The Korean Economic Review
Volume 30, Number 2, Winter 2014, 301-331.

Square Density Weighted Average Derivatives Estimation of Single Index Models

Myung Jae Sung*

This paper proposes an average derivatives estimator for index coefficients under a single index model, which does not require restrictive conditions such as zero boundary density or density trimming that are often adopted in previous studies including Powell, Stock, and Stoker (1989, PSSE) and Härdle and Stoker (1989, HSE), among others. Coefficients are consistently estimable by nonparametric mean regression with square density weighted average derivatives (SWADE). Relaxed requirements for SWADE allow more general applications. The asymptotic distribution of SWADE is equivalent in precision to the aforementioned average derivatives estimators (PSSE and HSE). Monte Carlo simulations show that SWADE outperforms HSE in finite sample but is slightly and weakly outweighed by PSSE. These imply that SWADE allows more flexible applications with relaxed distributional characteristics than PSSE and HSE at the expense of slightly deteriorated behavior in finite sample.

JEL Classification: C12, C13, C14
Keywords: Index Coefficients, Square Density Weighting, Average Derivatives, Kernel, Nonparametric

I. Introduction

This paper proposes a nonparametric method of estimating coefficient vector consistently under a single index model framework, which does not require restrictive conditions such as zero boundary density or density trimming that are often adopted in previous studies including Powell, Stock, and Stoker (1989) and Härdle and Stoker (1989), among others. Coefficients are consistently estimable by nonparametric mean regression with square density weighted average derivatives.

Let Y denote a dependent variable and X independent variables with the

Received: Jan. 3, 2014. Revised: Sept. 12, 2014. Accepted: Sept. 25, 2014.

* Professor, School of Economics, Hongik University, 94 Wausan-ro, Mapo-gu, Seoul 121-791, Korea, Phone: +82-2-320-1814, E-mail: mjaesung@hongik.ac.kr. I would like to thank two anonymous referees for helpful comments and suggestions. Any remaining error, if any, is of course mine.

regression function of $E(Y \mid X) = G(X)$. Consider a model where the conditional expectation is explained by the single index such that

$$G(X) = g(X'\beta) \tag{1}$$

where the random variable $X \in \ddot{\mathrm{X}}(\subset R^p)$ is generated from an absolutely continuous distribution and $g: R^1 \to R^1$ is a (possibly unknown) real-valued function. The vector X has no intercept, as is embedded in $g(\cdot)$. Alternatively, the single index model can be written in the following form

$$Y = g(X'\beta) + \varepsilon \tag{2}$$

where $\varepsilon \equiv Y - G(X) = Y - g(X'\beta)$. The regression model specified as Equation (1), alternatively written as Equation (2), is attractive for its wide range of applications such as linear models, 'Tobit' models, binary choice models, duration models, reduced forms of simultaneous equations models, index models, etc.[1]

In many semiparametric regression analyses, single index models are quite often adopted and simple to estimate the parameter vector β. If $g(\cdot)$ is smooth enough, β can be consistently estimable. Studies by Powell, Stock, and Stoker (1989, PSS hereafter), Härdle and Stoker (1989, HS hereafter), Stoker (1991), Park (1990), Han (1987a,b), Ichimura (1993), etc. have focused on this moment condition and contributed to estimating β consistently. Furthermore, Xia (2006) proposed alternative estimators: outer product of gradients estimator and minimum average variance estimator, by lowering the dimension of the kernel smoothing in the aforementioned studies.

Identification of β in Equation (1), under the single index model framework,[2] requires some regularity conditions on $G(\cdot)$ or $g(\cdot)$ unless $g(\cdot)$ is affine.

[1] Equation (2) also applies to a generalized regression model specified as $Y = h(X'\beta + u)$ for some $h: R^1 \to R^1$, if certain conditions are imposed on (X, u) or $h(\cdot)$. An interesting finding from this argument is that identification of β under the index model framework crucially depends on the relation between regressors and error terms. An example is the case where X and u are statistically independent. Then, $E(Y \mid X)$ can be written as a function of $X'\beta$ and Equation (2) is applicable for $\varepsilon = Y - E(Y \mid X) = Y - h^*(X'\beta)$ for some real-valued function $h^* R^1 \to R^1$. However, if X and u are not statistically independent, the conditional mean is generally not a function of a single index. There exists an exception: if ε is dependent on X only through $X'\beta$, the conditional expectation also becomes a function of $X'\beta$.

[2] There have been a number of studies involved in model checking of single index models. They are Fan and Li (1996), Aït-Sahalia, Bickel, and Stoker (2001), Xia, Li, Tong, and Zhang (2004), and Stute and Zhu (2005), among others. Fan and Li (1996) and Aït-Sahalia et al. (2001) proposed tests of single index models based on residual sum of squares. Xia et al. (2004) and Stute and Zhu (2005), together with Aït-Sahalia et al. (2001), proposed goodness-of-fit type test statistics. However, all of these suffer either the slower convergence rates than $N^{-1/2}$ or unknown asymptotic bias to be estimated (see Aït-Sahalia et al., 2001).

Coefficient vector, β , can be identified by differentiating the conditional expectations. Suppose that $G(\cdot)$ is continuously differentiable and that $\nabla G(X) \neq 0$. Then, the derivatives of conditional expectations are proportional to β :

$$\frac{\partial E(Y \mid X)}{\partial X} = \nabla G(X) = g'(X^t \beta)\beta \,. \tag{3}$$

An average of Equation (3) weighted by any nonzero functions of X is also proportional to β . Define a new parameter of average derivatives weighted by square densities of the form

$$\delta \equiv E\left(f^2(X) \cdot \frac{\partial E(Y \mid X)}{\partial X} \right) = E(f^2(X) \cdot \nabla G(X)) = E(f^2(X) \cdot g'(X^t \beta))\beta \tag{4}$$

where $f(X)$ is the density of X. δ is proportional to β. In general, $E(f^2(X) \cdot g'(X^t \beta))$ is unknown. Thus, β can be identified only up to a multiplicative constant.[3] As discussed later, the choice of square density weighting is adopted to avoid density trimming or restrictive applications of zero density particularly on the boundary in the support of regressors, X. Average derivatives estimators (ADEs) established in previous studies by PSS (1989) and HS (1989) require zero boundary density conditions. The ADE proposed by HS (HSE) depends on the density condition which is assumed to be zero at the boundary of X. HSE inevitably involves density estimates in the denominators of the derivative estimates and, thus, may exhibit erratic behavior when the values of density estimates are very small particularly near the boundary of the support of X.[4] To avoid this problem, they introduce density trimming. PSS proposes a density-weighted ADE (PSSE) of coefficient vector. They introduce density-weighting to avoid potential erratic behavior of the derivatives. However, PSSE requires zero boundary density condition for consistent estimation. These ADEs may be restrictively applicable because they preclude the distributions with boundary densities bounded away from zero.[5]

Unlike other ADEs which require density trimming and zero boundary density as in PSS or HS, square density weighting is often useful by relaxing conditions other than zero boundary density or density trimming. The square density weighted

[3] Coefficient parameters are identified only up to scale through the ADEs, due to unknown function of conditional expectation $g(\cdot)$ specified in Equation (2). Thus, coefficient estimates are often scale-normalized in many empirical studies such that its first component is set to be unity without loss of generality.

[4] Refer to the second and third paragraphs for more in greater details on p. 988 of HS (1989).

[5] Refer to Assumption 2 of PSS (1989) and Appendix A.1 of HS (1989).

average derivatives are free from restrictive density conditions or density trimming embedded in denominators of the derivatives written in fractions because the denominators are totally removed by square density weighting. This is discussed more precisely in Section II.

Section II proposes a squared density weighted average derivatives estimator (SWADE) of β (or, equivalently δ) up to a multiplicative constant, denoted by $\hat{\delta}$.[6] This provides a nonparametric method of estimating δ, which differs from those studied by Ichimura, PSS, HS, Stoker, Han, and Xia. Section II also lists sufficient assumptions for the asymptotic normality of the estimator (SWADE). Section III presents the asymptotic behavior of SWADE; its $\sqrt{N}$ − consistency and asymptotic normal distribution. Precision of the estimators of β is also compared with the other ADEs proposed by PSS and HS. Section IV presents some concluding remarks.

A superscript ‘t’, a prime symbol, and ‘∇’ denote a transpose of a vector (or a matrix), a derivative, and a gradient throughout the paper, respectively.

II. Estimator and Assumptions

II.1. The Estimator

As shown in Equation (3), the derivatives of conditional expectations are proportional to β. Furthermore, the square density weighted derivatives $f^2(X)\nabla G(X) = f^2(X)g'(X^t\beta)\beta$ are also proportional to β. The scale of β is not identified in both cases, if $g'(\cdot)$ is unknown. Redefine the parameters as $\delta \equiv E(\delta(X))$.

Consider a nonparametric kernel estimator of the conditional expectation $E(Y \mid X_i) = G(X_i)$:

$$\hat{E}(Y \mid X_i) = \hat{G}(X_i) \equiv \frac{\hat{T}(X_i)}{\hat{f}(X_i)}$$

where $\hat{T}(X_i) \equiv \frac{1}{Nh^p} \Sigma_{j \neq i} K\left(\frac{X_i - X_j}{h}\right) Y_j$ and $\hat{f}(X_i) \equiv \frac{1}{Nh^p} \Sigma_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)$.

[6] This paper proposes an estimator for β up to scale in the form of average derivatives estimator (ADE). An alternative is a semiparametric M-estimator (e.g. Ichimura, 1993, or Chen et al., 2014). ADE may have fewer advantages than the semiparametric M-estimators, in particular efficiency and more restrictive assumptions such as the continuous differentiability of the function $g(\cdot)$ in Equation (1). However, according to Horowitz (2009), the semiparametric M-estimators generally have computational burden due to optimization processes, and, sometimes need density trimming (Ichimura and Todd, 2007).

$K(\cdot)$ is a weight function (or a kernel) and $h(>0)$ is a bandwidth sequence which declines at a suitable rate to zero as $N \to \infty$.[7] Differentiate $\hat{G}(X_i)$ with respect to X_i:

$$\nabla\hat{G}(X_i) = \frac{\frac{1}{N^2h^{2p+1}}\{\sum_{j\neq i}\sum_{k\neq i}\nabla K(\frac{X_i-X_j}{h})Y_jK(\frac{X_i-X_k}{h}) - \sum_{j\neq i}\sum_{k\neq i}K(\frac{X_i-X_j}{h})Y_j\nabla K(\frac{X_i-X_k}{h})\}}{\hat{f}^2(X_i)}$$

$$= \frac{\frac{1}{N^2h^{2p+1}}\sum_{j\neq i}\sum_{k\neq i}Y_j\{\nabla K(\frac{X_i-X_j}{h})K(\frac{X_i-X_k}{h}) - K(\frac{X_i-X_j}{h})\nabla K(\frac{X_i-X_k}{h})\}}{\hat{f}^2(X_i)}$$

The denominator $\hat{f}^2(X_i)$ of $\nabla\hat{G}(X_i)$ is a scalar. Thus, the numerator of $\nabla\hat{G}(X_i)$, i.e. $\hat{f}^2(X_i)\nabla\hat{G}(X_i)$, is proportional to the derivative estimate $\nabla\hat{G}(X_i)$. Take $\hat{\delta}(X_i) \equiv \hat{f}^2(X_i)\nabla\hat{G}(X_i)$ as the estimator of $\delta(X_i) = f^2(X_i)\nabla G(X_i)$. Observe that the terms in which $j = k$ are zero. Without loss of generality, drop these terms and let $\delta(X_i)$ be of the following form:

$$\hat{\delta}(X_i) = \frac{1}{N^2h^{2p+1}}\sum_{j\neq i}\sum_{\substack{k\neq i\\ k\neq j}} Y_j\left\{\nabla K\left(\frac{X_i-X_j}{h}\right)K\left(\frac{X_i-X_k}{h}\right)\right.$$
$$\left. -K\left(\frac{X_i-X_j}{h}\right)\nabla K\left(\frac{X_i-X_k}{h}\right)\right\}. \quad (5)$$

Let $q \equiv (Y, X, \varepsilon)$ where $\varepsilon \equiv Y - G(X)$; $E(\varepsilon\,|\,X) = 0$. By taking an average of $\delta(X)$ over N i.i.d. observations, we can propose a squared density weighted average derivatives estimator of δ as

$$\hat{\delta} \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\delta}(X_i) = \frac{1}{N}\sum_{i=1}^{N}\hat{f}^2(X_i)\nabla\hat{G}(X_i) = \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{\substack{k\neq i\\ k\neq j}}\hat{s}(q_i, q_j, q_k) \quad (6)$$

where $\hat{s} \equiv \hat{s}(q_i, q_j, q_k) = \hat{s}_1 + \hat{s}_2 + \hat{s}_3$,

$$\hat{s}_1 \equiv \frac{1}{6h^{2p+1}}(Y_j - Y_k)\left\{\nabla K\left(\frac{X_i-X_j}{h}\right)K\left(\frac{X_i-X_k}{h}\right) - K\left(\frac{X_i-X_j}{h}\right)\nabla K\left(\frac{X_i-X_k}{h}\right)\right\},$$

$$\hat{s}_2 \equiv \frac{1}{6h^{2p+1}}(Y_k - Y_i)\left\{\nabla K\left(\frac{X_j-X_k}{h}\right)K\left(\frac{X_j-X_i}{h}\right) - K\left(\frac{X_j-X_k}{h}\right)\nabla K\left(\frac{X_j-X_i}{h}\right)\right\},$$

$$\hat{s}_3 \equiv \frac{1}{6h^{2p+1}}(Y_i - Y_j)\left\{\nabla K\left(\frac{X_k-X_i}{h}\right)K\left(\frac{X_k-X_j}{h}\right) - K\left(\frac{X_k-X_i}{h}\right)\nabla K\left(\frac{X_k-X_j}{h}\right)\right\}.$$

[7] $K(\cdot)$ and h must satisfy a couple of conditions for the consistency and asymptotic normality of $\hat{\delta}$ to be proposed below. The conditions are discussed more in detail in Assumptions 4 and 5 in Section III.

II.2. Assumptions

Five assumptions for the proper asymptotic behavior of $\hat{\delta}$ are presented here. Assumptions 1 and 2 give the regularity conditions on X. Assumption 3 gives an identification condition for δ. Assumption 4 presents a higher order kernel for asymptotic bias reduction of the estimator. Assumption 5 shows the conditions the bandwidth sequence h must satisfy for the $\sqrt{N}-$consistency of $\hat{\delta}$.

Assumption 1:
(a) (X,Y) is a random pair which is continuous on R^{p+1} where $p \geq 2$. The underlying measure υ is a Lebesgue measure on R^{p+1}.
(b) $(X_i,Y_i), i=1,2,\cdots,N$ are a random sample from the population.

Assumption 2:
The density $f(X)$ of X is bounded and $(\lambda+1)-$times continuously differentiable with bounded derivatives in the support $\ddot{\mathrm{X}}$ for some λ satisfying $\lambda > 2p+1$.

Assumption 3:
$G: R^p \rightarrow R^1$ where $G(X)=E(Y\,|\,X)$ is bounded and $(\lambda+1)-$ times continuously differentiable with bounded derivatives.

Assumption 3 implies that $g: R^1 \rightarrow R^1$ is bounded and $(\lambda+1)-$times continuously differentiable with bounded derivatives where $G(X)=E(Y\,|\,X)=g(X'\beta)$.

Assumption 4:
Let the set of kernels, $K_{p,\lambda}$, be the class of all measurable bounded real-valued functions $K: R^p \rightarrow R^1$ such that $|K(\cdot)| < K^* < \infty$. For $u \in R^p$ and $(r_1, r_2, \cdots, r_p) \in N^p$, $K(\cdot)$ satisfies

(a) $\int K(u)du = 1$,
(b) $\int K^2(u)du < \infty$ and $\int \nabla K(u) \nabla K(u)^t du < \infty$,
(c) $K(u)=K(-u)$,
(d) $\nabla K(u) \rightarrow 0$ and $K(u)u \rightarrow 0$, as $u \rightarrow \pm\infty$,
(e) $\int K(u) u_1^{r_1} \cdots u_p^{r_p} du \begin{cases} = 0 & if \;\; 0 < r_1 + \cdots + r_p < \lambda \\ < \infty & if \;\; 0 < r_1 + \cdots + r_p = \lambda \end{cases}$.

Assumption 5:
$N^{1-e}h^{4p+2} \rightarrow \infty$ and $Nh^{2\lambda} \rightarrow 0$ as $N \rightarrow \infty$ for some $e > 0$.

ADEs are often simple to understand and easy to calculate. Note that the estimators derived by average derivatives require differentiability of (conditional) expectation function denoted by $g(\cdot)$ in Equation (2) and continuity of regressors. One of their disadvantages is that they are not applicable to discrete regressors because $g(\cdot)$ is not differentiable with respect to discrete variables. In this sense, all ADE-type estimators for single index models such as PSSE and HSE, of course, including SWADE proposed in this paper assume that regressors are continuous and $g(\cdot)$ is continuously differentiable. Otherwise, different methods need to be posed for identification of index coefficients. The estimation methods proposed by Ichimura (1993) or other M-estimators such as maximum score estimators need to be pursued.

III. Asymptotic Behavior of the Estimator for δ

III.1. Asymptotic Behavior

This section presents the asymptotic behavior of $\hat{\delta}$.

Define $\varphi(X) \equiv f(X)E(\varepsilon^2 \mid X)$ and $Z \equiv \delta(X) - \varepsilon f(X)\nabla f(X)$ where $\delta(X) = f^2(X)\nabla G(X)$ and

$$\begin{aligned}\Sigma \equiv Var(Z) &= E\{\delta(X)\delta(X)' + E(\varepsilon^2 \mid X)f^2(X)\nabla f(X)\nabla f(X)'\} - \delta\delta' \\ &= E\{\delta(X)\delta(X)' + \varphi(X)f(X)\nabla f(X)\nabla f(X)'\} - \delta\delta'\end{aligned}$$

$\hat{\delta}$ is $\sqrt{N}-$consistent for δ, and its asymptotic variance is the same as that of $3Z$. The following two theorems show its asymptotic normality and the consistency of variance matrix.

Theorem 1. Given Assumptions 1 through 5 as stated in Section II, $\sqrt{N}(\hat{\delta} - \delta)$ has a limiting normal distribution with zero mean and variance 9Σ:

$$\sqrt{N}(\hat{\delta} - \delta) \rightarrow_d N(0, 9\Sigma).$$

Generally speaking, the convergence rate of a nonparametric (kernel) conditional mean estimator (e.g. $\hat{G}(X)$) is $\sqrt{Nh^p}$ and that of its derivative is $\sqrt{Nh^{p+2}}$. This is because the derivative estimator [e.g. $\hat{f}^2(X)\nabla\hat{G}(X)$] is more slowly convergent. However, the convergence rate of $\hat{\delta}$ is $\sqrt{N}$. This is because $\nabla\hat{G}(X)$ [or $\hat{f}^2(X)\nabla\hat{G}(X)$] is taken to be averaged over the sample and because the projection of a U-statistic [see Lemma 2 in Appendix A] of the estimator accelerates the rate of convergence to $\sqrt{N}$, as is usual in i.i.d. sample averages.

We now propose a consistent estimator of the asymptotic variance. Let

$$\nabla\hat{f}(X_i) \equiv \frac{1}{Nh^{p+1}} \sum_{j\neq i} \nabla K\left(\frac{X_i - X_j}{h}\right),$$

$$\hat{\varphi}(X_i) \equiv \frac{1}{Nh^p} \sum_{j\neq i} \hat{\varepsilon}_j^2 K\left(\frac{X_i - X_j}{h}\right) \text{ where } \hat{\varepsilon}_j = Y_j - \hat{G}(X_j).$$

Also, let the variance estimator be

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}[\hat{\delta}(X_i)\hat{\delta}(X_i)' + \hat{\varphi}(X_i)\hat{f}(X_i)\nabla\hat{f}(X_i)\nabla\hat{f}(X_i)'] - \hat{\delta}\hat{\delta}'.$$

Theorem 2. Under Assumptions 1 through 5 in Section II, $\hat{\Sigma}$ is a consistent estimator of Σ.

III.2. Comparison of Precision

Throughout this section, the conditional expectation is assumed to be a function of single index, $X'\beta$, and average derivatives estimators for β are compared. Alternative estimators for β , other than δ , are derived from integrating $E[\nabla G(X)]$ in HS or $E[\nabla G(X)f(X)]$ in PSS and Stoker by parts and from setting the boundary density of X equal to zero:

$$E[\nabla G(X)] = G(X)f(X)\Bigg|_{X=-\infty}^{X=+\infty} - \int G(X)\nabla f(X)dX$$
$$= -E\left(G(X)\frac{\nabla f(X)}{f(X)}\right) = -E\left(Y\frac{\nabla f(X)}{f(X)}\right)$$
$$E[\nabla G(X)f(X)] = G(X)f^2(X)\Bigg|_{X=-\infty}^{X=+\infty} - 2E[G(X)\nabla f(X)] = -2E[Y\nabla f(X)]$$

since $f(\pm\infty) = 0$ and $|G(X)| < \infty$ for all X . Thus, the structures of their estimators depend on the boundary density condition. However, that of $\hat{\delta}$ does not depend on the density on the boundary of support. This is because $\hat{\delta}$ is derived from differentiating the kernel estimator of expectations and weighting it by the squared density of X , and because we avoid the method of integration by parts employed for identification of the index coefficient in PSS as well as in HS. Therefore, $\hat{\delta}$ is free from the boundary density condition and thus, we have not imposed any restrictions on the boundary density of X.

In regard to the precision of estimators for β, the estimator of HS (HSE) and those of PSS (PSSE) and Stoker (SE) are asymptotically equivalent to $r(X,Y)+o_p(N^{-1/2})$ and $\rho(X,Y)=2f(X)r(X,Y)+o_p(N^{-1/2})$, respectively, where $r(X,Y)=g^{*\prime}(X^t\beta)\beta-(Y-g^*(X^t\beta))\frac{\nabla f(X)}{f(X)}=\frac{Z}{f^2(X)}$ for Z defined earlier (see HS p.988, PSS p.1412 or Stoker p.104). The asymptotic variances of HSE and PSSE (or SE) are $Var[r(X,Y)]=Var(\frac{Z}{f^2(X)})$ and $Var[2\rho(X,Y)]=4Var(\frac{Z}{f^2(X)})$, respectively, while that of $\hat{\delta}$ is $Var(3Z)$; the only difference among those is the weights: 1 in HSE, $f(X)$ in PSSE (or SE), and $f^2(X)$ in $\hat{\delta}$. This does not make any qualitative difference in the measurement of precision as well as in the inference on the hypotheses about β.

The choice of weights sometimes has different effects on the classical U-statistic structure of estimator in deriving asymptotic properties. As noted in PSS (1989, in the last paragraph of p. 1424), density weighting gives a simpler U-statistic structure with double summations as shown in Equation (3.4) of PSS (1989), whilst square density weighting gives a more complicated U-statistic structure with triple summations as shown in Equation (6).

III.3. Finite Sample Behavior

III.3.1. Model Specifications

Despite their asymptotic equivalence of precision, the performance of PSE, HSE, and SWADE in finite sample could be different. This is investigated for the case where $p=3$ for many different models consistent with "single index" specification as shown in Equation (1).[8] This study replicates the models specified by PSS (1989):

$$Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

where $\alpha=0$ and $(\beta_1,\beta_2,\beta_3)=(1,2,3)$. As in PSS (1989), the models considered here are linear or binary choice models with homoskedastic or heteroskedastic errors. There are four combinations of models and errors. In addition, the simulations are repeated for two types of regressors; the one is for zero densities and the other for strictly positive densities, both at the boundaries of the support. Therefore, the simulations are conducted for the total number of eight different combinations of

[8] A slightly different case is also discussed in Appendix C for a simple linear regression, where the variance of error term is smaller than those of regressors such that $\varepsilon_i \sim N(0,0.02^2)$ where X has zero mean and unit variance. This is to compare the performance of ADEs under different model specifications with different error specification from the above case.

models, errors, and regressor types.

In the linear models, Y_i^* is observed as it is. In the binary choice models, the latent variable, Y_i^*, is observed as either zero or one such that $Y_i = 1[Y_i^* > 0]$ where an indicator function, $1[\cdot]$, is one when the argument is true and zero, otherwise.

For homoskedastic models, ε_i is assumed to be standard normal and independent of regressors. For heteroskedastic models, ε_i is multiplicatively heteroskedastic such that $\varepsilon_i = \sigma_i \cdot v_i$ where v_i is an i.i.d. standard normal sequence and $\sigma_i^2 = \exp(X_i^t \beta + k)$, where k is a constant chosen so that $E(\sigma_i^2) = 1$, given the distribution of $X_i = (x_{1i}, x_{2i}, x_{3i})$. Here, elements of X are denoted by lower case letters to distinguish from the i-th observations of X for $i = 1, 2, \cdots, N$ in the sample which is denoted by upper case letters in this paper.

The consistency and finite sample behavior of ADEs depend on the distributional characteristics of regressors. In particular, both PSSE and HSE require the condition of zero boundary densities of regressors. Otherwise, the coefficient parameter of interest cannot be identified and their proposed estimators are no longer consistent for the index parameter of interest. The consistency of SWADE does not depend on the boundary densities of regressors. In this sense, SWADE outweighs PSSE and HSE in terms of consistency of the estimators particularly when the zero boundary density condition is not satisfied. However, their finite sample behaviors may be different. In these regards, Monte Carlo experiments are conducted under the two distributional specifications of X.

In the first specification of regressors, X is assumed to have zero boundary densities as in PSS (1989). More specifically, it is assumed that x_{1i} is distributed Chi-square with degrees of freedom 2, that x_{2i} is standard normal, and that x_{3i} follows gamma distribution with parameters, theta=2 and alpha=18. x_{1i} and x_{3i} are standardized to have zero mean and unit variance.

In the second specification, the density of X is assumed to be strictly positive and, thus, bounded away from zero everywhere in the support. Note that the most important identification condition of zero boundary density is violated for PSSE and HSE in this specification. Assume that x_{1i} is distributed uniform such that $x_{1i}^* = u_{1i}$, and let $x_{2i}^* = u_{2i}^2$ and $x_{3i}^* = \frac{1}{u_{3i}+1}$ where $u_{li} \sim U(0,1)$ for $l = 1, 2, 3$, which denote mutually independent uniform distributions defined over the interval (0,1). $(x_{1i}^*, x_{2i}^*, x_{3i}^*)$ are also standardized.[9]

Suppose that (X_i, Y_i), $i = 1, 2, \cdots, 300$ are a random sample of size 300 where $p = 3$, and that the researcher cannot observe ε but only the realizations of (X, Y). The researcher knows that the conditional expectation of Y given X is a function of an index, $g(X^t \beta)$, for some unknown continuously differentiable

[9] Note that $E(x_{1i}^*) = 0.5$, $Var(x_{1i}^*) = \frac{1}{12}$, $E(x_{2i}^*) = \frac{1}{3}$, $Var(x_{2i}^*) = \frac{4}{45}$, $E(x_{3i}^*) = \ln(2)$, and $Var(x_{3i}^*) = \frac{1}{2} - (\ln(2))^2$.

$g(\cdot)$.

With the constructed samples, the parameter of interest β (that is, δ) is estimated by the proposed estimator, SWADE defined in Equation (6) for each of the above eight combinations of models, errors, and specifications of regressors. Furthermore, it is also estimated by the methods proposed by PSS and HS to compare their finite sample behaviors.[10] In addition, a couple of parametric regressions are also compared. They are ordinary least squares (OLS) and Probit estimators. The OLS estimator known as the best linear unbiased estimator (BLUE) is compared for all of eight combinations of models, errors, and specifications of regressors, to provide a baseline or a standard for comparison of behavior between estimators, although it is biased under the binary choice models. The Probit estimates are provided for binary choice models.

The above process is repeated four hundred times independently to construct empirical distributions of the aforementioned estimators.

Tables 1 through 4 report summary statistics for aforementioned ADEs, OLS and Probit estimators under the four different combinations of models and errors between linear and binary choice models and between homoskedastic and heteroskedastic errors under the first specification of regressors. Tables 5 through 8 report the same statistics under the second specification of X. As in PSS (1989), reported summary statistics include sample mean (MEAN), standard deviation (SD), root mean squared error (RMSE), quartiles (LQ, Median, and UQ), and median absolute error (MAE) for both specifications.

The aforementioned ADE methods can identify the parameter of interest only up to scale. For a legitimate comparison, the coefficient estimates need to be normalized. The parameter estimates are normalized/rescaled by dividing them by the average value of the first component of the parameter, δ, for simplicity of discussion.[11] Note that the normalized sample mean of the first component ($\hat{\beta}_1$ or

[10] The following higher order Gaussian kernel is used for all of ADEs in the Monte Carlo simulations:

$$K(u) \equiv \sum_{m=1}^{\lambda} \frac{a_m}{(\sqrt{2\pi})^p \, |b_m|^p} \exp\left[-\frac{u'u}{2b_m^2} \right]$$

where a_m and b_m ($m = 1,2,\cdots,\lambda$) satisfy the following two conditions: $\sum_{m=1}^{\lambda} a_m = 1$ and $\sum_{m=1}^{\lambda} a_m b_m^{2l} = 0$ for $l = 1,2,\cdots,\lambda - 1$. (See Bierens, 1987, p. 112.) Smoothing parameters are specified as follows. Since $p = 3$ in this example, set $\lambda = 8$ according to Assumption 2. We specify $b_m = m^{-1/2}$ and solve the above linear equations system for a_m. As in PSS (1989), let $h = 1$. This Gaussian kernel is also used in the simulations for the estimates of PSS and HS.

[11] If the parameter estimates are divided by the corresponding estimates of the first component (β_1, or δ_1), the absolute values of the estimates of the other components, (β_2, β_3) or (δ_2, δ_3), tend to be inflated and, thus, result in biased inference. Although the true parameter is ($\beta_1, \beta_2, \beta_3$) where $c_2 = \beta_2 / \beta_1$ and $c_3 = \beta_3 / \beta_1$, the averages of the estimates of the second and third elements divided by the estimate of the first element tend to be larger than c_2 and c_3 in its absolute values,

$\hat{\delta}_1$) is always unity, since the estimates are rescaled by the absolute values of the average estimates of the first component (β_1, i.e., δ_1).

III.3.2. Specification 1: Zero Boundary Densities

In this section, the simulation results for the regressors with zero boundary densities are discussed. The results are reported in Tables 1 through 4. Appropriately specified parametric regressions provide best performance in finite sample among all; the OLS estimates perform best for linear models and the Probit estimates for binary choice models regardless of error specifications. The OLS estimates are biased for binary choice models. Unlike the misspecification biases of the OLS estimates particularly for binary choice models, nonparametric ADEs including PSSE, HSE and SWADE are robust to diversified parametric specifications of single index models. However, their estimates have larger variability than those of other parametric estimators.

Specification of zero boundary densities of regressors together with other regularity conditions allows consistency of PSSE and HSE. However, HSE without density trimming seems to show relatively erratic behavior reflected by much larger values of the SD, RMSE, and MAE than those of other ADEs. In particular, HSE behaves very poorly for binary choice models. Both the variability measures and absolute mean deviations from the prespecified parameter values of HS estimates are largest among all ADEs without an exception throughout all combinations of models and errors. In this regard, HSE without density trimming is not sufficiently well-behaved in finite sample, compared with other ADEs. In this sense, SWADE outweighs HSE in terms of finite sample behavior.

Table 1 reports the simulation results for a homoskedastic linear model. The OLS estimator performs best among all; its SD, RMSE, quartile range (=UQ-LQ), and MAE are smallest. The average values of rescaled parameter estimates for the second and third components (β_2, β_3) (i.e., (δ_2, δ_3)) are close to the true parameter values (2,3) among all ADEs except for $\hat{\delta}_3$ of PSSE. SWADE is slightly poorer than PSSE in terms of variability measures; PSSE has smaller values of the SD, RMSE, quartile range, and MAE for $\hat{\delta}_1$ through $\hat{\delta}_3$. However, the absolute deviations of sample means from the prespecified parameter values of (δ_2, δ_3)=(2,3) are larger for PSSE than for SWADE, although the difference between them seems small. In this sense, the finite sample behavior seems indifferent between PSSE and SWADE under the homoskedastic linear model.

For a binary choice model with a homoskedastic error reported in Table 2, the Probit estimator performs best in finite sample, among all the consistent estimates,

respectively. To avoid this scale problem, the estimates should be divided by the average values of the first component estimates instead of its estimate itself. To maintain the signs of the estimates, only the absolute values of the averages are used.

in terms of absolute deviations of sample means from the prespecified parameter values of (1,2,3). It is remarkable that the finite sample behaviors of the OLS estimates are best among all, despite their asymptotic inconsistency for binary choice models unlike others. Although this appears weird, biased estimates may sometimes outperform consistent ones in a particular finite sample. This is quite possible particularly when the biases of biased estimators are smaller than those of consistent ones in particular finite samples. Note that the asymptotic biases of the former do not vanish unlike that of the latter. The biasedness of OLS estimates for binary choice models causes a noticeable increase in the variability of estimates; for example, the SD for δ_3 increases from 0.060 in Table 1 to 0.184 in Table 2. The shift from linear to binary choice models increases the absolute deviations of sample means of all average derivatives estimates from true parameter values of (1,2,3) as well as the variability measures of the SD, RMSE, and MAE, both significantly. The "MEAN" column indicates that the ADEs for β_2 are relatively close to their true value of 2 except for HSE; PSSE gives the smallest bias for β_2 (0.242=2.242-2). However, its median value is not as close to two as that of SWADE: 1.832 (SWADE) vs. 1.772 (PSSE). In addition, the variability measures for SWADE are significantly smaller than those for PSSE. For δ_3, SWADE shows the best results among all ADEs in terms of mean deviations as well as variability measures. However, the differences are not significantly noticeable. In some sense, SWADE seems almost as equivalent as PSSE in terms of finite sample behavior.

Simulation results for heteroskedastic linear and binary choice models are reported in Tables 3 and 4, respectively. The variability measures (SD, RMSE, and MAE) of ADEs for heteroskedastic errors are significantly smaller than those for homoskedastic errors, particularly for SWADE and PSSE.[12] The sample mean of SWADE for $(\beta_2, \beta_3) = (1.874, 2.937)$ is slightly closer to the true value (2,3) than that of PSSE (1.801,2.894) for linear models. However, the latter (2.219,3.360) is slightly closer than the former (2.271,3.409) for binary choice models. The variability measures for SWADE are slightly smaller than those for PSSE in linear models. This inequality is reversed for binary choice models between Tables 3 and 4. HSE without density trimming for heteroskedastic errors is not sufficiently well-behaved particularly for binary choice models; the mean deviations from true parameter value for (β_2, β_3) are noticeably large with much larger variability measures than other ADEs. As discussed earlier, this is probably because the erratic behavior is caused by the infinitesimal density estimates in the denominator of HSE particularly near the boundaries of regressors with zero density.

Based on the results of simulations, ADEs seem to yield similar parameter

[12] An example for this case can be found in PSS (1989). The SD and RMSE of PSSE are smaller for heteroskedastic errors than for homoskedastic ones both for binary choice models. See Tables II and IV of PSS (1989).

estimates except HSE; the average values of parameter estimates are close to the true parameter value, (1,2,3), and the biases are not noticeably large. The standard deviations of the estimates are slightly different over the three estimators. PSSE has smaller values of SD, RMSE, and MAE than SWADE for homoskedastic linear models. SWADE has smaller variability measures than PSSE in other models. HSE has larger values of SD, RMSE and MAE than PSSE as well as SWADE. Although its behavior for homoskedastic linear models in finite sample is slightly and weakly outweighed by that of PSSE, SWADE seems relatively well-behaved for models with heteroskedastic errors.

In sum, SWADE is asymptotically equivalent in precision to the established average derivatives estimators of PSSE and HSE. SWADE allows flexible applications with less restrictive distributional characteristics than PSSE and HSE at the expense of slightly deteriorated finite sample behavior particularly for homoskedastic errors. In other words, there exists a trade-off between wider applications and finite sample behavior.

[Table 1] Finite-sample Behavior of Estimators for Homoskedastic Linear Model: Specification 1

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	4.133	4.127	3.280	-1.782	1.207	3.548
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.127	4.090	4.087	3.178	-0.550	2.252	4.605
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.912	4.423	4.418	3.514	-0.187	2.836	5.879
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	4.066	4.061	3.224	-1.765	1.215	3.515
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.187	4.033	4.032	3.169	-0.283	2.319	4.734
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.546	4.138	4.169	3.386	0.696	3.479	6.511
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	6.611	6.603	5.230	-3.671	1.306	5.516
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.834	6.598	6.592	5.138	-2.264	1.899	5.939
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.112	6.680	6.672	5.372	-1.442	3.486	7.381
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.058	0.058	0.047	0.957	0.999	1.036
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.005	0.060	0.060	0.048	1.966	2.006	2.043
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.014	0.060	0.061	0.049	2.975	3.017	3.054

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X of which the boundary densities are zero, where $Y = X_1 + 2X_2 + 3X_3 + \varepsilon$ for $\varepsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\bar{\hat{\delta}}_1|$.

[Table 2] Finite-sample Behavior of Estimators for Homoskedastic Binary Choice Model: Specification 1

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	8.587	8.576	6.529	-4.125	0.762	6.063
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.268	8.415	8.409	6.483	-3.198	1.832	7.634
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.316	8.433	8.429	6.432	-1.876	2.998	8.368
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	9.619	9.607	7.555	-4.936	0.839	7.268
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.242	9.591	9.582	7.625	-4.396	1.839	8.316
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.336	9.012	9.007	7.126	-2.593	3.513	8.983
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	13.351	13.334	10.515	-7.090	0.914	9.780
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	-0.078	13.594	13.735	10.959	-9.704	0.035	8.763
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.539	12.682	12.678	10.069	-5.018	4.119	11.481
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.183	0.183	0.146	0.880	1.010	1.124
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.024	0.171	0.172	0.139	1.908	2.032	2.137
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.000	0.184	0.184	0.146	2.886	3.000	3.133
PROBIT	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.205	0.204	0.158	0.854	0.977	1.117
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.977	0.314	0.315	0.243	1.760	1.945	2.136
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.959	0.452	0.453	0.351	2.646	2.885	3.211

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X of which the boundary densities are zero, where $Y = 1[X_1 + 2X_2 + 3X_3 + \varepsilon > 0]$ for $\varepsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\bar{\hat{\delta}}_1|$.

[Table 3] Finite-sample Behavior of Estimators for Heteroskedastic Linear Model: Specification 1

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	2.367	2.364	1.778	-0.344	0.980	2.390
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.874	2.423	2.423	1.822	0.395	1.926	3.284
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.937	2.557	2.554	1.954	1.236	2.785	4.343
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	2.686	2.683	2.033	-0.525	1.062	2.524
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.801	2.615	2.619	1.971	0.323	1.848	3.520
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.894	2.812	2.811	2.19	1.136	2.651	4.783
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	3.132	3.128	2.461	-1.011	0.975	3.085
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.750	3.055	3.061	2.348	-0.126	1.759	3.701
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.767	3.278	3.282	2.539	0.587	2.731	4.715

OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.170	0.170	0.128	0.898	1.000	1.088
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.973	0.076	0.081	0.064	1.925	1.969	2.019
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.966	0.098	0.103	0.080	2.904	2.964	3.027

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X of which the boundary densities are zero, where $Y = X_1 + 2X_2 + 3X_3 + \varepsilon$ for $\varepsilon = \sigma \cdot \upsilon$, $\sigma = \sigma(X)$, and $\upsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\bar{\hat{\delta}}_1|$.

[Table 4] Finite-sample Behavior of Estimators for Heteroskedastic Binary Choice Model: Specification 1

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	10.950	10.936	7.907	-3.784	0.701	7.635
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.271	14.502	14.487	10.647	-5.484	0.085	9.538
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.409	18.967	18.948	14.161	-8.184	0.581	12.417
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	9.160	9.149	6.819	-3.810	0.935	6.832
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.219	12.986	12.972	9.866	-5.494	0.633	9.663
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.36	15.266	15.251	11.894	-6.637	1.701	12.521
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	18.649	18.626	14.142	-9.446	-0.191	12.127
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.464	19.667	19.648	14.917	-9.279	1.111	14.788
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	7.244	18.627	19.082	14.808	-5.327	5.339	17.821
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.208	0.208	0.164	0.865	1.008	1.133
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.192	0.181	0.264	0.221	2.077	2.196	2.312
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.290	0.192	0.347	0.298	3.158	3.286	3.422
PROBIT	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	4.803	4.797	0.948	0.244	0.486	0.796
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.078	9.519	9.508	1.758	0.663	1.116	1.634
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.130	14.145	14.128	2.608	1.023	1.681	2.500

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X of which the boundary densities are zero, where $Y = 1[X_1 + 2X_2 + 3X_3 + \varepsilon > 0]$ for $\varepsilon = \sigma \cdot \upsilon$, $\sigma = \sigma(X)$, and $\upsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\bar{\hat{\delta}}_1|$.

III.3.3. Specification 2: Boundary Densities Bounded away from Zero

Tables 5 through 8 report the simulation results of ADEs and other parametric estimates for the second specification of regressors where the density of X is bounded away from zero everywhere in the support. Note that SWADE is a consistent estimator for β up to scale. PSSE and HSE are not asymptotically consistent under this setup because the coefficient vector cannot be identified unless

$f(X)=0$ at $X \in d\Omega$ where $f(\cdot)$ is the density of X and $d\Omega$ denotes its boundary of the support.[13]

As reported in Tables 5 through 8, almost all estimates are better-behaved in the second specification of regressors than in the first specification, except for the OLS estimates for binary choice models. For example, the absolute mean deviation of SWADE for β_2 for homoskedastic linear models in Table 5 is merely 0.057 (=2-1.943); its corresponding value reported in Table 1 is 0.127 (=2.127-2). Furthermore, the former has smaller values of SD, RMSE, and MAE than the latter. This difference is probably because the density of X is strictly positive everywhere in the support and, so, because the erratic behavior potentially caused by very small boundary densities is avoidable in this case. In terms of mean deviations from the prespecified parameter value (1,2,3) and all variability measures, the OLS estimates are best for linear models and the Probit estimates for binary choice models, just like in the first specification of X. All of the ADEs, of course, including HSE are also well-behaved without an exception in all combinations of models and errors. PSSE seems to outweigh SWADE and HSE among ADEs; its mean deviations are very similar to those of SWADE and HSE. However, the variability measures for PSSE are surely smaller than those of SWADE as well as HSE in all models and errors.

Note that PSSE and HSE are not asymptotically consistent for β in the second regressor specification unlike SWADE. However, it is remarkably surprising that the finite sample biases of PSSE and HSE are almost as small as those of SWADE. This is often probable. As reported in Table 2, the OLS estimates are sufficiently well-behaved with very small deviations from the prespecified parameter values under the first specification of zero boundary densities of regressors; of course, the OLS estimates are not consistent for homoskedastic binary choice models. The biases of PSSE and HSE from boundary terms in integration by parts do not asymptotically vanish unless $f(X)=0$ at $X \in d\Omega$. However, the biases could be sufficiently small in some finite samples and sometimes comparable to those of consistent estimates such as SWADE; SWADE may encounter relatively large biases in finite samples, although the biases tend to vanish as the sample size increases to infinity.

The simulation results for eight combinations of models, errors, and regressor specifications suggest that SWADE is sufficiently well-behaved with more relaxed and, thus, relaxed distributional specification of regressors in finite sample among ADEs. For the first four combinations in the first specification of zero boundary densities of regressors, SWADE outperforms HSE without density trimming in finite sample. The finite sample behavior of SWADE is slightly and weakly inferior

[13] This is because the boundary terms in the integration by parts formula for the derivatives of regression function do not vanish unless the zero boundary density condition is satisfied. Refer to Equation (2.1) and Assumption 2 of PSS (1989) and Equation (3.1) and the second assumption of Assumption A.1 of HS (1989).

in some models and almost equivalent to PSSE in some other models; however, the difference is negligibly small in most cases. For all models and errors under the second specification of regressors, all of ADEs compared in this paper have shown qualitatively similar performance. In this sense, all of them are more or less qualitatively equivalent in terms of absolute mean deviations from specified parameter values and various variability measures (SD, RMSE, MAE and quartile range). Although the finite sample behaviors of all ADEs are mutually similar, PSSE and HSE are biased and asymptotically inconsistent under this setup. SWADE is a consistent estimator. In this respect, SWADE is theoretically superior to PSSE and HSE.

With these results for both specifications of regressors, SWADE is a sufficiently well-behaved consistent estimator for coefficient vector of interest under the single index models in finite sample. In particular, SWADE allows for wider applications of models and errors and, more importantly, more flexible distributional characteristics of regressors than other ADEs which require some restrictive conditions such as zero boundary densities in the support. In this sense, SWADE seems to behave appropriately in finite sample and relatively easier to apply without a concern of regressor specification.

[Table 5] Finite-sample Behavior of Estimators for Homoskedastic Linear Model: Specification 2

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.444	0.443	0.352	0.692	1.001	1.294
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.943	0.502	0.505	0.408	1.602	1.914	2.260
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.962	0.559	0.560	0.445	2.591	2.914	3.336
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.351	0.351	0.283	0.749	0.999	1.241
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.944	0.397	0.401	0.324	1.665	1.936	2.201
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.971	0.403	0.404	0.325	2.701	2.976	3.253
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.498	0.498	0.392	0.674	1.015	1.318
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.013	0.520	0.519	0.425	1.646	2.019	2.390
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.038	0.531	0.532	0.423	2.702	3.068	3.393
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.056	0.056	0.045	0.960	1.001	1.039
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.011	0.056	0.057	0.045	1.973	2.011	2.046
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.005	0.056	0.057	0.046	2.962	3.004	3.046

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X^* of which the boundary densities are zero, where $Y = X_1^* + 2X_2^* + 3X_3^* + \varepsilon$ for $\varepsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\bar{\hat{\delta}}_1|$.

[Table 6] Finite-sample Behavior of Estimators for Homoskedastic Binary Choice Model: Specification 2

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.990	0.989	0.770	0.342	0.972	1.585
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.032	1.144	1.143	0.866	1.269	1.930	2.641
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.855	1.159	1.166	0.924	2.056	2.821	3.542
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.821	0.820	0.651	0.471	0.999	1.583
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.996	0.907	0.905	0.710	1.434	1.902	2.566
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.828	0.940	0.954	0.766	2.193	2.756	3.445
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	1.028	1.026	0.824	0.317	1.105	1.673
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.029	1.104	1.103	0.898	1.284	2.001	2.769
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.823	1.089	1.102	0.893	2.089	2.741	3.552
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.187	0.186	0.147	0.872	1.008	1.116
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.219	0.204	0.299	0.249	2.077	2.224	2.369
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.551	0.162	0.575	0.551	3.443	3.556	3.660
PROBIT	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.215	0.215	0.167	0.845	0.988	1.116
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.023	0.333	0.333	0.251	1.785	1.980	2.191
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.026	0.473	0.473	0.354	2.679	2.963	3.262

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X^* of which the boundary densities are zero, where $Y = 1[X_1^* + 2X_2^* + 3X_3^* + \varepsilon > 0]$ for $\varepsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\overline{\hat{\delta}}_1|$.

[Table 7] Finite-sample Behavior of Estimators for Heteroskedastic Linear Model: Specification 2

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.432	0.431	0.344	0.704	0.998	1.300
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.971	0.544	0.544	0.429	1.581	1.962	2.309
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.087	0.607	0.613	0.484	2.656	3.042	3.464
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.366	0.366	0.293	0.739	1.002	1.238
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.965	0.413	0.414	0.329	1.680	1.963	2.246
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.076	0.431	0.437	0.339	2.791	3.052	3.352
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.520	0.520	0.404	0.663	0.983	1.335
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.023	0.548	0.548	0.436	1.654	2.043	2.382
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.108	0.543	0.553	0.441	2.752	3.121	3.468

OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.070	0.070	0.053	0.956	0.999	1.042
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	1.996	0.082	0.082	0.065	1.938	1.992	2.049
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	2.989	0.078	0.079	0.061	2.941	2.988	3.036

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X^* of which the boundary densities are zero, where $Y = X_1^* + 2X_2^* + 3X_3^* + \varepsilon$ for $\varepsilon = \sigma \cdot \upsilon$, $\sigma = \sigma(X)$, and $\upsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\overline{\hat{\delta}}_1|$.

[Table 8] Finite-sample Behavior of Estimators for Heteroskedastic Binary Choice Model: Specification 2

		TRUE	MEAN	SD	RMSE	MAE	LQ	Median	UQ
SWADE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.879	0.878	0.704	0.364	1.006	1.543
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.129	1.192	1.198	0.908	1.300	1.985	2.755
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.125	1.375	1.379	1.071	2.184	2.935	3.939
PSSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.726	0.725	0.576	0.528	1.018	1.483
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.126	0.910	0.917	0.727	1.477	2.127	2.712
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.155	1.024	1.035	0.792	2.463	3.065	3.806
HSE	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	1.033	1.031	0.814	0.379	1.038	1.700
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.134	1.070	1.077	0.857	1.351	2.147	2.829
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.127	1.163	1.168	0.935	2.265	3.137	3.908
OLS	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.185	0.185	0.147	0.876	1.009	1.124
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.257	0.210	0.332	0.279	2.112	2.254	2.387
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.758	0.173	0.778	0.758	3.635	3.762	3.866
PROBIT	$\hat{\beta}_1$ (i.e., $\hat{\delta}_1$)	1	1.000	0.484	0.484	0.368	0.671	0.938	1.228
	$\hat{\beta}_2$ (i.e., $\hat{\delta}_2$)	2	2.063	0.912	0.913	0.679	1.453	1.916	2.512
	$\hat{\beta}_3$ (i.e., $\hat{\delta}_3$)	3	3.176	1.325	1.335	0.987	2.291	3.012	3.795

Notes: The simulations are repeated 400 times independently with the sample size of 300 for X^* of which the boundary densities are zero, where $Y = 1[X_1^* + 2X_2^* + 3X_3^* + \varepsilon > 0]$ for $\varepsilon = \sigma \cdot \upsilon$, $\sigma = \sigma(X)$, and $\upsilon \sim N(0,1)$.

Coefficient estimates are divided by the absolute value of the sample mean of the first component, $|\overline{\hat{\delta}}_1|$.

IV. Conclusion

This paper proposes a consistent estimator, $\hat{\delta}$, of coefficient vector (β) up to scale under the single index model framework, using the average derivatives estimation methods as in PSS (1989) and HS (1989), among others. The basic differences from their estimators are the weighting functions: 1 for HSE, $f(X)$ for PSSE, and $f^2(X)$ for squared density weighted average derivatives estimator (SWADE) proposed here. SWADE of δ exhibits the standard properties of an i.i.d. sample average: $\sqrt{N}-$consistency with an asymptotic normal distribution since the estimator can be approximated as a linear function (i.e., the average) of i.i.d. derivatives of conditional mean functions.

An attractive feature of SWADE is that it is not necessary to assume that the regressors have zero density on the boundary in the support, which is required for PSSE or HSE. PSSE and HSE are often powerful and easy to apply in empirical studies, but sometimes restrictive in application for some cases particularly where the density of regressors is everywhere bounded away from zero in the support. In this case, PSSE and HSE which rely on zero boundary density conditions are not applicable. In addition, HSE requires density trimming to ensure $\sqrt{N}-$ consistency and an asymptotic normal distribution by avoiding erratic behavior of the estimates near the boundary in the support of regressors. Fortunately, however, SWADE introduces square density weighting and, so, requires no such restrictions. In this sense, SWADE is more widely applicable with fewer restrictions.

SWADE proposed in this study is asymptotically equivalent in precision to other ADEs, that is, PSSE and HSE for example. Monte Carlo simulations indicate that, for all model/error specifications between linear and binary choice models and between homoskedastic and heteroskedastic errors, parametric regressions certainly outperform nonparameteric ADEs including SWADE in finite sample. Among the average derivatives estimators, the finite sample behavior of SWADE is slightly outweighed by that of PSSE in most cases except the heteroskedastic binary choice model in which the sample mean of SWADE is slightly closer to the true parameter value than that of PSSE. HSE is not sufficiently well-behaved in finite sample unlike other ADEs.

These imply that SWADE allows more flexible applications with relaxed distributional characteristics than PSSE and HSE at the expense of slightly deteriorated behavior in finite sample.

Appendix A: Preliminary Lemmas

Preliminary lemmas for the theorems are presented here. Lemma 1 gives useful techniques to reduce asymptotic bias of $\hat{\delta}$. Lemmas 2 and 3 provide the projections of several U-statistics to derive the asymptotic distribution of $\hat{\delta}$. Lemma 4 shows the uniform convergence rates of several nonparametric estimates.

Lemma 1. If $u = O(\frac{1}{h})$ and if $H(\cdot)$ is $(\lambda+1)-$times continuously differentiable with bounded derivatives,
(a) $\int K(u)H(X-uh)du = H(X)+O(h^{\lambda})$,
(b) $h^{-1}\int \nabla K(u)H(X\pm uh)du = \mp\nabla H(X)+O(h^{\lambda})$.

Proof:
(a) For $n \in N^1$ and $p<\infty$, define $R_{np} \equiv \{(r_1,\cdots,r_p) \mid \sum_{m=1}^{p} r_m = n\}$ and $\Pi_n(\cdot) \equiv \frac{\partial^p H(\cdot)}{\partial u_1^{r_1}\cdots\partial u_p^{r_p}}$. By Taylor expansion of $H(\cdot)$ around $h=0$,

$$\begin{aligned}\int K(u)H(X-uh)du = {} & H(X)\int K(u)du \\ & +\sum_{m=1}^{\lambda-1}(-1)^m\left(\frac{h^m}{m!}\right)\Pi_m(X)\int\sum_{R_{(m+1)p}} u_1^{r_1}\cdots u_p^{r_p}K(u)du \\ & +(-1)^{\lambda}\left(\frac{h^{\lambda}}{\lambda!}\right)\int\sum_{R_{\lambda p}} u_1^{r_1}\cdots u_p^{r_p}\Pi_{\lambda}(X-uh^*)K(u)du\end{aligned}$$

for some $h^* \in [0,h]$. The second term is zero and the last term is $O(h^{\lambda})$ by Assumption 4. So, the above is $H(X)+O(h^{\lambda})$, since $\int K(u)du = 1$.

(b) $h^{-1}K(u) \to 0$ as $h \to 0$ since $u = (h^{-1})$ and $K(u)u \to 0$ as $u \to \pm\infty$ by Assumption 4(d). Thus, by integrating by parts, the result follows as

$$\begin{aligned}h^{-1}\int \nabla K(u)H(X\pm uh)du &= h^{-1}K(u)H(X\pm uh)\Big|_{u=-\infty}^{u=+\infty} \\ &\quad -h^{-1}\int K(u)\nabla H(X\pm uh)(\pm h)du \\ &= 0 \mp \int K(u)\nabla H(X\pm uh)du \\ &= \mp\nabla H(X)+O(h^{\lambda}). \qquad \leftarrow \text{ by (a)}\end{aligned}$$

Lemma 2. For $\{q_i\}$ an i.i.d. sample, consider a modified m-th order U-statistic of the form

$$U_N = \begin{pmatrix} N \\ m \end{pmatrix}^{-1} \sum \hat{s}(q_{i_1}, \cdots, q_{i_m})$$

where the sum is taken over the $\binom{N}{m}$ combinations of m distinct elements $(i_1, \cdots, i_m)$ from the set $(1, \cdots, N)$ for the "kernel" $\hat{s}(\cdot)$ symmetric in its m arguments. Define the "projection" as

$$\hat{U}_N = \hat{\theta} + \frac{m}{N} \sum_{i=1}^{N} (\hat{r}(q_i) - \hat{\theta})$$

where $\hat{r}(q_i) \equiv E[\hat{s}(\cdot) \,|\, q_i]$ and $\hat{\theta} \equiv E[\hat{r}(q_i)]$. Then,

$$\text{if } E\left\|\hat{s}(\cdot)^2\right\| = o(N), \qquad \text{(a) } U_N = \hat{U}_N + o_p(N^{-1/2}),$$

$$\text{if } E\left\|\hat{s}(\cdot)^2\right\| = o\left(\frac{N}{h^{p+2}}\right), \quad \text{(b) } U_N = \hat{\theta} + o_p(1).$$

Proof: The result directly follows from Lemma 2.1 of Lee (1988) and Lemma A.3 of Ahn (1995).

The next lemma shows projections of U-statistics related to $\hat{\delta}$. Recall $Z_i = \delta(X_i) - \varepsilon_i f(X_i) \nabla f(X_i)$.

Lemma 3. Suppose that the estimator for the slope parameter of interest, that is, β (or, equivalently δ in regression models specified in Equation (4)) is proposed as $\hat{\delta}$ defined in Equation (5). Then,

$$\hat{\delta} = \frac{3}{N} \sum_{i=1}^{N} Z_i - 2\delta + o_p(N^{-1/2}).$$

Proof:

$\hat{\delta}$ can be written as $\hat{\delta} = \binom{N}{m}^{-1} \sum_{1 \le i < j < k \le N} \hat{s}(q_i, q_j, q_k) + o(N^{-1/2})$ for $m = 3$; $\hat{s}(q_i, q_j, q_k)$ is symmetric in its arguments. Observe that $E\|\hat{s}\|^2 = E\|\hat{s}(q_i, q_j, q_k)\|^2$ $= O(h^{-(4p+2)}) = o(N)$ since $Nh^{4p+2} \to \infty$ as $N \to \infty$. Thus, by Lemma 2(a) for $\hat{\theta} = E(\hat{s})$ and $\hat{r}(q_i) = E(\hat{s} \,|\, q_i)$,

$$\hat{\delta} = \frac{3}{N} \sum_{i=1}^{N} E(\hat{s} \,|\, q_i) - 2E(\hat{s}) + o_p(N^{-1/2}).$$

Hence, it suffices to show that $E(\hat{s} \,|\, q_i) = E(\hat{s}_1 \,|\, q_i) + E(\hat{s}_2 \,|\, q_i) + E(\hat{s}_3 \,|\, q_i) =$

$Z_i + o(N^{-1/2})$ and that $E(\hat{s}) = \delta + o(N^{-1/2})$ where

$$E(\hat{s}_1 \,|\, q_i) = \frac{1}{3} f^2(X_i)\nabla G(X_i) + o(N^{-1/2})$$

$$E(\hat{s}_2 \,|\, q_i) = E(\hat{s}_3 \,|\, q_i) = \frac{1}{3} f^2(X_i)\nabla G(X_i) - \frac{1}{2}\varepsilon_i f(X_i)\nabla f(X_i) + o(N^{-1/2})$$

so,

$$E(\hat{s} \,|\, q_i) = f^2(X_i)\nabla G(X_i) - \varepsilon_i f(X_i)\nabla f(X_i) + o(N^{-1/2}) = Z_i + o(N^{-1/2})$$
$$E(\hat{s}) = E(Z_i) + o(N^{-1/2}) = \delta + o(N^{-1/2}) .$$

Proofs for only $E(\hat{s}_1 \,|\, q_i)$ and $E(\hat{s}_2 \,|\, q_i)$ are presented below. The case for $E(\hat{s}_3 \,|\, q_i)$ can be shown similarly. For $\frac{X_i - X_j}{h} = u$ and $\frac{X_i - X_k}{h} = v$,

$$\begin{aligned}
E(\hat{s}_1 \,|\, q_i) &= \frac{1}{6h}\int [\nabla K(u)K(v) - K(u)\nabla K(v)] G(X_i - uh) f(X_i - uh) f(X_i - vh) du dv \\
&\quad - \frac{1}{6h}\int [\nabla K(u)K(v) - K(u)\nabla K(v)] f(X_i - uh) G(X_i - vh) f(X_i - vh) du dv \\
&= \frac{1}{6}\{ f^2(X_i)\nabla G(X_i) + G(X_i) f(X_i)\nabla f(X_i) - G(X_i) f(X_i)\nabla f(X_i) \\
&\quad - G(X_i) f(X_i)\nabla f(X_i) + G(X_i) f(X_i)\nabla f(X_i) + f^2(X_i)\nabla G(X_i) \} + O(h^{\lambda}) \\
&\qquad \leftarrow \text{ by Lemma 1(a), (b)} \\
&= \frac{1}{3} f^2(X_i)\nabla G(X_i) + o(N^{-1/2}) . \quad \leftarrow \ O(h^{\lambda}) = o(N^{-1/2}) \text{ by Assumption 5.}
\end{aligned}$$

For $\frac{X_j - X_i}{h} = u$ and $\frac{X_j - X_k}{h} = v$,

$$\begin{aligned}
E(\hat{s}_2 \,|\, q_i) &= \frac{1}{6h}\int (G[X_i + (u - v)h] - Y_i)\nabla K(v) K(u) f(X_i + uh) f[X_i + (u - v)h] du dv \\
&\quad - \frac{1}{6h}\int (G[X_i + (u - v)h] - Y_i) K(v)\nabla K(u) f(X_i + uh) f[X_i + (u - v)h] du dv \\
&= \frac{1}{6}\{ f^2(X_i)\nabla G(X_i) + G(X_i) f(X_i)\nabla f(X_i) - Y_i f(X_i)\nabla f(X_i) \} \\
&\quad - \frac{1}{6}\{ -f^2(X_i)\nabla G(X_i) - 2G(X_i) f(X_i)\nabla f(X_i) - Y_i[-2 f(X_i)\nabla f(X_i)] \} + O(h^{\lambda}) \\
&\qquad \leftarrow \text{ by Lemma 1(a) and (b)} \\
&= \frac{1}{3} f^2(X_i)\nabla G(X_i) - \frac{1}{2}\varepsilon_i f(X_i)\nabla f(X_i) + o(N^{-1/2}) . \quad \leftarrow \ Y_i = G(X_i) + \varepsilon_i .
\end{aligned}$$

Define $T(X) \equiv E(Y \,|\, X) f(X) = G(X) f(X) = \int Y f_{XY}(X,Y) dY$ and $\varphi_N(X) \equiv \frac{1}{Nh^p}\sum_{i=1}^{N} K(\frac{X - X_i}{h})\varepsilon_i^2$ where $f_{XY}(X,Y)$ is a joint density of (X,Y) . Note that

$\varphi_N(X)$ is similar to $\hat{\varphi}_N(X)$: the only difference between them is ε_i^2 instead of $\hat{\varepsilon}_i^2$. The following lemma shows uniform convergence rates of several estimators.

Lemma 4. For a compact $X \subset \ddot{X}$ in which $inf_{X\in\ddot{X}} f(X) > 0$,

(a) $\sup_{X\in\ddot{X}} |\hat{f}(X) - f(X)| = O((N^{1-e}h^{p})^{-1/2})$

(b) $\sup_{X\in\ddot{X}} |\hat{T}(X) - T(X)| = O((N^{1-e}h^{p})^{-1/2})$

(c) $\sup_{X\in\ddot{X}} |\hat{G}(X) - G(X)| = O((N^{1-e}h^{p})^{-1/2})$

(d) $\sup_{X\in\ddot{X}} |\varphi_N(X) - \varphi(X)| = O((N^{1-e}h^{p})^{-1/2})$

(e) $\sup_{X\in\ddot{X}} |\nabla\hat{f}(X) - \nabla f(X)| = O((N^{1-e}h^{p+2})^{-1/2})$

(f) $\sup_{X\in\ddot{X}} |\nabla\hat{T}(X) - \nabla T(X)| = O((N^{1-e}h^{p+2})^{-1/2})$

(g) $\sup_{X\in\ddot{X}} |\nabla\hat{G}(X) - \nabla G(X)| = O((N^{1-e}h^{p+2})^{-1/2})$.

Proof: (a), (b), and (c) follow from Theorem 3 of Collomb and Härdle (1986) or Lemma 1 of AM. Substituting ε_i^2 for Y_i into (b) implies (d) [cf. Corollary of AM]. (e) and (f) follow from (A.1b) and (A.1d) in the proof of Theorem 1 of Stoker (1991) (see pp. 109-110). (g) follows from the arguments below (notations abbreviated for simplicity):

$$|\nabla\hat{G} - \nabla G| = \left|\left(\frac{\nabla\hat{T}}{\hat{f}} - \frac{\hat{T}\nabla\hat{f}}{\hat{f}^2}\right) - \left(\frac{\nabla T}{f} - \frac{T\nabla f}{f^2}\right)\right| \le \left|\frac{\nabla\hat{T}}{\hat{f}} - \frac{\nabla T}{f}\right| + \left|\frac{\hat{T}\nabla\hat{f}}{\hat{f}^2} - \frac{T\nabla f}{f^2}\right|.$$

Hence,

$$\begin{aligned}\left|\frac{\nabla\hat{T}}{\hat{f}} - \frac{\nabla T}{f}\right| &\le \left|\frac{1}{\hat{f}\times f}\right| \times |\nabla\hat{T}f - \nabla T\hat{f}| \\ &\le \left|\frac{1}{\hat{f}\times f}\right| \times (|f| \times |\nabla\hat{T} - \nabla T| + |\nabla T| \times |f - \hat{f}|) \\ &= O((N^{1-e}h^{p+2})^{-1/2}), \qquad \leftarrow \text{ by (a) and (f)}\end{aligned}$$

and

$$\begin{aligned}\left|\frac{\hat{T}\nabla\hat{f}}{\hat{f}^2} - \frac{T\nabla f}{f^2}\right| &\le \left|\frac{1}{\hat{f}\times f}\right|^2 \times |\hat{T}\nabla\hat{f}f^2 - T\nabla f\hat{f}^2| \\ &\le \left|\frac{1}{\hat{f}\times f}\right|^2 \times (|\nabla\hat{f}f^2| \times |\hat{T} - T| + |Tf^2| \times |\nabla\hat{f} - \nabla f|\end{aligned}$$

$$\begin{aligned} &+|T\nabla f|\times|f^2-\hat{f}^2|) \\ &=O((N^{1-e}h^{p+2})^{-1/2}), \qquad \leftarrow \text{ by (a), (b) and (e)} \end{aligned}$$

since all derivatives are bounded and $|f^2-\hat{f}^2|=|f-\hat{f}|\times|f+\hat{f}|$.

Appendix B: Proofs of Theorems

Proof of Theorem 1:

The first term $\frac{3}{N}\sum_{i=1}^{N} Z_i$ of $\hat{\delta}$ in Lemma 3 is an i.i.d. average and $\lim_{N\to\infty} E[\sqrt{N}(\hat{\delta}-\delta)]=0$ and $\lim_{N\to\infty} Var[\sqrt{N}(\hat{\delta}-\delta)]=9Var(Z)=9\Sigma$. Therefore, the results follow by the Lindeberg-Levy Central Limit Theorem.

Proof of Theorem 2:

(i) By Lemma 4(a) and (e), $\hat{f}(X)$ and $\nabla\hat{f}(X)$ are uniformly consistent to $f(X)$ and $\nabla f(X)$, respectively.

(ii) $sup|\hat{\varepsilon}-\varepsilon|=sup|\hat{G}-G(X)|=O((N^{1-e}h^p)^{-1/2})=o(1)$ by Lemma 4(c).

(iii) By the triangular inequality,

$$sup|\hat{\varphi}(X)-\varphi(X)|\le sup|\hat{\varphi}(X)-\varphi_N(X)|+sup|\varphi_N(X)-\varphi(X)|$$

Hence, $sup|\hat{\varphi}(X)-\varphi(X)|=o(1)$ since $sup|\varphi_N(X)-\varphi(X)|=O((N^{1-e}h^p)^{-1/2})$ $=o(1)$ by Lemma 4(d) and

$$\begin{aligned} sup|\hat{\varphi}(X)-\varphi_N(X)| &\le \left|\frac{1}{Nh^p}\sum_{i=1}^{N} K\left(\frac{X-X_i}{h}\right)(\hat{\varepsilon}_i^2-\varepsilon_i^2)\right| \\ &\le h^{-p} sup|K(\cdot)|\times sup|\hat{\varepsilon}_i-\varepsilon_i|\times sup|\hat{\varepsilon}_i+\varepsilon_i| \\ &= h^{-p}\times O(1)\times O((N^{1-e}h^p)^{-1/2})\times O(1) \\ &= O((N^{1-e}h^{3p})^{-1/2})\times o(1). \qquad \leftarrow \text{ by (ii)} \end{aligned}$$

(iv) $\hat{\delta}(X)=\hat{f}^2(X)\nabla\hat{G}(X)$ is uniformly consistent for $\delta(X)=f^2(X)\nabla G(X)$, since $\hat{f}(X)$ and $\nabla\hat{G}(X)$ are uniformly consistent for $f(X)$ and $\nabla G(X)$ by Lemma 4(a) and (g), respectively. To verify this, consider

$$\begin{aligned} |\hat{\delta}(X)-\delta(X)| &= |\hat{f}^2(X)\nabla\hat{G}(X)-f^2(X)\nabla G(X)| \\ &\le |\hat{f}^2(X)\nabla\hat{G}(X)-\hat{f}^2(X)\nabla G(X)|+|\hat{f}^2(X)\nabla G(X)-f^2(X)\nabla G(X)| \\ &\le |\hat{f}^2(X)|\times|\nabla\hat{G}(X)-\nabla G(X)|+|\nabla G(X)|\times|\hat{f}(X)-f(X)| \\ &\times|\hat{f}(X)+f(X)|=o(1). \end{aligned}$$

By (i) through (iv), $\hat{\Sigma}$ is consistent for Σ:

$$\frac{1}{N}\Sigma_{i=1}^{N}[\hat{\varphi}(X_i)\hat{f}(X_i)\nabla\hat{f}(X_i)\nabla\hat{f}(X_i)'] \rightarrow_p \frac{1}{N}\Sigma_{i=1}^{N}[\varphi(X_i)f(X_i)\nabla f(X_i)\nabla f(X_i)']$$
$$\rightarrow_{SLLN} E[\varphi(X_i)f(X_i)\nabla f(X_i)\nabla f(X_i)']$$

and

$$\frac{1}{N}\Sigma_{i=1}^{N}[\hat{\delta}(X_i)\hat{\delta}(X_i)'] - \delta\delta' \rightarrow_p \frac{1}{N}\Sigma_{i=1}^{N}[\delta(X_i)\delta(X_i)'] - \delta\delta'$$
$$\rightarrow_{SLLN} E[\delta(X_i)\delta(X_i)'] - \delta\delta',$$

so, $\hat{\Sigma} \rightarrow \Sigma$.

Appendix C: Additional Monte Carlo Simulations

Here, finite sample behaviors of ADEs are investigated for a higher dimension of regressors, $p = 3$, under a linear model with a homoskedastic error having smaller variance, compared with those of regressors. Consider a case in which random numbers for (X, ε) of size five hundred are independently drawn from normal distributions, such that $x_j \sim N(0,1)$ for $j = 1,2,3$ and $\varepsilon \sim N(0, 0.02^2)$ where $X = (x_1, x_2, x_3)$. A slightly larger sample size is chosen here than the cases discussed in Section III.3. Let the following equation be the relation between Y and other variates:

$$Y = x_1 + 2x_2 + 3x_3 + \varepsilon.$$

The same higher order Gaussian kernel is used as in Section III.3 except for some parameter and bandwidth values: set $\lambda = 1$ for $p = 3$ by Assumption 2 and $h_N = cN^{-1/15}$ for some constant c. Without loss of generality, let $c = 1$. Simulations are independently repeated a thousand times to construct empirical distributions of the aforementioned PSSE, HSE, and SWADE.

Just like those in Section III.3, the average derivatives estimates are rescaled by dividing them by the average value of the first component of the parameter estimates for simplicity of comparison. This is shown in the third block in Table A.1. The first and the second blocks in Table A.1 present the original estimates as well as the estimates rescaled by their own first component estimates, respectively.

As shown in Table A.1, all of the three methods of average derivatives estimators yield similar results: the average values of parameter estimates are very close to the true parameter value, (1,2,3), and the biases are negligible, regardless of estimators. However, the standard deviations of the estimates are slightly different over the three estimators. HSE has the largest standard deviations (0.415,0.447,0.512), SWADE has the smallest (0.309,0.377,0.484), and those of PSSE are in-between (0.345,0.397,0.472). HSE is outweighed by SWADE and PSSE in terms of standard deviations. The differences of standard deviations between SWADE and PSSE are seemingly significant in the statistical sense for the first and the second elements: those of SWADE are smaller than PSSE. The inequality direction is reversed for the standard deviation of the third elements between the two; however, the difference is not as large, 0.484 (SWADE) versus 0.482 (PSSE). From the results of simulations, SWADE is mostly preferable in finite sample in the above setup, to the extent that the standard deviations of the estimates are concerned, although its precision is asymptotically equivalent to PSSE and HSE.

In what follows, the simulation results discussed above are briefly compared with those in Section III.3. The standard deviation of error term is assumed to be much smaller in this section (0.02) than those (1) of homoskedastic linear and binary choice models in Section III.3. Other differences are the dimension of regressors and sample size.

The most noticeable difference in simulation results is the signs of HSE estimates; the sample means of HSE estimates are all negative in Section III.3, but all positive here. The drastic change in the signs of HSE estimates may be caused by the significantly different levels of error variance. A larger variance may inflate the biases stemming from larger variations from error terms.

Another noticeable change can be found in the behavior of SWADE. It behaves more nicely under the linear model with smaller error variance, and less nicely under the models with larger error variance. SWADE is slightly outweighed by PSSE in most cases under the models with larger error variance. However, they are almost equivalent in terms of finite sample behavior under the model with smaller error variance. From this, we can infer that the precision of SWADE in finite sample may be blurred by the increased variations in error terms. Variations in error terms are often sensitive to model specifications or choice of regressors. In this sense, despite its advantages over distributional constraints, SWADE could be relatively more sensitive or vulnerable, for example, to omitted variables than PSSE.

However, the above discussion is indeterminate and needs more thorough investigation to confirm. Because the results can be case-dependent, this question is left for a further study.

[Table A.1] Monte Carlo Simulation Results: Comparison of Average Derivatives Estimates

Original Estimates ($\hat{\delta}$)						
	Averages			Std. Deviations		
Estimates	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3
SWADE	0.0115	0.02298	0.03439	0.00355	0.00433	0.00557
PSSE	0.0223	0.04483	0.06704	0.00769	0.00886	0.01052
HSE	-0.04011	-0.08066	-0.11962	0.01665	0.01794	0.02054
Estimates Normalized by Each of the First Elements ($\hat{\delta} / \hat{\delta}_1$)						
	Averages			Std. Deviations		
Estimates	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3
SWADE	1	2.19086	3.26875	0	0.83409	1.16137
PSSE	1	2.18964	3.27996	0	2.83231	3.99792
HSE	1	3.88982	6.45238	0	28.06116	59.81803
Estimates Normalized by the Average of the First Element Estimates ($\hat{\delta} / E(\hat{\delta}_1)$)						
	Averages			Std. Deviations		
Estimates	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3
SWADE	1	1.99794	2.99025	0.30885	0.37655	0.48444
PSSE	1	2.01006	3.00613	0.34502	0.39749	0.47189
HSE	1	2.01089	2.98235	0.41514	0.44715	0.5122

Note: The simulations are repeated 1,000 times independently with the sample size of 500.
The true model is $Y = x_1 + 2x_2 + 3x_3 + \varepsilon$. That is, $(\delta_1, \delta_2, \delta_3) = (1,2,3)$.

References

Ahn, H. (1995), "Nonparametric Two-Stage Estimation of Conditional Choice Probabilities in a Binary Choice Model under Uncertainty," *Journal of Econometrics*, Vol. 67, 337-378.

Aït-Sahalia, Y., P. J. Bickel, and T. M. Stoker (2001), "Goodness-of-fit Tests for Kernel Regression with an Application to Option Implied Volatilities," *Journal of Econometrics*, Vol. 105, 363-412.

Bierens, H. J. (1987), "Kernel Estimators of Regression Functions," in Truman F. Bewley, eds., Advances in Econometrics: Fifth World Congress, Cambridge University Press, 99-144.

Chen, L., S. Lee and M. J. Sung (2014), "Maximum Score Estimation with Nonparametrically Generated Regressors," *The Econometrics Journal*, Vol. 17, Issue 3, 271-300.

Collomb, G. and W. Härdle (1986), "Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis and Prediction: Kernel Regression Estimation from Dependent Observations," *Stochastic Process and Their Applications*, Vol. 23, 77-89.

Fan, Y. and Q. Li (1996), "Consistent Model Specification Test-Omitted Variables and Semiparametric Functional Forms," *Econometrica*, Vol. 64, 865-890.

Han, A. K. (1987a), "A Nonparametric Analysis of Transformations," *Journal of Econometrics*, Vol. 35, 191-209.

Han, A. K. (1987b), "Nonparametric Analysis of a Generalized Regression Model, The Maximum Rank Correlation Estimator," *Journal of Econometrics*, Vol. 35, 303-316.

Härdle, W. and T. M. Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, Vol. 84, 986-995.

Horowitz, J. (2009), Semiparametric and Nonparametric Methods in Econometrics, Springer Series in Statistics, Springer-Dordrecht, Heidelberg, London, and New York.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, Vol. 58, 71-120.

Ichimura, H. and P. E. Todd (2007), "Implementing Nonparametric and Semiparametric Estimators," *Handbook of Econometrics*, 6(Part B), 5369-468.

Lee, B. (1988), " Nonparametric Tests Using a Kernel Estimation Method," Doctoral Dissertation, Department of Economics, University of Wisconsin-Madison.

Park, J. (1990), "Nonparametric Estimation of the Generalized Regression Model: Generalized Accelerated Failure Time (GAFT) Model and GAFT-type Competing Risk Model," Doctoral Dissertation, Department of Economics, University of Wisconsin-Madison.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, Vol. 57, 1403-1430.

Stoker, T. M. (1991), "Equivalence of Direct and Indirect Estimators of Average Derivatives," in W. Barnett, J. Powell, and G. Tauchen, eds., Nonparametric and Semiparametric Methods in Econometrics and Statistics, City Cambridge: Cityplace

Cambridge University Press.

Stute, W. and L. Zhu (2005), “Nonparametric Checks for Single-index Models,” *The Annals of Statistics*, Vol. 33, No. 3, 1048-1083.

Xia, Y. (2006), “Asymptotic Distributions for Two Estimators of the Single-index Model,” *Econometric Theory*, Vol. 22, No. 6, 1112-1137.

Xia, Y, W. K. Li, H. Tong, and D. Zhang (2004), “A Goodness-of-fit Test for Single-index Models,” *Statistics Sinica*, Vol. 14, 1-39.