SEMIPARAMETRIC MODELS WITH SINGLE-INDEX NUISANCE PARAMETERS

KYUNGCHUL SONG

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.

KEY WORDS. Sample selection model; conditional median restrictions; matching estimators; maximum score estimation; cube-root asymptotics; generated regressors

JEL Classfication. C12, C14, C51.
AMS Classification. 62G07, 62G09, 62G10

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.

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

Date: July 22, 2013.

1For 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).
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 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.

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.

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.

---

2For 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.
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 or semiparametric 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 $W \equiv (W_1, \ldots, W_L)^T \in \mathbb{R}^L$, $S$ is a $d_S \times d_\varphi$ random matrix, and $X \in \mathbb{R}^d$ is a random vector, where all three random quantities $W$, $S$, and $X$, are assumed to be observable. We let $X = [X_1^T, X_2^T]^T \in \mathbb{R}^{d_1 + d_2}$, where $X_1$ is a continuous random vector and $X_2$ is a discrete random vector taking values from $\{x_1, \ldots, x_M\}$. Let $\Theta \subset \mathbb{R}^d$ be the space of a nuisance parameter $\theta_0$ that is known to be identified. Denote $U_{\theta} \equiv F_\theta(X^T \theta)$, where $F_\theta$ is the CDF of $X^T \theta$. We assume that $X^T \theta$ is a continuous random variable for all $\theta$ in a neighborhood of $\theta_0$. Given an observed binary variable $D \in \{0, 1\}$, we define

\begin{align}
\mu_\theta(U_{\theta}) & \equiv \mathbb{E}[W|U_{\theta}, D = 1], \\
\text{and when } \theta = \theta_0, \text{ we simply write } \mu_0(U_0), \text{ where } U_0 \equiv F_{\theta_0}(X^T \theta_0). \text{ 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, \ldots, M, \text{ let } S_m \text{ be the support of } X_1 \{X_2 = x_m, D = 1\}, \text{ and } S_W \text{ be the support of } W, \text{ and let } \varphi : S_W \to \mathbb{R}^{d_\varphi} \text{ be a known map that is twice continuously differentiable with bounded derivatives on the interior of the}.
\end{align}
support of $\mathbb{E}[W|X, D = 1]$. Then we define a map $a : \Theta \rightarrow \mathbb{R}^d$ by
\begin{equation}
(2.2) \quad a(\theta) \equiv \mathbb{E} [S \cdot \varphi(\mu_\theta(U_\theta))|D = 1], \ \theta \in \Theta.
\end{equation}

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 $\beta_0$, is identified as follows:
\begin{equation}
(2.3) \quad \beta_0 = H(a(\theta_0), b_0),
\end{equation}

where $H : \mathbb{R}^d \times \mathbb{R}^{d_0} \rightarrow \mathbb{R}^d$ is a map that is fully known, continuously differentiable in the first argument, and $b_0$ is a $d_0$ dimensional parameter that does not depend on $\theta_0$ and is consistently estimable. We will see examples of $\beta_0$ shortly.

Throughout this paper, we assume that there is an estimator $\hat{\theta}$ for $\theta_0$ which is either $\sqrt{n}$-consistent or $\sqrt[3]{n}$-consistent. A natural estimator of $\beta_0$ is obtained by
\begin{equation}
\hat{\beta} \equiv H(\hat{a}(\hat{\theta}), \hat{b}),
\end{equation}

where $\hat{a}(\theta)$ is an estimator of $a(\theta)$ and $\hat{b}$ is a consistent estimator of $b_0$. The estimator $\hat{\beta}$ can be obtained by using nonparametric estimation of conditional expectation $\mathbb{E}[W|U_\theta, D = 1]$. For future reference, we denote
\begin{equation}
\tilde{\beta} \equiv H(\hat{a}(\theta_0), \hat{b}),
\end{equation}

an infeasible estimator using $\theta_0$ 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
\begin{equation}
(2.4) \quad \sqrt{n}(\tilde{\beta} - \beta_0) \overset{d}{\rightarrow} N(0, V),
\end{equation}

it follows that
\begin{equation}
\sqrt{n}(\hat{\beta} - \beta_0) \overset{d}{\rightarrow} N(0, V).
\end{equation}

This result is very convenient, because the computation of the asymptotic variance matrix $V$ in $[2.4]$ 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 $Z \in \mathbb{R}^d$, a vector of covariates, as follows:
\begin{equation}
Y^* = Z^\top \beta_0 + v,
\end{equation}

The conditions for the identification of $\theta_0$ in many examples of semiparametric models is already established. See Horowitz (2009). The identification of $\theta_0$ in this paper’s context arises often in binary choice models. See Chapter 4 of Horowitz (2009) for the identification analysis for the binary choice models.
where \( v \) denotes an unobserved factor that affects the outcome. The econometrician observes \( Y^\ast \) only when a selection indicator \( D \in \{0, 1\} \) assumes number one, so that as for observed outcome \( Y \), we write

\[
Y = Y^\ast \cdot D.
\]

This paper specifies \( D \) as follows:

(2.5)

\[
D = 1\{X^\top \theta_0 > \varepsilon\},
\]

where \( \varepsilon \) is an unobserved component, and \( \theta_0 \in \mathbb{R}^d \) 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:

(2.6)

\( \varepsilon \) is independent of \( X \).

Condition (2.6) is mainly used to identify \( \theta_0 \) through a single-index restriction \( \mathbb{E}[D|X] = \mathbb{E}[D|X^\top \theta_0] \) or a parametric restriction \( \mathbb{E}[D|X] = F(X^\top \theta_0) \) for some known CDF \( F \).

Define for \( \theta \) in a neighborhood of \( \theta_0 \),

\[
S_{ZZ}(\theta) \equiv \mathbb{E}\left[ZZ^\top|D = 1\right] - \mathbb{E}\left[Z \cdot \mathbb{E}[Z^\top|U_0, D = 1]|D = 1\right]
\]

and

\[
S_{ZY}(\theta) \equiv \mathbb{E}[ZY|D = 1] - \mathbb{E}\left[Z \cdot \mathbb{E}[Y|U_0, D = 1]|D = 1\right].
\]

We consider the following assumptions.

**Assumption SSO:**

(i) \( (\varepsilon, v) \) 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 \( \varepsilon \) given \( X \).

(iii) The smallest eigenvalue of \( S_{ZZ}(\theta_0) \) is bounded away from zero uniformly over \( \theta \) in a neighborhood of \( \theta_0 \).

Assumption SSO(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 SSO(ii) which substantially relaxes Condition (2.6). The relaxation allows the individual components of \( X \) and \( \varepsilon \) to be stochastically dependent. Assumption SSO(iii) is slightly stronger than the usual condition that \( S_{ZZ}(\theta_0) \) is invertible.

Under Assumptions SSO(i) and (iii), we can write the equation for observed outcomes as a partial linear regression model, and follow Robinson (1988) to identify \( \beta_0 \) as

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

once \( \theta_0 \) is identified. To see that this \( \beta_0 \) is a special case of (2.3), let \( 1_2 \) be a \( 2 \times 1 \) vector of ones, \( S = 1_2 \otimes Z \), (the notation \( \otimes \) represents the Kronecker product of matrices) and

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.
define \( W = [Y; Z^\top] \), \( \mu(U_\theta) = \mathbb{E}[W|U_\theta, D = 1] \),

\[
\begin{align*}
a(\theta) &= \mathbb{E}[S \cdot \varphi(\mu(U_\theta))|D = 1], \quad \text{and} \\
b_0 &= \mathbb{E}[S \cdot W|D = 1],
\end{align*}
\]

where \( \varphi : \mathbb{R}^{d_Z+1} \rightarrow \mathbb{R}^{d_Z+1} \) 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 \( \theta \). 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 \( \beta_0 \) is a special case of (2.3).

### 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 \( \sigma \)-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. The econometrician observes \( Z \) and \( Y = Y_1Z + Y_0(1 - Z) \). The parameter of interest is \( \beta_0 = \mathbb{E}[Y_1 - Y_0|Z = 1] \), i.e., the treatment effect on the treated. We assume that \( X \) is a vector of observed covariates in \( \mathbb{R}^d \) and

\[
Z = 1 \{ X^\top \theta_0 \geq \varepsilon \},
\]

The common notation for the treatment status is \( D \), but the treatment status does not play the same role as \( D \) in (2.1). Hence we choose a different notation, \( Z \), here.
where \( \varepsilon \) denotes the unobserved factor that affects the treatment status, and \( \theta_0 \) is an unknown parameter. Define \( U_\theta = F_\theta(X^\top \theta) \), where \( F_\theta \) 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 \( \eta \leq P(U_0) \leq 1 - \eta \).

(iii) \( \text{Med}(\varepsilon|X) = 0 \).

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 \(\varepsilon\) and \(X\) are independent.

Under Assumption SM0, we can identify \( \beta_0 \) as follows:

\[
\beta_0 = \mathbf{E}[Y - \mathbf{E}[Y|U_0, Z = 0]|Z = 1].
\]

It is not immediately seen that \( \beta_0 \) can be written in the form of (2.3), 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 (2.3), rewrite \( \beta_0 \) as

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

Define \( W = [Y(1 - Z), Z]^\top \), and write \( \mu_\theta(U_\theta) = \mathbf{E}[W|U_\theta] \). Let

\[
b_0 = \left[ \frac{\mathbf{E}[YZ]}{P\{Z = 1\}} \right], \quad \text{and} \quad a(\theta) = \mathbf{E}[Z\varphi(\mu_\theta(U_\theta))],
\]

where \( \varphi : \mathbb{R} \times (0, 1) \rightarrow \mathbb{R} \) is defined to be \( \varphi(x, z) = x/(1 - z) \) for \((x, z) \in \mathbb{R} \times (0, 1)\). Note that \( b_0 \) does not depend on \( \theta \). Let \( H : \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R} \) be defined as \( H(a, b) = b_1 - b_2 \cdot a \), where \( a \in \mathbb{R} \) and \( b = [b_1, b_2]^\top \in \mathbb{R}^2 \). Then, we can write \( \beta_0 = H(a(\theta_0), b_0) \), i.e., in the form of (2.3) 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 \( \sqrt{n}\)-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 (2.2) is very smooth in \( \theta \) in a neighborhood of \( \theta_0 \). More

---

Song (2012) considered explicitly the case where \(X\) has both discrete and continuous components. Certainly, the result holds when \(X\) has only continuous components, because the discrete component can be replaced by a constant. However, the result requires that \(X\) have at least one continuous component.
specifically, under regularity conditions, there exist $C > 0$ and $\varepsilon \in (0, 1/2]$ such that for each $\eta \in (0, \varepsilon)$,

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

where $B(\theta_0; \eta)$ denotes the $\eta$-ball around $\theta_0$, 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 $\eta$ is 2 (not 1), which says that the map $a$ is very smooth in a neighborhood of $\theta_0$.

To see how this result serves our purpose, we write

$$\begin{align*}
\|\hat{\beta} - \hat{\beta}\| &= \|H(\hat{a}(\hat{\theta}), \hat{b}) - H(\hat{a}(\theta_0), \hat{b})\| \\
&\leq C\|\hat{a}(\hat{\theta}) - \hat{a}(\theta_0)\| + o_P(\|\hat{a}(\hat{\theta}) - \hat{a}(\theta_0)\|),
\end{align*}$$

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

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

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. For this, standard arguments use empirical process theory. (e.g. Andrews (1994).) To use empirical process theory, the crucial step is to obtain the asymptotic linear representation of $\sqrt{n}\{\hat{a}(\hat{\theta}) - a(\theta)\}$, that is,

$$\sqrt{n}\{\hat{a}(\hat{\theta}) - a(\theta)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(\theta) + o_P(1)$$

for some $\psi_i(\theta)$ such that $E\psi_i(\theta) = 0$ and $\psi_i(\theta)$'s are i.i.d. across $i$'s, where $o_P(1)$ is uniform local around $\theta_0$, and apply the maximal inequality to obtain the stochastic equicontinuity of the empirical process (i.e. the sum on the right hand side.) To see how this works intuitively, note that

$$\sqrt{n}\{\{\hat{a}(\hat{\theta}) - a(\theta)\} - \{\hat{a}(\theta_0) - a(\theta_0)\}\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\psi_i(\theta) - \psi_i(\theta_0)] + o_P(1).$$

Hence the asymptotic variance of the left hand side is equal to

$$E (\psi_i(\theta) - \psi_i(\theta_0))^2.$$ 

As $\theta \to \theta_0$, the variance goes to zero, under minor regularity conditions for $\psi_i(\theta)$. Hence we obtain

$$\{\hat{a}(\hat{\theta}) - a(\theta)\} - \{\hat{a}(\theta_0) - a(\theta_0)\} = o_P(n^{-1/2})$$

as $\theta \to \theta_0$. The actual argument is more complex than this, because what we have here is $\hat{\theta} \to_P \theta_0$, not $\theta \to \theta_0$. For this, standard arguments use empirical process theory to fill this gap. See Andrews (1994) for more details. The crucial step here is the derivation

---

7 The inequality in (2.8) is obtained without assuming differentiability of $\mu_\theta(u) = E[W(U_\theta = u)]$ in $\theta \in \Theta$. The low level conditions for the bound in (2.8) are provided in Lemma A1 in Appendix.

8 See the derivation (3.8) in Andrews (1994) for example.
of the asymptotic linear representation in (2.9). The asymptotic linear representation based on a symmetrized nearest neighborhood estimator that is proved in the Appendix.

As for $B_n$, the result in (2.8) implies that with probability approaching one,

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

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 (2.10) is $o_P(1/\sqrt{n})$. Therefore, we conclude that

$$||\hat{\beta} - \beta_0|| = o_P(1/\sqrt{n}).$$

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.

Let us now consider (2.8). Since $a(\theta)$ depends on $\theta$ only through the $\sigma$-field generated by $X^T\theta$, proving (2.8) is far from obvious. The sketch of the proof in Song (2012) is as follows. First, we consider a discretized version of $U_\theta$ and $U_0$, say, $U_\theta^\Delta$ and $U_0^\Delta$, where $\Delta > 0$ represents the size of steps in the grid points which $U_\theta^\Delta$ and $U_0^\Delta$ take values from. For simplicity, we let $d_\varphi = L = 1$, $\varphi$ be an identity map, and $D = 1$. Then for $\theta$ in a neighborhood of $\theta_0$, $|a(\theta) - a(\theta_0)|$ is bounded by

$$\begin{align*}
|E [S \cdot \{E[W|U_\theta] - E[W|U_\theta^\Delta]\}] | &+ |E [S \cdot \{E[W|U_\theta^\Delta] - E[W|U_0]\}] | \\
+ |E [S \cdot \{E[W|U_\theta^\Delta] - E[W|U_0^\Delta]\}] | \\
= |A_{1n}| + |A_{2n}| + |A_{3n}|, \text{ say.}
\end{align*}$$

Since $\sigma(U_\theta^\Delta) \subset \sigma(U_0)$ (where $\sigma(U_\theta^\Delta)$ and $\sigma(U_\theta)$ denoting the $\sigma$-fields of $U_\theta^\Delta$ and $U_\theta$), we use the law of iterated conditional expectations to write $A_{1n}$ as

$$E [E[S|U_\theta] \cdot \{E[W|U_\theta] - E[W|U_\theta^\Delta]\}]$$

$$= E \left[ (E[S|U_\theta] - E[S|U_\theta^\Delta]) \cdot \{E[W|U_\theta] - E[W|U_\theta^\Delta]\} \right].$$

Therefore, by Cauchy-Schwarz inequality, we find that

$$|A_{1n}| \leq \sqrt{E \left[ (E[S|U_\theta] - E[S|U_\theta^\Delta])^2 \right] \sqrt{E \left[ (E[W|U_\theta] - E[W|U_\theta^\Delta])^2 \right]}}.$$

Note that as $\Delta$ becomes smaller, $E[S|U_\theta]$ and $E[S|U_\theta^\Delta]$ become closer. In fact, under regularity conditions, it can be shown that

$$\sqrt{E \left[ (E[W|U_\theta] - E[W|U_\theta^\Delta])^2 \right]} \leq C\Delta \text{ and}$$

$$\sqrt{E \left[ (E[S|U_\theta] - E[S|U_\theta^\Delta])^2 \right]} \leq C\Delta,$$

for some $C > 0$. (See Lemma 3 of Song (2012).) Hence from (2.11), we conclude that

$$|A_{1n}| \leq C_1\Delta^2,$$

for some $C_1 > 0$. Applying the same arguments to $A_{3n}$, we also obtain that

$$|A_{3n}| \leq C_2\Delta^2,$$

for some $C_2 > 0$. 
Now as for \(A_{2n}\), one can expect that \(\mathbb{E}[W|U^\Delta_{\theta}]\) and \(\mathbb{E}[W|U^\Delta_{0}]\) will become close to each other as \(\theta\) becomes closer to \(\theta_0\). How fast they become closer to each other will also depend on how fine the discretization is. In fact, using Lemma 4 of Song (2012), one can show that for some \(C_3 > 0\),

\[
|A_{2n}| \leq C_3 \Delta ||\theta - \theta_0||,
\]

whenever \(||\theta - \theta_0|| \leq C_4 \Delta\) for some constant \(C_4 > 0\). Hence \(\mathbb{E}[W|U^\Delta_{\theta}]\) and \(\mathbb{E}[W|U^\Delta_{0}]\) become closer to each other faster when \(\Delta\) is smaller. (Note that the bound requires that \(||\theta - \theta_0|| \leq C_4 \Delta\), and hence one cannot make \(\Delta\) smaller without making \(\theta\) closer to \(\theta_0\).) Thus, we collect the results to obtain the following bound:

\[
(2.12) \quad |a(\theta) - a(\theta_0)| \leq C_5 \Delta^2 + C_6 \Delta ||\theta - \theta_0||,
\]

for some constants \(C_5 > 0\) and \(C_6 > 0\). Since one can take the discretization step \(\Delta\) arbitrarily small as long as the condition \(||\theta - \theta_0|| \leq C_4 \Delta\) is satisfied, we take \(\Delta = C_4^{-1}||\theta - \theta_4||\) to obtain the last bound in (2.12) as

\[
C_5 C_4^{-2} ||\theta - \theta_0||^2 + C_6 C_4^{-1} ||\theta - \theta_0||^2 \leq C ||\theta - \theta_0||^2,
\]

where \(C = C_5 C_4^{-2} + C_6 C_4^{-1}\). Thus we obtain the desired result

\[
|a(\theta) - a(\theta_0)| \leq C ||\theta - \theta_0||^2,
\]

for some \(C > 0\), and hence (2.8) \(^9\)

### 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 \(\theta_0\) 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 \(K_h(u) \equiv K(u/h)/h\), \(K : \mathbb{R} \to \mathbb{R}\) 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).

Define

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

\(^9\)The heuristics here uses the assumption that \(\varphi\) is a linear map. When \(\varphi\) is nonlinear yet twice continuously differentiable, we can linearize it to obtain a similar bound. (See Song (2012) for details.)
The estimator \( \hat{a}(\theta) \) 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:

\[
(3.2) \quad \hat{\beta} = H(\hat{a}(\theta), \hat{b}).
\]

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) \( X^T \theta \) 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 \( \{x^T \theta : \theta \in B(\theta_0; \varepsilon), x \in S_m\} \) 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) \( n^{1/2}h^3 + n^{-1/2}h^{-2}(- \log h) \rightarrow 0 \).

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 \( X^T \theta \) at the expense of a more complicated exposition. Assumption C2 allows \( \hat{\theta} \) to be either \( \sqrt{n} \)-consistent or \( \sqrt{m} \)-consistent. Assumption C3 concerns the kernel and the bandwidth. Assumption C3(i) is satisfied, for example, by a quartic kernel: \( K(u) = (15/16)(1 - u^2)^2\{||u|| \leq 1\} \).

**Assumption G1**: (i) For \( p \geq 4 \), \( \sup_{x \in S_X} E[||W||^p|X = x] + \sup_{x \in S_X} E[||S||^p|X = x] < \infty \).
(ii) (a) \( \varphi \) is twice continuously differentiable on the interior of the support of \( E[W|X] \) with derivatives bounded on the support of \( E[W|X] \), 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 \( m = 1, \ldots, M \), both \( E[S|X_1 = \cdot, (X_2, D) = (x_m, 1)] \) and \( E[W|X_1 = \cdot, (X_2, D) = (x_m, 1)] \) are Lipschitz continuous.
(ii) \( E[W|U_0 = \cdot] \) 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 \( \varphi \) 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 \( \sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} N(0, V) \) for some positive definite matrix \( V \). Then \( \sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} N(0, V) \).

**Remarks 1**: The asymptotic covariance matrix \( V \) in Theorem 1 is the same asymptotic covariance matrix that one would have obtained had \( \theta_0 \) 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 \( \sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} N(0, V) \) 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 (2.3). Writing \( \beta_0 \) into the form (2.3) 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 \( \beta_0 \) 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 \( p \geq 4 \), \( \sup_{x \in S_X} E [||Y||^p | X = x] + \sup_{x \in S_X} E [||Z||^p | X = x] < \infty \).
For some positive definite matrix $V$ in the conditional expectations. We also let $\tilde{\beta}$ be $\hat{\beta}$ except that we have a single-index $V$.

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

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

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 $p \geq 4$, $\sup_{x \in S_x} \mathbb{E}[|Y|^p | X = x] < \infty$.

Assumption SM2: (i) For each $m = 1, \cdots, M$, both $P\{Z = 1|X_1 = \cdot, X_2 = x_m\}$ and $\mathbb{E}[Y|X_1 = \cdot, X_2 = x_m]$ are Lipschitz continuous.

(ii) $P\{Z = 1|U_0 = \cdot\}$ and $\mathbb{E}[Y|U_0 = \cdot]$ are twice continuously differentiable with derivatives bounded uniformly over $\theta \in B(\theta_0; \varepsilon)$ with some $\varepsilon > 0$. 

\text{SEMIPARAMETRIC MODELS} 13
To construct an estimator of the average treatment effect on the treated based on the single-index matching, we first define

$$\hat{\mu}_{(1-Z)Y}(\hat{U}_k) = \frac{\sum_{i=1}^{n}(1 - Z_i)Y_iK_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^{n}K_h(\hat{U}_i - \hat{U}_k)}$$

and

$$\hat{P}(\hat{U}_k) = \frac{\sum_{i=1}^{n}Z_iK_h(\hat{U}_i - \hat{U}_k)}{\sum_{i=1}^{n}K_h(\hat{U}_i - \hat{U}_k)}.$$

Then, the sample analogue principle suggests

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

If we define

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

we can rewrite the estimator as

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

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 $\hat{\beta}$ be $\tilde{\theta}$ except that $\hat{\theta}$ is replaced by $\theta_0$. Suppose that Assumptions C1–C3 and SM1-SM2 hold and that $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_{SM})$ 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 $\mu_d(U_0) = E[Y|U_0, Z = d]$ and $P_d = P\{Z = d\}$ for $d \in \{0, 1\}$,

$$V_{SM} = E \left[ (Y - \mu_1(U_0))^2 | Z = 1 \right] / P_1$$

$$+ E \left[ (Y - \mu_0(U_0))^2 P^2(U_0)/(1 - P(U_0))^2 | Z = 0 \right] (1 - P_1) / P_1^2$$

$$+ Var(\mu_1(U_0) - \mu_0(U_0)|Z = 1) / P_1.$$

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

$$Z_i = U_{1i} - (\eta_{1i}/2)1$$

and

$$X_i = U_{2i} - \eta_i/2,$$

where $U_{1i}$ is an i.i.d. random vector in $\mathbb{R}^3$ constituted by independent random variables with uniform distribution on $[0, 1]$, $1$ is a 3 dimensional vector of ones, $U_{2i}$ and $\eta_i$ are random vectors in $\mathbb{R}^k$ 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 $\eta_{1i}$ is the first component of $\eta_i$. Then, the selection mechanism is defined as

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

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

We compare the performances of the two estimators of $\beta_0$, $\hat{\beta}(\hat{\theta})$ ("Plug-in $\hat{\theta}$") and $\hat{\beta}(\theta_0)$ ("Plug-in $\theta_0$") in terms of mean absolute deviation (MAE) and mean squared error (MSE). Bandwidths for the estimation of $E[Y_i|X_i^\top \theta_0, D_i = 1]$ and $E[Z_i|X_i^\top \theta_0, D_i = 1]$ 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 $\beta$ 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 $\beta_0$. It shows that the performance of the estimators remains similar regardless of whether $\theta_0$ 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).
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 $\beta_0$. 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 1: The Performance of the Estimators in Terms of MAE and RMSE (Specification A)

| Specification | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ w/o BC | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ w/o BC |
|---------------|--------------------|-------------------------------|--------------------|-------------------------------|
|               | $k = 3$            | $k = 6$                        | $k = 3$            | $k = 6$                        |
| $n = 300$     | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. AN      | 0.6911             | 2.2533                        | 0.6458             | 1.9634                        |
| Spec. AT      | 0.7345             | 2.5437                        | 0.6913             | 2.2408                        |
| $n = 500$     | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. AN      | 0.5327             | 1.3428                        | 0.4965             | 1.1615                        |
| Spec. AT      | 0.5658             | 1.5154                        | 0.5308             | 1.3256                        |
| $n = 1000$    | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. AN      | 0.3766             | 0.6693                        | 0.3475             | 0.5707                        |
| Spec. AT      | 0.3981             | 0.7455                        | 0.3734             | 0.6570                        |

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

| Specification | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ w/o BC | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ w/o BC |
|---------------|--------------------|-------------------------------|--------------------|-------------------------------|
|               | $k = 3$            | $k = 6$                        | $k = 3$            | $k = 6$                        |
| $n = 300$     | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. BN      | 0.6707             | 2.1182                        | 0.7017             | 2.3237                        |
| Spec. BT      | 0.7209             | 2.4474                        | 0.7563             | 2.7029                        |
| $n = 500$     | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. BN      | 0.5260             | 1.2997                        | 0.5504             | 1.4298                        |
| Spec. BT      | 0.5649             | 1.5078                        | 0.5867             | 1.6318                        |
| $n = 1000$    | MAE                | RMSE                          | MAE                | RMSE                          |
| Spec. BN      | 0.3870             | 0.7040                        | 0.4174             | 0.8148                        |
| Spec. BT      | 0.4053             | 0.7721                        | 0.4230             | 0.8444                        |
Table 3: The Performance of the Confidence Intervals (Specification A)

| Specification | Nom. Cov. Prob. | $k = 3$ | $k = 6$ |
|---------------|----------------|--------|--------|
|               |                | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ |
| Spec. AN      | 99%            | 0.9877 0.9811 | 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 |
| 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 |
| 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)

| Specification | Nom. Cov. Prob. | $k = 3$ | $k = 6$ |
|---------------|----------------|--------|--------|
|               |                | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ | Plug-in $\theta_0$ | Plug-in $\hat{\theta}$ |
| 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 |
| 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 |
| 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 $\theta_0$ 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 $\theta_0$ 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:

$$h_i = \beta_0 + \log(w_i)\beta_1 + Z_{2i}\beta_4 + \varepsilon_i \text{ and}$$

$$D_i = 1 \{X_i^T\theta_0 \geq \eta_i\},$$

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.

| Specification | 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 income     | 24066.3     | 20086.7   | 220 | 147970 | 24519.8   | 19927.9   | 220 | 147970 |
| husband’s labor income  | 50821.9     | 40895.6   | 52  | 212480 | 49236.1   | 38190.5   | 52  | 212480 |
| household’s total income| 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 | 1464.1     | 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 $\theta_0$ 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 $\varepsilon_i$ 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 \( \sqrt{n}\{\hat{\beta}(\hat{\theta}_{\text{probit}}) - \hat{\beta}(\hat{\theta}_{\text{mx.scr}})\} \), \( \hat{\theta}_{\text{probit}} \) denoting a probit estimator of \( \theta_0 \) and \( \hat{\theta}_{\text{mx.scr}} \) 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 \(\eta \in (0, \varepsilon]\),

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

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 \(\mathbb{R}\)-valued. Let \(\hat{\mu}_\theta\) and \(\hat{U}_{\theta, i}\) be \(\hat{\mu}\) and \(\hat{U}_i\) except
that \( \hat{\theta} \) is replaced by \( \theta \), and let \( \mu_\theta \) and \( U_{\theta,i} \) be \( \mu \) and \( U_{0,i} \) except that \( \theta_0 \) is replaced by \( \theta \), where \( U_{0,i} = F_{\theta_0}(X_i^T \theta_0) \). First, observe that

\[
\hat{a}(\theta) - a(\theta) = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i S_i \cdot \left\{ \varphi(\hat{\mu}_\theta(U_{\theta,i})) - \varphi(\mu(\theta,U_{\theta,i})) \right\} \\
+ \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} \left\{ D_i S_i \cdot \varphi(\mu(\theta,U_{\theta,i})) - \mathbb{E}[D_i S_i \cdot \varphi(\mu(\theta,U_{\theta,i}))] \right\} \\
+ \left\{ \frac{1}{n} \frac{1}{\sum_{i=1}^{n} D_i} - \mathbb{P}\{D_i = 1\} \right\} \mathbb{E}[D_i S_i \cdot \varphi(\mu(\theta,U_{\theta,i}))] \\
\equiv A_{1n}(\theta) + A_{2n}(\theta) + A_{3n}(\theta), \text{ say.}
\]

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

\[
\frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i S_i \cdot \varphi'(\mu(\theta,U_{\theta,i})) \left\{ \hat{\mu}_\theta(U_{\theta,i}) - \mu(\theta,U_{\theta,i}) \right\} \\
+ \frac{1}{2 \sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i S_i \cdot \varphi''(\mu(\theta,U_{\theta,i})) \left\{ \hat{\mu}_\theta(U_{\theta,i}) - \mu(\theta,U_{\theta,i}) \right\}^2
\]

\[
\equiv B_{1n}(\theta) + B_{2n}(\theta), \text{ say,}
\]

where \( \mu^*_\theta(U_{\theta,i}) \) lies on the line segment between \( \hat{\mu}_\theta(U_{\theta,i}) \) and \( \mu(\theta,U_{\theta,i}) \). Let \( 1_{n,i} = 1\{ |U_{0,i} - 1| > h/2 \} \), and write \( B_{2n}(\theta) \) as

\[
\frac{1}{2 \sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i S_i \cdot \varphi''(\mu(\theta(U_{\theta,i}))) \left\{ \hat{\mu}_\theta(U_{\theta,i}) - \mu(\theta(U_{\theta,i})) \right\}^2 1_{n,i} \\
+ \frac{1}{2 \sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i S_i \cdot \varphi''(\mu^*_\theta(U_{\theta,i})) \left\{ \hat{\mu}_\theta(U_{\theta,i}) - \mu^*_\theta(U_{\theta,i}) \right\}^2 \{1 - 1_{n,i}\}
\]

\[
\equiv C_{1n}(\theta) + C_{2n}(\theta), \text{ say.}
\]

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)),

\[
(7.1) \quad \max_{1 \leq i \leq n} \sup_{\theta(B(\theta_0;\varepsilon))} \left| \hat{U}_{\theta,i} - U_{\theta,i} \right| = O_P(1/\sqrt{n}) \quad \text{and} \\
\max_{1 \leq i \leq n} \sup_{\theta(B(\theta_0;\varepsilon))} \left| U_{\theta,i} - U_{0,i} \right| = O_P(c_n).
\]

By Assumption C3(ii), \( c_n h^{-1} \to 0 \). As for \( C_{1n}(\theta) \), we bound \( \left| \hat{\mu}_\theta(U_{\theta,i}) - \mu(\theta,U_{\theta,i}) \right| \) by (using (7.1))

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

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 (7.1), and that \( \mu(\theta) \) is continuously differentiable with a bounded derivative (see Assumption G2(ii)). Since
\( \varphi''(\cdot) \) is continuous and bounded on the support of \( E[W|X, D = 1] \) (Assumptions C1(i) and G1(ii)(a)), \( \sup_{\theta \in B(\theta_0; c_n)} |C_{1n}(\theta)| = O_P(n^{-1}h^{-2}(\log n)) \). As for \( C_{2n}(\theta) \), we bound \( \sup_{\theta \in B(\theta_0; c_n)} |\hat{\mu}_0(\hat{U}_{\theta,i}) - \mu_0(U_{\theta,i})|(1 - 1_{n,i}) = O_P(h + n^{-1/2}) \), 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 \( C > 0 \),

\[
\sup_{\theta \in B(\theta_0;c_n)} |C_{2n}(\theta)| \leq C \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\mu}_0(\hat{U}_{\theta,i}) - \mu_0(U_{\theta,i}) \right\}^2 |1 - 1_{n,i}|
\]

\[
\leq C \max_{1 \leq j \leq n} \sup_{\theta \in B(\theta_0;c_n)} \left\{ \hat{\mu}_0(\hat{U}_{\theta,j}) - \mu_0(U_{\theta,j}) \right\}^2 |1 - 1_{n,j}|
\]

\[
= O_P(h^2 + n^{-1}) \cdot O_P(h) = o_P(n^{-1/2}).
\]

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 \( P_1 = P\{D_i = 1\} \),

\[
\frac{1}{P_1} \cdot \frac{1}{n} \sum_{i=1}^{n} D_i E\left[ S_i \cdot \varphi'(\mu_0(U_{\theta,i}))|U_{\theta,i}, D_i = 1 \right] \{ \varphi(W_i) - \mu_0(U_{\theta,i}) \} + O_P(1/\sqrt{n}),
\]

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

\[
E\left[ S_i \cdot \varphi(\mu_0(U_{\theta,i}))|D_i = 1 \right] \cdot \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} \{ P\{D = 1\} - D_i \} + O_P(1/\sqrt{n})
\]

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

\[
\frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} D_i E\left[ S_i \cdot \varphi'(\mu_0(U_{\theta,i}))|U_{\theta,i}, D_i = 1 \right] \{ \varphi(W_i) - \mu_0(U_{\theta,i}) \}
\]

\[
+ \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} \{ D_i S_i \cdot \varphi(\mu_0(U_{\theta,i})) - E[D_i S_i \cdot \varphi(\mu_0(U_{\theta,i}))] \}
\]

\[
+ E\left[ S_i \cdot \varphi(\mu_0(U_{\theta,i}))|D_i = 1 \right] \cdot \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} \{ P\{D = 1\} - D_i \} + o_P(1),
\]

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

\[
(7.2) \quad \sup_{\theta \in B(\theta_0;c_n)} |\sqrt{n}\{ \hat{\alpha}(\theta) - \alpha(\theta) \}| = O_P(1)
\]

and

\[
\sup_{\theta \in B(\theta_0;c_n)} |\sqrt{n}\{ \hat{\alpha}(\theta) - \alpha(\theta) \} - (\hat{\alpha}(\theta_0) - \alpha(\theta_0))\} | = o_P(1).
\]
(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 $||\hat{\theta} - \theta_0|| = O_P(c_n)$ and $\sqrt{n}c_n^2 \to 0$ as $n \to \infty$. We write

$$
\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}\{H(\hat{a}(\theta), \hat{b}) - H(a(\theta), \hat{b})\} = H_1(a(\theta_0), b_0)\sqrt{n}\{\hat{a}(\theta) - \hat{a}(\theta_0)\} + o_P(1)
$$

$$
= H_1(a(\theta_0), b_0)\sqrt{n}\{\hat{a}(\theta) - a(\theta_0) + a(\theta_0)\}
+ H_1(a(\theta_0), b_0)\sqrt{n}\{\hat{a}(\theta) - a(\theta)\} + o_P(1) \equiv D_{1n} + D_{2n}, \text{ say.}
$$

The second equality uses the first statement of (7.2) and continuity of $H_1(\cdot, \cdot)$ (Assumption G1(ii)(b)). By the second statement of (7.2), $D_{1n} = o_P(1)$. As for $D_{2n}$, we apply Lemma A1 to deduce that

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

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. 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 $(S, W, X) \in \mathbb{R}^{d_S + d_W + d_X}$. Let $\mathcal{S}_S, \mathcal{S}_X$ and $\mathcal{S}_W$ be the supports of $S, X$, and $W$ respectively. Let $\Lambda$ be a class of $\mathbb{R}$-valued functions on $\mathbb{R}^{d_X}$ with a generic element denoted by $\lambda$. We also let $\Phi$ and $\Psi$ be classes of real functions on $\mathbb{R}$ and $\mathbb{R}^{d_S}$ with generic elements $\varphi$ and $\psi$ and let $\hat{\varphi}$ and $\hat{\psi}$ be their envelopes. Let $L_p(P)$, $p \geq 1$, be the space of $L_p$-bounded functions: $||f||_p \equiv \{\int |f(x)|^p dP(dx)\}^{1/p} < \infty$, and for a space of functions $\mathcal{F} \subset L_p(P)$ for $p \geq 1$, let $N(\mathcal{F}, ||\cdot||_p)$ denote the bracketing number of $\mathcal{F}$ with respect to the norm $||\cdot||_p$, i.e., the smallest number $r$ such that there exist $f_1, \cdots, f_r$ and $\Delta_1, \cdots, \Delta_r \in L_p(P)$ such that $||\Delta_i||_p < \epsilon$ 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 $N(\mathcal{F}, ||\cdot||_\infty)$ to be the bracketing number of $\mathcal{F}$ with respect to the sup norm $||\cdot||_\infty$, where for any real map $f$ on $\mathbb{R}^{d_X}$, we define $||f||_\infty = \sup_{z \in \mathbb{R}^{d_X}} |f(z)|$. For any norm $||\cdot||$ which is equal to $||\cdot||_p$ or $||\cdot||_\infty$, we define $N(\mathcal{F}, ||\cdot||)$ to be the covering number of $\mathcal{F}$, i.e., the smallest number of $\epsilon$-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 $g_{\varphi, \lambda}(u) \equiv \mathbb{E}[\varphi(W) | U_\lambda = u]$ and $g_{\psi, \lambda}(u) \equiv \mathbb{E}[\psi(S) | U_\lambda = u]$.

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:

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

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

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 \( G \equiv \{ g_{\varphi,\lambda}(\cdot) : (\varphi, \lambda) \in \Phi \times \Lambda \} \). We begin with the following lemma that establishes the bracketing entropy bound for \( G \) with respect to \( || \cdot ||_q \), \( q \geq 1 \).

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

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

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

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

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

**Proof of Lemma B1**: For each \( \eta \in (0,1] \), we define \( g_{\varphi,\lambda,\eta}(u) = E[\varphi(W)Q_{\eta}(U_{\lambda} - u)] \), where \( Q_{\eta}(u) = Q(u/\eta)/\eta \), and \( Q(u) = 1 \) \( u \in [-1/2, 1/2] \). For each \( \eta \in (0,1] \), we let \( G_{\eta} = \{ g_{\varphi,\lambda,\eta}(\cdot) : (\varphi, \lambda) \in \Phi \times \Lambda \} \). First, we show that for all \( \eta \in (0,1] \), and for all \( \varepsilon > 0 \) and \( p \geq 1 \),

\[
N_{||}(1 + 8C_{\Lambda}G_{\Phi} \cdot \varepsilon/\eta, G_{\eta}, || \cdot ||_\infty) \leq N_{||}(\varepsilon, \Phi, || \cdot ||_p) \cdot N_{||}(\varepsilon, \Lambda, || \cdot ||_\infty).
\]

Fix \( \eta \in (0,1], \varepsilon > 0 \), and \( p \geq 1 \), and choose two sets of brackets \( \{ \varphi_j, \Delta_{\varphi,j} \}_{j=1}^{N_{\Phi}} \) and \( \{ \lambda_j, \Delta_{\Lambda,j} \}_{j=1}^{N_{\Lambda}} \) that constitute \( \varepsilon \)-brackets for \( \Phi \) and \( \Lambda \), with respect to \( || \cdot ||_p \) and \( || \cdot ||_\infty \) respectively, where \( N_{\Phi} = N_{||}(\varepsilon, \Phi, || \cdot ||_p) \) and \( N_{\Lambda} = N_{||}(\varepsilon, \Lambda, || \cdot ||_\infty) \). Define \( g_{jk,\eta}(u) = E[\varphi_j(W)Q_{\eta}(U_{\lambda_k} - u)] \). For any \( g_{\varphi,\lambda,\eta} \in G_{\eta}, \) we can choose the pairs \( (\varphi_j, \Delta_{\varphi,j}) \) and \( (\lambda_k, \Delta_{\Lambda,k}) \) such that

\[
|\varphi(w) - \varphi_j(w)| \leq \Delta_{\varphi,j}(w), \text{ for all } w \in S_W \text{ and } \lambda(x) - \lambda_k(x) \leq \Delta_{\Lambda,k}(x), \text{ for all } x \in S_X
\]

and \( ||\Delta_{\varphi,j}||_p \leq \varepsilon \) and \( ||\Delta_{\Lambda,k}||_\infty \leq \varepsilon \). Note that for \( u \in [0,1] \),

\[
|g_{\varphi,\lambda,\eta}(u) - g_{jk,\eta}(u)| \leq E[\Delta_{\varphi,j}(W) \cdot |Q_{\eta}(U_{\lambda} - u)|] + E[\tilde{\varphi}(W) \cdot |Q_{\eta}(U_{\lambda} - u) - Q_{\eta}(U_{\lambda} - u)|].
\]
Certainly, $E[\Delta_{\phi,j}(W) \cdot |Q_\eta(U_\lambda - u)|] \leq E[\Delta_{\phi,j}(W)] / \eta$. As for the second term, let $\Delta_{\lambda,\lambda_k} = |U_\lambda - U_{\lambda_k}|$ and bound it by

$$\frac{1}{\eta} G_{\phi} \cdot P \{ \eta/2 - \Delta_{\lambda,\lambda_k} \leq |U_\lambda - u| \leq \eta/2 + \Delta_{\lambda,\lambda_k} \} \leq \frac{8}{\eta} C_{\phi} G_{\phi} ||\Delta_{\lambda,\lambda_k}||_\infty,$$

because $\Delta_{\lambda,\lambda_k} \leq 4C_{\lambda} ||\lambda,\lambda||_\infty$. Take $\Delta_{jk,\eta}(u) \equiv E[\Delta_{\phi,j}(W)] / \eta + 8C_{\lambda} G_{\phi} ||\Delta_{\lambda,\lambda_k}||_\infty / \eta$ (a constant function), so that $||\Delta_{jk,\eta}||_\infty \leq (1 + 8C_{\phi} G_{\phi}) \epsilon / \eta$. Hence take $\{g_{jk,\eta}, \Delta_{jk,\eta}\}_{j,k=1}^{N_{\Phi},N_{\lambda}}$ to be $(1 + 8C_{\phi} G_{\phi}) \epsilon / \eta$-brackets of $G_{\eta}$, affirming (7.3).

We turn to $G$. For any $(\phi, \lambda) \in \Phi \times \Lambda$, we obtain that for $u \in [0, 1]$,

$$|g_{\phi,\lambda}(u) - g_{\phi,\lambda}(u)| \leq |E[\{g_{\phi,\lambda}(U_\lambda) - g_{\phi,\lambda}(u)\} \cdot Q_\eta(U_\lambda - u)]| + |g_{\phi,\lambda}(u) E[Q_\eta(U_\lambda - u) - 1]| \leq C_L \eta + G_{\phi} \cdot b_\eta(u),$$

where $b_\eta(u) \equiv \frac{1}{2} (1 \{u \in (1 - \eta/2, 1)\} + 1 \{u \in [0, \eta/2]\})$. The second inequality follows by change of variables applied to the leading term. The last inequality follows because for all $\eta \in (0, 1]$,

$$|1 - EQ_\eta(U_\lambda - u)| = \left| 1 - \frac{1}{\eta} \int_0^1 1\{|v - u| \leq \eta/2\} dv \right| = \left| 1 - \frac{1}{\eta} \int_{|u - u/2, u + u/2| \cap [0, 1]} dv \right| \leq b_\eta(u).$$

Fix $\epsilon > 0$ and $q \geq 1$. We select $\eta = \epsilon^{q/(q + 1)}$ and take $((1 + 8C_{\phi} G_{\phi}) \epsilon / \eta$)-brackets $\{g_{jk,\eta}, \Delta_{jk,\eta}\}_{j,k=1}^{N_{\Phi},N_{\lambda}}$ that cover $G_{\eta}$ (with respect to $|| \cdot ||_\infty$) with $N_{\Phi} = N_{[]}^{[]} (\epsilon, \Phi, |i| \cdot ||p||)$ and $N_{\lambda} = N_{[]}^{[]} (\epsilon, \Lambda, |i| \cdot ||\infty||)$. We define

$$\tilde{\Delta}_{jk,\eta}(u) \equiv \Delta_{jk,\eta}(u) + C_L \eta + G_{\phi} \cdot b_\eta(u).$$

Then, certainly,

$$||\tilde{\Delta}_{jk}||_q \leq (1 + 8C_{\phi} G_{\phi}) \epsilon / \eta + C_L \eta + G_{\phi} \cdot ||b_\eta||_q$$

$$\leq (1 + 8C_{\phi} G_{\phi}) \epsilon / \eta + (C_L + G_{\phi}/2) \eta^{1/q}. $$

Therefore, the set $\{g_{jk,\eta}, \Delta_{jk,\eta}\}_{j,k=1}^{N_{\Phi},N_{\lambda}}$ forms the set of $C_{\phi} \epsilon^{q/(q + 1)}$-brackets for $G$ with respect to $|| \cdot ||_q$. This gives the desired entropy bound for $G$. ■

We are prepared to present the uniform Bahadur representation of $\nu_n(\lambda, \varphi, \psi)$. Let $A_n \equiv \{\lambda \in \Lambda : ||\lambda - \lambda_0||_\infty \leq c_n\}$, 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 $\{x_1, \ldots, x_M\}$. We make the following assumptions.

**Assumption B1**: (i) For some $C > 0$, $p \geq q > 4$, $b_\psi \in (0, q/(q - 1))$, and $b_\phi \in (0, q/((q + 1)(q - 1)))$,

$$\log N_{[]}^{[]} (\epsilon, \Phi, |i| \cdot ||p||) < C\epsilon^{-b_\phi} \text{ and } \log N_{[]}^{[]} (\epsilon, \Psi, |i| \cdot ||p||) < C\epsilon^{-b_\phi},$$

and $E[\tilde{\varphi}(W)^p] + E[\tilde{\psi}(S)^p] + \sup_{x \in S_X} E[\tilde{\varphi}(W)|X = x] + \sup_{x \in S_X} E[\tilde{\psi}(W)|X = x] < \infty$. 


For all $\lambda \in C_1$:

First, we show the following:

bounded away from zero on the interior of its support uniformly over $\nu$.

For $q > 4$ in (i) and for some $b_\Lambda \in (0, q/((q + 1)(q - 1))]$ and $C > 0$,

$$\log N_{\nu}(\varepsilon, \Lambda, \| \cdot \|_{\infty}) \leq C \varepsilon^{-b_\Lambda},$$

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**: $E[|\varphi(W)|U_{\lambda} = \cdot]$ 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 $\nu_n$.

**Lemma B2**: Suppose that Assumptions B1-B3 hold. Then,

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

**Proof of Lemma B2**: To make the flow of the arguments more visible, the proof proceeds by making certain claims and proving them at the end of the proof. Without loss of generality, assume that the support of $K$ is contained in $[-1, 1]$. Throughout the proofs, the notation $E_i$ indicates conditional expectation given $(W_i, S_i, X_i)$.

Let $\hat{\rho}_{\varphi, \lambda, i}(t) \equiv (n-1)^{-1} \sum_{j=1, j \neq i}^{n} K_h(U_{n, j} - t) \varphi(W_j), \xi_{1n}(u) \equiv \int_{[u/h,(1-u)/h]\cap[-1,1]} K(v)dv$, and

$$\Delta_{i}^{\varphi,:\psi}(\lambda) \equiv g_{\varphi, \lambda}(U_{n, i}) \{ \varphi(W_i) - g_{\varphi, \lambda}(U_{n, i}) \}.$$

We write $\hat{g}_{\varphi, \lambda, i}(U_{n, i}) - g_{\varphi, \lambda}(U_{n, i})$ as

$$R_{1i}(\lambda, \varphi) = \frac{\hat{\rho}_{\varphi, \lambda, i}(U_{n, i}) - g_{\varphi, \lambda}(U_{n, i})\hat{f}_{\lambda, i}(U_{n, i})}{\xi_{1n}(U_{n, i})} + \frac{[\hat{\rho}_{\varphi, \lambda, i}(U_{n, i}) - g_{\varphi, \lambda}(U_{n, i})\hat{f}_{\lambda, i}(U_{n, i})](\xi_{1n}(U_{n, i}) - \hat{f}_{\lambda, i}(U_{n, i}))}{\hat{f}_{\lambda, i}(U_{n, i})\xi_{1n}(U_{n, i})} = R_{1i}^{A}(\lambda, \varphi) + R_{1i}^{B}(\lambda, \varphi).$$

Put $\pi \equiv (\lambda, \varphi, \psi)$ and $\Pi_n \equiv \Lambda \times \Phi \times \Psi$, and write

$$\nu_n(\pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(S_i) R_{1i}^{A}(\lambda, \varphi) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(S_i) R_{1i}^{B}(\lambda, \varphi) \equiv r_{1n}^{A}(\pi) + r_{1n}^{B}(\pi), \ \pi \in \Pi_n.$$

First, we show the following:

**C1**: $\sup_{\pi \in \Pi_n} |r_{1n}^{B}(\pi)| = o_P(1).$
We turn to $r^A_1(\pi)$, which we write as
\[
\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K^\lambda_{ij} + \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} \{K^\lambda_{n,ij} - K^\lambda_{ij}\} = R_{1n}(\pi) + R_{2n}(\pi),
\]
where $\psi_{n,\lambda,i} \equiv \psi(S_i)/\xi_{1n}(U_{\lambda,i})$, $\Delta \phi_{\lambda,ij} \equiv \phi(W_j) - g_{\phi,\lambda}(U_{\lambda,i})$, $K^\lambda_{n,ij} \equiv K_h(U_{n,\lambda,j} - U_{n,\lambda,i})$, and $K^\lambda_{ij} \equiv K_h(U_{\lambda,j} - U_{\lambda,i})$. We will now show that
\[
\sup_{\pi \in \Pi_n} |R_{2n}(\pi)| \to_p 0.
\]
Later we show the following later. We will now show that
\[
\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij} d_{\lambda,ij} + \frac{1}{2(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} d_{\lambda,ij}^2 K''_{h,ij} = A_{1n}(\pi) + A_{2n}(\pi),
\]
where $K'_{h,ij} \equiv h^{-2} \partial K(t)/\partial t$ at $t = (U_{\lambda,i} - U_{\lambda,j})/h$ and $K''_{h,ij} \equiv h^{-3} \partial^2 K(t)/\partial t^2 |_{t=t_{ij}}$ with $t_{ij} \equiv (1 - a_{ij})(U_{\lambda,i} - U_{\lambda,j}) + a_{ij}(U_{n,\lambda,i} - U_{n,\lambda,j})/h$, for some $a_{ij} \in [0, 1]$. Later we will show the following:

C2: $\sup_{\pi \in \Pi_n} |A_{2n}(\pi)| = o_P(1)$.

We turn to $A_{1n}(\pi)$ which we write as
\[
\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij} \delta^\lambda_{ij} - \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij} \delta^\lambda_{ij} = B_{1n}(\pi) + B_{2n}(\pi),
\]
where $B_{1n}(\pi)$ as
\[
\frac{1}{n-1} \sum_{j=1}^n \left[ \frac{1}{\sqrt{n}} \sum_{i=1, i \neq j} \{\psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij} - E_j [\psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij}]\} (U_{n,\lambda,j} - U_{\lambda,j}) \right. + \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n E_j [\psi_{n,\lambda,i} \Delta \phi_{\lambda,ij} K'_{h,ij}] (U_{n,\lambda,j} - U_{\lambda,j}) = C_{1n}(\pi) + C_{2n}(\pi),
\]
As for $C_{1n}(\pi)$, we show the following later.

C3: $\sup_{\pi \in \Pi_n} |C_{1n}(\pi)| = o_P(1)$. 

We deduce a similar result for $B_{2n}(\pi)$, so that we write

\[(7.6) A_{1n}(\pi) = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} E_j \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right] (U_{n,\lambda,j} - U_{\lambda,j}) \]

\[= \frac{1}{(n-1)\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1, i\neq j}^{n} E_i \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right] (U_{n,\lambda,i} - U_{\lambda,i}) + o_P(1) \]

\[-D_{1n}(\pi) - D_{2n}(\pi) + o_P(1), \text{ say.} \]

Now, we show that $D_{1n}(\pi)$ and $D_{2n}(\pi)$ cancel out asymptotically. As for $D_{1n}(\pi)$, using Hoeffding's decomposition and taking care of the degenerate $U$-process (e.g. see C3 and its proof below),

\[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_{0}^{1} E \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right] (1\{U_{\lambda,i} \leq u\} - u) \, du + o_P(1), \]

uniformly over $\pi \in \Pi_n$. Similarly, as for $D_{2n}(\pi)$, we can write it as

\[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{0}^{1} E \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right] (1\{U_{\lambda,j} \leq u\} - u) \, du + o_P(1), \]

uniformly over $\pi \in \Pi_n$. Note that $E \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right]$ does not depend on a particular choice of the pair $(i,j)$ as long as $i \neq j$, so that

\[E \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K'_{h,ij} \right] = E \left[ \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,j} K'_{h,jj} \right], \]

whenever $j \neq i$. Hence $D_{1n}(\pi) = D_{2n}(\pi) + o_P(1)$, uniformly over $\pi \in \Pi_n$, and that

\[\sup_{\pi \in \Pi_n} |A_{1n}(\pi)| = o_P(1), \text{ which, together with (C2), completes the proof of (7.4).} \]

It remains to show that

\[(7.7) \sup_{\pi \in \Pi_n} \left| R_{1n}(\pi) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{\lambda,i}(U_{n,i}) \left\{ \varphi(W_i) - g_{\varphi,\lambda}(U_{n,i}) \right\} \right| = o_P(1). \]

We define $q_{n,ij}^{\pi} = \psi_{n,\lambda,i} \Delta_{\varphi,\lambda,ij} K_{ij}^{\lambda}$ and write $R_{1n}(\pi)$ as

\[(7.8) \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} q_{n,ij}^{\pi}. \]

Let $\rho_{n,ij}^{\pi} = q_{n,ij}^{\pi} - E_i[q_{n,ij}^{\pi}] - E_j[q_{n,ij}^{\pi}] + E[q_{n,ij}^{\pi}]$ and define

\[u_n(\pi) \equiv \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \rho_{n,ij}^{\pi}. \]

Then, $\{u_n(\pi) : \pi \in \Pi_n\}$ is a degenerate $U$-process on $\Pi_n$. We write (7.8) as

\[(7.9) \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\{ E_i[q_{n,ij}^{\pi}] + E_j[q_{n,ij}^{\pi}] - E[q_{n,ij}^{\pi}] \right\} + u_n(\pi). \]

We will later show the following two claims.

C4: $\sup_{\pi \in \Pi_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ E_i[q_{n,ij}^{\pi}] - E[q_{n,ij}^{\pi}] \} \right| = o_P(1)$.
Then the proof of Lemma B2 is completed by showing the following.

\[ 30 \quad \text{K. Song} \]

\[ (7.11) \]

C5: \( \sup_{x \in \Pi_n} |u_n(\pi)| = o_P(1) \).

We conclude from these claims that uniformly over \( \pi \in \Pi_n \),

\[ \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} q_{n,ij}^\pi = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} E_j[q_{n,ij}^\pi] + o_P(1). \]

Then the proof of Lemma B2 is completed by showing the following.

C6: \( \sup_{x \in \Pi_n} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (E_j[q_{n,ij}^\pi] - g_{\varphi,\lambda}(U_{\lambda,i})\{\varphi(W_j) - g_{\varphi,\lambda}(U_{\lambda,i})\}) \right| = o_P(1). \)

**Proof of C1:** From the proof of Lemma A3 of Song (2009) (by replacing \( \lambda \) with \( F_\lambda \circ \lambda \) there), it follows that

\[ \max_{1 \leq i \leq n} \sup_{\lambda \in \Lambda} \sup_{x \in \mathbb{R}^d} |F_{n,\lambda,i}(x) - F_\lambda(x)| = O_P(n^{-1/2}), \]

where \( F_{n,\lambda,i}(\bar{\lambda}) \equiv \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} 1 \{ \lambda(X_j) \leq \bar{\lambda} \} \). We bound \( \max_{1 \leq i \leq n} \sup_{\lambda \in \Lambda} |f_{\lambda,i} - f_{\lambda,i}(U_{\lambda,i})| \) by

\[ \max_{1 \leq i \leq n} \sup_{\lambda \in \Lambda} \sup_{u \in [0,1]} |f_{\lambda,i}(u) - \xi_{1n}(u)| + \max_{1 \leq i \leq n} \sup_{\lambda \in \Lambda} |\xi_{1n}(U_{\lambda,i}) - \xi_{1n}(U_{\lambda,\lambda})| \]

\[ = \ O_P(n^{-1/2}h^{-1} \sqrt{\log h}) + O_P(n^{-1/2}h^{-1}) = O_P(n^{-1/2}h^{-1} \sqrt{\log h}), \]

using (29) of Song (2009). Hence, uniformly over \( 1 \leq i \leq n \) and over \( \lambda \in \Lambda \),

\[ |\hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - \hat{\varphi}_{\lambda,i}(U_{\lambda,i})g_{\varphi,\lambda}(U_{\lambda,i})| \]

\[ \leq |\hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - \xi_{1n}(U_{\lambda,i})g_{\varphi,\lambda}(U_{\lambda,i})| + |g_{\varphi,\lambda}(U_{\lambda,i})||\xi_{1n}(U_{\lambda,i}) - \hat{\varphi}_{\lambda,i}(U_{\lambda,i})| \]

\[ = |\hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - \xi_{1n}(U_{\lambda,i})g_{\varphi,\lambda}(U_{\lambda,i})| + O_P(n^{-1/2}h^{-1} \sqrt{\log h}). \]

As for the leading term, we apply (23) of Song (2009) and Assumption B3 to deduce that uniformly over \( 1 \leq i \leq n \) and over \( (\lambda, \varphi) \in \Lambda \times \Phi \),

\[ |\hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - \xi_{1n}(U_{\lambda,i})g_{\varphi,\lambda}(U_{\lambda,i})| (1 - 1_{n,\lambda,i}) = O_P(h) \]

\[ |\hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - \xi_{1n}(U_{\lambda,i})g_{\varphi,\lambda}(U_{\lambda,i})| 1_{n,\lambda,i} = O_P(h^2 + n^{-1/2}h^{-1} \sqrt{\log h}), \]

where \( 1_{n,\lambda,i} = 1 \{ |1 - U_{n,\lambda,i}| > h \} \). Also, observe that (e.g. see arguments after (29) of Song (2009))

\[ \xi_{1n}(u) \geq 1/2 \text{ for all } u \in [0,1]. \]

Therefore, we bound \( |r_{1n}^\pi(\pi)| \) by

\[ \frac{C_1}{\sqrt{n}} \sum_{i=1}^{n} \left| \hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - g_{\varphi,\lambda}(U_{\lambda,i})\xi_{1n}(U_{\lambda,i})\right| (1 - 1_{n,\lambda,i}), \]

\[ + \frac{C_2}{\sqrt{n}} \sum_{i=1}^{n} \left| \hat{\varphi}_{\lambda,i}(U_{\lambda,i}) - g_{\varphi,\lambda}(U_{\lambda,i})\xi_{1n}(U_{\lambda,i})\right| \]
for some $C_1, C_2 > 0$, uniformly over $\pi \in \Pi_n$ with large probability. From (7.11) and (7.12), the first term is equal to $O_P(n^{-1/2} h^{-2} (\log h)) = o_P(1)$ uniformly over $1 \leq i \leq n$. As for the second term, from (7.11) and (7.12), it is bounded with large probability by

$$C n^{-1/2} \sqrt{-\log h} \cdot \frac{1}{n} \sum_{i=1}^{n} 1 \{ |1 - U_{0,i}| \leq C h \} = O_P(n^{-1/2} h \sqrt{-\log h}) = o_P(n^{-1/2}),$$

for some $C > 0$ with large probability. Hence (C1) is established.

**Proof of C2:** Let $\tilde{\Delta}_{ij} \equiv \tilde{\varphi}(W_i) + \sup_{x \in S_x} E[\tilde{\varphi}(W_j)]|X_j = x]$. Since $\max_{1 \leq i, j \leq n} \sup_{\lambda \in \Lambda} d_{\lambda,ij}^2 = O_P(n^{-1})$ by (7.10), with large probability, we bound $|A_{2n}(\pi)|$ by the absolute value of (7.14)

$$C \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left( \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right) - E \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] \right] + \sqrt{\frac{n}{n}} \left[ n \right] E \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] + o_P(1).

Using the standard $U$ statistics theory, we bound the leading term by

$$\left| \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left( \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right) - E \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] \right] + \sqrt{\frac{n}{n}} \left[ n \right] E \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] + o_P(1).

The expected value of the sum above is bounded by

$$\frac{C}{n} \sqrt{Var \left( E_i \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] \right]} + \frac{C}{n} \sqrt{Var \left( E_j \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right] \right]} = O(n^{-1} h^{-5/2}) = o(1),$$

so that the leading term of (7.14) is $o(1)$. On the other hand, $\frac{C}{n} \sqrt{E \left[ \tilde{\psi}_i \tilde{\Delta}_{ij} \big| K_{h_{ij}}'' \right]} = O(n^{-1/2} h^{-2}) = o(1).

**Proof of C3:** Let $K'(u) \equiv (\partial K(u)/\partial u)1\{u \in (0, 1)\}$. Then $K'(\cdot/h)$ is uniformly bounded and bounded variation. Let $K_\Lambda \equiv \{ K'(u - \sigma(\cdot))/h : (\sigma, u) \in I \times [0, 1] \}$, where $I \equiv \{(F_\lambda \circ \lambda)(x) : \lambda \in \Lambda \}$. Also, let $W_\Lambda \equiv \{ \xi_{1n}(\sigma(\cdot)) : \sigma \in I \}$. Take $p \geq 4$ as in Assumption B1. By Lemma A1 of Song (2009) and Assumptions B1(ii) and B3(iv),

$$(7.15) \quad \log N_\| (\varepsilon, K_\Lambda, \| \cdot \|_p) \leq \log N(C \varepsilon, I, \| \cdot \|_\infty) + C/\varepsilon \leq C \varepsilon^{-b_\Lambda},$$

for some $C > 0$. Let $\xi_{\pi, u}(S_i, X_i) \equiv \psi_{n,\lambda, i}K'(u - U_{\lambda, i})/h$. We bound $\sup_{\pi \in \Pi_n} |C_{1n}(\pi)| \leq h^{-2} G_1 \cdot V_1 + h^{-2} G_2 \cdot V_2$, where

$$V_1 \equiv \sup_{(\pi, u) \in \Pi_n \times [0, 1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \xi_{\pi, u}(S_i, X_i) - E \xi_{\pi, u}(S_i, X_i) \right) \right|$$

and

$$V_2 \equiv \sup_{(\pi, u) \in \Pi_n \times [0, 1]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \xi_{\pi, u}(S_i, X_i) g_{\varphi, \lambda}(U_{\lambda, i}) - E \xi_{\pi, u}(S_i, X_i) g_{\varphi, \lambda}(U_{\lambda, i}) \right) \right|.$$
where $G_{1n} \equiv \sup_{\lambda \in \Lambda_n} \frac{1}{n} \sum_{j=1}^{n} \tilde{\varphi}(W_j)|\delta_j^\lambda|$ and $G_{2n} \equiv \sup_{\lambda \in \Lambda_n} \frac{1}{n} \sum_{j=1}^{n} |\delta_j^\lambda|$. Define

\[ F_1 \equiv \{ \xi_{\pi,u}(\cdot,\cdot) : (\pi,u) \in \Pi_n \times [0,1] \} \] and

\[ F_2 \equiv \{ \xi_{\pi,u}(\cdot,\cdot) : (\varphi,\lambda) \in \Sigma \} \].

Then, $F_1 \subset (\Psi/W_\Lambda) \cdot \mathcal{K}_1 \Lambda$ and $F_2 \subset (\Psi/W_\Lambda) \cdot \mathcal{K}_1 \Lambda \cdot \mathcal{H}$, where

\[ \mathcal{H} \equiv \{(\varphi,\lambda,\sigma) : (\varphi,\lambda) \in \Phi \times \Lambda \}. \]

By Assumption B1(i) and Lemma B1, $\log N_{\mathcal{H}}(\varepsilon,\mathcal{H},||\cdot||_q) \leq C \varepsilon^{-(q+1)(b_\Psi \vee b_\Lambda)}$. Combining this with (7.15), the entropy bound for $\Psi$ in Assumption B1(i), (7.15), and the fact that using the maximal inequality of Pollard (1989) and the fact that $\sup_{x \in S} \mathbb{E}[\tilde{\varphi}(W)|X=x] < \infty$ and $\sup_{x \in S} \mathbb{E}[\tilde{\varphi}(W)|X=x] < \infty$, we find that

$$\log N_{\mathcal{H}}(\varepsilon,F_1,||\cdot||_2) \leq C \varepsilon^{-(b_\Psi \vee b_\Lambda)} \quad \text{and} \quad \log N_{\mathcal{H}}(\varepsilon,F_2,||\cdot||_2) \leq C \varepsilon^{-(b_\Psi \vee (q+1)(b_\Psi \vee b_\Lambda))}.$$ 

From (7.13), we take an envelope of $F_1$ as $\xi_1(s,x) = 2\tilde{\psi}(s)||K'||\infty$ and an envelope of $F_2$ as $\xi_2(s,x) = 2\tilde{\psi}(s)||K'||\infty \cdot \sup_{x \in S} \mathbb{E}[\tilde{\varphi}(W)|X=x]$. Certainly, $||\xi_1||_p < \infty$ and $||\xi_2||_p < \infty$ by Assumption B1(i). Using the maximal inequality of Pollard (1989) and the fact that $b_\Psi \vee [(q+1)(b_\Psi \vee b_\Lambda)] < 2$ (Assumptions B1(i) and B1(ii)(a)), we find that $V_{1n} = O_P(1)$ and $V_{2n} = O_P(1)$.

By the fact that $\max_{1 \leq j \leq n} \sup_{\lambda \in \Lambda_n} |\delta_j^\lambda| = O_P(n^{-1/2})$, we also deduce that $G_{1n} = O_P(n^{-1/2})$ and $G_{2n} = O_P(n^{-1/2})$. The desired result follows because $O_P(n^{-1/2}h^{-2}) = o_P(1)$.

**Proof of C4:** Observe that $E_i[q_{n,i}^\pi]$ is equal to $\zeta_{\pi}(S_i, X_i)$, where for $\pi \in \Pi_n$,

$$\zeta_{\pi}(S_i, X_i) = \frac{\psi(S_i)}{\xi_{\Lambda_n}(U_{\lambda,i})} \int_0^1 \{g_{\varphi,\lambda}(u) - g_{\varphi,\lambda}(U_{\lambda,i})\} K_h(u - U_{\lambda,i})du.$$ 

Define $F_3 = \{ \zeta_{\pi}(\cdot,\cdot) : \pi \in \Pi_n \}$. Take $p \geq 4$ as in Assumption B1. Then, similarly as in the proof of (C3), we can show that

$$\log N_{\mathcal{H}}(\varepsilon,F_3,||\cdot||_q) \leq C \varepsilon^{-(b_\Psi \vee (q+1)(b_\Psi \vee b_\Lambda))}$$

for some $C > 0$. With $\sigma_{\lambda}(x) \equiv (F_\Lambda \circ \lambda)(x)$, we take

$$\tilde{\zeta}(s,x) = 2\tilde{\psi}(s) \sup_{(\varphi,\lambda) \in \Phi \times \Lambda} \left[ \int_0^1 \{g_{\varphi,\lambda}(u) - g_{\varphi,\lambda}(\sigma_{\lambda}(x))\} K_h(u - \sigma_{\lambda}(x))du \right]$$

as an envelope of $F_3$. Observe that $E[\tilde{\zeta}(S_i, X_i)^2]$ is bounded by

$$4 \sup_{x \in S} \mathbb{E}[\tilde{\psi}(S)|X=x] \cdot \int_0^1 \sup_{(\varphi,\lambda) \in \Phi \times \Lambda} \left[ \int_0^1 \{g_{\varphi,\lambda}(t_2) - g_{\varphi,\lambda}(t_1)\} K_h(t_2 - t_1)dt_2 \right]^2 dt_1.$$ 

By change of variables, the integral inside the bracket becomes

$$\int_{[-t_1/h,(1-t_1)/h] \cap [-1,1]} \{g_{\varphi,\lambda}(t_1 + ht_2) - g_{\varphi,\lambda}(t_1)\} K_h(t_2)dt_2.$$ 

The result is replicated in Theorem A.2 of van der Vaart (1996).
After tedious algebra (using the symmetry of $K$), we can show that the outside integral in (7.17) is $O(h^5)$. Therefore, $\|\tilde{c}\|_2 = O(h^{3/2})$ as $n \to \infty$. Applying this and the maximal inequality of Pollard (1989),

$$
E \left[ \sup_{\pi \in \Pi_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (E_i[q_{n,i,j}] - E[q_{n,i,j}]) \right| \right] \leq C \int_0^{Ch^{3/2}} \sqrt{1 + \log N_{||}(\varepsilon, F_3, ||.||_2)} d\varepsilon
$$

for some $C > 0$. By (7.16), the last term is of order $O(h^{(3/2) \times (1 - (b_\Psi \vee [(q + 1)\{b_\Phi \vee b_\Lambda\}]) / 2)} \sqrt{-\log h})$. Since $b_\Psi \vee [(q + 1)\{b_\Phi \vee b_\Lambda\}] < 2$, we obtain the wanted result.

**Proof of C5:** Let us define $\tilde{J}_n = \{hq_n^\pi(\cdot, \cdot) : \pi \in \Pi\}$, where $q_n^\pi(Z_i, Z_j) = q_{n,i,j}$, $Z_i = (S_i, W_i, X_i)$, and $q_{n,i,j}$ is defined prior to (7.8). Take $q > 4$ as in Assumption B1. Using similar arguments as in (C3), we can show that for some $C > 0$, $\log N_{||}(\varepsilon, \tilde{J}_n, ||.||_q) \leq C \varepsilon^{-(b_\Psi \vee [(q + 1)\{b_\Phi \vee b_\Lambda\}])}$ for all $\varepsilon > 0$. By Assumption B1, $(b_\Psi \vee [(q + 1)\{b_\Phi \vee b_\Lambda\}])^{-1/2} < 1$. Then, from the proof of C3,

$$
\int_0^1 \left\{ \log N_{||}(\varepsilon, \tilde{J}_n, ||.||_q) \right\}^{(1-1/q)} d\varepsilon \leq C \int_0^1 \varepsilon^{-(b_\Psi \vee [(q + 1)\{b_\Phi \vee b_\Lambda\}])} d\varepsilon < \infty
$$

for some $C > 0$. Furthermore, as in the proof of C3, we can take an envelope of $\tilde{J}_n$ that is $L_q$-bounded. By Theorem 1 of Turki-Moalla (1998), p.878, for some small $\varepsilon > 0$,

$$
h \sup_{\pi \in \Pi_n} |u_{1n}(\pi)| = O_P(n^{1/2 - (1-1/q)+\varepsilon}) = O_P(n^{-1/2+1/q+\varepsilon}).
$$

Therefore, $\sup_{\pi \in \Pi_n} |u_{1n}(\pi)| = O_P(n^{-1/2+1/q+\varepsilon}h^{-1}) = o_P(1)$ by taking small $\varepsilon > 0$ and using Assumption B2(ii) and the fact that $q > 4$. Hence the proof is complete.

**Proof of C6:** For $i \neq j$, we write

$$
\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( E_j[q_{n,i,j}] - g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(W_j) - g_{\psi,\lambda}(U_{\lambda,j})\} \right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( E_j \left[ \left( \frac{q_{n,i,j}}{\xi_{1n}(U_{\lambda,i})} - \frac{1}{\xi_{1n}(U_{\lambda,i})} g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(W_j) - g_{\psi,\lambda}(U_{\lambda,j})\} \right) \right] \right)
$$

$$
- E \left[ \left( \frac{\xi_{1n}(U_{\lambda,i}) - 1}{\xi_{2n}(U_{\lambda,i})} \right) \right] \times \frac{1}{\sqrt{n}} \sum_{j=1}^{n} g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(W_j) - g_{\psi,\lambda}(U_{\lambda,j})\}
$$

$$
\equiv E_{1n}(\pi) - E_{2n}(\pi), \text{ say.}
$$

We focus on $E_{1n}$ first. Note that

(7.18) $$
E \left[ \sup_{\pi \in \Pi_n} \left( E_j[q_{n,i,j}] - g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(W_j) - g_{\psi,\lambda}(U_{\lambda,j})\} \right)^2 \right]
$$

$$
= \int_0^1 \sup_{\pi \in \Pi_n} \left\{ \int_0^1 A_{n,\pi}(t_1, t_2, y)dt_1 \right\}^2 dF_{Y,\lambda}(y, t_2),
$$
where $\int \cdot dF_{Y,\lambda}$ denotes the integration with respect to the joint distribution of $(W_i, U_{\lambda,i})$ and

$$A_{n,\pi}(t_1, t_2, y) = \frac{1}{\xi_{1n}(t_1)} (g_{\varphi,\lambda}(t_1) (\varphi(y) - g_{\varphi,\lambda}(t_1)) K_h(t_1 - t_2) - g_{\varphi,\lambda}(t_2) (\varphi(y) - g_{\varphi,\lambda}(t_2))).$$

After some tedious algebra, we can show that the last term in Eq. (7.18) is $O(h^3)$ (see the proof of C4). Using the similar arguments as in the proof of (C4), we find that

$$\sup_{\pi \in \Pi_n} |E_{1n}(\pi)| = o_P(1).$$

We turn to $E_{2n}$. It is not hard to see that for all $\lambda \in \Lambda_n$,

$$\mathbb{E} \left[ \left( \frac{\xi_{1n}(U_{\lambda,i}) - 1}{\xi_{1n}(U_{\lambda,i})} \right) \right] = \mathbb{E} \left[ \left( \frac{\xi_{1n}(U_{0,i}) - 1}{\xi_{1n}(U_{0,i})} \right) \right] = o(1), \text{ as } h \to 0.$$

The first equality follows because $U_{\lambda,i}$ follows a uniform distribution on $[0, 1]$ for all $\lambda \in \Lambda$. Furthermore, we have

$$\sup_{\pi \in \Pi_n} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (g_{\varphi,\lambda}(U_{\lambda,j}) (\varphi(W_j) - g_{\varphi,\lambda}(U_{\lambda,j})) \right| = O_P(1),$$

using bracketing entropy conditions for $\Psi$, $\Phi$, and $\Lambda_n$, Lemma B1, and the maximal inequality of Pollard (1989). Therefore, we find that $\sup_{\pi \in \Pi_n} |E_{2n}(\pi)| = o_P(1)$. $lacksquare$

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

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

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

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

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

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

(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

$$\sup_{(\lambda, \varphi, \psi) \in \Lambda \times \Phi \times \Psi} \left| \nu_{n,0}(\lambda, \varphi, \psi) - \frac{1}{\sqrt{n}P_1} \sum_{i=1}^{n} (1 - D_i) P(U_{\lambda,i}) g_{\varphi,\lambda,1}(U_{\lambda,i}) \varepsilon_{\varphi,\lambda,0,i} \right| = o_P(1).$$
**Proof of Lemma B3:** (i) Write

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

where \( g_{\varphi,\lambda}^{[d]}(u) = E[\varphi(W_i) 1\{D_i = d\} | U_{\lambda, i} = u] \), \( g_{\lambda}^{[d]}(u) = P\{D_i = d | U_{\lambda, i} = u\} \),

\[ \hat{g}_{\varphi,\lambda, i}^{[d]}(u) = \frac{1}{(n-1)f_{\lambda,d}(u)} \sum_{j=1; j \neq i}^{n} \varphi(W_j) 1\{D_j = d\} K_h(U_{n,\lambda, j} - u), \]

and

\[ \hat{g}_{\lambda, i}^{[d]}(u) = \frac{1}{(n-1)f_{\lambda,d}(u)} \sum_{j=1; j \neq i}^{n} 1\{D_j = d\} K_h(U_{n,\lambda, j} - u). \]

Using the arguments in the proof of Lemma B2, we can write

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

\[ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) \left\{ D_i \varphi(W_i) - g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) \right\} \]

\[ + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) \left\{ g_{\lambda}^{[1]}(U_{\lambda, i}) - D_i \right\} + o_P(1) \]

\[ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) D_i \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} \right\} + o_P(1). \]

Note that

\[ E \left[ \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} D_i \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} \right\} \right] \]

\[ = E \left[ E \left[ \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} \right\} | U_{\lambda, i}, D_i = 1 \right] \right] \]

\[ = E \left[ g_{\varphi,\lambda}^{[1]}(U_{\lambda, i}) E \left[ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[1]}(U_{\lambda, i})}{g_{\lambda}^{[1]}(U_{\lambda, i})} | U_{\lambda, i}, D_i = 1 \right] \right] = 0. \]
Therefore, using similar arguments in the proof of Lemma B2, we conclude that uniformly over \((\lambda, \varphi, \psi) \in \Lambda_n \times \Phi \times \Psi,\)

\[
\nu_{n,1}(\lambda, \varphi, \psi) = \frac{1}{P_1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[1]}(U_{\lambda,i})} D_i \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} + o_P(1).
\]

(ii) Write

\[
\nu_{n,0}(\lambda, \varphi, \psi) = \frac{1}{n} \sum_{i=1}^{n} \psi(S_i) D_i \left\{ \frac{g_{\varphi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{n,\lambda,i})} - \frac{g_{\psi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\}.
\]

Using the arguments in the proof of Lemma B2, we can write

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(S_i) D_i \left\{ \frac{g_{\varphi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{n,\lambda,i})} - \frac{g_{\psi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(S_i) D_i \left\{ \frac{g_{\varphi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{n,\lambda,i})} - \frac{g_{\psi,\lambda}^{[0]}(U_{n,\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} + o_P(1)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \left\{ (1 - D_i) \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\}
\]

\[
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \left\{ \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} - (1 - D_i) \right\} + o_P(1)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \left(1 - D_i\right) \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} + o_P(1).
\]

Note that \(E \left[ \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} (1 - D_i) \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} \right] \) is equal to

\[
\mathcal{E} \left[ E \left[ \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} \vert U_{\lambda,i}, D_i = 0 \right] \right]
\]

\[
= \mathcal{E} \left[ \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \mathcal{E} \left[ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \vert U_{\lambda,i}, D_i = 0 \right] \right] = 0.
\]

From similar arguments in the proof of Lemma B2, uniformly over \((\lambda, \varphi, \psi) \in \Lambda_n \times \Phi \times \Psi,\)

\[
\nu_{n,0}(\lambda, \varphi, \psi) = \frac{1}{P_1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g_{\psi,\lambda}^{[1]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} (1 - D_i) \left\{ \varphi(W_i) - \frac{g_{\varphi,\lambda}^{[0]}(U_{\lambda,i})}{g_{\lambda}^{[0]}(U_{\lambda,i})} \right\} + o_P(1).
\]
8. Derivation of Asymptotic Covariance Matrices

8.1. Example 1: Sample Selection Model with Conditional Median Restrictions. Let \( u_{Z,i} = Z_i - \mathbf{E}[Z_i | U_{0,i}, D_i = 1] \) and \( u_{Y,i} = Y_i - \mathbf{E}[Y_i | U_{0,i}, D_i = 1] \). Let \( \tilde{S}_{ZZ} \), \( \tilde{S}_{ZY} \), \( \tilde{\mu}_Z(u) \) and \( \tilde{\mu}_Y(u) \) be \( \hat{S}_{ZZ} \), \( \hat{S}_{ZY} \), \( \hat{\mu}_Z(u) \) and \( \hat{\mu}_Y(u) \) except that \( \hat{\theta} \) is replaced by \( \theta_0 \). Also, let \( \tilde{U}_i \) be \( \hat{U}_i \) except that \( \hat{\theta} \) is replaced by \( \theta_0 \). Write \( \tilde{S}_{ZZ} - S_{ZZ} = B_{1n} + B_{2n} \), where

\[
B_{1n} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( \tilde{u}_{Z,i} \tilde{u}_{Z,i}^\top - u_{Z,i} u_{Z,i}^\top \right)
\]

\[
B_{2n} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{Z,i}^\top - \mathbf{E}[u_{Z,i} u_{Z,i}^\top | D_i = 1] \right),
\]

and \( \tilde{u}_{Z,i} = Z_i - \tilde{\mu}_Z(\tilde{U}_i) \) and \( \tilde{u}_{Y,i} = Y_i - \tilde{\mu}_Y(\tilde{U}_i) \). Under regularity conditions, we can write \( B_{1n} \) as

\[
\frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i u_{Z,i} \{ \tilde{u}_{Z,i} - u_{Z,i} \}^\top + \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \{ \tilde{u}_{Z,i} - u_{Z,i} \} u_{Z,i}^\top + o_P(1/\sqrt{n}).
\]

Using Lemma B3(i), we find that both the sums above are equal to \( o_P(1/\sqrt{n}) \). Following similar arguments for \( \tilde{S}_{ZY} - S_{ZY} \), we conclude that

\[
(8.1) \quad \tilde{S}_{ZZ} - S_{ZZ} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{Z,i}^\top - \mathbf{E}[u_{Z,i} u_{Z,i}^\top | D_i = 1] \right) + o_P(1/\sqrt{n})
\]

\[
\tilde{S}_{ZY} - S_{ZY} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{Y,i} - \mathbf{E}[u_{Z,i} u_{Y,i} | D_i = 1] \right) + o_P(1/\sqrt{n}).
\]

First write \( \tilde{\beta} - \beta_0 = A_{1n} + A_{2n} \), where

\[
A_{1n} \equiv \{ \tilde{S}_{ZZ}^{-1} - S_{ZZ}^{-1} \} \tilde{S}_{ZY} \quad \text{and} \quad A_{2n} \equiv S_{ZZ}^{-1} \{ \tilde{S}_{ZY} - S_{ZY} \}.
\]

From (8.1), \( ||[\tilde{S}_{ZY} : \tilde{S}_{ZZ}] - [S_{ZY} : S_{ZZ}]|| = O_P(1/\sqrt{n}) \). Hence

\[
A_{1n} = S_{ZZ}^{-1} (S_{ZZ} - \tilde{S}_{ZZ}) S_{ZZ}^{-1} \tilde{S}_{ZY} = S_{ZZ}^{-1} (S_{ZZ} - \tilde{S}_{ZZ}) S_{ZZ}^{-1} \tilde{S}_{ZY} + o_P(1/\sqrt{n})
\]

\[
= S_{ZZ}^{-1} (S_{ZZ} - \tilde{S}_{ZZ}) \beta_0 + o_P(1/\sqrt{n})
\]

because \( \beta_0 = S_{ZZ}^{-1} S_{ZY} \). Therefore, from (8.1), \( \sqrt{n} \{ \tilde{\beta} - \beta_0 \} \) is equal to

\[
(8.2) \quad -S_{ZZ}^{-1} \left( \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{Z,i}^\top - \mathbf{E}[u_{Z,i} u_{Z,i}^\top | D_i = 1] \right) \right) \beta_0
\]

\[
+ S_{ZZ}^{-1} \left( \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{Y,i} - \mathbf{E}[u_{Z,i} u_{Y,i} | D_i = 1] \right) \right) + o_P(1/\sqrt{n}).
\]

As for the first term, observe that \( u_{Z,i}^\top \beta_0 = u_{Y,i} - u_{v,i} \), where \( u_{v,i} = v_i - \mathbf{E}[v_i | U_{0,i}, D_i = 1] \). From this, we find that

\[
(u_{Z,i} u_{Z,i}^\top - \mathbf{E}[u_{Z,i} u_{Z,i}^\top | D_i = 1]) \beta_0
\]

\[
= (u_{Z,i} u_{Y,i} - \mathbf{E}[u_{Z,i} u_{Y,i} | D_i = 1]) - (u_{Z,i} u_{v,i} - \mathbf{E}[u_{Z,i} u_{v,i} | D_i = 1]).
\]
Plugging this into the first sum of (8.2), we find that \( \tilde{\beta} - \beta_0 = S_{ZZ}^{-1} \xi_n + o_P(1/\sqrt{n}) \), where
\[
\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \left( u_{Z,i} u_{v,i} - E \left[ u_{Z,i} u_{v,i} | D_i = 1 \right] \right).
\]
Therefore the asymptotic variance is obtained through Assumption SS0 (i).

8.2. Example 2: Single-Index Matching Estimators of Treatment Effects on the Treated. Define \( \hat{\mu}(\tilde{U}_i) \) to be \( \tilde{\mu}(U_i) \) except that \( \tilde{\theta} \) is replaced by \( \theta_0 \). Write \( \sqrt{n}(\tilde{\beta} - \beta_0) \) as
\[
\frac{1}{n} \sum_{i=1}^{n} Z_i \left( Y_i - \mu_0(U_{0,i}) \right) - \beta_0 \right) \]
\[
+ \frac{1}{n} \sum_{i=1}^{n} Z_i \left( \mu_0(U_{0,i}) - \hat{\mu}(\tilde{U}_i) \right) \equiv A_{1n} + A_{2n}, \text{ say.}
\]
Let \( P_d \equiv P\{Z = d\}, d \in \{0, 1\} \). As for \( A_{1n} \), it is not hard to see that
\[
A_{1n} = \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} Z_i \left( Y_i - \mu_1(U_{0,i}) \right) + \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} Z_i \left( \mu_1(U_{0,i}) - \mu_0(U_{0,i}) - \beta_0 \right)
\]
\[
= \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} Z_i \left( Y_i - \mu_1(U_{0,i}) \right) + \frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} Z_i \left( \mu_1(U_{0,i}) - \mu_0(U_{0,i}) - \beta_0 \right) + o_P(1).
\]
As for \( A_{2n} \), we apply Lemma B3(ii) to deduce that it is equal to
\[
\frac{1}{P_1 \sqrt{n}} \sum_{i=1}^{n} \frac{(1 - Z_i) P(U_{0,i})}{1 - P(U_{0,i})} \left( \mu_0(U_{0,i}) - Y_i \right) + o_P(1).
\]
Combining \( A_{1n} \) and \( A_{2n} \), we find that \( \sqrt{n}(\tilde{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \gamma_i + o_P(1) \), where
\[
\gamma_i = \frac{Z_i \varepsilon_{1,i}}{P_1} - \frac{(1 - Z_i) P(U_{0,i}) \varepsilon_{0,i}}{(1 - P(U_{0,i})) P_1} + \frac{1}{P_1} Z_i \left( \mu_1(U_{0,i}) - \mu_0(U_{0,i}) - \beta_0 \right),
\]
and \( \varepsilon_{d,i} = Y_i - \mu_d(U_{0,i}), d \in \{0, 1\} \). Hence \( V_{SM} = E \gamma_i \gamma_i^T \).

9. Acknowledgements

Part of the result in this paper was circulated in a draft titled "Bootstrapping Semi-parametric Models with Single-Index Nuisance Parameters.” I thank Xiaohong Chen, Stefan Hoderlein, Simon Lee, Frank Schoerheide and seminar participants at the Greater New York Econometrics Colloquium at Princeton University for valuable comments for the earlier version of this paper. 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.
References

[1] Abrevaya, J. and J. Huang, 2005. On the bootstrap of the maximum score estimator. Econometrica 73, 1175-2204.
[2] Andrews, D. W. K. 1994. Empirical Process Methods in Econometrics. Handbook of Econometrics, Vol 4, ed. R. F. Engle and D. McFadden, 2247-2293.
[3] Chen, S. and S. Khan, 2003. Semiparametric estimation of a heteroskedastic sample selection model. Econometric Theory 19, 1040-1064.
[4] 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.
[5] Das, M., W. K. Newey, and F. Vella, 2003. Nonparametric estimation of sample selection models. Review of Economic Studies 70, 33-58.
[6] Dehejia, R. and S. Wahba, 1998. Propensity score matching methods for nonexperimental causal studies. NBER Working Paper No. 6829.
[7] Fan, Y. and Q. Li, 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865-890.
[8] Fix, E. and J. L. Hodges, 1951. Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4.
[9] Gallant, R. and D. Nychka, 1987. Semi-nonparametric maximum likelihood estimation, Econometrica 55, 363-390.
[10] Hahn, J. and G. Ridder, 2010. The asymptotic variance of semi-parametric estimators with generated regressors. Working paper.
[11] Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315-331.
[12] Härdle, W., P. Hall and H. Ichimura, 1993. Optimal semiparametric estimation in single index models. Annals of Statistics 21, 1, 157-178.
[13] Härdle, W., P. and A. B. Tsybacov, 1993. How sensitive are average derivatives. Journal of Econometrics 58, 31-48.
[14] Heckman, J. J., 1974. Shadow prices, market wages and labor supply,” Econometrica 42, 679-694.
[15] Heckman, J. J., H. Ichimura, J. Smith, H., and P. Todd, 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017-1098.
[16] 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.
[17] Heckman, J. J., H. Ichimura, and P. Todd, 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261-294.
[18] Hirano, K., G. Imbens, and G. Ridder, 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161-1189.
[19] Horowitz, J. L. 2009. Semiparametric and Nonparametric Methods in Econometrics. Springer-Verlag.
[20] 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.
[21] 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.
[22] Ichimura, H. 1993. Semiparametric least squares, SLS and weighted SLS estimation of single Index Models. Journal of Econometrics 58, 71-120.
[23] Kim, J. and D. Pollard, 1990. Cube root asymptotics. Annals of Statistics 18, 191-219.
[24] Klein, R. W. and R. H. Spady, 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 2, 387-421.
[25] Li, Q. and J. M. Wooldridge, 2002. Semiparametric estimation of partial linear models for dependent data with generated regressors. Econometric Theory 18, 625-645.
[26] Mammen, E., C. Rothe, and M. Schienle, 2012. Nonparametric regression with non-parametrically generated covariates. Annals of Statistics 40, 1132-1170.
[27] Manski, C. F., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3, 205-228.
[28] 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.
[29] Newey, W. K., Powell, J. and F. Vella, 1999. Nonparametric estimation of triangular simultaneous equation models. Econometrica 67, 565-603.
[30] Newey, W. K., Powell, J. and J. Walker, 1990. Semiparametric estimation of selection models: some empirical results. American Economic Review 80, 324-8.
[31] Pollard, D., 1989. A maximal inequality for sums of independent processes under a bracketing condition. Unpublished manuscript.
[32] 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.
[33] Powell, J., Stock, J. and T. Stoker, 1989. Semiparametric estimation of index coefficients. Econometrica 57, 6, 1403-1430.
[34] Rilstone, P., 1996. Nonparametric estimation of models with generated regressors. International Economic Review 37, 299-313.
[35] Robinson, P., 1987. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. Econometrica 55, 875-891.
[36] Robinson, P., 1988. Root-N consistent nonparametric regression. Econometrica 56, 931-954.
[37] Song, K., 2009. Testing conditional independence using Rosenblatt transforms. Annals of Statistics 37, 4011-4045.
[38] Song, K., 2012. On the smoothness of conditional expectation functionals. Statistics and Probability Letters 82, 1028-1034.
[39] Sperlich, S., 2009. A note on non-parametric estimation with predicted values. Econometrics Journal 12, 382-395.
[40] Stoker, T., 1986. Consistent estimation of scaled coefficients. Econometrica 54, 1461-1481.
[41] Stone, C. J., 1977. Consistent nonparametric regression. Annals of Statistics 5, 595-645.
[42] Stute, W., 1984. Asymptotic normality of nearest neighbor regression function estimates. Annals of Statistics 12, 917-926.
[43] Stute, W. and L. Zhu, 2005. Nonparametric checks for single-index models. Annals of Statistics 33, 1048-1083.
[44] Turki-Moalla, K., 1998. Rates of convergence and law of the iterated logarithm for U-processes. Journal of Theoretical Probability 11, 869-906.
[45] van der Vaart, A. W., 1996. New Donsker classes. Annals of Probability 24, 2128-2140.
[46] 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.