Semiparametric Models with Single-Index Nuisance Parameters

Kyungchul Song[1]

Department of Economics, University of British Columbia

August 3, 2012

Abstract

In many semiparametric models, the parameter of interest is identified through conditional expectations, where the conditioning variable involves a single-index that is estimated in the first step. Among the examples are sample selection models and propensity score matching estimators. When the first-step estimator follows cube-root asymptotics, no method of analyzing the asymptotic variance of the second step estimator exists in the literature. This paper provides nontrivial sufficient conditions under which the asymptotic variance is not affected by the first step single index estimator regardless of whether it is root-n or cube-root consistent. The finding opens a way to simple inference procedures in these models. Results from Monte Carlo simulations show that the procedures perform well in finite samples.

Keywords: Sample selection model; conditional median restrictions; matching estimators; maximum score estimation; cube-root asymptotics; generated regressors;

JEL Classifications: C12, C14, C51.

[1] Part of the result in this paper was circulated in a draft titled "Bootstrapping Semiparametric Models with Single-Index Nuisance Parameters." I thank Xiaohong Chen, Stefan Hoderlein, Simon Lee, Frank Schorfheide and seminar participants at the Greater New York Econometrics Colloquium at Princeton University for valuable comments. I also thank Flavio Cunha for providing me with the data sets that I used for the empirical study, and an associate editor and two referees for comments that led to a substantial improvement of the paper. All errors are mine. Address correspondence to: Kyungchul Song, Department of Economics, University of British Columbia, 997-1873 East Mall, Vancouver, BC, V6T 1Z1, Canada.

1 Introduction

Many empirical studies use a number of covariates to deal with the problem of endogeneity. Using too many covariates in nonparametric estimation, however, tends to worsen the quality of the empirical results significantly. A promising approach in this situation is to introduce a single-index restriction so that one can retain flexible specification while avoiding the curse of dimensionality. The single-index restriction has long attracted attention in the literature.[2]

Most literatures deal with a single-index model as an isolated object, whereas empirical researchers often need to use the single-index specification in the context of estimating a larger model. An example is a structural model in labor economics that requires a prior estimation of components such as wage equations. When single-index components are nuisance parameters that are plugged into the second step estimation of a finite dimensional parameter of interest, the introduction of single-index restrictions does not improve the convergence rate of the estimated parameter of interest which already achieves the parametric rate of $\sqrt{n}$. Nevertheless, the use of a single-index restriction in such a situation still has its own merits. After its adoption, the model requires weaker assumptions on the nonparametric function and on the kernel function. This merit becomes prominent when the nonparametric function is defined on a space of a large dimension and stronger conditions on the nonparametric function and higher-order kernels are required. (See Hristache, Juditsky and Spokoiny (2001) for more details.)

This paper focuses on semiparametric models, where the parameter of interest is identified through a conditional expectation function and the conditioning variable involves a single-index with an unknown finite dimensional nuisance parameter. We assume that there is a consistent first step estimator of this nuisance parameter. In this situation, a natural procedure is a two step estimation, where one estimates the single-index first, and uses it

[2] For example, Klein and Spady (1993) and Ichimura (1993) proposed M-estimation approaches to estimate the single-index, and Stoker (1986) and Powell, Stock and Stoker (1989) proposed estimation based on average derivatives. See also Härdle, Hall, and Ichimura (1993), Härdle and Tsybakov (1993), Horowitz and Härdle (1996), Fan and Li (1996) and Hristache, Juditsky and Spokoiny (2001).

to estimate the parameter of interest in the second step. Among the examples are sample selection models and propensity score matching estimators. The examples will be discussed in detail later.

A distinctive feature of the framework of this paper is that the first step estimator of a single-index is allowed to be either $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent. The latter case of $\sqrt[3]{n}$-consistent single index estimators is particularly interesting, for the framework includes new models that have not been studied in the literature, such as the sample selection model with conditional median restrictions, or propensity score matching estimators with conditional median restrictions. These conditional median restrictions often lead to a substantial relaxation of the existing assumptions that have been used in the literature.[3]

Dealing with the case of a nuisance parameter that follows cube-root asymptotics of Kim and Pollard (1990) in two step estimation is challenging. In typical two step estimation, the asymptotic variance of the second step estimator involves an additional term due to the first step estimation of the single-index component (e.g. Newey and McFadden (1994).) Unless this term is shown to be negligible, one needs to compute this additional term by first finding the asymptotic linear representation of the first step estimator. However, in the case of a first step estimator that follows cube-root asymptotics, there does not exist such an asymptotic linear representation.

The main contribution of this paper is to provide a set of conditions under which the first step estimator, regardless of whether it is $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent, does not have an impact on the asymptotic variance of the second step estimator. This result is convenient, because under these conditions, one can simply compute the asymptotic variance as if one knows the true nuisance parameter in the single-index.

[3] For example, the semiparametric sample selection model in Newey, Powell and Walker (1990) assumes that the error term in the selection equation is independent of observed covariates. Also, parametric specifications of propensity scores in the literature of program evaluations (such as logit or probit specifications) assume that the error term in the program participation equation is independent of observed covariates. (See Heckman, Ichimura, Smith and Todd (1998) for example.) In these situations, the assumption of the conditional median restriction is a weaker assumption because it allows for stochastic dependence between the error term and the observed covariates.

The result of this paper is based on a recent finding by the author (Song (2012)) which offers generic conditions under which conditional expectation functionals are very smooth. This smoothness is translated in our situation into insensitivity of the parameter of interest at a local perturbation of the single-index nuisance parameter.

To illustrate the usefulness of the result, this paper applies it to new semiparametric models such as semiparametric sample selection models with conditional median restrictions, and single-index matching estimators with conditional median restrictions. This paper offers procedures to obtain estimators and asymptotic variance formulas for the estimators.

This paper presents and discusses results from Monte Carlo simulation studies. The main focus of these studies lies on whether the asymptotic negligibility of the first step estimator's impact remains in force in finite samples. For this, it is investigated whether the estimators and the confidence sets based on the proposed asymptotic covariance matrix formula performs reasonably well in finite samples. Simulation results demonstrate clearly that they do so.

The main result of this paper is closely related to the literature of so-called *generated regressors* in nonparametric or semiparametric models. For example, Newey, Powell, and Vella (1999) and Das, Newey, and Vella (2003) considered nonparametric estimation of simultaneous equation models. Li and Wooldridge (2002) analyzed partial linear models with generated regressors when the estimated parameters in the generated regressors are $\sqrt{n}$-consistent. Rilstone (1996) and Sperlich (2009) studied nonparametric function estimators that involve generated regressors. Recent contributions by Hahn and Ridder (2010) and Mammen, Rothe, and Schienle (2012) offer a general analysis of the issue with generated regressors in nonparametric models. None of these papers considered generated regressors with coefficient estimators that follow cube-root asymptotics.

The paper is organized as follows. The paper defines the scope, introduces examples, and explains the main idea of this paper in the next section. Then Section 3 presents the formal result of the asymptotic distribution theory, and discusses their implications for exemplar models. Section 4 discusses Monte Carlo simulation results, and Section 5 presents an empirical illustration based on a simple female labor supply model. Some technical proofs are found in the Appendix.

2 The Scope, Examples, and the Main Idea

2.1 The Scope of the Paper

Let us define the scope of the paper. Suppose that $W \equiv (W_1, \cdots, W_L)^\top \in \mathbf{R}^L$, S is a $d_S \times d_\varphi$ random matrix, and $X \in \mathbf{R}^d$ is a random vector, where all three random quantities W, S, and X, are assumed to be observable. We let $X = [X_1^\top, X_2^\top]^\top \in \mathbf{R}^{d_1+d_2}$, where X_1 is a continuous random vector and X_2 is a discrete random vector taking values from $\{x_1, \cdots, x_M\}$. Let $\Theta \subset \mathbf{R}^d$ be the space of a nuisance parameter θ_0 that is known to be identified. Denote $U_\theta \equiv F_\theta(X^\top\theta)$, where F_θ is the CDF of $X^\top\theta$. We assume that $X^\top\theta$ is a continuous random variable for all θ in a neighborhood of θ_0. Given an observed binary variable $D \in \{0,1\}$, we define

$$\mu_\theta(U_\theta) \equiv \mathbf{E}\left[W|U_\theta, D=1\right], \tag{1}$$

and when $\theta = \theta_0$, we simply write $\mu_0(U_0)$, where $U_0 \equiv F_{\theta_0}(X^\top\theta_0)$. The support of a random vector is defined to be the smallest closed set in which the random vector takes values with probability one. For $m = 1, \cdots, M$, let $\mathcal{S}_m$ be the support of $X1\{X_2 = x_m, D = 1\}$. and $\mathcal{S}_W$ be the support of W, and let $\varphi : \mathcal{S}_W \to \mathbf{R}^{d_\varphi}$ be a known map that is twice continuously differentiable with bounded derivatives on the interior of the support of $\mathbf{E}[W|X, D=1]$. Then we define a map $a : \Theta \to \mathbf{R}^{d_S}$ by

$$a(\theta) \equiv \mathbf{E}\left[S \cdot \varphi(\mu_\theta(U_\theta))|D=1\right], \ \theta \in \Theta. \tag{2}$$

The general formulation admits the case without conditioning on $D = 1$ in which case it suffices to put $D = 1$ everywhere.

This paper focuses on semiparametric models where the parameter of interest, denoted by β_0, is identified as follows:

$$\beta_0 = H(a(\theta_0), b_0), \tag{3}$$

where $H : \mathbf{R}^{d_S} \times \mathbf{R}^{d_b} \to \mathbf{R}^{d_\beta}$ is a map that is fully known, continuously differentiable in the first argument, and b_0 is a d_b dimensional parameter that does not depend on θ_0 and is consistently estimable. We will see examples of β_0 shortly.

Throughout this paper, we assume that there is an estimator $\hat{\theta}$ for θ_0 which is either $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent. A natural estimator of β_0 is obtained by

$$\hat{\beta} \equiv H(\hat{a}(\hat{\theta}), \hat{b}),$$

where $\hat{a}(\theta)$ is an estimator of $a(\theta)$ and $\hat{b}$ is a consistent estimator of b_0. The estimator $\hat{a}(\theta)$ can be obtained by using nonparametric estimation of conditional expectation $\mathbf{E}\left[W|U_\theta, D = 1\right]$. For future reference, we denote

$$\tilde{\beta} \equiv H(\hat{a}(\theta_0), \hat{b}),$$

an infeasible estimator using θ_0 in place of $\hat{\theta}$. When $\hat{\theta}$ is $\sqrt[3]{n}$-consistent, it is not clear whether $\sqrt{n}(\hat{\beta} - \beta)$ will be asymptotically normal. In fact, it is not even clear whether $\hat{\beta}$ will be $\sqrt{n}$-consistent.

The main contribution of this paper is to provide conditions under which, whenever $\hat{\theta} = \theta_0 + O_P(n^{-1/3})$ and

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V), \tag{4}$$

it follows that

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V).$$

This result is very convenient, because the computation of the asymptotic variance matrix

V in (4) can be done, following the standard procedure.

2.2 Examples

2.2.1 Example 1: Sample Selection Models with Conditional Median Restrictions

Let Y^* be an outcome variable which is related to $Z \in \mathbf{R}^{d_Z}$, a vector of covariates, as follows:

$$Y^* = Z^\top \beta_0 + v,$$

where v denotes an unobserved factor that affects the outcome. The econometrician observes Y^* only when a selection indicator $D \in \{0, 1\}$ assumes number one, so that as for observed outcome Y, we write

$$Y = Y^* \cdot D.$$

This paper specifies D as follows:

$$D = 1\{X^\top \theta_0 > \varepsilon\}, \tag{5}$$

where ε is an unobserved component, and $\theta_0 \in \mathbf{R}^d$ an unknown parameter.

Sample selection models and their inference procedures have been extensively studied in the literature. The early generation of these models impose parametric distributional assumptions on the unobserved components (Heckman (1974)). Gallant and Nychka (1987), Cosslett (1990) and many others (e.g. Newey, Powell, and Walker (1990), and Das, Newey and Vella (1999)) analyzed semiparametric or nonparametric models that do not require parametric distributional assumptions. A common feature for these various models is the following assumption:[4]

$$\varepsilon \text{ is independent of } X. \tag{6}$$

Condition (6) is mainly used to identify θ_0 through a single-index restriction $\mathbf{E}[D|X] = \mathbf{E}[D|X^\top\theta_0]$ or a parametric restriction $\mathbf{E}[D|X] = F(X^\top\theta_0)$ for some known CDF F. Define for θ in a neighborhood of θ_0,

$$\begin{aligned} S_{ZZ}(\theta) &\equiv \mathbf{E}\left[ZZ^\top|D=1\right] - \mathbf{E}\left[Z \cdot \mathbf{E}[Z^\top|U_\theta, D=1]|D=1\right] \text{ and} \\ S_{ZY}(\theta) &\equiv \mathbf{E}[ZY|D=1] - \mathbf{E}\left[Z \cdot \mathbf{E}[Y|U_\theta, D=1]|D=1\right]. \end{aligned}$$

We consider the following assumptions.

Assumption SS0 : (i) (ε, v) is conditionally independent of Z given $X^\top\theta_0$.
(ii) $Med(\varepsilon|X) = 0$, a.e., where $Med(\varepsilon|X)$ denotes the conditional median of ε given X.
(iii) The smallest eigenvalue of $S_{ZZ}(\theta)$ is bounded away from zero uniformly over θ in a neighborhood of θ_0.

Assumption SS0(i) is a form of an index exogeneity condition. Such an assumption has been used in various forms in the literature (e.g. Powell (1994).) The distinctive feature of this model stems from Assumption SS0(ii) which substantially relaxes Condition (6). The relaxation allows the individual components of X and ε to be stochastically dependent. Assumption SS0(iii) is slightly stronger than the usual condition that $S_{ZZ}(\theta_0)$ is invertible.

Under Assumptions SS0(i) and (iii), we can write the equation for observed outcomes as a partial linear regression model, and follow Robinson (1988) to identify β_0 as

$$\beta_0 = S_{ZZ}^{-1}(\theta_0) \cdot S_{ZY}(\theta_0),$$

once θ_0 is identified. To see that this β_0 is a special case of (3), let $\mathbf{1}_2$ be a 2×1 vector of

[4] An exception to this assumption is Chen and Khan (2003) who considered semiparametric sample selection models with conditional heteroskedasticity, and proposed a three-step estimation procedure.

ones, $S = \mathbf{1}_2 \otimes Z$, (the notation $\otimes$ represents the Kronecker product of matrices) and define $W = [Y; Z^\top]$, $\mu(U_\theta) = \mathbf{E}[W|U_\theta, D = 1]$,

$$\begin{aligned} a(\theta) &= \mathbf{E}\left[S \cdot \varphi(\mu(U_\theta))|D = 1\right], \text{ and} \\ b_0 &= \mathbf{E}\left[S \cdot W|D = 1\right], \end{aligned}$$

where $\varphi : \mathbf{R}^{d_Z+1} \to \mathbf{R}^{d_Z+1}$ is an identity map. Note that $a(\theta)$ and b_0 are $2d_Z \times (d_Z + 1)$ matrices. Furthermore, b_0 does not depend on θ. Given $2d_Z \times (d_Z + 1)$ matrices a and b, we denote a_{22} and b_{22} to be the $d_Z \times d_Z$ lower-right sub-blocks of a and b, and denote a_{11} and b_{11} to be the $d_Z \times 1$ upper-left sub-blocks of a and b. Then define

$$H(a, b) = (b_{22} - a_{22})^{-1}(b_{11} - a_{11}),$$

whenever $b_{22} - a_{22}$ is invertible. We can reformulate the identification result as follows:

$$\beta_0 = H(a(\theta_0), b_0),$$

which shows that β_0 is a special case of (3).

2.2.2 Example 2: Single-Index Matching Estimators of Treatment Effects on the Treated

Among various estimators of treatment effects used in the studies on program evaluations, matching estimators have been widely studied and used. (See Dehejia and Wahba (1998) and Heckman, Ichimura and Todd (1997, 1998) and references therein for matching methods in general.) While many studies of econometric methodologies use nonparametric specification of the propensity score (e.g. Hahn (1998), Hirano, Imbens and Ridder (2003)), a single-index restriction on the propensity score can be useful in avoiding curse of dimensionality.

When the propensity score is specified by logit or probit assumptions, the propensity score is strictly increasing in the single-index. In general, when the propensity score satisfies a single-index restriction and is a strictly increasing function of the single-index, identification of the average treatment effects on the treated through propensity score matching is equivalent to the identification through single-index matching, because the σ-field generated by the propensity score is the same as that generated by the single-index.

In the current example, we develop what this paper calls a *single-index matching estimator*. The main merit of the single-index matching estimators is that the estimator does not require a parametric distributional assumption on the propensity score, while avoiding the curse of dimensionality. The distinctive feature of the estimator as a result of this paper's framework is that the single-index component is allowed to be estimable only at the cube-root rate. Such a case is relevant when the assumption of independence between the observed component and the unobserved component in the propensity score is relaxed into the assumption of conditional median independence.

Let Y_1 and Y_0 be potential outcomes of treated and untreated individuals and $Z \in \{0, 1\}$ the treatment status, where $Z = 1$ for the status of treatment and $Z = 0$ for the status of non-treatment.[5] The parameter of interest is $\beta_0 = \mathbf{E}[Y_1 - Y_0|Z = 1]$, i.e., the treatment effect on the treated. We assume that X is a vector of covariates in $\mathbf{R}^d$ and

$$Z = 1\left\{X^\top\theta_0 \geq \varepsilon\right\},$$

where ε denotes the unobserved factor that affects the treatment status, and θ_0 is an unknown parameter. Define $U_\theta = F_\theta(X^\top\theta)$, where F_θ is the CDF of $X^\top\theta$ and is assumed to be strictly increasing, and we write simply $U_0 = U_{\theta_0}$. We also define the propensity score $P(U_\theta) = P\{Z = 1|U_\theta\}$.

Assumption SM0 : (i) $\mathbf{E}[Y_0|U_0, Z = 0] = \mathbf{E}[Y_0|U_0, Z = 1]$.

(ii) There exists $\eta > 0$ such that $\eta \leq P(U_0) \leq 1 - \eta$.

[5] The common notation for the treatment status is D, but the treatment status does not play the same role as D in (1). Hence we choose a different notation, Z, here.

(iii) $Med(\varepsilon|X) = 0$.

The first condition in Assumption SM0(i) is weaker than the unconfoundedness assumption, i.e., the assumption of conditional independence between (Y_1, Y_0) and Z given U_0, and as noted by Heckman, Ichimura, and Todd (1997), this assumption together with Assumption SM0(ii) suffices for identification of the average treatment effect on the treated. Assumption SM0(ii) requires that the propensity score is away from 0 and 1. The new feature of the model is Assumption SM0(iii) which says that the conditional median of the observed component in the propensity score is zero once X is conditioned on. This condition is much weaker than the common assumption that ε and X are independent.

Under Assumption SM0, we can identify β_0 as follows:

$$\beta_0 = \mathbf{E}\left[Y_1 - \mathbf{E}[Y_0|U_0, Z = 0]|Z = 1\right]. \tag{7}$$

It is not immediately seen that β_0 can be written in the form of (3), because the conditioning on $Z = 0$ in the inner conditional expectation is different from the conditioning on $Z = 1$ in the outer conditional expectation. To write it in the form of (3), rewrite β_0 as

$$\frac{\mathbf{E}\left[YZ\right]}{P\{Z = 1\}} - \frac{1}{P\{Z = 1\}} \cdot \mathbf{E}\left[Z \cdot \frac{\mathbf{E}[Y(1 - Z)|U_0]}{1 - P(U_0)}\right],$$

where $Y \equiv ZY_1 + (1 - Z)Y_0$ is an observable quantity. Define $W = [Y(1 - Z), Z]^\top$, and write $\mu_\theta(U_\theta) = \mathbf{E}[W|U_\theta]$. Let

$$b_0 = \begin{bmatrix} \mathbf{E}\left[YZ\right]/P\{Z = 1\} \\ 1/P\{Z = 1\} \end{bmatrix}, \text{ and } a(\theta) = \mathbf{E}\left[Z\varphi\left(\mu_\theta(U_\theta)\right)\right],$$

where $\varphi : \mathbf{R} \times (0, 1) \to \mathbf{R}$ is defined to be $\varphi(x, z) = x/(1 - z)$ for $(x, z) \in \mathbf{R} \times (0, 1)$. Note that b_0 does not depend on θ. Let $H : \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}$ be defined as $H(a, b) = b_1 - b_2 \cdot a$, where $a \in \mathbf{R}$ and $b = [b_1, b_2]^\top \in \mathbf{R}^2$. Then, we can write $\beta_0 = H\left(a(\theta_0), b_0\right)$, i.e., in the form of (3) with D there simply replaced by 1.

2.3 A Heuristic Summary of the Main Idea

As previously mentioned, the main contribution of this paper is to provide nontrivial sufficient conditions under which the first step estimator error of $\hat{\theta}$ does not affect the asymptotic distribution of $\hat{\beta}$, regardless of whether $\hat{\theta}$ is $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent. The development is based on the finding due to Song (2012) that under regularity conditions that are to be made precise later, the function $a(\theta)$ defined in (2) is very smooth in θ in a neighborhood of θ_0. More specifically, under regularity conditions, there exist $C > 0$ and $\varepsilon \in (0, 1/2]$ such that for each $\eta \in (0, \varepsilon]$,

$$\sup_{\theta \in B(\theta_0;\eta)} \|a(\theta) - a(\theta_0)\| \leq C\eta^2, \tag{8}$$

where $B(\theta_0;\eta)$ denotes the η-ball around θ_0, i.e., $B(\theta_0;\eta) \equiv \{\theta \in \Theta : ||\theta - \theta_0|| < \eta\}$. The novel feature of the above bound lies in the fact that the exponent of η is 2 (not 1), which says that the map a is *very smooth* in a neighborhood of θ_0.

To see how this result serves our purpose, we write

$$||\hat{\beta} - \tilde{\beta}|| = ||H(\hat{a}(\hat{\theta}), \hat{b}) - H(\hat{a}(\theta_0), \hat{b})|| \leq C||\hat{a}(\hat{\theta}) - \hat{a}(\theta_0)|| + o_P(1),$$

by the continuous differentiability of map H. As for the last term, observe that

$$\begin{aligned} \hat{a}(\hat{\theta}) - \hat{a}(\theta_0) &= \hat{a}(\hat{\theta}) - a(\hat{\theta}) - \{\hat{a}(\theta_0) - a(\theta_0)\} \\ &\quad + a(\hat{\theta}) - a(\theta_0) \\ &\equiv A_n + B_n, \text{ say.} \end{aligned}$$

As long as $||\hat{\theta} - \theta_0|| = o_P(1)$, the term A_n can be shown to be $o_P(1/\sqrt{n})$ using the standard arguments of stochastic equicontinuity. As for B_n, the result in (8) implies that with probability approaching one,

$$||a(\hat{\theta}) - a(\theta_0)|| \leq \sup_{\theta \in \Theta: ||\theta - \theta_0|| \leq \eta_n} \|a(\theta) - a(\theta_0)\| \leq C\eta_n^2, \text{ for some } C > 0, \tag{9}$$

if $||\hat{\theta} - \theta_0|| \leq \eta_n$ with probability approaching one. If $\hat{\theta} = \theta_0 + O_P(n^{-1/3})$, we find that by taking $\eta_n = n^{-1/3} \log n$, the left-end term of (9) is $o_P(1/\sqrt{n})$. Therefore, we conclude that

$$||\hat{\beta} - \tilde{\beta}|| = o_P(1/\sqrt{n}).$$

This implies that if $\sqrt{n}(\tilde{\beta} - \beta_0)$ has an asymptotic normal distribution, the quantity $\sqrt{n}(\hat{\beta} - \beta_0)$ has the same asymptotic normal distribution.

To see intuition behind (8), we first let $d_\varphi = L = 1$, φ be an identity map, and $D = 1$ for simplicity. Then for θ in a neighborhood of θ_0, $a(\theta) - a(\theta_0)$ is written as

$$\begin{aligned}
& \mathbf{E}\left[S \cdot \{\mathbf{E}[W|U_\theta] - \mathbf{E}[W|U_0]\}\right] \\
= \; & \mathbf{E}\left[\mathbf{E}[S|U_\theta, U_0] \cdot \{\mathbf{E}[W|U_\theta] - \mathbf{E}[W|U_\theta, U_0]\}\right] \\
& + \mathbf{E}\left[\mathbf{E}[S|U_\theta, U_0] \cdot \{\mathbf{E}[W|U_\theta, U_0] - \mathbf{E}[W|U_0]\}\right] \\
= \; & \mathbf{E}\left[\{\mathbf{E}[S|U_\theta, U_0] - \mathbf{E}[S|U_\theta]\} \cdot \{\mathbf{E}[W|U_\theta] - \mathbf{E}[W|U_\theta, U_0]\}\right] \\
& + \mathbf{E}\left[\{\mathbf{E}[S|U_\theta, U_0] - \mathbf{E}[S|U_0]\} \cdot \{\mathbf{E}[W|U_\theta, U_0] - \mathbf{E}[W|U_0]\}\right].
\end{aligned}$$

The last equality uses the law of iterated conditional expectations. Therefore, by Cauchy-Schwarz inequality, we find that

$$\begin{aligned}
|a(\theta) - a(\theta_0)| \; \leq \; & \sqrt{\mathbf{E}(\mathbf{E}[S|U_\theta, U_0] - \mathbf{E}[S|U_\theta])^2}\sqrt{\mathbf{E}(\mathbf{E}[W|U_\theta] - \mathbf{E}[W|U_\theta, U_0])^2} \\
& + \sqrt{\mathbf{E}(\mathbf{E}[S|U_\theta, U_0] - \mathbf{E}[S|U_0])^2}\sqrt{\mathbf{E}(\mathbf{E}[W|U_\theta, U_0] - \mathbf{E}[W|U_0])^2}.
\end{aligned}$$

Once we show that for some $C > 0$,

$$\mathbf{E}(\mathbf{E}[S|U_\theta, U_0] - \mathbf{E}[S|U_\theta])^2 \leq C||\theta - \theta_0||^2, \tag{10}$$

and so on, we obtain the result of (8). However, finding the bound in (10) is a nontrivial task. (See Song (2012) for details.) [6]

3 Inference and Asymptotic Theory

3.1 General Asymptotic Theory

Let us consider an estimation method of $\hat{\beta}$. Instead of putting forth high level conditions, we choose a specific estimator $\hat{\beta}$ and provide low level conditions. Suppose that we are given an estimator $\hat{\theta}$ of θ_0 such that $\hat{\theta} = \theta_0 + O_P(n^{-1/3})$ (Assumption C2 below). Assume that $\{(X_i, W_i, S_i, D_i)\}_{i=1}^n$ is a random sample from the joint distribution of (X, W, S, D). Let $\hat{U}_k \equiv \frac{1}{n}\sum_{i=1}^n 1\{X_i^\top \hat{\theta} \leq X_k^\top \hat{\theta}\}$ and define

$$\hat{\mu}(\hat{U}_k) \equiv \frac{\sum_{i=1}^n D_i W_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n D_i K_h(\hat{U}_i - \hat{U}_k)},$$

as an estimator of $\mu_0(U_0)$, where $K_h(u) \equiv K(u/h)/h$, $K : \mathbf{R} \to \mathbf{R}$ is a kernel function, and h is a bandwidth parameter. The estimator is a symmetrized nearest neighborhood (SNN) estimator. Symmetrized nearest neighborhood estimation is a variant of nearest neighborhood estimation originated by Fix and Hodges (1951), and analyzed and expanded by Stone (1977). Robinson (1987) introduced k-nearest neighborhood estimation in semiparametric models in the estimation of conditional heteroskedasticity of unknown form. The symmetrized nearest neighborhood estimation that this paper uses was proposed by Yang (1981) and further studied by Stute (1984).

[6] The heuristics here uses the assumption that φ is a linear map. When φ is nonlinear yet twice continuously differentiable, we can linearize it to obtain a similar bound. (See Song (2012) for details.)

Define

$$\hat{a}(\hat{\theta}) \equiv \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \{S_i \cdot \varphi(\hat{\mu}(\hat{U}_i))\}. \tag{11}$$

The estimator $\hat{a}(\hat{\theta})$ is a sample analogue of $a(\theta_0)$, where the conditional expectations are replaced by the nonparametric estimators and unconditional expectations by the sample mean.

Suppose that we are given a consistent estimator $\hat{b}$ of b_0 (Assumption G1(iii) below). Then, our estimator takes the following form:

$$\hat{\beta} = H(\hat{a}(\hat{\theta}), \hat{b}). \tag{12}$$

We make the following assumptions. The assumptions are divided into two groups. The first group of assumptions (denoted by Assumptions C1-C3) are commonly assumed throughout the examples when we discuss them later. On the other hand, the second group of assumptions (denoted by Assumptions G1-G2) are the ones for which sufficient conditions will be provided later when we discuss the examples.

Assumption C1 : (i) For some $\varepsilon > 0$, $P\{D = 1|U_0 = u\} > \varepsilon$ for all $u \in [0, 1]$.
(ii) There exists $\varepsilon > 0$ such that for each $\theta \in B(\theta_0; \varepsilon)$, (a) $X^\top\theta$ is continuous and its conditional density function given $D = 1$ is bounded uniformly over $\theta \in B(\theta_0; \varepsilon)$ and bounded away from zero on the interior of its support uniformly over $\theta \in B(\theta_0; \varepsilon)$, and (b) the set $\{x'\theta : \theta \in B(\theta_0; \varepsilon), x \in S_m\}$ is an interval of finite length for all $1 \leq m \leq M$.

Assumption C2 : $||\hat{\theta} - \theta_0|| = O_P(n^{-1/3})$.

Assumption C3 : (i) $K(\cdot)$ is bounded, nonnegative, symmetric, compact supported, twice continuously differentiable with bounded derivatives on the interior of the support, and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^3 + n^{-1/2}h^{-2}(-\log h) \to 0$.

Assumption C1(ii)(a) excludes the case where $\theta_0 = 0$. Assumption C1(ii)(b) is satisfied when S_m is bounded and convex. We can weaken Assumption C1(ii) by replacing it with certain tail conditions of X or $X^\top\theta$ at the expense of a more complicated exposition. Assumption C2 allows $\hat{\theta}$ to be either $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent. Assumption C3 concerns the kernel and the bandwidth. Assumption C3(i) is satisfied, for example, by a quartic kernel: $K(u) = (15/16)(1-u^2)^2 1\{|u| \leq 1\}$.

Assumption G1 : (i) For $p \geq 4$, $\sup_{x \in \mathcal{S}_X} \mathbf{E}\left[||W||^p|X = x\right] + \sup_{x \in \mathcal{S}_X} \mathbf{E}\left[||S||^p|X = x\right] < \infty$.
(ii) (a) φ is twice continuously differentiable on the interior of the support of $\mathbf{E}[W|X]$ with derivatives bounded on the support of $\mathbf{E}[W|X]$, and (b) there exists $\eta > 0$ such that for all $b \in B(b_0;\eta)$, $H(\cdot, b)$ is continuously differentiable at $a = a(\theta_0)$ and the derivative $\partial H(a,b)/\partial a$ is continuous at $(a(\theta_0), b_0)$.
(iii) $\hat{b} = b_0 + o_P(1)$.

Assumption G2 : (i) For each $m = 1, \cdots, M$, both $\mathbf{E}[S|X_1 = \cdot, (X_2, D) = (x_m, 1)]$ and $\mathbf{E}[W|X_1 = \cdot, (X_2, D) = (x_m, 1)]$ are Lipschitz continuous.
(ii) $\mathbf{E}[W|U_\theta = \cdot]$ is twice continuously differentiable with derivatives bounded uniformly over $\theta \in B(\theta_0;\varepsilon)$ with some $\varepsilon > 0$.

Assumption G1(i) requires moment conditions with $p \geq 4$. Assumption G1(ii) is easy to check, because φ is explicitly known in many examples. Assumption G1(iii) requires that $\hat{b}$ be a consistent estimator of b_0. As we will see from the examples, a $\sqrt{n}$-consistent estimator for b_0 is typically available. The smoothness conditions in Assumptions G2(i) and (ii) are often used in the literature of nonparametric estimation.

Theorem 1: *Suppose that Assumptions C1–C3 and G1-G2 hold and that* $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V)$ *for some positive definite matrix* V. *Then*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V).$$

REMARKS 1 : The asymptotic covariance matrix V in Theorem 1 is the same asymptotic covariance matrix that one would have obtained had θ_0 been used instead of $\hat{\theta}$. Therefore, the estimation error in $\hat{\theta}$ does not affect the asymptotic distribution of $\hat{\beta}$. When the nuisance parameter estimator $\hat{\theta}$ is $\sqrt{n}$-consistent, such a phenomenon has been observed to arise in other contexts (e.g. Song (2009)). To the best of the author's knowledge, there has not been a literature that shows a similar phenomenon even when $\hat{\theta}$ is $n^{1/3}$-consistent.

2: The computation of V such that $\sqrt{n}(\tilde{\beta} - \beta) \overset{d}{\to} N(0, V)$ can be done using the standard procedure. (e.g. Newey and McFadden (1994)). Section 3.2 below derives the asymptotic covariance matrix V for the examples in Section 2.2. For the derivation, one does not need to rely on the form (3). Writing β_0 into the form (3) is done only to ensure that Theorem 1 is applicable.

3: The proof of Theorem 1 uses a Bahadur representation of sample linear functionals of SNN estimators that is established in the appendix. In fact, the representation can also be used to derive the asymptotic covariance matrix V, and is useful in various specification tests or estimation for semiparametric models.

4: Theorem 1 implies that there exists a simple bootstrap procedure for $\hat{\beta}$ that is asymptotically valid, even if the first-step estimator $\hat{\theta}$ follows cube-root asymptotics. This is interesting given that nonparametric bootstrap fails for $\hat{\theta}$. (Abrevaya and Huang (2005)). Since there is no clear advantage of using this bootstrap over the asymptotic covariance matrix of Theorem 1, this paper omits the details.

3.2 Examples Revisited

In this section, we revisit the examples discussed in Section 2.2. In each example, we first provide sufficient conditions that yield Assumptions G1-G2. (Recall that Assumptions C1-C3 are made commonly in these examples.) Then we show how we construct an estimator of β_0 in detail. Finally, we present the asymptotic distribution of $\hat{\beta}$ along with the explicit asymptotic covariance matrix formula.

3.2.1 Example 1: Sample Selection Models with Conditional Median Restrictions

In this example, Assumptions G1-G2 are translated into the following conditions.

Assumption SS1 : For $p \geq 4$, $\sup_{x \in \mathcal{S}_X} \mathbf{E}\left[|Y|^p | X = x\right] + \sup_{x \in \mathcal{S}_X} \mathbf{E}\left[||Z||^p | X = x\right] < \infty$.

Assumption SS2 : (i) For each $m = 1, \cdots, M$, both $\mathbf{E}[Y | X_1 = \cdot, (X_2, D) = (x_m, 1)]$ and $\mathbf{E}[Z | X_1 = \cdot, (X_2, D) = (x_m, 1)]$ are Lipschitz continuous.
(ii) $\mathbf{E}[Y | U_\theta = \cdot]$ and $\mathbf{E}[Z | U_\theta = \cdot]$ are twice continuously differentiable with derivatives bounded uniformly over $\theta \in B(\theta_0; \varepsilon)$ with some $\varepsilon > 0$.

Since φ that constitutes $a(\theta)$ is an identity map in this example, Assumption G1(ii)(a) is already fulfilled by Assumptions SS1 and SS2(ii).

Let us consider an estimator of β_0 in the sample selection models in Example 1. With $\hat{U}_k$ as defined previously, let

$$\hat{\mu}_Y(\hat{U}_k) \equiv \frac{\sum_{i=1}^n D_i Y_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n D_i K_h(\hat{U}_i - \hat{U}_k)} \text{ and } \hat{\mu}_Z(\hat{U}_k) \equiv \frac{\sum_{i=1}^n D_i Z_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n D_i K_h(\hat{U}_i - \hat{U}_k)}. \tag{13}$$

Using $\hat{\mu}_Y(\hat{U}_k)$ and $\hat{\mu}_Z(\hat{U}_k)$, we define

$$\begin{aligned}
\hat{S}_{ZZ} &\equiv \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n (Z_i - \hat{\mu}_Z(\hat{U}_i))(Z_i - \hat{\mu}_Z(\hat{U}_i))^\top D_i \text{ and} \\
\hat{S}_{ZY} &\equiv \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n (Z_i - \hat{\mu}_Z(\hat{U}_i))(Y_i - \hat{\mu}_Y(\hat{U}_i)) D_i,
\end{aligned}$$

which are estimated versions of $S_{ZZ}(\theta_0)$ and $S_{ZY}(\theta_0)$. An estimator of β_0 is given by

$$\hat{\beta} \equiv \hat{S}_{ZZ}^{-1} \cdot \hat{S}_{ZY}. \tag{14}$$

This is the estimator proposed by Robinson (1988), except that we have a single-index in the conditional expectations. We also let $\tilde{\beta}$ be $\hat{\beta}$ except that $\hat{\theta}$ is replaced by θ_0.

Suppose that Assumptions C1–C3 and SS1-SS2 hold and that $\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V_{SS})$ for some positive definite matrix V_{SS}. Then by Theorem 1,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_{SS}).$$

The computation of V_{SS} can be done in a standard manner. Under regularity conditions, the asymptotic variance V_{SS} takes the following form: $V_{SS} = S_{ZZ}^{-1}(\theta_0)\Omega S_{ZZ}^{-1}(\theta_0)$ with $\sigma^2(U_0) \equiv Var(v|U_0, D = 1)$,

$$\Omega \equiv \mathbf{E}\left[\sigma^2(U_0)(Z - \mathbf{E}[Z|D = 1, U_0])(Z - \mathbf{E}[Z|D = 1, U_0])^\top |D = 1\right]/P_1.$$

The derivation can be obtained by using the Bahadur representation in the appendix (Lemma B3).

3.2.2 Example 2: Single-Index Matching Estimators of Treatment Effects on the Treated

We introduce translations of Assumptions G1-G2 in this example.

Assumption SM1 : For $p \geq 4$, $\sup_{x \in \mathcal{S}_X} \mathbf{E}\left[|Y|^p | X = x\right] < \infty$.

Assumption SM2 :(i) For each $m = 1, \cdots, M$, both $P\{Z = 1|X_1 = \cdot, X_2 = x_m\}$ and $\mathbf{E}[Y|X_1 = \cdot, X_2 = x_m]$ are Lipschitz continuous.
(ii) $P\{Z = 1|U_\theta = \cdot\}$ and $\mathbf{E}[Y|U_\theta = \cdot]$ are twice continuously differentiable with derivatives bounded uniformly over $\theta \in B(\theta_0; \varepsilon)$ with some $\varepsilon > 0$.

To construct an estimator of the average treatment effect on the treated based on the single-index matching, we first define

$$\begin{aligned}\hat{\mu}_{(1-Z)Y}(\hat{U}_k) &\equiv \frac{\sum_{i=1}^n (1-Z_i)Y_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n K_h(\hat{U}_i - \hat{U}_k)} \text{ and} \\ \hat{P}(\hat{U}_k) &\equiv \frac{\sum_{i=1}^n Z_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n K_h(\hat{U}_i - \hat{U}_k)}.\end{aligned}$$

Then, the sample analogue principle suggests

$$\hat{\beta} = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i \left\{ Y_i - \frac{\hat{\mu}_{(1-Z)Y}(\hat{U}_k)}{1 - \hat{P}(\hat{U}_k)} \right\}.$$

If we define

$$\hat{\mu}(\hat{U}_k) = \frac{\sum_{i=1}^n (1-Z_i)Y_i K_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^n (1-Z_i) K_h(\hat{U}_i - \hat{U}_k)},$$

we can rewrite the estimator as

$$\hat{\beta} = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i \{Y_i - \hat{\mu}(\hat{U}_k)\}.$$

This takes precisely the same form as the propensity score matching estimators of Heckman, Ichimura, and Todd (1998), except that instead of propensity score matching, the estimator uses single-index matching.

As before, we let $\tilde{\beta}$ be $\hat{\beta}$ except that $\hat{\theta}$ is replaced by θ_0. Suppose that Assumptions C1–C3 and SM1-SM2 hold and that $\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, V_{SM})$ for some positive definite matrix V_{SM}. Then, by Theorem 1,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_{SM}).$$

Under regularity conditions, the asymptotic variance V_{SM} takes the following form: with

$\mu_d(U_0) = \mathbf{E}[Y|U_0, Z = d]$ and $P_d = P\{Z_i = d\}$ for $d \in \{0, 1\}$,

$$\begin{aligned} V_{SM} &= \mathbf{E}\left[(Y - \mu_1(U_0))^2 |Z = 1\right]/P_1 \\ &\quad +\mathbf{E}\left[(Y - \mu_0(U_0))^2 P^2(U_0)/(1 - P(U_0))^2|Z = 0\right](1 - P_1)/P_1^2 \\ &\quad +Var\left(\mu_1(U_0) - \mu_0(U_0)|Z = 1\right)/P_1. \end{aligned}$$

4 A Monte Carlo Simulation Study

In this section, we present and discuss some Monte Carlo simulation results. We consider the following data generating process. Let

$$Z_i = U_{1i} - (\eta_{1i}/2)\mathbf{1} \text{ and } X_i = U_{2i} - \eta_i/2,$$

where U_{1i} is an i.i.d. random vector in $\mathbf{R}^3$ constituted by independent random variables with uniform distribution on $[0, 1]$, $\mathbf{1}$ is a 3 dimensional vector of ones, U_{2i} and η_i are random vectors in $\mathbf{R}^k$ with entries equal to i.i.d. random variables of uniform distribution on $[0, 1]$. The dimension k is chosen from $\{3, 6\}$. The random variable η_{1i} is the first component of η_i. Then, the selection mechanism is defined as

$$D_i = 1\{X_i^\top \theta_0 + \varepsilon_i \geq 0\},$$

where ε_i follows the distribution of $T_i \cdot \varphi(X_i^\top \theta_0) + e_i$, with $e_i \sim N(0, 1)$, and T_i and $\varphi(\cdot)$ are chosen as follows:

Specification AN: $T_i \sim N(0, 1),\ \varphi = 2\Phi(z^2 + |z|)$

Specification AT: $T_i \sim t_1,\ \varphi = 2\Phi(z^2 + |z|)$,

Specification BN: $T_i \sim N(0, 1) + e_i,\ \varphi = \exp(z - 1)$

Specification BT: $T_i \sim t_1 + e_i,\ \varphi = \exp(z - 1)$, where t_1 denotes the t-distribution with degree of freedom 1 and Φ denotes the standard normal CDF.

Hence the selection mechanism has errors that are conditionally heteroskedastic, and in the case of DGPs with AT and BT, heavy tailed. We define the latent outcome Y_i^* as follows:

$$Y_i^* = Z_i^\top \beta_0 + v_i,$$

where $v_i \sim (2\zeta_i + e_i) \times 2\Phi\left((X_i^\top \theta_0)^2 + |X_i^\top \theta_0|\right)$ and $\zeta_i \sim N(0,1)$ independent of the other random variables. Therefore, v_i in the outcome equation and ε_i in the selection equation are correlated, so that the data generating process admits the sample selection bias. We set θ_0 to be the vector of 2's and $\beta_0 = [2,2,2]^\top$. In the simulation studies we estimated θ_0 by using the maximum score estimation to obtain $\hat{\theta}$.

We compare the performances of the two estimators of β_0, $\hat{\beta}(\hat{\theta})$ ("Plug-in $\hat{\theta}$") and $\hat{\beta}(\theta_0)$ ("Plug-in θ_0") in terms of mean absolute deviation (MAE) and mean squared error (MSE). Bandwidths for the estimation of $\mathbf{E}[Y_i|X_i^\top \theta_0, D_i = 1]$ and $\mathbf{E}[Z_i|X_i^\top \theta_0, D_i = 1]$ were chosen separately using a least-squares cross-validation method. If the role of the sample selection bias were already marginal, the estimation error effect of $\hat{\theta}$ would be small accordingly, preventing us from discerning the negligibility of the estimation error effect of $\hat{\theta}$ from the negligible sample selection bias. Hence, we also report the results from the estimation of β that ignores the sample selection bias (w/o BC: Without (Sample Selection) Bias Correction).

Table 1 reports the average of MAEs and MSEs of estimators for the individual components of β_0. It shows that the performance of the estimators remains similar regardless of whether θ_0 is used or $\hat{\theta}$ is used. When the sample size is increased from 300 to 500, the estimators perform better as expected. The negligibility of the effect of the estimation error in $\hat{\theta}$ is not due to inherently weak sample selection bias, as it is evident when we compare the results with those from the estimators that ignore the sample selection bias (w/o BC).

Table 1: The Performance of the Estimators in Terms of MAE and RMSE (Specification A)

			$k = 3$			$k = 6$		
	Specification		Plug-in θ_0	Plug-in $\hat{\theta}$	w/o BC	Plug-in θ_0	Plug-in $\hat{\theta}$	w/o BC
$n = 300$	Spec. AN	MAE	0.6911	0.6909	0.7837	0.6458	0.6453	0.6436
		RMSE	2.2533	2.2539	2.8485	1.9634	1.9621	1.9479
	Spec. AT	MAE	0.7345	0.7351	0.7729	0.6913	0.6918	0.6739
		RMSE	2.5437	2.5457	2.7955	2.2408	2.2446	2.1316
$n = 500$	Spec. AN	MAE	0.5327	0.5328	0.6717	0.4965	0.4966	0.5122
		RMSE	1.3428	1.3432	2.0406	1.1615	1.1620	1.2270
	Spec. AT	MAE	0.5658	0.5654	0.6360	0.5308	0.5310	0.5316
		RMSE	1.5154	1.5134	1.8833	1.3256	1.3263	1.3318
$n = 1000$	Spec. AN	MAE	0.3766	0.3765	0.5765	0.3475	0.3475	0.3880
		RMSE	0.6693	0.6696	1.4206	0.5707	0.5707	0.6993
	Spec. AT	MAE	0.3981	0.3980	0.5182	0.3734	0.3734	0.3982
		RMSE	0.7455	0.7449	1.2094	0.6570	0.6572	0.7368

Table 2: The Performance of the Estimators in Terms of MAE and RMSE (Specification B)

			$k = 3$			$k = 6$		
	Specification		Plug-in θ_0	Plug-in $\hat{\theta}$	w/o BC	Plug-in θ_0	Plug-in $\hat{\theta}$	w/o BC
$n = 300$	Spec. BN	MAE	0.6707	0.6710	1.4572	0.7017	0.7020	1.4565
		RMSE	2.1182	2.1211	8.0981	2.3237	2.3277	8.1755
	Spec. BT	MAE	0.7209	0.7212	1.2461	0.7563	0.7561	1.2164
		RMSE	2.4474	2.4489	6.3876	2.7029	2.7050	6.2357
$n = 500$	Spec. BN	MAE	0.5260	0.5260	1.4355	0.5504	0.5515	1.4331
		RMSE	1.2997	1.2990	7.2972	1.4298	1.4350	7.3290
	Spec. BT	MAE	0.5649	0.5649	1.1998	0.5867	0.5870	1.1574
		RMSE	1.5078	1.5069	5.5043	1.6318	1.6338	5.2330
$n = 1000$	Spec. BN	MAE	0.3870	0.3869	1.4258	0.4174	0.4184	1.4317
		RMSE	0.7040	0.7040	6.6726	0.8148	0.8190	6.7706
	Spec. BT	MAE	0.4053	0.4052	1.1746	0.4230	0.4235	1.1291
		RMSE	0.7721	0.7709	4.7835	0.8444	0.8469	4.5057

In Tables 3 and 4, the finite sample coverage probabilities of the confidence sets based on the asymptotic normal distribution are reported. These tables report only the coverage probabilities of the first component of the estimators of β_0. The performance of the remaining components was similar. In Table 3, the results were obtained from Specification A, and in Table 4, from Specification B. Recall that Specification B is associated with a more severe selection bias than Specification A as we saw from Tables 1-2.

Table 3: The Performance of the Confidence Intervals (Specification A)

			$k = 3$		$k = 6$	
	Specification	Nom. Cov. Prob.	Plug-in θ_0	Plug-in $\hat{\theta}$	Plug-in θ_0	Plug-in $\hat{\theta}$
$n = 300$	Spec. AN	99%	0.9877	0.9881	0.9875	0.9877
		95%	0.9449	0.9444	0.9487	0.9468
		90%	0.8938	0.8930	0.8972	0.8947
	Spec.AT	99%	0.9858	0.9866	0.9871	0.9861
		95%	0.9416	0.9398	0.9429	0.9427
		90%	0.8866	0.8878	0.8900	0.8900
$n = 500$	Spec. AN	99%	0.9886	0.9891	0.9879	0.9880
		95%	0.9480	0.9484	0.9456	0.9453
		90%	0.8988	0.8980	0.8957	0.8978
	Spec. AT	99%	0.9859	0.9864	0.9878	0.9883
		95%	0.9445	0.9449	0.9466	0.9472
		90%	0.8940	0.8950	0.8935	0.8958
$n = 1000$	Spec. AN	99%	0.9891	0.9884	0.9909	0.9910
		95%	0.9463	0.9468	0.9513	0.9508
		90%	0.8956	0.8964	0.9030	0.9031
	Spec. AT	99%	0.9858	0.9862	0.9898	0.9899
		95%	0.9435	0.9447	0.9488	0.9474
		90%	0.8953	0.8969	0.8993	0.9005

Table 4: The Performance of the Confidence Intervals (Specification B)

			$k = 3$		$k = 6$	
	Specification	Nom. Cov. Prob.	Plug-in θ_0	Plug-in $\hat{\theta}$	Plug-in θ_0	Plug-in $\hat{\theta}$
$n = 300$	Spec. BN	99%	0.9832	0.9829	0.9824	0.9823
		95%	0.9334	0.9340	0.9339	0.9345
		90%	0.8852	0.8837	0.8777	0.8796
	Spec.BT	99%	0.9851	0.9853	0.9833	0.9827
		95%	0.9382	0.9388	0.9377	0.9384
		90%	0.8898	0.8903	0.8844	0.8851
$n = 500$	Spec. BN	99%	0.9834	0.9833	0.9809	0.9805
		95%	0.9353	0.9357	0.9295	0.9289
		90%	0.8825	0.8841	0.8770	0.8744
	Spec. BT	99%	0.9844	0.9844	0.9828	0.9826
		95%	0.9399	0.9402	0.9360	0.9363
		90%	0.8829	0.8837	0.8838	0.8815
$n = 1000$	Spec. BN	99%	0.9797	0.9794	0.9765	0.9770
		95%	0.9264	0.9276	0.9156	0.9143
		90%	0.8679	0.8680	0.8511	0.8505
	Spec. BT	99%	0.9859	0.9858	0.9833	0.9830
		95%	0.9364	0.9367	0.9323	0.9310
		90%	0.8811	0.8820	0.8760	0.8750

First, observe that the performances between the estimator using true parameter θ_0 and the estimator using its estimator $\hat{\theta}$ is almost negligible in finite samples, as expected from the asymptotic theory. This is true regardless of whether we use three or six covariates. In

Specification B, the confidence sets tend to undercover the true parameter, as compared to Specification A. Nevertheless, the difference in coverage probabilities between the estimator using θ_0 and the estimator using $\hat{\theta}$ is still very negligible. Finally, the performances do not show much difference with regard to the heavy tailedness of the error distribution in the selection equation, as seen from comparing results between Specifications AN and AT or between Specifications BN and BT.

5 Empirical Application: Female Labor Supply from NLSY79

In this section, we illustrate the proposal of this paper by estimating a simple female labor supply model:

$$\begin{aligned} h_i &= \beta_0 + \log(w_i)\beta_1 + Z_{2i}^{\top}\beta_4 + \varepsilon_i \text{ and} \\ D_i &= 1\left\{X_i^{\top}\theta_0 \geq \eta_i\right\}, \end{aligned}$$

where h_i denotes hours that the i-th female worker worked, w_i her hourly wage, Z_{2i} denotes other demographic variables. The following table shows different specifications that this study used.

Table 5 : Variables used for Z_{2i} and X_i

		Variables Used
Specification I	Z_{2i}	nonwife income, age, schooling # kids w/ age 0-5, # kids w/ age 6-18,
	X_i	mother and father's schooling age, schooling
Specification II	Z_{2i}	nonwife income, age, schooling # kids w/ age 0-5, # kids w/ age 6-18,
	X_i	mother and father's schooling age, schooling, and household income
Specification III	Z_{2i}	nonwife income, age, schooling # kids w/ age 0-5, # kids w/ age 6-18,
	X_i	mother and father's schooling age, schooling, household income, and age&school interaction

The data sets were taken from NLSY79 for the 1998 round. The data set used in this study contains 960 female workers, after eliminating the individuals with missing values for demographic variables used in this study. In particular, the data set excluded those with missing values for household's total income, because we cannot compute nonwife income or nonlabor income from the data sets. The income variables are in 1998 dollars. The following table offers summary statistics of the variables, and also compares them before and after the selection process.

Table 6 : Summary Statistics of Variables

		Whole	Sample	$(n = 1268)$		Selected	Sample	$(n = 960)$
	mean	std dev	min	max	mean	std dev	min	max
mother's schooling	11.94	2.41	0	20	11.95	2.35	0	19
father's schooling	12.25	3.13	0	20	12.29	3.11	1	20
wife's labor incme	24066.3	20086.7	220	147970	24519.8	19927.9	220	147970
hsbd's labor incme	50821.9	40895.6	52	212480	49236.1	38190.5	52	212480
hshld's total incme	69603.0	45415.0	10	244343	70986.7	45085.1	700	244343
employment status	0.74	0.44	0	1	0.75	0.43	0	1
wife's schooling	13.8	2.39	6	20	13.9	2.31	7	20
wife's age	36.9	2.23	33	41	36.9	2.22	33	41
wife's hours	1447.4	966.1	0	6708	1.464.1	948.1	0	5200
husband's age	39.4	5.09	26	70	39.24	4.92	27	62
husband's schooling	13.8	2.61	3	20	13.7	2.50	3	20
# kids w/ age 0-5	0.46	0.73	0	4	0.47	0.73	0	4
# kids w/ age 6-18	1.33	1.11	0	6	1.32	1.08	0	5

In this study, we focus on how the estimates of coefficients in the outcome equation vary across different specifications of X_i and different methods of estimating θ_0 in the participation equation. We estimated the model using three different estimation methods. The first method is OLS, ignoring sample selection. The second method is Heckman's two step approach assuming the joint normality of the errors in the outcome and selection equations. The third method employs a semiparametric approach through the formulation of partial linear model and following the procedure of Robinson (1988). As for the third method, this study considered two different methods of estimating the coefficients to X_i : probit and maximum score estimation.

The results are shown in Tables 6-9. (The covariates in Z_i and X_i were appropriately rescaled to ensure numerical stability.) First, nonwife income and the number of young and old children play a significant role in determining the labor supply of female workers. This result is robust through different model specifications, although the significance of nonwife income is somewhat reduced when one incorporates sample selection. The negative effect of the number of children is conspicuous, with the effect of the number of young children stronger than that of old children. In contrast, the significant role that the female worker's age and schooling appear to play in the case of OLS or Heckman's two step procedure with the joint normality assumption disappears or is substantially reduced, when one moves to a semiparametric model.

Finally, it is interesting to observe that within the framework of a semiparametric model, the effects of log wage and nonwife income on labor participation are shown to be more significant in the case of using maximum score estimation than in the case of using a probit estimator $\hat{\theta}$. This appears to be some evidence against the assumption that X_i and ε_i in the selection equation are independent. While a formal testing procedure seems appropriate, a direct test comparing the estimates are not available in the literature as far as the author is concerned. In particular, the standard Hausman type test will not have an asymptotically exact size because when the probit estimator and the maximum score estimator are both consistent, the asymptotic distribution of $\sqrt{n}\{\hat{\beta}(\hat{\theta}_{probit})-\hat{\beta}(\hat{\theta}_{mx.scr})\}$, $\hat{\theta}_{probit}$ denoting a probit estimator of θ_0 and $\hat{\theta}_{mx.scr}$ a maximum score estimator, will be degenerate. The latter degeneracy is a major implication of Theorem 1 in this paper.

Table 7: Estimation of Female Labor Participation (Specification I)
(In the parentheses are standard errors.)

	OLS	Mill's Ratio	Semiparametric	Model
			$\hat{\theta}$ w/ Probit	$\hat{\theta}$ w/Mx. Scr
Log Wage	−37.500	−41.857	−46.764	−50.458
	(40.890)	(21.520)	(60.550)	(60.799)
Nonwife Income	−0.0248	−0.0251	−0.0197	−0.0217
	(0.0080)	(0.0042)	(0.0111)	(0.0111)
Young Children	−0.1491	−0.1558	−0.1748	−0.1761
	(0.0424)	(0.0223)	(0.0390)	(0.0395)
Old Children	−0.1217	−0.1233	−0.1310	−0.1305
	(0.0258)	(0.0137)	(0.0251)	(0.0249)
Age	0.0466	0.0438	−0.0024	0.0097
	(0.0046)	(0.0030)	(0.0131)	(0.0163)
Schooling	0.0368	0.0382	0.0158	0.0227
	(0.0120)	(0.0078)	(0.0145)	(0.0228)

Table 8: Estimation of Female Labor Participation (Specification II)
(In the parentheses are standard errors.)

	OLS	Mill's Ratio	Semiparametric	Model
			$\hat{\theta}$ w/ Probit	$\hat{\theta}$ w/Mx. Scr
Log Wage	−37.500	−37.800	−50.691	−212.93
	(40.890)	(2.9290)	(59.971)	(65.328)
Nonwife Income	−0.0248	−0.0247	−0.0222	−0.0937
	(0.0080)	(0.0007)	(0.0110)	(0.0172)
Young Children	−0.1491	−0.1500	−0.1724	−0.1543
	(0.0424)	(0.0030)	(0.0385)	(0.0374)
Old Children	−0.1217	−0.1219	−0.1305	−0.1140
	(0.0258)	(0.0018)	(0.0252)	(0.0234)
Age	0.0466	0.0462	−0.0097	0.0199
	(0.0046)	(0.0004)	(0.0131)	(0.0123)
Schooling	0.0368	0.0369	0.0124	−0.0226
	(0.0120)	(0.0010)	(0.0143)	(0.0132)

Table 9: Estimation of Female Labor Participation (Specification III)
(In the parentheses are standard errors.)

	OLS	Mill's Ratio	Semiparametric	Model
			$\hat{\theta}$ w/ Probit	$\hat{\theta}$ w/Mx. Scr
Log Wage	-37.500	-49.653	-48.206	-145.471
	(40.890)	(83.915)	(59.261)	(61.158)
Nonwife Income	-0.0248	-0.0218	-0.0226	-0.0689
	(0.0080)	(0.0245)	(0.0110)	(0.0137)
Young Children	-0.1491	-0.1766	-0.1735	-0.1557
	(0.0424)	(0.0809)	(0.0382)	(0.0366)
Old Children	-0.1217	-0.1301	-0.1291	-0.1176
	(0.0258)	(0.0477)	(0.0250)	(0.0241)
Age	0.0466	0.0079	-0.0056	0.0028
	(0.0046)	(0.0142)	(0.0122)	(0.0120)
Schooling	0.0368	0.0244	0.0163	-0.0112
	(0.0120)	(0.0387)	(0.0144)	(0.0133)

6 Conclusion

This paper focuses on semiparametric models where the identified parameter involves conditional expectations with a single-index as a conditioning variable. This paper offers a set of sufficient conditions under which the first step estimator of the single-index does not have a first order impact on the asymptotic distribution of the second step estimator. The remarkable aspect of the result is that the asymptotic negligibility of the first step estimator holds even when the estimator follows cube-root asymptotics. This asymptotic negligibility is also demonstrated through Monte Carlo simulation studies. The usefulness of this procedure is illustrated by an empirical study of female labor supply using an NLSY79 data set.

7 The Appendix

7.1 Proof of Theorem 1

Lemma A1: *Suppose that Assumptions C1, G1(i)(ii), and G2(i) hold. Then, there exist* $C > 0$ *and* $\varepsilon > 0$ *such that for each* $\eta \in (0, \varepsilon]$,

$$\sup_{\theta \in \mathbf{R}^d : ||\theta - \theta_0|| \leq \eta} \|a(\theta) - a(\theta_0)\| \leq C\eta^2.$$

Proof: The result follows from Theorem 1 of Song (2012). See Example 1 there. ■

Proof of Theorem 1: Without loss of generality, let K be $[-1/2, 1/2]$-supported. It suffices to focus on the case where H is $\mathbf{R}$-valued. Let $\hat{\mu}_\theta$ and $\hat{U}_{\theta,i}$ be $\hat{\mu}$ and $\hat{U}_i$ except that $\hat{\theta}$ is replaced by θ, and let μ_θ and $U_{\theta,i}$ be μ and $U_{0,i}$ except that θ_0 is replaced by θ, where $U_{0,i} = F_{\theta_0}(X_i^\top \theta_0)$. First, observe that

$$\begin{aligned} \hat{a}(\theta) - a(\theta) &= \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n D_i S_i \cdot \left\{ \varphi(\hat{\mu}_\theta(\hat{U}_{\theta,i})) - \varphi(\mu_\theta(U_{\theta,i})) \right\} \\ &\quad + \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n \left\{ D_i S_i \cdot \varphi(\mu_\theta(U_{\theta,i})) - \mathbf{E}\left[D_i S_i \cdot \varphi(\mu_\theta(U_{\theta,i})) \right] \right\} \\ &\quad + \left\{ \frac{1}{\frac{1}{n}\sum_{i=1}^n D_i} - \frac{1}{P\{D_i = 1\}} \right\} \mathbf{E}\left[D_i S_i \cdot \varphi(\mu_\theta(U_{\theta,i})) \right] \\ &\equiv A_{1n}(\theta) + A_{2n}(\theta) + A_{3n}(\theta), \text{ say.} \end{aligned}$$

We write $A_{1n}(\theta)$ as

$$\begin{aligned} & \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n D_i S_i \cdot \varphi'(\mu_\theta(U_{\theta,i})) \left\{ \hat{\mu}_\theta(\hat{U}_{\theta,i}) - \mu_\theta(U_{\theta,i}) \right\} \\ & + \frac{1}{2\sum_{i=1}^n D_i} \sum_{i=1}^n D_i S_i \cdot \varphi''(\mu_\theta^*(U_{\theta,i})) \left\{ \hat{\mu}_\theta(\hat{U}_{\theta,i}) - \mu_\theta(U_{\theta,i}) \right\}^2 \\ \equiv\ & B_{1n}(\theta) + B_{2n}(\theta), \text{ say,} \end{aligned}$$

where $\mu_\theta^*(U_{\theta,i})$ lies on the line segment between $\hat{\mu}_\theta(\hat{U}_{\theta,i})$ and $\mu_\theta(U_{\theta,i})$. Let $1_{n,i} = 1\{|U_{0,i} - 1| >$

$h/2\}$, and write $B_{2n}(\theta)$ as

$$\begin{aligned}
&\frac{1}{2\sum_{i=1}^n D_i}\sum_{i=1}^n D_i S_i \cdot \varphi''(\mu_\theta^*(U_{\theta,i}))\left\{\hat{\mu}_\theta(\hat{U}_{\theta,i})-\mu_\theta(U_{\theta,i})\right\}^2 1_{n,i} \\
&+\frac{1}{2\sum_{i=1}^n D_i}\sum_{i=1}^n D_i S_i \cdot \varphi''(\mu_\theta^*(U_{\theta,i}))\left\{\hat{\mu}_\theta(\hat{U}_{\theta,i})-\mu_\theta(U_{\theta,i})\right\}^2 \{1-1_{n,i}\} \\
\equiv\ & C_{1n}(\theta)+C_{2n}(\theta), \text{ say.}
\end{aligned}$$

For small $\varepsilon > 0$ and a positive sequence $c_n > 0$ such that $0 < c_n n^{1/4} \to 0$ (see, e.g., the proof of Lemma A3 of Song (2009)),

$$\begin{aligned}
\max_{1\leq i\leq n}\sup_{\theta\in B(\theta_0;\varepsilon)}\left|\hat{U}_{\theta,i}-U_{\theta,i}\right| &= O_P(1/\sqrt{n}) \text{ and} \\
\max_{1\leq i\leq n}\sup_{\theta\in B(\theta_0;c_n)}\left|U_{\theta,i}-U_{0,i}\right| &= O_P(c_n).
\end{aligned} \tag{15}$$

By Assumption C3(ii), $c_n h^{-1} \to 0$. As for $C_{1n}(\theta)$, we bound $\left|\hat{\mu}_\theta(\hat{U}_{\theta,i})-\mu_\theta(U_{\theta,i})\right| 1_{n,i}$ by (using (15))

$$\sup_{u\in[h/2,1-h/2]}|\hat{\mu}_\theta(u)-\mu_\theta(u)|+\left|\mu_\theta(\hat{U}_{\theta,i})-\mu_\theta(U_{\theta,i})\right| 1_{n,i}$$

from some large n on. The first term is $O(n^{-1/2}h^{-1}\sqrt{\log n})$ (e.g. see Lemma A4 of Song (2009)), and the second term is $O_P(1/\sqrt{n})$, both uniformly over $1 \leq i \leq n$ and over $\theta \in B(\theta_0; c_n)$. The latter rate $O_P(1/\sqrt{n})$ stems from (15), and that $\mu_\theta(\cdot)$ is continuously differentiable with a bounded derivative (see Assumption G2(ii)). Since $\varphi''(\cdot)$ is continuous and bounded on the support of $\mathbf{E}[W|X, D = 1]$ (Assumptions C1(i) and G1(ii)(a)), $\sup_{\theta\in B(\theta_0;\varepsilon)}|C_{1n}(\theta)| = O_P(n^{-1}h^{-2}(\log n))$. As for $C_{2n}(\theta)$, we bound $\sup_{\theta\in B(\theta_0;c_n)}|\hat{\mu}_\theta(\hat{U}_{\theta,i}) - \mu_\theta(U_{\theta,i})|(1-1_{n,i}) = O_P(h + n^{-1/2})$, uniformly over $1 \leq i \leq n$. The rate $O_P(h)$ here is the convergence rate at the boundary (e.g. see Lemma A4 of Song (2009)). Therefore, for some

$C > 0$,

$$\begin{aligned}\sup_{\theta \in B(\theta_0;c_n)} |C_{2n}(\theta)| &\leq C\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_\theta(\hat{U}_{\theta,i}) - \mu_\theta(U_{\theta,i})\right\}^2 |1 - 1_{n,i}| \\ &\leq C \max_{1\leq j\leq n} \sup_{\theta\in B(\theta_0;c_n)} \left\{\hat{\mu}_\theta(\hat{U}_{\theta,j}) - \mu_\theta(U_{\theta,j})\right\}^2 |1 - 1_{n,j}| \\ &\quad \cdot\frac{1}{n}\sum_{i=1}^{n}|1 - 1_{n,i}| \\ &= O_P(h^2 + n^{-1}) \cdot O_P(h) = o_P(n^{-1/2}).\end{aligned}$$

We conclude $\sup_{\theta\in B(\theta_0;c_n)} |A_{1n}(\theta) - B_{1n}(\theta)| = o_P\left(n^{-1/2}\right)$. As for $B_{1n}(\theta)$, we apply Lemma B3 below to deduce that with $P_1 \equiv P\{D_i = 1\}$,

$$\frac{1}{P_1}\cdot\frac{1}{n}\sum_{i=1}^{n} D_i\mathbf{E}\left[S_i \cdot \varphi'(\mu_\theta(U_{\theta,i}))|U_{\theta,i}, D_i = 1\right]\{\varphi(W_i) - \mu_\theta(U_{\theta,i})\} + o_P(1/\sqrt{n}),$$

uniformly over $\theta \in B(\theta_0; c_n)$. (With the help of Lemma B1 below, one can check that Assumptions B1-B3 in Lemma B3 are satisfied by the conditions of the theorem here.)

We turn to $A_{3n}(\theta)$, which we write as

$$\mathbf{E}\left[S_i \cdot \varphi(\mu_\theta(U_{\theta,i}))|D_i = 1\right] \cdot \frac{1}{P_1 n}\sum_{i=1}^{n}\{P\{D = 1\} - D_i\} + o_P(1/\sqrt{n})$$

uniformly over $\theta \in B(\theta_0; c_n)$. Combining the results so far, we find that $\sqrt{n}\{\hat{a}(\theta) - a(\theta)\}$ is equal to

$$\begin{aligned}&\frac{1}{P_1\sqrt{n}}\sum_{i=1}^{n} D_i\mathbf{E}\left[S_i \cdot \varphi'(\mu_\theta(U_{\theta,i}))|U_{\theta,i}, D_i = 1\right]\{\varphi(W_i) - \mu_\theta(U_{\theta,i})\} \\ &+\frac{1}{P_1\sqrt{n}}\sum_{i=1}^{n}\{D_iS_i \cdot \varphi(\mu_\theta(U_{\theta,i})) - \mathbf{E}\left[D_iS_i \cdot \varphi(\mu_\theta(U_{\theta,i}))\right]\} \\ &+\mathbf{E}\left[S_i \cdot \varphi(\mu_\theta(U_{\theta,i}))|D_i = 1\right] \cdot \frac{1}{P_1\sqrt{n}}\sum_{i=1}^{n}\{P\left\{D = 1\right\} - D_i\} + o_P(1),\end{aligned}$$

uniformly over $\theta \in B(\theta_0; c_n)$. From this uniform linear representation of $\sqrt{n}\{\hat{a}(\theta) - a(\theta)\}$, it is not hard to show that

$$\begin{aligned}\sup_{\theta \in B(\theta_0;c_n)} |\sqrt{n}\{\hat{a}(\theta) - a(\theta)\}| &= O_P(1) \text{ and} \\ \sup_{\theta \in B(\theta_0;c_n)} \left|\sqrt{n}\{\hat{a}(\theta) - a(\theta) - (\hat{a}(\theta_0) - a(\theta_0))\}\right| &= o_P(1).\end{aligned} \tag{16}$$

(For this, we can use Lemma B1 below to obtain an entropy bound for the class of functions indexing the processes in the linear representation of $\sqrt{n}\{\hat{a}(\theta) - a(\theta)\}$. Details are omitted.)

Let $H_1(a,b) = \partial H(a,b)/\partial a$, and let the sequence c_n chosen above be such that $||\hat{\theta} - \theta_0|| = O_P(c_n)$. We write

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \tilde{\beta}) &= \sqrt{n}\{H(\hat{a}(\hat{\theta}), \hat{b}) - H(\hat{a}(\theta_0), \hat{b})\} \\ &= H_1(a(\theta_0), b_0)^\top \sqrt{n}\{\hat{a}(\hat{\theta}) - \hat{a}(\theta_0)\} + o_P(1) \\ &= H_1(a(\theta_0), b_0)^\top \sqrt{n}\{\hat{a}(\hat{\theta}) - a(\hat{\theta}) - \hat{a}(\theta_0) + a(\theta_0)\} \\ &\quad + H_1(a(\theta_0), b_0)^\top \sqrt{n}\{a(\hat{\theta}) - a(\theta_0)\} + o_P(1) \equiv D_{1n} + D_{2n}, \text{ say.}\end{aligned}$$

The second equality uses the first statement of (16) and continuity of $H_1(\cdot,\cdot)$ (Assumption G1(ii)(b)). By the second statement of (16), $D_{1n} = o_P(1)$. As for D_{2n}, we apply Lemma A1 to deduce that

$$|H_1(a(\theta_0), b_0)^\top \sqrt{n}\{a(\hat{\theta}) - a(\theta_0)\}| = O_P(n^{1/2} \times c_n^2) = o_P(1).$$

Thus we obtain the desired result. ■

7.2 Bahadur Representation of Sample Linear Functionals of SNN Estimators

In this section, we present a Bahadur representation of sample linear functionals of SNN estimators that is uniform over function spaces. (The proofs are found in the supplemental note to this paper that is available from the author's website.) In a different context, Stute and Zhu (2005) obtained a related result that is not uniform.

Suppose that we are given a random sample $\{(S_i, W_i, X_i)\}_{i=1}^n$ drawn from the distribution of a random vector $(S, W, X) \in \mathbf{R}^{d_S+1+d_X}$. Let $\mathcal{S}_S, \mathcal{S}_X$ and $\mathcal{S}_W$ be the supports of S, X, and W respectively. Let Λ be a class of $\mathbf{R}$-valued functions on $\mathbf{R}^{d_X}$ with a generic element denoted by λ. We also let Φ and Ψ be classes of real functions on $\mathbf{R}$ and $\mathbf{R}^{d_S}$ with generic elements φ and ψ and let $\tilde{\varphi}$ and $\tilde{\psi}$ be their envelopes. Let $L_p(P),\ p \geq 1$, be the space of L_p-bounded functions: $||f||_p \equiv \{\int |f(x)|^p P(dx)\}^{1/p} < \infty$, and for a space of functions $\mathcal{F} \subset L_p(P)$ for $p \geq 1$, let $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_p)$ denote the bracketing number of $\mathcal{F}$ with respect to the norm $||\cdot||_p$, i.e., the smallest number r such that there exist $f_1, \cdot\cdot\cdot, f_r$ and $\Delta_1, \cdot\cdot\cdot, \Delta_r \in L_p(P)$ such that $||\Delta_i||_p < \varepsilon$ and for all $f \in \mathcal{F}$, there exists $1 \leq i \leq r$ with $||f_i - f||_p < \Delta_i/2$. Similarly, we define $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_\infty)$ to be the bracketing number of $\mathcal{F}$ with respect to the sup norm $||\cdot||_\infty$, where for any real map f on $\mathbf{R}^{d_X}$, we define $||f||_\infty = \sup_{z\in\mathbf{R}^{d_X}} |f(z)|$. For any norm $||\cdot||$ which is equal to $||\cdot||_p$ or $||\cdot||_\infty$, we define $N(\varepsilon, \mathcal{F}, ||\cdot||)$ to be the covering number of $\mathcal{F}$, i.e., the smallest number of ε-balls that cover $\mathcal{F}$. Letting $F_\lambda(\cdot)$ be the CDF of $\lambda(X)$, we denote $U_\lambda \equiv F_\lambda(\lambda(X))$. Define $g_{\varphi,\lambda}(u) \equiv \mathbf{E}[\varphi(W)|U_\lambda = u]$ and $g_{\psi,\lambda}(u) \equiv \mathbf{E}[\psi(S)|U_\lambda = u]$.

Let $U_{n,\lambda,i} \equiv \frac{1}{n-1}\sum_{j=1,j\neq i}^n 1\{\lambda(X_j) \leq \lambda(X_i)\}$ and consider the estimator:

$$\hat{g}_{\varphi,\lambda,i}(u) \equiv \frac{1}{(n-1)\hat{f}_{\lambda,i}(u)} \sum_{j=1,j\neq i}^n \varphi(W_j) K_h\left(U_{n,\lambda,j} - u\right),$$

where $\hat{f}_{\lambda,i}(u) \equiv (n-1)^{-1}\sum_{j=1,j\neq i}^n K_h(U_{n,\lambda,j} - u)$. The semiparametric process of focus takes the following form:

$$\nu_n(\lambda, \varphi, \psi) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(S_i)\left\{\hat{g}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_\varphi(U_{\lambda,i})\right\},$$

with $(\lambda, \varphi, \psi) \in \Lambda \times \Phi \times \Psi$.

The main focus in this section is on establishing an asymptotic linear representation of

$\nu_n(\lambda, \varphi, \psi)$. The critical element in the proof is to bound the size of the class of conditional expectation functions $\mathcal{G} \equiv \{g_{\varphi,\lambda}(\cdot) : (\varphi, \lambda) \in \Phi \times \Lambda\}$. We begin with the following lemma that establishes the bracketing entropy bound for $\mathcal{G}$ with respect to $||\cdot||_q$, $q \geq 1$.

Lemma B1 : *Suppose that the density f_λ of $\lambda(X)$ is bounded uniformly over $\lambda \in \Lambda$. Furthermore, assume that there exists an envelope $\tilde{\varphi}$ for Φ such that $G_\Phi \equiv \sup_{x \in \mathcal{S}_X} \mathbf{E}[\tilde{\varphi}(W)|X = x] < \infty$, and that for some $C_L > 0$,*

$$\sup_{\varphi \in \Phi} \sup_{\lambda \in \Lambda} |g_{\varphi,\lambda}(u_1) - g_{\varphi,\lambda}(u_2)| \leq C_L|u_1 - u_2|, \textit{ for all } u_1, u_2 \in [0,1].$$

Then for all $\varepsilon > 0$, $q \geq 1$, and $p \geq 1$,

$$N_{[]}(C_\Phi \varepsilon^{1/(q+1)}, \mathcal{G}, ||\cdot||_q) \leq N_{[]}(\varepsilon, \Phi, ||\cdot||_p) \cdot N_{[]}(\varepsilon, \Lambda, ||\cdot||_\infty),$$

where $C_\Phi \equiv 1 + 8C_\Lambda G_\Phi + C_L + G_\Phi/2$ and $C_\Lambda \equiv \sup_{\lambda \in \Lambda} \sup_{v \in \mathbf{R}} f_\lambda(v)$.

We are prepared to present the uniform Bahadur representation of $\nu_n(\lambda, \varphi, \psi)$. Let $\Lambda_n \equiv \{\lambda \in \Lambda : ||\lambda - \lambda_0||_\infty \leq c_n\}$, where $0 < c_n n^{1/4} \to 0$. We let $X = [X_1^\top, X_2^\top]^\top$, where X_1 is a continuous random vector and X_2 is a discrete random vector taking values in a finite set $\{x_1, \cdot\cdot\cdot, x_M\}$. We make the following assumptions.

Assumption B1 : (i) For some $C > 0$, $p \geq q > 4$, $b_\Psi \in (0, q/(q-1))$, and $b_\Phi \in (0, q/\{(q+1)(q-1)\})$,

$$\log N_{[]}(\varepsilon, \Phi, ||\cdot||_p) < C\varepsilon^{-b_\Phi} \text{ and } \log N_{[]}(\varepsilon, \Psi, ||\cdot||_p) < C\varepsilon^{-b_\Psi}, \text{ for each } \varepsilon > 0,$$

and $\mathbf{E}[\tilde{\varphi}(W)^p] + \mathbf{E}[\tilde{\psi}(S)^p] + \sup_{x \in \mathcal{S}_X} \mathbf{E}[\tilde{\varphi}(W)|X = x] + \sup_{x \in \mathcal{S}_X} \mathbf{E}[\tilde{\psi}(W)|X = x] < \infty$.
(ii) (a) For $q > 4$ in (i) and for some $b_\Lambda \in (0, q/\{(q+1)(q-1)\})$ and $C > 0$, $\log N_{[]}(\varepsilon, \Lambda, ||\cdot||_\infty) \leq C\varepsilon^{-b_\Lambda}$, for each $\varepsilon > 0$.
(b) For all $\lambda \in \Lambda$, the density $f_\lambda(\cdot)$ of $\lambda(X)$ is bounded uniformly over $\lambda \in \Lambda$ and bounded

away from zero on the interior of its support uniformly over $\lambda \in \Lambda$.

Assumption B2 : (i) $K(\cdot)$ is symmetric, nonnegative, compact supported, twice continuously differentiable with bounded derivatives, and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^3 + n^{-1/2}h^{-2}(-\log h) \to 0$ as $n \to \infty$.

Assumption B3 : $\mathbf{E}[\varphi(W)|U_\lambda = \cdot]$ is twice continuously differentiable with derivatives bounded uniformly over $(\lambda, \varphi) \in B(\lambda_0; \varepsilon) \times \Phi$ with some $\varepsilon > 0$.

The following lemma offers a uniform representation of ν_n.

Lemma B2 : *Suppose that* Assumptions B1-B4 *hold. Then,*

$$\sup_{(\lambda,\varphi,\psi)\in\Lambda_n\times\Phi\times\Psi}\left|\nu_n(\lambda,\varphi,\psi) - \frac{1}{\sqrt{n}}\sum_{i=1}^n g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(W_i) - g_{\varphi,\lambda}(U_{\lambda,i})\}\right| = o_P(1).$$

Let $D_i \in \{0,1\}$ be a binary random variable and for $d \in \{0,1\}$, define $g_{\varphi,\lambda,d}(u) \equiv \mathbf{E}[\varphi(W_i)|U_{\lambda,i} = u, D_i = d]$ and $g_{\psi,\lambda,d}(u) \equiv \mathbf{E}[\psi(S_i)|U_{\lambda,i} = u, D_i = d]$. Consider the estimator:

$$\hat{g}_{\varphi,\lambda,d}(U_{n,\lambda,i}) \equiv \frac{1}{(n-1)\hat{f}_{\lambda,d}(U_{n,\lambda,i})}\sum_{j=1,j\neq i}^n \varphi(W_j)1\{D_j = d\}K_h\left(U_{n,\lambda,j} - U_{n,\lambda,i}\right),$$

where $\hat{f}_{\lambda,d}(U_{n,\lambda,i}) \equiv (n-1)^{-1}\sum_{j=1,j\neq i}^n 1\{D_j = d\}K_h(U_{n,\lambda,j} - U_{n,\lambda,i})$. Similarly as before, we define

$$\nu_{n,d}(\lambda,\varphi,\psi) \equiv \frac{\sqrt{n}}{\sum_{i=1}^n D_i}\sum_{i=1}^n \psi(S_i)D_i\left\{\hat{g}_{\varphi,\lambda,d}(U_{n,\lambda,i}) - g_{\varphi,\lambda,d}(U_{\lambda,i})\right\},$$

with $(\lambda,\varphi,\psi) \in \Lambda \times \Phi \times \Psi$. The following lemma presents variants of Lemma B2.

Lemma B3 **:** *Suppose that* Assumptions B1-B3 *hold, and let* $P_d \equiv P\{D = d\}$, *and* $\varepsilon_{\varphi,\lambda,d,i} \equiv \varphi(W_i) - g_{\varphi,\lambda,d}(U_{\lambda,i})$, $d \in \{0,1\}$.
(i) *If there exists* $\varepsilon > 0$ *such that* $P\{D_i = 1|U_{\lambda,i} = u\} \geq \varepsilon$ *for all* $(u,\lambda) \in [0,1] \times \Lambda$, *then*

$$\sup_{(\lambda,\varphi,\psi)\in\Lambda_n\times\Phi\times\Psi}\left|\nu_{n,1}(\lambda,\varphi,\psi) - \frac{1}{\sqrt{n}P_1}\sum_{i=1}^n D_i g_{\psi,\lambda,1}(U_{\lambda,i})\varepsilon_{\varphi,\lambda,1,i}\right| = o_P(1).$$

(ii) *If there exists* $\varepsilon > 0$ *such that* $P\{D_i = 1 | U_{\lambda,i} = u\} \in [\varepsilon, 1-\varepsilon]$ *for all* $(u,\lambda) \in [0,1] \times \Lambda$, *then*

$$\sup_{(\lambda,\varphi,\psi)\in\Lambda_n\times\Phi\times\Psi} \left| \nu_{n,0}(\lambda,\varphi,\psi) - \frac{1}{\sqrt{n}P_1}\sum_{i=1}^{n} \frac{(1-D_i)P(U_{\lambda,i})g_{\psi,\lambda,1}(U_{\lambda,i})}{1-P(U_{\lambda,i})} \varepsilon_{\varphi,\lambda,0,i} \right| = o_P(1).$$

References

[1] Abrevaya, J. and J. Huang, 2005. On the bootstrap of the maximum score estimator. Econometrica 73, 1175-2204.

[2] Chen, S. and S. Khan, 2003. Semiparametric estimation of a heteroskedastic sample selection model. Econometric Theory 19, 1040-1064.

[3] Cosslett, S., 1990. Semiparametric estimation of a regression model with sample selectivity. In W. A. Barnett et. al., eds., Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics. Cambridge: Cambridge University Press, 1990.

[4] Das, M., W. K. Newey, and F. Vella, 2003. Nonparametric estimation of sample selection models. Review of Economic Studies 70, 33-58.

[5] Dehejia, R. and S. Wahba, 1998. Propensity score matching methods for nonexperimental causal studies. NBER Working Paper No. 6829.

[6] Fan, Y. and Q. Li, 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865-890.

[7] Fix, E. and J. L. Hodges, 1951. Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4.

[8] Gallant, R. and D. Nychka, 1987. Semi-nonparametric maximum likelihood estimation, Econometrica 55, 363-390.

[9] Hahn, J. and G. Ridder, 2010. The asymptotic variance of semi-parametric estimators with generated regressors. Working paper.

[10] Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315-331.

[11] Härdle, W., P. Hall and H. Ichimura, 1993. Optimal semiparametric estimation in single index models. Annals of Statistics 21, 1, 157-178.

[12] Härdle, W., P. and A. B. Tsybacov, 1993. How sensitive are average derivatives. Journal of Econometrics 58, 31-48.

[13] Heckman, J. J., 1974. Shadow prices, market wages and labor supply," Econometrica 42, 679-694.

[14] Heckman, J. J., H. Ichimura, J. Smith, H., and P. Todd, 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017-1098.

[15] Heckman, J. J., H. Ichimura, and P. Todd, 1997. Matching as an econometric evaluation estimator : evidence from evaluating a job training programme. Review of Economic Studies, 64, 605-654.

[16] Heckman, J. J., H. Ichimura, and P. Todd, 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261-294.

[17] Hirano, K., G. Imbens, and G. Ridder, 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161-1189.

[18] Horowitz, J. L. and W. Härdle, 1996. Direct semiparametric estimation of single-index models with discrete covariates, Journal of the American Statistical Association 91, 1632-1640.

[19] Hristache, M., A. Juditsky and V. Spokoiny, 2001. Direct estimation of the index coefficient in a single-index model. Annals of Statistics 29, 595-623.

[20] Ichimura, H, 1993. Semiparametric least squares, SLS and weighted SLS estimation of single Index Models. Journal of Econometrics 58, 71-120.

[21] Kim, J. and D. Pollard, 1990. Cube root asymptotics. Annals of Statistics 18, 191-219.

[22] Klein, R. W. and R. H. Spady, 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 2, 387-421.

[23] Li, Q. and J. M. Wooldrige, 2002. Semiparametric estimation of partial linear models for dependent data with generated regressors. Econometric Theory 18, 625-645.

[24] Mammen, E., C. Rothe, and M. Schienle, 2012. Nonparametric regression with nonparametrically generated covariates. Annals of Statistics 40, 1132-1170.

[25] Manski, C. F., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3, 205-228.

[26] Newey, W. K. and D. McFadden, 1994. Large sample estimation and hypothesis testing. Handbook of Econometrics, Vol 4, ed. R. F. Engle and D. McFadden, 2111-2245.

[27] Newey, W. K., Powell, J. and F. Vella, 1999. Nonparametric estimation of triangular simultaneous equation models. Econometrica 67, 565-603.

[28] Newey, W. K., Powell, J. and J. Walker, 1990. Semiparametric estimation of selection models: some empirical results. American Economic Review 80, 324-8.

[29] Pollard, D., 1989. A maximal inequality for sums of independent processes under a bracketing condition. Unpublished manuscript.

[30] Powell, J., 1994. Estimation of semiparametric models. In The Handbook of Econometrics, Vol. IV, ed. by R. F. Engle and D. L. McFadden, Amsterdam: North-Holland.

[31] Powell, J., Stock, J. and T. Stoker, 1989. Semiparametric estimation of index coefficients. Econometrica 57, 6, 1403-1430.

[32] Rilstone, P., 1996. Nonparametric estimation of models with generated regressors. International Economic Review 37, 299-313.

[33] Robinson, P., 1987. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. Econometrica 55, 875-891.

[34] Robinson, P., 1988. Root-N consistent nonparametric regression. Econometrica 56, 931-954.

[35] Song, K., 2009. Testing conditional independence using Rosenblatt transforms. Annals of Statistics 37, 4011-4045.

[36] Song, K., 2012. On the smoothness of conditional expectation functionals. Statistics and Probability Letters 82, 1028-1034.

[37] Sperlich, S., 2009. A note on non-parametric estimation with predicted values. Econometrics Journal 12, 382-395.

[38] Stoker, T., 1986. Consistent estimation of scaled coefficients. Econometrica 54, 1461-1481.

[39] Stone, C. J., 1977. Consistent nonparametric regression. Annals of Statistics 5, 595-645.

[40] Stute, W., 1984. Asymptotic normality of nearest neighbor regression function estimates. Annals of Statistics 12, 917-926.

[41] Stute, W. and L. Zhu, 2005. Nonparametric checks for single-index models. Annals of Statistics 33, 1048-1083.

[42] Turki-Moalla, K., 1998. Rates of convergence and law of the iterated logarithm for U-processes. Journal of Theoretical Probability 11, 869-906.

[43] van der Vaart, A. W., 1996. New Donsker classes. Annals of Probability 24, 2128-2140.

[44] Yang, S., 1981. Linear functionals of concomitants of order statistics with application to nonparametric estimation of regression function. Journal of the American Statistical Association 76, 658-662.