# Two-way Exclusion Restrictions in Models with Heterogeneous Treatment Effects: Supplementary Materials

Shenglong Liu[◇] Ismael Mourifié[†] Yuanyuan Wan[‡]

This is the supplementary material for Liu, Mourifie, and Wan (2019). In Section 1, we state the assumptions made in the main text for the convenience of readers. In Section 2, we provide asymptotic results of our estimation when $Z$ is continuous. We discuss in details why "mother's education" is not a valid IV for our data set in Section 3. Additional empirical and simulation results are included in Section 4 and Section 5, respectively.

## 1. Assumptions in the Main Text

**Assumption 1 (Exclusion restrictions).** *(i) The variable $S$ is excluded from the observed treatment, i.e. $D = \vartheta(X, Z, \varsigma)$ for some unknown measurable functions $\vartheta$ and random vector $\varsigma$. (ii) The variable $Z$ does not enter $f_d(S, X)$ for each $d \in \mathscr{D} \equiv \{0, 1, \cdots, T\}$.*

**Assumption 2 (Independence).** $(U, \varsigma) \perp S|X, Z$*, where* $U = (U_0, U_1, \cdots, U_T)'$.

**Assumption 3 (Differentiability).** *$S$ is continuous. Let $\mathscr{S}_x$ be the support of $S$ conditional on $X = x$. Then for each $x \in \mathscr{X}$, $f_d(\cdot, x)$ for $d = 0, 1, \cdots, T$ is continuously differentiable in the interior of $\mathscr{S}_x$.*

**Assumption 4.** $\{(Y_i, D_i, X_i, S_i, Z_i)\}_{i=1}^n$ *are i.i.d. observations.*

**Assumption 5.** *The support of the conditional distribution of $Z|(S, X) = (s, x)$ does not depend on $(s, x)$. Furthermore, $\mathbb{V}[\boldsymbol{\pi}_0(x, Z)]$ is positive definite.*

**Assumption 6.** *The bandwidth $h$ is chosen such that $h \propto n^{-\frac{1}{6+(d_s+d_x)-\delta}}$ for some $0 < \delta < 1$.*

**Assumption 7.** *(i) The conditional density of $(S, X)$ given $Z = z$ is bounded away from 0 and has bounded first-order derivative over its compact support for each $z \in \mathscr{Z}$. (ii) $\boldsymbol{\pi}(\cdot)$ and $\mathbb{E}[Y|W = \cdot]$ are $q+1$ times continuously differentiable for some $q \geq 2$. (iii) There exists some $\nu > 2$ such that $\mathbb{E}\|U\|^{\nu}$ is finite.*

**Assumption 8.** *The symmetric kernel $K(\cdot)$ has support $[-1, 1]$, integrates to one, and is continuously differentiable.*

## 2. Estimation When $Z$ is Continuous

The first stage estimation is similar to the discrete case. Let $d_m = d_x + d_z + d_s$ and $d_\pi = d_x + d_z$: hence $d_m$ and $d_\pi$ are the dimensions of the arguments in $m$ and $\pi$, respectively. Define:

$$\hat{\boldsymbol{\alpha}}^m = \operatorname{argmin}_{\boldsymbol{\alpha}} \frac{1}{2n} \sum_{i=1}^n \boldsymbol{K}_s \left( \frac{S_i - s}{h} \right) \boldsymbol{K}_x \left( \frac{X_i - x}{h} \right) \boldsymbol{K}_z \left( \frac{Z_i - z}{h} \right) \times \left( Y_i - \mathcal{P}(W_i - w, \boldsymbol{\alpha}, p) \right)^2,$$

and for $d = 1, 2, \cdots, T$,

$$\hat{\boldsymbol{\alpha}}^{\pi_d} = \operatorname{argmin}_{\boldsymbol{\alpha}} \frac{1}{2n} \sum_{i=1}^n \boldsymbol{K}_x \left( \frac{X_i - x}{h} \right) \boldsymbol{K}_z \left( \frac{Z_i - z}{h} \right) \times \left( 1\{D_i = d\} - \mathcal{P}((X_i - x, Z_i - z), \boldsymbol{\alpha}, p) \right)^2.$$

Then analogously to the discrete chase, we define our estimator $\hat{m}(s, x, z)$ to be the estimated coefficient corresponding to the linear term $(S_i - s)$ in the first regression and $\hat{\pi}_d(x, z)$ to be the coefficient associated with the constant term in the second regression.

We state the assumptions for the continuous case as below. These assumptions strengthen those for the discrete case and are needed to derive the uniform Bahadur representation (Kong, Linton, and Xia, 2010) of the first stage estimators $\hat{m}$ and $\hat{\boldsymbol{\pi}}$. With the Bahadur representation, we can approximate the estimator $\hat{\beta}(s, x)$ by a U-statistics, from which we derive its limiting distribution.

**Assumption 9.** *The bandwidths $h$ satisfies (i) $h \to 0$, $nh^{d_m+2} \to \infty$; (ii) $nh^{d_\pi+2(p+1)} \to 0$; (iii) $nh^{d_\pi+d_z} \to \infty$ in polynomial rates.*

**Assumption 10.** *(i) The joint density $g_w$ of $W$ is bounded away from $0$ and has bounded first order derivative over its bounded support $\mathscr{W}$. (ii) The conditional density $g_{w|u}$ of $W$ given $U = u$ exists and is bounded for any $u$ in its bounded support $\mathscr{U}$. (iii) $\mathbb{E}[Y|W = \cdot]$ and $\Pr(D = d|X = x, Z = z)$, for each $d = 0, 1, \cdots, T$, are $q+2$ times continuously differentiable for some $q \geq p$.*

Assumption 9-(i) and (ii) are the bandwidth conditions to apply the uniform Bahadur representation (Kong, Linton, and Xia, 2010) to the first stage estimators for $m$ and $\boldsymbol{\pi}$, respectively. Assumption 9-(ii) also plays a role of under-smoothing and eliminates the first stage bias. Assumption 9-(iii) ensures the cross-product remainder terms of the Bahadur representations of $\hat{m}$ and $\hat{\boldsymbol{\pi}}$ are negligible for the second stage estimation. It is implied by Assumption 9-(i) and (ii) when $d_z \leq d_s + 2$. The intuition of Assumption 9-(iii) is as following: the rate of convergence of the cross-product of the remainder terms from the first stage depends on the dimension of $Z$ and $X$ and the rate of the second stage estimator depends on $S$ and $X$. Therefore, for the cross-product of the remainder terms to be negligible, the dimension of $Z$ can not be too high compared with $S$. In the case where all variables are univariate, continuous and the degree of polynomial is chosen to be $p = 2$, the rate condition is satisfied if we choose $h = n^{-r}$ for some $r \in (1/8, 1/5)$. Note that using $p = 2$ to estimate the first-order derivative of a function with three arguments, the optimal rate is $n^{-1/9}$. Hence, the required choice of $r \in (1/8, 1/5)$ is effectively under-smoothing. Assumption 10 requires the model admits enough smoothness, depending on the dimension of the arguments of the unknown functions.

**Proposition 1.** *Let $(s, x)$ be an interior point of the joint support of $(S, X)$. Suppose that Assumptions 1 to 5 and 8 to 10 are satisfied, then $\hat{\beta}(s, x) \xrightarrow{p} \beta(s, x)$ and furthermore,*

$$\sqrt{nh^{d_x+d_s+2}} \left\{ \hat{\beta}(s, x) - \beta(s, x) \right\} \xrightarrow{d} N(0, V^{-1}\Omega_m V^{-1}),$$

*where $V = \mathbb{E}[\boldsymbol{\pi}(x, Z_i)\boldsymbol{\pi}'(x, Z_i)]$ and $\Omega_m$ is defined in Equation (A.5).*

Proof. *Let* $\boldsymbol{K}_w = \boldsymbol{K}_s \times \boldsymbol{K}_x \times \boldsymbol{K}_z$ *and also write* $\boldsymbol{K}_{t,h}(\cdot)$ *for* $\boldsymbol{K}_t(\cdot/h)$, $t \in \{s, x, z\}$. *By Lemma 1 (notation defined therein), we have uniformly in* $w$,

$$\begin{aligned}\hat{m}(w) &= m(w) \\ &\quad \underbrace{- \frac{1}{nh^{d_m+1}} \Sigma_{n,m}^{-(s,\cdot)} H_n^{-1} \sum_{i=1}^{n} \boldsymbol{K}_{s,h}(S_i - s) \boldsymbol{K}_{x,h}(X_i - x) \boldsymbol{K}_{z,h}(Z_i - z) \epsilon_i^m \mu^m(W_i - w)}_{\eta_{m,n}(w)} \\ &\quad + \underbrace{O_p(h^p)}_{r_{m,1}} + \underbrace{O_p\left(\frac{\log n}{nh^{d_m+1}}\right)}_{r_{m,2}}. \end{aligned} \tag{1}$$

*To save notation we suppress the subscript of* $\pi_d$, $d = 1, \cdots, T$ *and use* $\pi$ *to denote a generic element in the vector* $\boldsymbol{\pi}_0 = [\pi_1, ..., \pi_T]'$; *likewise we use* $\hat{\pi}$ *to denote a generic element in* $\hat{\boldsymbol{\pi}}$. *Lemma 2 shows that uniformly in* $(x, z)$,

$$\begin{aligned}\hat{\pi}(x,z) = \pi(x,z) &\underbrace{- \frac{1}{nh^{d_\pi}} \Sigma_{n,\pi}^{-(1,\cdot)} H_n^{-1} \sum_{i=1}^{n} \boldsymbol{K}_{x,h}(X_i - x) \boldsymbol{K}_{z,h}(Z_i - z) \epsilon_i^\pi \mu^\pi(\tilde{W}_i - \tilde{w})}_{\eta_{\pi,n}(x,z)} \\ &+ \underbrace{O_p(h^{p+2})}_{r_{\pi,1}} + \underbrace{O_p\left(\frac{\log n}{nh^{d_\pi}}\right)}_{r_{\pi,2}}. \end{aligned} \tag{2}$$

*Recall that our estimator is defined as*

$$\hat{\beta}(s,x) = \left(\frac{1}{n} \sum \hat{\boldsymbol{\pi}}(x, Z_i) \hat{\boldsymbol{\pi}}(x, Z_i)'\right)^{-1} \left(\frac{1}{n} \sum \hat{\boldsymbol{\pi}}(x, Z_i) \hat{m}(s, x, Z_i)\right).$$

*First consider the denominator; it is easy to see that under the assumptions of Proposition 1 and the representation in Equation* (2),

$$\frac{1}{n} \sum_i \hat{\boldsymbol{\pi}}(x, Z_i) \hat{\boldsymbol{\pi}}'(x, Z_i) \xrightarrow{p} \mathbb{E}[\boldsymbol{\pi}(x, Z_i) \boldsymbol{\pi}'(x, Z_i)] \equiv V.$$

*For the numerator, it follows from Lemma 5 that* $\eta_{n,m}$ *and* $\eta_{n,\pi}$ *are* $o_p(1)$ *and applying the law of large number, we have*

$$\frac{1}{n} \sum \hat{\boldsymbol{\pi}}(x, Z_i) \hat{m}(s, x, Z_i) \xrightarrow{p} \mathbb{E}[\boldsymbol{\pi}(x, Z_i) m(s, x, Z_i)]$$

*The consistency of the estimator follows.*

*Regarding the limiting distribution, we consider the following decomposition,*

$$\begin{aligned}\frac{1}{n}\sum \hat{\boldsymbol{\pi}}(x,Z_i)\hat{m}(s,x,Z_i) - \mathbb{E}[\boldsymbol{\pi}(x,Z_i)m(s,x,Z_i)] \\ = \left(\frac{1}{n}\sum \hat{\boldsymbol{\pi}}(x,Z_i)\hat{m}(s,x,Z_i) - \frac{1}{n}\sum \boldsymbol{\pi}(x,Z_i)m(s,x,Z_i)\right) \\ + \left(\frac{1}{n}\sum \boldsymbol{\pi}(x,Z_i)m(s,x,Z_i) - \mathbb{E}[\boldsymbol{\pi}(x,Z_i)m(s,x,Z_i)]\right)\end{aligned}$$

*The second term is standard and is of order $O_p(1/\sqrt{n})$. It remains to deal with the first term. For notational simplicity, we write $\eta_{m,n}(s,x,Z_i)$ as $\eta_{m,n}(Z_i)$, and write $\eta_{\pi,n}(x,Z_i)$ for $\eta_{\pi,n}(Z_i)$.*

$$\begin{aligned}&\frac{1}{n}\sum_i \hat{\boldsymbol{\pi}}(x,Z_i)\hat{m}(s,x,Z_i) - \frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)m(s,x,Z_i) = \frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\{\eta_{m,n}(Z_i)+r_{m,1}+r_{m,2}\} \\ &+\frac{1}{n}\sum_i m(s,x,Z_i)\{\eta_{\pi,n}(Z_i)+r_{\pi,1}+r_{\pi,2}\} + \frac{1}{n}\sum_i \{\eta_{m,n}(Z_i)+r_{m,1}+r_{m,2}\}\{\eta_{\pi,n}(Z_i)+r_{\pi,1}+r_{\pi,2}\} \\ &= \frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\eta_{m,n}(Z_i) + \frac{1}{n}\sum_i m(s,x,Z_i)\eta_{\pi,n}(Z_i) + \frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\{r_{m,1}+r_{m,2}\} \\ &+ \frac{1}{n}\sum_i m(s,x,Z_i)\{r_{\pi,1}+r_{\pi,2}\} + \frac{1}{n}\sum_i \{\eta_{m,n}(Z_i)+r_{m,1}+r_{m,2}\}\{\eta_{\pi,n}(Z_i)+r_{\pi,1}+r_{\pi,2}\}.\end{aligned}$$

*The third and fourth RHS terms are of order smaller than $O_p(1/\sqrt{nh^{d_x+d_s+2}})$ by Lemma 4. The last RHS term is of order smaller than $O_p(1/\sqrt{nh^{d_x+d_s+2}})$ by Lemma 7. By Lemma 3,*

$$\sqrt{nh^{d_x+d_s+2}}\left\{\frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\eta_{m,n}(s,x,Z_i) + \frac{1}{n}\sum_i m(s,x,Z_i)\eta_{\pi,n}(x,Z_i)\right\} \xrightarrow{d} N(0,\Omega_m).$$

*It then follows that*

$$\sqrt{nh^{d_x+d_s+2}}\left\{\frac{1}{n}\sum \hat{\boldsymbol{\pi}}(x,Z_i)\hat{m}(s,x,Z_i) - \mathbb{E}[\boldsymbol{\pi}(x,Z_i)m(s,x,Z_i)]\right\} \xrightarrow{d} N(0,\Omega_m),$$

*or alternatively,*

$$\sqrt{nh^{d_x+d_s+2}}\left\{\hat{\beta}(s,x) - \beta(s,x)\right\} \xrightarrow{d} N(0, V^{-1}\Omega_m V^{-1}).$$

□

We can see from Proposition 1 that the convergence rate of $\hat{m}$, instead of $\hat{\pi}_d$, determines the convergence rate of $\hat{\beta}$ because $m$ is the first-order derivative of a conditional expectation. If $m$ has a higher degree of smoothness, then $\hat{\beta}$ will converge faster. Also, under Assumption 5, the dimension of $Z$ does not affect the convergence rate of $\hat{\beta}$ since $Z$ is averaged out in the second stage with respect to its marginal (empirical) distribution. The factor $d_s + d_x$ reflects the fact that the estimand $\beta$ is a function evaluated at a $d_s + d_x$-dimensional vector $(s, x)$; the factor 2 in the power of $h$ reflects that $m$ is the first-order derivative of the function $\mathbb{E}[Y|W = w]$ with respect to $s$.

For inference, we again propose to estimate the asymptotic variance by plugging-in consistent estimator $\widehat{V}$ and $\widehat{\Omega}_m$, respectively. As before, we provide the formula for $\widehat{\Omega}_m$ (Equation B.2) and an example in Appendix B.

## 3. Validity of Mother's Education as an IV in our data set

In this section, we test the necessary implications of LATE-validity assumption when using mother's education as an instrument. Following similar derivations in Mourifié and Wan (2017, Equation 1), the LATE assumptions imply the following four testable necessary conditions, that is, for any $A \subseteq \mathscr{Y}$,

$$\begin{cases} \mathbb{P}(Y \in A, D = 2|Z = 0) \leq \mathbb{P}(Y \in A, D = 2|Z = 1) \leq \mathbb{P}(Y \in A, D = 2|Z = 2) \\ \mathbb{P}(Y \in A, D = 0|Z = 2) \leq \mathbb{P}(Y \in A, D = 0|Z = 1) \leq \mathbb{P}(Y \in A, D = 0|Z = 0) \end{cases}. \tag{3}$$

Mourifié and Wan (2017) show that each of the inequality constraints can be rewritten as a conditional moment inequality and the null hypothesis that all inequalities in (3) hold can be written as

$$H_0 : \theta_0 \equiv \sup_{y \in \mathscr{Y}, j=1,\cdots,4} \theta(y, j) \leq 0, \quad H_1 : \theta_0 > 0, \tag{4}$$

where $\theta(y, j)$, $j = 1, \cdots, 4$, represent a conditional moment inequality. We test the null hypothesis in (4) at the province level. Mourifié and Wan (2017) propose using the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013), which we follow here. We used the "clrtest" Stata command of Chernozhukov, Kim, Lee, and Rosen (2013) to conduct the test and also use the "clrbound" command to calculate the lower bound of the confidence set of $\theta_0$. The results are reported in Table 1. We can see that the test rejects the null hypothesis in (3) or (4) for a significant portion of provinces, meaning that the dataset under analysis here shows strong evidences against the use of mother's education as a conventional IV to estimate the LATE.[1] Notice that in the rejected cases, the lower boundaries of the confidence interval of $\theta_0$ are all above zero, as demonstrated in the right panel of the table. In Table 1, the unknown conditional expectations are estimated with the local regression method. As a robustness check, we also conduct the tests using sieve estimation and the results are qualitatively similar and reported in Table 2.

## 4. Additional Empirical Results

In this section, we conduct a few robustness checks for our empirical results.

### *4.1. Adding Covariates.*

So far our analysis uses the whole sample. We also estimate the model using subsamples based on gender, ethnic group (Han and minority), and age (below or above the median age of the whole sample).[2] Although insignificant in some subsamples, the above-mentioned

[1] Since we are testing the hypothesis for 31 provinces, it is desirable to ensure that the Family-wise Error Rate (FWER) is controlled at targeted levels. We conducted the test again at 0.1% significance level and found rejections for Hubei, Guangdong, Chongqing, and Xizang. By the multiple testing procedure of Holm (1979), we can conclude that our test rejects the null hypothesis with FWER be controlled by no more than 5%.

[2] The age can be viewed here as a proxy for experiences, which unfortunately we do not observe in this dataset.

Table 1: Testing LATE Assumption by Province (Local Regression)

| Province | Sample Size | Test (clrtest) | | | Lower Bound of CI (clrbound) | | |
|---|---|---|---|---|---|---|---|
| | | 90% | 95% | 99% | 90% | 95% | 99% |
| Beijing | 2,476 | NR[a] | NR | NR | -0.0000049 | -0.0000054 | -0.0000061 |
| Tianjin | 4,762 | NR | NR | NR | -0.0021587 | -0.0022713 | -0.0024809 |
| Hebei | 7,108 | NR | NR | NR | -0.0003199 | -0.0004228 | -0.0006254 |
| Shanxi | 7,872 | NR | NR | NR | -0.0003136 | -0.0004318 | -0.0006733 |
| Neimeng | 2,807 | NR | NR | NR | -0.0002477 | -0.0002662 | -0.0003040 |
| Liaoning | 4,286 | **R** | NR | NR | 0.0000625 | -0.0000813 | -0.0003704 |
| Jilin | 5,018 | NR | NR | NR | -0.0005115 | -0.0005857 | -0.0006417 |
| Heilongjiang | 3,901 | NR | NR | NR | -0.0008943 | -0.0009552 | -0.0010810 |
| Shanghai | 4,401 | **R** | **R** | **R** | 0.0028629 | 0.0020509 | 0.0002967 |
| Jiangsu | 5,284 | NR | NR | NR | -0.0000001 | -0.0000060 | -0.0000131 |
| Zhejiang | 3,894 | NR | NR | NR | -0.0000880 | -0.0000932 | -0.0001027 |
| Anhui | 4,902 | NR | NR | NR | -0.0000531 | -0.0000604 | -0.0000733 |
| Fujian | 3,278 | NR | NR | NR | -0.0001924 | -0.0002106 | -0.0002422 |
| Jiangxi | 3,777 | NR | NR | NR | -0.0001948 | -0.0002129 | -0.0002543 |
| Shandong | 7,737 | NR | NR | NR | -0.0001029 | -0.0001114 | -0.0001313 |
| Henan | 6,666 | NR | NR | NR | -0.0003481 | -0.0003628 | -0.0003927 |
| Hubei | 5,467 | **R** | **R** | **R** | 0.0039942 | 0.0033046 | 0.0019484 |
| Hunan | 6,769 | NR | NR | NR | -0.0000923 | -0.0000973 | -0.0001074 |
| Guangdong | 25,652[b] | **R** | **R** | **R** | 0.0006007 | 0.0005268 | 0.0003795 |
| Guangxi | 4,846 | **R** | **R** | NR | 0.0001403 | 0.0000723 | -0.0000673 |
| Hainan | 2,902 | NR | NR | NR | -0.0002162 | -0.0002688 | -0.0003758 |
| Chongqing | 3631 | **R** | **R** | **R** | 0.0003058 | 0.0002450 | 0.0001268 |
| Sichuan | 6,347 | **R** | NR | NR | 0.0000219 | -0.0000136 | -0.0000824 |
| Guizhou | 3,797 | NR | NR | NR | -0.0000029 | -0.0000031 | -0.0000036 |
| Yunnan | 13,696 | NR | NR | NR | -0.0000022 | -0.0000024 | -0.0000028 |
| Xizang | 2,510 | **R** | **R** | **R** | 0.0000000 | 0.0000000 | 0.0000000 |
| Shanxi | 8,124 | NR | NR | NR | -0.0003338 | -0.0004210 | -0.0005974 |
| Gansu | 7,196 | NR | NR | NR | -0.0000273 | -0.0000458 | -0.0000836 |
| Qinghai | 2,451 | **R** | **R** | **R** | 0.0006214 | 0.0004856 | 0.0002012 |
| Ningxia | 1,521 | NR | NR | NR | -0.0000216 | -0.0000288 | -0.0000418 |
| Xinjiang | 3,380 | NR | NR | NR | -0.0000946 | -0.0001081 | -0.0001312 |

a. "R" stands for rejection of LATE assumptions and "NR" stands for no rejection.
b. For provinces with more than 8000 observations, we choose a random subsample of $8,000$

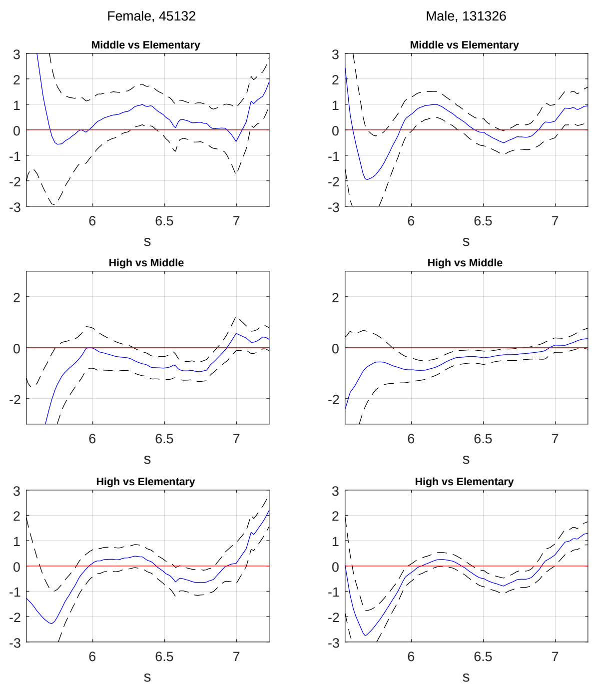
Table 2: Testing LATE Assumption by Province (series estimation)

| Province | Sample Size | Test (clrtest) | | | Lower Bound of CI (clrbound) | | |
|---|---|---|---|---|---|---|---|
| | | 90% | 95% | 99% | 90% | 95% | 99% |
| Beijing | 2,476 | **R** | NR | NR | 0.0004902 | -0.0000099 | -0.0000124 |
| Tianjin | 4,762 | NR | NR | NR | -0.0027308 | -0.0029010 | -0.0032281 |
| Hebei | 7,108 | NR | NR | NR | -0.0006457 | -0.0006887 | -0.0007800 |
| Shanxi | 7,872 | NR | NR | NR | -0.0007102 | -0.0007668 | -0.0008840 |
| Neimeng | 2,807 | NR | NR | NR | -0.0004630 | -0.0004917 | -0.0005436 |
| Liaoning | 4,286 | NR | NR | NR | -0.0002511 | -0.0004393 | -0.0008444 |
| Jilin | 5,018 | NR | NR | NR | -0.0006048 | -0.0006349 | -0.0006952 |
| Heilongjiang | 3,901 | NR | NR | NR | -0.0009264 | -0.0009834 | -0.0011067 |
| Shanghai | 4,401 | **R** | **R** | NR | 0.0021831 | 0.0012057 | -0.0000013 |
| Jiangsu | 5,284 | NR | NR | NR | -0.0002047 | -0.0002246 | -0.0002701 |
| Zhejiang | 3,894 | NR | NR | NR | -0.0000119 | -0.0000134 | -0.0000163 |
| Anhui | 4,902 | NR | NR | NR | -0.0000662 | -0.0000711 | -0.0000814 |
| Fujian | 3,278 | **R** | **R** | NR | 0.0046661 | 0.0021814 | -0.0002473 |
| Jiangxi | 3,777 | NR | NR | NR | -0.0004043 | -0.0004401 | -0.0005066 |
| Shandong | 7,737 | NR | NR | NR | -0.0001303 | -0.0001394 | -0.0001574 |
| Henan | 6,666 | NR | NR | NR | -0.0002255 | -0.0002392 | -0.0002703 |
| Hubei | 5,467 | **R** | **R** | **R** | 0.0001907 | 0.0001295 | 0.0000017 |
| Hunan | 6,769 | NR | NR | NR | -0.0002038 | -0.0002231 | -0.0002601 |
| Guangdong | 25,652 | **R** | NR | NR | 0.0000374 | -0.0000117 | -0.0000192 |
| Guangxi | 4,846 | **R** | **R** | **R** | 0.0008004 | 0.0006593 | 0.0003694 |
| Hainan | 2,906 | **R** | **R** | **R** | 0.0050706 | 0.0041392 | 0.0021845 |
| Chongqing | 3,631 | **R** | **R** | **R** | 0.0002169 | 0.0001490 | 0.0000044 |
| Sichuan | 6,347 | NR | NR | NR | -0.0001346 | -0.0001783 | -0.0002720 |
| Guizhou | 3,797 | **R** | NR | NR | 0.0000127 | -0.0000054 | -0.0000064 |
| Yunnan | 13,696 | NR | NR | NR | -0.0000175 | -0.0000280 | -0.0000451 |
| Xizang | 2,510 | NR | NR | NR | -0.0000003 | -0.0000003 | -0.0000004 |
| Shanxi | 8,124 | NR | NR | NR | -0.0002387 | -0.0003214 | -0.0004953 |
| Gansu | 7,196 | **R** | NR | NR | 0.0000574 | -0.0000181 | -0.0000215 |
| Qinghai | 2,451 | **R** | **R** | NR | 0.0004750 | 0.0003183 | -0.0000097 |
| Ningxia | 1,521 | NR | NR | NR | -0.0000366 | -0.0000438 | -0.0000564 |
| Xinjiang | 3,380 | NR | NR | NR | -0.0004457 | -0.0005053 | -0.0006315 |

"R" stands for rejection and "NR" stands for no rejection.

pattern exists overall. For instance, the same result holds and is significant for both males and females. It holds and is significant for populations below the median age of the sample (27 years old), and holds but is less significant for populations older than 27. It also holds for both Han but is not significant for minority groups.[3] Please see Figures 1 to 3 for details.

Figure 1: Estimates and 95% CI by gender, $c = -0.01$




### *4.2. Re-categorizing to Binary Treatment*

To further examine if our result is robust, we estimate the model by re-categorizing the education levels into a binary treatment and a binary outcome exclusion variable, that is, $\tilde{D} = 0$ (or $\tilde{Z} = 0$) for elementary school and below, and $\tilde{D} = 0$ (or $\tilde{Z} = 1$) for middle school and above. The results for the whole sample are reported in Section 4.2 and we can see that the same pattern exists and is significant. Estimation results based on subsamples are collected in Figure 5 of Section 4. Except that the result is not significant for subsample of minorities (which is the case in triple-valued $D$ and $Z$ as well), we see the same pattern over other subsamples.

### *4.3. Smoothing Constants.*

We also use under-smoothing constant $c = 0$ and $c = -0.03$ and the results are plotted in Figure 6. They are qualitatively similar to those report in the main text.

[3]One possible reason is that minority groups face different bars at entrance exams in China.

Figure 2: Estimates and 95% CI by ethnic group, $c = -0.01$

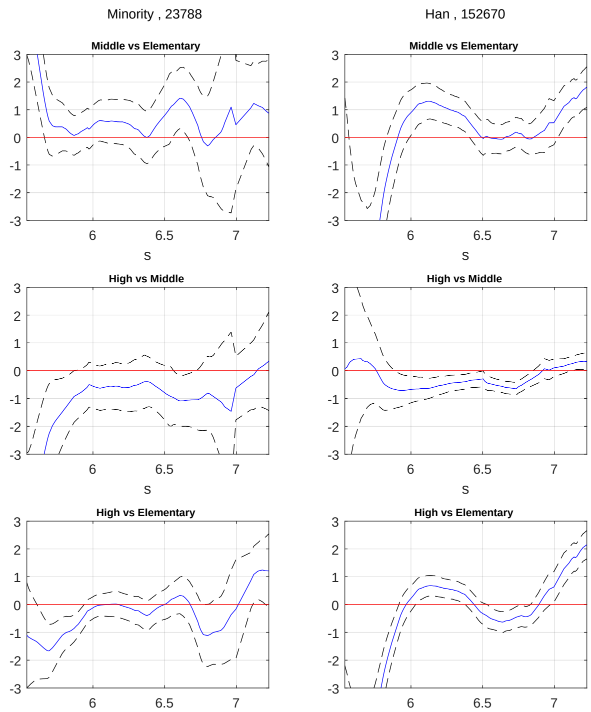



Table 3: Education ($\tilde{D}$) and Mother's Education ($\tilde{Z}$): Re-categorizing

| $\tilde{D}$ | $\tilde{Z}$ = 0 | $\tilde{Z}$ = 1 | Total |
|---|---|---|---|
| 0 | 36,153 | 1,496 | 37,649 |
| 1 | 96,495 | 42,314 | 138,809 |
| Total | 132,648 | 42,810 | 176,458 |

Figure 3: Estimates and 95% CI by age, $c = -0.01$

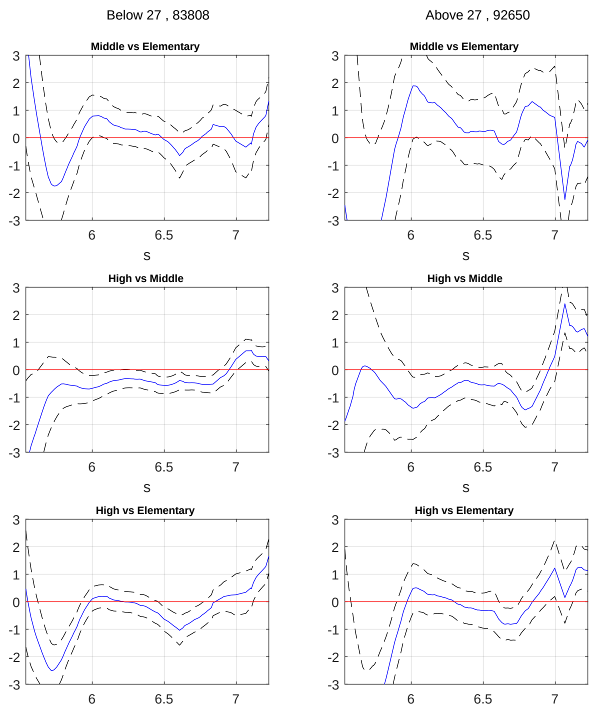



Figure 4: Binary Case, Whole Sample, $c = -0.01$

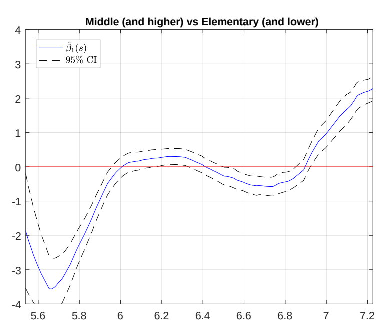


Figure 5: Estimates and 95% CI,, Binary Case, Subsamples, $c = -0.01$

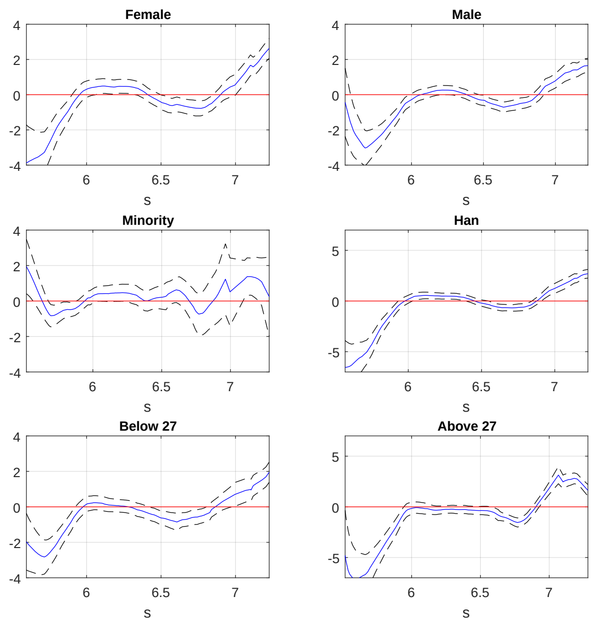
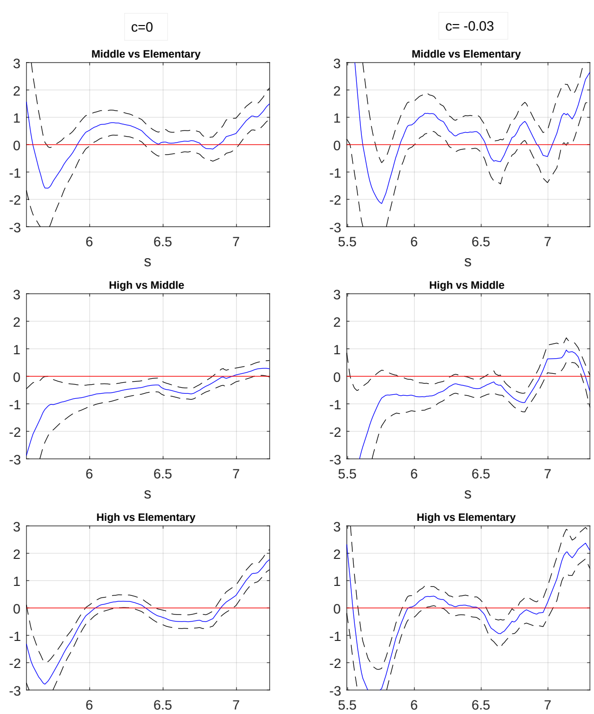



Figure 6: Estimates and 95% CI, by smoothing level



*4.4. Internal Migration*

. The internal labor migration has always been an important factor to consider in research on China's development and inequality, see e.g. Ha, Yi, Yuan, and Zhang (2016) and references therein. In our sample 4.58% individuals have not lived in their Hukou address for the past six months from the survey date, so we consider these individuals as a subsample of internal migrants (across different prefectures). Although internal migration is not the primary focus our analysis, as a robustness check, we conduct our analysis again by excluding this subsample and the result is the same qualitatively and very similar quantitatively (hence omitted here).

## 5. Simulation Results

In this section we provide some numerical examples to investigate the finite sample performance of our estimator. We consider a binary $D$ and triple-valued $Z$. Let $\boldsymbol{e} = (e_1, e_2, e_3, e_4)' \in \mathbb{R}^4$ follow a multivariate normal distribution with zero mean and covariance matrix given by

$$\begin{pmatrix} 1.0 & 0.5 & 0.3 & 0 \\ 0.5 & 1.0 & 0.3 & 0 \\ 0.3 & 0.3 & 1.0 & 0.3 \\ 0 & 0 & 0.3 & 1.0 \end{pmatrix}.$$

Let $\varsigma \sim N(0,1)$. Let $U_0 = 0.1(e_1 - \varsigma)$, $U_1 = 0.1(e_2 + \varsigma)$, $Z = \mathbf{1}\{0.3 \le \Phi(e_3) \le 0.7\} + 2\mathbf{1}\{\Phi(e_3) > 0.7\}$ and $S = 0.02 \times \text{Ceil}[200\Phi(e_4) - 100]$, where $\Phi(\cdot)$ is the standard normal CDF and Ceil$[\cdot]$ is the ceiling function which returns the smallest integer that is no less than its argument. Finally, let $D = \mathbf{1}\{Z + \varsigma - 0.5 > 0\}$. As constructed, $(U_0, U_1)$ is correlated with $\varsigma$ and $Z$, which implies the endogeneity of treatment $D$ and invalidity of $Z$ as a conventional instrumental variable. In the meantime, $S$ is independent with $(U_0, U_1, \varsigma)$ but correlated with $Z$.[4]

We normalize $f_0(s) = 0$ and hence $Y_0 = U_0$. We specify

$$f_1(s) = 0.3e^{-4(s+1)^2} + 0.7e^{-16(s-1)^2}, \quad Y_1 = f_1(S) + U_1$$

In this DGP, the conditional average treatment effect (for $S = s$) is given by $\Delta(s) = f_1(s) - f_0(s)$ (as a function of $s$). Its derivative $\beta_1(s)$ is the parameter of interest:

$$\beta_1(s) = -2.4(s+1)e^{-4(s+1)^2} - 22.4(s-1)e^{-16(s-1)^2}.$$

Figure 7 plots $\Delta(s)$ and $\beta_1(s)$ , respectively.

We consider five sample sizes: $n = 1000 \times 2^k$, for $k \in \{0, 1, 2, 3, 4\}$. We use the Epanechnikov kernel. Since we estimate first-order derivative in the first stage, we use the second-order polynomial ($p = 2$), as recommended by Fan and Gijbels (1996). To the best of our knowledge, there are few results available for choosing the bandwidth optimally in the two-stage nonparametric estimation context that we consider here. Hence, we follow the "Rule of Thumb" bandwidth $h^{ROT}$ proposed in Fan and Gijbels (1996, Section 4.1). In our

[4] In this design, the support of $S$ is actually a fine grid on the unit interval. For comparison, we also estimate $m(s,z)$ by treating $S$ as an ordered discrete random variable, with the latter we apply the method of Li and Racine (2004) and Li, Racine and Wooldridge (2009). It turns out that both methods give very similar results, probably because the grids are fine enough.

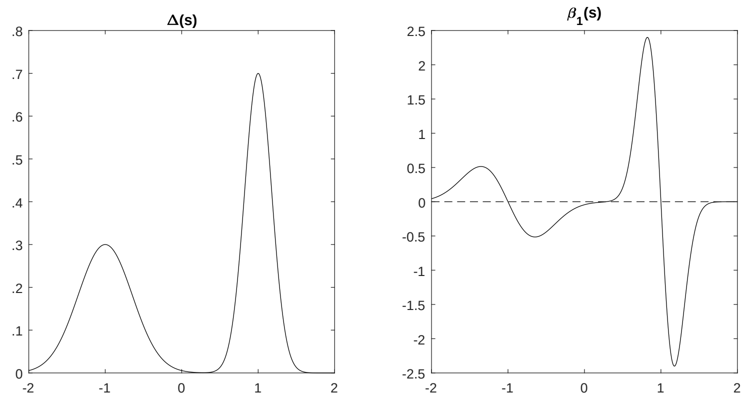



Figure 7: True functions

simulation, $h^{ROT} \propto n^{-\frac{1}{2p+3}} = n^{-\frac{1}{7}}$. We also consider different levels of under-smoothing: $h = h^{ROT} n^c$, where $c \in \{0, -0.01, -0.03\}$. Here $c = 0$ corresponds to no under-smoothing. For each given sample size and under-smoothing level, we consider $1,000$ replications. For each replication, we estimate $\beta_1(s)$ over 200 uniformly spread grids on the support of $S$. We report the mean squared error (MSE) for $\hat{\beta}_1(s)$ with $s \in \{-1.8, -1, 0, 1, 1.8\}$, which corresponds to the $5\%, 25\%, 50\%, 75\%$ and $95\%$ quantiles of $S$, respectively.

Performance of our estimator when choosing $c = -0.01$ is reported in Tables 4 to 6. Figure 8 plots estimates and pointwise confidence band for two random samples of size 2000 and 16000, respectively. First, for each $s$ and $c$, we can see that the MSE decreases as sample size increases, as expected. The MSEs are relatively larger when $s$ is close to the boundary of the support ($s = \pm 1.9$) or the second-order derivative is larger in absolute values ($s = 1$), which is also not surprising. As we increase the level of under-smoothing, we observe the overall pattern that the variance increases and bias decreases. It appears that the trade-off between the bias and variance carries through from the first-stage estimation to the second stage, although the average magnitude of the variance is much larger than the bias. When comparing MSEs across difference sample sizes, we can see that when the sample size increases from 1000 to 2000, the MSEs overall decrease by a greater factor than what our theory predicts ($\propto 2^{4/7} \approx 1.5$). This is possibly because sample sizes are not large enough to fully show asymptotic behavior. If we look at larger sample sizes, we would observe that the factor by which the MSEs decrease is roughly in line with the $n^{2/7}$ convergence rate.

To investigate the performance of the confidence intervals, we calculate the confidence intervals for $\beta_1(s)$, $s \in \{-1.8, -1, 0, 1, 1.8\}$ and their coverage frequencies for the true values at three nominal level $90\%, 95\%$ and $99\%$. As we can see from Table 9, the finite sample coverage frequencies are quite close to nominal levels, especially in larger sample sizes. Similar to the estimation, the performance of confidence intervals are better when $s$ is away from the boundaries of its support. To investigate the precision of the confidence intervals, we test $H_0 : \beta_1(-0.5)$ against $H_1 : \beta_1(-0.5) \neq 0$ by checking if 0 is contained in the confidence interval of $\beta_1(-0.5)$. The rejection frequencies are reported in Table 10, which shows that our test has stronger power against the false null hypothesis since the true value is $\beta_1(-0.5) \approx -0.44$. We also examined other $s$ values and obtained the same qualitative

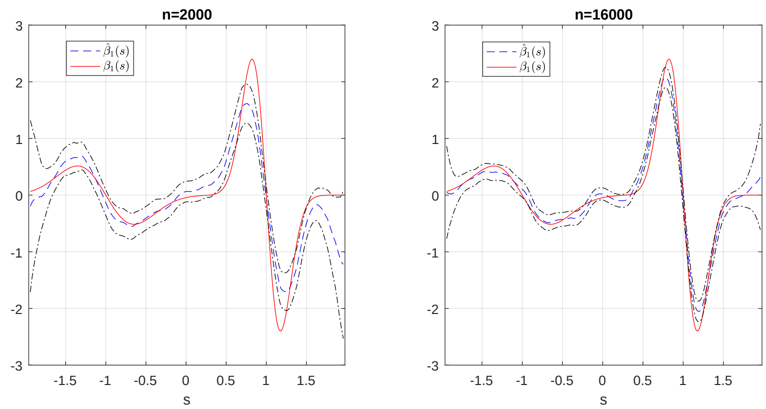



Figure 8: Estimates and Confidence Bands

results.

We also investigate the performance of the pointwise confidence intervals. For this, we calculate the confidence intervals for $\beta_1(s)$, $s \in \{-1.8, -1, 0, 1, 1.8\}$ and their coverage frequencies for the true values at three nominal level $90\%, 95\%$ and $99\%$. As we can see from Table 9, the finite sample coverage frequencies are quite close to nominal levels, especially in larger sample sizes. Similar to the estimation, the performance of confidence intervals are better when $s$ is away from the boundaries of its support. To investigate the precision of the confidence intervals, we test $H_0 : \beta_1(-0.5)$ against $H_1 : \beta_1(-0.5) \neq 0$ by checking if 0 is contained in the confidence interval of $\beta_1(-0.5)$. The rejection frequencies are reported in Table 10, which shows that our test has stronger power against the false null hypothesis since the true value is $\beta_1(-0.5) \approx -0.44$. We also examined other $s$ values and obtained the same qualitative results.

## Appendix A. Auxiliary Lemmas for Proving Proposition 1

This appendix collects auxiliary lemmas for proving Proposition 1. We define some notation first. For the purpose of exposition, we define notation for estimation of $m$. The notation for estimation of $\pi_d$ is similar. For $j = 0, 1, \cdots, p$, let $N_j$ be the number of $d_m$-dimensional vectors $\underline{r}$ such that $|\underline{r}| = j$. Arrange all such vectors in the lexicographical order with the first one is $(0, 0, \cdots, j)$ and last one is $(j, \cdots, 0, 0)$. Let $\tau_j$ be the one to one mapping from an order to the associated vector. For example, $\tau_j(1) = (0, 0, \cdots, j)$, $\tau_j(2) = (0, 1, \cdots, j-1) \cdots$, and $\tau_j(N_j) = (j, 0, \cdots, 0)$. For $j = 0, \cdots, p$, let $\nu_{n,m,\underline{j}}(w) = \int u^{\underline{j}} \boldsymbol{K}_w(u) g_w(w + hu) du$,[5] where here $u$ is a $d_m \times 1$ vector and $u^{\underline{j}}$ stands for the product of powers of elements of $u$ such that the sum of power index equals to $j$. Let $\Sigma_{n,m}$ be a

[5]The general definition in KLX is that $\nu_{n,m,\underline{j}}(w) = \int \boldsymbol{K}_w(u) u^{\underline{j}} g(w + hu) f_w(w + hu) du$, where in KLX's notation, $f$ is the joint density of $W$ and function $g$ is defined as in KLX-Equation. A7 Since we use quadratic loss function, $g(\cdot) \equiv 1$ in our case.

Table 4: Performance of $\hat{\beta}_1(s)$, $c = -0.01$

| | $s$ | $n$ 1000 | 2000 | 4000 | 8000 | 16000 |
|---|---|---|---|---|---|---|
| MSE | -1.8 | 0.3355 | 0.151 | 0.0689 | 0.0408 | 0.019 |
| | -1 | 0.0162 | 0.0091 | 0.0063 | 0.0044 | 0.003 |
| | 0 | 0.0121 | 0.009 | 0.0051 | 0.0033 | 0.0025 |
| | 1 | 0.0775 | 0.0134 | 0.008 | 0.0054 | 0.0047 |
| | 1.8 | 0.2902 | 0.1074 | 0.0547 | 0.0277 | 0.0125 |
| Bias | -1.8 | 0.0058 | -0.013 | -0.0114 | 0.0004 | -0.0067 |
| | -1 | -0.0085 | 0.0037 | -0.0016 | -0.0019 | -0.0027 |
| | 0 | -0.0222 | -0.0221 | -0.0139 | -0.0128 | -0.0109 |
| | 1 | -0.1333 | -0.0006 | -0.001 | -0.0039 | -0.0019 |
| | 1.8 | -0.057 | -0.0673 | -0.0446 | -0.0134 | -0.0068 |
| Variance | -1.8 | 0.3355 | 0.1509 | 0.0687 | 0.0408 | 0.0189 |
| | -1 | 0.0161 | 0.0091 | 0.0063 | 0.0044 | 0.003 |
| | 0 | 0.0116 | 0.0085 | 0.005 | 0.0031 | 0.0024 |
| | 1 | 0.0597 | 0.0134 | 0.008 | 0.0054 | 0.0047 |
| | 1.8 | 0.287 | 0.1028 | 0.0528 | 0.0275 | 0.0124 |

symmetric matrix

$$\Sigma_{n,m} = \begin{pmatrix} \Sigma_{n,m,0,0} & \Sigma_{n,m,0,1} & \cdots & \Sigma_{n,m,0,p} \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma_{n,m,p,0} & \Sigma_{n,m,p,1} & \cdots & \Sigma_{n,m,p,p} \end{pmatrix},$$

where $\Sigma_{n,m,i,j}$ is an $N_i$ by $N_j$ matrix whose $(\ell, k)$ element is $\nu_{n,m,\tau_i(\ell)+\tau_j(k)}$. So $\Sigma_{n,m,0,0}$ is the $(1,1)_{th}$ element of the matrix $\Sigma_{n,m}$ and equals to $\nu_{n,m,\underline{0}}(w)$; $\Sigma_{n,m,0,1}$ is a $1 \times d_m$ vector contains terms of $\nu_{n,m,\underline{1}}$ corresponding to each variable in vector $u$; $\Sigma_{n,m,1,1}$ is a $d_m \times d_m$ matrix which contains elements constructed from $\nu_{n,m,\underline{2}}$ where each elements contains interaction terms from two variables from the vector $u$ etc.. Let $\Sigma_m$ be defined as similar to $\Sigma_{n,m}$ with $\nu_{m,\underline{j}} = g_w(w) \int \boldsymbol{K}_w(u) u^{\underline{j}} du$ replacing $\nu_{n,m,\underline{j}}$. Clearly $\Sigma_m = \Sigma_{n,m} + o(1)$ given $h \downarrow 0$, as shown in KLX Lemma 8.

Let $M(w) = \mathbb{E}[Y|W = w]$. Let $\alpha^m_{\underline{r}}(w)$ be a vector of $|\underline{r}|$-th order partial derivative of $M$ evaluated at $w$ with the position of each term in the vector being arranged in the same lexicographical order as described above. For example, for $w = (s, x, z)$,

$$\alpha^m_{\underline{1}}(w) = \left(\frac{\partial M(w)}{\partial z}, \frac{\partial M(w)}{\partial x}, \frac{\partial M(w)}{\partial s}\right)'_{3\times 1}$$

and

$$\alpha^m_{\underline{2}}(w) = \left(\frac{\partial^2 M(w)}{\partial z^2}, \frac{\partial^2 M(w)}{\partial z \partial x}, \cdots, \frac{\partial^2 M(w)}{\partial s \partial x}, \frac{\partial^2 M(w)}{\partial s^2}\right)'_{6\times 1}$$

Let $\boldsymbol{\alpha}^m(w)$ be the stacked vector of $\alpha^m_{\underline{r}}(w)$ of which $0 \le |\underline{r}| \le p$ for some $p \ge 1$ based on the order that is increasing in $|\underline{r}|$. Wherever it causes no confusion, we will simply write $\alpha_{\underline{r}}$ for

Table 5: Performance of $\hat{\beta}_1(s)$, $c = 0$

| | $s$ | $n$ = 1000 | 2000 | 4000 | 8000 | 16000 |
|---|---|---|---|---|---|---|
| MSE | -1.8 | 0.2623 | 0.1189 | 0.0703 | 0.0354 | 0.0196 |
| | -1 | 0.0131 | 0.0078 | 0.0057 | 0.0038 | 0.0025 |
| | 0 | 0.012 | 0.0068 | 0.0043 | 0.0031 | 0.0019 |
| | 1 | 0.0697 | 0.0108 | 0.0062 | 0.0044 | 0.0032 |
| | 1.8 | 0.2924 | 0.1003 | 0.0583 | 0.0283 | 0.0132 |
| Bias | -1.8 | -0.0297 | -0.0041 | -0.0069 | -0.0034 | -0.001 |
| | -1 | -0.0154 | -0.0007 | -0.001 | -0.0017 | -0.0011 |
| | 0 | -0.0301 | -0.0295 | -0.0203 | -0.016 | -0.0121 |
| | 1 | -0.116 | -0.0015 | 0.0002 | 0.003 | 0.0008 |
| | 1.8 | -0.0977 | -0.0809 | -0.0678 | -0.0387 | -0.0289 |
| Variance | -1.8 | 0.2614 | 0.1189 | 0.0702 | 0.0353 | 0.0196 |
| | -1 | 0.0129 | 0.0078 | 0.0057 | 0.0038 | 0.0025 |
| | 0 | 0.0111 | 0.006 | 0.0039 | 0.0028 | 0.0017 |
| | 1 | 0.0563 | 0.0108 | 0.0062 | 0.0044 | 0.0032 |
| | 1.8 | 0.2829 | 0.0937 | 0.0537 | 0.0268 | 0.0124 |

$\alpha_{\underline{r}}(w)$ and write $\boldsymbol{\alpha}^m$ for $\boldsymbol{\alpha}^m(w)$. Let $\mu^m_{\underline{r}}(w)$ be a vector of polynomials of $w$ with a typically element equals to $w^{\underline{r}}$ for some $0 \le |\underline{r}| \le p$ and all the terms in $\mu^m_{\underline{r}}(w)$ are arranged in the same lexicographical order as above. Let $\mu^m(w)$ be the stacked vector of $\mu^m_{\underline{r}}$ increasing in $|\underline{r}|$. Note that Taylor expansion leads to the approximation that $M(w) \approx \sum_{0\le|\underline{r}|\le p} \frac{1}{\underline{r}!}\alpha^m_{\underline{r}} \cdot \mu^m_{\underline{r}}(w)$, where "$\cdot$" represents the inner product of two vectors.

Let $H_n$ be a diagonal with the same number of rows as the dimension of $\mu^m$, with diagonal entries being $h^{|\underline{r}|}$ for $0 \le |\underline{r}| \le p$ and arranged in the same lexicographical order. Let $W_p$ be another diagonal matrix with diagonal entries be $\underline{r}!$ for $0 \le |\underline{r}| \le p$ and arranged in the same lexicographical order.

Lemmas 1 and 2 adopt the results from Kong, Linton, and Xia (2010, KLX).

**Lemma 1.** *Under Assumptions 1 to 5 and 8 to 10, we have*

$$\sup_{w\in\mathscr{W}} |h\{\hat{m}(w) - m(w)\} - m^*_n(w)| = O_p\left(\frac{\log n}{nh^{d_m}}\right),$$

*where $m^*_n(w)$ is the Bahadur representation of $\hat{m} - m$:*

$$m^*_n(w) = -\frac{1}{nh^{d_m}}\Sigma^{-(s,\cdot)}_{n,m} H^{-1}_n \sum_{i=1}^{n} \boldsymbol{K}_{s,h}(S_i - s)\boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z)$$
$$\times \left[Y_i - \sum_{0\le|\underline{r}|\le p} \frac{1}{|\underline{r}|!}\alpha_{\underline{r}} \cdot \mu^m_{\underline{r}}(W_i - w)\right] \mu^m(W_i - w) \quad \text{(A.1)}$$

Table 6: Performance of $\hat{\beta}_1(s)$, $c = -0.03$

| | | $n$ | | | | |
|---|---|---|---|---|---|---|
| | $s$ | 1000 | 2000 | 4000 | 8000 | 16000 |
| MSE | -1.8 | 0.2835 | 0.1459 | 0.0719 | 0.038 | 0.0194 |
| | -1 | 0.0238 | 0.0144 | 0.0107 | 0.0071 | 0.0051 |
| | 0 | 0.0171 | 0.0108 | 0.008 | 0.0057 | 0.0038 |
| | 1 | 0.0989 | 0.0197 | 0.0135 | 0.0103 | 0.0078 |
| | 1.8 | 0.3096 | 0.1102 | 0.055 | 0.0284 | 0.0145 |
| Bias | -1.8 | 0.0043 | -0.0195 | 0.0107 | 0.0065 | -0.0019 |
| | -1 | -0.0109 | -0.0029 | -0.0041 | 0.0003 | -0.0021 |
| | 0 | -0.0142 | -0.0139 | -0.0116 | -0.0097 | -0.0056 |
| | 1 | -0.158 | -0.0027 | -0.003 | -0.0027 | 0.0019 |
| | 1.8 | -0.0405 | -0.0429 | -0.0194 | -0.0091 | -0.004 |
| Variance | -1.8 | 0.2835 | 0.1455 | 0.0718 | 0.038 | 0.0194 |
| | -1 | 0.0236 | 0.0144 | 0.0107 | 0.0071 | 0.0051 |
| | 0 | 0.0169 | 0.0106 | 0.0078 | 0.0056 | 0.0038 |
| | 1 | 0.074 | 0.0197 | 0.0135 | 0.0103 | 0.0078 |
| | 1.8 | 0.308 | 0.1083 | 0.0546 | 0.0284 | 0.0145 |

*where $\Sigma_{n,m}^{-(s,\cdot)}$ is a row from $\Sigma_{n,m}^{-1}$ which corresponds to the linear term of $S$.*[6] *Furthermore, Let $\epsilon_i^m = Y_i - \mathbb{E}[Y_i|W_i]$ and*

$$\eta_{m,n}(w) = -\frac{1}{nh^{d_m+1}}\Sigma_{n,m}^{-(s,\cdot)}H_n^{-1}\sum_{i=1}^{n}\boldsymbol{K}_{s,h}(S_i - s)\boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z)\epsilon_i^m\mu^m(W_i - w),$$

*then*

$$\sup_{w\in\mathscr{W}}|\hat{m}(w) - m(w) - \eta_{m,n}(w)| \leq O_p\left(\frac{\log n}{nh^{d_m+1}}\right) + O_p(h^p). \tag{A.2}$$

Proof. *Since the loss function is quadratic, to show the first displayed equation, we apply the results stated in KLX (Equation 13, pp1536). We take $\lambda_1 = 1, \lambda_2 = 1/2$ and verify KLX conditions A1-A7. Then Equation* (A.1) *holds by noticing that the partial derivative $m$ that we are estimating corresponds to the fourth element of $\boldsymbol{\alpha}^m$, the fourth diagonal element of $H_n$ is $1/h$ and the fourth diagonal element of $W_p$ is 1.*

*KLX-A1 part 1 holds since we consider the quadratic loss function. KLX-A1 part 2 holds by Assumption 10-(i). KLX-A2 holds again because we consider quadratic loss function, hence the first order derivative is linear. KLX-A3 is the assumption on kernels and it is satisfied by Assumption 8. KLX-A4 is the smoothness assumption on the joint distribution of $(S, X, Z)$, it holds by Assumption 10-(i). KLX-A5 is the smoothness assumption on $m$ and is satisfied by Assumption 10-(ii). KLX-A7 part 1 is ensured by Assumption 10-(ii) and part 2 holds since we have i.i.d. observations.*

*It remains to verify KLX-A6. For two sequences $a_b$ and $b_n$, we use $a_n \succ b_n$ to denote $b_n/a_n \xrightarrow{p} 0$. Analogously define "$\prec$". First, since $nh^{d_m} \succ nh^{d_m+2}$ and $nh^{d_m+2} \to \infty$*

[6] *For example, if all the variables are scalar-valued, then it is the fourth row of $\Sigma_{n,m}^{-1}$.*

Table 7: Coverage frequency for $\beta_1(s)$, Size

| $s$ | $n$ | $c=0$ 90% | $c=0$ 95% | $c=0$ 99% | $c=-0.01$ 90% | $c=-0.01$ 95% | $c=-0.01$ 99% | $c=-0.03$ 90% | $c=-0.03$ 95% | $c=-0.03$ 99% |
|---|---|---|---|---|---|---|---|---|---|---|
| $s=-1.8$ | 1000 | 0.84 | 0.901 | 0.958 | 0.812 | 0.894 | 0.964 | 0.818 | 0.889 | 0.961 |
| | 2000 | 0.863 | 0.931 | 0.979 | 0.853 | 0.921 | 0.976 | 0.843 | 0.916 | 0.98 |
| | 4000 | 0.883 | 0.945 | 0.99 | 0.878 | 0.935 | 0.981 | 0.87 | 0.928 | 0.986 |
| | 8000 | 0.891 | 0.942 | 0.984 | 0.885 | 0.942 | 0.995 | 0.889 | 0.932 | 0.987 |
| | 16000 | 0.884 | 0.945 | 0.984 | 0.891 | 0.953 | 0.987 | 0.915 | 0.946 | 0.985 |
| $s=-1$ | 1000 | 0.885 | 0.942 | 0.98 | 0.884 | 0.934 | 0.983 | 0.896 | 0.934 | 0.981 |
| | 2000 | 0.898 | 0.938 | 0.988 | 0.895 | 0.952 | 0.986 | 0.897 | 0.947 | 0.981 |
| | 4000 | 0.91 | 0.947 | 0.992 | 0.896 | 0.948 | 0.985 | 0.909 | 0.945 | 0.988 |
| | 8000 | 0.899 | 0.942 | 0.988 | 0.898 | 0.94 | 0.99 | 0.913 | 0.951 | 0.994 |
| | 16000 | 0.889 | 0.957 | 0.997 | 0.902 | 0.952 | 0.998 | 0.916 | 0.96 | 0.992 |
| $s=0$ | 1000 | 0.864 | 0.929 | 0.988 | 0.87 | 0.943 | 0.986 | 0.883 | 0.921 | 0.978 |
| | 2000 | 0.865 | 0.925 | 0.983 | 0.898 | 0.927 | 0.984 | 0.881 | 0.94 | 0.99 |
| | 4000 | 0.892 | 0.949 | 0.986 | 0.875 | 0.937 | 0.987 | 0.898 | 0.947 | 0.992 |
| | 8000 | 0.863 | 0.932 | 0.988 | 0.896 | 0.943 | 0.982 | 0.909 | 0.938 | 0.988 |
| | 16000 | 0.882 | 0.935 | 0.99 | 0.9 | 0.942 | 0.991 | 0.917 | 0.948 | 0.987 |
| $s=1$ | 1000 | 0.899 | 0.947 | 0.99 | 0.883 | 0.948 | 0.988 | 0.901 | 0.931 | 0.991 |
| | 2000 | 0.886 | 0.957 | 0.993 | 0.894 | 0.95 | 0.985 | 0.889 | 0.945 | 0.994 |
| | 4000 | 0.914 | 0.945 | 0.992 | 0.902 | 0.947 | 0.991 | 0.902 | 0.95 | 0.99 |
| | 8000 | 0.902 | 0.934 | 0.989 | 0.897 | 0.954 | 0.985 | 0.911 | 0.945 | 0.99 |
| | 16000 | 0.909 | 0.951 | 0.989 | 0.9 | 0.944 | 0.989 | 0.893 | 0.959 | 0.99 |
| $s=1.8$ | 1000 | 0.828 | 0.897 | 0.973 | 0.854 | 0.916 | 0.976 | 0.842 | 0.895 | 0.964 |
| | 2000 | 0.847 | 0.911 | 0.962 | 0.851 | 0.923 | 0.97 | 0.854 | 0.908 | 0.981 |
| | 4000 | 0.876 | 0.922 | 0.983 | 0.881 | 0.926 | 0.982 | 0.88 | 0.932 | 0.983 |
| | 8000 | 0.877 | 0.937 | 0.983 | 0.899 | 0.944 | 0.987 | 0.886 | 0.95 | 0.988 |
| | 16000 | 0.864 | 0.956 | 0.993 | 0.886 | 0.949 | 0.983 | 0.895 | 0.952 | 0.985 |

*by Assumption 9-(i), the first condition in the Equation A.2 of KLX-A6 is satisfied. The second condition in the Equation A.2 holds because we take $\lambda_2 = 1/2$ and the assumption that $nh^{d_m+2(p+1)} \prec nh^{d_\pi+2(p+1)}$ and $nh^{d_\pi+2(p+1)} \to 0$ by Assumption 9-(ii). To verify the third condition in Equation A.2, let $\gamma_n = nh^{d_m}/\log n$, then using KLX's notation in their Equation A.1, we have for some $M > 2$,*

$$d_n = \gamma_n^{-1-\frac{1}{4}+\frac{1}{2}} \log n = \gamma_n^{-\frac{3}{4}} \log n, \quad r(n) = \gamma_n^{\frac{1}{4}}$$

$$M_n^{(1)} = M\gamma_n^{-1}, \quad M_n^{(2)} = M^{\frac{1}{4}}\gamma_n^{-\frac{1}{2}}.$$

*Hence we have*

$$\begin{aligned} n^{-1}\{r(n)\}^{\nu_2/2} d_n \log n \{M_n^{(2)}\}^{-1} &= M^{-1/4}\{\log n\}^2 n^{-1}\gamma_n^{\frac{\nu_2}{8}-\frac{3}{4}+\frac{1}{2}} = M^{-1/4}\{\log n\}^2 n^{-1}\gamma_n^{\frac{\nu_2-2}{8}} \\ &= M^{-1/4}\{\log n\}^{2-\frac{\nu_2-2}{8}} n^{-1}(nh^{d_m})^{\frac{\nu_2-2}{8}}, \end{aligned}$$

*where we can take $\nu_2 \leq \nu_1$ and $\nu_1$ be large enough, then the above quantity diverges to infinity. This is ensured by Assumption 10-(ii). Equation A.3 and A.4 of KLX-A6 are*

Table 8: Testing $H_0 : \beta_1(-0.5) = 0$ vs $H_0 : \beta_1(-0.5) \neq 0$, Power

| | Rejection Frequencies | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $c = 0$ | | | $c = -0.01$ | | | $c = -0.03$ | | |
| $n$ | 10% | 5% | 1% | 10% | 5% | 1% | 10% | 5% | 1% |
| 1000 | 0.878 | 0.923 | 0.806 | 0.850 | 0.915 | 0.775 | 0.739 | 0.827 | 0.630 |
| 2000 | 0.972 | 0.990 | 0.971 | 0.961 | 0.986 | 0.944 | 0.904 | 0.941 | 0.848 |
| 4000 | 1.000 | 1.000 | 0.998 | 0.996 | 1.000 | 0.997 | 0.983 | 0.991 | 0.971 |
| 8000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.999 | 0.997 |
| 16000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

*satisfied since by the i.i.d. observation Assumption 4, the mixing coefficient $\gamma[k] = 0$ for all $k \geq 1$.*

*Next we verify Equation* (A.2) *in the statement of this Lemma. By definition of $\eta_{m,n}$, we can write $m_n^*(w)$ as*

$$m_n^*(w) = h\eta_{m,n}(w) + \frac{1}{nh^{d_m}} \Sigma_{n,m}^{-(s,\cdot)} H_n^{-1} \sum_{i=1}^{n} \boldsymbol{K}_{s,h}(S_i - s)\boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z) \times \left[ \mathbb{E}[Y_i|W_i] - \sum_{0 \leq |\underline{r}| \leq p} \frac{1}{|\underline{r}|!} \alpha_{\underline{r}} \cdot \mu_{\underline{r}}^m (W_i - w) \right] \mu^m(W_i - w).$$

*To show the second RHS term is of order $O_p(h^{p+1})$, it is sufficient to show that the following term is of order $O(h^{p+1})$ uniformly in $w$:*

$$e_n \equiv \frac{1}{h^{d_m}} \mathbb{E} \left| \boldsymbol{K}_{s,h}(S_i - s)\boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z)\tilde{\mu}^m(W_i - w) \left[ \mathbb{E}[Y_i|W_i] - \sum_{0 \leq |\underline{r}| \leq p} \frac{1}{|\underline{r}|!} \alpha_{\underline{r}} \cdot \mu_{\underline{r}}^m (W_i - w) \right] \right|,$$

*where $\tilde{\mu}^m(W_i - w) = \Sigma_{n,m}^{-(4,\cdot)}(w) H_n^{-1} \mu^m(W_i - w)$. For a generic vector $C_i$, let $u_c = \frac{C_i - c}{h}$.*

*Conducting changing variable we have*

$$e_n = \int |\boldsymbol{K}_s(u_s)\boldsymbol{K}_x(u_x)\boldsymbol{K}_z(u_z)\tilde{\mu}^m(hu_w)| \times \left| M(w + hu_w) - \sum_{0 \leq |\underline{r}| \leq p} \frac{1}{|\underline{r}|!} \alpha_{\underline{r}} \cdot \mu_{\underline{r}}^m (hu_w) \right| g_w(hu_w + w) du_w,$$

*where $M(\cdot) = \mathbb{E}[Y|W = \cdot]$ and $g_w$ is the density of $W$. $e_n$ is of order $O(h^{p+1})$ since $M$ is $p + 1$ times continuously differentiable, $g_w$ is uniformly bounded and the kernel function is bounded with finite support. Hence we have uniformly over $w$*

$$|h^{-1} m_n^*(w) - \eta_{m,n}(w)| = O_p(h^p).$$

Table 9: Coverage frequency for $\beta_1(s)$, Size

| | | $c=0$ | | | $c=-0.01$ | | | $c=-0.03$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $s$ | $n$ | 90% | 95% | 99% | 90% | 95% | 99% | 90% | 95% | 99% |
| | 1000 | 0.84 | 0.901 | 0.958 | 0.812 | 0.894 | 0.964 | 0.818 | 0.889 | 0.961 |
| | 2000 | 0.863 | 0.931 | 0.979 | 0.853 | 0.921 | 0.976 | 0.843 | 0.916 | 0.98 |
| $s=-1.8$ | 4000 | 0.883 | 0.945 | 0.99 | 0.878 | 0.935 | 0.981 | 0.87 | 0.928 | 0.986 |
| | 8000 | 0.891 | 0.942 | 0.984 | 0.885 | 0.942 | 0.995 | 0.889 | 0.932 | 0.987 |
| | 16000 | 0.884 | 0.945 | 0.984 | 0.891 | 0.953 | 0.987 | 0.915 | 0.946 | 0.985 |
| | 1000 | 0.885 | 0.942 | 0.98 | 0.884 | 0.934 | 0.983 | 0.896 | 0.934 | 0.981 |
| | 2000 | 0.898 | 0.938 | 0.988 | 0.895 | 0.952 | 0.986 | 0.897 | 0.947 | 0.981 |
| $s=-1$ | 4000 | 0.91 | 0.947 | 0.992 | 0.896 | 0.948 | 0.985 | 0.909 | 0.945 | 0.988 |
| | 8000 | 0.899 | 0.942 | 0.988 | 0.898 | 0.94 | 0.99 | 0.913 | 0.951 | 0.994 |
| | 16000 | 0.889 | 0.957 | 0.997 | 0.902 | 0.952 | 0.998 | 0.916 | 0.96 | 0.992 |
| | 1000 | 0.864 | 0.929 | 0.988 | 0.87 | 0.943 | 0.986 | 0.883 | 0.921 | 0.978 |
| | 2000 | 0.865 | 0.925 | 0.983 | 0.898 | 0.927 | 0.984 | 0.881 | 0.94 | 0.99 |
| $s=0$ | 4000 | 0.892 | 0.949 | 0.986 | 0.875 | 0.937 | 0.987 | 0.898 | 0.947 | 0.992 |
| | 8000 | 0.863 | 0.932 | 0.988 | 0.896 | 0.943 | 0.982 | 0.909 | 0.938 | 0.988 |
| | 16000 | 0.882 | 0.935 | 0.99 | 0.9 | 0.942 | 0.991 | 0.917 | 0.948 | 0.987 |
| | 1000 | 0.899 | 0.947 | 0.99 | 0.883 | 0.948 | 0.988 | 0.901 | 0.931 | 0.991 |
| | 2000 | 0.886 | 0.957 | 0.993 | 0.894 | 0.95 | 0.985 | 0.889 | 0.945 | 0.994 |
| $s=1$ | 4000 | 0.914 | 0.945 | 0.992 | 0.902 | 0.947 | 0.991 | 0.902 | 0.95 | 0.99 |
| | 8000 | 0.902 | 0.934 | 0.989 | 0.897 | 0.954 | 0.985 | 0.911 | 0.945 | 0.99 |
| | 16000 | 0.909 | 0.951 | 0.989 | 0.9 | 0.944 | 0.989 | 0.893 | 0.959 | 0.99 |
| | 1000 | 0.828 | 0.897 | 0.973 | 0.854 | 0.916 | 0.976 | 0.842 | 0.895 | 0.964 |
| | 2000 | 0.847 | 0.911 | 0.962 | 0.851 | 0.923 | 0.97 | 0.854 | 0.908 | 0.981 |
| $s=1.8$ | 4000 | 0.876 | 0.922 | 0.983 | 0.881 | 0.926 | 0.982 | 0.88 | 0.932 | 0.983 |
| | 8000 | 0.877 | 0.937 | 0.983 | 0.899 | 0.944 | 0.987 | 0.886 | 0.95 | 0.988 |
| | 16000 | 0.864 | 0.956 | 0.993 | 0.886 | 0.949 | 0.983 | 0.895 | 0.952 | 0.985 |

*Therefore, by triangular inequality,*

$$\sup_{w\in\mathscr{W}} |\{\hat{m}(w)-m(w)\}-\eta_{m,n}(w)| \leq \sup_{w\in\mathscr{W}} |\hat{m}(w)-m(w)-h^{-1}m_n^*(w)| + \sup_{w\in\mathscr{W}} |h^{-1}m_n^*(w)-\eta_{m,n}(w)| = O_p\left(\frac{\log n}{nh^{d_m+1}}\right)+O_p(h^p),$$

*which establishes Equation* (A.2). □

Let $\tilde{W}=(X,Z)$ and $\tilde{\mathscr{W}}$ be its support. Again, to simplify notation, we use $\pi$ to denote a generic element from the vector $\boldsymbol{\pi}_0=[\pi_1,...,\pi_T]'$. We define other notation in a similar way as we define them for estimation $m$, with $\pi$ replacing $m$. For example, $\Sigma_{n,\pi}$ and $\Sigma_\pi$ are two matrices defined analogously to $\Sigma_{n,m}$ and $\Sigma_m$, with matrix dimension adjusted accordingly.

**Lemma 2.** *Suppose that Assumptions 1 to 5 and 8 to 10 hold, then uniformly over* $\tilde{\mathscr{W}}$,

$$\hat{\pi}(x,z)-\pi(x,z)-\pi_n^*(x,z)=O_p\left(\frac{\log n}{nh^{d_\pi}}\right),$$

Table 10: Testing $H_0 : \beta_1(-0.5) = 0$ vs $H_0 : \beta_1(-0.5) \neq 0$, Power

| | Rejection Frequencies | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $c = 0$ | | | $c = -0.01$ | | | $c = -0.03$ | | |
| $n$ | 10% | 5% | 1% | 10% | 5% | 1% | 10% | 5% | 1% |
| 1000 | 0.878 | 0.923 | 0.806 | 0.850 | 0.915 | 0.775 | 0.739 | 0.827 | 0.630 |
| 2000 | 0.972 | 0.990 | 0.971 | 0.961 | 0.986 | 0.944 | 0.904 | 0.941 | 0.848 |
| 4000 | 1.000 | 1.000 | 0.998 | 0.996 | 1.000 | 0.997 | 0.983 | 0.991 | 0.971 |
| 8000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.999 | 0.997 |
| 16000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

*where $\pi_n^*$ is the Bahadur representation of $\hat{\pi} - \pi$:*

$$\pi_n^*(x,z) = -\frac{1}{nh^{d_\pi}} \Sigma_{n,\pi}^{-(1,\cdot)} H_n^{-1} \sum_{i=1}^n \boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z) \times \left[ 1\{D_i = d\} - \sum_{0\leq|\underline{r}|\leq p} \frac{1}{|\underline{r}|!}\alpha_{\underline{r}} \cdot \mu^\pi(\widetilde{W}_i - \widetilde{w}) \right] \mu^\pi(\tilde{W}_i - \tilde{w}), \quad \text{(A.3)}$$

*where $\Sigma_{n,\pi}^{-(1,\cdot)}$ is the first row $\Sigma_{n,\pi}^{-1}$. Furthermore, let $\epsilon_i^\pi = \boldsymbol{1}\{D_i = d\} - \Pr(D_i = d|X_i, Z_i)$ and let*

$$\eta_{\pi,n}(x,z) = -\frac{1}{nh^{d_\pi}} \Sigma_{n,\pi}^{-(1,\cdot)} H_n^{-1} \sum_{i=1}^n \boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{z,h}(Z_i - z)\epsilon_i^\pi \mu^\pi(\tilde{W}_i - \tilde{w}),$$

*then*

$$\sup_{(x,z)\in\tilde{\mathscr{W}}} |\hat{\pi}(x,z) - \pi(x,z) - \eta_{n,\pi}(x,z)| \leq O_p\left(\frac{\log n}{nh^{d_\pi}}\right) + O_p(h^{p+2})). \quad \text{(A.4)}$$

Proof. *We verify KLX-A6; the other assumptions (A1-A5 and A7) and the rest of the proof can be verified analogously to Lemma 1. Since $nh^{d_\pi} \succ nh^{d_m+2} \to \infty$ at polynomial rate, the first condition in the Equation A.2 of KLX-A6 is satisfied. The second condition in the Equation A.2 holds because we take $\lambda_2 = 1/2$ and the assumption that $nh^{d_\pi+2(p+1)} \to 0$ by Assumption 9-ii. To verify the third condition in the Equation A.2, let $\gamma_n = nh^{d_\pi}/\log n$, as before, we have for some $M > 2$,*

$$d_n = \gamma_n^{-1-\frac{1}{4}+\frac{1}{2}} \log n = \gamma_n^{-\frac{3}{4}} \log n, \quad r(n) = \gamma_n^{\frac{1}{4}}$$

$$M_n^{(1)} = M\gamma_n^{-1}, \quad M_n^{(2)} = M^{\frac{1}{4}}\gamma_n^{-\frac{1}{2}}$$

*Hence we have*

$$n^{-1}\{r(n)\}^{\nu_2/2} d_n \log n \{M_n^{(2)}\}^{-1} = M^{-1/4}\{\log n\}^2 n^{-1}\gamma_n^{\frac{\nu_2}{8}-\frac{3}{4}+\frac{1}{2}} = M^{-1/4}\{\log n\}^2 n^{-1}\gamma_n^{\frac{\nu_2-2}{8}} = M^{-1/4}\{\log n\}^{2-\frac{\nu_2-2}{8}} n^{-1}(nh^{d_\pi})^{\frac{\nu_2-2}{8}}.$$

*Note that $\mathbb{E}|\epsilon_i^\pi|^{\nu_1} < \infty$ for any $\nu_1$ since $\epsilon_i^\pi$ is bounded; then the above quantity diverges to* *infinity by letting $\nu_2$ arbitrarily large. Finally, note that the bias is of order $h^{p+2}$ by KLX Proposition 2.* $\square$

**Lemma 3.** *Let $g_z$ be the density of $Z$ and $g_w$ be the density of $W$. Let $\eta_{m,n}$ and $\eta_{\pi,n}$ be as defined in Equations (1) and (2), respectively. Suppose that the assumptions of Proposition 1 are satisfied, then*

$$\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\eta_{m,n}(s,x,Z_i) \xrightarrow{d} N(0,\Omega_m).$$

*where $\Omega_m$ is a $T\times T$ positive definite matrix*

$$\Omega_m(s,x) = \int \left\{\sigma_m^2(x,s,Z_1)\left(\int \Gamma\left(Z_1;u_s,u_x,-u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2 \right. \\ \left. \times \boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)g_z^2(Z_1)\boldsymbol{\pi}(x,Z_1)\boldsymbol{\pi}'(x,Z_1)\right\} g_w(s,x,Z_1)du_x du_s dZ_1, \quad \text{(A.5)}$$

*$\sigma_m^2(w) = \mathbb{V}[\epsilon^m|W=w]$ and $\Gamma(\cdot)$ is defined in Equation (A.6).*

*For a generic element $\pi$ in $\boldsymbol{\pi}_0 = [\pi_1,...,\pi_T]'$, there is*

$$\frac{\sqrt{nh^{d_\pi-d_z}}}{n}\sum_i m(s,x,Z_i)\eta_{\pi,n}(x,Z_i) \xrightarrow{d} N(0,\Omega_\pi),$$

*$\Omega_\pi$ is a positive scalar such that*

$$\Omega_\pi \equiv \int \left\{\sigma_\pi^2(x,Z_1)\left(\int \Psi(Z_1;u_x,-u_z)\boldsymbol{K}_z(u_z)du_z\right)^2 K^2(u_x)g_z^2(Z_1)m^2(s,x,Z_1)\right\} g_w(s,x,Z_1)du_x dZ_1.$$

*and $\sigma_\pi^2(x,z) = \mathbb{V}[\epsilon^\pi|(X,Z)=(x,z)]$ and $\Psi(\cdot)$ is defined in Equation (A.9).*

Proof. *Recall that*

$$\eta_{m,n}(w) = -\frac{1}{nh^{d_m+1}}\Sigma_{n,m}^{-(s,\cdot)}H_n^{-1}\sum_{i=1}^n \boldsymbol{K}_{s,h}(S_i-s)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{z,h}(Z_i-z)\epsilon_i^m\mu^m(W_i-w),$$

*where $\epsilon_i^m = Y_i - \mathbb{E}[Y_i|W_i]$. Let $d_z$ be the dimension of $Z$, then*

$$\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\eta_{m,n}(s,x,Z_i) \\ = -\frac{1}{n\sqrt{h^{d_z}}\sqrt{nh^{d_m}}}\sum_i\sum_{j\neq i}\boldsymbol{\pi}(x,Z_i)\Sigma_{n,m}^{-(s,\cdot)}(Z_i)\boldsymbol{K}_{z,h}(Z_j-Z_i)\boldsymbol{K}_{x,h}(X_j-x)\boldsymbol{K}_{s,h}(S_j-s)\epsilon_j^m H_n^{-1}\mu^m(W_j-(s,x,Z_i)') \\ \underbrace{-\frac{\boldsymbol{K}_z(0)}{n\sqrt{h^{d_z}}\sqrt{nh^{d_m}}}\sum_i \boldsymbol{\pi}(x,Z_i)\Sigma_{n,m}^{-(s,\cdot)}(Z_i)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{s,h}(S_i-s)\epsilon_i^m H_n^{-1}\mu^m(W_i-(s,x,Z_i)')}_{B1}.$$

*where we abbreviate $\Sigma_{n,m}^{-(s,\cdot)}(s,x,Z_i)$ as $\Sigma_{n,m}^{-(s,\cdot)}(Z_i)$. Since the second term $B1$ is asymptoti-*

*cally negligible, we only focus on the first term. Note that the vector*

$$H_n^{-1}\mu^m(W_j-(s,x,Z_i)')=\left(1,\frac{Z_j-Z_i}{h},\frac{X_j-x}{h},\frac{S_j-s}{h},\left(\frac{Z_j-Z_i}{h}\right)^2,\cdots,\left(\frac{S_j-s}{h}\right)^p\right)'.$$

*then we can write in short hand*

$$\Sigma_{n,m}^{-(s,\cdot)}(Z_i)H_n^{-1}\mu^m(W_j-(s,x,Z_i)')=\Gamma_n\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right).$$

*Since* $\Sigma_{n,m}^{-(s,\cdot)}(w)$ *converges uniformly to* $\Sigma_m^{-(s,\cdot)}(w)$*, it follows that*

$$\begin{aligned}\Sigma_m^{-(s,\cdot)}(Z_i)H_n^{-1}\mu^m(W_j-(s,x,Z_i)')&\equiv\Gamma\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)\\&=\Gamma_n\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)+o(1).\end{aligned}\quad\text{(A.6)}$$

*By defining*

$$\psi_{ij}^*=\boldsymbol{\pi}(x,Z_i)\boldsymbol{K}_{z,h}(Z_j-Z_i)\boldsymbol{K}_{x,h}(X_j-x)\boldsymbol{K}_{s,h}(S_j-s)\epsilon_j^m\Gamma\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right),$$

*and* $\psi_{ij}=\frac{1}{2}(\psi_{ij}^*+\psi_{ji}^*)$*, we can write*

$$\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i\boldsymbol{\pi}(x,Z_i)\eta_{m,n}(Z_i)\overset{d}{\approx}-\frac{\sqrt{n}}{n^2}\sum_i\sum_{j\neq i}\frac{\psi_{ij}}{\sqrt{h^{d_z}}\sqrt{h^{d_m}}},$$

*So we can approximate the objective of analysis by a U-statistics.*

*It is easy to verify that* $\mathbb{E}[\psi_{ij}]=0$ *since* $\mathbb{E}[\epsilon_j^m|W_i]=\mathbb{E}[\epsilon_j^m|W_j]=0$*. To derive the limiting distribution, it remains to find the variance. Let* $\bar{\psi}_1=\mathbb{E}[\psi_{12}|W_1,Y_1]=\frac{1}{2}\{\mathbb{E}[\psi_{12}^*|W_1,Y_1]+\mathbb{E}[\psi_{21}^*|W_1,Y_1]\}$*; based on the standard U-statistics asymptotic result, the limiting variance is* $4\mathbb{V}(\bar{\psi}_1)/(h^{d_z+d_m})$.

*By law of iterated expectation, i.i.d. observation assumption and* $\mathbb{E}[\epsilon_2^m|W_2]=0$*, we have*

$$\mathbb{E}[\psi_{12}^*|W_1,Y_1]=\mathbb{E}[\mathbb{E}[\psi_{12}^*|W_2,W_1,Y_1]|W_1,Y_1]=0.$$

*Therefore,*

$$
\begin{aligned}
&2\bar{\psi}_1 = \mathbb{E}[\psi^*_{21}|W_1, Y_1] \\
&= \mathbb{E}\left[\boldsymbol{\pi}(x, Z_2)\boldsymbol{K}_{z,h}(Z_2 - Z_1)\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\epsilon_1^m\Gamma\left(Z_2; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, \frac{Z_1 - Z_2}{h}\right)|W_1, Y_1\right] \\
&= \epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\mathbb{E}\left[\boldsymbol{\pi}(x, Z_2)\Gamma\left(Z_2; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, \frac{Z_1 - Z_2}{h}\right)\boldsymbol{K}_{z,h}(Z_2 - Z_1)|W_1, Y_1\right] \\
&\overset{(i)}{=} \epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\int \boldsymbol{\pi}(x, Z_2)\Gamma\left(Z_2; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, \frac{Z_1 - Z_2}{h}\right)\boldsymbol{K}_{z,h}(Z_2 - Z_1)g_z(Z_2)dZ_2 \\
&\overset{(ii)}{=} h^{d_z}\epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\int \boldsymbol{\pi}(x, Z_1 + hu_z)\Gamma\left(Z_1 + hu_z; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, -u_z\right)\boldsymbol{K}_z(u_z)g_z(Z_1 + hu_z)du_z \\
&\overset{(iii)}{=} h^{d_z}\epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\left\{\boldsymbol{\pi}(x, Z_1)g_z(Z_1)\int \Gamma\left(Z_1; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, -u_z\right)\boldsymbol{K}_z(u_z)du_z + o(h)\right\},
\end{aligned}
$$

*where (i) holds because i.i.d. observations; (ii) holds by changing variable $u_z = (Z_2 - Z_1)/h$, and (iii) holds by the continuous differentiability of the integrand (implied by Assumption 10) and the assumption that the support of the kernel is bounded. So the dominant term of $2\bar{\psi}_1$ is*

$$
h^{d_z}\epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\boldsymbol{\pi}(x, Z_1)g_z(Z_1)\int \Gamma\left(Z_1; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, -u_z\right)\boldsymbol{K}_z(u_z)du_z.
$$

*Since $\mathbb{E}[\bar{\psi}_1] = 0$, then up to the negligible terms, we have*

$$
\begin{aligned}
&4\mathbb{V}(\bar{\psi}_1) = 4\mathbb{E}[\bar{\psi}_1\bar{\psi}_1'] = h^{2d_z}\mathbb{E}\left[\left(\epsilon_1^m\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\boldsymbol{g}_z(Z_1)\int \Gamma\left(Z_1; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, -u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)\right] + o(1) \\
&\overset{(i)}{=} h^{2d_z}\mathbb{E}\left[\sigma_m^2(W_1)\left(\boldsymbol{K}_{x,h}(X_1 - x)\boldsymbol{K}_{s,h}(S_1 - s)\boldsymbol{g}_z(Z_1)\int \Gamma\left(Z_1; \frac{S_1 - s}{h}, \frac{X_1 - x}{h}, -u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)\right] + o(1) \\
&\overset{(ii)}{=} \left(h^{2d_z + d_x + d_s}\right)\int\left\{\sigma_m^2(x, s, Z_1)\boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)g_z^2(Z_1)\left(\int \Gamma\left(Z_1; u_s, u_x, -u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)\right\}g_w(s, x, Z_1)du_xdu_sdZ_1 + o(1) \\
&\overset{(iii)}{=} \left(h^{d_z + d_m}\right)\int\left\{\sigma_m^2(x, s, Z_1)\boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)g_z^2(Z_1)\left(\int \Gamma\left(Z_1; u_s, u_x, -u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)\right\}g_w(s, x, Z_1)du_xdu_sdZ_1 + o(1)
\end{aligned} \tag{A.7}
$$

*where (i) holds by taking the conditional expectation of $(\epsilon_1^m)^2$ given $W_1$; (ii) holds by changing variable $u_x = (X_1 - x)/h$, $u_s = (S_1 - s)/h$ and ingoing higher order terms, and (iii) holds because $d_m = d_s + d_z + d_x$.*

*Then we know that*

$$
4\mathbb{V}\left(\frac{\bar{\psi}_1}{\sqrt{h^{d_m + d_z}}}\right) = \Omega_m(s, x), \tag{A.8}
$$

*where*

$$
\begin{aligned}
\Omega_m(s, x) = \int\Bigg\{&\sigma_m^2(s, x, Z_1)\left(\int \Gamma\left(Z_1; u_s, u_x, -u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2 \\
&\times\boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)g_z^2(Z_1)\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)\Bigg\}g_w(s, x, Z_1)du_xdu_sdZ_1,
\end{aligned}
$$

*where the $\Gamma$ term is defined in Equation* (A.6). *By the standard U statistics theory, we have* *(by abbreviating $\Omega_m(s,x)$ as $\Omega_m$)*

$$\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\eta_{m,n}(Z_i) \xrightarrow{d} N(0,\Omega_m).$$

*Following similar argument, we can show that*

$$\frac{\sqrt{nh^{d_\pi-d_z}}}{n}\sum_i m(x,s,Z_i)\eta_{\pi,n}(Z_i) \xrightarrow{d} N(0,\Omega_\pi),$$

*where*

$$\Omega_\pi \equiv \int \left\{\sigma_\pi^2(x,Z_1)\Psi^2(Z_1;u_x,-u_z)\boldsymbol{K}_x^2(u_x)g_z^2(Z_1)m^2(s,x,Z_1)\right\} g_w(s,x,Z_1)du_x dZ_1.$$

*where the shorthand term $\Psi$ is defined such that*

$$\begin{aligned}&\Sigma_\pi^{-(\cdot)}(Z_i)H_n^{-1}\mu^\pi((X_j,Z_j)-(x,Z_i)') \equiv \Psi\left(Z_i;\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)\\ &= \Psi_n\left(Z_i;\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)+o(1) = \Sigma_{n,\pi}^{-(\cdot)}(Z_i)H_n^{-1}\mu^\pi((X_j,Z_j)-(x,Z_i)')+o(1) \qquad \text{(A.9)}\end{aligned}$$

*Since $d_m+2>d_\pi$, it follows that $\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i m(x,s,Z_i)\eta_{\pi,n} \xrightarrow{p} 0$. For the same reason, the asymptotic covariance between $\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i \boldsymbol{\pi}(x,Z_i)$ and $\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i m(x,s,Z_i)\eta_{\pi,n}$ converges in probability to zero as well. This establishes the result.* □

**Lemma 4.** *Let $\kappa_n = \sqrt{nh^{d_m+2-d_z}}$. Suppose that the assumptions of Proposition 1 are satisfied, then*

$$\frac{1}{n}\sum_i m(s,x,Z_i)\{r_{\pi,1}+r_{\pi,2}\} = o_p(\kappa_n^{-1}), \quad \frac{1}{n}\sum_i \boldsymbol{\pi}(x,Z_i)\{r_{m,1}+r_{m,2}\} = o_p(\kappa_n^{-1}).$$

Proof. *The first equality holds because $\frac{1}{n}\sum_i m(s,x,Z_i) = O_p(1/\sqrt{n}) = o(\kappa_n^{-1})$, and the fact that $r_{\pi,1}+r_{\pi,2} = o_p(1)$ and does not depend on $i$. The second equality holds analogously.* □

**Lemma 5.** *Suppose that the assumptions of Proposition 1 are satisfied and let $\eta_{m,n}(w)$ and $\eta_{\pi,n}(x,z)$ be as defined in Equations (1) and (2), then for a generic element $\pi \in \boldsymbol{\pi}_0$, then there exists $\lambda_m$ and $\lambda_\pi$ such that*

$$\frac{\sqrt{nh^{d_m+2-d_z}}}{n}\sum_i \eta_{m,n}(s,x,Z_i) \xrightarrow{d} N(0,\lambda_m).; \quad \frac{\sqrt{nh^{d_\pi-d_z}}}{n}\sum_i \eta_{\pi,n}(x,Z_i) \xrightarrow{d} N(0,\lambda_\pi).$$

Proof. *It follows from the same argument as in Lemma 3 by replacing $\boldsymbol{\pi}(x,Z_i)$ and $m(x,s,Z_i)$ with 1, respectively.* □

**Lemma 6.** *Let $\kappa_n = \sqrt{nh^{d_m+2-d_z}}$. Suppose that the assumptions of Proposition 1 are satisfied, then $(r_{\pi,1}+r_{\pi,2})(r_{m,1}+r_{m,2}) = o(\kappa_n^{-1})$.*

Proof. *Note that*

$$\kappa_n(r_{\pi,1}+r_{\pi,2})(r_{m,1}+r_{m,2}) = \kappa_n \left(O_p\left(h^p\right) + O_p\left(\frac{\log n}{nh^{d_m+1}}\right)\right) \times \left(O_p\left(h^{p+2}\right) + O_p\left(\frac{\log n}{nh^{d_\pi}}\right)\right)$$
$$= \kappa_n O_p(h^{2p+2}) + \kappa_n O_p\left(\frac{h^{p+2}\log n}{nh^{d_m+1}}\right) + \kappa_n O_p\left(\frac{h^p \log n}{nh^{d_\pi}}\right) + \kappa_n O_p\left(\frac{\log n}{n^2 h^{d_m+1+d_\pi}}\right).$$

*The first right hand side term is* $\sqrt{nh^{d_m+6-d_z+4p}} = \sqrt{nh^{d_\pi+2+2p}}\sqrt{h^{2p+d_s-d_z}} \xrightarrow{p} 0$ *by Assumption 9-(ii) and Assumption 10-(iii). The second RHS term is of order* $\log n\sqrt{\frac{h^{2p+4}nh^{d_m+2-d_z}}{n^2h^{2d_m+2}}} = \log n\sqrt{\frac{h^{2p+4}}{nh^{d_m+d_z}}} \prec \log n\sqrt{\frac{h^{d_z-d_s}}{nh^{d_m+d_z}}} = \log n\sqrt{\frac{1}{nh^{d_m+d_s}}} \xrightarrow{p} 0$*, where* $x \prec y$ *denote* $x/y \xrightarrow{p} 0$*. The third RHS term is of order* $\log n\sqrt{\frac{h^{2p}nh^{d_m+2-d_z}}{n^2h^{2d_\pi}}} = \log n\sqrt{\frac{h^{2p}h^{d_s+2}}{nh^{d_\pi+d_z}}} \xrightarrow{p} 0$ *by Assumption 9-(iii). For the fourth RHS term, it is of order* $(\log n)^2\sqrt{\frac{nh^{d_m+2-d_z}}{n^4h^{2d_m+2+2d_\pi}}} = (\log n)^2\sqrt{\frac{1}{n^3h^{d_m+2d_\pi+d_z}}} = (\log n)^2\sqrt{\frac{1}{nh^{d_\pi+d_z}nh^{d_\pi}nh^{d_m}}} \xrightarrow{p} 0$ *by Assumption 9-(i) and (ii).*

**Lemma 7.** *Let* $\kappa_n = \sqrt{nh^{d_m+2-d_z}}$*. Suppose that the assumptions of Proposition 1 are satisfied, then*

$$T_n \equiv \frac{1}{n}\sum_i \{\eta_{m,n}(s,x,Z_i) + r_{m,1} + r_{m,2}\}\{\eta_{\pi,n}(x,Z_i) + r_{\pi,1} + r_{\pi,2}\} = o_p\left(\kappa_n^{-1}\right).$$

Proof. $T_n$ *can be decomposed as the following four terms,*

$$T_n = \frac{1}{n}\sum_i \eta_{m,n}(s,x,Z_i)\eta_{\pi,n}(x,Z_i) + (r_{\pi,1}+r_{\pi,2})\frac{1}{n}\sum_i \eta_{m,n}(s,x,Z_i)$$
$$+ (r_{m,1}+r_{m,2})\frac{1}{n}\sum_i \eta_{\pi,n}(x,Z_i) + (r_{\pi,1}+r_{\pi,2})(r_{m,1}+r_{m,2}). \quad \text{(A.10)}$$

*The RHS4 is dealt with by Lemma 6. The RHS2 and RHS3 are of order* $o_p\left(\kappa_n^{-1}\right)$ *by Lemma 5 and the fact that the* $r$ *terms converge to zero (in probability). It remains to verify RHS1 of Equation* (A.10) *is also of order* $o_p(\kappa_n^{-1})$*. Let*

$$U_n \equiv \frac{1}{n}\sum_i \eta_{m,n}(s,x,Z_i)\eta_{\pi,n}(x,Z_i)$$
$$= \frac{1}{n^3h^{d_m+1}h^{d_\pi}} \times \sum_i \left(\sum_j \boldsymbol{K}_{z,h}(Z_j-Z_i)\boldsymbol{K}_{x,h}(X_j-x)\boldsymbol{K}_{s,h}(S_j-s)\epsilon_j^m\Gamma\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)\right)$$
$$\times \left(\sum_t \boldsymbol{K}_{z,h}(Z_t-Z_i)\boldsymbol{K}_{x,h}(X_t-x)\epsilon_t^\pi\Psi\left(Z_i;\frac{X_t-x}{h},\frac{Z_t-Z_i}{h}\right)\right)$$
$$= \frac{1}{n^3h^{d_m+1}h^{d_\pi}}\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^n \xi_{ijt}^*, \quad \text{(A.11)}$$

*where*

$$\xi^*_{ijt} = \Gamma\left(Z_i; \frac{S_j - s}{h}, \frac{X_j - x}{h}, \frac{Z_j - Z_i}{h}\right)\Psi\left(Z_i; \frac{X_t - x}{h}, \frac{Z_t - Z_i}{h}\right)$$
$$\times \boldsymbol{K}_{s,h}(S_j - s)\boldsymbol{K}_{x,h}(X_j - x)\boldsymbol{K}_{x,h}(X_t - x)\boldsymbol{K}_{z,h}(Z_j - Z_i)\boldsymbol{K}_{z,h}(Z_t - Z_i)\epsilon^m_j\epsilon^\pi_t,$$

*and* $\Psi\left(Z_i; \frac{X_t-x}{h}, \frac{Z_t-Z_i}{h}\right)$ *is defined analogously as* $\Gamma\left(Z_i; \frac{S_j-s}{h}, \frac{X_j-x}{h}, \frac{Z_j-Z_i}{h}\right)$. *We decompose* $U_n$ *into five parts:*

$$U_{1n} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j\neq i}^n \sum_{t\neq j, t\neq i}^n \xi^*_{ijt},$$

$$U_{2n} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j\neq i}^n \sum_{t=j}^n \xi^*_{ijt} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j\neq i}^n \xi^*_{ijj},$$

$$U_{3n} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j=i}^n \sum_{t\neq j}^n \xi^*_{ijt} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{t\neq i}^n \xi^*_{iit},$$

$$U_{4n} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j\neq i}^n \sum_{t=i}^n \xi^*_{ijt} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i=1}^n \sum_{j\neq i}^n \xi^*_{iji},$$

$$U_{5n} = \frac{1}{n^3 h^{d_m+1} h^{d_\pi}} \times \sum_{i}^n \xi^*_{iii},$$

*so* $U_n = U_{1n} + U_{2n} + U_{3n} + U_{4n} + U_{5n}$. *We will show that all these terms are asymptotically negligible.*

***Part 1:*** $\boldsymbol{U_{1n}}$. *We write*

$$U_{1n} = \frac{n(n-1)(n-2)\sqrt{h^{d_m+d_\pi}}}{n^3 h^{d_m+1} h^{d_\pi}} \times \underbrace{\frac{1}{n(n-1)(n-2)} \times \sum_{i=1}^n \sum_{j\neq i}^n \sum_{t\neq j, t\neq i}^n \psi^*_{ijt}}_{\tilde{U}_{1n}},$$

*where* $\psi^*_{ijt} = \frac{\xi^*_{ijt}}{\sqrt{h^{d_m+d_\pi}}}$. *The* $U_{1n}$ *term is proportional to a third order U-statistics with kernel function* $\psi^*_{ijt}$. *Let* $\psi_{ijt}$ *be a symmetric transformation of* $\psi^*_{ijt}$, *that is,* $\psi_{ijt} = \frac{1}{6}\sum_p \psi^*_{ijt}$ *where* $\sum_p$ *is the sum over all permutations of* $i, j, t$. *Write* $H_i = (W_i, Y_i, D_i)$. *It is staightforward to calculate that* $\mathbb{E}[\psi^*_{123}|H_1] = \mathbb{E}[\psi^*_{123}|H_2] = \mathbb{E}[\psi^*_{123}|H_3] = 0$, *which implies that* $\mathbb{E}[\psi_{ijt}|H_1] = 0$ *as well as* $\mathbb{E}[U_{1n}] = 0$. *Hence* $\tilde{U}_{1n}$ *is a degenerated U-statistics. In the mean time,* $\mathbb{E}[\psi^2_{123}]$

*is of the same order of* $\mathbb{E}[(\psi^*_{123})^2]$, *which is*

$$\mathbb{E}[(\psi^*_{123})^2] = \frac{1}{h^{d_m+d_\pi}}\mathbb{E}[(\xi^*_{123})^2]$$
$$= \frac{1}{h^{d_m+d_\pi}}\int \Gamma^2\left(Z_1;\frac{S_2-s}{h},\frac{X_2-x}{h},\frac{Z_2-Z_1}{h}\right)\Psi^2\left(Z_1;\frac{X_3-x}{h},\frac{Z_3-Z_1}{h}\right)\boldsymbol{K}^2_{s,h}(S_2-s)\boldsymbol{K}^2_{x,h}(X_2-x)$$
$$\times\boldsymbol{K}^2_{x,h}(X_3-x)\boldsymbol{K}^2_{z,h}(Z_2-Z_1)\boldsymbol{K}^2_{z,h}(Z_3-Z_1)(\epsilon^m_2\epsilon^\pi_3)^2\overline{g}(Z_1,S_2,X_2,Z_2,X_3,Z_3)d(Z_1,S_2,X_2,Z_2,X_3,Z_3)$$
$$\approx\int \Gamma^2(Z_1;u_s,u_{x2},u_{z2})\Psi^2(Z_1;u_{x3},u_{z3})\boldsymbol{K}^2_s(u_s)\boldsymbol{K}^2_x(u_{x2})\boldsymbol{K}^2_x(u_{x3})\boldsymbol{K}^2_z(u_{z2})\boldsymbol{K}^2_z(u_{z3})$$
$$\times(\sigma_m(s,x,Z_1)\sigma_\pi(x,Z_1))^2\overline{g}(Z_1,s,x,Z_1,x,Z_1)d(Z_1,u_s,u_{x2},u_{z2},u_{x3},u_{z3}),$$

*where we apply changing of variable:* $(S_2-s)/h=u_s$, $(X_2-x)/h=u_{x2}$, $(X_3-x)/h=u_{x3}$, $(Z_2-Z_1)/h=u_{z2}$, $(Z_3-Z_1)/h=u_{z3}$ *and use law of iterated expectation, it is straightforward to see that the above term is finite. Then by Serfling (1980, Theorem, Chapter 5.5.2),* $n\tilde{U}_{1n}\xrightarrow{d}3\mathscr{Y}$, *where* $\mathscr{Y}$ *is an infinite weighted sum of* $\chi^2$ *distributions. So the order of* $U_{1n}$ *is*

$$U_{1n}\sim O_p\left(\frac{n(n-1)(n-2)\sqrt{h^{d_m+d_\pi}}}{n^3h^{d_m+1}h^{d_\pi}}\times\frac{1}{n}\right)=O_p\left(\frac{1}{\sqrt{n^2h^{d_\pi+d_m+2}}}\right).$$

*By Assumption 9-iii,* $nh^{d_z+d_\pi}\succ nh^{d_z+d_\pi+1}\to\infty$, *hence,*

$$\kappa_nU_{1n}\sim\sqrt{nh^{d_m+2-d_z}}\times O_p\left(\frac{1}{\sqrt{n^2h^{d_\pi+d_m+2}}}\right)\sim O_p\left(\sqrt{\frac{1}{nh^{d_z+d_\pi}}}\right)=o_p(1).$$

***Part 2:*** $\boldsymbol{U_{2n}}$. *Wherever causes no confusion, we will write* $\Gamma_{ij}=\Gamma\left(Z_i;\frac{S_j-s}{h},\frac{X_j-x}{h},\frac{Z_j-Z_i}{h}\right)$ *and* $\Psi_{it}=\Psi\left(Z_i;\frac{X_t-x}{h},\frac{Z_t-Z_i}{h}\right)$. *Now we analyze the* $U_{2n}$ *term, which we can write as*

$$U_{2n}=\frac{1}{n^3h^{d_m+1}h^{d_\pi}}\times\sum_{i=1}^n\sum_{j\neq i}\Gamma_{ij}\Psi_{ij}\boldsymbol{K}_{s,h}(S_j-s)\boldsymbol{K}^2_{x,h}(X_j-x)\boldsymbol{K}^2_{z,h}(Z_j-Z_i)\epsilon^m_j\epsilon^\pi_j$$
$$=\frac{h^\tau n(n-1)}{n^3h^{d_m+1}h^{d_\pi}}\times\underbrace{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\underbrace{\frac{1}{h^\tau}\Gamma_{ij}\Psi_{ij}\boldsymbol{K}_{s,h}(S_j-s)\boldsymbol{K}^2_{x,h}(X_j-x)\boldsymbol{K}^2_{z,h}(Z_j-Z_i)\epsilon^m_j\epsilon^\pi_j}_{\phi^*_{ij}}}_{\tilde{U}_{2n}}$$

*where* $\tau=\frac{1}{2}d_s+d_z+\frac{1}{2}d_x$. *Now we analyze* $\tilde{U}_{2n}$, *which is a second order statistics with kernel function* $\phi^*_{ij}$. *Let* $\phi_{ij}$ *be the symmetric transformation of* $\phi^*_{ij}$. *To calculate the variance of*

$\tilde{U}_{2n}$, *we need to calculate the variance of* $\mathbb{E}[\phi_{ij}|H_1] = \frac{1}{2}\mathbb{E}[\phi^*_{12}|H_1] + \frac{1}{2}\mathbb{E}[\phi^*_{21}|H_1]$.

$$\mathbb{E}[\phi^*_{12}|H_1] = \frac{1}{h^\tau}\int \Gamma_{12}\Psi_{12}\boldsymbol{K}_{s,h}(S_2-s)\boldsymbol{K}^2_{x,h}(X_2-x)\boldsymbol{K}^2_{z,h}(Z_2-Z_1)\mathbb{E}[\epsilon^m_2\epsilon^\pi_2|W_2]g_w(W_2)dW_2$$
$$= \frac{h^{d_m}}{h^\tau}\int \Gamma\left(Z_1;u_s,u_x,u_z\right)\Psi\left(Z_1;u_x,u_z\right)\boldsymbol{K}_s(u_s)\boldsymbol{K}^2_x(u_x)\boldsymbol{K}^2_z(u_z)\sigma_{m\pi}(s,x,Z_1+hu_z)g_w(s,x,Z_1+hu_z)du_sdu_xdu_z$$
$$= h^{\frac{1}{2}d_s+\frac{1}{2}d_x}\int \Gamma\left(Z_1;u_s,u_x,u_z\right)\Psi\left(Z_1;u_x,u_z\right)\boldsymbol{K}_s(u_s)\boldsymbol{K}^2_x(u_x)\boldsymbol{K}^2_z(u_z)\sigma_{m\pi}(s,x,Z_1+hu_z)g_w(s,x,Z_1+hu_z)du_sdu_xdu_z$$

*where we apply the changing of variable* $(S_2-s)/h = u_s$, $(X_2-x)/h = u_x$, $(Z_2-Z_1)/h = u_z$, *and* $\sigma_{m\pi}(s,x,z) = \mathbb{E}[\epsilon^m\epsilon^\pi|W=w]$. *It is not difficult to see that* $\mathbb{E}(\mathbb{E}[\phi^*_{12}|H_1])^2 \downarrow 0$.

*Next, we look at* $\mathbb{E}[\phi^*_{21}|H_1]$ *and apply* $(Z_1-Z_2)/h = u_z$,

$$\mathbb{E}[\phi^*_{21}|H_1] = \frac{1}{h^\tau}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}^2_{x,h}(X_1-x)\int \Gamma_{21}\Psi_{21}\boldsymbol{K}^2_{z,h}(Z_1-Z_2)\mathbb{E}[\epsilon^m_1\epsilon^\pi_1|W_1]g_z(Z_2)dZ_2$$
$$= \frac{h^{d_z}}{h^\tau}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}^2_{x,h}(X_1-x)\int \Gamma\left(Z_1-hu_z;\frac{S_1-s}{h},\frac{X_1-x}{h},u_z\right)\Psi\left(Z_1-hu_z;\frac{X_1-x}{h},u_z\right)\boldsymbol{K}^2_z(u_z)\mathbb{E}[\epsilon^m_1\epsilon^\pi_1|W_1]g_z(Z_1-hu_z)du_z$$
$$\approx \frac{h^{d_z}}{h^\tau}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}^2_{x,h}(X_1-x)\sigma_{m\pi}(W_1)g_z(Z_1)\int \Gamma\left(Z_1;\frac{S_1-s}{h},\frac{X_1-x}{h},u_z\right)\Psi\left(Z_1;\frac{X_1-x}{h},u_z\right)\boldsymbol{K}^2_z(u_z)du_z$$

*where we apply* $(Z_1-Z_2)/h = u_z$. *Therefore,*

$$\mathbb{E}(\mathbb{E}[\phi^*_{21}|H_1])^2 =$$
$$\approx \frac{h^{2d_z}}{h^{2\tau}}\int \boldsymbol{K}^2_{s,h}(S_1-s)\boldsymbol{K}^4_{x,h}(X_1-x)\sigma^2_{m\pi}(W_1)g^2_z(Z_1)\left(\int \Gamma\left(Z_1;\frac{S_1-s}{h},\frac{X_1-x}{h},u_z\right)\Psi\left(Z_1;\frac{X_1-x}{h},u_z\right)\boldsymbol{K}^2_z(u_z)du_z\right)^2 g_z(Z_1)dZ_1$$
$$\approx \frac{h^{2d_z+d_x+d_s}}{h^{2\tau}}\int \boldsymbol{K}^2_s(u_s)\boldsymbol{K}^4_x(u_x)\sigma^2_{m\pi}(s,x,Z_1)g^2_z(Z_1)\left(\int\int \Gamma\left(Z_1;u_s,u_x,u_z\right)\Psi\left(Z_1;u_x,u_z\right)\boldsymbol{K}^2_z(u_z)du_z\right)^2 g_z(Z_1)dZ_1,$$

*which is of order* $O(1)$ *the last equality holds because* $\tau = \frac{1}{2}d_s + d_z + \frac{1}{2}d_x$. *So we have* $\mathbb{E}(\mathbb{E}[\phi_{21}|H_1])^2 = \max\{\mathbb{E}(\mathbb{E}[\phi^*_{21}|H_1])^2, \mathbb{E}(\mathbb{E}[\phi^*_{12}|H_1])^2\} = O(1)$. *Therefore,*

$$\sqrt{n}\left(\tilde{U}_{2n} - \mathbb{E}[\tilde{U}_{2n}]\right) = O_p(1).$$

*It remains to analyze the order of* $\mathbb{E}[\tilde{U}_{2n}]$.

$$\mathbb{E}[\phi^*_{21}] = \mathbb{E}[\mathbb{E}[\phi^*_{21}|H_1]] =$$
$$= \frac{h^{d_z}}{h^\tau}\int \boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}^2_{x,h}(X_1-x)\sigma_{m\pi}(W_1)g_z(Z_1)\int \Gamma\left(Z_1;\frac{S_1-s}{h},\frac{X_1-x}{h},u_z\right)\Psi\left(Z_1;\frac{X_1-x}{h},u_z\right)\boldsymbol{K}^2_z(u_z)du_zdW$$
$$= h^{\frac{1}{2}d_s+\frac{1}{2}d_x}\int \boldsymbol{K}^2_x(u_x)\boldsymbol{K}_s(u_s)\sigma_{m\pi}(s,x,Z_1)g_z(Z_1)\Gamma\left(Z_1;u_s,u_x,u_z\right)\Psi\left(Z_1;u_x,u_z\right)\boldsymbol{K}^2_z(u_z)du_zdu_xdu_sdZ_1$$

*In the mean time, we can show that*

$$\mathbb{E}[\phi^*_{12}] = \mathbb{E}[\mathbb{E}[\phi^*_{12}|H_1]] =$$
$$= \frac{h^{d_z}}{h^\tau}\int \boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}^2_{x,h}(X_1-x)\sigma_{m\pi}(W_1)g_z(Z_1)\int \tilde{\Gamma}(u_z,S_1,X_1)\tilde{\Psi}(u_z,X_1)\boldsymbol{K}^2_z(u_z)du_zdW_1$$
$$= h^{\frac{1}{2}d_s+\frac{1}{2}d_x}\int \Gamma\left(Z_1;u_s,u_x,u_z\right)\Psi\left(Z_1;u_x,u_z\right)\boldsymbol{K}_s(u_s)\boldsymbol{K}^2_x(u_x)\boldsymbol{K}^2_z(u_z)\sigma_{m\pi}(s,x,Z_1+hu_z)g_w(s,x,Z_1+hu_z)du_sdu_xdu_zd$$

*So* $\mathbb{E}[\tilde{U}_{2n}] = O(h^{d_m}/h^{\tau}) = h^{\frac{1}{2}d_s+\frac{1}{2}d_x}$. *Now we can conclude that*

$$\tilde{U}_{2n} = \underbrace{\tilde{U}_{2n} - \mathbb{E}[\tilde{U}_{2n}]}_{=O_p\left(\frac{1}{\sqrt{n}}\right)} + \underbrace{\mathbb{E}[\tilde{U}_{2n}]}_{=O\left(h^{\frac{1}{2}d_s+\frac{1}{2}d_x}\right)}.$$

*Therefore,*

$$\kappa_n U_{2n} \sim \frac{h^{\tau}\sqrt{nh^{d_m+2-d_z}}}{nh^{d_m+1+d_\pi}} \times O_p\left(\frac{1}{\sqrt{n}}\right) + \frac{h^{\tau}\sqrt{nh^{d_m+2-d_z}}}{nh^{d_m+1+d_\pi}} \times O\left(h^{\frac{1}{2}d_s+\frac{1}{2}d_x}\right)$$
$$= O_p\left(\frac{1}{nh^{d_\pi}}\right) + O_p\left(\sqrt{\frac{h^{d_s}}{nh^{d_\pi+d_z}}}\right) = o_p(1).$$

*The RHS is* $o(1)$ *because Assumption 9-(iii).*

***Part 3:*** $\boldsymbol{U_{3n}}$. *Let* $k_0 = K(0)$. *Recall that*

$$U_{3n} = \frac{k_0}{n^3h^{d_m+1}h^{d_\pi}} \times \sum_{i=1}^{n}\sum_{t\neq i}^{n} \underbrace{\Gamma_{ii}\Psi_{it}\boldsymbol{K}_{s,h}(S_i-s)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{x,h}(X_t-x)\boldsymbol{K}_{z,h}(Z_t-Z_i)\epsilon_i^m\epsilon_t^\pi}_{\zeta_{ij}^*}.$$

*Let* $\zeta_{ij} = \frac{1}{2}(\zeta_{ij}^* + \zeta_{ji}^*)$. *So* $U_{3n} = \frac{k_0}{n^3h^{d_m+1}h^{d_\pi}} \times \sum_{i=1}^n\sum_{t\neq i}^n \zeta_{ij}$ *is proportional to a U-statistics. Now consider*

$$\mathbb{E}[\zeta_{12}^*|H_1] = \mathbb{E}[\Gamma_{11}\Psi_{12}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}_{x,h}(X_1-x)\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{z,h}(Z_2-Z_1)\epsilon_1^m\epsilon_2^\pi|H_1]$$
$$= \Gamma_{11}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}_{x,h}(X_1-x)\epsilon_1^m\mathbb{E}[\Psi_{12}\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{z,h}(Z_2-Z_1)\epsilon_2^\pi|H_1]$$
$$= \Gamma_{11}\boldsymbol{K}_{s,h}(S_1-s)\boldsymbol{K}_{x,h}(X_1-x)\epsilon_1^m\mathbb{E}[\Psi_{12}\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{z,h}(Z_2-Z_1)\mathbb{E}[\epsilon_2^\pi|H_1,W_2]|H_1] = 0$$

*Likewise,*

$$\mathbb{E}[\zeta_{21}^*|H_1] = \mathbb{E}[\Gamma_{22}\Psi_{21}\boldsymbol{K}_{s,h}(S_2-s)\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{x,h}(X_1-x)\boldsymbol{K}_{z,h}(Z_1-Z_2)\epsilon_2^m\epsilon_1^\pi|H_1]$$
$$= \boldsymbol{K}_{x,h}(X_1-x)\epsilon_1^\pi\mathbb{E}[\Gamma_{22}\Psi_{21}\boldsymbol{K}_{s,h}(S_2-s)\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{z,h}(Z_1-Z_2)\epsilon_2^m|H_1]$$
$$= \boldsymbol{K}_{x,h}(X_1-x)\epsilon_1^\pi\mathbb{E}[\Gamma_{22}\Psi_{21}\boldsymbol{K}_{s,h}(S_2-s)\boldsymbol{K}_{x,h}(X_2-x)\boldsymbol{K}_{z,h}(Z_1-Z_2)\mathbb{E}[\epsilon_2^m|H_1,W_2]|H_1] = 0.$$

*Therefore,*

$$\mathbb{E}[\zeta_{12}|W_1,Y_1] = 0 \Rightarrow \mathbb{E}[\zeta_{ij}] = 0.$$

*So we can conclude that* $U_{3n}$ *is proportional to a degenerate U-statistics. It remains to find the order of the variance:* $\mathbb{E}[\zeta_{12}^2] = \frac{1}{4}\mathbb{E}\left(\zeta_{21}^* + \zeta_{12}^*\right)^2$.

*Consider* $\mathbb{E}(\zeta_{21}^*)^2$ *first.*

$$\mathbb{E}(\zeta_{21}^*)^2 = \int \Gamma_{11}^2\Psi_{12}^2\boldsymbol{K}_{s,h}^2(S_1-s)\boldsymbol{K}_{x,h}^2(X_1-x)\boldsymbol{K}_{x,h}^2(X_2-x)\boldsymbol{K}_{z,h}^2(Z_2-Z_1)$$
$$\times \sigma_m^2(W_1)\sigma_\pi^2(W_2)g(W_1)g(W_2)dW_1dW_2.$$

*Apply changing variable routine it is easy to see that* $\mathbb{E}(\zeta_{21}^*)^2 = O(h^{d_m+d_\pi})$. *Likewise,*

*$\mathbb{E}(\zeta_{12}^*)^2 = O(h^{d_m+d_\pi})$. Therefore, $\mathbb{V}(\zeta_{12}) = O(h^{d_m+d_\pi})$. As the consequence, it follows that*

$$\frac{1}{n(n-1)}\frac{1}{\sqrt{h^{d_m+d_\pi}}}\sum_{i=1}^{n}\sum_{t\neq i}^{n}\zeta_{ij} = O_p\left(\frac{1}{n}\right),$$

*or equivalently,*

$$U_{3n} = \frac{k_0(n-1)\sqrt{h^{d_m+d_\pi}}}{n^2 h^{d_m+1}h^{d_\pi}}\frac{1}{n(n-1)}\frac{1}{\sqrt{h^{d_m+d_\pi}}}\sum_{i=1}^{n}\sum_{t\neq i}^{n}\zeta_{ij} = O\left(\frac{1}{n^2\sqrt{h^{d_m+d_\pi+2}}}\right) = o(\kappa_n^{-1}).$$

***Part 4: $U_{4n}$***. *Recall that*

$$U_{4n} = \frac{k_0}{n^3 h^{d_m+1}h^{d_\pi}}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Gamma_{ij}\Psi_{ii}\boldsymbol{K}_{s,h}(S_j-s)\boldsymbol{K}_{x,h}(X_j-x)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{z,h}(Z_j-Z_i)\epsilon_j^m\epsilon_i^\pi,$$

*Following the same argument for $U_{3n}$ that $U_{4n} = o(\kappa_n^{-1})$.*

***Part 5: $U_{5n}$***. *Now we consider $U_{5n}$, then*

$$\begin{aligned} U_{5n} &= \frac{k_0^2}{n^3 h^{d_m+1}h^{d_\pi}}\sum_{i=1}^{n}\Gamma_{ii}\Psi_{ii}\boldsymbol{K}_{s,h}(S_i-s)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{x,h}(X_i-x)\epsilon_i^m\epsilon_i^\pi \\ &= \frac{k_0^2}{n^2 h^{d_\pi+1}}\left\{\frac{1}{nh^{d_m}}\sum_{i=1}^{n}\Gamma_{ii}\Psi_{ii}\boldsymbol{K}_{s,h}(S_i-s)\boldsymbol{K}_{x,h}(X_i-x)\boldsymbol{K}_{x,h}(X_i-x)\epsilon_i^m\epsilon_i^\pi\right\}, \end{aligned}$$

*the term in the bracket converges in probably to a finite constant by law of large number and applying usual changing variable trick, so $U_{5n} = O_p(\frac{1}{n^2h^{d_\pi+1}}) = o(\kappa_n^{-1})$.*

***Part 6: $U_n$***. *Combine Part 1–pat 5, we can conclude that $U_n = o(\kappa_n^{-1})$. Hence the conclusion of the lemma holds.* □

## Appendix B. Auxiliary Results for Section 2

Recall that

$$\Omega_m(s,x) = \int\left\{\sigma_m^2(s,x,Z_1)\left(\int\Gamma\left(Z_1;u_s,u_x,-u_z\right)\boldsymbol{K}_z(u_z)du_z\right)^2 \times\boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)g_z^2(Z_1)\boldsymbol{\pi}(x,Z_1)\boldsymbol{\pi}'(x,Z_1)\right\}g_w(s,x,Z_1)du_xdu_sdZ_1.$$

We first look at the term $\int\Gamma\left(Z_1;u_s,u_x,-u_z\right)\boldsymbol{K}(u_z)du_z$. Recall that $\Sigma_m$ is a matrix

$$\Sigma_m = \begin{pmatrix} \Sigma_{m,0,0} & \Sigma_{m,0,1} & \cdots & \Sigma_{m,0,p} \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma_{m,p,0} & \Sigma_{m,p,1} & \cdots & \Sigma_{m,p,p} \end{pmatrix},$$

in which $\Sigma_{m,i,j}$ is an $N_i$ by $N_j$ matrix whose $(\ell,k)$ element is $\nu_{m,\tau_i(\ell)+\tau_j(k)}$, where $\nu_{m,\underline{j}} = g_w(w)\int\boldsymbol{K}_w(u)u^{\underline{j}}du$. Therefore,

$$\Sigma_m = g_w(w)\Lambda_m,$$

where $\Lambda_m$ is defined analogously as $\Sigma_m$ but with a typical element being $\lambda_{m,\underline{j}} = \int \boldsymbol{K}_w(u)u^{\underline{j}}du$. Hence, the corresponding row of $\Sigma_m^{-1}$, $\Sigma_m^{-(s,\cdot)}(w) = g_w^{-1}(w)\Lambda_m^{-(s,\cdot)}$, where $\Lambda_m^{-(s,\cdot)}$ is the corresponding row of the inverse of $\Lambda_m$ and is just a vector of constants. Let it be $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_{\bar{N}})$, where $\bar{N} = \sum_{j=1}^{p} N_j$. By definition of $\Gamma$ (see Equation (A.6)),

$$\int \Gamma\left(Z_1; u_s, u_x, -u_z\right) \boldsymbol{K}_z(u_z)du_z = g_w^{-1}(s, x, Z_1) \times \int \boldsymbol{\lambda} \cdot \left(1, -u_z, u_x, u_s, \cdots, (u_s)^p\right)' \boldsymbol{K}_z(u_z)du_z \quad \text{(B.1)}$$

Note that the integral part of the right hand side is an known function of $(u_s, u_x)$, for which we denote as $\varpi(u_s, u_x)$. Given this, the formula for $\Omega_m$ can be simplified as

$$\Omega_m = \int \varpi^2(u_s, u_x)\boldsymbol{K}_x^2(u_x)\boldsymbol{K}_s^2(u_s)du_x du_s \times \int \left\{ \frac{\sigma_m^2(s, x, Z_1)\boldsymbol{\pi}(x, Z_1)\boldsymbol{\pi}'(x, Z_1)}{g_{S,X|Z}(s, x|Z = Z_1)} \right\} g_z(Z_1)dZ_1.$$

The first term on the right hand side is a constant (let us denote it by $c_K$) can be directly calculated based on the kernel and the second term can be estimated by sample analogs and plug-in estimators.

$$\widehat{\Omega}_m = \frac{c_K}{n} \sum_i \frac{\hat{\sigma}_m^2(s, x, Z_i)\hat{\boldsymbol{\pi}}(x, Z_i)\hat{\boldsymbol{\pi}}'(x, Z_i)}{\hat{g}_{S,X|Z}(s, x|Z_i)}, \quad \text{(B.2)}$$

where

$$\hat{\sigma}_m^2(s, x, z) = \frac{\sum_i (Y_i - \hat{Y}_i)\boldsymbol{K}_{w,h}(W_i - w)}{\text{Tr}\left(\mathbb{W}_j^* - \mathbb{W}_j^*\mathbb{X}_j^*\left(\mathbb{X}_j^{*\prime}\mathbb{W}_j^*\mathbb{X}_j^*\right)^{-1}\mathbb{X}_j^{*\prime}\mathbb{W}_j\right)},$$

and $\mathbb{X}_j^*$ and $\mathbb{W}_j^*$ are defined analogously to $\mathbb{X}_j$ and $\mathbb{W}_j$ defined in **??**.

**Example 1.** *Consider the case in which $S, X$ and $Z$ are all one-dimensional. Suppose $p = 2$ then we have $\bar{N} = N_0 + N_1 + N_2 = 10$. If we use triangular kernel (for each variable), that is, $K(u) = (1 - |u|)\mathbf{1}\{|u| \leq 1\}$, then $\int u^r K(u)du = 0$ for odd $r$, and $\int u^2 K(u)du = \frac{1}{6}$, $\int u^4 K(u)du = \frac{1}{15}$.*

$$\Lambda_m = \begin{pmatrix} 1 & 0 & 0 & 0 & \frac{1}{6} & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 & \frac{1}{15} & 0 & 0 & \frac{1}{36} & 0 & \frac{1}{36} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{36} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{36} & 0 & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 & \frac{1}{36} & 0 & 0 & \frac{1}{15} & 0 & \frac{1}{36} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{36} & 0 \\ \frac{1}{6} & 0 & 0 & 0 & \frac{1}{36} & 0 & 0 & \frac{1}{36} & 0 & \frac{1}{15} \end{pmatrix}_{10\times 10}$$

*Calculating the inverse of $\Lambda_m$ we found that the only nonzero element in the fourth row of $\Lambda_m^{-1}$ is the fourth element, that is, $\boldsymbol{\lambda} = (0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0)$. Therefore, $\int \Gamma\left(Z_1; u_s, u_x, -u_z\right) K(u_z)du_z$*

*is simply*

$$\int \Gamma\left(Z_1; u_s, u_x, -u_z\right) K(u_z) du_z = g_w^{-1}(s, x, Z_1) \times 6u_s.$$

*In this case,* $\varpi(u_s, u_x) = 6u_s$ *and therefore,*

$$c_K = \int_{-1}^{1} 36u_s^2(1 - |u_s|)^2 du_s \times \int_{-1}^{1} (1 - |u_x|)^2 du_x = \frac{8}{5}.$$

*The estimator for* $\Omega_m$ *is therefore*

$$\widehat{\Omega}_m = \frac{8}{5n} \sum_i \frac{\hat{\sigma}_m^2(s, x, Z_i) \hat{\boldsymbol{\pi}}(x, Z_i) \hat{\boldsymbol{\pi}}'(x, Z_i)}{\hat{g}_{S,X|Z}(s, x|Z_i)}.$$

**Corollary 1.** *Let* $\widehat{\Omega}_m$ *be defined as in Equation* (B.2) *and* $\widehat{V}$ *be defined as in* **??**, *and let* $\hat{g}_W(w)$ *be a uniformly consistent estimator for the joint density of* $W$ *at* $(s, x, z)$, *and suppose* $\sigma_m^2(s, x, z)$ *is constant in a local neighborhood of* $(s, x, z)$, *then*

$$\left(\widehat{V}^{-1} \widehat{\Omega}_m \widehat{V}^{-1}\right)^{-\frac{1}{2}} \sqrt{n h^{d_x + d_s + 2}} (\hat{\beta}(s, x) - \beta(s, x)) \xrightarrow{d} N(0, I).$$

## References


Chernozhukov, V., W. Kim, S. Lee, and A. M. Rosen (2013): "Implementing Intersection Bounds in Stata," *working paper*.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Ha, W., J. Yi, Y. Yuan, and J. Zhang (2016): "The dynamic effect of rural-to-urban migration on inequality in source villages: System GMM estimates from rural China," *China Economic Review*, 37, 27–39.

Holm, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26(05), 1529–1564.

Liu, S., I. Mourifie, and Y. Wan (2019): "Two-way Exclusion Restrictions in Models with Heterogeneous Treatment Effects," Discussion paper.

Mourifié, I., and Y. Wan (2017): "Testing Local Average Treatment Effect Assumptions," *Review of Economics and Statistics*, 99(2), 305–313.

Serfling, R. J. (1980): *Approximation theorems of mathematical statistics*. John Wiley & Sons.