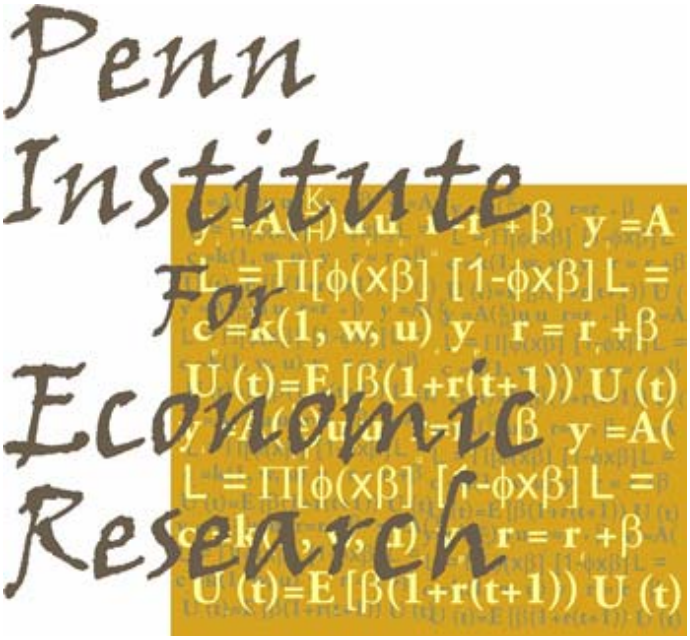


Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

# *PIER Working Paper 10-026*

"Bootstrapping Semiparametric Models with Single-Index Nuisance Parameters"
Second Version

by

Kyungchul Song

http://ssrn.com/abstract=1652202

# Bootstrapping Semiparametric Models with Single-Index Nuisance Parameters

Kyungchul Song[1]
*Department of Economics, University of Pennsylvania*

August 2, 2010

**Abstract**

This paper considers models of conditional moment restrictions that involve nonparametric functions of single-index nuisance parameters. This paper proposes a bootstrap method of constructing confidence sets which has the following three merits. First, the bootstrap is valid even when the single-index estimator follows cube-root asymptotics. Second, the bootstrap method accommodates conditional heteroskedasticity. Third, the bootstrap does not require re-estimation of the single-index component for each bootstrap sample. The method is built on this paper's general finding that as far as the single-index is a conditioning variable of a conditional expectation, the influence of the estimated single-indices in these models is asymptotically negligible. This finding is shown to have a generic nature through an analysis of Fréchet derivatives of linear functionals of conditional expectations. Some results from Monte Carlo simulations are presented and discussed.

*Keywords*: Semiparametric conditional moment restrictions; single-index restrictions; cube root asymptotics; bootstrap;

*JEL Classifications:* C12, C14, C51.

[1]A previous version of this paper was circulated under the title, "Two-Step Extremum Estimation with Estimated Single-Indices." I thank Xiaohong Chen, Stefan Hoderlein, Simon Lee, Frank Schorfheide and seminar participants at the Greater New York Econometrics Colloquium at Princeton University for valuable comments. I would also like to express my gratitude to two referees for comments that led to a substantial improvement of the paper. All errors are mine. Address correspondence to Kyungchul Song, Department of Economics, University of Pennsylvania, 528 McNeil Bldg, 3718 Locust Walk, Philadelphia, PA 19104-6297.

# 1 Introduction

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

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

This paper focuses on semiparametric conditional moment restrictions where the restrictions contain nonparametric functions of single-indices that are identified and estimated prior to the estimation of the parameter of interest. The restrictions allow the single-indices to follow cube-root asymptotics. Numerous examples belong to this class of restrictions. For example, a sample selection model where the selection equation error satisfies a conditional median restriction belongs to the framework of this paper. In such a situation, one may estimate the single-index in the selection equation using maximum score estimation. Other examples include models of single-index exogeneity, where the instrumental variable takes the form of a single-index that is to be estimated in the first step.

This paper considers two-step estimation, estimating the single-index component in the first step and then estimating the parameter of interest in the second step. Then the main concern is whether the first-step estimation error leaves its mark on the asymptotic distribution of the second step estimator. The analysis is typically based on the asymptotic linear

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

representation of estimated parameters. (See Newey (1994) for a systematic exposition regarding this analysis.) However, this approach does not apply when the first step parameter follows cube-root asymptotics, and as far as the author is concerned, there is no literature that formally studies this problem. Furthermore, when one attempts to make bootstrap-based inference, it is not clear what method of bootstrap will deliver the wanted result. As is well-known (Abrevaya and Huang (2005)), the method of bootstrap fails for estimators that follow cube-root asymptotics.

This paper proposes a bootstrap method for the parameters of interest in this situation. The method has three advantages. First, the bootstrap procedure is valid even when the single-index component follows cube-root asymptotics. This is interesting in the light of the result from Abrevaya and Huang (2005). This paper's result affirms that as far as the single-index is a nuisance parameter that is a conditioning variable of a conditional expectation, there is a valid bootstrap procedure for the parameter of interest even when the single-index estimator follows cube-root asymptotics. Second, the bootstrap method accommodates conditional heteroskedasticity. Note that conditional heteroskedascity is natural for models under conditional moment restrictions. Third, the bootstrap method does not require re-estimation of the single-index component or the nonparametric function for each bootstrap sample. Hence it is computationally attractive when the dimension of the single-index coefficient vector is large and its estimation involves numerical optimization. This is indeed the case when the single-index is estimated through maximum score estimation and the number of covariates is large. Therefore, the bootstrap method in this paper can be conveniently used for models that involve nonparametric estimators of cube-root converging single-indices.

The result of this paper is built on a general finding that when the single-index enters as a conditioning variable of a conditional expectation, the influence of the estimated single-index is asymptotically negligible even if it follows cube-root asymptotics. To place this phenomenon in the perspective of Newey (1994), this paper considers functionals that involve conditional expectations where the conditioning variable involves an unknown parameter. It is shown that in this situation, the first order Fréchet derivative of the functional with respect to the unknown parameter is zero. This means that there is no first order influence of the estimator in the conditioning variable on an estimator of any functional of the conditional expectation. This result may have interesting consequences in a broader context than that studied in this paper.

For the sake of concreteness, this paper establishes a uniform Bahadur representation of symmetrized nearest neighborhood (SNN) estimators over function spaces. Symmetrized nearest neighborhood estimators do not suffer from the random denominator problem and

hence do not require a trimming sequence. Based on the uniform representation result, this paper offers lower level conditions for the asymptotic theory of this paper. A Bahadur representation of SNN estimators was originally established by Stute and Zhu (2005) who established a non-uniform result in the context of testing single-index restrictions. In particular, Stute and Zhu (2005) showed that the first order effect of a $\sqrt{n}$-converging single-index estimator is asymptotically negligible. This paper puts their finding in the perspective of semiparametric estimation and shows that the phenomenon of the asymptotic negligibility of the estimated single-index arises even when the single-index component has a cube-root rate. The uniform Bahadur representation is also useful for many other purposes, for example, for analyzing various semiparametric specification tests.

There are many researches that study models with estimated regressors. 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 estimators that involve predicted regressors. While the last two papers are related to this paper, the set-up of this paper is different. The asymptotic behavior of the nonparametric estimator of the predicted regressors is not a major concern here because the nonparametric part is a nuisance parameter in this paper's set-up. The main concern is centered on the inference about the finite dimensional parameter of interest when the semiparametric nuisance parameter involves a nonparametric function and a single-index that potentially follows cube-root asymptotics.

The paper is organized as follows. In the next section, we define the scope of this paper by introducing models of semiparametric conditional moment restrictions and motivate the models with examples that are relevant in the literature. Section 3 proposes a new bootstrap-based inference method for the models and offers the main result that establishes the asymptotic validity of the bootstrap procedure under general conditions. Some heuristics behind the results are also provided. Section 4 investigates whether the proposed bootstrap procedure performs well in finite samples by using Monte Carlo simulations. Section 5 concludes. The Appendix introduces a general lemma about continuity of functionals of conditional expectations in parameters constituting the conditioning variable. The appendix also presents a general uniform Bahadur representation of SNN estimators which can be useful for other purposes.

# 2 Semiparametric Conditional Moment Restrictions

This paper focuses on the following form of semiparametric conditional moment restrictions. For $j = 1, \cdots, J+1$, let $\lambda_j(x) = \lambda_j(x; \theta_0)$, where $\lambda_j(\cdot)$ is a real function known up to $\theta_0 \in \mathbf{R}^{d_\Theta}$. For example, $\lambda_j(x; \theta_0) = x_j^\top \theta_{0,j}$, where $x_j$ and $\theta_{0,j}$ are conformable subvectors of $x$ and $\theta_0$. Another example is $\lambda_j(x; \theta_0) = \exp(x_j^\top \theta_{0,j})/\{1 + \exp(x_j^\top \theta_{0,j})\}$. Given observable i.i.d. random vectors $X_i \in \mathbf{R}^{d_X}$, $Y_i \in \mathbf{R}^J$, and observable i.i.d. binary random variables $D_i \in \{0, 1\}$, we define

$$\lambda_{i,j} = \lambda_j(X_i) \text{ and } \mu_{i,j} = \mathbf{E}\left[Y_{i,j} | \lambda_{i,j}, D_i = 1\right],$$

where $Y_{i,j}$ is the $j$-th entry of $Y_i$. Let $\mu_i = (\mu_{i,1}, \cdots, \mu_{i,J})^\top$. Then we assume that the parameter of interest $\beta_0 \in \mathbf{R}^{d_\beta}$ is identified through the following restriction:

$$\mathbf{E}\left[\rho(V_i, \mu_i; \beta_0) | D_i = 1, W_i\right] = 0, \tag{1}$$

where $W_i = (W_{1,i}, \lambda_{i,J+1})$, $(V_i, W_{1,i}) \in \mathbf{R}^{d_V + d_{W_1}}$ is an observable random vector and $\rho(\cdot, \cdot; \beta_0) : \mathbf{R}^{d_V + J} \to \mathbf{R}$ is known up to $\beta_0 \in B \subset \mathbf{R}^{d_\beta}$. Throughout this paper, we assume that $\theta_0$ is identified before one imposes the conditional moment restriction in (1). Hence it suffices that the restriction in (1) identifies the parameter $\beta_0$ only. The function $\rho(\cdot, \cdot; \beta_0)$ is called the *generalized residual function* which is a generalized version of the residual from the linear regression models. The random variable $\lambda_{i,j} : \mathbf{R}^{d_X} \to \mathbf{R}$ is a single-index of $X_i$, and the distributions of $\lambda_{i,j}$'s are assumed to be absolutely continuous with respect to the Lebesgue measure.

This paper's situation is such that the parameter of main interest is $\beta_0$ and the parameter $\theta_0$ in the single-index is a nuisance parameter. The primary focus of this paper is on the inference of $\beta_0$ when $\theta_0$ is estimated at the rate of $n^{1/2}$ or $n^{1/3}$. Note that $W_i$ is allowed to depend on an unknown continuous single index $\lambda_{i,J+1}$. This feature is relevant when the IV exogeneity takes the form of *single-index exogeneity*, where the instrumental variable takes the form of a single-index.

**Example 1 (Sample Selection Model with a Median Restriction) :** Consider the following model:

$$\begin{aligned} Y_i &= \beta_0^\top W_{1,i} + v_i \text{ and} \\ D_i &= 1\{\lambda_i \geq \varepsilon_i\}, \end{aligned}$$

where $\lambda_i = X_i^\top \theta_0$. The variable $Y_i$ denotes the latent outcome and $W_{1,i}$ a vector of covariates

that affect the outcome. The binary $D_i$ represents the selection of the vector $(Y_i, W_{1,i})$ into the observed data set, so that $(Y_i, W_{1,i})$ is observed only when $D_i = 1$. The incidence of selection is governed by a single index $\lambda_i$ of covariates $X_i$. The variables $v_i$ and $\varepsilon_i$ represent unobserved heterogeneity in the individual observation. The exclusion restriction here requires that $W_{1,i}$ is not measurable with respect to the $\sigma$-field generated by $\lambda_i$.

The variable $\varepsilon_i$ is permitted to be correlated with $X_i$ but $Med(\varepsilon_i|X_i) = 0$. And $W_{1,i}$ is independent of $(v_i, \varepsilon_i)$ conditional on the index $\lambda_i$ in the selection mechanism. This involves the median restriction and the single-index exogeneity. The assumptions of the model are certainly weaker than the common requirement that $(W_{1,i}, X_i)$ be independent of $(v_i, \varepsilon_i)$. (e.g. Heckman (1990), Newey, Powell, and Walker (1990).) More importantly, this model does not assume that $X_i$ is independent of $\varepsilon_i$ in the selection equation or of $v_i$ in the outcome equation. Hence we cannot use the characterization of the selection bias through the propensity score $P\{D_i = 1|\lambda_i\}$ as has often been done in the literature of semiparametric extension of the sample selection model. (e.g. Powell (1989), Ahn and Powell (1993), Chen and Khan (2003), and Das, Newey and Vella (2003)).

From the method of Robinson (1988), the identification of $\beta_0$ still follows if the matrix

$$\mathbf{E}\left[(X_i - \mathbf{E}[X_i|D_i = 1, \lambda_i])(X_i - \mathbf{E}[X_i|D_i = 1, \lambda_i])^\top | D_i = 1\right]$$

is positive definite. In this case, we can write for the observed data set $(D_i = 1)$

$$Y_i = \beta_0^\top W_{1,i} + \tau(\lambda_i) + u_i,$$

where $u_i$ satisfies that $\mathbf{E}[u_i|D_i = 1, W_{1,i}, \lambda_i] = 0$ and $\tau$ is an unknown nonparametric function. This model can be estimated by using the method of Robinson (1988). Let $\mu_{Y,i} = \mathbf{E}[Y_i|D_i = 1, \lambda_i]$, and $\mu_{W_1,i} = \mathbf{E}[W_{1,i}|D_i = 1, \lambda_i]$. Then, we consider a conditional moment restriction:

$$\mathbf{E}\left[\{Y_i - \mu_{Y,i}\} - \beta_0^\top \{W_{1,i} - \mu_{W_1,i}\} | D_i = 1, W_{1,i}, \lambda_i\right] = 0.$$

By putting

$$\rho(V_i, \mu_i; \beta_0) = \{Y_i - \mu_{Y,i}\} - \beta_0^\top \{W_{1,i} - \mu_{W_1,i}\})$$

and $W_i = (W_{1,i}^\top, \lambda_i)^\top$, we find that this model belongs to the model of semiparametric conditional moment restrictions.

One may estimate $\theta_0$ in $\lambda_0$ using maximum score estimation in the first step and use it in the second step estimation of $\beta_0$. Then the remaining question is concerned with the effect of the first step estimator of $\theta_0$ which follows cube root asymptotics upon the estimator of $\beta_0$.

Note that the identification of $\theta_0$ does not stem from a direct imposition of single-index restrictions on $\mathbf{E}[Y_i|D_i = 1, X_i = \cdot]$ and $\mathbf{E}[Z_i|D_i = 1, X_i = \cdot]$. The identification follows from the use of auxiliary data set $((D_i = 0), X_i)$ in the sense of Chen, Hong, and Tarozzi (2008). Such a model of "single-index selectivity bias" has a merit of avoiding a strong exclusion restriction and has early precedents. See Powell (1989), Newey, Powell, and Walker (1990), and Ahn and Powell (1993). ■

**Example 2 (Models with a Single-Index Instrumental Variable) :** Consider the following model:

$$\begin{aligned} Y_i &= Z_i^\top \beta_0 + \varepsilon_i, \text{ and} \\ D_i &= 1\{\lambda_i \geq \eta_i\}, \end{aligned}$$

where $\lambda_i = X_i^\top \theta_0$ and $\varepsilon_i$ and $\eta_i$ satisfy that $\mathbf{E}[\varepsilon_i|\lambda_i] = 0$ and $Med(\eta_i|X_i) = 0$. The data set $(D_i, X_i)$ plays the role of an auxiliary data set in Chen, Hong, and Tarozzi (2008) and enables us to identify the single-index $\lambda_i$ that plays the role of the instrumental variable (IV). However, the IV exogeneity condition is weaker than the conventional one because the exogeneity is required only of the single-index $X_i^\top \theta_0$ not of the whole vector $X_i$. In other words, some of the elements of the vector $X_i$ are allowed to be correlated with $\varepsilon_i$. Furthermore, $X_i$ is not required to be independent of $\eta_i$ as long as it maintains the conditional median restriction. This conditional median restriction enables one to identify $\theta_0$ and in consequence $\beta_0$.

We consider the following conditional moment restriction:

$$\mathbf{E}\left[Y_i - Z_i^\top \beta_0 | \lambda_i\right] = 0.$$

In this case, $\rho(V_i, \mu_i; \beta_0) = Y_i - Z_i^\top \beta_0$ and $W_i = \lambda_i$. Hence there is no nonparametric component $\mu$ in the generalized residual function.

We can first estimate $\lambda_i$ and then estimate $\beta_0$ by plugging in these estimates into a sample version of the conditional moment restriction. Again, when $\theta_0$ is estimated using maximum score estimation, the main question is how we can analyze the estimator's effect on the estimation of $\beta_0$. ■

# 3 Inference

## 3.1 Estimators and Asymptotic Distributions

This paper considers a two-step procedure where one estimates the single-index parameter $\theta_0$ first, and using this estimator, estimates $\beta_0$ in the second step. Suppose that we have obtained a consistent estimator $\hat{\theta}$ of $\theta$. For this, one may use estimation methods in the literature of single-index restrictions (e.g. Ichimura (1993), Hristache, Juditsky and Spokoiny (2001).) When the single-index is involved in a selection equation with a conditional median restriction, one may obtain $\hat{\theta}$ through maximum score estimation. All we require for our purpose is that the rate of convergence of the estimator $\hat{\theta}$ is either $n^{-1/2}$ or $n^{-1/3}$ (Assumption 2 below).

Given the estimator $\hat{\theta}$, we let $\hat{\lambda}_{i,j} = \lambda_j(X_i; \hat{\theta})$. As for $\mu$, this paper considers symmetrized nearest neighborhood (SNN) estimation. Let $\hat{u}_{k,j} = \frac{1}{n}\sum_{i=1}^n 1\{\hat{\lambda}_{i,j} \leq \hat{\lambda}_{k,j}\}$ and $\hat{\mu}_k = [\hat{\mu}_{k,1}, \cdots, \hat{\mu}_{k,J}]^\top$, where

$$\hat{\mu}_{k,j} = \frac{\sum_{i=1}^n D_i Y_{i,j} K_h\left(\hat{u}_{i,j} - \hat{u}_{k,j}\right)}{\sum_{i=1}^n D_i K_h\left(\hat{u}_{i,j} - \hat{u}_{k,j}\right)}, \tag{2}$$

and $K_h(u) = K(u/h)/h$ and $K : \mathbf{R} \to \mathbf{R}$ is a kernel function. The estimator $\hat{\mu}_{k,j}$ is a SNN estimator proposed by Yang (1981) and studied by Stute (1984). The probability integral transform of $\lambda_{i,j}$ turns its density into a uniform density on $[0, 1]$. (Recall that we assume that the distribution of $\lambda_{i,j}$ is absolutely continuous throughout this paper.) Using the probability integral transform obviates the need to introduce a trimming sequence. The trimming sequence is often required to deal with the random denominator problem (e.g. Ichimura (1993) and Klein and Spady (1993)), but there is not much practical guidance for its choice. The use of the probability integral transform eliminates such a nuisance altogether.

We introduce an estimator of $\beta_0$. For any vectors $x$ and $y$ in $\mathbf{R}^{d_W}$, we write $x \leq y$ to mean that $x_j \leq y_j$ for all $j = 1, \cdots, d_W$, where $x_j$'s and $y_j$'s are entries of $x$ and $y$ respectively. We define

$$\hat{\beta} = \underset{\beta \in B}{\operatorname{argmin}} \sum_{k=1}^n D_k \left\{ \sum_{i=1}^n D_i \rho(V_i, \hat{\mu}_i; \beta) 1\{\hat{W}_i \leq \hat{W}_k\} \right\}^2,$$

where $\hat{W}_i = (W_{1,i}, \hat{\lambda}_{i,J+1})$. The estimation method is similar to the proposal by Domínguez and Lobato (2004). While they considered weakly dependent observations in contrast to the i.i.d. set-up of this paper, their model does not involve single-index components that are estimated in the first step. Let $\Theta(\delta) \equiv \{\theta \in \mathbf{R}^{d_\Theta} : ||\theta - \theta_0|| < \delta\}$.

**Assumption 1 :** (i) $\{(V_i, X_i, Y_i, W_i, D_i)\}_{i=1}^n$ is a random sample.
(ii) $\mathbf{E}[\rho(V_i, \mu_i; \beta) D_i | W_i] = 0$ a.s. iff $\beta = \beta_0$ and $\beta_0$ belongs to the interior of a compact set

$B$.
(iii) $\rho(v,\mu;\beta)$ as a function of $(\beta,\mu) \in B \times \mathbf{R}^J$ is twice continuously differentiable with the first order derivatives $\rho_\beta$ and $\rho_\mu$ and the second order derivatives $\rho_{\beta\beta}$, $\rho_{\beta\mu}$ and $\rho_{\mu\mu}$ such that $\mathbf{E}[\sup_{\beta\in B}||\tilde{\rho}(V_i,\mu_i;\beta)||^p] < \infty$, $p > 2$, for all $\tilde{\rho} \in \{\rho, \rho_\beta, \rho_\mu, \rho_{\beta\beta}, \rho_{\beta\mu}\}$.
(iv) For some $M > 0$ and $p > 8$, $\mathbf{E}[||Y_i||^p] < M$, $\mathbf{E}[||\rho_\mu(V_i,\mu_i;\beta_0)||^p] < M$, and

$$\mathbf{E}[\sup_{(\beta,\bar{\mu})\in B\times[-M,M]}||\rho_{\mu\mu}(V_i,\bar{\mu};\beta)||^q] < \infty, \ q > 8. \tag{3}$$

**Assumption 2 :** The estimator $\hat{\theta}$ satisfies that $||\hat{\theta} - \theta_0|| = O_P(n^{-r})$ with $r = 1/2$ or $1/3$.

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

Assumption 1 is standard in many models of conditional moment restrictions. The condition $\mathbf{E}[||\rho_\mu(V_i,\mu_i;\beta_0)||^p] < M$ and (3) in Assumption 1(iv) are trivially satisfied when $\rho(v,\mu;\beta)$ is linear in $\mu$ as in Examples 1 and 2. Assumption 3(i) is satisfied, for example, by a quartic kernel: $K(u) = (15/16)(1-u^2)^2 1\{|u| \leq 1\}$. The bandwidth condition in Assumption 3(ii) does not require undersmoothing; it is satisfied by any $h = n^{-s}$ with $q/(6q-2) < s < q/(4q+2)$. There are other assumptions that are of more technical character. These assumptions (named Assumption A) and discussions are found in the appendix.

**Theorem 1 :** *Suppose that* Assumptions 1-3 *and* Assumption A (in the Appendix) *hold. Then,*

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d \left(\int \dot{H}(w)\dot{H}(w)^\top dF_{W|D=1}(w)\right)^{-1} \int \dot{H}(w)\zeta(w)dF_{W|D=1}(w),$$

*where* $\dot{H}(w) = \mathbf{E}[\rho_\beta(V_i,\mu_i;\beta_0)D_i 1\{W_i \leq w\}]$, $F_{W|D=1}$ *is the conditional CDF of* $W_i$ *given* $D_i = 1$, $\zeta$ *is a centered Gaussian process on* $\mathbf{R}^{d_W}$ *that has a covariance kernel given by* $C(w_1,w_2) = \mathbf{E}\left[\xi_i(w_1)\xi_i(w_2)D_i\right]$ *with* $\xi_i(w) = \rho(V_i,\mu_i;\beta_0)1\{W_i \leq w\} - r_i(w)$,

$$r_i(w) = \sum_{j=1}^{J} \mathbf{E}[1\{W_i \leq w\}\rho_{\mu,j}(V_i,\mu_i;\beta_0)|\lambda_{i,j}, D_i = 1]\left(Y_{i,j} - \mu_{i,j}\right) \tag{4}$$

*and* $\rho_{\mu,j}(V_i,\mu_i;\beta_0)$ *is the* $j$*-th entry of* $\rho_\mu(V_i,\mu_i;\beta_0)$.

Compared with the asymptotic covariance matrix of Domínguez and Lobato (2004), the asymptotic covariance matrix contains additional terms $r_i(w)$. This is due to the nonparametric estimation error in $\hat{\mu}$. The asymptotic covariance matrix remains the same regardless

of whether we use the estimated indices $\hat{\lambda}_{i,j}$ or the true indices $\lambda_{i,j}$. This is true even if $\hat{\theta}$ follows cube root asymptotics. The following subsection offers heuristic arguments behind this phenomenon.

## 3.2 Some Heuristics

For simplicity, assume that $\lambda(X_i;\theta) = X_i^\top\theta$, $D_i = 1$ for all $i = 1, \cdots, n$, and the generalized residual takes the form of

$$\rho(V_i, \mu_i; \beta_0) = \beta_0 - \mathbf{E}\left[Y_i | X_i^\top \theta_0\right],$$

where $\beta_0 \in \mathbf{R}$. Furthermore, we assume that the moment condition

$$\mathbf{E}\left[\rho(V_i, \mu_i; \beta_0) Z_i\right] = 0$$

identifies $\beta_0$ for a certain instrumental variable $Z_i$, where we normalize $\mathbf{E}Z_i = 1$. Then $\beta_0$ is identified as $\beta_0 = \Gamma(\theta_0)$ where

$$\Gamma(\theta) = \mathbf{E}\left[\mathbf{E}\left[Y_i | X_i^\top\theta\right] Z_i\right].$$

The first order effect of the estimation of $\theta_0$ on that of $\beta_0$ is determined by the way $\Gamma(\theta)$ behaves as we perturb $\theta$ around $\theta_0$. (e.g. See Newey (1994).)

Under certain regularity conditions for the conditional density of $Y_i$ given $X_i^\top\theta$, we can show that (see the appendix for details)

$$|\Gamma(\theta_1) - \Gamma(\theta_2)| = O(||\theta_1 - \theta_2||^2). \tag{5}$$

In other words, $\Gamma(\theta)$ is fairly insensitive to the perturbation in $\theta$. (Note that the order is not $O(||\theta_1 - \theta_2||)$ but $O(||\theta_1 - \theta_2||^2)$.) Roughly speaking, when $\hat{\theta}$ is within a $n^{-1/3}$-neighborhood of $\theta_0$, $\Gamma(\hat{\theta})$ is within a $n^{-2/3}$-neighborhood of $\theta_0$. This means that $\sqrt{n}(\Gamma(\hat{\theta}) - \Gamma(\theta_0)) \to_P 0$, even if $\hat{\theta}$ has the cube-root convergence rate. Therefore, there is no estimation error effect from $\hat{\theta}$.

The result in (5) can be seen intuitively as follows. To simplify the notations, we write $\Lambda_{1,i} = X_i^\top\theta_1$ and $\Lambda_{2,i} = X_i^\top\theta_2$. First, using the law of iterated conditional expectations,

$$\begin{aligned}
\Gamma(\theta_1) - \Gamma(\theta_2) &= \mathbf{E}\left[Z_i \left\{\mathbf{E}\left[Y_i|\Lambda_{1,i}\right] - \mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right\}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[Z_i|\Lambda_{1,i}, \Lambda_{2,i}\right] \left\{\mathbf{E}\left[Y_i|\Lambda_{1,i}\right] - \mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right\}\right]
\end{aligned}$$

By adding and subtracting terms, we rewrite the above as

$$\begin{aligned}
&\mathbf{E}\left[\left(\mathbf{E}\left[Z_i|\Lambda_{1,i},\Lambda_{2,i}\right]-\mathbf{E}\left[Z_i|\Lambda_{2,i}\right]\right)\left(\mathbf{E}\left[Y_i|\Lambda_{1,i}\right]-\mathbf{E}\left[Y_i|\Lambda_{1,i},\Lambda_{2,i}\right]\right)\right] \\
&+\mathbf{E}\left[\left(\mathbf{E}\left[Z_i|\Lambda_{1,i},\Lambda_{2,i}\right]-\mathbf{E}\left[Z_i|\Lambda_{2,i}\right]\right)\left(\mathbf{E}\left[Y_i|\Lambda_{1,i},\Lambda_{2}\right]-\mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right)\right] \\
&+\mathbf{E}\left[\mathbf{E}\left[Z_i|\Lambda_{2,i}\right]\left\{\mathbf{E}\left[Y_i|\Lambda_{1,i}\right]-\mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right\}\right].
\end{aligned} \tag{6}$$

The last expectation is equal to

$$\mathbf{E}\left[\left\{\mathbf{E}\left[Z_i|\Lambda_{2,i}\right]-\mathbf{E}\left[Z_i|\Lambda_{1,i},\Lambda_{2,i}\right]\right\}\left\{\mathbf{E}\left[Y_i|\Lambda_{1,i}\right]-\mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right\}\right]$$

because $\mathbf{E}\left[\mathbf{E}\left[Z_i|\Lambda_{1,i},\Lambda_{2,i}\right]\left\{\mathbf{E}\left[Y_i|\Lambda_{1,i}\right]-\mathbf{E}\left[Y_i|\Lambda_{2,i}\right]\right\}\right]=0$. Hence if for $S_i=Y_i$ or $Z_i$,

$$\begin{aligned}
\mathbf{E}\left[S_i|\Lambda_{1,i}\right]-\mathbf{E}\left[S|\Lambda_{1,i},\Lambda_{2,i}\right] &\approx O(||\theta_1-\theta_2||) \text{ and} \\
\mathbf{E}\left[S_i|\Lambda_{1,i}\right]-\mathbf{E}\left[S_i|\Lambda_{2,i}\right] &\approx O(||\theta_1-\theta_2||),
\end{aligned} \tag{7}$$

all the components in the sum of (6) are $O(||\theta_1-\theta_2||^2)$. Therefore $\Gamma(\theta)$ is insensitive to the first order perturbation of $\theta$. This analysis carries over even when $\lambda$ is an infinite dimensional parameter taking values in a function space, say, $\Lambda$, as long as certain regularity conditions for conditional densities are maintained. A detailed version of this result is presented in the appendix.

It should be noted that the asymptotic negligibility result relies on the particular structure where the single-index $\lambda_i$ (here $\lambda_i=X_i^\top\theta$) enters as a conditioning variable of a conditional expectation. For example, Ahn and Powell (1993) and Chen and Khan (2003) use generated regressors to estimate the main parameter of interest. In their cases, the generated regressors do not enter as a conditioning variable of a conditional expectation, but enter as part of a weighting matrix. Hence the phenomenon of asymptotic negligibility of the generated regressor does not arise. Another example that is worth attention is the case where one employs density weighting in the estimation using the density of the single-index. In this case, the asymptotic negligibility of the estimated single-index does not arise either. For instance, the model of Li and Wooldridge (2002) involves a generated regressor as a conditioning variable of conditional expectation, and as shown in Theorem 2.1 in their paper, there exists a first order effect of generated regressors in the asymptotic theory. This result appears to stand in contradiction to the result of this paper. To see this closely, observe that Li and Wooldridge (2002) considers the following partial linear model (Eq. (4) on page 627):

$$Y_i=X_i^\top\gamma+m(\eta_i)+u_i$$

where $m$ is an unknown function and $\eta_i = S_i - Z_i^\top \alpha$ with $\alpha$ being an unknown parameter. The parameter of interest is $\gamma$. Following Robinson (1988) and applying density weighting as in Powell, Stock and Stoker (1988), Li and Wooldridge estimate $\gamma$ based on the following identification strategy:

$$\gamma = \mathbf{E}\left[(X_i - \mathbf{E}(X_i|\eta_i))(X_i - \mathbf{E}(X_i|\eta_i))^\top f^2(\eta_i)\right]^{-1} \mathbf{E}\left[(Y_i - \mathbf{E}(Y_i|\eta_i))(X_i - \mathbf{E}(X_i|\eta_i))^\top f^2(\eta_i)\right],$$

where $f$ denotes the density of $\eta_i$. The asymptotic variance of their least squares estimator of $\gamma$ involves an additional term due to the use of $\hat{\eta}_i = S_i - Z_i^\top \hat{\alpha}$ in place of $\eta_i$. Precisely speaking, this additional term stems from the use of density weighting. The density weighting makes $\gamma$ depend on the variable $\eta_i$ outside the conditional expectations $\mathbf{E}(X_i|\eta_i)$ and $\mathbf{E}(Y_i|\eta_i)$. One can show that this additional term disappears when one takes the density weighting $f$ to be a constant 1.

## 3.3 Bootstrap Procedure

While one can construct confidence sets for $\beta_0$ based on the asymptotic theory, the estimation of the asymptotic covariance matrix is complicated. requiring a choice of multiple bandwidths. This paper proposes a bootstrap method that is easy to use and robust to conditional heteroskedasticity. The proposal is based on the wild bootstrap of Wu (1986). (See also Liu (1988).)

First, we find a consistent estimator $\hat{r}_i(w)$ of $r_i(w)$ defined in Theorem 1. As for the estimator $\hat{r}_i(w)$, we assume the following:

**Assumption 4 :** $\sup_{w \in \mathbf{R}^{d_W}} \max_{1 \le i \le n} |\hat{r}_i(w) - r(w)| = o_P(1)$.

Conditions for the uniform consistency of a nonparametric estimator is well-known in the literature (e.g. Hansen (2008)). Then, define $\hat{r}_{ik} = \hat{r}_i(\hat{W}_k)$ and

$$\hat{\rho}_{lk}(\beta) = 1\{\hat{W}_l \le \hat{W}_k\}\rho(V_l, \hat{\mu}_l; \beta),$$

where $\hat{\mu}_i$ is a first step estimator defined in (2). This paper suggests the following bootstrap procedure.

**Step 1 :** For $b = 1, \cdots, B$, draw i.i.d. $\{\omega_{i,b}\}_{i=1}^n$ from a two-point distribution assigning masses $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}-1)/(2\sqrt{5})$ to the points $-(\sqrt{5}-1)/2$ and $(\sqrt{5}+1)/2$.

**Step 2 :** Compute $\{\hat{\beta}_b^* : b = 1, \cdots, B\}$ by

$$\hat{\beta}_b^* = \underset{\beta \in B}{\operatorname{argmin}} \sum_{k=1}^{n} D_k \left\{ \sum_{l=1}^{n} D_l \left[ \left( \hat{\rho}_{lk}(\hat{\beta}) - \hat{\rho}_{lk}(\beta) \right) + \omega_{l,b} \left\{ \hat{\rho}_{lk}(\hat{\beta}) + \hat{r}_{lk} \right\} \right] \right\}^2$$

and use the bootstrap distribution of $\sqrt{n}(\hat{\beta}_b^* - \hat{\beta})$ in place of the finite sample distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ for inferences.

The bootstrap procedure is computationally very simple. The estimator $\hat{\mu}_i$ is stored once and repeatedly used for each bootstrap sample. In other words, we do not have to re-estimate $\theta_0$ for each bootstrap sample. This computational merit is prominent when the dimension of the parameter $\theta_0$ is large and one has to resort to a numerical optimization algorithm for its estimation as in the case of maximum score estimation.

The bootstrap procedure is a modified version of a wild bootstrap procedure which is typically used in the context of semiparametric specification tests (e.g. Härdle and Mammen (1993), Whang (2000), Delgado and González Manteiga (2001), Song (2009).) The main modification of the procedure is that it includes the additional term $\hat{r}_{lk}$ in the bootstrap procedure. This inclusion is made to induce the first order estimation error effect of $\hat{\mu}$ in the bootstrap estimation problem. If there were a further estimation error effect from $\hat{\lambda}$, we would have to induce this further effect in the bootstrap to ascertain validity of the bootstrap procedure. When $\hat{\lambda}$ follows cube-root asymptotic theory, it is not clear how one can accomplish this. Now, since the result of Theorem 1 has established that there is no estimation effect from $\hat{\lambda}$ even if it follows cube-root asymptotics, we do not need to induce the estimation error effect in the bootstrap as far as bootstrap validity is concerned. This is the main reason why the bootstrap still works even if it has an estimator of the nuisance parameter that converges at the rate of $n^{-1/3}$. Following the conventional notations, we denote $\rightarrow_{d^*}$ to indicate the convergence of bootstrap distributions conditional on $\{(V_i, X_i, Y_i, W_i, D_i)\}_{i=1}^n$.

**Theorem 2 :** *Suppose that* Assumptions 1-3 *and* Assumption A (in the Appendix) *hold. Then,*

$$\sqrt{n}(\hat{\beta}_b^* - \hat{\beta}) \rightarrow_{d^*} \left( \int \dot{H}(w)\dot{H}(w)^\top dF_{W|D=1}(w) \right)^{-1} \int \dot{H}(w)\zeta(w) dF_{W|D=1}(w) \text{ in } P$$

*where* $\dot{H}$ *and* $\zeta$ *are as in* Theorem 1.

Theorem 2 shows that the bootstrap procedure is asymptotically valid. As we explained above, the main reason that this bootstrap procedure works is due to the fact that there

is no first order estimation effect from $\hat{\theta}$. It is expected that the same phenomenon will carry over to the situation where the observations are weakly dependent, or even where the function $\lambda(\cdot)$ is a nonparametric function. In fact, the results of Theorems 1 and 2 stem from the result of continuity of functionals of conditional expectations. (See Section 3.2 above and Section 6.1 below in the Appendix.) This continuity result does not rely on the i.i.d. assumption of the observations. Furthermore, the result is established in a general set-up where $\lambda$ is a nonparametric function. A full development in these extensions is left to a future research.

In the following we revisit the two examples that we discussed before and see how the bootstrap procedure applies.

**Example 1 (Continued):** Let $e_i = Y_i - \mu_{Y,i} - \beta_0^\top \left(W_{1,i} - \mu_{W_1,i}\right)$ and $\mu_i = [\mu_{Y,i}, \mu_{W_1,i}]^\top$. After some algebra, we find that $r_i(w)$ defined in (4) is equal to $-F_i(w) \cdot e_i$, where $F_i(w) = \mathbf{E}\left[D_i 1\{W_i \leq w\} | \lambda_i\right]$. We construct estimator $\hat{e}_i$ of $e_i$ by using estimators $\hat{\mu}_{Y,i}$, $\hat{\mu}_{W_1,i}$ as in (2) and $\hat{\beta}$, and define

$$\hat{F}_i(w) = \frac{\sum_{j=1}^n D_j 1\{W_j \leq w\} K_h\left(\hat{u}_j - \hat{u}_i\right)}{\sum_{j=1}^n D_j K_h\left(\hat{u}_j - \hat{u}_i\right)},$$

where $\hat{u}_i = \frac{1}{n}\sum_{k=1}^n 1\{\hat{\lambda}_k \leq \hat{\lambda}_i\}$. Finally, let

$$A_{ik} = \hat{e}_i \cdot \left(1\{\hat{W}_i \leq \hat{W}_k\} + \omega_{i,b}\left(1\{\hat{W}_i \leq \hat{W}_k\} - \hat{F}_i(w)\right)\right).$$

Then the bootstrap version of the estimator $\hat{\beta}$ is defined as

$$\hat{\beta}_b^* = \underset{\beta \in B}{\operatorname{argmin}} \sum_{k=1}^n D_k \left\{ \sum_{l=1}^n D_l T_{Y,lk} - \beta^\top \left( \sum_{l=1}^n D_l T_{X,lk} \right) \right\}^2,$$

where

$$T_{Y,lk} = \{A_{lk} - \{Y_l - \hat{\mu}_{Y,l}\}\} D_l \text{ and } T_{X,lk} = (W_{1,l} - \hat{\mu}_{W_1,l}) D_l.$$

Since the form is least squares estimation, the solution $\hat{\beta}_b^*$ is explicit as follows. Let $T_X$ be the $n \times d_{W_1}$ vector whose $k$-th row is given by $\sum_{l=1}^n T_{X,lk}^\top$ and let $T_Y$ be the $n \times 1$ vector whose $k$-th entry is given by $\sum_{l=1}^n T_{Y,lk}$. Then,

$$\hat{\beta}_b^* = \left(T_X^\top T_X\right)^{-1} T_X^\top T_Y.$$

Note that for each $b = 1, \cdots, B$, it suffices to use the same estimators, $\hat{\mu}_{Y,i}$, $\hat{\mu}_{W_1,i}$, and only change $\omega_{i,b}$ in the definition of $A_{ik}$.

**Example 2 (Continued):** In this example, as for $r_i(w)$ defined in Theorem 1, $r_i(w) = 0$.

Hence let $\hat{\theta}$ be the maximum score estimation of $\theta_0$ and $\hat{W}_k = X_k^\top \hat{\theta}$. We define

$$\hat{\beta}_b^* = \underset{\beta \in B}{\operatorname{argmin}} \sum_{k=1}^{n} \left\{ \left[ \sum_{l=1}^{n} T_{Y,lk} - \beta^\top \left( \sum_{l=1}^{n} T_{X,lk} \right) \right] \right\}^2,$$

where

$$\begin{aligned} T_{Y,lk} &= 1\{\hat{W}_l \leq \hat{W}_k\}\{\hat{\beta}^\top Z_l + \omega_{l,b}(\{Y_l - Z_l^\top \hat{\beta}\})\} \text{ and} \\ T_{X,lk} &= Z_l 1\{\hat{W}_l \leq \hat{W}_k\}. \end{aligned}$$

Then, the solution is explicit as $\hat{\beta}_b^* = \left(T_X^\top T_X\right)^{-1} T_X^\top T_Y$ similarly as before when we define $T_X$ be the $n \times d_Z$ vector whose $k$-th row is given by $\sum_{l=1}^{n} T_{X,lk}^\top$ and let $T_Y$ be the $n \times 1$ vector whose $k$-th entry is given by $\sum_{l=1}^{n} T_{Y,lk}$.

# 4 A Monte Carlo Simulation Study

## 4.1 The Performance of the Estimator

In this section, we present and discuss some Monte Carlo simulation results. Based on the sample selection model in Example 1, we consider the following data generating process. Let

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

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

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

where $\varepsilon_i$ follows the distribution of $2T_i \times \frac{1}{d_X} \sum_{k=1}^{d_X} \Phi\left(X_{ik}^2 + |X_{ik}|\right) + \zeta_i$, $\zeta_i \sim N(0,1)$, $\Phi$ denoting the standard normal distribution function, and $T_i$ is chosen as follows:

DGP A1: $T_i \sim N(0,1)$ or
DGP A2: $T_i \sim t$ distribution with degree of freedom 1.

Hence the selection mechanism has errors that are conditionally heteroskedastic, and in the case of DGP A2, heavy tailed. Then, we define the latent outcome $Y_i^*$ as follows:

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

where $v_i \sim (a\zeta_i + e_i) \times \Phi\left(Z_i^2 + |Z_i|\right)$ with $e_i \sim N(0,1)$. 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. The degree of the sample selection bias varies depending on the choice of $a$. This simulation study considered $a \in \{1, 2\}$. We set $\theta_0$ to be the vector of 2's and $\beta_0 = 2$. In the simulation studies we estimated $\theta_0$ by using the maximum score estimation to obtain $\hat{\theta}$.

There are four combinations, depending on whether $\theta_0$ is assumed to be known (TR) or estimated through maximum score estimation (ES) and depending on whether SNN estimation was used (NN) or usual kernel estimation was used (KN). For the latter case, we used the standard normal PDF as a kernel. Bandwidths for the estimation of $\mathbf{E}[Y_i|X_i^\top\theta_0, D_i = 1]$ and $\mathbf{E}[Z_i|X_i^\top\theta_0, D_i = 1]$ were chosen separately using a least-squares cross-validation method. If the role of the sample selection bias were already marginal, the estimation error effect of $\hat{\theta}$ would be small accordingly, preventing us from discerning the negligibility of the estimation error effect of $\hat{\theta}$ from the negligible sample selection bias. Hence, we also report the results from the estimation of $\beta$ that ignores the sample selection bias (W-SBC: Without Sample Selection Bias Correction).

Table 1 shows the performance of the estimators. The results show that the performance of the estimators does not change significantly as we increase the number of covariates from 3 to 6. This indicates that the quality of the second step estimator $\hat{\beta}$ is robust to the quality of the first step estimator $\hat{\theta}$. This fact is shown more clearly when we compare the performance of the estimator (TR) that uses $\theta_0$ and the estimator (ES) that uses $\hat{\theta}$. The performance does not show much difference between these two estimators. The performance of the SNN estimator appears slightly better than the kernel estimator. When the sample size was increased from 200 to 500, the estimator's performance improved as expected. In particular the improvement in terms of RMSE is conspicuous.

The negligibility of the effect of the estimation error in $\hat{\theta}$ is not due to inherently weak sample selection bias. This is evident when we compare the results with those from the estimators that ignore the sample selection bias (W-SBC). Comparing Table 1 with Table 2, we observe that the sample selection bias increases when we enhance the correlation between $\varepsilon_i$ and $v_i$ by increasing $a = 1$ to $a = 2$. Nevertheless, the difference between the performance of the estimators using $\theta_0$ and that of the estimators using $\hat{\theta}$ continues to be marginal.

Table 1: The Performance of the Estimators in Terms of MAE and RMSE: $a = 1$

| | | $k$ | | NN-TR | NN-ES | KN-TR | KN-ES | W-SBC |
|---|---|---|---|---|---|---|---|---|
| | | 3 | MAE | 0.4304 | 0.4329 | 0.4337 | 0.4414 | 0.6039 |
| | DGP A1 | | RMSE | 0.2967 | 0.2984 | 0.3014 | 0.3108 | 0.5764 |
| | | 6 | MAE | 0.4079 | 0.4084 | 0.4065 | 0.4201 | 0.5487 |
| $n = 200$ | | | RMSE | 0.2654 | 0.2678 | 0.2644 | 0.2820 | 0.4628 |
| | | 3 | MAE | 0.4439 | 0.4473 | 0.4443 | 0.4583 | 0.6067 |
| | DGP A2 | | RMSE | 0.3095 | 0.3144 | 0.3119 | 0.3285 | 0.5848 |
| | | 6 | MAE | 0.4176 | 0.4115 | 0.4254 | 0.4188 | 0.5483 |
| | | | RMSE | 0.2738 | 0.2681 | 0.2727 | 0.2756 | 0.4766 |
| | | 3 | MAE | 0.2709 | 0.2705 | 0.2764 | 0.2781 | 0.4395 |
| | DGP A1 | | RMSE | 0.1134 | 0.1128 | 0.1182 | 0.1192 | 0.2990 |
| | | 6 | MAE | 0.2553 | 0.2551 | 0.2566 | 0.2615 | 0.3586 |
| $n = 500$ | | | RMSE | 0.1039 | 0.1042 | 0.1050 | 0.1086 | 0.2026 |
| | | 3 | MAE | 0.2683 | 0.2676 | 0.2707 | 0.2739 | 0.4482 |
| | DGP A2 | | RMSE | 0.1150 | 0.1150 | 0.1162 | 0.1209 | 0.3138 |
| | | 6 | MAE | 0.2631 | 0.2636 | 0.2626 | 0.2689 | 0.3692 |
| | | | RMSE | 0.1073 | 0.1083 | 0.1078 | 0.1122 | 0.2117 |
| | | 3 | MAE | 0.2138 | 0.2125 | 0.2198 | 0.2234 | 0.3906 |
| | DGP A1 | | RMSE | 0.0715 | 0.0705 | 0.0752 | 0.0775 | 0.2298 |
| | | 6 | MAE | 0.2064 | 0.2055 | 0.2067 | 0.2107 | 0.2916 |
| $n = 800$ | | | RMSE | 0.0674 | 0.0666 | 0.0675 | 0.0700 | 0.1313 |
| | | 3 | MAE | 0.2166 | 0.2176 | 0.2198 | 0.2225 | 0.3846 |
| | DGP A2 | | RMSE | 0.0728 | 0.0735 | 0.0754 | 0.0771 | 0.2279 |
| | | 6 | MAE | 0.2154 | 0.2142 | 0.2118 | 0.2203 | 0.2903 |
| | | | RMSE | 0.0717 | 0.0717 | 0.0703 | 0.0755 | 0.1351 |

Table 2: The Performance of the Estimators in Terms of MAE and RMSE: $a = 2$

| | | $k$ | | NN-TR | NN-ES | KN-TR | KN-ES | W-SBC |
|---|---|---|---|---|---|---|---|---|
| | | 3 | MAE | 0.6572 | 0.6533 | 0.6613 | 0.6735 | 1.0337 |
| | DGP A1 | | RMSE | 0.6726 | 0.6725 | 0.6896 | 0.7130 | 1.6586 |
| | | 6 | MAE | 0.6485 | 0.6523 | 0.6545 | 0.6665 | 0.8734 |
| $n = 200$ | | | RMSE | 0.6743 | 0.6814 | 0.6890 | 0.7056 | 1.1978 |
| | | 3 | MAE | 0.6674 | 0.6729 | 0.6764 | 0.6807 | 1.0113 |
| | DGP A2 | | RMSE | 0.7108 | 0.7192 | 0.7280 | 0.7362 | 1.6308 |
| | | 6 | MAE | 0.6680 | 0.6651 | 0.6722 | 0.6762 | 0.9180 |
| | | | RMSE | 0.7057 | 0.7066 | 0.7139 | 0.7235 | 1.3084 |
| | | 3 | MAE | 0.4208 | 0.4225 | 0.4336 | 0.4388 | 0.7630 |
| | DGP A1 | | RMSE | 0.2769 | 0.2778 | 0.2922 | 0.2987 | 0.8835 |
| | | 6 | MAE | 0.4100 | 0.4089 | 0.4114 | 0.4161 | 0.5696 |
| $n = 500$ | | | RMSE | 0.2640 | 0.2628 | 0.2653 | 0.2713 | 0.5052 |
| | | 3 | MAE | 0.4516 | 0.4501 | 0.4571 | 0.4644 | 0.7815 |
| | DGP A2 | | RMSE | 0.3214 | 0.3188 | 0.3287 | 0.3385 | 0.9258 |
| | | 6 | MAE | 0.4220 | 0.4214 | 0.4186 | 0.4300 | 0.5756 |
| | | | RMSE | 0.2816 | 0.2818 | 0.2806 | 0.2927 | 0.5243 |
| | | 3 | MAE | 0.3441 | 0.3448 | 0.3551 | 0.3584 | 0.6857 |
| | DGP A1 | | RMSE | 0.1873 | 0.1880 | 0.2003 | 0.2052 | 0.6763 |
| | | 6 | MAE | 0.3264 | 0.3255 | 0.3258 | 0.3325 | 0.4642 |
| $n = 800$ | | | RMSE | 0.1678 | 0.1674 | 0.1688 | 0.1747 | 0.3388 |
| | | 3 | MAE | 0.3425 | 0.3417 | 0.3480 | 0.3532 | 0.6838 |
| | | | RMSE | 0.1845 | 0.1839 | 0.1911 | 0.1966 | 0.6855 |
| | | 6 | MAE | 0.3340 | 0.3352 | 0.3362 | 0.3414 | 0.4721 |
| | | | RMSE | 0.1761 | 0.1783 | 0.1785 | 0.1841 | 0.3520 |

## 4.2 The Performance of the Bootstrap Procedure

In this subsection, we investigate the bootstrap procedure, using the same model as before. Table 2 contains finite sample coverage probabilities for the four types of estimators. When the sample size was 200, the bootstrap coverage probability is smaller than the nominal ones. When the sample size was 500, the bootstrap methods perform reasonably well.

It is worth noting that the performance difference between the case with true parameter $\theta_0$ (TR) and the case with the estimated parameter $\theta_0$ (ES) is almost negligible. This again affirms the robustness of the bootstrap procedure to the quality of the first step estimator $\hat{\theta}$.

Table 3: The Performance of the Proposed Bootstrap Method

| | | $k$ | Nom. Cov. Prob. | NN-TR | NN-ES | KN-TR | KN-ES |
|---|---|---|---|---|---|---|---|
| | | | 99% | 0.9815 | 0.9785 | 0.9825 | 0.9775 |
| | | 3 | 95% | 0.9355 | 0.9360 | 0.9380 | 0.9300 |
| | DGP A1 | | 90% | 0.8835 | 0.8815 | 0.8795 | 0.8755 |
| | | | 99% | 0.9825 | 0.9845 | 0.9800 | 0.9495 |
| | | 6 | 95% | 0.9355 | 0.9380 | 0.9405 | 0.9050 |
| $n = 200$ | | | 90% | 0.8885 | 0.8920 | 0.8915 | 0.8560 |
| | | | 99% | 0.9835 | 0.9830 | 0.9830 | 0.9765 |
| | | 3 | 95% | 0.9425 | 0.9490 | 0.9465 | 0.9330 |
| | DGP A2 | | 90% | 0.9025 | 0.8985 | 0.9005 | 0.8730 |
| | | | 99% | 0.9810 | 0.9835 | 0.9875 | 0.9255 |
| | | 6 | 95% | 0.9415 | 0.9415 | 0.9440 | 0.8800 |
| | | | 90% | 0.8945 | 0.8935 | 0.9015 | 0.8330 |
| | | | 99% | 0.9910 | 0.9905 | 0.9875 | 0.9900 |
| | | 3 | 95% | 0.9395 | 0.9440 | 0.9400 | 0.9470 |
| | DGP A1 | | 90% | 0.8980 | 0.8990 | 0.8960 | 0.8900 |
| | | | 99% | 0.9885 | 0.9885 | 0.9880 | 0.9860 |
| | | 6 | 95% | 0.9480 | 0.9445 | 0.9495 | 0.9440 |
| $n = 500$ | | | 90% | 0.8890 | 0.8945 | 0.8975 | 0.8890 |
| | | | 99% | 0.9900 | 0.9885 | 0.9905 | 0.9880 |
| | | 3 | 95% | 0.9485 | 0.9440 | 0.9425 | 0.9395 |
| | DGP A2 | | 90% | 0.8920 | 0.8850 | 0.8870 | 0.8920 |
| | | | 99% | 0.9880 | 0.9880 | 0.9885 | 0.9860 |
| | | 6 | 95% | 0.9435 | 0.9455 | 0.9480 | 0.9435 |
| | | | 90% | 0.8970 | 0.9005 | 0.8965 | 0.8855 |

Likewise, the performance is also similar across different numbers of covariates 3 and 6. It is interesting to note that the estimator NN-ES appears to perform slightly better than KN-ES. This may be perhaps due to the fact that the probability integral transform in the SNN estimation has an effect of reducing further the estimation error in $\hat{\theta}$. A more definite answer would require an analysis of the second order effect of $\hat{\theta}$. Finally, the bootstrap performance does not show much difference with regard to the heavy tailedness of the error distribution in the selection equation.

# 5 Empirical Application: Female Labor Supply

In this section, we illustrate the bootstrap procedure of this paper drawing on a well-known study of female labor supply. The model and the data sets are taken from Mroz (1987) that contain demographic characteristics of 753 married female workers in the United States. As for the hours equation and the labor participation equation, we consider the following:

$$\begin{aligned} h_i &= \beta_0 + \log(w_i)\beta_1 + Z_{2i}\beta_2 + Z_{3i}^\top \beta_4 + \varepsilon_i \text{ and} \\ D_i &= 1\left\{X_i^\top \theta_0 \geq \eta_i\right\}, \end{aligned}$$

where $h_i$ denotes hours that the $i$-th female worker worked (divided by $10^3$), $w_i$ her hourly wage, $Z_{2i}$ nonwife income of the household that the female worker belongs (divided by 10) and $Z_{3i}$ a vector of other demographic variables.

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. As for variables to be included in $X_i$, we take as common background variables such as unemployment rate in the county, parents' schooling, variables related to the number of children, and nonwife income. We consider the following specifications of $X_i$ in the participation equation:

| | |
|---|---|
| Specification I | : background variables plus variables of labor market experiences |
| Specification II | : background variables plus variables of age and schooling |
| Specification III | : all the variables in Specifications I and II. |

The variables in $X_i$ are also appropriately rescaled.

We estimated the model assuming two situations for $\eta_i$ : one with the assumption that the conditional median of $\eta_i$ given $X_i$ is zero, and the other with the assumption that $\eta_i$ and $X_i$ are independent, $\eta_i$ following a normal distribution. For the former model, we used maximum score estimation to estimate $\theta_0$ and for the latter, probit estimation. As for the estimation of $\beta_0, \cdots, \beta_4$, we employ the estimation method of partial linear models of Robinson (1988).

The results are shown in Tables 4-6. First, it appears that the results do not show much difference between those using kernel estimation and those using SNN estimation. This result appears due to the fact that estimation errors in $\hat{\theta}$ do not affect $\hat{\beta}$ in the first order asymptotic approximation. Also estimation through probit estimation or maximum score estimation does not appear to produce much difference for most coefficients. Second, there seems to be more variation across different specifications of $X_i$ for certain variables such as coefficient estimates of the number of young children and nonwife income, in particular

Table 4: Estimation of Female Labor Participation (Specification I)

(In the parentheses are bootstrap standard errors.)

| | Probit Estimation | | Maximum Score Estimation | |
|---|---|---|---|---|
| | SNN | Kernel Estimation | SNN | Kernel Estimation |
| Log wage | 0.0870 | 0.1096 | 0.2245 | 0.2225 |
| | (0.1309) | (0.1257) | (0.1449) | (0.1443) |
| Nonwife Income | 0.0324 | 0.0299 | 0.0787 | 0.0916 |
| | (0.1075) | (0.1039) | (0.1059) | (0.0807) |
| Young Children | 0.0559 | 0.0724 | −0.5609 | −0.5351 |
| | (0.2413) | (0.2471) | (0.2023) | (0.1988) |
| Old Children | −0.0904 | −0.0887 | −0.0865 | −0.0876 |
| | (0.0647) | (0.0645) | (0.0604) | (0.0560) |
| Age | 0.0222 | −0.0204 | −0.1320 | −0.1319 |
| | (0.1173) | (0.1171) | (0.0836) | (0.0833) |
| Education | 0.0065 | 0.0101 | −0.0112 | −0.0105 |
| | (0.0485) | (0.0486) | (0.0478) | (0.0467) |

Table 5: Estimation of Female Labor Participation (Specification II)

(In the parentheses are bootstrap standard errors.)

| | Probit Estimation | | Maximum Score Estimation | |
|---|---|---|---|---|
| | SNN | Kernel Estimation | SNN | Kernel Estimation |
| Log wage | 0.1313 | 0.1378 | 0.1966 | 0.2081 |
| | (0.1521) | (0.1584) | (0.1758) | (0.1807) |
| Nonwife Income | 0.0085 | 0.0655 | −0.0025 | 0.1928 |
| | (0.1663) | (0.1016) | (0.1682) | (0.1546) |
| Young Children | −0.6462 | −0.3990 | −0.4318 | −0.4598 |
| | (0.6747) | (0.3435) | (0.2015) | (0.2057) |
| Old Children | −0.1188 | −0.1044 | −0.1417 | −0.3298 |
| | (0.0552) | (0.0514) | (0.2261) | (0.1958) |
| Age | −0.0227 | −0.1571 | −0.1832 | −0.2667 |
| | (0.2280) | (0.1216) | (0.1263) | (0.1153) |
| Education | −0.0078 | 0.0558 | −0.0223 | −0.0476 |
| | (0.1198) | (0.0609) | (0.0572) | (0.0526) |

Table 6: Estimation of Female Labor Participation (Specification III)
(In the parentheses are bootstrap standard errors.)

| | Probit Estimation | | Maximum Score Estimation | |
|---|---|---|---|---|
| | SNN | Kernel Estimation | SNN | Kernel Estimation |
| Log wage | 0.0913 | 0.1297 | 0.1665 | 0.1793 |
| | (0.1192) | (0.1249) | (0.1241) | (0.1242) |
| Nonwife Income | 0.1200 | 0.0748 | −0.0118 | −0.0323 |
| | (0.0909) | (0.0892) | (0.0861) | (0.0865) |
| Young Children | 0.3549 | −0.3887 | −0.3198 | −0.2929 |
| | (0.2646) | (0.2852) | (0.2239) | (0.2212) |
| Old Children | −0.0834 | −0.0881 | −0.0906 | −0.0892 |
| | (0.0546) | (0.0544) | (0.0555) | (0.0553) |
| Age | −0.0117 | −0.0183 | −0.0306 | −0.0442 |
| | (0.1052) | (0.1041) | (0.1031) | (0.1097) |
| Education | −0.1298 | −0.1031 | −0.0338 | −0.0248 |
| | (0.0435) | (0.0442) | (0.0416) | (0.0422) |

between Specification I and Specifications II and III. Third, the variation across different specifications of $X_i$ appears less prominent in the case of maximum score estimation than in the case of probit estimation.

In summary, the results of the empirical exercise suggest that for most coefficient estimates of the outcome equation, the specification of the participation equation does not make much difference, except for certain variables, and the results appear more robust to the various different specification of $X_i$ in the case of maximum score estimation. Part of this robustness seems to be due to the first order robustness of estimates of $\beta$ to the noise in the estimation of the participation equation.

# 6 Conclusion

This paper considers a semiparametric conditional moment restriction that contains conditional expectations of single-index conditioning variables. This paper shows that the influence of the first step index estimators on the estimator of the parameter of interest is asymptotically negligible in this situation. An analysis was performed in terms of the Fréchet derivatives of a relevant class of functionals. Hence this phenomenon appears to have a generic nature. This result enables this paper to develop a bootstrap procedure that

is asymptotically valid in the presence of first step single-index estimators following cube root asymptotics. The simulation studies confirm that the method performs reasonably well.

As mentioned in the main text, it is expected that the results of this paper extend to the case of the single-index $\lambda_j(x)$ being a nonparametric function This situation often arises in the literature of program evaluations where the single-index component corresponds to a propensity score. It also appears that the result extends to the case of weakly dependent observations. However, the extension may be more than an obvious corollary from the result of this paper, because this paper heavily draws on the empirical process theory that applies to the i.i.d. observations.

# 7 Appendix

## 7.1 Continuity of Linear Functionals of Conditional Expectations

Conditional expectations that involve unknown parameters in the conditioning variable frequently arise in semiparametric models. Continuity of conditional expectations with respect to such parameters plays a central role in the asymptotic analysis. In this section, we provide a generic, primitive condition that yields such continuity. Let $X \in \mathbf{R}^{d_X}$ be a random vector with support $\mathcal{S}_X$ and let $\Lambda$ be a class of $\mathbf{R}$-valued functions on $\mathbf{R}^{d_X}$ with a generic element denoted by $\lambda$.

Fix $\lambda_0 \in \Lambda$ and let $f_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2)$ denote the conditional density function of a random vector $Y \in \mathbf{R}^{d_Y}$ given $(\lambda_0(X), \lambda(X)) = (\bar{\lambda}_1, \bar{\lambda}_2)$ with respect to a $\sigma$-finite measure, say, $w_\lambda(\cdot|\bar{\lambda}_1, \bar{\lambda}_2)$. Note that we do not assume that $Y$ is continuous as we do not require that $w_\lambda(\cdot|\bar{\lambda}_1, \bar{\lambda}_2)$ is the Lebesgue measure. Let $\mathcal{S}_Y$ be the support of $Y$ and let $\mathcal{S}_\lambda$ be that of $(\lambda_0(X), \lambda(X))$. We define $||\cdot||$ to be the Euclidean norm in $\mathbf{R}^J$ and $||\cdot||_\infty$ to be the sup norm: $||f||_\infty = \sup_{x \in \mathcal{S}_X} |f(x)|$.

**Definition A :** (i) $\mathcal{P}_Y \equiv \{f_\lambda(y|\cdot,\cdot) : (\lambda, y) \in \Lambda \times \mathcal{S}_Y\}$ is *regular for* $\tilde{\varphi} : \mathbf{R}^{d_Y} \to \mathbf{R}^J$, if for each $\lambda \in \Lambda$ and $(\bar{\lambda}_1, \bar{\lambda}_2) \in \mathcal{S}_\lambda$,

$$\sup_{(\tilde{\lambda}_1, \tilde{\lambda}_2) \in \mathcal{S}_\lambda : |\bar{\lambda}_1 - \tilde{\lambda}_1| + |\bar{\lambda}_2 - \tilde{\lambda}_2| \leq \delta} \left| f_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2) - f_\lambda(y|\tilde{\lambda}_1, \tilde{\lambda}_2) \right| < C_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2)\delta, \ \delta \in [0, \infty) \quad (8)$$

where $C_\lambda(\cdot|\bar{\lambda}_1, \bar{\lambda}_2) : \mathcal{S}_Y \to \mathbf{R}$ is such that for some $C > 0$ that does not depend on $\lambda$,

$$\sup_{(\bar{\lambda}_1, \bar{\lambda}_2) \in \mathcal{S}_\lambda} \int ||\tilde{\varphi}(y)|| C_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2) w_\lambda(dy|\bar{\lambda}_1, \bar{\lambda}_2) < C.$$

(ii) When $\mathcal{P}_Y$ is regular for an identity map, we say simply that it is *regular*.

The regularity condition is a type of an equicontinuity condition for functions $f_\lambda(y|\cdot,\cdot)$, $(y,\lambda) \in \mathcal{S}_Y \times \Lambda$. Roughly speaking a set of conditional densities are regular when the response of a conditional density function to a small perturbation in the conditioning variable is small uniformly over $\lambda \in \Lambda$. The condition does not require that the conditional density function be continuous in the parameter $\lambda \in \Lambda$, which is cumbersome to check in many situations. (Note that the perturbation on the right-hand side of (8) is concerned with a "fixed" function $f_\lambda(y|\cdot,\cdot)$, not across different density functions with different $\lambda$'s.)

When $f_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2)$ is continuously differentiable in $(\bar{\lambda}_1, \bar{\lambda}_2)$ with a derivative that is bounded uniformly over $\lambda \in \Lambda$ and $\tilde{\varphi}(Y)$ has a bounded support, $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$. Alternatively suppose that there exists $C > 0$ such that for each $\lambda \in \Lambda$ and $(\bar{\lambda}_1, \bar{\lambda}_2) \in \mathcal{S}_\lambda$,

$$\sup_{(\tilde{\lambda}_1,\tilde{\lambda}_2)\in\mathcal{S}_\lambda : |\bar{\lambda}_1-\tilde{\lambda}_1|+|\bar{\lambda}_2-\tilde{\lambda}_2|\leq\delta} \left| \frac{f_\lambda(y|\tilde{\lambda}_1, \tilde{\lambda}_2)}{f_\lambda(y|\bar{\lambda}_1, \bar{\lambda}_2)} - 1 \right| \leq C\delta,$$

and $\mathbf{E}[||\tilde{\varphi}(Y)|||X] < C$. Then $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$. The regularity condition for $\mathcal{P}_Y$ yields the following Lemma A1 as an important consequence.

**Lemma A1 :** *Suppose that $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$ an envelope of $\Phi$ and $\Phi$ is a class of $\mathbf{R}^J$-valued functions on $\mathbf{R}^{d_Y}$. Then, for each $\lambda \in \Lambda$, $\varphi \in \Phi$, and $x \in \mathcal{S}_X$,*

$$\begin{aligned} ||\mu_\varphi(x;\lambda_0,\lambda) - \mu_\varphi(x;\lambda)|| &\leq C|\lambda(x) - \lambda_0(x)|, \text{ and} \\ ||\mu_\varphi(x;\lambda_0,\lambda) - \mu_\varphi(x;\lambda_0)|| &\leq C|\lambda(x) - \lambda_0(x)|, \end{aligned}$$

*where*

$$\begin{aligned} \mu_\varphi(x;\lambda) &= \mathbf{E}[\varphi(Y)|\lambda(X) = \lambda(x)] \text{ and} \\ \mu_\varphi(x;\lambda_0,\lambda) &= \mathbf{E}[\varphi(Y)|(\lambda_0(X),\lambda(X)) = (\lambda_0(x),\lambda(x))] \end{aligned}$$

*and $C$ does not depend on $\lambda, \lambda_0$, $x$, or $\varphi$.*

Lemma A1 shows that the conditional expectations are continuous in the parameter $\lambda$ in the conditioning variable. This result is similar to Lemma A2(ii) of Song (2008). (See also Lemma A5 of Song (2009).)

We introduce an additional random vector $Z \in \mathbf{R}^{d_Z}$ with a support $\mathcal{S}_Z$. Let $\Psi$ be a class of $\mathbf{R}^J$-valued functions on $\mathbf{R}^{d_Z}$ with a generic element denoted by $\psi$ and its envelope by $\tilde{\psi}$. As before, we fix $\lambda_0 \in \Lambda$, let $h_\lambda(z|\bar{\lambda}_1, \bar{\lambda}_2)$ denote the conditional density function of $Z$ given $(\lambda_0(X), \lambda(X)) = (\bar{\lambda}_1, \bar{\lambda}_2)$ with respect to a $\sigma$-finite measure, and define $\mathcal{P}_Z \equiv \{h_\lambda(z|\cdot,\cdot) :$

$(\lambda, z) \in \Lambda \times \mathcal{S}_Z\}$. Suppose that the parameter of interest takes the form of

$$\Gamma_{\varphi,\psi}(\lambda) = \mathbf{E}\left[\mu_\varphi(X;\lambda)^\top \psi(Z)\right].$$

We would like to analyze continuity of $\Gamma_{\varphi,\psi}(\lambda)$ in $\lambda \in \Lambda$. When $\mathcal{P}_Y$ and $\mathcal{P}_Z$ are regular, we obtain the following result.

**Lemma A2 :** *Suppose that $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$ and $\mathcal{P}_Z$ is regular for $\tilde{\psi}$. Then, there exists $C > 0$ such that for each $\lambda$ in $\Lambda$,*

$$\sup_{(\varphi,\psi)\in\Phi\times\Psi} |\Gamma_{\varphi,\psi}(\lambda) - \Gamma_{\varphi,\psi}(\lambda_0)| \leq C||\lambda - \lambda_0||_\infty^2.$$

*Therefore, the first order Fréchet derivative of $\Gamma_{\varphi,\psi}(\lambda)$ at $\lambda_0 \in \Lambda$ is equal to zero.*

Lemma A2 says that the functional $\Gamma_{\varphi,\psi}(\lambda)$ is not sensitive to the first order perturbation of $\lambda$ around $\lambda_0$. In view of Newey (1994), Lemma A2 suggests that in general, there is no estimation effect of $\hat{\lambda}$ on the asymptotic variance of the estimator $\hat{\Gamma}_{\varphi,\psi}(\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^n \hat{\mu}_\varphi(X_i;\hat{\lambda})^\top \psi(Z_i)$, where $\hat{\mu}_\varphi(X_i;\lambda)$ denotes a nonparametric estimator of $\mu_\varphi(X_i;\lambda)$.

**Proof of Lemma A1 :** We proceed in a similar manner as in the proof of Lemma A5 of Song (2009). We show only the first statement because the proof is almost the same for the second statement.

Choose $x \in \mathcal{S}_X$ and $\lambda_1 \in \Lambda$ and let $\delta \equiv |\bar{\lambda}_1 - \bar{\lambda}_0|$, where $\bar{\lambda}_0 \equiv \lambda_0(x)$ and $\bar{\lambda}_1 \equiv \lambda_1(x)$. We write $\mu_\varphi(\bar{\lambda}_1, \bar{\lambda}_0) = \mu_\varphi(x;\lambda_1,\lambda_0)$ and $\mu_\varphi(\bar{\lambda}_0) = \mu_\varphi(x;\lambda_0)$. Let $P_{0,\varphi}$ be the conditional distribution of $(\varphi(Y), X)$ given $\lambda_0(X) = \bar{\lambda}_0$ and let $\mathbf{E}_{0,\varphi}$ denote the expectation under $P_{0,\varphi}$. Let $A_j \equiv 1\{|\lambda_j(X) - \bar{\lambda}_j| \leq 3\delta\}$, $j = 0, 1$. Note that $\mathbf{E}_{0,\varphi}[A_0] = 1$ and $\mathbf{E}_{0,\varphi}[A_1] = 1$. Let $\tilde{\mu}_\varphi(\bar{\lambda}_j, \bar{\lambda}_0) \equiv \mathbf{E}_{0,\varphi}\left[\varphi(Y)A_j\right] / \mathbf{E}_{0,\varphi}[A_j] = \mathbf{E}_{0,\varphi}\left[\varphi(Y)A_j\right]$, $j = 0, 1$. Then,

$$\begin{aligned} \left\|\mu_\varphi(\bar{\lambda}_1,\bar{\lambda}_0) - \mu_\varphi(\bar{\lambda}_0)\right\| &\leq \left\|\mu_\varphi(\bar{\lambda}_1,\bar{\lambda}_0) - \tilde{\mu}_\varphi(\bar{\lambda}_1,\bar{\lambda}_0)\right\| + \left\|\tilde{\mu}_\varphi(\bar{\lambda}_1,\bar{\lambda}_0) - \mu_\varphi(\bar{\lambda}_0)\right\| \\ &= (I) + (II), \text{ say.} \end{aligned}$$

Let us turn to $(I)$. By the definition of conditional expectation,

$$\tilde{\mu}_\varphi(\bar{\lambda}_1,\bar{\lambda}_0) = \int_{\bar{\lambda}_1-3\delta}^{\bar{\lambda}_1+3\delta} \mu_\varphi(\bar{\lambda},\bar{\lambda}_0) dF_{\lambda_1}(\bar{\lambda}|\bar{\lambda}_0),$$

where $F_{\lambda_1}(\cdot|\bar{\lambda}_0)$ is the conditional CDF of $\lambda_1(X_i)$ given $\lambda_0(X_i) = \bar{\lambda}_0$. Note that

$$\left\|\mu_\varphi(\bar{\lambda}_1,\bar{\lambda}_0) - \tilde{\mu}_\varphi(\bar{\lambda}_1,\bar{\lambda}_0)\right\| \leq \sup_{v\in[-3\delta,3\delta]:(\bar{\lambda}_1+v,\bar{\lambda}_0)\in\mathcal{S}_{\lambda_1}} \left\|\mu_\varphi(\bar{\lambda}_1+v,\bar{\lambda}_0) - \mu_\varphi(\bar{\lambda}_1,\bar{\lambda}_0)\right\|$$

because $\int_{\bar{\lambda}_1-3\delta}^{\bar{\lambda}_1+3\delta} dF_{\lambda_1}(\bar{\lambda}|\bar{\lambda}_0) = \mathbf{E}_{0,\varphi}[A_1] = 1$. The last term above is bounded by

$$\begin{aligned} & \sup_{v\in[-3\delta,3\delta]:(\bar{\lambda}_1+v,\bar{\lambda}_0)\in\mathcal{S}_{\lambda_1}} \int_{\mathcal{S}_Y} ||\tilde{\varphi}(y)|| \left| f_{\lambda_1}(y|\bar{\lambda}_1+v,\bar{\lambda}_0) - f_{\lambda_1}(y|\bar{\lambda}_1,\bar{\lambda}_0)\right| w_{\lambda_1}(dy|\bar{\lambda}_1,\bar{\lambda}_0) \\ \leq \;& \delta \int_{\mathcal{S}_Y} ||\tilde{\varphi}(y)|| C_{\lambda_1}(y|\bar{\lambda}_1,\bar{\lambda}_0) w_{\lambda_1}(dy|\bar{\lambda}_1,\bar{\lambda}_0) \leq C\delta. \end{aligned}$$

Let us turn to $(II)$ which we write as

$$||\mathbf{E}_{0,\varphi}\left[\varphi(Y)A_1\right] - \mathbf{E}_{0,\varphi}\left[\varphi(Y)\right]|| = ||\mathbf{E}_{0,\varphi}\left[VA_1\right]||,$$

where $V \equiv \varphi(Y) - \mathbf{E}_{0,\varphi}\left[\varphi(Y)\right]$ because $\mathbf{E}_{0,\varphi}\left[A_1\right] = 1$. The term $(II)$ is equal to

$$\begin{aligned} & \left\| \int_{\bar{\lambda}_1-3\delta}^{\bar{\lambda}_1+3\delta} \mathbf{E}\left[VA_1|\lambda_1(X)=\bar{\lambda},\lambda_0(X)=\bar{\lambda}_0\right] dF_{\lambda_1}(\bar{\lambda}|\bar{\lambda}_0)\right\| \\ = \;& \left\| \int_{\bar{\lambda}_1-3\delta}^{\bar{\lambda}_1+3\delta} \mathbf{E}\left[V|\lambda_1(X)=\bar{\lambda},\lambda_0(X)=\bar{\lambda}_0\right] dF_{\lambda_1}(\bar{\lambda}|\bar{\lambda}_0)\right\| \end{aligned}$$

which is bounded by $C\delta$, similarly as before. This implies that $(II) \leq C\delta$. ■

**Proof of Lemma A2 :** Let $\mu_{\varphi,\lambda}(x) = \mu_\varphi(x;\lambda)$ and $\mu_{\varphi,0}(x) = \mu_\varphi(x;\lambda_0)$. Similarly define $\mu_{\psi,\lambda}(x) = \mu_\psi(x;\lambda)$ and $\mu_{\psi,0}(x) = \mu_\psi(x;\lambda_0)$, where $\mu_\psi(x;\lambda) = \mathbf{E}[\psi(Z)|\lambda(X)=\lambda(x)]$. First write

$$\begin{aligned} & \mathbf{E}\left[\psi(Z)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] = \mathbf{E}\left[\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] \\ = \;& \mathbf{E}\left[\left(\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]-\mu_{\psi,0}(X)\right)^\top\left(\mu_{\varphi,\lambda}(X)-\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]\right)\right] \\ & +\mathbf{E}\left[\left(\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]-\mu_{\psi,0}(X)\right)^\top\left(\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]-\mu_{\varphi,0}(X)\right)\right] \\ & +\mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] \\ = \;& \mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] + O(||\lambda-\lambda_0||_\infty^2) \end{aligned}$$

by applying Lemma A1 to the first two expectations on the right-hand side of the first equality. The last expectation is equal to

$$\begin{aligned} & \mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]\right\}\right] \\ & +\mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]-\mu_{\varphi,0}(X)\right\}\right] \\ = \;& \mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]\right\}\right] \\ = \;& \mathbf{E}\left[\left\{\mu_{\psi,0}(X)-\mu_{\psi,\lambda}(X)\right\}^\top\left\{\mu_{\varphi,\lambda}(X)-\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]\right\}\right]. \end{aligned}$$

Applying Lemma A1 again, the last expectation is equal to $O(||\lambda-\lambda_0||_\infty^2)$. Hence we conclude that

$$\mathbf{E}\left[\psi(Z)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right]=O(||\lambda-\lambda_0||_\infty^2),$$

affirming the claim that the Fréchet derivative is equal to zero. ■

## 7.2 Assumptions on Regularity of Conditional Densities

We collect the conditions for Theorem 1 that have a technical character. Let $S_{i,j}$ be the $j$-th entry of $S_i$, where $S_i=\rho_\mu(V_i,\mu_i;\beta_0)$ and let $u_{i,j}=F_j(\lambda_j(X_i))$, where $F_j$ denotes the CDF of $\lambda_j(X_i)$. Define $Z_{i,j}=(S_{i,j},W_{1,i},u_{i,J+1})$ if $u_{i,J+1}\neq u_{i,j}$ and $Z_{i,j}=(S_{i,j},W_{1,i})$ if $u_{i,J+1}=u_{i,j}$. We set $\tilde{\psi}$ to be such that $\tilde{\psi}(Z_{i,j})=|S_{i,j}|$. Define $f_{\theta,j}(y|u_0,u_1)$ to be the conditional density of $Y_{i,j}$ given $(u_{i,j},u_{\theta,i,j})=(u_0,u_1)$ with respect to a $\sigma$-finite measure, where $u_{\theta,i,j}=F_{\theta,j}(\lambda_j(X_i;\theta))$ and $F_{\theta,j}$ is the CDF of $\lambda_j(X_i;\theta)$. Similarly define $h_{\theta,j}(z|u_0,u_1)$ to be the conditional density of $Z_{i,j}$ given $(u_{i,j},u_{\theta,i,j})=(u_0,u_1)$ with respect to a $\sigma$-finite measure. Let $\mathcal{S}_{Y,j}$ and $\mathcal{S}_{Z,j}$ be the supports of $Y_{i,j}$ and $Z_{i,j}$,

$$\begin{aligned}\mathcal{P}_{Y,j}(\delta) &\equiv \{f_{\theta,j}(y|\cdot,\cdot):(\theta,y)\in\Theta(\delta)\times\mathcal{S}_{Y,j}\}\text{ and}\\ \mathcal{P}_{Z,j}(\delta) &\equiv \{h_{\theta,j}(z|\cdot,\cdot):(\theta,z)\in\Theta(\delta)\times\mathcal{S}_{Z,j}\}.\end{aligned}$$

**Assumption A :** For each $j=1,\cdots,J+1$, there exist $\delta_j>0$ and $C_j>0$ such that
(i) for each $j=1,\cdots,J+1$,

$$|F_{\theta_1,j}(\lambda_j(x;\theta_1))-F_{\theta_2,j}(\lambda_j(x;\theta_2))|\leq C_j||\theta_1-\theta_2||,\text{ for all }\theta_1,\theta_2\in\Theta(\delta_j),$$

(ii) for each $j=1,\cdots,J$, $\mathcal{P}_{Y,j}(\delta_j)$ is regular and $\mathcal{P}_{Z,j}(\delta_j)$ is regular for $\tilde{\psi}$, and
(iii) for each $j=1,\cdots,J$, (a) $\sup_{u\in[0,1]}\mathbf{E}[|Y_{i,j}||u_{i,j}=u]<\infty$, and (b) $\mathbf{E}[Y_{i,j}|u_{i,j}=\cdot]$ is twice continuously differentiable with bounded derivatives.

Assumption A(i) is a regularity condition for the index function $\lambda_j(\cdot;\theta)$. Some sufficient conditions for the regularity of $\mathcal{P}_{Y,j}(\delta_j)$ were discussed after Lemma A1. The regularity of $\mathcal{P}_{Z,j}(\delta_j)$ in Assumption A(ii) can be replaced by a lower level sufficient condition in more specific contexts. Note that in the case of the sample selection model in Example 1, $J=2$, $u_{i,1}=u_{i,2}=u_{i,3}$, and in the case of the model with the single-index instrument in Example 2, $J=1$, $u_{i,1}=u_{i,2}$. In both cases, $S_i$ is a constant vector of $-1$'s. Hence $\mathcal{P}_{Z,j}(\delta_j)$ becomes regular, for instance, if the conditional density function of $W_{1,i}$ given $(u_{i,1},u_{\theta,i,1})=(u_0,u_1)$ is continuously differentiable in $(u_0,u_1)$ with a derivative uniformly bounded over $\theta\in\Theta(\delta_j)$ and $W_{1,i}$ has a bounded support.

## 7.3 Proofs of the Main Results

Throughout the proofs, the notation $C$ denotes a positive constant that may assume different values in different contexts. Let $L_p(P)$, $p \geq 1$, be the space of $L_p$-bounded functions: $||f||_p := \{\int |f(x)|^p P(dx)\}^{1/p} < \infty$, and for a space of functions $\mathcal{F} \subset L_p(P)$ for $p \geq 1$, let $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_p)$, the bracketing number of $\mathcal{F}$ with respect to the norm $||\cdot||_p$, to be 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 $||\Delta_i||_p < \varepsilon$ and for all $f \in \mathcal{F}$, there exists $i \leq r$ with $||f_i - f||_p < \Delta_i/2$. Similarly, we define $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_\infty)$ to be the bracketing number of $\mathcal{F}$ with respect to the sup norm $||\cdot||_\infty$. For any norm $||\cdot||$ which is equal to $||\cdot||_p$ or $||\cdot||_\infty$, we define $N(\varepsilon, \mathcal{F}, ||\cdot||)$ to be the covering number of $\mathcal{F}$, i.e. the smallest number of $\varepsilon$-balls that cover $\mathcal{F}$.

**Proof of Theorem 1 :** Write $\mu(x) = \mu(x; \lambda_0)$ and $\hat{\mu}(x) = \hat{\mu}(x; \hat{\lambda})$ and as in Section 7.2, introduce notations $u_{i,j} = F_j(\lambda_j(X_i))$ and $u_{\theta,i,j} = F_{\theta,j}(\lambda_j(X_i;\theta))$. Put briefly, $\hat{1}_{il} = 1\{\hat{W}_i \leq \hat{W}_l\}$ and $1_{il} = 1\{W_i \leq W_l\}$ and

$$\begin{aligned}
\rho_i(\beta) &= \rho(V_i, \mu_i; \beta),\ \rho_{\mu,i}(\beta) = \rho_\mu(V_i, \mu_i; \beta), \\
\hat{\rho}_i(\beta) &= \rho(V_i, \hat{\mu}_i; \beta), \text{ and } \hat{\rho}_{\beta,i}(\beta) = \rho_\beta(V_i, \hat{\mu}_i; \beta).
\end{aligned}$$

We first show the consistency of $\hat{\beta}$. Let $Q(\beta) = \int \{\mathbf{E}[\rho_i(\beta) 1\{W_i \leq w\} D_i]\}^2 dF_{W,D=1}(w)$,

$$\begin{aligned}
\hat{Q}(\beta) &= \frac{1}{n}\sum_{l=1}^n D_l \left\{ \frac{1}{n}\sum_{i=1}^n D_i \hat{\rho}_i(\beta) \hat{1}_{il} \right\}^2 \text{ and} \\
\tilde{Q}(\beta) &= \frac{1}{n}\sum_{l=1}^n D_l \left\{ \frac{1}{n}\sum_{i=1}^n D_i \rho_i(\beta) 1_{il} \right\}^2,
\end{aligned}$$

where $F_{W,D=1}(w) = P\{W_i \leq w, D_i = 1\}$. Let $F_{n,\theta,j}(\bar{\lambda}) = \frac{1}{n}\sum_{i=1}^n 1\{\lambda_j(X_i;\theta) \leq \bar{\lambda}\}$ and $F_{\theta,j}(\bar{\lambda}) = P\{\lambda_j(X_i;\theta) \leq \bar{\lambda}\}$, and let $\hat{g}_j(u) = \sum_{i=1}^n Y_{ji} D_i K_h(\hat{u}_{i,j} - u) / \{\sum_{i=1}^n D_i K_h(\hat{u}_{i,j} - u)\}$ and

$$g_j(u) = \mathbf{E}[Y_{i,j} | u_{i,j} = u, D_i = 1],$$

Note that $||\hat{\mu} - \mu||_\infty \equiv \sup_{x \in \mathbf{R}^{d_X}} ||\hat{\mu}(x) - \mu(x)||$ is bounded by the maximum over $j = 1, \cdots, J+1$ of

$$\sup_{u \in [0,1]} |\hat{g}_j(u) - g_j(u)| + \sup_{x \in \mathbf{R}^{d_X}} |g_j(F_{n,\hat{\theta},j}(\lambda_j(x;\hat{\theta}))) - g_j(F_j(\lambda_j(x;\theta_0)))|. \tag{9}$$

The first term is $o_P(1)$ as in the proof of Lemma A4 of Song (2009) and the second term is $O_P(||\hat{\theta} - \theta_0||)$ (e.g. see the proof of Lemma A3 of Song (2009).) Therefore, $||\hat{\mu} - \mu||_\infty = o_P(1)$.

Now,

$$\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^{n} D_i\left\{\hat{\rho}_i(\beta)-\rho_i(\beta)\right\}\hat{1}_{il}\right| \leq \frac{||\hat{\mu}-\mu||_\infty}{n}\sum_{i=1}^{n}\sup_{\beta\in B}||\rho_\mu(V_i,\mu_i;\beta)|| \tag{10}$$
$$+\frac{||\hat{\mu}-\mu||_\infty^2}{2n}\sum_{i=1}^{n}\sup_{(\beta,\bar{\mu})\in B\times[-M,M]}||\rho_{\mu\mu}(V_i,\bar{\mu};\beta)||,$$

with probability approaching one for large $M$ such that $||\mu||_\infty < M$. The last term is $o_P(1)$ by Assumption 1(iv).

Note also that from large $n$ on,

$$\mathbf{E}\left(\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^{n} D_i\rho_i(\beta)\left(\hat{1}_{il}-1_{il}\right)\right|\right) \tag{11}$$
$$\leq \frac{1}{n}\sum_{i=1,i\neq l}^{n}\left\{\mathbf{E}\left[\sup_{\beta\in B}|\rho_i(\beta)|^2\right]\right\}^{1/2}\sqrt{P\{u_{l,J+1}-\Delta_n < u_{i,J+1}\leq u_{l,J+1}+\Delta_n\}},$$

where $\Delta_n = \max_{1\leq i\leq n}\sup_{\theta\in B(\theta_0,\delta_n)}||\hat{u}_{\theta,i,J+1}-u_{i,J+1}||$, $\delta_n = n^{-1/3+\varepsilon}$, with small $\varepsilon > 0$, and $\hat{u}_{\theta,i,J+1} = \frac{1}{n}\sum_{j=1,j\neq i}^{n} 1\{\lambda_{J+1}(X_j;\theta)\leq\lambda_{J+1}(X_i;\theta)\}$. Similarly as in the proof of Lemma A3 of Song (2009), $\Delta_n = O_P(\delta_n)$, so that the last term in (11) is $o(1)$. From (10) and (11),

$$\hat{Q}(\beta) = \tilde{Q}(\beta) + o_P(1), \text{ uniformly in } \beta\in B.$$

Since $\rho(v,\mu(x);\beta)$ is Lipschitz in $\beta$ with an $L_p$-bounded coefficient, $p > 2$, and $B$ is compact, the uniform convergence of $\tilde{Q}(\beta)$ to $Q(\beta)$ follows by the standard procedure. Hence $\sup_{\beta\in B}|\hat{Q}(\beta)-Q(\beta)| = o_P(1)$. As in Domínguez and Lobato (2004), this yields the consistency of $\hat{\beta}$.

Now, using the first order condition of the extremum estimation and the mean value theorem,

$$\sqrt{n}(\hat{\beta}-\beta_0) = G_n(\hat{\beta},\hat{\mu},\{\hat{W}_l\})^{-1}\sqrt{n}\xi_n(\hat{\beta},\hat{\mu},\{\hat{W}_l\}),$$

where, with $\bar{\beta}$ lying between $\hat{\beta}$ and $\beta_0$,

$$G_n(\hat{\beta},\hat{\mu},\{\hat{W}_l\}) = \frac{1}{n}\sum_{l=1}^{n} D_l\left\{\frac{1}{n}\sum_{i=1}^{n} D_i\hat{\rho}_{\beta,i}(\hat{\beta})\hat{1}_{il}\right\}\left\{\frac{1}{n}\sum_{i=1}^{n} D_i\hat{\rho}_{\beta,i}(\bar{\beta})^\top\hat{1}_{il}\right\} \text{ and}$$
$$\xi_n(\hat{\beta},\hat{\mu},\{\hat{W}_l\}) = \frac{1}{n}\sum_{l=1}^{n} D_l\left\{\frac{1}{n}\sum_{i=1}^{n} D_i\hat{\rho}_{\beta,i}(\hat{\beta})\hat{1}_{il}\right\}\left\{\frac{1}{n}\sum_{i=1}^{n} D_i\hat{\rho}_i(\beta_0)\hat{1}_{il}\right\}.$$

Using consistency of $\hat{\beta}$ and following similar steps in (10) and (11), we can show that

$G_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\})$ is equal to

$$G_n(\beta_0, \mu, \{W_l\}) + o_P(1) = \int \dot{H}(w)\dot{H}(w)^\top dF_{W,D=1}(w) + o_P(1),$$

by the law of large numbers. We turn to the analysis of $\sqrt{n}\xi_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\})$. Let $\mu_{\theta,i,j} = \mathbf{E}[Y_{i,j}|\lambda_j(X_i;\theta), D_i = 1]$ and $\mu_{\theta,i} = [\mu_{\theta,i,1}, \cdots, \mu_{\theta,i,J}]^\top$. Write

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\hat{\rho}_i(\beta_0)\hat{1}_{il} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\left\{\rho(V_i, \hat{\mu}_i;\beta_0) - \rho(V_i, \mu_{\hat{\theta},i};\beta_0)\right\}\hat{1}_{il} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\left\{\rho(V_i, \mu_{\hat{\theta},i};\beta_0) - \rho(V_i, \mu_i;\beta_0)\right\}\hat{1}_{il} \\
&= A_{1n} + A_{2n}, \text{ say.}
\end{aligned}$$

We first deal with $A_{1n}$ which we write as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\rho_\mu(V_i, \mu_{\hat{\theta},i};\beta_0)^\top\hat{1}_{il}\left(\hat{\mu}_i - \mu_{\hat{\theta},i}\right) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\sum_{r=1}^J\sum_{s=1}^J \rho_{\mu_r\mu_s}(V_i, \bar{\mu}_i;\beta_0)\hat{1}_{il}\left(\hat{\mu}_{i,r} - \mu_{\hat{\theta},i,r}\right)\left(\hat{\mu}_{i,s} - \mu_{\hat{\theta},i,s}\right) \\
= \; & B_{1n} + B_{2n}, \text{ say,}
\end{aligned}$$

where $\bar{\mu}_i$ lies between $\hat{\mu}_i$ and $\mu_{\hat{\theta},i}$. We deal with $B_{2n}$ first. By Hölder inequality, for $q > 4$ in Assumption 1(iv),

$$\begin{aligned}
\mathbf{E}\left[|B_{2n}|\right] &\leq C\sqrt{n}\left\{\mathbf{E}[\sup_{\bar{\mu}\in[-M,M]}||\rho_{\mu\mu}(V_i, \bar{\mu};\beta_0)||^q]\right\}^{1/q} \\
&\quad \times\left\{\int_{\mathcal{S}_X}\left|\left(\hat{\mu}_r(x) - \mu_{\hat{\theta},r}(x)\right)\left(\hat{\mu}_s(x) - \mu_{\hat{\theta},s}(x)\right)\right|^{\frac{q}{q-1}} dP_X(x)\right\}^{\frac{q-1}{q}},
\end{aligned}$$

where $\mu_{\theta,j}(x) = \mathbf{E}\left[Y_{i,j}|\lambda_j(X_i;\theta) = \lambda_j(x;\theta)\right]$. Note that $\mathbf{E}[\sup_{\bar{\mu}\in[-M,M]}||\rho_{\mu\mu}(V_i, \bar{\mu};\beta_0)||^q] < \infty$ and

$$\begin{aligned}
&\int_{\mathcal{S}_X}\left|\left(\hat{\mu}_r(x) - \mu_{\hat{\theta},r}(x)\right)\left(\hat{\mu}_s(x) - \mu_{\hat{\theta},s}(x)\right)\right|^{\frac{q}{q-1}} dP_X(x) \qquad (12) \\
\leq \; & \int_{\mathcal{D}_{1n}}\left|\left(\hat{\mu}_r(x) - \mu_{\hat{\theta},r}(x)\right)\left(\hat{\mu}_s(x) - \mu_{\hat{\theta},s}(x)\right)\right|^{\frac{q}{q-1}} dP_X(x) \\
&+ \int_{\mathcal{D}_{2n}}\left|\left(\hat{\mu}_r(x) - \mu_{\hat{\theta},r}(x)\right)\left(\hat{\mu}_s(x) - \mu_{\hat{\theta},s}(x)\right)\right|^{\frac{q}{q-1}} dP_X(x),
\end{aligned}$$

where $\mathcal{D}_{1n} = \{x : |F_{n,\hat{\theta},i}(\lambda(x;\hat{\theta})) - 1| > h/2\}$ and $\mathcal{D}_{2n} = \{x : |F_{n,\hat{\theta},i}(\lambda(x;\hat{\theta})) - 1| \leq 2h\}$. Using the steps in (9) and in the proof of Lemma A4 of Song (2009), the first term is bounded by

$$\sup_{u\in[0,1]:|u-1|>h/2} \left| \left(\hat{g}_r(u) - g_{\hat{\theta},s}(u)\right) \left(\hat{g}_s(u) - g_{\hat{\theta},s}(u)\right) \right|^{\frac{q}{q-1}} + O_P(\{n^{-1/2}w_n\}^{\frac{q}{q-1}}) = O_P(w_n^{\frac{2q}{q-1}})$$

where $w_n = n^{-1/2}h^{-1}\sqrt{-\log h} + h^2$ and $g_{\theta,r}(u) = \mathbf{E}[Y_{i,r}|F_{\theta,r}(\lambda_r(X_i;\theta)) = u]$. Similarly, the last term in (12) is bounded by $C\int_{u\in[0,1]:|u-1|\leq 2h} \hat{D}(u)du$, where $\hat{D}(u)$ is equal to

$$\left| \left(\hat{g}_r(u) - g_{\hat{\theta},r}(u)\right) \left(\hat{g}_s(u) - g_{\hat{\theta},s}(u)\right) \right|^{\frac{q}{q-1}} + O_P(\{n^{-1/2}h\}^{\frac{q}{q-1}}).$$

When $|u-1| \leq 2h$, $|(\hat{g}_r(u) - g_{\hat{\theta},r}(u))(\hat{g}_s(u) - g_{\hat{\theta},s}(u))|^{\frac{q}{q-1}} = O_P(h^{\frac{2q}{q-1}})$ uniformly over such $u$'s. (See Lemma A4 of Song (2009).) The Lebesgue measure of such $u$'s is $O(h)$. Hence the last integral in (12) is $O_P(h^{(3q-1)/(q-1)})$. We conclude that $B_{2n} = O_P(n^{1/2}\{w_n^2 + h^{3-1/q}\}) = o_P(1)$ by the condition for bandwidths.

We turn to $B_{1n}$. Suppose that $\hat{\lambda}_{i,J+1} \leq \hat{\lambda}_{l,J+1}$. Then, $u_{\hat{\theta},i,J+1} \leq u_{\hat{\theta},l,J+1}$. Exchanging the roles of $i$ and $l$, we find that if $\hat{\lambda}_{i,J+1} \geq \hat{\lambda}_{l,J+1}$, $u_{\hat{\theta},i,J+1} \geq u_{\hat{\theta},l,J+1}$. Therefore, letting $W_{\theta,i} = (W_{1,i}, u_{\theta,i,J+1})$, we write $1\{\hat{W}_i \leq \hat{W}_l\} = 1\{W_{\theta,i} \leq W_{\theta,l}\}$. Using this, we deduce that

$$B_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \rho_\mu(V_i, \mu_{\hat{\theta},i}; \beta_0)^\top 1\{W_{\hat{\theta},i} \leq W_{\hat{\theta},l}\} \left(\hat{\mu}_i - \mu_{\hat{\theta},i}\right).$$

Choose any $\delta_n \to 0$ such that $\sqrt{n}\delta_n^2 \to 0$ and $n^{1/3}\delta_n \to \infty$, and define

$$\tilde{\nu}_n(\theta, \bar{x}, \bar{w}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta,\bar{x},\bar{w}}(V_i, X_i, D_i, W_{\theta,i})^\top \left(\hat{\mu}_i - \mu_{\theta,i}\right), \ (\theta, x, w) \in B(\theta_0, \delta_n) \times \mathcal{S}_X \times \mathcal{S}_{W_1},$$

where $\psi_{\theta,\bar{x},\bar{w}}(v, x, w) = \rho_\mu(v, \mu_\theta(x); \beta_0) t_{\theta,\bar{x},\bar{w}}(x, d, w)$ and

$$t_{\theta,\bar{x},\bar{w}}(x, d, w) = 1\{w \leq \bar{w}\}1\{d = 1\}1\{F_{\theta,J+1}(\lambda_{J+1}(x;\theta)) \leq F_{\theta,J+1}(\lambda_{J+1}(\bar{x};\theta))\}.$$

Consider $\mathcal{H}_n = \{1\{F_{\theta,J+1}(\lambda_{J+1}(\cdot;\theta)) \leq F_{\theta,J+1}(\lambda_{J+1}(\bar{x};\theta))\} : (\theta, \bar{x}) \in B(\theta_0, \delta_n) \times \mathcal{S}_X\}$. Since the indicator functions are bounded and of bounded variation, we apply Lemma A1 of Song (2009) and Assumption 3(i) to deduce that

$$\log N_{[]}(\varepsilon, \mathcal{H}_n, ||\cdot||_q) \leq C \log \varepsilon + C/\varepsilon, \text{ for } \varepsilon > 0. \tag{13}$$

By Lemma A1 and Assumption 3(i),

$$\left\|\rho_\mu(v,\mu_{\theta_1}(x);\beta_0)-\rho_\mu(v,\mu_{\theta_2}(x);\beta_0)\right\| \leq C\text{sup}_{\bar{\mu}\in[-M,M]}\left\|\rho_{\mu\mu}(v,\bar{\mu};\beta_0)\right\| \times \|\theta_1-\theta_2\|.$$

Therefore, using this, (3) and (13), we conclude that for $\Psi = \{\psi_{\theta,\bar{x},\bar{w}} : (\theta,\bar{x},\bar{w}) \in B(\theta_0,\delta_n)\times \mathcal{S}_X \times \mathcal{S}_{W_1}\}$,

$$\log N_{[]}(\varepsilon,\Psi,||\cdot||_q) \leq C\log\varepsilon + C/\varepsilon, \text{ for } \varepsilon>0. \tag{14}$$

After some algebra (e.g. see the proof of (Step 1) in the proof of Lemma B1 below), we find that $\tilde{\nu}_n(\theta,\bar{x},\bar{w})$ is equal to

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^n D_i \sum_{j=1}^J \mathbf{E}\left[\psi_{\theta,\bar{x},\bar{w},j}(V_i,X_i,D_i,W_{1,i})|u_{\theta,i,j},D_i=1\right]\left(Y_{i,j}-\mu_{\theta,i,j}\right)+o_P(1) \\ = & \frac{1}{\sqrt{n}}\sum_{i=1}^n D_i \sum_{j=1}^J \mathbf{E}\left[\psi_{0,\bar{x},\bar{w},j}(V_i,X_i,D_i,W_{1,i})|u_{i,j},D_i=1\right]\left(Y_{i,j}-\mu_{i,j}\right)+o_P(1), \end{aligned}$$

uniformly over $(\theta,\bar{x},\bar{w})\in B(\theta_0,\delta_n)\times\mathcal{S}_X\times\mathcal{S}_{W_1}$, where $\psi_{\theta,\bar{x},\bar{w},j}$ denotes the $j$-th component of $\psi_{\theta,\bar{x},\bar{w},j}$ and $\psi_{0,\bar{x},\bar{w},j}=\psi_{\theta_0,\bar{x},\bar{w},j}$. (For the equality above, see the proof of (Step 2) in the proof of Lemma B1 below.) Therefore, we conclude that

$$A_{1n}=\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\sum_{j=1}^J \mathbf{E}\left[\psi_{0,\bar{x},\bar{w},j}(V_i,X_i,D_i,W_{1,i})|u_{i,j},D_i=1\right]_{(\bar{x},\bar{w})=(X_l,W_{1,l})}\left(Y_{i,j}-\mu_{i,j}\right)+o_P(1).$$

We turn to $A_{2n}$ which we write as

$$A_{2n}=\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i\psi_{\hat{\theta},X_l,W_{1l}}(V_i,X_i,D_i,W_{1,i})^\top\left(\mu_{\hat{\theta},i}-\mu_i\right).$$

Using previous arguments yielding (14), we can establish a similar bracketing entropy bound for $\mathcal{F}_n=\{\psi_{\theta,\bar{x},\bar{w}}(\cdot,\cdot,\cdot)\left(\mu_\theta(\cdot)-\mu(\cdot)\right) : (\theta,\bar{x},\bar{w})\in B(\theta_0,\delta_n)\times\mathcal{S}_X\times\mathcal{S}_{W_1}\}$. Following the usual stochastic equicontinuity arguments and using Lemma A1, Lemma A2 and Assumption 3(i), we deduce that

$$\begin{aligned} |A_{2n}| &\leq \text{sup}_{(\theta,\bar{x},\bar{w})}\left|\sqrt{n}\mathbf{E}\left[\psi_{\theta,\bar{x},\bar{w}}(V_i,X_i,D_i,W_{1,i})\left(\mu_{\theta,i}-\mu_i\right)\right]\right|+o_P(1) \\ &\leq \sqrt{n}\text{sup}_{(\theta,\bar{x},\bar{w})}\left|\mathbf{E}\left[\psi_{0,\bar{x},\bar{w}}(V_i,X_i,D_i,W_{1,i})\left\{\mu_{\theta,i}-\mu_i\right\}\right]\right| \\ &\quad +O(\sqrt{n}\delta_n^2)+o_P(1)=O(\sqrt{n}\delta_n^2)+o_P(1)=o_P(1), \end{aligned}$$

where the supremum is over $(\theta,\bar{x},\bar{w})\in B(\theta_0,\delta_n)\times\mathcal{S}_X\times\mathcal{S}_{W_1}$. Therefore, letting $\rho_{\mu,i,j}(\beta_0)$ be

the $j$-th entry of $\rho_{\mu,i}(\beta_0)$ and

$$\begin{aligned} z_n(w) &\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\rho_i(\beta_0)1\{W_i \le w\} \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{j=1}^{J} \mathbf{E}\left[\rho_{\mu,i,j}(\beta_0)1\{W_i \le w\}|u_{i,j}, D_i = 1\right]\left(Y_{i,j} - \mu_{i,j}\right), \end{aligned}$$

and collecting the results of $A_{1n}$ and $A_{2n}$, we write

$$\sqrt{n}\xi_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\}) = \frac{1}{n}\sum_{l=1}^{n} D_l \left\{\frac{1}{n}\sum_{i=1}^{n} D_i\rho_\beta(V_i, \hat{\mu}_i; \hat{\beta})\hat{1}_{il}\right\} z_n(W_l) + o_P(1).$$

Since $\sup_{w\in\mathbf{R}^{d_W}}|z_n(w)| = O_P(1)$, using (10) and (11) again, we conclude that

$$\sqrt{n}\xi_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\}) = \frac{1}{n}\sum_{l=1}^{n} D_l\dot{H}(W_l)z_n(W_l) + o_P(1).$$

The wanted result now follows by applying the weak convergence of $z_n$ to $\zeta$ and the continuous mapping theorem (e.g. Theorem 18.11 of van der Vaart (1998).) ■

**Proof of Theorem 2 :** First, define $m(\beta; w) \equiv \mathbf{E}\left[\rho_l(\beta)1\{W_l \le w\}D_l\right]$,

$$\begin{aligned} \hat{m}_b(\beta; \hat{W}_k) &\equiv \frac{1}{n}\sum_{l=1}^{n} D_l\left[\left\{\hat{\rho}_l(\hat{\beta}) - \hat{\rho}_l(\beta)\right\}\hat{1}_{lk} + \omega_{l,b}\left\{\rho_l(\hat{\beta})\hat{1}_{lk} + \hat{r}_{lk}\right\}\right], \text{ and} \\ \tilde{m}_b(\beta; W_k) &\equiv \frac{1}{n}\sum_{l=1}^{n} D_l\left[\left\{\rho_l(\beta_0) - \rho_l(\beta)\right\}1_{lk} + \omega_{l,b}\left\{\rho_l(\beta_0)1_{lk} + r_{lk}\right\}\right], \end{aligned}$$

where $r_{lk} = r_l(W_k)$. Then, we introduce

$$\hat{Q}_b^*(\beta) \equiv \frac{1}{n}\sum_{k=1}^{n} D_k\hat{m}_b(\beta; \hat{W}_k)^2 \text{ and } \tilde{Q}_b^*(\beta) \equiv \frac{1}{n}\sum_{k=1}^{n} D_k\tilde{m}_b(\beta; W_k)^2.$$

We first show that the bootstrap estimator is consistent conditional on $\mathcal{G}_n \equiv \{(V_i, Y_i, X_i, W_{1,i})\}_{i=1}^n$ in probability. (Following the conventions, we use notations $O_{P^*}$ and $o_{P^*}$ that indicate conditional stochastic convergences given $\mathcal{G}_n$.) Define

$$\begin{aligned} \tilde{Q}(\beta) &\equiv \int \left(\mathbf{E}\left[\rho_l(\beta)1\{W_l \le w\}D_l\right]\right)^2 dF_{W,D=1}(w) \\ &\quad + \int \mathbf{E}\left[D_l\{\rho_l(\beta_0) + r_l(w)\}^2\right] dF_{W,D=1}(w). \end{aligned}$$

Then it is not hard to show that uniformly over $\beta \in B$,

$$\tilde{Q}_b^*(\beta) = \tilde{Q}(\beta) + o_{P^*}(1) \text{ in } P.$$

For consistency of $\hat{\beta}_b^*$, it suffices to show that

$$\sup_{\beta \in B} |\hat{Q}_b^*(\beta) - \tilde{Q}_b^*(\beta)| = o_{P^*}(1) \text{ in } P. \tag{15}$$

We write

$$\left|\hat{Q}_b^*(\beta) - \tilde{Q}_b^*(\beta)\right| \leq \max_{1 \leq k \leq n} \left|\hat{m}_b(\beta; \hat{W}_k) - \tilde{m}_b(\beta; W_k)\right| \frac{1}{n} \sum_{k=1}^n \left|\hat{m}_b(\beta; \hat{W}_k) + \tilde{m}_b(\beta; W_k)\right|.$$

As for the last sum, note that

$$\begin{aligned}
& \mathbf{E}\left[\frac{1}{n}\sum_{k=1}^n \left(\hat{m}_b(\beta; \hat{W}_k) + \tilde{m}_b(\beta; W_k)\right)^2 |\mathcal{G}_n\right] \\
\leq\ & \frac{C}{n}\sum_{k=1}^n \left(\frac{1}{n}\sum_{l=1}^n D_l \left\{\hat{\rho}_l(\hat{\beta}) - \hat{\rho}_l(\beta)\right\} \hat{1}_{lk}\right)^2 + \frac{C}{n}\sum_{k=1}^n \left(\frac{1}{n}\sum_{l=1}^n D_l \left\{\rho_l(\hat{\beta}) - \rho_l(\beta)\right\} 1_{lk}\right)^2 \\
& + \frac{C}{n}\sum_{k=1}^n \frac{1}{n}\sum_{l=1}^n D_l \left\{\rho_l(\beta_0) 1_{lk} + r_{lk}\right\}^2 + \frac{C}{n}\sum_{k=1}^n \frac{1}{n}\sum_{l=1}^n D_l \left\{\rho_l(\hat{\beta})\hat{1}_{lk} + \hat{r}_{lk}\right\}^2.
\end{aligned}$$

The all four terms are $O_{P^*}(1)$ in $P$, and hence for (15), it suffices to show that

$$\sup_{\beta \in B} \max_{1 \leq k \leq n} \left|\hat{m}_b(\beta; \hat{W}_k) - \tilde{m}_b(\beta; W_k)\right| = o_{P^*}(1) \text{ in } P. \tag{16}$$

First, we write

$$\hat{m}_b(\beta; \hat{W}_k) - \tilde{m}_b(\beta; W_k) = \frac{1}{n}\sum_{l=1}^n D_l \left[\left\{\hat{\rho}_l(\hat{\beta}) - \hat{\rho}_l(\beta)\right\} \hat{1}_{lk} - \left\{\rho_l(\beta_0) - \rho_l(\beta)\right\} 1_{lk}\right] + \eta_{n,k}, \tag{17}$$

where

$$\eta_{n,k} \equiv \frac{1}{n}\sum_{l=1}^n \omega_{l,b} D_l \left[\rho_l(\hat{\beta})\hat{1}_{lk} - \rho_l(\beta_0) 1_{lk}\right] + \frac{1}{n}\sum_{l=1}^n \omega_{l,b} D_l \left[\hat{r}_{lk} - r_{lk}\right]. \tag{18}$$

It is not hard to show that the first sum in (17) is $o_P(1)$ uniformly in $(\beta, k) \in B \times \{1, \cdots, n\}$ using the similar arguments in the proof of Theorem 1. We show that $\max_{1 \leq k \leq n} \mathbf{E}[\eta_{n,k}^2 | \mathcal{G}_n] =$

$o_P(1)$. For a future use, we show a stronger statement:

$$\max_{1\leq k\leq n}\sqrt{\mathbf{E}[\eta_{n,k}^2|\mathcal{G}_n]}=o_P(n^{-1/2}). \tag{19}$$

Using the fact that $\omega_{l,b}$ is a bounded, mean-zero random variables independent of the data, we find that

$$\begin{aligned}&\mathbf{E}\left[\left(\frac{1}{n}\sum_{l=1}^n\omega_{l,b}D_l\left[\rho_l(\hat{\beta})\hat{1}_{lk}-\rho_l(\beta_0)1_{lk}\right]\right)^2|\mathcal{G}_n\right]\\ =\;&\frac{1}{n^2}\sum_{l=1}^n D_l\left[\rho_l(\hat{\beta})\hat{1}_{lk}-\rho_l(\beta_0)1_{lk}\right]^2.\end{aligned}$$

Following the proof of Theorem 1, we can show that the last sum is $o_P(n^{-1/2})$ uniformly over $1\leq k\leq n$. We focus on the last sum in the definition of $\eta_{n,k}$ in (18). Note that

$$\mathbf{E}\left[\left|\frac{1}{n}\sum_{l=1}^n\omega_{l,b}D_l(\hat{r}_{lk}-r_{lk})\right|^2|\mathcal{G}_n\right]\leq\frac{1}{n^2}\sum_{l=1}^n||\hat{r}_{lk}-r_{lk}||^2=o_P(n^{-1})$$

uniformly over $1\leq k\leq n$, by Assumption 4. Therefore, we obtain (19). This yields the following:

$$\sup_{(\beta,w)\in B\times\mathbf{R}^{d_W}}\max_{1\leq k\leq n}|\hat{m}_b(\beta;\hat{W}_k)-\tilde{m}_b(\beta;W_k)|=o_{P^*}(1)\text{ in }P.$$

From this, we deduce (16) and that $\hat{\beta}_b^*=\beta_0+o_{P^*}(1)$ in $P$. Clearly, $\hat{\beta}_b^*=\hat{\beta}+o_{P^*}(1)$ in $P$, because $\hat{\beta}$ is consistent.

Now, we turn to the bootstrap distribution of $\hat{\beta}_b^*$. As in the proof of Theorem 1, we can write

$$\sqrt{n}\{\hat{\beta}_b^*-\hat{\beta}\}=G_n^*(\hat{\beta},\hat{\mu},\{\hat{W}_l\})^{-1}\sqrt{n}\xi_n^*(\hat{\beta}_b^*,\hat{\mu},\{\hat{W}_l\}),$$

where

$$\begin{aligned}G_n^*(\hat{\beta}_b^*,\hat{\mu},\{\hat{W}_l\})&=\frac{1}{n}\sum_{l=1}^n D_l\left\{\frac{1}{n}\sum_{i=1}^n D_i\rho_\beta(V_i,\hat{\mu}_i;\hat{\beta}_b^*)\hat{1}_{il}\right\}\left\{\frac{1}{n}\sum_{i=1}^n D_i\rho_\beta^\top(V_i,\hat{\mu}_i;\bar{\beta}_b^*)\hat{1}_{il}\right\}\text{ and}\\ \xi_n^*(\hat{\beta}_b^*,\hat{\mu},\{\hat{W}_l\})&=\frac{1}{n}\sum_{l=1}^n D_l\left\{\frac{1}{n}\sum_{i=1}^n D_i\rho_\beta(V_i,\hat{\mu}_i;\hat{\beta}_b^*)\hat{1}_{il}\right\}\left\{\frac{1}{n}\sum_{i=1}^n D_i\omega_{i,b}\left\{\rho_i(\hat{\beta})\hat{1}_{ik}+\hat{r}_{ik}\right\}\right\},\end{aligned}$$

and $\bar{\beta}_b^*$ lies between $\hat{\beta}_b^*$ and $\hat{\beta}$. Again, similarly as in the proof of Theorem 1, we can show

that

$$\begin{aligned} G_n^*(\hat{\beta}_b^*, \hat{\mu}, \{\hat{W}_l\}) &= G_n(\beta_0, \mu, \{W_l\}) + o_{P^*}(1) \text{ in } P \\ &= \int \dot{H}(w)\dot{H}(w)^\top dF_{W,D=1}(w) + o_P(1) + o_{P^*}(1) \text{ in } P. \end{aligned}$$

Note that the only difference here is that we have $\hat{\beta}_b^*$ in place of $\hat{\beta}$. However, $\hat{\beta}_b^*$ is consistent for $\beta_0$ just as $\hat{\beta}$ is, yielding the first equality in the above.

As for $\xi_n^*(\hat{\beta}_b^*, \hat{\mu}, \{\hat{W}_l\})$, note that by (19),

$$\begin{aligned} \sqrt{n}\xi_n^*(\hat{\beta}_b^*, \hat{\mu}, \{\hat{W}_l\}) &= \frac{1}{\sqrt{n}} \sum_{k=1}^n D_k \left\{ \frac{1}{n} \sum_{i=1}^n D_i \rho_\beta(V_i, \hat{\mu}_i; \hat{\beta}_b^*) \hat{1}_{ik} \right\} \\ &\quad \times \left\{ \frac{1}{n} \sum_{i=1}^n D_i \omega_{i,b} \left\{ \rho_i(\beta_0) 1_{ik} + r_{ik} \right\} \right\} + o_{P^*}(1) \text{ in } P. \end{aligned}$$

Similarly as in the proof of Theorem 2, the leading term above is equal to

$$\frac{1}{n} \sum_{k=1}^n D_k \dot{H}(W_k) \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n D_i \omega_{i,b} \left\{ \rho_i(\beta_0) 1_{ik} + r_{ik} \right\} \right\} + o_{P^*}(1) \text{ in } P.$$

Let $\Gamma_n(f) = \frac{1}{n} \sum_{i=1}^n f(W_i) D_i$ and $\Gamma(f) = \int f(w) dF_{W,D=1}(w)$. Choose any sequence $f_n : \mathbf{R}^{d_W} \to \mathbf{R}^k$ such that $\sup_w ||f_n(w) - f(w)|| \to 0$, for some $f$ such that $\mathbf{E}\left[||f(W_i)||D_i\right] < \infty$. Then we have

$$\begin{aligned} \Gamma_n(f_n) - \Gamma(f) &= \frac{1}{n} \sum_{i=1}^n \left( f_n(W_i) - f(W_i) \right) D_i + \frac{1}{n} \sum_{i=1}^n f(W_i) D_i - \mathbf{E}\left[ f(W_i) D_i \right] \\ &= o(1) + o_{a.s.}(1), \end{aligned}$$

by the strong law of large numbers. Let

$$F_n(w; \mathcal{G}_n) = \frac{1}{\sqrt{n}} \sum_{l=1}^n \omega_{l,b} D_l \left[ \rho_l(\beta_0) 1\{W_l \leq w\} + r_l(w) \right] \times \dot{H}(w).$$

Now, by the conditional multiplier central limit theorem of Ledoux and Talagrand (1988), conditional on almost every sequence in $\mathcal{G}_\infty$,

$$F_n(\cdot; \mathcal{G}_n) \Longrightarrow \zeta.$$

Therefore, by the almost sure representation theorem (e.g. Theorem 6.7 of Billingsley

(1999)), there is a sequence $\tilde{F}_n(\cdot)$ such that $\tilde{F}_n(\cdot)$ is distributionally equivalent to $F_n(\cdot)$ and $\tilde{F}_n(\cdot) \to_{a.s.} \zeta$ conditional on almost every sequence $\mathcal{G}_n$. Then, by the previous arguments, conditional on almost every sequence $\{S_l\}_{l=1}^n$, we have

$$\Gamma_n(\tilde{F}_n(\cdot; \mathcal{G}_n)) \to_{a.s.} \int \zeta(w)\dot{H}(w)dF_{W,D=1}(w).$$

Hence the proof is complete. ■

## 7.4 Uniform Representation of Sample Linear Functionals of SNN Estimators

In this section, we present a uniform representation of sums of SNN estimators that is uniform over function spaces. Stute and Zhu (2005) obtained a non-uniform result in a different form. Their proof uses the oscillation results for smoothed empirical processes. Since we do not have such a result under the generality assumed in this paper, we take a different approach here.

Suppose that we are given a random sample $\{(Z_i, X_i, Y_i)\}_{i=1}^n$ drawn from the distribution of a random vector $S = (Z, X, Y) \in \mathbf{R}^{d_Z+d_X+J}$. Let $\mathcal{S}_Z, \mathcal{S}_X$ and $\mathcal{S}_Y$ be the supports of $Z, X$, and $Y$ respectively. Let $\Lambda$ be a class of $\mathbf{R}$-valued functions on $\mathbf{R}^{d_X}$ with generic elements denoted by $\lambda$. We also let $\Phi$ and $\Psi$ be classes of real functions on $\mathbf{R}^J$ and $\mathbf{R}^{d_Z}$ with generic elements $\varphi$ and $\psi$. We fix $\lambda_0 \in \Lambda$ such that $\lambda_0(X)$ is a continuous random variable. Then we focus on $g_\varphi(u) = \mathbf{E}[\varphi(Y)|U = u]$, where $U = F_0(\lambda_0(X))$ and $F_0(\cdot)$ is the CDF of $\lambda_0(X)$. Similarly, we define $g_\psi(u) = \mathbf{E}[\psi(Z)|U = u]$. Letting $F_\lambda(\cdot)$ be the CDF of $\lambda(X)$, we denote $U_\lambda = F_\lambda(\lambda(X))$. We define $f_\lambda(y|u_0, u_1)$ and $h_\lambda(z|u_0, u_1)$ to be the conditional densities of $Y$ given $(U, U_\lambda) = (u_0, u_1)$ and $Z$ given $(U, U_\lambda) = (u_0, u_1)$ with respect to some $\sigma$-finite measures, and let

$$\begin{aligned} \mathcal{P}_Y &\equiv \{f_\lambda(y|\cdot, \cdot) : (\lambda, y) \in \Lambda_n \times \mathcal{S}_Y\} \text{ and} \\ \mathcal{P}_Z &\equiv \{h_\lambda(z|\cdot, \cdot) : (\lambda, y) \in \Lambda_n \times \mathcal{S}_Z\}. \end{aligned}$$

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

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

where $\hat{f}_{\lambda,i}(u) = (n-1)^{-1}\sum_{j=1,j\neq i}^n K_h(U_{n,\lambda,j} - u)$. Introduce $\Lambda_n = \{\lambda \in \Lambda : ||F_\lambda \circ \lambda - F_0 \circ \lambda_0||_\infty \leq n^{-b}\}$ for $b \in (1/4, 1/2]$. The semiparametric process of focus takes the following

form:

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

with $(\lambda, \varphi, \psi) \in \Lambda_n \times \Phi_n \times \Psi_n$.

**Assumption B1 :** (i) Classes $\Phi$ and $\Psi$ for some $C > 0,\ p > 8$, and $b_\Psi,\ b_\Phi \in (0, 6/5)$,

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

and envelopes $\tilde{\varphi}$ and $\tilde{\psi}$ satisfy that $\mathbf{E}[|\tilde{\varphi}(Y)|^p] < \infty$ and $\mathbf{E}[|\tilde{\psi}(Z)|^p] < \infty$, and $\sup_{u\in[0,1]} \mathbf{E}[|\tilde{\varphi}(Y)||U = u] < \infty$.
(ii) For $\Lambda_n^F = \{F_\lambda \circ \lambda : \lambda \in \Lambda_n\}$, some $b_\Lambda \in (0, 1)$ and $C > 0$,

$$\log N(\varepsilon, \Lambda_n^F, ||\cdot||_\infty) \leq C\varepsilon^{-b_\Lambda}, \text{ for each } \varepsilon > 0.$$

**Assumption B2 :** (i) $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$ and $\mathcal{P}_Z$ is regular for $\tilde{\psi}$.
(ii) $g_\varphi(\cdot)$ is twice continuously differentiable with derivatives bounded uniformly over $\varphi \in \Phi$.

**Assumption B3 :** (i) $K(\cdot)$ is symmetric, compact supported, twice continuously differentiable with bounded derivatives, and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^{3-1/p} + n^{-1/2}h^{-2-1/p}(-\log h) \to 0$.

The following lemma offers a uniform representation of $\nu_n$.

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

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

*Furthermore, the representations remain the same when we replace $\nu_n(\lambda, \varphi, \psi)$ by $\nu_n(\lambda_0, \varphi, \psi)$.*

**Proof of Lemma B1 :** To make the flow of the arguments more visible, the proof proceeds by making certain claims which involve extra arguments and are proved 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 $\mathbf{E}_{S_i}$ indicates the conditional expectation given $S_i$.

Let $g_{\varphi,\lambda}(u) \equiv \mathbf{E}[\varphi(Y)|U_\lambda = u]$ and $g_{\psi,\lambda}(u) \equiv \mathbf{E}[\psi(Z)|U_\lambda = u]$. Define

$$\Delta_i^{\varphi,\psi}(\lambda) \equiv g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(Y_i) - g_{\varphi,\lambda}(U_{\lambda,i})\}.$$

The proof proceeds in the following two steps.

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

**Step 2 :** $\sup_{(\lambda,\varphi,\psi)\in\Lambda_n\times\Phi\times\Psi}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\Delta_i^{\varphi,\psi}(\lambda)-\Delta_i^{\varphi,\psi}(\lambda_0)\right\}\right|=o_P(1)$.

Then the wanted statement follows by chaining Steps 1 and 2.

**Proof of Step 1 :** Define $\hat{\rho}_{\varphi,\lambda,i}(t)\equiv(n-1)^{-1}\sum_{j=1,j\neq i}^n K_h(U_{n,\lambda,j}-t)\varphi(Y_j)$ and write $\hat{g}_{\varphi,\lambda,i}(U_{n,\lambda,i})-g_{\varphi,\lambda}(U_{\lambda,i})$ as

$$\begin{aligned}R_{1i}(\lambda,\varphi) &\equiv \frac{\hat{\rho}_{\varphi,\lambda,i}(U_{n,\lambda,i})-g_{\varphi,\lambda}(U_{\lambda,i})\hat{f}_{\lambda,i}(U_{n,\lambda,i})}{f_\lambda(U_{\lambda,i})}\\ &\quad+\frac{[\hat{\rho}_{\varphi,\lambda,i}(U_{n,\lambda,i})-g_{\varphi,\lambda}(U_{\lambda,i})\hat{f}_{\lambda,i}(U_{n,\lambda,i})](f_\lambda(U_{\lambda,i})-\hat{f}_{\lambda,i}(U_{n,\lambda,i}))}{\hat{f}_{\lambda,i}(U_{n,\lambda,i})f_\lambda(U_{\lambda,i})}\\ &= R_{1i}^A(\lambda,\varphi)+R_{1i}^B(\lambda,\varphi),\ \text{say.}\end{aligned}$$

where $f_\lambda(u)=1\{u\in[0,1]\}$. Put $\pi=(\lambda,\varphi,\psi)$ and $\Pi_n=\Lambda_n\times\Phi\times\Psi$, and write

$$\begin{aligned}\nu_n(\pi) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(Z_i)R_{1i}^A(\lambda,\varphi)+\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(Z_i)R_{1i}^B(\lambda,\varphi)\\ &= r_{1n}^A(\pi)+r_{1n}^B(\pi),\ \pi\in\Pi_n,\ \text{say.}\end{aligned}$$

From the proof of Lemma A3 of Song (2009) (by replacing $\lambda$ and $\lambda_0$ with $F_\lambda\circ\lambda$ there and using Assumption B1(ii)), it follows that

$$\max_{1\leq i\leq n}\sup_{\lambda\in\Lambda_n}\sup_{x\in\mathbf{R}^{d_X}}|F_{n,\lambda,i}(\lambda(x))-F_\lambda(\lambda(x))|=O_P(n^{-1/2}),\tag{20}$$

where $F_{n,\lambda,i}(\bar{\lambda})=\frac{1}{n-1}\sum_{j=1,j\neq i}^n 1\{\lambda(X_j)\leq\bar{\lambda}\}$. Using (20) and employing similar arguments around (12) in the proof of Theorem 1, we can show that $\sup_{\pi\in\Pi_n}\left|r_{1n}^B(\pi)\right|=o_P(1)$.

We turn to $r_{1n}^A(\pi)$, which we write as

$$\begin{aligned}&\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\psi_i\Delta_{\varphi,\lambda,ij}K_{ij}^\lambda+\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\psi_i\Delta_{\varphi,\lambda,ij}\{K_{n,ij}^\lambda-K_{ij}^\lambda\}\\ &= R_{1n}(\pi)+R_{2n}(\pi),\ \text{say,}\end{aligned}$$

where $\psi_i=\psi(Z_i)$, $\Delta_{\varphi,\lambda,ij}=\varphi(Y_j)-g_{\varphi,\lambda}(U_{\lambda,i})$, $K_{n,ij}^\lambda=K_h(U_{n,\lambda,j}-U_{n,\lambda,i})$ and $K_{ij}^\lambda=K_h(U_{\lambda,j}-U_{\lambda,i})$. We will now show that

$$\sup_{\pi\in\Pi_n}|R_{2n}(\pi)|\to_P 0.\tag{21}$$

Let $\delta_i^\lambda = U_{n,\lambda,i} - U_{\lambda,i}$ and $d_{\lambda,ji} = \delta_j^\lambda - \delta_i^\lambda$ and write $R_{2n}(\pi)$ as

$$\begin{aligned} & \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} d_{\lambda,ji} + \frac{1}{2(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} d_{\lambda,ji}^2 K''_{h,ij} \\ = \;& A_{1n}(\pi) + A_{2n}(\pi), \text{ say,} \end{aligned}$$

where $K'_{h,ij} = h^{-2}\partial K(t)/\partial t$ at $t = (U_{\lambda,i} - U_{\lambda,j})/h$ and

$$K''_{h,ij} = h^{-3}\partial^2 K(t)/\partial t^2$$

at $t = \{(1-a_{ij})(U_{\lambda,i} - U_{\lambda,j}) + a_{ij}(U_{n,\lambda,i} - U_{n,\lambda,j})\}/h$, for some $a_{ij} \in [0,1]$. Later we will show the following:

**C1 :** $\sup_{\pi\in\Pi_n} |A_{2n}(\pi)| = o_P(1)$.

We turn to $A_{1n}(\pi)$ which we write as

$$\begin{aligned} & \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} \delta_j^\lambda - \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} \delta_i^\lambda \quad (22) \\ = \;& B_{1n}(\pi) + B_{2n}(\pi), \text{ say.} \end{aligned}$$

Write $B_{1n}(\pi)$ as (up to $O(n^{-1})$)

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^{n} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} - \mathbf{E}\left[ \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} | U_{\lambda,j} \right] \right\} \right] (U_{n,\lambda,j} - U_{\lambda,j}) \\ & + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbf{E}\left[ \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} | U_{\lambda,j} \right] (U_{n,\lambda,j} - U_{\lambda,j}) = C_{1n}(\pi) + C_{2n}(\pi), \text{ say.} \end{aligned}$$

As for $C_{1n}(\pi)$, we show the following later.

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

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

$$\begin{aligned} A_{1n}(\pi) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbf{E}\left[ \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} | U_{\lambda,j} \right] (U_{n,\lambda,j} - U_{\lambda,j}) \quad (23) \\ & \quad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{E}\left[ \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} | U_{\lambda,i} \right] (U_{n,\lambda,i} - U_{\lambda,i}) + o_P(1) \\ &= D_{1n}(\pi) - D_{2n}(\pi) + o_P(1), \text{ say.} \end{aligned}$$

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),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,j}=u_1\right](1\{U_{\lambda,i}\le u_1\}-u_1)\,du_1+o_P(1).$$

Using the symmetry of $K$, we deduce that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,j}=u_1\right](1\{U_{\lambda,i}\le u_1\}-u_1)\,du_1\\
=\;&\frac{1}{h^2\sqrt{n}}\sum_{i=1}^{n}\int_0^1\int_0^1 g_{\psi,\lambda}(u_2)\left\{g_{\varphi,\lambda}(u_1)-g_{\varphi,\lambda}(u_2)\right\}K'\left(\frac{u_1-u_2}{h}\right)du_2\,(1\{U_{\lambda,i}\le u_1\}-u_1)\,du_1\\
=\;&\frac{1}{h^2\sqrt{n}}\sum_{i=1}^{n}\int_0^1\int_0^1 g_{\psi,\lambda}(u_2)\left\{g_{\varphi,\lambda}(u_2)-g_{\varphi,\lambda}(u_1)\right\}K'\left(\frac{u_2-u_1}{h}\right)du_2\,(1\{U_{\lambda,i}\le u_1\}-u_1)\,du_1.
\end{aligned}$$

As for $D_{2n}(\pi)$, we also observe that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,i}=u_1\right](1\{U_{\lambda,j}\le u_1\}-u_1)\,du_1\\
=\;&\frac{1}{h^2\sqrt{n}}\sum_{i=1}^{n}\int_0^1\int_0^1 g_{\psi,\lambda}(u_1)\left\{g_{\varphi,\lambda}(u_2)-g_{\varphi,\lambda}(u_1)\right\}K'\left(\frac{u_2-u_1}{h}\right)du_2\,(1\{U_{\lambda,j}\le u_1\}-u_1)\,du_1.
\end{aligned}$$

Write the sum above as

$$\begin{aligned}
&\frac{1}{h^2\sqrt{n}}\sum_{j=1}^{n}\int_0^1\int_0^1 g_{\psi,\lambda}(u_2)\left\{g_{\varphi,\lambda}(u_2)-g_{\varphi,\lambda}(u_1)\right\}K'\left(\frac{u_2-u_1}{h}\right)du_2\,(1\{U_{\lambda,j}\le u_1\}-u_1)\,du_1\\
&+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^1\phi_n(u_1;\pi)\,(1\{U_{\lambda,j}\le u_1\}-u_1)\,du_1,
\end{aligned}$$

where

$$\phi_n(u_1;\pi)=\frac{1}{h^2}\int_0^1\left\{g_{\psi,\lambda}(u_1)-g_{\psi,\lambda}(u_2)\right\}\left\{g_{\varphi,\lambda}(u_2)-g_{\varphi,\lambda}(u_1)\right\}K'\left(\frac{u_2-u_1}{h}\right)du_2.$$

Note that $\sup_{\pi\in\Pi_n}|\phi_n(u_1;\pi)|=O(h)$ by using the first order differentiability of $g_{\psi,\lambda}$ and

$g_{\varphi,\lambda}$. Therefore,

$$\sup_{\pi\in\Pi_n}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^1\phi_n(u_1;\pi)\left(1\{U_{\lambda,j}\le u_1\}-u_1\right)du_1\right|=o_P(1).$$

We conclude that $D_{1n}(\pi)=D_{2n}(\pi)+o_P(1)$ uniformly over $\pi\in\Pi_n$, and that $\sup_{\pi\in\Pi_n}|A_{1n}(\pi)|=o_P(1)$, which, together with (C1), completes the proof of (21).

It suffices for (Step 1) to show that

$$\sup_{\pi\in\Pi_n}\left|R_{1n}(\pi)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(Y_i)-g_{\varphi,\lambda}(U_{\lambda,i})\}\right|=o_P(1). \tag{24}$$

We define $q^{\pi}_{n,ij}\equiv q^{\pi}_n(S_i,S_j)\equiv\psi_i\Delta_{\varphi,\lambda,ij}K^{\lambda}_{ij}$ and write $R_{1n}(\pi)$ as

$$\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}q^{\pi}_{n,ij}. \tag{25}$$

Let $\rho^{\pi}_{n,ij}\equiv\rho^{\pi}_n(S_i,S_j)\equiv q^{\pi}_{n,ij}-\mathbf{E}_{S_i}[q^{\pi}_{n,ij}]-\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]+\mathbf{E}[q^{\pi}_{n,ij}]$ and define

$$u_n(\pi)\equiv\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\rho^{\pi}_{n,ij}.$$

Then, $\{u_n(\cdot),\pi\in\Pi_n\}$ is a degenerate $U$-process. We write (25) as

$$\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left\{\mathbf{E}_{S_i}[q^{\pi}_{n,ij}]+\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]-\mathbf{E}[q^{\pi}_{n,ij}]\right\}+u_n(\pi). \tag{26}$$

We will later show the following two claims.

**C3 :** $\sup_{\pi\in\Pi_n}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\mathbf{E}_{S_i}[q^{\pi}_{n,ij}]-\mathbf{E}[q^{\pi}_{n,ij}]\}|=o_P(1)$.
**C4 :** $\sup_{\pi\in\Pi_n}|u_n(\pi)|=o_P(1)$.

We conclude from these claims that

$$\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}q^{\pi}_{n,ij}=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]+o_P(1).$$

Then the proof of Step 1 is completed by showing the following.

**C5:** $\sup_{\pi\in\Pi_n}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]-g_{\psi,\lambda}(U_{\lambda,j})\{\varphi(Y_j)-g_{\varphi,\lambda}(U_{\lambda,j})\}\right)\right|=o_P(1)$.

**Proof of C1 :** First observe that $\max_{1\leq i,j\leq n}\sup_{\lambda\in\Lambda_n}||d^2_{\lambda,ji}|| = O_P(n^{-1})$ by (20). Let $b \in (1/4, 1/2]$ be as defined in the definition of $\Lambda_n$. Let $\tilde{\Delta}_{ij} = \tilde{\varphi}(Y_i) + \mathbf{E}[\tilde{\varphi}(Y_j)|U_j] + Mn^{-b}$. With large probability along with large $M > 0$, we bound $|A_{2n}(\pi)|$ by

$$\frac{Cn^{-1}}{2(n-1)\sqrt{n}}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\left|\tilde{\psi}_i\tilde{\Delta}_{ij}K''_{h,ij}\right| \leq \frac{1}{\sqrt{n}}\frac{C}{2n(n-1)h^3}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\left|\tilde{\psi}_i\tilde{\Delta}_{ij}\right|1_{ij},$$

where $1_{ij} = 1\left\{|U_i - U_j| \leq h + Cn^{-b}\right\}$. We bound the last term again by

$$\frac{1}{\sqrt{n}}\frac{C}{2n(n-1)h^3}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\left\{\left|\tilde{\psi}_i\tilde{\Delta}_{ij}\right|1_{ij} - \mathbf{E}\left[\left|\tilde{\psi}_i\tilde{\Delta}_{ij}\right|1_{ij}\right]\right\} + \frac{C\mathbf{E}\left[\left|\tilde{\psi}_i\tilde{\Delta}_{ij}\right|1_{ij}\right]}{2h^3\sqrt{n}}.$$

The leading term is $O_P(n^{-1}h^{-3}) = o_P(n^{-1/2}h^{-3/2}) = o_P(1)$ using the standard $U$ statistics theory. Through using Hölder inequality, we find that the second term is equal to $O(n^{-1/2}h^{-2-1/p}) = o(1)$.

**Proof of C2 :** Note that $K'(\cdot/h)$ is uniformly bounded and bounded variation. Let $\mathcal{K}_{1,\Lambda} = \{K'(\sigma(\cdot)/h) : \sigma \in \mathcal{I}_n\}$, where $\mathcal{I}_n = \{\sigma_{\lambda,u} : (\lambda, u) \in \Lambda_n \times [0,1]\}$ and $\sigma_{\lambda,u}(x) = (F_\lambda \circ \lambda)(x) - u$. By Lemma A1 of Song (2009) and Assumption B1(ii),

$$\log N_{[]}(\varepsilon, \mathcal{K}_{1,\Lambda}, ||\cdot||_p) \leq \log N(C\varepsilon, \mathcal{I}_n, ||\cdot||_\infty) + C/\varepsilon \leq C\varepsilon^{-b_\Lambda}. \tag{27}$$

Using (27) and following standard arguments, we can show that

$$\begin{aligned}
&\max_{1\leq j\leq n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij} - \mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,j}, U_j\right]\right\}\right| \\
\leq\ & \frac{1}{h^2}\sup_{(\pi,k)\in\Pi_n\times\mathcal{K}_{1,\Lambda}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\psi_i\Delta_{\varphi,\lambda,ij}k(X_j) - \mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}k(X_j)|U_{\lambda,j}, U_j\right]\right\}\right| = O_P(h^{-2}).
\end{aligned}$$

By the fact that $\max_{1\leq j\leq n}||\delta^\lambda_j|| = O_P(n^{-1/2})$, the wanted result follows because $O_P(n^{-1/2}h^{-2}) = o_P(1)$.

**Proof of C3 :** First we note that

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\pi\in\Pi_n}\left|\mathbf{E}_{S_i}[q^\pi_{n,ij}]\right|^2\right] \\
\leq\ & \int_0^1\left\{g^2_{\tilde{\psi},\lambda_0}(t_1) + Cn^{-2b}\right\}\sup_{(\varphi,\lambda)\in\Phi\times\Lambda_n}\left[\int_0^1\{g_{\varphi,\lambda}(t_2) - g_{\varphi,\lambda}(t_1)\}K_h(t_2 - t_1)dt_2\right]^2 dt_1.
\end{aligned} \tag{28}$$

By change of variables, the integral inside the bracket becomes

$$\int_{\{-t_1/h\}\vee(-1)}^{(1-t_1)/h\}\wedge 1} \{g_{\varphi,\lambda}(t_1+ht_2)-g_{\varphi,\lambda}(t_1)\}K(t_2)dt_2.$$

After tedious algebra, we can show that the expectation in (28) is $O(h^3)$. This implies that we take an envelope, say, $J$ of the class $\mathcal{J}_n \equiv \{h\mathbf{E}[q^\pi_{n,ij}|S_i=\cdot] : \pi \in \Pi_n\}$ such that $||J||_2 = O(h^{3/2+1})$ as $n\to\infty$. Similarly as in the proof of C2, note that $K(\cdot/h)$ is uniformly bounded and bounded variation. Let $\mathcal{K}_\Lambda = \{K(\sigma(\cdot)/h) : \sigma \in \mathcal{I}_n\}$. Then by Lemma A1 of Song (2009), for any $p \geq 1$,

$$\log N_{[]}(\varepsilon, \mathcal{K}_\Lambda, ||\cdot||_p) \leq \log N(\varepsilon, \mathcal{I}_n, ||\cdot||_\infty) + C/\varepsilon \leq C\varepsilon^{-b_\Lambda}. \tag{29}$$

Let us define $\tilde{\mathcal{J}}_n = \{hq^\pi_n(\cdot,\cdot) : \pi \in \Pi_n\}$, where $q^\pi_n(\cdot\cdot,\cdot)$ is defined prior to (25). Observe that for any $\lambda_1, \lambda_2 \in \Lambda_n$,

$$\begin{aligned} \|g_{\varphi,\lambda_1}(F_{\lambda_1}(\lambda_1(\cdot))) - g_{\varphi,\lambda_2}(F_{\lambda_2}(\lambda_2(\cdot)))\|_\infty &\leq C||(F_{\lambda_1}\circ\lambda_1) - (F_{\lambda_2}\circ\lambda_2)||_\infty \text{ and} \\ \|g_{\psi,\lambda_1}(F_{\lambda_1}(\lambda_1(\cdot))) - g_{\psi,\lambda_2}(F_{\lambda_2}(\lambda_2(\cdot)))\|_\infty &\leq C||(F_{\lambda_1}\circ\lambda_1) - (F_{\lambda_2}\circ\lambda_2)||_\infty, \end{aligned} \tag{30}$$

by Lemma A1. From this and using the fact that $\mathcal{K}_\Lambda$ is uniformly bounded, it is easy to show that

$$\log N_{[]}(\varepsilon, \tilde{\mathcal{J}}_n, ||\cdot||_{p/2}) \leq \log N_{[]}(\varepsilon/C, \Phi, ||\cdot||_p) + \log N_{[]}(\varepsilon/C, \Psi, ||\cdot||_p) + C\varepsilon^{-b_\Lambda}. \tag{31}$$

Therefore, $\log N_{[]}(\varepsilon, \tilde{\mathcal{J}}_n, ||\cdot||_{p/2}) \leq C\varepsilon^{-(b_\Phi \vee b_\Psi \vee b_\Lambda)}$. Using this result, we obtain that

$$\log N_{[]}(\varepsilon, \mathcal{J}_n, ||\cdot||_{p/2}) \leq C\varepsilon^{-(b_\Phi \vee b_\Psi \vee b_\Lambda)}.$$

Then by the maximal inequality of Pollard (1989) (e.g. Theorem A.2 of van der Vaart (1996)),

$$\begin{aligned} & \mathbf{E}\left[\sup_{\pi\in\Pi_n}\left|\frac{h}{\sqrt{n}}\sum_{i=1}^n \left\{\mathbf{E}_{S_i}[q^\pi_{n,il}] - \mathbf{E}[q^\pi_{n,il}]\right\}\right|\right] \\ \leq\ & C\int_0^{O(h^{(3/2)+1})} \sqrt{1+\log N_{[]}(\varepsilon, \mathcal{J}_n, ||\cdot||_2)}d\varepsilon = O(h^{(5/2)\times\{1-(b_\Phi\vee b_\Psi\vee b_\Lambda)/2\}}) = o(h), \end{aligned}$$

because $(b_\Phi \vee b_\Psi \vee b_\Lambda) < 6/5$. Hence we obtain the wanted result.

**Proof of C4 :** Since $p > 8$, we can take $\Delta \in (0, 1/6)$ and $\eta = 1/4 + \Delta/2$ such that

$n^{-\eta+\Delta/2}h^{-1} \to 0$, $\eta + 1/2 \leq 1 - 1/p$ and $(b_\Phi \vee b_\Psi \vee b_\Lambda)(1/2+\eta) < 1$. Then, from the proof of C3,

$$\int_0^1 \left\{\log N_{[]}(\varepsilon, \tilde{\mathcal{J}}_n, ||\cdot||_{p/2})\right\}^{(1/2+\eta)} d\varepsilon \leq \int_0^1 C\varepsilon^{-(b_\Phi \vee b_\Psi \vee b_\Lambda)\{1/2+\eta\}} d\varepsilon < \infty.$$

By Theorem 1 of Turki-Moalla (1998), p.878,

$$h \sup_{\pi \in \Pi_n} |u_{1n}(\pi)| = o_P(n^{1/2-(1/2+\eta)+\Delta/2}) = o_P(n^{-\eta+\Delta/2}).$$

Therefore, $\sup_{\pi\in\Pi_n} |u_{1n}(\pi)| = o_P(n^{-\eta+\Delta/2}h^{-1}) = o_P(n^{-1/4}h^{-1}) = o_P(1)$. Hence the proof is complete.

**Proof of C5** : We consider the following:

$$\begin{aligned} & \mathbf{E}\left[\sup_{\pi\in\Pi_n} \left\{\mathbf{E}_{S_j}[q^\pi_{n,ij}] - g_{\psi,\lambda}(U_{\lambda,j})\{\varphi(Y_j) - g_{\varphi,\lambda}(U_{\lambda,j})\}\right\}^2\right] \qquad (32) \\ = & \int \sup_{\pi\in\Pi_n} \left\{\int_0^1 A_{n,\pi}(t_1, t_2, y) dt_1\right\}^2 dF_{Y,\lambda}(y, t_2), \end{aligned}$$

where $\int \cdot dF_{Y,\lambda}$ denotes the integration with respect to the joint distribution of $(Y_i, U_{\lambda,i})$ and

$$\begin{aligned} A_{n,\pi}(t_1, t_2, y) &= g_{\psi,\lambda}(t_1)\{\varphi(y) - g_{\varphi,\lambda}(t_1)\}K_h(t_1 - t_2) \\ & \quad - g_{\psi,\lambda}(t_2)\{\varphi(y) - g_{\varphi,\lambda}(t_2)\}. \end{aligned}$$

After some tedious algebra, we can show that the last term in (32) is $O(h^3)$ (see the proof of C3). Following the proof of C3 similarly, we can obtain the wanted result.

**Proof of Step 2 :** The proof is based on standard arguments of stochastic equicontinuity (Andrews (1994)). For the proof, it suffices to show that the class

$$\mathcal{G} = \{g_{\psi,\lambda}(F_\lambda(\lambda(\cdot)))\{\varphi(\cdot) - g_{\varphi,\lambda}(F_\lambda(\lambda(\cdot)))\} : (\lambda, \varphi, \psi) \in \Lambda_n \times \Phi \times \Psi\}$$

has a finite integral bracketing entropy with an $L_{2+\varepsilon}(P)$-bounded envelope for some $\varepsilon > 0$. Using (30) and standard arguments, we find that

$$\log N_{[]}(\varepsilon, \mathcal{G}, ||\cdot||_{p/2}) \leq C\varepsilon^{-(b_\Phi \vee b_\Psi \vee b_\Lambda)}.$$

Since $b_\Phi \vee b_\Psi \vee b_\Lambda < 2$, the wanted bracketing integral entropy condition follows. We take

an envelope as

$$F_M(x,y) = \{g_{\tilde{\psi},\lambda_0}(F_0(\lambda_0(x))) + Mn^{-b}\}\{\tilde{\varphi}(y) + g_{\tilde{\varphi},\lambda_0}(F_0(\lambda_0(x))) + Mn^{-b}\}$$

for some large $M$. Clearly, this function $F_M$ is $L_{2+\varepsilon}(P)$-bounded by Assumption B1. Therefore, the process

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\Delta_i^{\varphi,\psi}(\lambda) - \Delta_i^{\varphi,\psi}(\lambda_0) - \mathbf{E}\left[\Delta_i^{\varphi,\psi}(\lambda) - \Delta_i^{\varphi,\psi}(\lambda_0)\right]\right\}$$

is stochastically equicontinuous in $(\lambda,\varphi,\psi) \in \Lambda_n \times \Phi \times \Psi$. (See e.g. Theorem 4 of Andrews (1994)). Since $\Lambda_n$ is a shrinking neighborhood of $\lambda_0$ and $\mathbf{E}[\Delta_i^{\varphi,\psi}(\lambda) - \Delta_i^{\varphi,\psi}(\lambda_0)] = 0$, we obtain the wanted result. ■

Let $D_i \in \{0,1\}$ be a binary random variable and define $g_\varphi(u,1) = \mathbf{E}[\varphi(Y_i)|U_i = u, D_i = 1]$ and $g_\psi(u,1) = \mathbf{E}[\psi(Z_i)|U_i = u, D_i = 1]$. Consider the estimator:

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

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

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

with $(\lambda,\varphi,\psi) \in \Lambda_n \times \Phi_n \times \Psi_n$. The following lemma is an extension of Lemma B1. Note that when $D_i = 1$ for all $i$, the result reduces to Lemma B1. The result is in fact a corollary to Lemma B1.

**Lemma B2 :** *Suppose that* Assumptions B1-B3 *hold and that* $sup_{u\in[0,1]}\mathbf{E}[D_i|U_i = u] > 0$. *Then,*

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

*Furthermore, the result remains the same when we replace* $\nu_n(\lambda,\varphi,\psi,1)$ *by* $\nu_n(\lambda_0,\varphi,\psi,1)$.

**Proof :** Write

$$\nu_n(\lambda,\varphi,\psi,1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Z_i)D_i\left\{\frac{\hat{g}^{[1]}_{\varphi,\lambda,i}(U_{n,\lambda,i})}{\hat{g}^{[2]}_i(U_{n,\lambda,i})} - \frac{g^{[1]}_\varphi(U_i)}{g^{[2]}(U_i)}\right\},$$

where $g_{\varphi}^{[1]}(u) = \mathbf{E}[\varphi(Y_i)D_i|U_i = u]$, $g^{[2]}(u) = \mathbf{E}[D_i|U_i = u]$,

$$\begin{aligned}
\hat{g}_{\varphi,\lambda,i}^{[1]}(u) &= \frac{1}{(n-1)\hat{f}_{\lambda,i}(u)} \sum_{j=1,j\neq i}^{n} \varphi(Y_j)D_j K_h\left(U_{n,\lambda,j} - u\right), \\
\hat{g}_{i}^{[2]}(u) &= \frac{1}{(n-1)\hat{f}_{\lambda,i}(u)} \sum_{j=1,j\neq i}^{n} D_j K_h\left(U_{n,\lambda,j} - u\right).
\end{aligned}$$

Using the arguments in the proof of Lemma B1, we can write

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i)D_i \left\{ \frac{\hat{g}_{\varphi,\lambda,i}^{[1]}(U_{n,\lambda,i})}{\hat{g}_{i}^{[2]}(U_{n,\lambda,i})} - \frac{g_{\varphi}^{[1]}(U_i)}{g^{[2]}(U_i)} \right\} \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\psi(Z_i)D_i}{g^{[2]}(U_i)} \left\{ \hat{g}_{\varphi,\lambda,i}^{[1]}(U_{n,\lambda,i}) - g_{\varphi}^{[1]}(U_i) \right\} \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i)D_i \frac{g_{\varphi}^{[1]}(U_i)}{(g^{[2]}(U_i))^2} \left\{ g^{[2]}(U_i) - \hat{g}_{i}^{[2]}(U_{n,\lambda,i}) \right\} + o_P(1).
\end{aligned}$$

By applying Lemma B1 to both terms, we obtain the wanted result. ■

# References


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

[2] Ahn, H. and C. F. Manski, 1993. Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations. Journal of Econometrics 56, 291-321.

[3] Ahn, H. and J. L. Powell, 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58, 3-29.

[4] Andrews, D. W. K, 1994. Empirical process methods in econometrics. In The Handbook of Econometrics, Vol. IV, ed. by R. F. Engle and D. L. McFadden, Amsterdam: North-Holland.

[5] Billingsley, 1999. Convergence of Probability Measures. John Wiley & Sons, New York.

[6] Buchinsky, M. and J. Hahn, 1998. An alternative estimator for the censored quantile regression model. Econometrica 66, 653-671.

[7] Chen, X., H. Hong, and A. Tarozzi, 2008. Semiparametric efficiency in GMM models with auxiliary data set. Annals of Statistics 36, 808-843.

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

[9] Chen, X., O. Linton, and I. van Keilegom, 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71, 1591-1608.

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

[11] Domínguez, M. A. and I. M. Lobato, 2004. Consistent estimation of models defined by conditional moment restrictions. Econometrica 72, 1601-1615.

[12] Escanciano, J-C. and K. Song, 2008. Testing single-index restrictions with a focus on average derivatives. Working paper.

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

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

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

[16] Heckman, J. J., 1990. Varieties of selection bias. American Economic Review 80, 313-328.

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

[18] 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.

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

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

[21] Ledoux, M. and M. Talagrand, 1988. Un critère sur les petite boules dans le théorème limite central. Probability Theory and related Fields 77, 29-47.

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

[23] Liu, R. Y., 1988. Bootstrap procedures under some non i.i.d. models. Annals of Statistics 16, 1696-1708.

[24] Mroz, T., 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica **55**, 765-799.

[25] 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.

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

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

[28] Powell, J., 1989. Semiparametric estimation of bivariate latent variable models. Manuscript, University of Wisconsin Madison.

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

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

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

[32] Song, K., 2008. Uniform convergence of series estimators over function spaces. Econometric Theory 24, 1463-1499.

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

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

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

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

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

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

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

[40] van der Vaart, A. W., 1998. Asymptotic Statistics, Cambridge University Press, New York.

[41] van der Vaart, A. W. and J. A. Wellner, 1996. Weak Convergence and Empirical Processes, Springer-Verlag, New York.

[42] Wu, C. F. J., 1986. Jacknife, bootstrap and other resampling methods in regression analysis. Annals of Statistics 14, 1261-1295.

[43] 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.