A Note on Average Derivative Estimation*

Ping Yu
University of Auckland

First Version: January 2014
This Version: January 2014

Abstract

This paper studies asymptotic properties of the direct estimator of average derivatives when the density of regressors is not zero on the boundary of its support. It is shown that the estimator cannot achieve the $\sqrt{n}$ convergence rate while does not suffer from the curse of dimensionality either. We also show that the average derivative estimator associated with a discrete regressor can achieve the $\sqrt{n}$ convergence rate. The differences in the convergence rates are attributed to whether the average derivative or the average level is estimated. We further study the information content of the index structure in a special single index model (i.e., the linear regression) where the parameter of average derivative is usually motivated.

KEYWORDS: average derivative, U-statistic, index structure, local polynomial estimator, curse of dimensionality, least squares estimator
JEL-CLASSIFICATION: C13, C14, C21

*Email: p.yu@auckland.ac.nz

1 Introduction

Average derivatives are useful parameters for semiparametric index models and nonparametric demand analysis. For the former, see Stoker (1986, 1991a), and for the latter, see Härdle et al. (1991). Suppose

$$y = m(x) + \varepsilon, E[\varepsilon|x] = 0,$$

where $x \in \mathbb{R}^d$ has a density $f(x)$, and then the parameter of interest is

$$\delta = E\left[m'(x)\right]. \tag{1}$$

When $m(x) = g(x'\beta)$ as in the single index model,

$$\delta = E\left[\partial g(x'\beta)/\partial(x'\beta)\right]\beta, \tag{2}$$

which is the scaled β.

A key observation in estimating δ is that if $f(x) = 0$ on the boundary of x values, then integration by parts in (1) gives

$$\delta = E\left[y\ell(x)\right] = E[m(x)\ell(x)], \tag{3}$$

where $\ell(x) = -f'(x)/f(x)$ is the negative log-density derivative. Estimators based on this observation are usually labeled as *indirect* estimators. Stoker (1986) provides an indirect estimator by assuming that $f(x)$ takes a parametric form. His estimator has a variant which takes the form of an instrumental variable estimator. Härdle and Stoker (1989) develop the counterpart estimators when $f(x)$ is nonparametric. Stoker (1991b) provides two other *direct* estimators, and shows that all four estimators are first-order equivalent. Since $f(x) = 0$ on the boundary of x's support, we need to trim the estimator of $f(x)$ in the estimation of δ. To avoid this problem, Powell et al. (1989) suggest to estimate the weighted average of derivatives with the weight equal to $f(x)$. Powell and Stoker (1996) derive the optimal bandwidth for this density-weighted average. This optimal bandwidth is smaller than the optimal pointwise bandwidth, so undersmoothing (and higher-order kernel) is required to avoid the bias problem in the pointwise estimation of $f'(x)$. As an alternative, Cattaneo et al. (2014) develop asymptotics that are robust to the bandwidth (and kernel) selection.

When $f(x) \neq 0$ on the boundary of x's support, a weight $w(x)$ is usually put on the derivatives such that $w(x)f(x) = 0$ on the boundary of x's support. Newey and Stoker (1993) discuss the semiparametric efficiency bound for this parameter and show that the estimators in Härdle and Stoker (1989) and Stoker (1991b) are efficient if properly adjusted. They also develop the efficiency bound for β in the single index model. This efficiency bound is generally lower than that of δ because the index includes more information on the derivative of $m(x)$.

It remains unknown what the asymptotic properties of a direct estimator of δ based on averaging the estimates of $m'(x)$ would be if $f(x) \neq 0$. It is only known that the convergence rate of such an estimator is slower than $\sqrt{n}$. To understand this result, we repeat the discussion following Assumption 3.1 in Newey and Stoker (1993). Suppose x is a scalar and uniformly distributed on $[0, 1]$. Then

$$E\left[m'(x)\right] = \int_0^1 m'(x)dx = m(1) - m(0),$$

which cannot be estimated in $\sqrt{n}$ rate since the semiparametric variance bound is infinite in this case. In this paper, we study the asymptotic properties of such a direct estimator of δ. We consider both the cases

with continuous x and with a discrete regressor in x. It turns out that the δ estimator in the former case has a slower-than-$\sqrt{n}$ convergence rate but does not suffer from the curse of dimensionality and the δ estimator in the latter case can achieve $\sqrt{n}$ consistency. Such results are developed in Section 2. Furthermore, since δ is often motivated in the single index model as in (2), we discuss in Section 3 the information content of the index structure in a simple single index model - the linear regression. Finally, Section 4 concludes. All proofs are contained in a technical appendix.

2 Estimation of the Average Derivative

When all regressors are continuous, we consider the direct estimator of δ in (1). When some regressors are discrete, we assume only one regressor is discrete and this regressor is binary. Now,

$$y = m(D, x) + \varepsilon, E[\varepsilon|D, x] = 0,$$

where D can take only two values, 0 and 1, and x is continuous. The counterpart of the average derivative of D is

$$\Delta = E\left[m(1, x) - m(0, x)\right] \equiv E\left[\Delta(x)\right],$$

which is reminiscent of the average treatment effects (ATE) under unconfoundedness.

2.1 When x is Continuous

Given the definition of δ, we can estimate it by its sample analog,

$$\widehat{\delta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{b}(x_i),$$

where $\widehat{b}(x_i)$ is the local polynomial estimator (LPE) of $\partial E[y_i|x_i]/\partial x'$.

To define $(\widehat{a}(x_i), \widehat{b}(x_i)')'$ with $\widehat{a}(x_i)$ being the LPE of $E[y_i|x_i]$, we first define the generalized kernel function below. The following discussion on the generalized kernel function is borrowed from Müller (1991).

Definition: *$k_h(\cdot,\cdot)$ is called a* univariate generalized kernel function of order p *if $k_h(u,t) = 0$ if $u > t$ or $u < t-1$ and for all $t \in [0,1]$,*

$$\int_{t-1}^{t} u^j k_h(u,t)du = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } 1 \le j \le p-1. \end{cases}$$

To simplify the construction of $k_h(u,t)$, we put the following constraints on the support of x.

Assumption S: $(y, x')' \in \mathbb{R} \times \mathcal{X} \subset \mathbb{R}^{d+1}$, and $\mathcal{X} = [0,1]^d$.

The restriction of the support of x as $[0,1]^d$ is not really restrictive, up to some monotone transformations such as the empirical percentile transformation. The assumption of the compactness of $\mathcal{X}$ is necessary for imposing positivity on $f(x)$ on the boundary of $\mathcal{X}$. Now, define $k_+(\cdot,\cdot)$ and $k_-(\cdot,\cdot)$ as follows:

(i) The support of $k_-(x,r)$ is $[-1,r]\times[0,1]$ and the support of $k_+(x,r)$ is $[-r,1]\times[0,1]$.

(ii) $k_-(\cdot,r) \in \mathcal{M}_p\left([-1,r]\right)$ and $k_+(\cdot,r) \in \mathcal{M}_p\left([-r,1]\right)$.

(iii) $k_+(x,r) = k_-(-x,r)$.

where

$$\mathcal{M}_p([a,b]) = \left\{ g \in \text{Lip}([a,b]), \int_a^b x^j g(x)dx = \left\{ \begin{array}{ll} 1, & \text{if } j=0, \\ 0, & \text{if } 1 \le j \le p-1 \end{array} \right\} \right.,$$

and $\text{Lip}([a,b])$ denotes the space of Lipschitz continuous functions on $[a,b]$. Then

$$k_h(u,t) = \left\{ \begin{array}{ll} \frac{1}{h}k\left(\frac{u}{h}\right), & \text{if } h \le t \le 1-h, \\ \frac{1}{h}k_+\left(\frac{u}{h}, \frac{t}{h}\right), & \text{if } 0 \le t \le h, \\ \frac{1}{h}k_-\left(\frac{u}{h}, \frac{1-t}{h}\right), & \text{if } 1-h \le t \le 1, \end{array} \right. \tag{4}$$

where

$$k(\cdot) = k_+(\cdot,1) = k_-(\cdot,1) \in \mathcal{M}_p([-1,1]).$$

We can show that $k_h(u,t)$ is a generalized kernel function of order p. Further, we construct a multivariate generalized kernel function of order p by the product of univariate generalized kernel functions of order p. Since the LPE is used, we only need $k_h(u,t)$ to be a second-order kernel function. Formally,

Assumption K: $k_h(u,t)$ takes the form of (4) with $p=2$.

Define

$$K^x_{h,ij} = \prod_{l=1}^d k_h(x_{lj} - x_{li}, x_{li}) \equiv K^x_h(x_j - x_i, x_i) \equiv \frac{1}{h^d} K^x\left(\frac{x_j - x_i}{h}, x_i\right).$$

Then the LPE $(\widehat{a}(x_i), \widehat{b}(x_i)')'$ is the first $(d+1)$ elements of the solution to

$$\min_{\beta} \sum_{j=1, j\neq i}^n \left[y_j - (x_j' - x_i')^{S_p}\beta\right]^2 K^x_{h,ij},$$

where $p \ge 1$, for a row (column) vector $\xi \in \mathbb{R}^d$, $\xi^{S_p} = (\xi^{S(\nu)})_{\nu \in \{0,\cdots,p\}}$ is a row (column) vector, $\xi^{S(\nu)} = (\xi^s)_{|s|=\nu}$ is a row (column) vector of length $(\nu+d-1)!/\nu!(d-1)!$, $s = (s_1, \cdots, s_d)$ is a vector with all its elements being nonnegative integers, the norm of s is defined as $|s| \equiv s_1 + \cdots s_d$, and $\xi^s = \xi_1^{s_1} \cdots \xi_d^{s_d} / (s_1! \cdots, s_d!)$. For convenience, we assume that $\{(s_1, \cdots, s_d)\}$ in the definition of ξ^{S_p} are ordered lexicographically. x_i is excluded in the construction of $\widehat{b}(x_i)$ to let $\widehat{\delta}$ take the form of a U-statistic instead of a V-statistic. From Theorem 3 of Heckman et al. (1998), $\widehat{a}(x_i)$ and $\widehat{b}(x_i)$ are asymptotically linear, so $\widehat{\delta}$ is indeed asymptotically a U-statistic. To guarantee this linear approximation, we need the following assumptions.

Assumption F: The density of x, $f(x)$, is continuously differentiable on $\mathcal{X}$, and $0 < \underline{f} \le f(x) \le \overline{f} < \infty$ for $x \in \mathcal{X}$.

$f(x)$ is bounded away from zero to avoid trimming the low-density area of $\mathcal{X}$ as in Heckman et al. (1998). This is also the main difference of this paper from the existing literature of the average derivative estimation. To impose restrictions on $m(x)$ and the bandwidth, we define the smoothness index s of $m(x)$ as follows: $m(x)$ is s-times continuously differentiable and its s-th derivative satisfies Hölder's condition.

Assumption M: $m(x)$ has the smoothness index $s > d$ for all $x \in \mathcal{X}$.

Assumption H: $h \to 0$, $nh^{p+1+d/2} \to [0,\infty)$, $nh^{d+2} \to 0$, and $nh^d / \ln n \to \infty$.

This assumption is parallel to Assumption 3 of Heckman et al. (1998). The assumption that $nh^{p+1+d/2} \to [0,\infty)$ is to guarantee the asymptotic bias in the asymptotic distribution of $\widehat{\delta}$ is finite. Assumption of Heckman et al. (1998) also requires that $nh^{2s} \to [0,\infty)$. Given Assumption M, $s \geq d+1$, so this assumption is implied by $nh^{d+2} \to 0$. To derive the asymptotic distribution of $\widehat{\delta}$, we also need some restrictions on the conditional moment of ε:

Assumption E: $E[\varepsilon^4|x]$ is uniformly bounded on $x \in \mathcal{X}$.

The following theorem states the asymptotic distribution of $\widehat{\delta}$. To ease our exposition, we define some notations here. M is the square matrix of size $\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ with the l-th row, t-th column "block" being

$$\int u^{S(l)}\left(u'\right)^{S(t)} K(u)\, du, 0 \leq l,t \leq p,$$

where $K(u) = \prod_{l=1}^{d} k(u_l)$. B is a $\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ by $(p+d)!/(p+1)!(d-1)!$ matrix with the l-th block being

$$\int u^{S(l)}\left(u'\right)^{S(p+1)} K(u)\, du.$$

$\mathbf{e}_l$ is a $\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ by 1 vector with the lth element being 1 and all other elements being 0, $l = 1,\cdots,d+1$. $(\mathbf{0}, I_d, \mathbf{0})$ is a $d \times \sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ matrix with the first zero matrix being a column vector and I_d being an identity matrix of size d. $m^{(p+1)}(x)$ is a $(p+d)!/(p+1)!(d-1)!$ by 1 vector of the partial derivatives of $m(x)$.

Theorem 1 *Under Assumptions E, F, H, K, M and S,*

$$nh^{1+d/2}\left(\widehat{\delta} - \delta - h^p\left(\mathbf{0}, I_d, \mathbf{0}\right) M^{-1} B \cdot E[m^{(p+1)}(x)]\right) \xrightarrow{d} N\left(0, \int \sigma^2(x)dx \cdot \Sigma\right),$$

where $\sigma^2(x) = E[\varepsilon^2|x]$*, and the* (l,t) *element of* Σ *is*

$$\int \left(\mathbf{e}_{l+1}' M^{-1} u^{S_p} K(u)\right)\left(\mathbf{e}_{t+1}' M^{-1} u^{S_p} K(u)\right) du,$$

for $l,t = 1,\cdots,d$.

Since $nh^{d+2} \to 0$, $nh^{1+d/2} = o(\sqrt{n})$, i.e., the convergence rate of $\widehat{\delta}$ is slower than $\sqrt{n}$. If minimizing the asymptotic mean squared error (AMSE) of $\widehat{\delta}$, we have

$$h = O\left(n^{-\frac{2}{2p+2+d}}\right),$$

and the square root of the AMSE (RAMSE) of $\widehat{\delta}$ is $O\left(n^{-\frac{2p}{2p+2+d}}\right)$. When d is large, this RAMSE is large, so it seems that $\widehat{\delta}$ suffers from the curse of dimensionality. This is not the case. Actually, from Section 2.5 of Yu (2013), we expect that the convergence rate of $\widehat{\delta}$ is $\sqrt{n}h$ rather than $nh^{1+d/2} = \sqrt{n}h\sqrt{nh^d}$. If the convergence rate is $\sqrt{n}h$, $\widehat{\delta}$ does not suffer from the curse of dimensionality. Our convergence rate is faster than $\sqrt{n}h$ given that $nh^d \to \infty$, so we expect $\widehat{\delta}$ does not suffer from the curse of dimensionality in practice. To be specific, recall that the optimal RAMSE of the one-dimensional slope estimation is of $O\left(n^{-\frac{p}{2p+2+1}}\right)$, so when $p \geq \frac{d}{2} - 2$, the convergence rate of $\widehat{\delta}$ is not slower than that in the one-dimensional slope estimation. Usually, $d \leq 6$, so even the local linear estimator will achieve a convergence rate at least as fast as the-one dimensional slope estimator. The faster convergence rate is due to the fact that $\widehat{\delta} - \delta$ can actually be represented as a

degenerate U-statistic. To appreciate why this can happen, note that $\int \mathbf{e}_{l+1}' M^{-1} u^{S_p} K(u)\, du = 0$, so the first-order projection of U-statistic does not contribute to the asymptotic distribution.

Recall also that the slope estimation at a single point on $\mathcal{X}$ has the optimal RAMSE of $O\left(n^{-\frac{p}{2p+2+d}}\right)$, so the average derivative estimator has a much faster convergence rate than the pointwise derivative estimator. To be specific, let $d = 1$ and $p = 1$. Then the former has the optimal RAMSE of $O\left(n^{-\frac{2}{5}}\right)$, which is the same as in the one-dimensional level estimation, and the latter has the optimal RAMSE of $O\left(n^{-\frac{1}{5}}\right)$.

The bias and variance of $\widehat{\delta}$ take the form of integrated bias and variance of $\widehat{b}(x_i)$ at all x_i. This is understandable from the construction of $\widehat{\delta}$. The asymptotic variance matrix in the theorem can be consistently estimated by its sample analog. Such results are standard in the literature, so are omitted here.

2.2 When x Contains a Discrete Regressor

From Hahn (1998), Δ can be estimated in $\sqrt{n}$ rate, and the asymptotic variance bound is

$$E\left[\frac{\sigma_1^2(x)}{p(x)} + \frac{\sigma_0^2(x)}{1-p(x)} + (\Delta(x) - \Delta)^2\right],$$

where $p(x) = E[D|x]$ is the propensity score, and $\sigma_d^2(x) = Var(y|x, D = d)$, $d = 0, 1$, is the conditional variance in each treatment state. So different from the case with continuous covariates, we can estimate the counterpart of the average derivative of a discrete covariate in $\sqrt{n}$ rate. The existing methods of estimating Δ in the literature are summarized in Imbens (2004).

A natural question here is that why Δ can be estimated in $\sqrt{n}$ rate while δ cannot. Basically, this is because Δ is an average of levels $m(1, x)$ and $m(0, x)$, while δ is an average of slopes $m'(x)$. It is well known that slopes are harder to estimate than levels. To appreciate this result, note that when δ can be expressed as an average of levels (see (3)), it can also be estimated in $\sqrt{n}$ rate.

The key assumption to achieve the $\sqrt{n}$ convergence rate in the estimation of Δ is that D cannot be perfectly predicted by x. Otherwise, the variance bound diverges to infinity. In the treatment effects literature, this assumption is called the *overlapping* or *matching* assumption, that is, for each x, there must be some individuals treated and some untreated. To understand this result, assume the treatment effect is constant for each x. In other words,

$$m(D, x) = D\Delta + m(x). \tag{5}$$

This is the partially linear model considered in Robinson (1988). It implies that $\partial m(1, x)/\partial x = \partial m(0, x)/\partial x$. From the Theorem in Robinson (1988), when D can be perfectly predicted by x, Δ cannot be identified.[1]

Finally, we close this section by some discussions on the efficiency bound of Hahn (1998). It is well known that when the model is homoskedastic, that is, $E[\varepsilon^2|x, D] = \sigma^2$, the asymptotic variance bound for Δ in (5) is $\sigma^2/E[p(x)(1 - p(x))]$, which is the same as predicted by Hahn (1998). However, when $\sigma^2(x, D) = E[\varepsilon^2|x, D]$ depends on $(x', D)'$ or only depends on x, the efficiency bound predicted by Hahn (1998) is not the same as that in Chamberlain (1992). Hahn's bound is

$$E\left[\frac{\sigma_1^2(x)}{p(x)} + \frac{\sigma_0^2(x)}{1-p(x)}\right],$$

[1] Strictly speaking, to identify Δ in (5), we require only that D cannot be perfectly predicted by x at some values of x, i.e., we allow that D can be perfectly predicted by x at some values of x. However, for the general model, we require that D cannot be perfectly predicted by x at all values of x.

which reduces to $E\left[\frac{\sigma^2(x)}{p(x)(1-p(x))}\right]$ when $\sigma_1^2(x)=\sigma_0^2(x)=\sigma^2(x)$. While Chamberlain's bound is

$$1\Big/\left\{E\left[\frac{D}{\sigma^2(x,D)}\right]-E\left[E\left[\frac{D}{\sigma^2(x,D)}\middle|x\right]^2\Big/E\left[\frac{1}{\sigma^2(x,D)}\middle|x\right]\right]\right\}=1\Big/E\left[\frac{1}{\frac{\sigma_1^2(x)}{p(x)}+\frac{\sigma_0^2(x)}{1-p(x)}}\right],$$

which reduces to $1\Big/E\left[\frac{p(x)(1-p(x))}{\sigma^2(x)}\right]$ when $\sigma_1^2(x)=\sigma_0^2(x)=\sigma^2(x)$. By Jensen's inequality, the former is larger than the latter. This is because Chamberlain's bound is based on the conditional moment restrictions, while Hahn's bound is based on the general setup of the treatment model. In other words, in the constant treatment model, more information is available, and such information is unavailable in the general treatment model.

3 Information Content in the Index Structure

As mentioned in the introduction, the parameter of average derivative is usually motivated in the single index model. In this section, we study the information content of the index structure in the single index model. From the last section, δ cannot be estimated in $\sqrt{n}$ rate. However, when an index structure is imposed, e.g., $m(x)=g(x'\beta)$, Ichimura (1993) shows that β can be estimated in $\sqrt{n}$ rate up to a scale factor. To understand the information content of the index structure, we consider in this section a special single index model, i.e., the linear regression, where the scale factor is 1.

In linear regression,

$$y_i=\beta_0+x_i'\beta_1+\varepsilon_i\equiv\mathbf{x}_i'\beta+\varepsilon_i, E[\varepsilon_i|x_i]=0,$$

where $\mathbf{x}_i=(1,x_i')'\in\mathbb{R}^{d+1}$ and $\beta=(\beta_0,\beta_1')'$. The most popular estimator is the ordinary least squares estimator (LSE), combined with the heteroskedasticity-robust standard error, although the weighted least squares estimator is more efficient under the conditional moment restriction. It is well known that the LSE of β, $\widehat{\beta}_{LSE}\equiv(\widehat{\beta}_{0,LSE},\widehat{\beta}_{1,LSE}')'$, has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\beta}_{LSE}-\beta\right)\xrightarrow{d}N\left(0,E\left[\mathbf{x}\mathbf{x}'\right]^{-1}E\left[\mathbf{x}\mathbf{x}'\varepsilon^2\right]E\left[\mathbf{x}\mathbf{x}'\right]^{-1}\right).$$

Although this result is standard, we try to answer two seemingly naive questions in this paper. First, why can $\widehat{\beta}_{LSE}$ achieve the $\sqrt{n}$ rate given that $\widehat{\delta}$ cannot. Second, why $\widehat{\beta}_{1,LSE}$, as a slope estimator, has the same convergence rate as $\widehat{\beta}_{0,LSE}$ which is an intercept (or level) estimator. Our answer is that this is because the LSE uses the index structure in the conditional mean of y.

Since $\delta=\beta_1$ in linear regression, $\widehat{\delta}$ is estimating β_1. Since $\beta_0=E[y]-E[x]'\beta_1$, it can be estimated as

$$\widehat{\beta}_0^{(1)}=\overline{y}-\overline{x}'\widehat{\delta},$$

where $\overline{x}$ and $\overline{y}$ are the sample averages of x and y. Of course, given that $\beta_0=E[E[y|x]-x'\beta]$, it can also be estimated as

$$\widehat{\beta}_0^{(2)}=\frac{1}{n}\sum_{i=1}^n\left(\widehat{a}(x_i)-x_i'\widehat{\delta}\right)\equiv\widehat{A}-\overline{x}'\widehat{\delta},$$

where $\widehat{A}=n^{-1}\sum_{i=1}^n\left(\widehat{a}(x_i)-x_i'\widehat{\delta}\right)$. Neither $\widehat{\beta}_0^{(1)}$ and $\widehat{\beta}_0^{(2)}$ uses the index structure of linear regression.[2] If

[2] Of course, the additive structure of $m(x)$ is used.

we use the index structure, we can regress $\widehat{a}(x_i)$ on $\mathbf{x}_i$ to get an estimator of β,

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i \widehat{a}(x_i)\right).$$

The following theorem states the asymptotic distributions of $\left(\widehat{\beta}_0^{(1)}, \widehat{\beta}_0^{(2)}\right)$ and $\widetilde{\beta}$.

Theorem 2 *Under Assumptions E, F, H, K, and S, $\widehat{\beta}_0^{(1)} - \beta_0$ and $\widehat{\beta}_0^{(2)} - \beta_0$ have the same asymptotic distribution. When $E[x] \neq 0$, their asymptotic distribution is the same as $-E[x]'\left(\widehat{\delta} - \delta\right)$; when $E[x] = 0$, their asymptotic distribution is the same as $\overline{\varepsilon}$. The asymptotic distribution of $\widetilde{\beta}$ is the same as that of $\widehat{\beta}_{LSE}$.*

We do not need assumption M since $m(x)$ is infinitely differentiable in linear regression. Also, assumption E can be relaxed as $E[\varepsilon^{2+\alpha}|x]$ is uniformly bounded on $x \in \mathcal{X}$ for some $\alpha > 0$.

From this theorem, the asymptotic properties of the intercept estimators are indeed different from those of the slope estimator. Usually, we use β_0 to summarize the level information of $m(x)$, so it is natural to assume $E[x] = 0$ or center x_i at $\overline{x}$. In this case, the intercept estimators have the same asymptotic distribution as $\overline{\varepsilon}$, i.e., they achieve the $\sqrt{n}$ convergence rate even though the index structure is not imposed. This result is much expected from the discussion in Section 2.2. When $E[x] \neq 0$, the convergence rate of the intercept estimators is contaminated by the slower convergence rate of $\widehat{\delta}$. It is also interesting to observe that using either $\overline{y}$ or $\widehat{A}$ to estimate $E[y]$ does not affect the asymptotic distribution of the intercept estimators.

When $E[x] = 0$, note that $\widehat{\beta}_{0,LSE}$ has the same asymptotic distribution as $\overline{\varepsilon}$, so the index structure does not provide any extra information for the intercept. On the contrary, when the index structure is imposed, $\widetilde{\beta}_1$ has a faster convergence rate than $\widehat{\delta}$ and its asymptotic distribution is the same as $\widehat{\beta}_{1,LSE}$. So the index structure mainly contains information about the global derivatives of $m(x)$.

4 Conclusion

In this paper, we study asymptotic properties of the direct estimator of average derivatives when the density of regressors is not zero on the boundary of its support. We show that the estimator cannot achieve the $\sqrt{n}$ convergence rate; nevertheless, it does not suffer from the curse of dimensionality either. We also show that the average derivative estimator associated with a discrete regressor can achieve the $\sqrt{n}$ convergence rate. The differences in the convergence rates can happen because the former is estimating an average of slopes while the latter is estimating an average of levels. Given that the parameter of average derivative is usually motivated in the single index model, we further study the information content of the index structure in a special single index model - the linear regression, and show that the index structure mainly contains information about derivatives rather than levels.

The main purpose of this paper is to attract more attentions on this obvious problem in the average derivative estimation, so only basic analyses are conducted. There are many interesting questions unsolved in this paper. For example, what is the optimal (or minimax) rate of convergence of δ given that it cannot be estimated in $\sqrt{n}$ rate? Is the bootstrap valid for the inference of $\widehat{\delta}$?[3]

[3]For the indirect estimator of δ, the bootstrap validity has already been discussed in Nishiyama and Robinson (2005).

References

Cattaneo, M.D., R.K. Crump and M. Jansson, 2014, Small Bandwidth Asymptotics for Density-Weighted Average Derivatives, *Econometric Theory*, 30, 176-200.

Chamberlain, G., 1992, Efficiency Bounds for Semiparametric Regression, *Econometrica*, 60, 567-596.

Hahn, J., 1998, On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, *Econometrica*, 66, 315-331.

Hall, P., 1984, Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, *Journal of Multivariate Analysis*, 14, 1-16.

Härdle, W., W. Hildenbrank, and M. Jerison, 1991, Empirical Evidence on the Law of Demand, *Econometrica*, 59, 1525-1549.

Heckman, J.J., H. Ichimura and P. Todd, 1998, Matching as an Econometric Evaluation Estimator, *Review of Economic Studies*, 65, 261–294.

Ichumura, H., 1993, Semiparametric Least Squares Estimation of Single Index Models (SLS) and Weighted SLS Estimation of Single Index Models, *Journal of Econometrics*, 58, 72-120.

Imbens, G.W., 2004, Nonparametric Estimaiton of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86, 4-29.

Müller, H.-G., 1991, Smooth Optimum Kernel Estimators Near Endpoints, *Biometrika*, 78, 521-530.

Newey, W. and T.M. Stoker, 1993, Efficiency of Weighted Average Derivative Estimators and Index Models, *Econometrica*, 61, 1199-1223.

Nishiyama, Y. and P.M. Robinson, 2005, The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives, *Econometrica*, 73, 903-948.

Powell, J.L., J.H. Stock and T.M. Stocker, 1989, Semiparametric Estimation of Index Coefficients, *Econometrica*, 57, 1403-1430.

Powell, J.L., and T.M. Stoker, 1996, Optimal Bandwidth Choice for Density-Weighted Averages, *Journal of Econometrics*, 75, 291-316.

Robinson, P.M., 1988, Root-N-Consistent Semiparametric regression, *Econometrica*, 56, 931-954.

Stoker, T.M., 1986, Consistent Estimation of Scaled Coefficients, *Econometrica*, 1461-1481.

Stoker, T.M., 1991a, *Lectures on Semiparametric Econometrics*, CORE Lecture Series, CORE Fundation, Univerisité Catholique de Louvain, Louvain.

Stoker, T.M., 1991b, Equivalence of Direct, Indirect, and Slope Estimators of Average Derivatives, *Nonparametric and Semiparametric Methods in Economtrics and Statistics*, ed. by: W.A. Barnett, J.L. Powell, and G.E. Tauchen, Cambridge, MA: Cambridge University Press, 99-118.

Yu, P., 2013, Threshold Regression with Endogeneity, mimeo, University of Auckland.

Appendix: Proofs

Proof of Theorem 1. First derive the mathematical formula for $\widehat{\delta}$. Following Appendix A.1 of Heckman et al. (1998), we have

$$\begin{aligned}
\left(\widehat{a}(x_i), \widehat{b}'(x_i)\right)' &= (I_{d+1}, \mathbf{0}) \left(X_i' W_i X_i\right)^{-1} X_i' W_i Y \\
&= (I_{d+1}, \mathbf{0}) H^{-1} \left(H^{-1} X_i' W_i X_i H^{-1}\right)^{-1} H^{-1} X_i' W_i Y \\
&= (I_{d+1}, \mathbf{0}) H^{-1} \left(Z_i' W_i Z_i\right)^{-1} Z_i' W_i Y \\
&= (I_{d+1}, \mathbf{0}) H^{-1} \left(\frac{1}{n} \sum_{j=1, j\neq i}^{n} z_j^i z_j^{i\prime} w_j^i\right)^{-1} \left(\frac{1}{n} \sum_{j=1, j\neq i}^{n} z_j^i w_j^i y_j\right) \\
&\equiv (I_{d+1}, \mathbf{0}) H^{-1} \left(M_i\right)^{-1} r_i,
\end{aligned}$$

where

$$\begin{aligned}
X_i &= \begin{pmatrix} (x_1' - x_i')^{S_p} \\ \vdots \\ (x_{i-1}' - x_i')^{S_p} \\ \mathbf{0} \\ (x_{i+1}' - x_i')^{S_p} \\ \vdots \\ (x_n' - x_i')^{S_p} \end{pmatrix}_{n \times \sum_{\nu=0}^{p} (\nu+d-1)!/\nu!(d-1)!}, Y = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ 0 \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \\
H &= \operatorname{diag}\left\{1, hI_d, \cdots, hI_{(p+d-1)!/p!(d-1)!}\right\}, Z_i = X_i H^{-1}, z_j^{i\prime} = (x_j' - x_i')^{S_p} H^{-1}, \\
W_i &= \operatorname{diag}\left\{K_h^x\left(x_1 - x_i, x_i\right), \cdots, K_h^x\left(x_n - x_i, x_i\right)\right\} = \operatorname{diag}\left\{w_1^i, \cdots, w_n^i\right\}_{n\times n}.
\end{aligned}$$

Now,

$$\widehat{\delta} = \frac{1}{n} \sum_{i=1}^{n} \widehat{b}(x_i).$$

Given assumptions E, F and H, we can apply the arguments in Theorem 3 of Heckman et al. (1998) to have

$$\sqrt{n}h\left(\widehat{\delta} - \overline{\delta}\right) = \frac{h^{p+1}}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{0}, I_d, \mathbf{0}) \left(\overline{M}_i\right)^{-1} r_i^m + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{0}, I_d, \mathbf{0}) \left(\overline{M}_i\right)^{-1} r_i^{\varepsilon} + o_p(1)$$

where $\overline{\delta} = n^{-1} \sum_{i=1}^{n} m(x_i)$, $\overline{M}_i$ is the square matrix of size $\sum_{\nu=0}^{p} (\nu+d-1)!/\nu!(d-1)!$ with the l-th row, t-th column "block" being, for $0 \leq l, t \leq p$,

$$\int (u')^{S(l)\prime} (u')^{S(t)} K^x(u_x, x_i) f(x_i + uh) du,$$

r_i^m is a $\sum_{\nu=0}^{p} (\nu+d-1)!/\nu!(d-1)!$ by 1 vector with the t-th block being

$$\int u^{S(t)} \left[(u')^{S(p+1)} m^{(p+1)}(x_i)\right] K^x(u, x_i) f(x_i) du,$$

and

$$r_i^{\varepsilon} = \frac{1}{n} \sum_{j=1, j\neq i}^{n} z_j^i w_j^i \varepsilon_j.$$

The terms associated with r_i^m will contribute to the bias and the terms associated with r_i^{ε}, which is a second-order U-statistic, will contribute to the variance.

First, analyze the bias.

$$\begin{aligned} & E\left[\frac{1}{n}\sum_{i=1}^{n} (\mathbf{0}, I_d, \mathbf{0}) \left(\overline{M}_i\right)^{-1} r_i^m\right] \\ = \quad & (\mathbf{0}, I_d, \mathbf{0}) \int \left(\overline{M}_i\right)^{-1} r_i^m f(x_i) dx_i \to (\mathbf{0}, I_d, \mathbf{0}) M^{-1} B \cdot E[m^{(p+1)}(x)], \end{aligned}$$

where M and B are defined in the main text. Note here that the kernel K^x is replaced by K because the data in the h neighborhood of the boundary of $\mathcal{X}$ can be neglected asymptotically. Also, we can calculate the variance of this term is $O\left(\frac{1}{n}\right) = o(1)$, so it converges in probability to its expectation.

Second, analyze the variance. Take the lth element of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{0}, I_d, \mathbf{0}) \left(\overline{M}_i\right)^{-1} r_i^{\varepsilon}$, $l = 1, \cdots, d$, which is asymptotically equivalent to

$$\frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathbf{e}_{l+1}' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i \varepsilon_j.$$

From Lemma 8.4 of Newey and McFadden (1994), this U-statistic is asymptotically equivalent to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_n(x_i, \varepsilon_i)$, where

$$m_n(x_j, \varepsilon_j) = E\left[\mathbf{e}_{l+1}' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i \varepsilon_j \middle| x_j, \varepsilon_j\right] = \varepsilon_j \int \mathbf{e}_{l+1}' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i f(x_i) dx_i.$$

We apply the Liapunov central limit theorem to derive its asymptotic distribution. It is standard to check that the Liapunov condition is satisfied, so we concentrate on calculating its asymptotic variance. Note that

$$E\left[\varepsilon_j^2 \left(\int \mathbf{e}_{l+1}' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i f(x_i) dx_i\right)^2\right] = E\left[\varepsilon_j^2 \left(\int \mathbf{e}_{l+1}' M^{-1} u^{S_p} K(u) \, du\right)^2\right] + o(1) \to 0,$$

where the last equality is because $\int \mathbf{e}_{l+1}' M^{-1} u^{S_p} K(u) \, du = 0$. In other words, the associated U-statistic is degenerate.

To derive the asymptotic distribution of $\widehat{\delta} - \overline{\delta}$, we apply Theorem 1 of Hall (1984). Define

$$U_n = \sum_{i=1}^{n} \sum_{j>i} H_n(Z_i, Z_j),$$

where $Z_i = (x_i, \varepsilon_i)$, and $H_n(Z_i, Z_j) = h^{d/2} \left[\mathbf{e}_{l+1}' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i \varepsilon_j + \mathbf{e}_{l+1}' \left(\overline{M}_j\right)^{-1} z_i^j w_i^j \varepsilon_i\right]$. It is not hard to check that $E[H_n^2(Z_1, Z_2)] < \infty$ for each n. We now check

$$\frac{E\left[G_n^2(Z_1, Z_2)\right] + n^{-1} E[H_n^4(Z_1, Z_2)]}{E^2[H_n^2(Z_1, Z_2)]} \to 0 \text{ as } n \to \infty, \tag{6}$$

where $G_n(Z_1,Z_2)=E[H_n(Z_3,Z_1)H_n(Z_3,Z_2)|Z_1,Z_2]$.

$$\begin{aligned}
G_n(Z_1,Z_2) &= h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_3\right)^{-1}z_1^3w_1^3\varepsilon_1+\mathbf{e}_{l+1}'\left(\overline{M}_1\right)^{-1}z_3^1w_3^1\varepsilon_3\right)\left(\mathbf{e}_{l+1}'\left(\overline{M}_3\right)^{-1}z_2^3w_2^3\varepsilon_2+\mathbf{e}_{l+1}'\left(\overline{M}_2\right)^{-1}z_3^2w_3^2\varepsilon_3\right)\middle|Z_1,Z_2\right]\\
&= h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_3\right)^{-1}z_1^3w_1^3\varepsilon_1\right)\left(\mathbf{e}_{l+1}'\left(\overline{M}_3\right)^{-1}z_2^3w_2^3\varepsilon_2\right)\middle|Z_1,Z_2\right]\\
&= h^d\varepsilon_1\varepsilon_2h^{-d}O\left(K\left(\frac{x_2-x_1-uh}{h}\right)\right),
\end{aligned}$$

so

$$E\left[G_n^2(Z_1,Z_2)\right]=O(h^d).$$

Since

$$E[H_n^4(Z_1,Z_2)]=O(h^{-d}),E^2[H_n^2(Z_1,Z_2)]=O(1),$$

(6) is satisfied under Assumption H. So by Theorem 1 of Hall (1984),

$$U_n\Bigg/\sqrt{\frac{1}{2}n^2E\left[H_n^2(Z_1,Z_2)\right]}=\frac{\sqrt{n^3}h^{d/2}\sqrt{n}h\left(\widehat{\overline{\delta}}_l-\overline{\delta}_l\right)}{\sqrt{n^2\sigma^2\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)^2du}}\xrightarrow{d}N(0,1),$$

where

$$\begin{aligned}
\frac{1}{2}n^2E\left[H_n^2(Z_1,Z_2)\right] &= \frac{1}{2}n^2h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\varepsilon_2+\mathbf{e}_{l+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\varepsilon_1\right)^2\right]\\
&= \frac{1}{2}n^2h^d\left\{E\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\varepsilon_2\right)^2\right]+E\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\varepsilon_1\right)^2\right]\right\}\\
&\longrightarrow \frac{1}{2}n^2h^d\frac{2\int\sigma^2(x)dx}{h^d}\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)^2du\\
&= n^2\int\sigma^2(x)dx\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)^2du.
\end{aligned}$$

That is,

$$\sqrt{n}h\sqrt{nh^d}\left(\widehat{\overline{\delta}}_l-\overline{\delta}_l\right)\xrightarrow{d}N\left(0,\int\sigma^2(x)dx\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)^2du\right).$$

Also, for $l\neq t$,

$$\begin{aligned}
& h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\varepsilon_2+\mathbf{e}_{l+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\varepsilon_1\right)\left(\mathbf{e}_{t+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\varepsilon_2+\mathbf{e}_{t+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\varepsilon_1\right)\right]\\
= & h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\right)\left(\mathbf{e}_{t+1}'\left(\overline{M}_1\right)^{-1}z_2^1w_2^1\right)\varepsilon_2^2\right]+h^dE\left[\left(\mathbf{e}_{l+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\right)\left(\mathbf{e}_{t+1}'\left(\overline{M}_2\right)^{-1}z_1^2w_1^2\right)\varepsilon_1^2\right]\\
= & 2\int\sigma^2(x)dx\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)\left(\mathbf{e}_{t+1}'M^{-1}u^{S_p}K\left(u\right)\right)du,
\end{aligned}$$

so the asymptotic covariance between $\widehat{\overline{\delta}}_l$ and $\widehat{\overline{\delta}}_t$ is $\int\sigma^2(x)dx\int\left(\mathbf{e}_{l+1}'M^{-1}u^{S_p}K\left(u\right)\right)\left(\mathbf{e}_{t+1}'M^{-1}u^{S_p}K\left(u\right)\right)du$.

Finally, note that

$$nh^{1+d/2}\left(\widehat{\delta}-\delta\right)=nh^{1+d/2}\left(\widehat{\overline{\delta}}-\overline{\delta}\right)+nh^{1+d/2}\left(\overline{\delta}-\delta\right).$$

Since $\overline{\delta}-\delta=O_p(n^{-1/2})$, $nh^{1+d/2}\left(\overline{\delta}-\delta\right)=O_p\left(n^{1/2}h^{1+d/2}\right)=o_p(1)$ under Assumption H. As a result, $nh^{1+d/2}\left(\widehat{\delta}-\delta\right)$ has the same asymptotic distribution as $nh^{1+d/2}\left(\widehat{\overline{\delta}}-\overline{\delta}\right)$. ■

Proof of Theorem 2. We first discuss the asymptotic properties of $\widehat{A}$. From the proof in the last theorem,

$$\begin{aligned}\sqrt{n}\left(\widehat{A}-\overline{m}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}r_i^{\varepsilon}+o_p(1)\\ &= \frac{1}{\sqrt{n^3}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i\varepsilon_j+o_p(1),\end{aligned}$$

where $\overline{m}=\frac{1}{n}\sum_{i=1}^{n}E[y_i|x_i]=\beta_0+\overline{x}'\beta_1$, Note here that the linear representation does not include the bias term since higher-order derivatives of $E[y_i|x_i]$ are zero in linear regression. Correspondingly, the bias term in Theorem 1 does not appear in linear regression. From Lemma 8.4 of Newey and McFadden (1994), the U-statistic representation of $\sqrt{n}\left(\widehat{A}-\overline{m}\right)$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m_n(x_i,\varepsilon_i)$, where

$$m_n(x_j,\varepsilon_j)=E\left[\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i\varepsilon_j\middle|x_j,\varepsilon_j\right]=\varepsilon_j\int\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i f(x_i)dx_i.$$

By the Liapunov central limit theorem,

$$\sqrt{n}\left(\widehat{A}-\beta_0-\overline{x}'\beta_1\right)\xrightarrow{d}N(0,\sigma^2),$$

where $\sigma^2=E[\varepsilon^2]$. This asymptotic variance is from

$$E\left[\varepsilon_j^2\left(\int\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i f(x_i)dx_i\right)^2\right]=E\left[\varepsilon_j^2\left(\int\mathbf{e}_1'M^{-1}u^{S_p}K(u)\,du\right)^2\right]+o(1)\to\sigma^2,$$

where the last equality is from that fact that $\int\mathbf{e}_1'M^{-1}u^{S_p}K(u)\,du=\mathbf{e}_1'M^{-1}\int u^{S_p}K(u)\,du=1$.

Now we discuss the asymptotic properties of $\widehat{\beta}_0^{(1)}$ and $\widehat{\beta}_0^{(2)}$. $\widehat{\beta}_0^{(1)}=\overline{y}-\overline{x}'\widehat{\delta}$, so

$$\widehat{\beta}_0^{(1)}-\beta_0=-\overline{x}'\left(\widehat{\delta}-\beta_1\right)+\overline{\varepsilon}.$$

If $E[x]\neq 0$, its asymptotic distribution is the same as $-\overline{x}'\left(\widehat{\delta}-\beta_1\right)$ or $-E[x]'\left(\widehat{\delta}-\beta_1\right)$. If $E[x]=0$,

$$\sqrt{n}\left(\widehat{\beta}_0^{(1)}-\beta_0\right)=-\left(\sqrt{n}\overline{x}'\right)\left(\widehat{\delta}-\beta_1\right)+\sqrt{n}\overline{\varepsilon}=\sqrt{n}\overline{\varepsilon}+o_p(1).$$

Given that $\widehat{\beta}_0^{(2)}-\beta_0=\widehat{A}-\beta_0-\overline{x}'\widehat{\beta}_1=\left(\widehat{A}-\beta_0-\overline{x}'\beta_1\right)-\overline{x}'\left(\widehat{\beta}_1-\beta_1\right)$ and $\widehat{A}-\beta_0-\overline{x}'\beta_1$ has the same asymptotic distribution as $\overline{\varepsilon}$, $\widehat{\beta}_0^{(2)}-\beta_0$ has the same asymptotic distribution as $\widehat{\beta}_0^{(1)}-\beta_0$.

As to $\widetilde{\beta}$, note that

$$\begin{aligned}\sqrt{n}\left(\widetilde{\beta}-\beta\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\left(\widehat{a}(x_i)-\mathbf{x}_i\beta\right)\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\frac{1}{\sqrt{n^3}}\sum_{i=1}^{n}\mathbf{x}_i\sum_{j=1,j\neq i}^{n}\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i\varepsilon_j+o_p(1).\end{aligned}$$

We first derive the asymptotic distribution of $\frac{1}{\sqrt{n^3}}\sum_{i=1}^{n}\mathbf{x}_i\sum_{j=1,j\neq i}^{n}\mathbf{e}_1'\left(\overline{M}_i\right)^{-1}z_j^i w_j^i\varepsilon_j$, which is a second-order

U-statistic. From Lemma 8.4 of Newey and McFadden (1994), this U-statistic is asymptotically equivalent to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_n(x_i, \varepsilon_i)$, where

$$m_n(x_j, \varepsilon_j) = E\left[\mathbf{x}_i \mathbf{e}_1' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i \varepsilon_j \middle| x_j, \varepsilon_j\right] = \varepsilon_j \int \mathbf{x}_i \mathbf{e}_1' \left(\overline{M}_i\right)^{-1} z_j^i w_j^i f(x_i) dx_i.$$

We apply the Liapunov central limit theorem to derive its asymptotic distribution. It can be shown that

$$E\left[m_n(x_i, \varepsilon_i) m_n(x_i, \varepsilon_i)'\right] \longrightarrow E\left[\mathbf{x}_i \mathbf{x}_i' \varepsilon_i^2\right],$$

so the asymptotic distribution of $\widetilde{\beta}$ is the same as that of the LSE. ■