqrt{n}).$$

For (8), again arguing as in the continuous case, its leading term is

$$\frac{1}{n^2}\sum_{i\neq j}\bar{\mathbf{S}}_n^{-1}(V_i)[\boldsymbol{\kappa}_n(\hat{V}_j-\hat{V}_i)-\boldsymbol{\kappa}_n(V_j-V_i)]D_j\epsilon_jT_i$$

and using Hoeffding decomposition and the maximal inequality for U-processes of Chen and Kato (2020), it suffices to analyze

$$\mathbb{E}\big[\bar{\mathbf{S}}_n^{-1}(V_i)[\boldsymbol{\kappa}_n(\theta(x_0,Z_j,\beta)-\theta(X_i,Z_i,\beta))-\boldsymbol{\kappa}_n(V_j-V_i)]D_j\epsilon_jT_i\big]_{\beta=\hat{\beta}_n}.$$

By Taylor expansion,

$$\begin{aligned}&\mathbb{E}\big[\bar{\mathbf{S}}_n^{-1}(V_i)[\boldsymbol{\kappa}_n(\theta(x_0,Z_j,\beta)-\theta(X_i,Z_i,\beta))-\boldsymbol{\kappa}_n(V_j-V_i)]D_j\epsilon_jT_i\big]_{\beta=\hat{\beta}_n}\\&=b_n^{-1}\mathbb{E}\Big[\bar{\mathbf{S}}_n^{-1}(V_i)\boldsymbol{\kappa}_n^{(1)}(V_j-V_i)\frac{\partial\theta(X_j,Z_j,\beta_0)}{\partial\beta'}D_j\epsilon_jT_i\Big](\hat{\beta}_n-\beta_0)+O_{\mathbb{P}}(1/nb_n^2)\end{aligned}$$

where $\boldsymbol{\kappa}_n^{(1)}(u)=\boldsymbol{\kappa}^{(1)}(u/b_n)/b_n^{d_v}$ and $\boldsymbol{\kappa}^{(1)}(u)=\partial\boldsymbol{\kappa}(u)/\partial u'$. Recalling the notation from A.2.2, define $\mathbf{t}(u)=[u^{\pi(1)}\dots u^{\pi(Q)}]'$. Then,

$$\bar{\mathbf{S}}_n(v)=\int\mathbf{t}(u)\mathbf{t}(u)'K(u)duf_{V|X}(v|x_0)\rho_0+b_n\int\mathbf{t}(u)\mathbf{t}(u)'K(u)u'du\frac{\partial f_{V|X}(v|x_0)}{\partial v}\rho_0+O(b_n^2)$$

where the remainder term uses boundedness of the second derivative of $f_{V|X}(\cdot|x_0)$. Then,

$$\bar{\mathbf{S}}_n^{-1}(v)=\frac{\mathbf{S}_0^{-1}}{f_{V|X}(v|x_0)\rho_0}-\frac{b_n}{f_{V|X}^2(v|x_0)\rho_0}\mathbf{S}_0^{-1}\int\mathbf{t}(u)\mathbf{t}(u)'K(u)u'du\frac{\partial f_{V|X}(v|x_0)}{\partial v}\mathbf{S}_0^{-1}+O(b_n^2).$$

Letting $\alpha(v) = \tau(v)[f_{V|X}(v|x_0)\rho_0]^{-1}$ or $\alpha(v) = \tau(v)\tilde{u}'\partial f_{V|X}(v|x_0)/\partial v[f_{V|X}(v|x_0)^2\rho_0]^{-1}$ for some fixed $\tilde{u}$, we consider

$$
\begin{aligned}
&\mathbb{E}[\alpha(V_i)\boldsymbol{\kappa}_n^{(1)}(V_j - V_i)\frac{\partial\theta(X_j, Z_j, \beta_0)}{\partial\beta'}D_j\epsilon_j] \\
&= \int\int \boldsymbol{\kappa}^{(1)}(u)\big[\alpha(v - ub_n)f_V(v - ub_n)\big]_{v=\theta(x_0,z,\beta_0)}du\frac{\partial\theta(X_j, Z_j, \beta_0)}{\partial\beta'}\mathbb{E}[\epsilon|X = x_0, Z = z]f_{Z|X}(z|x_0)dz\rho_0 \\
&= \int \boldsymbol{\kappa}^{(1)}(u)du\int \big[\alpha(v)f_V(v)\big]_{v=\theta(x_0,z,\beta_0)}\frac{\partial\theta(X_j, Z_j, \beta_0)}{\partial\beta'}\mathbb{E}[\epsilon|X = x_0, Z = z]f_{Z|X}(z|x_0)dz\rho_0 \\
&\quad - b_n\int \boldsymbol{\kappa}^{(1)}(u)u'\frac{\partial\alpha(v)f_V(v)}{\partial v}\Big|_{v=\theta(x_0,z,\beta_0)+\tilde{u}}\frac{\partial\theta(X_j, Z_j, \beta_0)}{\partial\beta'}\mathbb{E}[\epsilon|X = x_0, Z = z]f_{Z|X}(z|x_0)dudz\rho_0.
\end{aligned}
$$

By evenness of K, $\mathbf{S}_0^{-1}\int \boldsymbol{\kappa}^{(1)}(u)du = \mathbf{0}$. Then, the leading term is

$$
-\mathbf{S}_0^{-1}\int \boldsymbol{\kappa}^{(1)}(u)u'du\frac{\partial}{\partial v}\Big[\frac{\tau(v)f_V(v)}{f_{V|X}(v|x_0)}\Big]_{v=\theta(x_0,z,\beta_0)}\frac{\partial\theta(x_0, z, \beta_0)}{\partial\beta'}\mathbb{E}[\epsilon|X = x_0, Z = z]f_{Z|X}(z|x_0)dz(\hat{\beta}_n - \beta_0)
$$

and (8) equals

$$
\frac{1}{n}\sum_{i=1}^n -\mathbf{e}_1'\mathbf{S}_0^{-1}\int \frac{\partial\boldsymbol{\kappa}(u)}{\partial u'}u'du\mathbb{E}\Big[\frac{\partial}{\partial v}\Big[\frac{\tau(V)f_V(V)}{f_{V|X}(V|x_0)}\Big]\frac{\partial\theta(x_0, Z, \beta_0)}{\partial\beta'}\epsilon\Big|X = x_0\Big]\varphi(W_i, X_i, Z_i) + o_{\mathbb{P}}(1/\sqrt{n}),
$$

which verifies (7) combined with the above derivation.

A.2.5 Proof of Theorems 3 and 5

For the continuous case, from the proof of Theorem 2, $\max_{1\le i\le n}|\hat{\mathbf{S}}_n^{-1}(x_0, \hat{V}_i) - \bar{\mathbf{S}}_n^{-1}(x_0, V_i)| = o_{\mathbb{P}}(1)$, $|\boldsymbol{\kappa}_n(X_i - x_0, \hat{V}_i - \hat{V}_j) - \boldsymbol{\kappa}_n(X_i - x_0, V_i - V_j)| \le K_n(X_i - x_0)\bar{K}_n(V_i - V_j)O_{\mathbb{P}}(n^{-1/2}b_n^{-1})$ uniformly over (i, j), $\sup_{|x-x_0|\le\delta, v\in\overline{\mathcal{V}}^{\delta/2}}|\hat{m}_n(x, v) - m_0(x, v)| = o_{\mathbb{P}}(1)$, and

$$
\sup_{|x-x_0|\le\delta, v\in\overline{\mathcal{V}}}\Big|\frac{1}{n}\sum_{j=1}^n \mathbf{e}_1'\bar{\mathbf{S}}_n(x_0, V_j)\boldsymbol{\kappa}_n(x - x_0, v - V_j)T_j - \mathbb{E}[\mathbf{e}_1'\bar{\mathbf{S}}_n(x_0, V)\boldsymbol{\kappa}_n(x - x_0, v - V)\tau(V)]\Big| = o_{\mathbb{P}}(1).
$$

Thus,

$$
\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \tau_0^{-2}\mathbb{E}\big[\mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(x_0, V)\boldsymbol{\kappa}_n(X_i - x_0, V_i - V)\tau(V)\big|X_i, V_i\big]^2\epsilon_i^2 + o_{\mathbb{P}}(1)
$$

and the second moment of the summand is bounded by a constant multiple of $b_n^{-4d_x}\mathbb{E}[\epsilon^4K^4([X - x_0]/b_n)] = O(b_n^{-3d_x})$ and the desired result follows.

For discrete X, first consider the term $\hat{\psi}_{in}$. From the proof of Theorem 3,

$$\max_{1\le i\le n}\Big|\frac{1}{n}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_j)\boldsymbol{\kappa}_n(\hat{V}_j-\hat{V}_i)T_j-\frac{f_V(V_i)\tau(V_i)}{f_{V|X}(V_i)\rho_0}\Big|=o_{\mathbb{P}}(1)$$

and

$$\max_{1\le i\le n}|\hat{\psi}_{in}-\psi_i|=o_{\mathbb{P}}(1),\qquad \psi_i=\frac{m_0(x_0,V_i)T}{\tau_0}-\mu(x_0)+\frac{\tau(V_i)f_V(V_i)}{f_{V|X}(V_i|x_0)\rho_0\tau_0}\mathbb{1}\{X_i=x_0\}\epsilon_i.$$

For $\hat{\mathbf{\Gamma}}_n$, the first term is

$$\begin{aligned}
&\frac{1}{n^2b_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_1'\hat{\mathbf{S}}_n^{-1}(\hat{V}_i)\boldsymbol{\kappa}_n^{(1)}(\hat{V}_j-\hat{V}_i)\frac{\partial\theta(x_0,Z_j,\hat{\beta}_n)}{\partial\beta'}D_jT_i\hat{\epsilon}_j\\
&=\frac{1}{n^2b_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(V_i)\boldsymbol{\kappa}_n^{(1)}(V_j-V_i)\frac{\partial\theta(x_0,Z_j,\beta_0)}{\partial\beta'}D_jT_i\epsilon_j+O_{\mathbb{P}}(n^{-1/2}b_n^{-2}+R_n)\\
&=b_n^{-1}\mathbb{E}[\mathbf{e}_1'\bar{\mathbf{S}}_n^{-1}(V_i)\boldsymbol{\kappa}_n^{(1)}(V_j-V_i)\frac{\partial\theta(x_0,Z_j,\beta_0)}{\partial\beta'}D_jT_i\epsilon_j]+O_{\mathbb{P}}\big((nb_n^2)^{-1/2}+(nb_n^{d_v+2})^{-1}\big)+o_{\mathbb{P}}(1)
\end{aligned}$$

where $R_n=b_n^q+(\log n/nb_n^{d_v+2})^{1/2}\max\{1,(\log n/n^{1-2/s}b_n^{d_v})^{1/2}\}$ and the second equality uses Hoeffding decomposition. The expectation term converges to the correct object as shown in the proof of Theorem 3. For the derivative estimator $\widehat{m}_n^{(1)}(x_0,v)$, from the analysis similar to the one for $\hat{m}_n(x_0,v)$, we have $\max_{1\le i\le n}|\widehat{m}_n^{(1)}(x_0,\hat{V}_i)-\partial m_0(x_0,\hat{V}_i)/\partial v|=O_{\mathbb{P}}(R_n)$ as R_n defined above. Then, $\hat{\mathbf{\Gamma}}_n\to_{\mathbb{P}}\mathbf{\Gamma}$ and

$$\sigma_n^2=\frac{1}{n}\sum_{i=1}^{n}\big(\psi_i+\mathbf{\Gamma}\varphi(W_i,X_i,Z_i)\big)^2+o_{\mathbb{P}}(1).$$

Thus, the desired result follows from the law of large numbers.

B Additional Identification Results

I illustrate wide applicability of the proxy method developed in the main paper by providing additional identification results in triangular models. The first model concerns multi-valued treatment without the monotonicity condition in the first-stage equation, and the second model involves group-level treatment with potential sorting into groups.

B.1 Multi-valued Treatment Effects

I consider the model

$$Y = g(X, \varepsilon)$$

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

where ε, η, ζ are unobserved, $\zeta \perp\!\!\!\perp (Z, \varepsilon)|\eta$, and $Z \perp\!\!\!\perp \varepsilon|\eta$. The conditional restrictions mean that ζ is an idiosyncratic term and η captures the dependence between X and ε. Here, I focus on the case of X taking discrete values with no natural ordering. For instance, X represents the college major chosen by a student and researchers are interested in the effect of college major on future earnings. With this structure, the process $h(\cdot)$ is often modelled as a discrete choice problem e.g.,

$$X = \arg\max_{x \in \mathcal{X}} \{\nu(x, Z, \eta, \zeta)\}$$

where $\mathcal{X}$ is the set of possible values, $\nu(x, \cdot)$ is some function representing the utility gained from choosing the alternative x, and ζ is an idiosyncratic shock to the utilities. Although this discrete choice structure is not necessary, it helps to ground the idea on concrete terms.

For identification, I assume availability of the proxy variable W for the first-stage unobserved heterogeneity η, which satisfies (i) $W \perp\!\!\!\perp (Z, \zeta)|\eta$ and (ii) η is bounded complete for W. With the discrete choice formulation, the unobserved heterogeneity η may represent the preference over alternatives $\mathcal{X}$. Although it is difficult to measure individual preferences, it may be reasonable to assume that observed covariates have substantial information about people's preferences. In such case, the rank condition for the proxy can be satisfied. Note that the completeness assumption restricts the dimension of η because for the condition to hold, the dimension of W needs to be at least as large as that of η.

The model considered here is different from the model of the main paper in two important aspects. Firstly, the unobserved heterogeneity in the outcome equation ε is unrestricted, except for some conditional independence assumptions. This feature allows for a very general model of the outcome determination. The second difference is that I impose availability of a proxy variable for the first-stage unobserved heterogeneity rather than for the one in the outcome equation. In settings where researchers are willing to impose some structure on the first-stage equation, the current

approach has the advantage that the imposed structures may help justify the proxy assumption. The above discussion on η representing individual preferences illustrates this point.

Now, I briefly sketch the identification argument. The conditional independence restriction $(X, Z) \perp\!\!\!\perp \varepsilon|\eta$ implies

$$\mathbb{E}[Y|X = x, Z] = \int \mathbb{E}[g(x, \varepsilon)|X = x, Z, \eta] f_{\eta|XZ}(\eta|x, Z) d\eta = \int \mathbb{E}[g(x, \varepsilon)|\eta] f_{\eta|XZ}(\eta|x, Z) d\eta$$

and the last integral indicates that if we are able to hold constant $f_{\eta|XZ}(\cdot|X, Z)$, then we would be able to identify partial effects of X on Y. That is, if $f_{\eta|XZ}(\cdot|x, z) = f_{\eta|XZ}(\cdot|\tilde{x}, \tilde{z})$, then the difference $\mathbb{E}[Y|X = x, Z = z] - \mathbb{E}[Y|X = \tilde{x}, Z = \tilde{z}]$ represents a ceteris paribus effect. The proxy method of this paper makes it possible to control for $f_{\eta|XZ}(\cdot|X, Z)$. Specifically,

$$f_{W|XZ}(w|x, z) = \int f_{W|XZ\eta}(w|x, z, \eta) f_{\eta|XZ}(\eta|x, z) d\eta = \int f_{W|\eta}(w|\eta) f_{\eta|XZ}(\eta|x, z) d\eta$$

where the first equality uses the law of total probability and the second uses the conditional independence restriction of the proxy variable. Then, the completeness assumption implies that there exists some mapping Ψ such that $\Psi(f_{W|XZ}(\cdot|x, z))(\eta) = f_{\eta|X,Z}(\eta|x, z)$, and thus, conditioning on $V = f_{W|XZ}(\cdot|X, Z)$ holds constant $f_{\eta|XZ}(\cdot|X, Z)$, which in turn enables identification of partial effects. Similarly to the main paper's result, the conditional proxy distribution is a valid control function in the current setting.

The above argument did not require any monotonicity condition on the first-stage equation. As previously mentioned, many existing results in triangular models impose some form of monotonicity condition on $h(\cdot)$ with respect to η, and thus, these existing results do not apply to the case where the endogenous variable is multi-valued and has no natural ordering. Recent developments that avoid monotonicity restrictions include Lee and Salanié (2018), who analyzed the marginal treatment effects by imposing availability of continuous instruments and specific structures on the first-stage equation. Unlike their paper, I focus on conditional distributional treatment effects such as average and quantiles. On the other hand, the proxy condition accommodates more general specifications for the first-stage equation than those imposed by Lee and Salanié.

B.2 Group-Level Treatment Effects

Next, I consider a model where partial effects of interest concern group-level variables. In the model, there exist individuals and groups, indexed by i and s, respectively. The outcome of interest, denoted by Y_{is}, is determined by

$$Y_{is} = g(X_s, \nu_i, u_s) \qquad i = 1, \ldots, N,\ s = 1, \ldots, S$$

where $g(\cdot)$ is an unknown function, X_s is observed group-level characteristics (i.e., group-level treatment), ν_i is a individual unobserved variable, and u_s represents unobserved group-level features. Also, we observe individual covariates W_i. In principle, W_i can enter the outcome equation, but as I focus on the effect of X_s on Y_{is}, I subsume W_i in ν_i as a subvector. In the sequel, I write $\varepsilon_{is} = (\nu_i, u_s)$.

In the model, the outcome is defined for all groups, but a researcher only observes the outcome variable for the group to which an individual belongs. That is, with $J_i \in \{1, \ldots, S\}$ representing the group to which agent i belongs, we only observe $Y_{iJ_i} = \sum_{s=1}^{S} Y_{is} \mathbb{1}\{J_i = s\}$. I model the group determination by

$$J_i = J(Z_1, \ldots, Z_S, \Theta_i, \zeta_{i1}, \ldots, \zeta_{iS})$$

where $J(\cdot)$ is a nonparametric function, Z_s denotes observed group-level characteristics important for group determination, Θ_i represents unobserved individual preference for different group features, and ζ_{is} is an idiosyncratic term.

To fix ideas, consider the example of education production functions. Here, Y denotes some student outcome of interest (e.g., test scores), X and Z represent observed school characteristics such as teacher quality, ν is student's academic motivation, u is unobserved school features, and Θ is student's preference for various school characteristics. Then, $J(\cdot)$ represents the "selection equation," which determines what school students attend, and $g(\cdot)$ is the education production function for academic performance. In this setting, identification of treatment effects is challenging as ν and Θ are correlated (i.e., highly motivated students value certain school characteristics) and selection into school is partly determined by Θ, creating dependence between X_{J_i} and ν_i (e.g., good quality schools attract highly motivated students).

The identification idea is to use student characteristics W as a proxy variable for unobserved preference Θ. In particular, analogous to the identification result discussed in the main paper, the school-level distribution of student characteristics $V = f_{W|XZ}(\cdot|X, Z)$ plays the role of a control function. To formalize this argument, I impose the following assumptions.

Identification Assumption

(i) $(X_s, Z_s, u_s)_{s=1}^S \perp\!\!\!\perp (W_i, \nu_i)|\Theta_i$, $(X_s, Z_s)_{s=1}^S \perp\!\!\!\perp (u_s)_{s=1}^S|\Theta_i$, and $(\zeta_{is})_{s=1}^S$ is independent of everything else conditional on Θ_i.

(ii) The following mapping, defined on the set of bounded and integrable functions,

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

is injective.

(iii) With respect to some σ-finite product measure, the distribution of $W_i \times \varepsilon_{is} \times Z_1 \times \cdots \times Z_S \times \Theta_i$ is absolutely continuous, the conditional density of Θ_i given Z_{J_i} is bounded, and the conditional density of W_i given Z_{J_i} is bounded and continuous.

(iv) The joint distribution of $(W_i, X_{J_i}, Z_{J_i}, \Theta_i, \varepsilon_{iJ_i})$ is identical across i, and the distribution of $(Y_{iJ_i}, W_i, X_{J_i}, Z_{J_i})$ is identifiable from the data.

(v) For some non-empty set $\mathcal{X}_0 \subset \text{supp}(X_s)$, the support of $f_{W|Z_J}(\cdot|Z_{J_i})$ conditional on $X_{J_i} = x$ equals the unconditional one for $x \in \mathcal{X}_0$.

The first part of (i), $(X_s, Z_s, u_s)_{s=1}^S \perp\!\!\!\perp (W_i, \nu_i)|\Theta_i$, roughly states that school characteristics and student variables are independent *before* selection into groups occurs. In particular, it does not impose $(X_{J_i}, Z_{J_i}, u_{J_i}) \perp\!\!\!\perp (W_i, \nu_i)$, where the subscript J_i denotes that they are observed *after* selection. To rationalize such independence condition, consider the following scenario: first, school features and student characteristics are drawn independently from some distributions, and then students determine which school to attend. This may be a reasonable model given that schools are established first and then households make decisions to relocate near their desired schools. Note that this independence requirement only needs to hold after conditioning on Θ_i, which allows

for more general settings than complete independence of school and student characteristics. For instance, it accommodates some dependence through multiple-stage selection (e.g., first select into a metropolitan area, and then select into a school district/neighborhood) as long as the first-stage selection only depends on Θ_i. In a sense, this conditional independence assumption is essential for analysis of selection bias because without such independence assumption, endogeneity would be present even if there were no selection.

The second part of (i), $(X_s, Z_s)_{s=1}^S \perp\!\!\!\perp (u_s)_{s=1}^S | \Theta_i$, requires that the unobserved school characteristics u_s be independent of other school features. This requirement means that a researcher has good measurements of school features that enter the education production function. Admittedly, this independence condition can be stringent. Yet, without such restriction, endogeneity would be present even if it were not for selection into groups, via dependence between X_s and u_s. Since I focus on the issue of selection bias, I maintain this assumption. The third part of (i) imposes that $(\zeta_{i1}, \ldots, \zeta_{iS})$ is a purely idiosyncratic term, independent of all the other variables.

The condition (ii) is a variant of the bounded completeness assumption. A sufficient condition for this injectivity requirement is that $\mathbb{P}[\mathbb{E}[m(\Theta)|W] = 0] = 1$ implies $\mathbb{P}[m(\Theta) = 0] = 1$ for any bounded function m and the ratio $f_{\Theta|XZ}/f_\Theta$ is bounded. For discussion of the completeness assumption, see the main paper. In this setting, if a researcher is willing to impose more structures on the selection equation, we can find some primitive conditions for (ii). For instance, Altonji and Mansfield (2018) posit that the group choice is determined by a random utility discrete choice model and in particular the preference variable Θ_i takes the form

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

where Γ is some conformable non-stochastic matrix and ω_i is an unobserved random vector. A classical result states that completeness holds if ω_i given W_i has a mean-zero normal distribution with a fixed non-singular covariance matrix and Γ is of full column rank. More generally, (1) $\omega_i \perp\!\!\!\perp W_i$, (2) the characteristic function of ω_i is non-zero everywhere, and (3) $\text{supp}(\Gamma W_i) = \mathbb{R}^d$ with $d = \dim(\Theta_i)$ are a set of sufficient conditions for bounded completeness (Mattner, 1993).

The condition (iii) collects mild regularity conditions, and (iv) concerns identical distribution and identifiability of the joint distribution of observed variables. A sufficient condition for

identifiability is that group-level variables (X_s, Z_s, u_s) are i.i.d. draws across s, individual-level variables (W_i, ν_i, Θ_i) are i.i.d. draws across i, the idiosyncratic shock ζ_{is} is i.i.d. across (i, s), and $(X_s, Z_s, u_s) \perp\!\!\!\perp (W_i, \nu_i, \Theta_i)$. This random sampling assumption is an easy-to-interpret sufficient condition, but we can accommodate a wide class of data generating processes. For instance, we can allow for school features (X_s, Z_s, u_s) to exhibit spatial dependence via mixing conditions or cluster structures.

The condition (v) is the common support condition. As discussed in the main paper, we can weaken this requirement by focusing on subpopulations. For simplicity, I consider the sufficient condition to identify the unconditional ASF.

With the above assumptions, I can use an argument similar to those for Theorem 1 in the main paper and in Section B.1 to show

$$\mathbb{E}[Y_{iJ_i}|X_{J_i} = x, V_i] = \mathbb{E}[g(x, \varepsilon_{is})|V_i]$$

where $V_i = f_{W|XZ}(\cdot|X_{J_i}, Z_{J_i})$. Thus, with the common support condition (v), the proxy method identifies the ASF

$$\mu(x) = \int g(x, e) f_\varepsilon(e) de, \qquad x \in \mathcal{X}_0$$

and this object can be used to recover average effects of group-level treatments.

C Additional Parameters of Interest

In this section, I discuss parameters of interest other than the ASF, namely, the quantile structural function (QSF, Imbens and Newey, 2009) and the local average response (LAR Altonji and Matzkin, 2005).

The QSF is the quantile of $g(x, \varepsilon, \zeta)$ for fixed x. Note that the distribution is computed under the marginal distribution of (ε, ζ). Under an appropriate support condition, identification of the QSF follows from the same argument as for Theorem 1 by replacing Y with $\mathbb{1}\{Y \leq y\}$ for $y \in \mathbb{R}$. Namely, it identifies $\mathbb{P}[g(x, \varepsilon, \zeta) \leq y]$ for fixed x and the left-inverse of this distribution function corresponds to the QSF. As in the ASF, the common support condition can be restrictive. One way to relax this assumption is to follow the partial identification approach of Imbens and Newey

(2009). Also, the flexible parametric modelling of Chernozhukov et al. (2020) and Newey and Stouli (2019) can be applied as described in the main paper.

For the LAR, assume X is continuously distributed and $g(x, \varepsilon, \zeta)$ is continuously differentiable in x with bounded derivatives. Then, the LAR is defined as

$$\mathbb{E}[g_x(x, \varepsilon, \zeta)|X = x], \qquad g_x(x, e, s) = \frac{\partial g(x, e, s)}{\partial x}.$$

Given the control function result $X \perp\!\!\!\perp \varepsilon|V$, this parameter is identified if for each $v \in \text{supp}(V|X = x)$, the support $\text{supp}(X|V = v)$ contains a neighborhood of x. This support condition corresponds to Assumption 2.2 of Altonji and Matzkin (2005).

D Bibliography

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

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

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

Chen, X. and K. Kato (2020): "Jackknife Multiplier Bootstrap: Finite Sample Approximations to the U-proces supremum with Applications," *Probability Theory and Related Fields*, 176, 1097–1163.

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

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

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

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

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

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric Estimation with Generated Covariates," *Econometric Theory*, 32, 1140–1177.

Masry, E. (1996): "Multivariate Regression Estimation Local Polynomial Fitting for Time Series," *Stochastic Processes and Their Applications*, 65, 81–101.

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

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

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

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

——— (1995): "Uniform Ratio Limit Theorems for Empirical Processes," *Scandinavian Journal of Statistics*, 22, 271–278.