TWO-WAY EXCLUSION RESTRICTIONS IN MODELS WITH HETEROGENEOUS TREATMENT EFFECTS

ISMAEL MOURIFIÉ† AND YUANYUAN WAN‡

UNIVERSITY OF TORONTO

ABSTRACT. This paper discusses identification and estimation of nonparametric structural functions in models with discrete endogenous regressors (treatment) and additive treatment-specific error terms. We do not make functional form assumption or impose shape restrictions on the selection equation, which is also allowed to contain multi-dimensional error vectors. We focus on applications in which there exists two-way exclusive variables: (i) an outcome-exclusive variable which affects the treatment but is excluded from the potential outcome equation, (ii) another treatment-exclusive exogenous variable which affects the potential outcome but excluded from the selection equation. We nonparametrically identify the derivative of conditional average treatment effect and (up to a location normalization) the conditional average treatment effect. We also propose an asymptotically normal two-step estimator.

Keywords: Two-way exclusion, nonparametric identification, heterogeneous treatment effect

JEL: C14, C21

This version: Tuesday 1st March, 2016

† Corresponding author. Department of Economics, University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada. ismael.mourifie@utoronto.ca.
‡ Department of Economics, University of Toronto, yuanyuan.wan@utoronto.ca.
1. Introduction

We study the following potential outcome model

\[
\begin{cases}
  Y = \sum_{d=0}^{T} 1[D = d]Y_d \\
  Y_d = f_d(S, X) + U_d \\
  D = \vartheta(X, Z, V)
\end{cases}
\]

(1)

where for \(d \in D = \{0, 1, 2, \cdots, T\}\), \(Y_d\) is the potential outcome when the treatment variable \(D\) is externally set to \(d\), \(U_d \in \mathbb{R}\) is the treatment specific error term, \(Y\) is the observed outcome, \(X \in \mathcal{X} \subset \mathbb{R}^{d_x}\) is a vector of covariates, \(S \in \mathcal{S} \subset \mathbb{R}^{d_s}\) is a treatment-exclusive vector that only enters the potential outcome equation, \(Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}\) is the outcome-exclusive vector which only enters the selection equation, and \(V\) is a vector (can potentially be infinite dimensional) of error terms. Without loss of generality, we focus on univariate \(S\) case (i.e. \(d_s = 1\)) in the rest of the paper.

We will discuss, for a given value of \((s, x)\), the identification of structural functions \(f_d(s, x)\) \(d \in \{0, ..., T\}\) and their derivatives with respect to the treatment exclusive variable \(\partial f_d(s, x) / \partial s\).\(^1\)

From them we can therefore identify \(\Delta_d(s, x) = f_d(s, x) - f_{d'}(s, x)\) and its derivative \(\beta_d(s, x) = \partial \Delta_d(s, x) / \partial s\). Both of them have interesting policy implications. For example, \(\Delta_d(s, x)\) can be interpreted as the conditional average treatment effect under the usual conditional mean independence assumption of the structural errors, i.e., \(E[U_d | S, X] = 0 \text{ a.s.}\), despite such an assumption is not needed for identification of \(\beta_d(s, x)\). We will discuss more examples and the relationship between \(\beta_d(s, x)\) and conventional parameters in more details in Sections 4.1 and 4.2.

To identify \(\beta_d(s, x)\), we do not make functional form assumptions or impose shape restrictions on either potential outcome function \(f_d\) or the selection function \(\vartheta\); we allow the error term \(V\) to be multi- or even infinite dimensional. We neither assume \(Z\) has a large support nor require it to be independent of \((U, V)\) conditional of other covariates. Instead, we focus on applications in which there exists a treatment-exclusive exogenous regressor \(S\), which affects the potential outcome but not the treatment (will be defined rigorously later). In our context, \(S\) and \(Z\) together forms a two-way exclusive restriction, which will play a determinant role in our identification for \(\beta_d(s, x)\). The

\(^1\) We focus on the case in which \(S\) is continuous and \(f_d\) is differentiable with respect to \(s\). As we shall see later, our identification results hold straightforwardly when \(S\) is discrete.
identification of $\beta_{d,d'}(s,x)$ also leads to the identification of $\Delta_{d,d'}(s,x)$ under an additional location normalization.

Similar to the conventional exclusion variable $Z$, the existence of $S$ is more natural in some applications than others. Recently, Eisenhauer, Heckman, and Vytlacil (2015) consider a generalized Roy model with limited information by the agent and discuss scenarios in which such two-way exclusive variables exist. For example, the outcome-exclusive variable $Z$ plays a similar role as the cost shifter of taking the treatment (in the terms of Eisenhauer, Heckman, and Vytlacil, 2015), and the treatment-exclusive variable $S$ likewise plays a similar role as the benefit shifter which is not perfectly foreseen at the time of treatment. The existence of $S$ variable is also often justifiable in two-step decision problems. In such models, an economic agent first chooses to engage in an activity (or in a regime) and then subsequently decides the level of his/her activity. If there is an outcome-relevant variable being realized between the two decisions, then it can be a candidate of the $S$ variable. For instance, in the studies of the returns of college education, the first decision of a high school graduate is to go to college or not, which of course determines the level of his/her future outcomes. However, there is often a significant time gap between college enrollment and the realization of returns to college education. During this period, many variables such as the contemporaneous local wage rate or the experience as used in Eisenhauer, Heckman, and Vytlacil (2015) will affect the present wage, however their realizations were unknown to the student when he/she made decision for college enrollment.

Our model is related to the literature of nonparametric instrumental variable regression where the error terms are additively separable, see for example Darolles, Fan, Florens, and Renault (2011), Newey, Powell, and Vella (1999), Newey and Powell (2003), Pinkse (2000) and Severini and Tripathi (2006), among others. Our model is different in that we consider a model with discrete regressors, whereas the before-mentioned papers consider continuous endogenous regressors. The model studied in Das (2005) and Florens and Malavolti (2002) is similar to ours but they assumed $U_d = U$ for all $d$, that is, the error terms are not treatment-specific, which rules out the “essential heterogeneity” in the treatment effect (see discussions in Heckman, Urzua, and Vytlacil, 2006). Following Heckman and Vytlacil (2005), we allow $U_d$ to be different across $d$, and therefore allow the treatment effect be

---

2Imbens and Newey (2009) allows non-separable and multi-dimensional errors in outcome equation, where they focus on continuous endogenous regressors and assume monotonicity in the selection equation.
heterogeneous even conditional on exogenous regressors. However, unlike Heckman and Vytlacil (2005), we impose no structure or monotonicity restriction on the treatment; but as a trade-off, we will need the existence of a treatment-exclusive regressor $S$.

Another way of taking into account heterogeneous treatment effect is to use a model with non-separable structural function. Thereby, a part on the literature has been focusing on nonseparable structural functions with discrete endogenous variables, see for instance Chesher (2005), Jun, Pinkse, and Xu (2011), Shaikh and Vytlacil (2011), Mourifié (2015) among many others. However, most of the literature imposes additional restrictions on the selection equation, such as threshold crossing model with a scalar-valued latent error term and achieve, in general, set identification results. ³ Using the exclusive variable $S$ and assuming the additive-separability structure of potential outcomes free us from making shape restrictions on the selection equation.

Our paper also contributes to the discussion on the relevance of the Instrumental Variable (IV) estimand in the presence of heterogeneous treatment response. Indeed, in presence of heterogeneous treatment effects, Imbens and Angrist (1994) showed that the (IV) estimand identifies the average treatment effect only for a subpopulation, namely compliers, under the monotonicity assumption imposed on the treatment. Heckman and Vytlacil (2005) pointed out that without the monotonicity restriction on the treatment, the IV estimand ceases to identify interesting policy effect parameters. Here we contribute to this discussion by showing in presence of this treatment-exclusive regressor, our proposed estimand identifies an interesting policy effect parameter without imposing any further restrictions on the treatment. Moreover, if some type of location normalizations can be justified in the empirical application under study, we can recover from our proposed estimand the well-known average treatment effect parameter for the entire population.

We propose an asymptotically normal two-step estimator for $\beta_{d,d'}(s,x)$. In the first step, we estimate the conditional choice probability of treatment and derivative of the conditional expectation of outcome with respect to the treatment-exclusive variable $S$ by local linear estimation method. The second step estimator utilizes the fitted value of first stage estimators and has an explicit form, which is analogous to the ordinary least square estimator. We show our estimator is asymptotically normal.

³A rare exception is Chernozhukov and Hansen (2005), who do not impose direct structure on the selection equation but require a type of monotone likelihood ratio condition.
and has the (nonparametric) convergence rate of Linton and Härdle (1996)-type marginal integration estimators.

The rest of the paper is organized as follows. We present our main identification result in Section 2. We propose the two-step estimator and derive its asymptotic properties in Section 3. We discuss some extensions in Section 4 and conclude the paper in Section 5.

2. IDENTIFICATION

In this section, we will discuss the identifying assumptions and lay out our identification strategy. To simplify notation, we take \( S \) to be scalar-valued; our results can be extended straightforwardly to the cases with vector-valued \( S \). Let \( W = (X, Z, S) \). We make the following assumptions.

**Assumption 1** (Exclusion restrictions). (i) The variable \( S \) is excluded from the observed treatment, i.e. \( D = \vartheta(X, Z, V) \) for some unknown vector of measurable functions \( \vartheta \) and random vector \( V \). (ii) The variable \( Z \) is excluded from the potential outcome equation, i.e. \( Y_d = f_d(S, X) + U_d \) for each \( d \in \mathcal{D} \).

This assumption is interpreted as follows: there is no direct causal effect of \( S \) on the treatment \( D \), and of \( Z \) on the potential outcome \( Y_d \). Notice there is nearly no restriction on how the selection equation \( D = \vartheta(X, Z, V) \) other than \( S \) being excluded. We allow \( V \) to be multi- or infinite dimensional and do not impose any monotonicity or threshold crossing conditions on \( \vartheta \).

**Assumption 2** (Independence). Let \( U = (U_0, U_1, \cdots, U_T)' \), then \( (U', V)' \perp S | (X, Z) \).

Assumption 2 only requires \( S \) to be independent with all the latent error terms when conditional on \( (X, Z) \). This assumption is also made in Eisenhauer, Heckman, and Vytlacil (2015, Assumption 1) for the “benefit shifter”. Note we only require \( Z \) be excluded from the potential outcome equation but allow correlation between \( (X, Z) \) and \( (U', V) \). Therefore, we allow \( Z \) to be an imperfect instrument (excluded from the potential outcomes but potentially correlated with their unobserved terms). This allows the variable \( Z \) to be easier to find than conventional instrumental variables used in this literature. Assumption 1-(i) and \( V \perp S | (X, Z) \) implies \( D \perp S | (X, Z) \), which is a testable restriction. Also note we do not require \( S \) and \( Z \) to be independent—either conditionally or unconditionally.

**Assumption 3** (Differentiability). Let \( \mathcal{S}_x \) be the support of \( S \) conditional on \( X = x \). Then for each \( x \in \mathcal{X} \), \( f_d(\cdot, x) \) for \( d = 0, 1, \cdots, T \) is differentiable in the interior of \( \mathcal{S}_x \).
Assumption 3 also implicitly requires \( S \) to be continuous. We make Assumption 3 just for simplification of notation. As will be clearer later, our identification and estimation strategy can be easily extended to the discrete \( S \) case.

We describe our identification strategy below. Without loss of generality, we take \( d' = 0 \) as the reference group. Model (1) can be equivalently written as follows:

\[
Y = \sum_{d=0}^{T} f_d(S, X)1\{D = d\} + \sum_{d=0}^{T} U_d1\{D = d\}
\]

\[
= f_0(S, X) + \sum_{d=1}^{T} (f_d(S, X) - f_0(S, X))1\{D = d\} + \sum_{d=1}^{T} (U_d - U_0)1\{D = d\} + U_0.
\]

We write \( \tilde{f}_d(s, x) = f_d(s, x) - f_0(s, x) \). Let \( W = (S, X', Z')' \) and \( \nu \) be the support of \( W \). Let \( \mathbb{E}[\cdot|w] \) represent the conditional expectation given \( W = w \), we have:

\[
\mathbb{E}[Y|w] = f_0(s, x) + \sum_{d=1}^{T} \tilde{f}(s, x)\mathbb{P}(D = d|w) + \mathbb{E} \left[ \sum_{d=1}^{T} (U_d - U_0)1\{D = d\} + U_0|w \right].
\]

(2)

Under Assumption 1, \( \mathbb{P}(D = d|w) = \mathbb{P}(D = d|x, z) \) for all \((x, z)\) and the last term of the right hand side does not depends on \( s \) too. By taking the derivative of the latter equation with respect to \( s \), we have under Assumptions 2 and 3:

\[
\frac{\partial \mathbb{E}[Y|w]}{\partial s} = \sum_{d=1}^{T} \beta_d(s, x)\mathbb{P}(D = d|x, z) + \beta_0(s, x),
\]

(3)

where \( \beta_0(s, x) \equiv \frac{\partial f_0(s, x)}{\partial s} \) and \( \beta_d(s, x) \equiv \frac{\partial \tilde{f}_d(s, x)}{\partial s} \) for \( d = 1, 2, \cdots, T \).

To simplify the notation, let \( m(w) = \partial \mathbb{E}[Y|w]/\partial s, \pi_d(x, z) = \mathbb{P}(D = d|x, z), \pi_0(x, z) = [\pi_1(x, z), \cdots, \pi_T(x, z)]', \pi(x, z) = [1, \pi_0(x, z)]', \) and \( \beta(s, x) = [\beta_0(s, x), \beta_1(s, x), \cdots, \beta_K(s, x)]' \). Therefore, Equation (3) can be rewritten as follows:

\[
m(s, x, z) = \pi(x, z)'\beta(s, x), \quad \forall (s, x, z) \in \nu
\]

(4)

Equation (4) is the key identifying equation. \( m \) and \( \pi \) are identified directly from the data. \( \beta \) is the parameter of interest. As shall be clear soon, the conditional variation of \( Z \) given \((S, X) = (s, x)\) provides identification power for \( \beta(s, x) \), which is summarized by the following Theorem.
**Theorem 1.** Let \((s, x)\) be a point from the joint support of \(S\) and \(X\). Under Assumptions 1 to 3 and provided that corresponding expectations exist, \(\beta(s, x)\) is identified if and only if the conditional variance \(\mathbb{V}[\pi_0(x, Z) | S = s, X = x]\) is positive definite. The identification equation is given by

\[
\beta(s, x) = \left\{ \mathbb{E}[\pi(x, Z) \pi(x, Z)' | S = s, X = x] \right\}^{-1} \mathbb{E}[\pi(x, Z)m(s, x, Z) | S = s, X = x].
\]

*(5)*

**Proof.** See Appendix A.1. \(\square\)

We have a few comments. First, the above Theorem 1 generalizes Das (2005, Theorem 2.1). The key difference is that Theorem 1 allows treatment-specific error terms by taking advantage of the treatment-exclusive regressor \(S\). Second, when \(Z\) is discrete, the identification of \(\beta(s, x)\) requires that conditional on \((S, X) = (s, x)\), the support of \(Z\) contains at least \(T + 1\) distinguish values: \(\{z_1, z_2, \ldots, z_{T+1}\}\). If the matrix \(\Pi(x) = [\pi(x, z_1)'; \pi(x, z_2)'; \ldots, \pi(x, z_{T+1})']\) has full rank (which is testable), then it is also sufficient for identification. Third, \(\beta\) itself is a parameter of empirical interest. It measures how average treatment effect changes with \(s\), for example, how the return of college education be affected by a variation of the local labor market conditions. To identify \(\beta\), we do not need to impose any location normalization on the distribution of \(U_d\). Lastly, our identification result holds as long as the final term on the right hand side of Equation (2) does not depend on \(s\); therefore the conditional independence between \(S\) and errors (Assumption 2) is sufficient but not necessary.

When \(S\) is discrete, for a given vector of \((s, \tilde{s}, x, z)\), we can define

\[
m(s, \tilde{s}, x, z) \equiv \mathbb{E}[Y|s, x, z] - \mathbb{E}[Y|\tilde{s}, x, z] = f_0(s, x) - f_0(\tilde{s}, x) + \sum_{d=1}^{T-1} (f_d(s, x) - f_0(s, x) + f_0(\tilde{s}, x)) \mathbb{P}(D = d|x, z),
\]

and analogously define \(\beta_d(s, \tilde{s}, x) \equiv f_d(s, x) - f_d(\tilde{s}, x) - f_0(s, x) + f_0(\tilde{s}, x)\). In this scenario, the same identification strategy carries through for the identification of \(\beta_d(s, \tilde{s}, x)\).

Under current assumptions, the structural function \(f_d(s, x)\) can only be identified up to location normalizations. The following corollary shows that if we impose a suitable location normalization, the well known average treatment effect (ATE) is identified.
Corollary 1. If there is $s^*$ such that $f_d(s^*, x) = f_{d'}(s^*, x)$ for $d \neq d'$ then $\Delta_{d,d'}(s, x) \equiv f_d(s, x) - f_{d'}(s, x)$ is identified

$$\Delta_{d,d'}(s, x) = \int_{s^*}^{s} (\beta_d(t, x) - \beta_{d'}(t, x)) dt.$$ 

Furthermore, if $\mathbb{E}[U_d | S, X] = 0$ a.s., then $\Delta_{d,d'}(s, x)$ represents the well-known average treatment effect (ATE) $\mathbb{E}[Y_d - Y_{d'} | S = s, X = x]$.

To illustrate further, let $s$ denote the years of experience as in Eisenhauer, Heckman, and Vytlacil (2015) and $d$ the years of education. We could expect that without any experience i.e., $s = 0$, two individuals with similar observable characteristics $x$ but only one year difference in the level of education may have little effect on expected potential labor income: $\mathbb{E}[Y_d | S = s, X = x] = \mathbb{E}[Y_{d+1} | S = s, X = x]$ or alternatively $f_d(0, x) = f_{d+1}(0, x)$ under $\mathbb{E}[U_d | S, X] = 0$ a.s.. For example, individuals with one or two years of college education but zero working experience may be evaluated in a similar way on the labor market. Notice that with the existence of treatment specific unobserved heterogeneity, those two individuals may still have different outcome realizations $Y_d - Y_{d+1} = U_d - U_{d+1} \neq 0$. Of course, if the gap in the years of education is wide, or one more year’s education upgrades the diploma to next level (e.g. from high school to college), such a location normalization would be less likely to hold.

If the empirical application has no natural location normalization, our identification strategy only leads to identify $f_d(s, x)$ up to an unknown constant with respect to $s$. Identification up to location normalization is common in this literature, see, for example Das, Newey, and Vella (2003) and Das (2005) among many others. For some applications the identification of $\beta_d(s, x)$ would be sufficient. Admittedly, there are also many cases in which the normalization constant (which is still a function of $x$) is important, as discussed in Heckman (1990). Under certain conditions it can be estimated by adapting Andrews and Schafgans (1998)’s methodology. We leave this question for future work.

3. Estimation

In this section, we propose an estimator for $\beta_d(s, x)$ for given $d, s$ and $x$.\(^4\) We make the following assumptions.

\(^4\) We take $X$ and $Z$ as continuous random variables. If $(X, Z)$ contains discrete regressors, we can also conduct the same analysis by conditioning on realizations of those regressors.
Assumption 4. The support of the conditional distribution of $Z|S, X = (s, x)$ does not depend on $(s, x)$. Furthermore, $V[\pi_0(x, Z)]$ is positive definite.

Assumption 4 holds when $Z$ and $(S, X)$ are independent and tends to hold in applications where $Z$ is discrete. We will discuss in Section 4 the estimation procedure without imposing it. Under Assumption 4, Equation (4) implies that for the given $(s, x)$, the following equation holds almost surely with respect to the marginal distribution of $Z$:

$$m(s, x, Z) = \pi(x, Z) \beta(s, x),$$

which in turn implies

$$\beta(s, x) = \{E[\pi(x, Z)\pi(x, Z)']\}^{-1}E[\pi(x, Z)m(s, x, Z)].$$  \hspace{1cm} (6)

Equation (6) suggests a two-step estimator for $\beta(s, x)$. In the first step, we estimate $\pi(x, \cdot)$ and $m(s, x, \cdot)$ nonparametrically; in the second step, we estimate $\beta(s, x)$ by plugging fitted values:

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

We choose the local linear estimator for the first step for following reasons. First, $m(s, x, \cdot)$ is the partial derivative of the conditional mean function of $Y$ given $W = (s, x, z)$ with respect to $s$, which can be conveniently estimated by the local linear method. Second, we will apply the uniform Bahadur representation results of Kong, Linton, and Xia (2010) to the first step estimators and derive the asymptotic distribution for the second step estimator. Although we use the local linear method, it shall be noted the model can also be estimated by other nonparametric methods, for example, sieve estimation.

We define some notations. For a generic random variable vector $\eta$, let $d_\eta$ be its dimension. Define $d_m = d_x + d_z + d_s = d_x + d_z + 1$, $d_\pi = d_x + d_z$. So $d_m$ and $d_\pi$ are the dimensions of the arguments in $m$ and $\pi$, respectively. Let $K(\cdot)$ be a univariate kernel function and $h$ a bandwidth. Let $r^m = (r_1, r_2, \cdots, r_{d_m})$ be a $d_m$-vector of nonnegative integers. Let $|r^m| = \sum_{j=1}^{d_m} r_j$. For the $d_m$-dimensional vector $w$ and a given vector $r^m$, let $\alpha_{r^m} \cdot w^{r^m}$ be a polynomial of $w$ of order $|r^m|$ with corresponding coefficients $\alpha_{r^m}$. We follow the convention that $w^{r^m} = 1$ when $|r^m| = 0$. For example, if $d_s = d_x = d_z = 1$, $w = (s, x, z)'$, and $r^m = (0, 1, 0)$, then $\alpha_{(0,1,0)} \cdot w^{(0,1,0)} = \alpha_{(0,1,0)} x$, a
linear term in $x$ with coefficient $\alpha_{(0,1,0)}$; if instead $\pi^m = (1, 0, 2)$, then $\alpha_{\pi^m} \cdot w^m = \alpha_{(1,0,2)}sz^2$. Let $\alpha^m$ be the stacked vector of $\alpha_{\pi^m}$ of which $0 \leq |\pi^m| \leq p$ for some $p \geq 1$ based on the order such that it is increasing in $|\pi^m|$, and for those with the same $|\pi^m|$, $\alpha_{(|\pi^m|, \ldots, 0)}$ goes first and $\alpha_{(0,0,\ldots, |\pi^m|)}$ goes last. Define\(^5\)

$$\hat{\alpha}^m = \arg\min_{\alpha^m} \frac{1}{2n} \sum_{i=1}^{n} K \left( \frac{S_i - s}{h} \right) K \left( \frac{X_i - x}{h} \right) K \left( \frac{Z_i - z}{h} \right) \times \left( Y_i - \sum_{0 \leq |\pi^m| \leq p} \alpha_{\pi^m} \cdot (W_i - w)^m \right)^2.$$ 

Let $\tilde{W} = (X, Z)$. For $d = 1, 2, \cdots, T$, let $r^{\pi_d}, \alpha^{\pi_d}$, and $\tilde{w}^{r^{\pi_d}}$ be analogously defined as $\pi^m$, $\alpha^m$, and $w^m$, respectively. We define

$$\hat{\alpha}^{\pi_d} = \arg\min_{\alpha^{\pi_d}} \frac{1}{2n} \sum_{i=1}^{n} K \left( \frac{X_i - x}{h} \right) K \left( \frac{Z_i - z}{h} \right) \times \left( 1 \{ D_i = d \} - \sum_{0 \leq |\pi^m| \leq p} \alpha_{\pi^m} \cdot (\tilde{W}_i - \tilde{w})^{\pi^m} \right)^2,$$

Lastly, we define our estimator for $m(w)$ and $\pi_d(x, z)$ as

$$\hat{m}(w) = \hat{\alpha}^m_{(1,0,\ldots,0)}(w), \; \hat{\pi}_d(x,z) = \hat{\alpha}^{\pi_d}_{(0,0,\ldots,0)}(x,z), \quad (7)$$

that is, $\hat{m}(w)$ is the second element of the vector $\hat{\alpha}^m$ and $\hat{\pi}_d(x, z)$ is the first element of $\hat{\alpha}^{\pi_d}$.

We make the following assumptions.

**Assumption 5.** $\{ (Y_i, D_i, X_i, S_i, Z_i) \}_{i=1}^{n}$ are i.i.d. observations.

**Assumption 6.** The symmetric kernel $K(\cdot)$ has support $[-1, 1]$, is continuous differentiable, and satisfies $K(0) = 0$.

**Assumption 7.** The bandwidths $h$ is a polynomial function of sample size $n$ and satisfies (i) $h \to 0$, $nh^{d_n+2} \to \infty$; (ii) $nh^{d_n+2(p+1)} \to 0$; (iii) $nh^{d_{n}+d_z} \to \infty$.

**Assumption 8.** (i) The joint density $g_w$ of $W$ is bounded away from 0 and has bounded first order derivative over its compact support $\mathcal{W}$. (ii) The conditional density $g_{w|u}$ of $W$ given $U$ exists and is bounded. (iii) $E[Y|W = \cdot]$ and $P(D = d|X = x, Z = z)$, for each $d = 0, 1, \cdots, T$, are $p + 1$ times continuously differentiable where $2(p + 1) > d_z$.

\(^5\)In practice, one can choose different kernel functions and/or bandwidth for the two local linear regressions.
Assumption 9. Let \( \epsilon_m = Y - \mathbb{E}[Y|W] \) and \( \gamma_n = nh^{d_m} / \log n \), then \( \mathbb{E} |\epsilon_m|^{\nu_1} < \infty \) for some \( \nu_1 > 0 \) such that \( n^{-1} \gamma_n^{(\nu_1-2)/8} \rightarrow \infty \) at polynomial rate.

Assumption 6 requires that the kernel admits enough smoothness. Assumption 7-(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 \( \pi \), respectively. Assumption 7-(ii) also plays a role of under-smoothing and eliminates the first stage bias. Assumption 7-(iii) ensures the cross-product remainder terms of the Bahadur representations of \( \hat{m} \) and \( \hat{\pi} \) is negligible for the second stage estimation. Assumption 7-(iii) is implied by Assumption 7-(i) and (ii) when \( d_z \leq d_s + 2 \). Assumption 8 requires the model admits enough smoothness, depending on the dimension of the arguments of the unknown functions. Both the smoothness condition and the compactness of the support are needed for the uniform Bahadur representation. In the case where all variables are univariate, continuous and the degree of smoothness \( p = 1 \), the rate condition is satisfied if we choose \( h = n^{-r} \) for some \( r \in (1/6, 1/5) \). Assumption 9 is a technical condition that ensures that the (population) regression residual \( e \equiv Y - \mathbb{E}[Y|W] \) has certain finite order moments-- also assumed in Kong, Linton, and Xia (2010). When \( U_d \) or \( Y_d \) has a bounded support, Assumption 9 holds for any \( \nu_1 > 0 \), which is why we only need this assumption for the estimation of \( m \) but not \( \pi_d \).

Theorem 2. Suppose Assumptions 1 to 9 are satisfied, then for each \((s, x)\),
\[
\sqrt{nh^{d_s+d_x+2}} \left\{ \hat{\beta}(s, x) - \beta(s, x) \right\} \xrightarrow{d} N(0, V^{-1}\Omega_m V^{-1}),
\]
where matrices \( V \) and \( \Omega_m \) depends on \((s, x)\) and are defined in Lemma 3.

Proof. See Appendix A.2.

We can see from Theorem 2 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 4, 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 \( s \).
4. ILLUSTRATIVE EXAMPLE, DISCUSSIONS AND EXTENSIONS

4.1. An illustrative example. In this subsection we will discuss illustratively a linear model for which our method consistently estimate the parameter of interest while the popular two stage least square estimator fails to do so even with valid instrumental variables.

Analyzing the wage gap between the immigrants and the native population has always been an important economic question. In an influential study, Borjas (1987) pointed out that the difficulty of analyzing the wage gap results from the endogeneity of the immigration decision. For example, researchers usually believe that an representative immigrant individual tends to have higher talent than a representative native individual, since the immigration is likely the result of an important selection process which is usually based on individual’s talents. To address such endogeneity issue, one can consider a potential outcome model:

\[
\begin{align*}
Y_1 &= \beta_1 X' + \alpha S + U_1 \\
Y_0 &= \beta_0 X' + U_0
\end{align*}
\]

(8)

where \(Y_1\) and \(Y_0\) are the potential wages of an individual had him/her be an immigrant worker or native worker, respectively. \(X\) is a vector of demographic characteristics, \(U_d, d = 0, 1\), is the unobserved idiosyncratic term, and \(S\) is a variable that measures how immigrants adapt to their new country labor market. \(S\) is usually determined by political or culture factors and is likely to be independent \(U_1\).\(^6\) Consider an individual with observed characteristics \(X = x\) and suppose that the error terms satisfies \(E[U_d|X, S] = 0\) a.s., then the conditional treatment effect of immigration on wage can be written as

\[
E[Y_1 - Y_0|X = x, S = s] = (\beta_1 - \beta_0)x + \alpha s.
\]

Therefore, the parameter \(\alpha\) measures how the variable \(S\) affects the wage gap between immigrants and natives, which is often of interest to policy makers. For example if \(\alpha\) is large, policy makers can be interested in improving the assimilation process by giving more language courses or labor market training to new immigrants.

\(^6\)Different types of measures could be constructed depending on the data availability, see discussions about “assimilation” in Borjas (1987), who used the earning growth of an immigrant cohort (relative to natives) evaluated at ten years after immigration as a such measure.
Suppose that there is a vector of instrumental variables $Z$ which is independent of structural error $U_1$ and $U_0$ but correlated with the decision to immigrate. Then one possible attempt to estimate $\alpha$ is to consider two stage least square with the structural equation be specified as

$$Y = \gamma_0 X' + \gamma_1 X'D + \gamma_2 S D + \epsilon,$$

where $D = 1$ indicates the individual is an immigrant worker and $D = 0$ for native worker. If the true model is given by Equation (8), then $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 - \beta_0$, $\gamma_2 = \alpha$ and $\epsilon = (U_1 - U_0)D + U_0$. However, even $Z$ is a valid instrument, the 2SLS can not consistently estimate $\alpha$ because $Z$ is correlated with $\epsilon$ through $D$. Even with the additional monotonicity assumption in the first stage regression, the 2SLS estimand identifies the local average treatment effect, which in general is different from $\alpha$—as will be further elaborated in Section 4.2. On the other hand, our method can identify and consistently estimate $\alpha$ without assuming $Z$ be independent with $(U_0, U_1)$ and the monotonicity.

4.2. **Binary $D$ and binary $Z$.** In this section we will further discuss the point we made in the previous section in the framework of binary $D$ and binary $Z$. We abstract away $X$ to simplify notation. We also assume that $\mathbb{E}[U_d|S] = 0$ a.s. in this subsection.

Under Assumptions 1 to 3, one can follow Heckman and Vytlacil (2005) and define treatment effect for individuals with characteristics $(s, u_0, u_1)$ as $\Delta(s, u_0, u_1) = f_1(s) - f_0(s) + u_1 - u_0$ and the conditional average treatment effect $\Delta^{ATE}(s) = \mathbb{E}[\Delta(s, U_1, U_0)|S = s] = f_1(s) - f_0(s)$, where the expectation is taken over the conditional distribution of $(U_1, U_0)$ given $S = s$. Similarly, one can also define the conditional local average treatment effect of Imbens and Angrist (1994) as $\Delta^{LATE}(s) = \mathbb{E}[\Delta(s, U_1, U_0)|S = s, D_1 = 1, D_0 = 0]$, which is indeed the mean gain to persons who would be induced to switch from $D = 0$ to $D = 1$ if $Z$ were manipulated externally from 0 to 1. It is straightforward to see that our $\beta_1(s)$ estimand identifies the derivative of the conditional average treatment effect $\Delta^{ATE}(s)$ under Assumptions 1 to 3.

---

7The first possible set of instruments can be variables that measures the intensities of humanitarian crisis happens in the city of origin of the potential immigrant. For example, the bombing activity in Raqqa in Syria, the earthquake intensity per city in Nepal, the proportion of Ebola cases per city in Liberia among many others. This set of instruments is unlikely to be imperfect. Alternatively, one can considers instruments such as the amount of financial help or transfers received by the potential immigrant relatives. This later set of instruments may not be valid for usual 2SLS but poses no issue for our identification strategy.
Next we discuss the instrumental Variable (IV) estimand (also known as the Wald estimand). The IV estimand is proposed by Holland (1988) to identify causal effect in the presence of homogeneous treatment effect. Here we consider a generalized version of the IV estimand by conditioning on $S = s$, that is,

$$
\beta^{IV}(s) = \frac{Cov[Y, \pi_0(Z)|S = s]}{Cov[D, \pi_0(Z)|S = s]}
$$

In the case with no treatment-specific heterogeneity (i.e., $U_1 = U_0 = U$), Das (2005) shows that under the conditional mean independence assumption, $\Delta^{ATE}(s)$ is identified by $\beta^{IV}(s)$. The intuition of Das (2005)’s result can be interpreted as follows. When $U_0 = U_1$, the individual treatment effect $\Delta(s, U_1, U_0) = f_1(s) - f_0(s)$ becomes a deterministic function of $s$ and hence is homogeneous across individuals who share the same observed characteristics $s$. As a consequence, $\beta^{IV}(s)$ identifies the $\Delta^{ATE}(s)$ by extending Holland (1988)’s argument to conditioning on of $S = s$. One can show that by inserting $Y = DY_1 + (1 - D)Y_0$ and $Y_d = f_d(S) + U_d$,

$$
\beta^{IV}(s) = \frac{Cov[(f_1(s) - f_0(s))D, \pi_0(Z)|S = s]}{Cov[D, \pi_0(x, Z)|S = s]} + \frac{Cov[(U_1 - U_0)D + U_0, \pi_0(Z)|S = s]}{Cov[D, \pi_0(Z)|S = s]}
$$

where the last term on the right hand side is zero since $U_1 = U_0$. However, it is evident from Equation (9) that when $U_1 \neq U_0$, the treatment effect becomes heterogeneous and $\beta^{IV}(s)$ can no longer identify $\Delta^{ATE}(s)$, as pointed out by Heckman and Vytlacil (2005).8

Furthermore, the derivative of $\beta^{IV}(s)$ does not identify the derivative of $\Delta^{ATE}(s)$ either—this is even when the two-way exclusive regressors $(S, Z)$ are available and $S$ being independent of $(U, V)$. There are two possible cases in which $\frac{d\beta^{IV}(s)}{ds}$ identifies $\frac{d\Delta^{ATE}(s)}{ds}$. One is when $U_1 = U_0$ and the other is when $S$ and $Z$ are independent. In both cases, the last term of Equation (9) is a flat function in $s$ and therefore disappears after taking derivatives.

To summarize, if we allow that $U_1 \neq U_0$, neither $\beta^{IV}(s)$ estimand nor its derivative identifies interesting policy parameters without making additional assumptions. Our $\beta(s)$ estimand, on the other hand, identifies the derivative of $\Delta^{ATE}(s)$ and hence identifies $\Delta^{ATE}(s)$ itself under a suitable location normalization by exploring the treatment-exclusive variable $S$.

8If one imposes an additional monotonicity assumption (i.e. conditional on $S = s$, $D_1 \geq D_0$ a.s.), then $\beta^{IV}(s)$ identifies $\Delta^{LATE}(s)$, see Imbens and Angrist (1994).
4.3. **Estimation without support** Assumption 4. Recall in Equation (4), we have

\[ m(x, z, s) = \pi(x, z)'\beta(s, x), \quad \forall (s, x, z) \in \mathcal{W}, \]

which implies

\[ \mathbb{E}[\pi(X, Z)m(X, Z, S)|X = x, S = s] = \mathbb{E}[\pi(X, Z)\pi(X, Z)|X = x, S = s]'\beta(s, x), \quad \forall (x, s). \tag{10} \]

In this case, we can define the second step estimator as

\[
\hat{\beta}(s, x) = \left( \frac{1}{nh^{d_i+d_s}} \sum_{i} \hat{\pi}(X_i, Z_i) \hat{\pi}(X_i, Z_i)'\Psi_h(X_i - x, S_i - s) \right)^{-1} \\
\times \left( \frac{1}{nh^{d_i+d_s}} \sum_{i} \hat{\pi}(X_i, Z_i)m(S_i, X_i, Z_i)\Psi_h(X_i - x, S_i - s) \right),
\]

where \( \Psi_h(\cdot) = \Psi(\cdot/h) \) is a kernel function.

4.4. **Estimation of the ATE derived in Corollary 1.** The ATE derived in Corollary 1 can be estimated using our above estimator of \( \beta_d \). Let \( f_s \) be the density of \( S \), then

\[
\Delta_{d,d'}(s, x) = \int_{s^*}^{s} (\beta_d(t, x) - \beta_{d'}(t, x))dt \\
= \int_{s^*}^{s} \frac{1}{f_s(t)} \frac{1}{\mathbb{E}} \left[ \mathbb{I}\{s^* \leq t \leq s\} (\beta_d(S_i, x) - \beta_{d'}(S_i, x)) \right] dt.
\]

Then \( \hat{\Delta}_{d,d'}(s, x) \) can be estimated by the sample analog and with a plug-in estimator \( \hat{\beta} \)

\[
\hat{\Delta}_{d,d'}(s, x) = \frac{1}{n} \sum_{i} \frac{1}{f_s(S_i)} \mathbb{I}\{s^* \leq S_i \leq s\} \times \left\{ \hat{\beta}_d(S_i, x) - \hat{\beta}_{d'}(S_i, x) \right\}.
\]

Alternatively, one can also estimate \( \beta_d(\cdot, x) - \beta_{d'}(\cdot, x) \) over fine grids of the interval \([s^*, s]\) and conduct numerical integration to obtain \( \hat{\Delta}_{d,d'}(s, x) \) to avoid the possible trimming.

5. **Conclusion**

In this paper we discussed identification and estimation of nonparametric structural function in heterogeneous treatment effect models with a discrete endogenous treatment. We focus on applications
in which the potential outcome function is additive-separable to a treatment-specific structural error and there exist two-way exclusion variables: this frees us from imposing any form of monotonicity assumptions or other functional forms on the selection equation. We also provided a two-stage nonparametric estimators for the parameter of interest.
REFERENCES


APPENDIX A. PROOFS TO THE MAIN RESULTS

A.1. Proof to Theorem 1. The idea of proof is similar to Das (2005). To show the sufficiency, the identification Equation (4) implies that almost surely with respect to the conditional distribution of \( Z \) given \((S, X) = (s, x)\),

\[
m(s, x, Z) = \pi(x, Z)\beta(s, x),
\]

which in turn implies

\[
\pi(x, Z)m(s, x, Z) = \pi(x, Z)\pi(x, Z)\beta(s, x).
\]

Taking expectation with respect to the conditional distribution of \( Z \) gives

\[
E[\pi(x, Z)m(s, x, Z)|S = s, X = x] = E[\pi(x, Z)\pi(x, Z)\beta(s, x)|S = s, X = x].
\]

Since \( \nabla[\pi_0(x, Z)|S = s, X = x] \) is positive definite and

\[
\nabla[\pi_0(x, Z)|S = s, X = x] = E[\pi(x, Z)\pi(x, Z)|S = s, X = x].
\]

we can obtain Equation (5) by pre-multiply \( \nabla^{-1} \) to both sides.

The necessity holds because when \( \nabla[\pi_0(x, Z)|S = s, X = x] \) does not have full rank, there are multiple \( \beta \) satisfying Equation (4) a.s. in \( Z|(S, X) = (s, x) \).

A.2. Proof to Theorem 2. Let \( K_h(\cdot) = K(\cdot/h) \). By Lemma 1 (notation defined therein), we have uniformly in \( w \),

\[
\hat{m}(w) = m(w) - \frac{1}{nh^{d+2}} \sum_{n,m}^{(2,2)} \sum_{i=1}^{n} K_h(S_i - s)K_h(X_i - x)K_h(Z_i - z)\epsilon_i^m(S_i - s)
\]

\[
\underbrace{\eta_{n,m}(w)} + O_p(h^{p+1}) + O_p \left( \frac{\log n}{nh^{d+1}} \right).
\]

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 \( \pi_0 = [\pi_1, \ldots, \pi_T] \); likewise we use \( \hat{\pi} \) to denote a generic element in \( \hat{\pi} \). Then Lemma 2 shows that uniformly in \((x, z)\),

\[
\hat{\pi}(x, z) = \pi(x, z) - \frac{1}{nh^{d+\pi}} \sum_{n,i}^{(1,1)} \sum_{i=1}^{n} K_h(X_i - x)K_h(Z_i - z)\epsilon_i^\pi
\]

\[
\underbrace{\eta_{\pi,n}(x, z)} + O_p(h^{p+1}) + O_p \left( \frac{\log n}{nh^{d+\pi}} \right).
\]
Recall that our estimator is defined as

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

First consider the denominator; it is easy to see that under the assumptions of Theorem 2 and the representation in Equation (12),

\[ \frac{1}{n} \sum \hat{\pi}(x, Z_i) \hat{\pi}'(x, Z_i) \xrightarrow{p} \mathbb{E}[\pi(x, Z_i) \pi'(x, Z_i)] = V. \]

For the numerator, we consider the following decomposition,

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

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 \( \eta_{\pi,n}(Z_i) \).

\[
\frac{1}{n} \sum \hat{\pi}(x, Z_i) \hat{m}(s, x, Z_i) - \frac{1}{n} \sum \pi(x, Z_i)m(s, x, Z_i) = \frac{1}{n} \sum \pi(x, Z_i) \{ \eta_{m,n}(Z_i) + r_{m,1} + r_{m,2} \} + \frac{1}{n} \sum \{ \eta_{\pi,n}(Z_i) + r_{m,1} + r_{m,2} \} \{ \eta_{\pi,n}(Z_i) + r_{\pi,1} + r_{\pi,2} \}
\]

The third and fourth RHS terms are of order smaller than \( O_p(1/\sqrt{nh^{d_1+d_2+2}}) \) by Lemma 4. The last RHS term is of order smaller than \( O_p(1/\sqrt{nh^{d_1+d_2+2}}) \) by Lemma 6. By Lemma 3,

\[
\sqrt{nh^{d_1+d_2+2}} \left\{ \frac{1}{n} \sum \pi(x, Z_i) \eta_{m,n} + \frac{1}{n} \sum m(s, x, Z_i) \eta_{\pi,n} \right\} \xrightarrow{d} N(0, \Omega_m).
\]

It then follows that

\[
\sqrt{nh^{d_1+d_2+2}} \left\{ \frac{1}{n} \sum \pi(x, Z_i) \hat{m}(s, x, Z_i) - \mathbb{E}[\pi(x, Z_i)m(s, x, Z_i)] \right\} \xrightarrow{d} N(0, \Omega_m),
\]

or alternatively,

\[
\sqrt{nh^{d_1+d_2+2}} \left\{ \hat{\beta}(s, x) - \beta(s, x) \right\} \xrightarrow{d} N(0, V^{-1} \Omega_m V^{-1}).
\]
APPENDIX B. AUXILIARY LEMMAS

Lemmas 1 and 2 adopt the results from Kong, Linton, and Xia (2010, KLX). We define some notation first. For the purpose of exposition, we define notation for estimation of \( m \). The notation for estimation of \( \tau_d \) is similar. For \( j = 0, 1, \ldots, p \), let \( N_j \) be the number of \( d_m \)-dimensional vectors \( \bar{x} \) such that \( |\bar{x}| = j \). Arrange all such vectors in the lexicographical order with the first one is \((j, 0, \ldots, 0)\) and last one is \((0, \ldots, 0, j)\). Let \( \tau_j \) be the one to one mapping from an order to the associated vector. For \( j = 0, \ldots, p \), let \( v_{n,m,j}(w) = \int K(u)u_j g(w + hu) f_w(w + hu) du \), where here \( u \) is a \( d_m \times 1 \) vector and \( u_j \) stands for the product of powers of elements of \( u \) such that the sum of power index equals to \( j \). The function \( g \) is defined as in KLX-Equation A7. Since we use quadratic loss function, \( g(w) = 1 \) for any \( w \in \mathcal{W} \) in our case. Let \( \Sigma_{n,m} \) be a symmetric matrix

\[
\Sigma_{n,m} = \begin{pmatrix}
\Sigma_{n,m,0,0} & \Sigma_{n,m,0,1} & \cdots & \Sigma_{n,m,0,p} \\
\Sigma_{n,m,1,0} & \Sigma_{n,m,1,1} & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{n,m,p,0} & \Sigma_{n,m,p,1} & \cdots & \Sigma_{n,m,p,p}
\end{pmatrix},
\]

where \( \Sigma_{n,m,i,j} \) is an \( N_j \) by \( N_j \) matrix whose \((\ell, k)\) element is \( v_{n,m,\tau_{j}(\ell)+\tau_{j}(k)} \). So \( \Sigma_{n,m,0,0} \) is the \((1, 1)\)th element of the matrix \( \Sigma_{n,m} \) and equals to \( v_{n,m,0} \Sigma(\mathcal{W}) \); \( \Sigma_{n,m,0,1} \) is a \( 1 \times d_m \) vector contains terms of \( v_{n,m,0} \) 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 \( v_{n,m,2} \) 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 \( v_{m,j} = f_w(w) \int K(u)u_j du \) replacing \( v_{n,m,j} \).

Lemma 1. Let \( \hat{m} \) be the estimator specified in Equation (7), then under Assumptions 5 to 9, we have

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

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

\[
m^*_n(w) = -\frac{1}{nh^{d_m}} \sum_{i=1}^{n} K_h(S_i - s) K_h(X_i - x) K_h(Z_i - z) \left[ Y_i - \sum_{0 \leq |z^m| \leq p} \alpha_{z^m} \cdot (W_i - w)^{z^m} \right] (S_i - s)
\]

where \( \eta^m = \Sigma_{n,m}^{(2,2)} \Sigma^{-1}_{n,m} \), \( d_m \) is dimension of \( w \), \( \alpha^m \) is a vector of values which minimizes the population objective function, \( \Sigma_{n,m}^{(2,2)} \) is the second diagonal element of \( \Sigma_{n,m}^{-1} \). Furthermore, uniformly over \( w \), \( m^*_n(w) \) can be written as

\[
m^*_n(w) = -\frac{1}{nh^{d_m+2}} \sum_{i=1}^{n} K_h(S_i - s) K_h(X_i - x) K_h(Z_i - z) e^m_i (S_i - s) + O_p(h^{p+1})
\]

where \( e^m_i = Y_i - \mathbb{E}[Y_i|W_i] \).

\(^9\)The matrices \( \Sigma_{n,m} \) and \( \Sigma_m \) are defined in the same way as \( S_{n,p} \) and \( S_p \) in KLX, respectively.
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.

KLX-A1 part 1 holds since we consider the quadratic loss function. KLX-A1 part 2 holds by Assumption 9. 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 6. KLX-A4 is the smoothness assumption on the joint distribution of $(S, X, Z)$, it holds by Assumption 8-(i). KLX-A5 is the smoothness assumption on $m$ and is satisfied by Assumption 8-(ii). KLX-A7 part 1 is ensured by Assumption 8-(ii) and part 2 holds since we have i.i.d. observations.

It remains to verify KLX-A6. For two sequences $a_n$ and $b_n$, we use $a_n \succ b_n$ to denote $b_n/a_n \xrightarrow{p} 0$. Similarly define “$\prec$”. First, since $nh^{d_m} \succ nh^{d_m+2}$ and $nh^{d_m+2} \to \infty$ by Assumption 7-(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 $nh^{d_m+2(p+1)} \prec nh^{d_m+2(p+1)}$ and $nh^{d_m+2(p+1)} \to 0$ by Assumption 7-(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 $M > 2$,

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

Hence we have

$$n^{-1} \{r(n)\}^{2/3} d_n \log n \{M_n^{(2)}\}^{-1} = M^{-1/4} \{\log n\}^2 n^{-1} \gamma_n^{-\frac{3}{4}} = M^{-1/4} \{\log n\}^2 n^{-1} \gamma_n^{-\frac{3}{4}}.$$ 

where we take $\nu_2 = \nu_1$ and $\nu_1$ is a constant satisfying Assumption 9, then the above quantity diverges to infinity. Equation A.3 and A.4 of KLX-A6 are satisfied since by the i.i.d. observation Assumption 5, the mixing coefficient $\gamma[k] = 0$ for all $k \geq 1$.

Next we show the displayed Equation (13) in the statement of this Lemma. By definition of $\eta_{m,n}$, we can write $m_n^*(w)$ as

$$m_n^*(w) = \eta_{m,n}(w) + \frac{1}{nh^{d_m}} \left( \sum_{i=1}^{n} K_h(S_i - s)K_h(X_i - x)K_h(Z_i - z) \right)$$ 

$$\times \left[ \mathbb{E}[Y_i|W_i] - \sum_{0 \leq l^m \leq p} \alpha_{l^m} \cdot (W_i - w)l^m \right] (S_i - s).$$

22
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 \( n_h \):

\[
e_n \equiv \frac{1}{h^{d \pi + 1}} \mathbb{E} \left[ K_h(S_i - s)K_h(X_i - x)K_h(Z_i - z)(S_i - s) \left[ \mathbb{E}[Y_i|W_i] - \sum_{0 \leq |\zeta| \leq p} \alpha_{\zeta}^m : (W_i - w)\zeta \right] \right]
\]

For a generic vector \( C_i \), let \( u_i = \frac{C_i - c}{n_h} \). Conducting changing variable we have

\[
e_n = \int |K(u_s)K(u_x)K(u_z)u_s| \times M(u_s + hs, u_x + hx, u_z + hz) - \sum_{0 \leq |\zeta| \leq p} \alpha_{\zeta}^m : (hu_w)\zeta \right| g_w(u_w + hw)du_idu_xdu_z,
\]

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

Let \( \Sigma_{n, \pi} \) and \( \Sigma_{\pi} \) be two \( (d_\pi + 1) \times (d_\pi + 1) \) matrices defined analogously to \( \Sigma_{n,m} \) and \( \Sigma_m \), respectively. Let \( \hat{W} = (X, Z) \) and \( \hat{\pi} \) be its support.

**Lemma 2.** Suppose that Assumptions 5 to 9 hold, then uniformly over \( \hat{\pi} \),

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

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

\[
\pi_n^*(x, z) = -\frac{1}{n h^{d_\pi}} \sum_{n, \pi}^{-(1, 1)} \sum_{i=1}^{n} K_h(X_i - x)K_h(Z_i - z) \left[ 1\{D_i = d\} - \sum_{0 \leq |\zeta| \leq p} \alpha_{\zeta}^m : (\hat{W}_i - w)\zeta \right]
\]

where \( \Sigma_{n, \pi}^{-1} \) is the first diagonal element of \( \Sigma_{n, \pi}^{-1} \). Furthermore, uniformly in \( (x, z) \), \( \pi_n^*(x, z) \) can be written as

\[
\pi_n^*(x, z) = -\frac{1}{n h^{d_\pi}} \sum_{n, \pi}^{-(1, 1)} \sum_{i=1}^{n} K_h(X_i - x)K_h(Z_i - z)e_i^\pi + O_p(h^{p+1}),
\]

where for a generic \( d \), \( e_i^\pi = 1\{D_i = d\} - \mathbb{P}(D_i = d|X_i, Z_i) \).

**Proof.** The proof follows the similar argument as in Lemma 1. We verify KLX-A6; the other assumptions (A1-A5 and A7) can be verified analogously to Lemma 1. Since \( nh^{d_\pi} \gg nh^{d_m + 2} \rightarrow \infty \) 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 \( nh^{d_\pi + 2(p+1)} \rightarrow 0 \) by Assumption 7-ii. To verify the third condition in the Equation A.2, let \( \gamma_m = nh^{d_\pi} \log n \), as before, we have for some \( M > 2 \),

\[
d_n = \gamma_m^{-\frac{1}{4} + \frac{1}{2} + \frac{1}{4}} \log n = \gamma_m^{-\frac{1}{4}} \log n, \quad r(n) = \gamma_n^{\frac{1}{2}}
\]
\[ M_n^{(1)} = M \gamma_n^{-1}, \quad M_n^{(2)} = M^{1/2} \gamma_n^{-1/2} \]

Hence we have

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

Note that \( \mathbb{E}[e^{\pi}]^{\nu_1} < \infty \) for any \( \nu_1 \) since \( e^{\pi} \) is bounded; then the above quantity diverges to infinity by letting \( \nu_2 \) arbitrarily large.

\[ \Box \]

**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 (13) and (14), respectively. Suppose that the assumptions of Theorem 2 are satisfied, then

\[ \frac{\sqrt{nh^d}}{n} \sum \pi(x, Z_1) \eta_{m,n}(s, x, Z_1) \overset{d}{\rightarrow} N(0, \Omega_m). \]

where \( \Omega_m \) is a \( T \times T \) positive definite matrix

\[ \Omega_m = \int \left\{ \sigma_m^2(x, s, Z_1) \left( \sum_{i=1}^{n} (2,2) \eta_{m,n}(s, x, Z_i) \right)^2 K^2(u_x)K^2(u_x)u^2_{x}g^2_{x}(Z_1)\pi(x, Z_1)\pi'(x, Z_1) \right\} g_w(x, s, Z_1)du_xdu_xdZ_1. \]

and \( \sigma_m^2(w) = \mathbb{V}[e^m|W = w]. \)

For a generic element \( \pi \) in \( \pi_0 = [\pi_1, ..., \pi_T]' \), there is

\[ \frac{\sqrt{nh^d}}{n} \sum \pi(x, s, Z_i) \eta_{\pi,n}(x, Z_i) \overset{d}{\rightarrow} N(0, \Omega_{\pi}), \]

\( \Omega_{\pi} \) is a positive scalar such that

\[ \Omega_{\pi} \equiv \int \left\{ \sigma_{\pi}^2(x, Z_1) \left( \sum_{i=1}^{n} (1,1) \eta_{\pi,n}(x, Z_i) \right)^2 K^2(u)g^2_{x}(Z_1)m^2(s, x, Z_1) \right\} g_w(s, x, Z_1)du_xdZ_1. \]

and \( \sigma_{\pi}^2(x, z) = \mathbb{V}[e^{\pi}|(X, Z) = (x, z)]. \)

**Proof.** Recall that

\[ \eta_{m,n}(w) = -\frac{1}{nh^d + 2} \sum_{i=1}^{n} \frac{w}{(2,2)} K_h(S_i - s)K_h(X_i - x)K_h(Z_i - z)e^{\pi}(S_i - s), \]
where \( e_m^n = Y_j - \mathbb{E}[Y_j | W_j] \). Let \( d_z \) be the dimension of \( Z \), then since \( K(0) = 0 \),

\[
\frac{\sqrt{n h^{d_u+2-d_z}}}{\sqrt{n h^{d_u+2}}} \sum_i \pi(x, Z_i) \eta_{m,n}(Z_i) = \frac{1}{n \sqrt{h^{d_z}}} \sum_i \sum_{j \neq i} \pi(x, Z_i) \hat{\Sigma}_{m2}(Z_i) K_h(Z_j - Z_i) K_h(X_j - x) K_h(S_j - s) e_m^n(S_j - s),
\]

where we abbreviate \( \Sigma_m^{-2,(s)}(s, x, Z_i) \) as \( \hat{\Sigma}_{m2}(Z_i) \); we also abbreviate \( \Sigma_m^{-2,(s)}(s, x, Z_i) \) as \( \Sigma_{m2}(Z_i) \). Since \( \hat{\Sigma}_{m2} \) are uniformly consistent by KLX-Lemma 8, we can replace it by \( \Sigma_{m2}(Z_i) \) in the rest of the analysis without affecting the limiting distribution of our estimator. Define

\[
\psi_{ij}^* = \pi(x, Z_i) \Sigma_{m2}(Z_i) K_h(Z_j - Z_i) K_h(X_j - x) K_h(S_j - s) e_m^n(S_j - s),
\]

and rewrite \( \psi_{ij} = \frac{1}{2} (\psi_{ij}^* + \psi_{ij}^{**}) \). Therefore we can write

\[
\frac{\sqrt{n h^{d_u+2-d_z}}}{\sqrt{n h^{d_u+2}}} \sum_i \pi(x, Z_i) \eta_{m,n}(Z_i) = -\frac{\sqrt{n}}{n^2} \sum_i \sum_{j \neq i} \psi_{ij} \sqrt{h^{d_z}} \sqrt{h^{d_u+2}},
\]

where the right hand side is proportional to a U-statistics. It is easy to verify that \( \mathbb{E}[\psi_{ij}] = 0 \) since \( \mathbb{E}[e_m^n | W_j] = \mathbb{E}[e_m^n | W_i] = 0 \). To derive the limiting distribution, it remains to find the variance. Let \( \bar{\psi}_1 = \mathbb{E}[\psi_{12} | W_i, Y_1] = \frac{1}{2} \{ \mathbb{E}[\psi_{12} | W_i, Y_1] + \mathbb{E}[\psi_{21} | W_i, Y_1] \} \), the limiting variance is \( 4 \mathbb{V}(\bar{\psi}_1) / (h^{d_z+d_u+2}) \).

Note first since \( \mathbb{E}[K_h(Z_2 - Z_1) e_m^n(S_2 - s) | W_i, Y_1] = \mathbb{E}[K_h(Z_2 - Z_1)(S_2 - s) e_m^n | W_2, W_i, Y_1] | W_i, Y_1] = \mathbb{E}[K_h(Z_2 - Z_1)(S_2 - s) e_m^n | W_2, W_i, Y_1] = 0 \), we have

\[
\mathbb{E}[\psi_{12} | W_i, Y_1] = \mathbb{E}[\pi(x, Z_1) \Sigma_{m2}(Z_1) K_h(Z_2 - Z_1) K_h(X_1 - x) K_h(S_1 - s) e_m^n(S_2 - s) | W_i, Y_1] = 0.
\]

and therefore

\[
2\bar{\psi}_1 = \mathbb{E}[\psi_{21} | W_i, Y_1]
\]

\[
= \mathbb{E}[\pi(x, Z_2) \Sigma_{m2}(Z_2) K_h(Z_2 - Z_1) K_h(X_1 - x) K_h(S_1 - s) e_m^n(S_1 - s) | W_i, Y_1]
\]

\[
= e_m^n(S_1 - s) K_h(X_1 - x) K_h(S_1 - s) \mathbb{E}[\pi(x, Z_2) \Sigma_{m2}(Z_2) K_h(Z_2 - Z_1) | W_i, Y_1]
\]

\[
= (i) e_m^n(S_1 - s) K_h(X_1 - x) K_h(S_1 - s) \int \pi(x, Z_2) \Sigma_{m2}(Z_2) K_h(Z_2 - Z_1) g_z(Z_2) dZ_2
\]

\[
= (ii) h^{d_z} e_m^n(S_1 - s) K_h(X_1 - x) K_h(S_1 - s) \int \pi(x, Z_2 + hu) \Sigma_{m2}(Z_2 + hu) K(u) g_z(Z_2 + hu) du
\]

\[
= (iii) h^{d_z} e_m^n(S_1 - s) K_h(X_1 - x) K_h(S_1 - s) \{ \pi(x, Z_1) \Sigma_{m2}(Z_1) g_z(Z_1) + o(h) \},
\]

25
where (i) holds because i.i.d. observations; (ii) holds by changing variable $u = (Z_2 - Z_1)/h$, and (iii) holds by the continuous differentiability of the integrand (implied by Assumption 8) and the assumption that the support of the kernel is bounded. So the dominant term is

$$2\hat{p}_1 \approx h^{d_x}e_1^m (S_1 - s) K(X_1 - x/h) K(S_1 - s/h) \pi(x, Z_1) g_z(Z_1) \Sigma_m(Z_1).$$

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

$$4\mathcal{V}(\hat{p}_1) = 4\mathbb{E}[\hat{p}_1 \hat{p}_1']$$

$$\approx h^{d_x} \mathbb{E} \left[ (e_1^m (S_1 - s) K((X_1-x)/h) K((S_1-s)/h) g_z(Z_1))^2 \pi(x, Z_1) \pi'(x, Z_1) \right]$$

$$\overset{\text{(i)}}{=} h^{2d_x} \mathbb{E} \left[ \sigma_m^2(W_1) ((S_1 - s) K((X_1-x)/h) K((S_1-s)/h) g_z(Z_1))^2 \pi(x, Z_1) \pi'(x, Z_1) \right]$$

$$\overset{\text{(ii)}}{=} h^{2d_x+2d_1} \int \left\{ \sigma_m^2(x, s, Z_1) \Sigma_m^2(Z_1) K^2(u_x) K^2(u_s) g_z^2(Z_1) \pi(x, Z_1) \pi'(x, Z_1) \right\} g_w(x, s, Z_1) du_s du_x dZ_1$$

$$\overset{\text{(iii)}}{=} \left( h^{2d_x+2d_1+d_2+d_z} \right) \int \left\{ \sigma_m^2(x, s, Z_1) \Sigma_m^2(Z_1) K^2(u_x) K^2(u_s) g_z^2(Z_1) \pi(x, Z_1) \pi'(x, Z_1) \right\} g_w(x, s, Z_1) du_s du_x dZ_1$$

where (i) holds by taking the conditional expectation of $e_1^m$ given $W_1$; (ii) holds by multiplying and dividing $h^{2d_x}$; (iii) holds by changing variable $u_x = (X_1 - x)/h$, $u_s = (S_1 - s)/h$, and ingoing higher order terms, and (iv) holds because $d_s = 1$ and $d_m = d_z + d_x + d_s$. Hence we know that

$$4\mathcal{V} \left( \frac{\hat{p}_1}{\sqrt{h^{d_x} h^{d_1}}} \right) = \Omega_m,$$

where

$$\Omega_m = \int \left\{ \sigma_m^2(x, s, Z_1) \Sigma_m^2(Z_1) K^2(u_x) K^2(u_s) g_z^2(Z_1) \pi(x, Z_1) \pi'(x, Z_1) \right\} g_w(x, s, Z_1) du_s du_x dZ_1.$$

By the standard U statistics theory, we have

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

Next we consider the term $\frac{\sqrt{n h^{d_m-d_z}}}{n} \sum_i m(x, s, Z_i) \eta_{s,n}(x, Z_i)$, we can write:

$$\frac{\sqrt{n h^{d_m+2-d_z}}}{n} \sum_i m(x, s, Z_i) \eta_{s,n}(Z_i)$$

$$= - \frac{1}{n \sqrt{h^{d_x} h^{d_m+2}}} \sum_{i \neq j} m(x, s, Z_i) \Sigma_{x1}(Z_i) K_h(Z_j - Z_i) K_h(X_j - X) K_h(S_j - S) e_j^\pi.$$
Following similar argument, we can show that
\[
\frac{\sqrt{n}h^{d+1-d_2}}{n} \sum_i m(x, s, Z_i) \eta_{\pi,n}(Z_i) \to N(0, \Omega_{\pi}),
\]
where
\[
\Omega_{\pi} \equiv \int \left\{ \sigma_{\pi}^2(x, Z_1) \Sigma_{\pi 1}(Z_1) \mathbb{K}^2(u) g^2_Z(Z_1) m^2(s, x, Z_1) \right\} g_w(s, x, Z_1) dudZ_1.
\]
where \(\sigma^2_{\pi}(x, z)\) is the conditional variance of \(1[D = d] - \mathbb{P}[D = d | X = x, Z = z]\).

Since \(d_m + 2 > d_\pi\), it follows that \(\frac{\sqrt{n}h^{d+m+1-2-d_2}}{n} \sum_i m(x, s, Z_i) \eta_{\pi,n} \to 0\). For the same reason, the asymptotic covariance between \(\frac{\sqrt{n}h^{d+m+1-2-d_2}}{n} \sum_i \pi(x, Z_i)\) and \(\frac{\sqrt{n}h^{d+2-2-d_2}}{n} \sum_i m(x, s, Z_i) \eta_{\pi,n}\) converges in probability to zero as well. This establishes the result. \(\square\)

**Lemma 4.** Let \(\kappa_n = \sqrt{n}h^{d+m+1-2-d_2}\). Suppose that the assumptions of Theorem 2 are satisfied, then
\[
\frac{1}{n} \sum_i m(s, x, Z_i) \{r_{\pi,1} + r_{\pi,2}\} = o_p(\kappa_n^{-1}),
\]
\[
\frac{1}{n} \sum_i \pi(x, Z_i) \{r_{m,1} + r_{m,2}\} = o_p(\kappa_n^{-1}).
\]

**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. \(\square\)

**Lemma 5.** Suppose that the assumptions of Theorem 2 are satisfied and let \(\eta_{m,n}(w)\) and \(\eta_{\pi,n}(x, z)\) be as defined in Equations (11) and (12), then for a generic element \(\pi \in \pi_0\), then there exists \(\lambda_m\) and \(\lambda_{\pi}\) such that
\[
\frac{\sqrt{n}h^{d+m+1-2-d_2}}{n} \sum_i \eta_{m,n}(s, x, Z_i) \to N(0, \lambda_m),
\]
\[
\frac{\sqrt{n}h^{d+2-2-d_2}}{n} \sum_i \eta_{\pi,n}(x, Z_i) \to N(0, \lambda_{\pi}).
\]

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

**Lemma 6.** Let \(\kappa_n = \sqrt{n}h^{d+m+1-2-d_2}\). Suppose that the assumptions of Theorem 2 are satisfied, then
\[
T_n \equiv \frac{1}{n} \sum_i \{\eta_{m,n}(s, x, Z_i) + r_{m,1} + r_{m,2}\} \{\eta_{\pi,n}(x, Z_i) + r_{\pi,1} + r_{\pi,2}\} = o_p(\kappa_n^{-1}).
\]

**Proof.** \(T_n\) can be decomposed as the following four terms,
\[
T_n = \frac{1}{n} \sum_i \eta_{m,n}(s, x, Z_i) \eta_{\pi,n}(x, Z_i) + (r_{\pi,1} + r_{\pi,2}) \frac{1}{n} \sum_i \eta_{m,n}(s, x, Z_i)
+ (r_{m,1} + r_{m,2}) \frac{1}{n} \sum_i \eta_{\pi,n}(x, Z_i) + (r_{\pi,1} + r_{\pi,2})(r_{m,1} + r_{m,2}).
\]

The RHS4 is of \(O_p \left( h^{p+1} \right) + O_p \left( \frac{\log n}{\sqrt{n}h^{d+m+1}} \right) \) times \(O_p \left( h^{p+1} \right) \) + \(O_p \left( \frac{\log n}{\sqrt{n}h^{d+m+1}} \right) \). It is easy to verify that it is of order \(o_p \left( \kappa_n^{-1} \right)\) under Assumption 7. 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 (15) is also of order \( o_p(\kappa_n^{-1}) \). We write \( \hat{\Sigma}_{m,\tau}(Z_i) = \Sigma_{n,m}^{(2,2)}(Z_i)\Sigma_{n,\tau}^{−(1,1)}(Z_i) \). As before, we replace \( \hat{\Sigma}_{m,\tau}(Z_i) \) by its population counterpart \( \Sigma_{m,\tau}(Z_i) \) in the following analyze since it is does not affect the asymptotic results. Since \( K(0) = 0, \)

\[
\bar{U}_n \equiv \frac{1}{n} \sum_{i} \eta_{m,n}(s, x, Z_i) \eta_{\tau,n}(x, Z_i) \approx \frac{1}{h^3I_d^m+2I_d^\tau} \times \sum_{i=1}^{n} \sum_{j \neq i} \sum_{t \neq i} \frac{\mathbf{\Sigma}_{m,\tau}(Z_i) K_h(S_j - s) K_h(X_j - x) K_h(X_i - x) K_h(Z_j - Z_i) K_h(Z_i - Z_i)(S_j - s) e_i^m e_j^\tau}{\hat{\psi}_{ijt}}.
\]

(16)

where we write \( \approx \) to denote same distribution in the limit since replacing \( \hat{\Sigma}_{m,\tau}(Z_i) \) by its population counterpart \( \Sigma_{m,\tau}(Z_i) \) does not affect the limiting distribution. Depending on whether \( t = j \) or not, we further decompose \( \bar{U}_n \) into two parts:

\[
\bar{U}_{1n} = \frac{1}{h^3I_d^m+2I_d^\tau} \times \sum_{i=1}^{n} \sum_{j \neq i} \sum_{t \neq i} \frac{\mathbf{\Sigma}_{m,\tau}(Z_i) K_h(S_j - s) K_h(X_j - x) K_h(X_i - x) K_h(Z_j - Z_i) K_h(Z_i - Z_i)(S_j - s) e_i^m e_j^\tau}{\hat{\psi}_{ijt}}
\]

\[
\bar{U}_{2n} = \frac{1}{h^3I_d^m+2I_d^\tau} \times \sum_{i=1}^{n} \sum_{j \neq i} \sum_{t = j} \frac{\mathbf{\Sigma}_{m,\tau}(Z_i) K_h(S_j - s) K_h(X_j - x) K_h(X_i - x) K_h(Z_j - Z_i) K_h(Z_i - Z_i)(S_j - s) e_i^m e_j^\tau}{\hat{\psi}_{ijt}}
\]

so \( \bar{U}_n = \bar{U}_{1n} + \bar{U}_{2n} \). Further write

\[
\bar{U}_{1n} = \frac{n(n - 1)(n - 2)\sqrt{I_d^m+I_d^\tau}}{n^3I_d^m+2I_d^\tau} \times \frac{1}{n(n - 1)(n - 2)} \times \sum_{i=1}^{n} \sum_{j \neq i} \sum_{t \neq i} \frac{\psi_{ijt}^n}{\sqrt{I_d^m+I_d^\tau}}
\]

where \( \psi_{ijt}^n = \frac{\hat{\psi}_{ijt}}{\sqrt{I_d^m+I_d^\tau}} \). The \( \bar{U}_{1n} \) term is proportional to a third order \( U \)-statistics with kernel function \( \psi_{ijt}^n \). Let \( \psi_{ijt} \) be a symmetric transformation of \( \psi_{ijt}^n \), that is, \( \psi_{ijt} = \frac{1}{6} \sum_{p} \psi_{ijt}^p \) where \( \sum_{p} \) is the sum over all permutations of \( i, j, t \). Write \( H_i = (W_i, Y_i, D_i) \). It is straightforward 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 \( U_{1n} \) is a degenerated \( U \)-statistics. In the mean time, \( \mathbb{E}[\psi_{123}^*] \) is of the same order of \( \mathbb{E}[(\psi_{123}^*)^2] \), which is

\[
\mathbb{E}[(\psi_{123}^*)^2] = \frac{1}{h^2I_d^m+d_s} \mathbb{E}[(\xi_{123}^*)^2] = \frac{h^2I_d^m+d_s}{h^2I_d^m+d_s} \times \int \sum_{m,\tau}(Z_1) K_h^2(S_2 - s) K_h^2(X_2 - x) K_h^2(X_3 - x) K_h^2(Z_2 - Z_1) K_h^2(Z_3 - Z_1)((S_2 - s) / h)^2(e_2^m e_3^\tau)^2 g(\ldots) d\ldots \approx \frac{h^2I_d^m+d_s}{h^2I_d^m+d_s} \int [::g(\ldots) d\ldots = \int [::g(\ldots) d\ldots < \infty
\]
where \( \mathcal{g}(\ldots) \) denote the joint density of all random variables. (i) holds by multiplying and dividing \( h^{2d_z} \) inside and outside of the integration, respectively; (ii) holds by conducting changing variables for \( S_2, X_2, X_3, Z_2 \) and \( Z_3 \), which gives the additional term \( h^{d_z+d_x+d_z+d_x} \). Then by Serfling (1980, Theorem, Chapter 5.5.2), \( nU_{1n} \xrightarrow{d} 3Y \), where \( \mathcal{Y} \) is an infinite weighted sum of \( \chi^2 \) distributions. So the order of \( U_{1n} \) is

\[
U_{1n} \sim O_p \left( \frac{n(n-1)(n-2)\sqrt{h^{2d_z+d_x}}}{n^3h^{d_m+2d_d}} \times \frac{1}{n} \right) = O_p \left( \frac{h^{d_z/2}}{n^{dh_{d_x}}+2} \right).
\]

Therefore by Assumption 7-iii,

\[
\kappa_n U_{1n} \sim h^{d_m+2} \times O_p \left( \frac{h^{d_z/2}}{n^{dh_{d_x}}+2} \right) \sim O_p \left( \sqrt{\frac{1}{n^{dh_{d_x}}+2}} \right) = o_p(1).
\]

where we use the fact that \( d_s = 1 \). It remains to analyze the \( U_{2n} \) term, which we can write as

\[
U_{2n} = \frac{1}{n^3h^{d_m+2d_d}} \times \sum_{i=1}^{n} \sum_{j \neq i} \Sigma_{m,n}(Z_i)K_h(S_j - s)K^2_h(X_j - x)(S_j - s) \epsilon^m_i \epsilon^m_j
\]

\[
= \frac{h^{d_z}}{h^{d_m+2d_d}} \times \frac{n(n-1)}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \Sigma_{m,n}(Z_i)K_h(S_j - s)K^2_h(X_j - x)(S_j - s) \epsilon^m_i \epsilon^m_j
\]

\[
\Phi_{ij} \quad \mathcal{U}_{2n}
\]

where \( \tau = \frac{3}{2}d_s + d_z + \frac{1}{2}d_x \). Now we analyze \( \mathcal{U}_{2n} \), 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 \( \Phi_{ij} \), we need to calculate the variance of \( \mathbb{E}[\Phi_{ij}|H_1] = \frac{1}{2} \mathbb{E}[\Phi_{ij}^2|H_1] + \frac{1}{2} \mathbb{E}[\Phi_{ij}^2|H_1] \).

\[
\mathbb{E}[\Phi_{ij}^2|H_1] = \frac{h^{d_z}}{h^{d_z}} \int \Sigma_{m,n}(Z_i)K_h(S_2 - s)K^2_h(X_2 - x) \mathbb{E}[\epsilon^m_i \epsilon^m_j|W_2]g_w(W_2)dw_2
\]

\[
= \frac{h^{d_z}}{h^{d_z}} \int \Sigma_{m,n}(Z_i)K_h(u_2)K^2(u_2)u_2 \mathbb{E}[\epsilon^m_i \epsilon^m_j|s, x, Z_1 + hu_2]g_w(s, x, Z_1 + hu_2)du_2du_2du_z = 0,
\]

where the last equality again holds by \( \int u_2K(u_2)du_2 = 0 \). Next, we look at

\[
\mathbb{E}[\Phi_{ij}^2|H_1] = \frac{h^{d_z}}{h^{d_z}} \int \Sigma_{m,n}(Z_i)K_h(S_i - s)K^2_h(X_i - x) \mathbb{E}[\epsilon^m_i \epsilon^m_j|W_2]g_w(Z_2)dz_2
\]

\[
= \frac{h^{d_z}}{h^{d_z}} \int \Sigma_{m,n}(Z_i - hu_2)K_h(S_i - s)K^2_h(X_i - x) \mathbb{E}[\epsilon^m_i \epsilon^m_j|Z_1 - hu_2]g_w(Z_1 - hu_2)du_2
\]

\[
= \frac{h^{d_z}}{h^{d_z}} \int \Sigma_{m,n}(Z_i - hu_2)K_h(S_i - s)K^2_h(X_i - x) \mathbb{E}[\epsilon^m_i \epsilon^m_j|Z_1]g_w(Z_1) \int K^2(u_2)du_2
\]
Therefore

\[
\mathbb{E}(\mathbb{E}[\phi_1|H_1])^2 
\approx \int \frac{h^{2d_x+2d_z}}{h^{2\tau}} K_h(S_1 - s) K_h^4(X_1 - x)(S_1 - s)/h^2 \mathbb{E}^2[e^{\tau}e_1^T|Z_1]g_x^2(Z_1) \left( \int K^2(u_x)du_x \right)^2 g_x(Z_1)dZ_1 
\approx \frac{h^{d_x+2d_z+d_z+d_z}}{h^{2\tau}} \int \sum_m \int \left( \int \mathbb{E}[\phi_1]^2 |Z_1| K^2(u_x) u_x g_x(Z_1) g_w(u_s, u_x, Z_1) du_x du_s du_z Z_1 \left( \int K^2(u_x)du_x \right)^2 \right) = O(1),
\]

where the last equality holds because \( \tau = \frac{3}{2}d_s + d_z + \frac{1}{2}d_x \). So we have

\[
\sqrt{n} (\bar{U}_{2n} - \mathbb{E}[\bar{U}_{2n}]) = O_p(1).
\]

It remains to analyze the order of \( \mathbb{E}[\bar{U}_{2n}] \). Since \( \mathbb{E}[\phi_{21}|H_1] = 0 \Rightarrow \mathbb{E}[\phi_{21}^\ast] = 0 \), therefore we have

\[
\mathbb{E}[\bar{U}_{2n}] = \frac{1}{2} \mathbb{E}[\phi_{12}^\ast] + \frac{1}{2} \mathbb{E}[\phi_{21}^\ast] = \frac{1}{2} \mathbb{E}[\phi_{21}^\ast] = \frac{h^{d_x}}{2h^{2\tau}} \int \sum_m \int \left( \int \mathbb{E}[\phi_1]^2 |Z_1| K^2(u_x) K^2(u_x) u_x g_x(Z_1) g_w(s, x, Z_1) du_x du_s du_z Z_1 + O(h) \right) = O \left( \frac{h^{d_x+d_m+1}}{h^{2\tau}} \right),
\]

where the last equality holds because \( \int u_x K(u_x)du_x = 0 \). Now we can conclude that

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

Therefore,

\[
\kappa_n U_{2n} \sim \frac{h^{\tau} \sqrt{n h^{d_m+2-d_x}}}{nh^{d_m+2+d_x}} \times O_p \left( \frac{1}{\sqrt{n}} \right) + \frac{h^{\tau} \sqrt{n h^{d_m+2-d_x}}}{nh^{d_m+2+d_x}} \times O \left( \frac{h^{d_x+d_m+1}}{h^{2\tau}} \right) = O_p \left( \frac{1}{nh^{d_x}} \right) + O_p \left( \sqrt{h^{3d_x+d_z+d_z} / nh^{d_m+2}} \right) = o_p(1).
\]

Since both \( U_{1n} \) and \( U_{2n} \) are of order \( o_p(\kappa_n^{-1}) \), so is \( U_n \). Therefore we can conclude that \( T_n \) is also of order \( o_p(\kappa_n^{-1}) \).