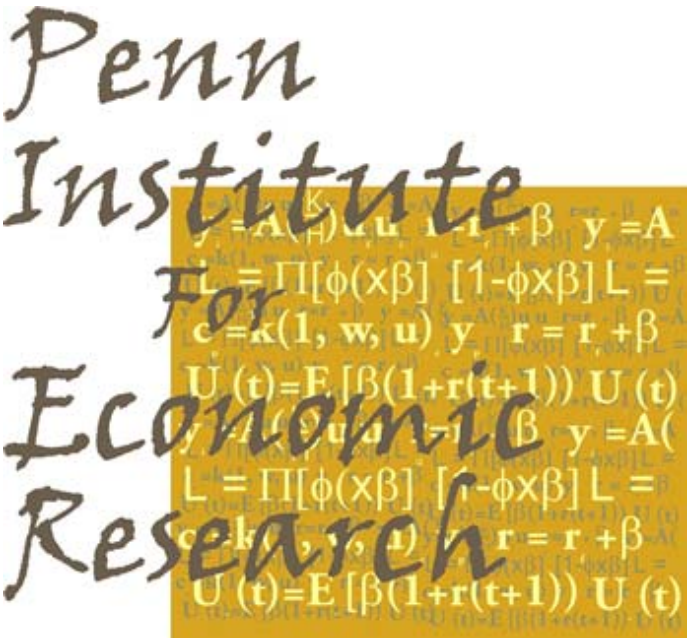


Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 09-012

"Two-Step Extremum Estimation with Estimated Single-Indices"

by

Kyungchul Song

http://ssrn.com/abstract=1360975

Two-Step Extremum Estimation with Estimated Single-Indices

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

February 16, 2009

Abstract

This paper studies two-step extremum estimation that involves the first step estimation of nonparametric functions of single-indices. First, this paper finds that under certain regularity conditions for conditional measures, linear functionals of conditional expectations are insensitive to the first order perturbation of the parameters in the conditioning variable. Applying this result to symmetrized nearest neighborhood estimation of the nonparametric functions, this paper shows that the influence of the estimated single-indices on the estimator of main interest is asymptotically negligible even when the estimated single-indices follow cube root asymptotics. As a practical use of this finding, this paper proposes a bootstrap method for conditional moment restrictions that are asymptotically valid in the presence of cube root-converging single-index estimators. Some results from Monte Carlo simulations are presented and discussed.

Keywords: two-step extremum estimation; single-index restrictions; cube root asymptotics; bootstrap;

JEL Classifications: C12, C14, C51.

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

[1]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. 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.

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. 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), Härdle, Hall and Ichimura (1993), Horowitz and Härdle (1996), and Hristache, Juditsky and Spokoiny (2001).

Most literatures have dealt with a single-index model as an isolated object, whereas researchers often use it as part of a larger model. This paper considers the following estimation framework. Let the parameter of interest $\beta_0 \in \mathbf{R}^d$ be identified as the unique maximizer of a population objective function :

$$\beta_0 = \text{argmax}_\beta Q(\beta, \mu_0(\cdot; \lambda_0)), \tag{1}$$

where $\mu_0(\cdot; \lambda_0) = (\mu_{0,1}(\cdot; \lambda_{0,1}), \cdots, \mu_{0,J}(\cdot; \lambda_{0,J}))^\top$ and

$$\mu_{0,j}(\cdot; \lambda_{0,j}) = \mathbf{E}[Y^{(j)} | \lambda_{0,j}(X) = \lambda_{0,j}(\cdot)]$$

with $Y^{(j)}$ being the j-th component of random vector $Y \in \mathbf{R}^J$ and X being a random vector in $\mathbf{R}^{d_X}$. The real function $\lambda_{0,j} : \mathbf{R}^{d_X} \to \mathbf{R}$ is a single-index of X. The distributions of $\lambda_{0,j}(X)$'s are assumed to be absolutely continuous.

We assume that μ_0 and λ_0 are identified and estimated prior to estimating β_0. The identification is ensured either through a single-index restriction imposed on an identified nonparametric function or through some auxiliary data set in the sense of Chen, Hong, and Tarozzi (2008). Then the estimator of β_0 can be constructed as

$$\hat{\beta} = \text{argmax}_\beta Q_n(\beta, \hat{\mu}(\cdot; \hat{\lambda})), \tag{2}$$

where $Q_n(\beta, \hat{\mu}(\cdot; \hat{\lambda}))$ is the sample objective function and $\hat{\mu}(\cdot; \hat{\lambda})$ is the nonparametric estimator of $\mu_0(\cdot; \lambda_0)$ using $\hat{\lambda}$, an estimator of λ_0. The function $\lambda_{0,j}$ is either a nonparametric function or a parametric function. In the latter case, the estimator $\hat{\lambda}_j$ is allowed to be either $\sqrt{n}$-consistent or $n^{1/3}$-consistent.

The main finding of this paper is that there is no estimation effect of $\hat{\lambda}$ upon the asymptotic variance matrix of $\hat{\beta}$ under certain regularity conditions. (See Theorem 1 below.) Newey (1994) explained how the first step estimators affect the asymptotic variance of the second step estimators. The influence of the first step estimators is represented through a pathwise derivative of the parameter of interest in the nuisance parameters. However, the

nature of the problem here is different in the sense that the nonparametric function $\mu_0(\cdot;\lambda_0)$ depends on λ_0 through the σ-field generated by $\lambda_0(X)$. Therefore, it is not immediately obvious to find the pathwise derivative of the parameter in λ_0. Note also that the usual analysis through an asymptotic linear representation of $\hat{\lambda}$ does not help either when $\hat{\lambda}$ follows cube root asymptotics because such a linear representation does not exist in this case.

First, the paper introduces regularity conditions for conditional measures and show that under these conditions, linear functionals of $\mu_0(\cdot;\lambda)$ have a zero Fréchet derivative in λ (Lemma 2). Using this result, the paper establishes a uniform Bahadur representation of sample linear functionals of the symmetrized nearest neighborhood (SNN) estimator (Lemma A1 in the Appendix). Through the uniform representation, it is shown that there is no estimation effect of $\hat{\lambda}$ upon the asymptotic variance of $\hat{\beta}$.

The asymptotic negligibility of the estimated single-index has broad implications for inference of various semiparametric models. Among other things, the result of this paper illuminates the asymptotic theory of estimators from certain models that have not appeared in the literature. Examples are a sample selection model with conditional median restrictions and models with single-index instrumental variables that are estimable at the rate of $n^{1/3}$. Second, there can be valid bootstrap methods for the inference of β_0 even when $\hat{\lambda}$ follows cube root asymptotics. This is interesting because bootstrap is known to fail for such $n^{1/3}$-converging estimators (Abrevaya and Huang (2005).) This paper proposes a bootstrap method in the special case of conditional moment restrictions.

A similar finding for $\sqrt{n}$-consistent single-index estimators has already appeared in Fan and Li (1996) in the context of testing semiparametric models. See also Stute and Zhu (2005) for a related result in testing single-index restrictions. These literatures deal with a special case where the single-index component is a parametric function with a $\sqrt{n}$-consistent estimator. This paper places in the broad perspective of extremum estimation the phenomenon of asymptotic negligibility of the estimated single-index and allows for the single-index estimator to be a $n^{1/3}$-consistent estimator or a nonparametric estimator. Let us conclude the introduction by discussing some examples.

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

$$\begin{aligned} Y &= \beta_0^\top W_1 + v \text{ and} \\ D &= 1\{\lambda_0(X) \geq \varepsilon\}, \end{aligned}$$

where $\lambda_0(X) = X^\top \theta_0$. The variable Y denotes the latent outcome and W_1 a vector of covariates that affect the outcome. The binary D represents the selection of the vector

(Y, W_1) into the observed data set, so that (Y, W_1) is observed only when $D = 1$. The incidence of selection is governed by a single index $\lambda_0(X)$ of covariates X. The variables v and ε represent unobserved heterogeneity in the individual observation.

The variable ε is permitted to be correlated with X but $Med(\varepsilon|X) = 0$. And W_1 is independent of (v, ε) conditional on the index $\lambda_0(X)$ in the selection mechanism. Therefore, the individual components of X can be correlated with v. The assumptions of the model are certainly weaker than the common requirement that (W_1, X) be independent of (v, ε). (e.g. Heckman (1990), Newey, Powell, and Walker (1990).) More importantly, this model does not assume that X is independent of unobserved component ε in the selection equation. Hence we cannot use the characterization of the selection bias through the propensity score $P\{D = 1|\lambda_0(X)\}$ 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 β_0 still follows if the matrix

$$\mathbf{E}\left[(X - \mathbf{E}[X|D = 1, \lambda_0(X)])(X - \mathbf{E}[X|D = 1, \lambda_0(X)])^\top | D = 1\right]$$

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

$$Y = \beta_0^\top W_1 + \tau(\lambda_0(X)) + u,$$

where u satisfies that $\mathbf{E}[u|D = 1, W_1, \lambda_0(X)] = 0$ and τ is an unknown nonparametric function. This model can be estimated by using the method of Robinson (1988). Let $\mu_Y(\cdot) = \mathbf{E}[Y|D = 1, \lambda_0(X) = \cdot]$, and $\mu_{W_1}(\cdot) = \mathbf{E}[W_1|D = 1, \lambda_0(X) = \cdot]$. Then, we consider a conditional moment restriction:

$$\mathbf{E}\left[\{Y - \mu_Y(\lambda_0(X))\} - \beta_0^\top \{W_1 - \mu_{W_1}(\lambda_0(X))\} | D = 1, W_1, \lambda_0(X)\right] = 0.$$

One may estimate θ_0 in λ_0 using maximum score estimation in the first step and use it in the second step estimation of β_0. Then the remaining question centers on the effect of the first step estimator of θ_0 which follows cube root asymptotics upon the estimator of β_0.

Note that the identification of θ_0 does not stem from a direct imposition of single-index restrictions on $\mathbf{E}[Y|D = 1, X = \cdot]$ and $\mathbf{E}[Z|D = 1, X = \cdot]$. The identification follows from the use of auxiliary data set $((D = 0), X)$ 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 Walk (1990), and Ahn and Powell (1993). ■

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

$$\begin{aligned} Y &= Z^\top \beta_0 + \varepsilon, \text{ and} \\ D &= 1\{\lambda_0(X) \geq \eta\}, \end{aligned}$$

where $\lambda_0(X) = X^\top \theta_0$ and ε and η satisfy that $\mathbf{E}[\varepsilon|\lambda_0(X)] = 0$ and $Med(\eta|X) = 0$. Therefore, the index $\lambda_0(X)$ 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 $X^\top \theta_0$ not the whole vector X. In other words, some of the elements of the vector X are allowed to be correlated with ε. Furthermore, X is not required to be independent of η as long as it maintains the conditional median restriction. This conditional median restriction enables one to identify θ_0 and in consequence β_0. Hence the data set (D, X) plays the role of an auxiliary data set in Chen, Hong, and Tarozzi (2008).

While there are many ways to estimate β_0, we consider the following conditional moment restriction:

$$\mathbf{E}\left[Y - \mathbf{E}[Z|\lambda_0(X)]^\top \beta_0 | \lambda_0(X)\right] = 0.$$

We can first estimate λ_0 and $\mathbf{E}[Z|\lambda_0(X)]$ and then estimate β_0 by plugging in these estimates into a sample version of the conditional moment restriction. ■

Example 3 (Models with Single-Index Restrictions) : There are numerous semiparametric models that contain nonparametric estimation of a function $\mathbf{E}[Y|X]$ in the first step. (e.g. Ahn and Manski (1993), Buchinsky and Hahn (1998), Hirano, Imbens, and Ridder (2003).) The finding of this paper enables one to employ the same asymptotic analysis in the literature when one imposes a single index restriction:

$$\mathbf{E}[Y|X] = m(X^\top \gamma_0)$$

for some unknown function m and parameter γ_0. We can estimate γ_0 using the methods of inference for single-index models and plug the estimator $\hat{\gamma}_0$ in the nonparametric estimation of m. The coefficient estimator $\hat{\gamma}$ is typically $\sqrt{n}$-consistent. Then the asymptotic analysis can be done as if we know the true index parameter γ_0, because the estimation error in $\hat{\gamma}_0$ does not affect the asymptotic variance of the parameter of interest. ■

Some models where an unknown nonparametric function $\lambda_0(\cdot)$ constitutes the conditioning variable of a conditional expectation have received attention in the literature.

Example 4 (Matching Estimators of Treatment Effects on the Treated) : Let Y_1

and Y_0 be potential outcomes of a treated and an untreated individuals and D the treatment status. The parameter of interest is $\mu_1 = \mathbf{E}[Y_1 - Y_0|D = 1]$, i.e., the treatment effect on the treated. Let $\lambda_0(X) = P\{D = 1|X\}$, where X is a vector of covariates. Under the condition:

$$\mathbf{E}[Y_0|\lambda_0(X), D = 0] = \mathbf{E}[Y_0|\lambda_0(X), D = 1], \tag{3}$$

we can identify (Heckman, Ichimura, and Todd (1997))

$$\mu_1 = \mathbf{E}\left[Y_1 - \mathbf{E}[Y_0|D = 0, \lambda_0(X)]|D = 1\right].$$

Therefore, the parameter of interest μ_1 involves a nonparametric function λ_0 in the conditioning variable. Then, following Heckman, Ichimura and Todd (1998), we can estimate μ_1 by

$$\hat{\mu}_1 = \frac{1}{\sum_{i=1}^n 1\{D_i = 1\}} \sum_{i=1}^n 1\{D_i = 1\} \left\{Y_{1i} - \hat{\mathbf{E}}[Y_{0i}|\hat{\lambda}(X_i), D = 0]\right\}, \tag{4}$$

where $\hat{\mathbf{E}}[Y_{0i}|\lambda(X_i), D = 0]$ is a nonparametric estimator of $\mathbf{E}[Y_{0i}|\lambda_0(X_i), D = 0]$ and $\hat{\lambda}(X)$ that of $\lambda_0(X)$. Therefore, it is important for the asymptotic variance of $\hat{\mu}_1$ to analyze the effect of estimation $\hat{\lambda}$. ■

The remainder of the paper has three sections. The first section exposits the main result of this paper and provides heuristics. The second section focuses on the case with conditional moment restrictions and proposes a valid bootstrap procedure in the presence of $n^{1/3}$-converging nuisance parameter estimators. The third section presents and discusses simulation results and the last section concludes. The appendix contains technical proofs of the main results and a general uniform Bahadur representation of sample linear functionals of SNN estimators.

2 The Main Results

2.1 A Motivating Example

To illustrate the main motivation of this paper, we present some simulation results from the following semiparametric model:

$$Y_i = Z_i\beta_0 + \gamma_0 f(X_i^\top \theta_0) + \varepsilon_i,$$

where $f(v)$ is unknown, $\mathbf{E}[\varepsilon_i|X_i^\top\theta_0, Z_i] = 0$ and θ_0 is identified and estimated using some other data sources. We first generated the following fictitious first step "estimator" with

varied noise levels:

$$\tilde{\theta}_k = \theta_{0,k} + a \times N(0,1),\ k = 1, 2,$$

with $a \in \{0.2, 0.4, 0.6, 1, 2, 3, 4\}$. We normalized the scale and defined $\hat{\theta} = \tilde{\theta}/||\tilde{\theta}||$ as the first step "estimator" of θ_0. Using Robinson's procedure, we can write the model as a semiparametric conditional moment restriction. Then, in the second step, we estimated β_0 from this restriction. (Details are found in Section 3.)

The data generating process used is as follows. We drew $\varepsilon_i,\ v_i,\ w_i,\ \varepsilon_{1,i}$ and $\varepsilon_{2,i}$ independently from $N(0,1)$ and defined

$$Z_i = v_i + w_i,\ \text{and } X_{k,i} = v_i + \varepsilon_{k,i},\ k = 1, 2.$$

We set $\theta_0 = [-0.5,\ 1]^\top$, $\gamma_0 = 0$, and $\beta_0 = 2$. The sample size was $n = 300$ and the Monte Carlo simulation number was 1000.

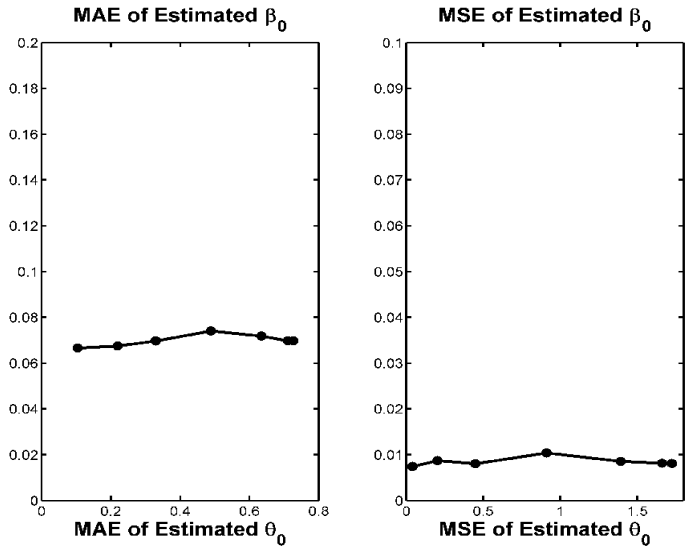
The results are shown in Figure 1 which plots the mean absolute error (MAE) and the mean squared error (MSE) of $\hat{\beta}$ against those of $\hat{\theta}$. The different points in the line represent results corresponding to the different choices of the noise level a. The results show that the quality of $\hat{\beta}$ is robust to that of $\hat{\theta}$, both in terms of MAE and MSE. The robustness of MSE of $\hat{\beta}$ against that of $\hat{\theta}$ is remarkable. This paper analyzes this phenomenon and reveals that it has a generic nature in a much broader context of extremum estimation. In particular, this robustness enables us to bootstrap $\hat{\beta}$ validly even when $\hat{\theta}$ follows cube root asymptotics in models of conditional moment restrictions.

2.2 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 this paper. 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 Λ be a class of $\mathbf{R}$-valued functions on $\mathbf{R}^{d_X}$ with a generic element denoted by λ.

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 σ-finite measure, say, $w_\lambda(\cdot|\bar{\lambda}_1, \bar{\lambda}_2)$. Note that we do not assume that Y is absolutely continuous as we do not require that $w_\lambda(\cdot|\bar{\lambda}_1, \bar{\lambda}_2)$ is a 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: $||f||_\infty = \sup_{x \in \mathcal{S}_X} |f(x)|$.

Figure 1: The Robustness of the Second Step Estimator



Definition 1 : (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 $(\bar{\lambda}_1, \bar{\lambda}_2) \in \mathcal{S}_\lambda$,

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

where $C_\lambda(\cdot|\bar{\lambda}_1,\bar{\lambda}_2) : \mathcal{S}_Y \to \mathbf{R}$ is such that for some $C > 0$,

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

(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 $f_\lambda(y|\cdot,\cdot)$, $(y,\lambda) \in \mathcal{S}_Y \times \Lambda$. Note that the condition does not require that the conditional density function be continuous in $\lambda \in \Lambda$, which is cumbersome to check in many situations. 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 important consequence. Define

$$\mu_\varphi(x;\lambda) = \mathbf{E}[\varphi(Y)|\lambda(X) = \lambda(x)],$$

where $\varphi \in \Phi$ with Φ being a class of $\mathbf{R}^J$-valued functions on $\mathbf{R}^{d_Y}$.

Lemma 1 : *Suppose that $\mathcal{P}_Y$ is regular for $\tilde{\varphi}$ an envelope of Φ. Then, for each $\lambda \in \Lambda$ 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 $\mu_\varphi(x;\lambda_0,\lambda) = \mathbf{E}[\varphi(Y)|(\lambda_0(X),\lambda(X)) = (\lambda_0(x),\lambda(x))]$ and C does not depend on λ, λ_0, x, or φ.

Lemma 1 shows that the conditional expectations are continuous in the parameter λ 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 $\mathcal{S}_Z$ and a class Ψ being a class of $\mathbf{R}^J$-valued functions on $\mathbf{R}^{d_Z}$ with a generic element denoted by ψ 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 σ-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 unexpected result.

Lemma 2 : *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 λ in Λ,*

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

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

Lemma 2 says that the functional $\Gamma_{\varphi,\psi}(\lambda)$ is not sensitive to the first order perturbation of λ around λ_0. In view of Newey (1994), Lemma 2 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 $\mu_\varphi(X_i;\lambda)$. We explore its implication in the broader context of extremum estimation.

2.3 The Main Result

In this subsection, we formalize the main results. Let us introduce high-level conditions for extremum estimation.

Condition A1 : There is an $\mathbf{R}^d$-valued random function $\xi_n(\mu)$ such that

$$q_n(t_n, \hat{\mu}(\cdot,\hat{\lambda})) - q(t_n, \mu_0(\cdot,\lambda_0)) - \xi_n(\hat{\mu}(\cdot,\hat{\lambda}))^\top t_n = o_P(||t_n||^2), \text{ for any } t_n \to 0,$$

where $q_n(t,\mu) = Q_n(\beta_0 + t, \mu) - Q_n(\beta_0, \mu)$ and $q(t) = Q(\beta_0 + t, \mu_0) - Q(\beta_0, \mu_0)$.

Condition A2 : For a nonsingular Ω, $q(t_n) = t_n'\Omega t_n + o(||t_n||^2)$, for any $t_n \to 0$.

Condition A1 is known as a stochastic differentiability condition (Pollard (1985)). This condition can be proved using stochastic equicontinuity arguments or the convexity lemma as in Pollard (1991). While the presence of $\hat{\mu}(\cdot,\hat{\lambda})$ may complicate the analysis, the procedure is standard. (Newey and McFadden (1994)). Under Conditions A1-A2, one can write (See e.g. the proof of Theorem 3.2.16 of van der Vaart and Wellner (1996))

$$\sqrt{n}(\hat{\beta} - \beta_0) = \Omega^{-1}\sqrt{n}\xi_n(\hat{\mu}(\cdot,\hat{\lambda})) + o_P(1).$$

To analyze the role of the estimation error in $\hat{\mu}(\cdot,\hat{\lambda})$ for the asymptotic distribution of $\hat{\beta}$, we need to investigate the right-hand side term. For this, we introduce the following assumptions.

Condition B1 : There exist a sequence of $d \times J$ random matrices $\{Z_i\}_{i=1}^n$ such that $\{Z_i\}_{i=1}^n$ is i.i.d. and

$$\sqrt{n}\{\xi_n(\hat{\mu}(\cdot,\hat{\lambda})) - \xi_n(\mu_0(\cdot,\lambda_0))\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\left\{\hat{\mu}(X_i;\hat{\lambda}) - \mu_0(X_i;\lambda_0)\right\} + o_P(1). \qquad (5)$$

Condition B2 : $\xi_n(\mu_0(\cdot,\lambda_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_2(S_i) + o_P(1)$, for some $\mathbf{R}^J$-valued function η_2 such that $\mathbf{E}\eta_2(S_i) = 0$ and $\mathbf{E}||\eta_2(S_i)||^2 < \infty$, where $\{S_i\}_{i=1}^n$ are i.i.d. random vectors.

Condition B1 can be checked through the usual linearization of the sample objective function. When $Q_n(\beta,\mu)$ is not differentiable in μ (in the sense of the usual pointwise differentiation), we can decompose the problem into that of linearization of $Q(\beta,\mu)$ in μ and

the oscillation property of $Q_n(\beta,\mu) - Q(\beta,\mu)$ in μ to obtain the above result. Condition B2 says that $\xi_n(\mu_0(\cdot,\lambda_0))$ is approximated as a normalized i.i.d. sum of mean zero random vectors.

The effect of $\hat{\mu}(\cdot,\hat{\lambda})$ on the asymptotic variance of $\hat{\beta}$ is revealed through the analysis of the right-hand side of (5). For the sake of specificity, we consider symmetrized neighborhood estimation of $\hat{\mu}$. Let $\hat{U}_k^{(j)} = \frac{1}{n}\sum_{i=1}^n 1\{\hat{\lambda}_j(X_i) \leq \hat{\lambda}_j(X_k)\}$ and $\hat{\mu}(X_k;\hat{\lambda}) = [\hat{\mu}_1(X_k;\hat{\lambda}_1),\cdots,\hat{\mu}_J(X_k;\hat{\lambda}_J)]^\top$, where

$$\hat{\mu}_j(X_k;\hat{\lambda}_j) = \frac{\sum_{i=1}^n Y_i^{(j)} K_h\left(\hat{U}_i^{(j)} - \hat{U}_k^{(j)}\right)}{\sum_{i=1}^n K_h\left(\hat{U}_i^{(j)} - \hat{U}_k^{(j)}\right)}, \tag{6}$$

and $Y_i^{(j)}$ is the j-th component of Y_i and $K_h(u) = K(u/h)/h$ and $K : \mathbf{R} \to \mathbf{R}$ is a kernel function. The estimator $\hat{\mu}_j$ is a symmetrized nearest neighborhood (SNN) estimator proposed by Yang (1981) and studied by Stute (1984). The probability integral transform of $\lambda_{0,j}(X)$ turns its density into a uniform density on $[0,1]$. 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.

Under regularity conditions, we can apply the uniform Bahadur representation theorem established in the appendix (Lemma A1) to show that

$$\sqrt{n}\xi_n(\hat{\mu}(\cdot;\hat{\lambda})) = \sqrt{n}\xi_n(\mu_0(\cdot;\lambda_0)) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_1(X_i) + o_P(1), \tag{7}$$

where $\eta_1(X_i) = [\eta_{1,1}(X_i),\cdots,\eta_{1,d}(X_i)]^\top$, $\eta_{1,k}(X_i) = \sum_{j=1}^J \mathbf{E}[Z_i^{(k,j)}|\lambda_{0,j}(X_i)](Y_i^{(j)} - \mu_{0,j}(X_i))$, and $Z_i^{(k,j)}$ is the (k,j)-th entry of Z_i. The second term involving $\eta_1(X_i)$ is due to the non-parametric estimation error in $\hat{\mu}$. However, the Bahadur representation remains the same regardless of whether we use λ_0 or $\hat{\lambda}$ in constructing $\hat{\mu}$. Using this result, we can prove the following (See Theorem 1 below.)

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, \Omega^{-1}\Sigma\Omega^{-1}), \tag{8}$$

where $\Sigma = \mathbf{E}\left[(\eta_1(X_i) + \eta_2(X_i))\,(\eta_1(X_i) + \eta_2(X_i))^\top\right]$. Hence the asymptotic covariance matrix remains the same with or without the estimation of λ_0.

We can place this phenomenon in the perspective of Lemma 2. By Condition B1, the

effect of $\hat{\lambda}$ upon $\hat{\beta}$ is revealed through the analysis of the following:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \left\{ \mu(X_i; \lambda) - \mu(X_i; \lambda_0) \right\}$$

with λ lying within a shrinking neighborhood Λ_n of λ_0. After subtracting its mean, the above sum becomes asymptotically negligible through stochastic equicontinuity in $\lambda \in \Lambda_n$, leaving

$$\sqrt{n}\mathbf{E}\left[Z_i \left\{ \mu(X_i; \lambda) - \mu(X_i; \lambda_0) \right\} \right].$$

By Lemma 2, the expectation above is $O(||\lambda - \lambda_0||_\infty^2)$, yielding that whenever $||\hat{\lambda} - \lambda_0||_\infty = o_P(n^{-1/4})$, the first order effect of $\hat{\lambda}$ disappears.

To formalize the result, let us introduce some notations and assumptions. Let Λ_j be a class of functions $\lambda_j : \mathbf{R}^{d_X} \to \mathbf{R}$ such that $P\{\hat{\lambda}_j \in \Lambda_j\} \to 1$ as $n \to \infty$, and $\Lambda_j(\delta) = \{\lambda_j \in \Lambda_j : ||F_{\lambda,j} \circ \lambda_j - F_{0,j} \circ \lambda_{0,j}||_\infty < \delta\}$, where $F_{0,j}$ and $F_{\lambda,j}$ are the cdfs of $\lambda_{0,j}(X)$ and $\lambda_j(X)$. For a class $\mathcal{F}$ of functions, let $N(\varepsilon, \mathcal{F}, ||\cdot||_\infty)$ be the covering number of $\mathcal{F}$ with respect to $||\cdot||_\infty$. (See van der Vaart and Wellner (1996) for details.) Denote $f_\lambda^{(j)}(y|u_0, u_1)$ to be the conditional density of $Y^{(j)}$ given $(U_0^{(j)}, U_\lambda^{(j)}) = (u_0, u_1)$ with respect to a σ-finite measure, where $U_0^{(j)} = F_{0,j}(\lambda_{0,j}(X))$ and $U_\lambda^{(j)} = F_{\lambda,j}(\lambda_j(X))$. Similarly, define $h_\lambda^{(k,j)}(z|u_0, u_1)$ to be the conditional density of $Z^{(k,j)}$ given $(U_0^{(j)}, U_\lambda^{(j)}) = (u_0, u_1)$ with respect to a σ-finite measure. Let $\mathcal{S}_Y^{(j)}$ be the support of $Y^{(j)}$ and $\mathcal{S}_Z^{(k,j)}$ be the support of $Z^{(k,j)}$, and define

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

Then, we introduce the following assumptions.

Assumption G1 : (i) For each $j = 1, \cdots, J$, (a) $||\hat{\lambda}_j - \lambda_{0,j}||_\infty = O_P(n^{-b})$, $b \in (-1/4, 1/2]$, and (b) for some $C_j > 0$,

$$|F_{0,j}(\bar{\lambda}_1) - F_{0,j}(\bar{\lambda}_2)| \leq C_j |\bar{\lambda}_1 - \bar{\lambda}_2|, \text{ for all } \bar{\lambda}_1, \bar{\lambda}_2 \in \mathbf{R}.$$

(ii) $\mathbf{E}[||Y_i||^p] < \infty$ and $\mathbf{E}[||Z_i||^p] < \infty$ for $p > 8$.

Assumption G2 : For $j = 1, \cdots, J$, there exists $\delta_j > 0$ such that
(i) for $b_j \in [0, 1)$ and $C_j > 0$, $\log N(\varepsilon, \Lambda_j^F, ||\cdot||_\infty) < C_j \varepsilon^{-b_j}$, where $\Lambda_j^F = \{F_{\lambda,j} \circ \lambda : \lambda \in \Lambda_j(\delta_j)\}$,
(ii) $\mathcal{P}_Y^{(j)}(\delta_j)$ and $\mathcal{P}_Y^{(k,j)}(\delta_j)$, $k = 1, \cdots, d$, are regular (in the sense of Definition 1), and
(iii) $\sup_{u \in [0,1]} \mathbf{E}[|Y^{(j)}||U_0^{(j)} = u] < \infty$, and $\mathbf{E}[Y^{(j)}|U_0^{(j)} = \cdot]$ is twice continuously differentiable with bounded derivatives.

Assumption G3 : (i) $K(\cdot)$ is symmetric, compact supported, twice continuously differentiable with bounded derivatives, $\int K(t)dt = 1$.
(ii) $n^{1/2}h^3 + n^{-1/2}h^{-2}(-\log h) \to 0$.

These assumptions are introduced to ensure the asymptotic representation in (7). Assumption G1(i) allows $\hat{\lambda}$ to converge at the rate of $n^{-1/3}$. Assumption G2 is a regularity condition for the index functions λ_j. Assumption G3(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 G3(ii) does not require undersmoothing; it is satisfied for any $h = n^{-s}$ with $1/6 < s < 1/4$.

Theorem 1 : *Suppose that* Conditions A1-A2 *and* B1-B2 *hold. Furthermore, suppose that* Assumptions G1-G3 *hold. Then, the asymptotic normality in* (8) *follows. Moreover, the asymptotic covariance matrix in* (8) *does not change when we replace* $\hat{\lambda}$ *by* λ_0.

In view of Newey (1994), the result of Lemma 2 suggests that the asymptotic negligibility of $\hat{\lambda}$ will not depend on the particular estimation method employed. Indeed, an analogous result in testing single-index restrictions was obtained by Escanciano and Song (2008) using series estimation.

Theorem 1 has an important implication for matching estimators based on a propensity score. Consider the set-up of Example 3 and the matching estimator

$$\hat{\mu}_1 = \frac{1}{\sum_{i=1}^n 1\{D_i = 1\}} \sum_{i=1}^n 1\{D_i = 1\} \left\{ Y_{1i} - \hat{\mu}(X_i; \hat{\lambda}) \right\},$$

where $\hat{\mu}(X_i; \hat{\lambda}) = \sum_{i=1}^n Y_{0i} K_h(\hat{U}_i - \hat{U}_k) / \sum_{i=1}^n K_h(\hat{U}_i - \hat{U}_k)$, $\hat{U}_k = \frac{1}{n}\sum_{i=1}^n 1\{\hat{\lambda}(X_i) \leq \hat{\lambda}(X_k)\}$, and $\hat{\lambda}(X)$ is a nonparametric estimator of the propensity score $\lambda_0(X) = P\{D = 1|X\}$. Then Theorem 1 tells us that the asymptotic variance of $\hat{\mu}_1$ remains the same if we replace $\hat{\lambda}$ by λ_0.

Another important implication is that there can exist a valid bootstrap method for estimating β_0 even when $\lambda(X_i) = \lambda(X_i; \theta_0)$, a parametric function, and a $\sqrt[3]{n}$-consistent estimator $\hat{\theta}$ of θ_0 is used in the first step estimation. We suggest one bootstrap method for models of conditional moment restrictions in the next section.

3 Bootstrap in Models of Conditional Moment Restrictions

In this section, we focus on conditional moment restrictions as a special case. For $j = 1, \cdots, J+1$, let $\lambda_{0,j}(x) = \lambda_j(x; \theta_0)$, known up to $\theta_0 \in \mathbf{R}^{d_\Theta}$. Let β_0 be identified through the

following restriction:

$$\mathbf{E}\left[\rho(V, \mu_0(X; \lambda_0); \beta_0)|W\right] = 0,$$

where $W = (W_1, \lambda_{0,J+1}(X))$, $(V, W_1, X) \in \mathbf{R}^{d_V + d_{W_1} + d_X}$ is an observable random vector and $\rho(v, \mu; \beta_0) : \mathbf{R}^{d_V + J} \to \mathbf{R}$ is known up to $\beta_0 \in B \subset \mathbf{R}^{d_\beta}$. The function $\mu_0(X; \lambda_0)$ is as defined in the introduction (below (1)). Note that W is allowed to depend on an unknown continuous single index $\lambda_{0,J+1}(X)$. This feature is relevant when the IV exogeneity takes the form of *single-index exogeneity*, where the instrumental variable takes a form of a single-index. Examples 1 and 2 in a preceding section belong to this framework.

Given the estimator $\hat{\theta}$, we let $\hat{\lambda}_j(\cdot) = \lambda_j(\cdot; \hat{\theta})$ and assume that $\mu_0(X; \lambda_0)$ is estimated by $\hat{\mu}(X; \hat{\lambda})$ in the first step as in (6). Then we estimate β_0 as follows:

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

where $\hat{W}_k = (W_{1k}, \hat{U}_k^{(J+1)})$ and $\hat{U}_k^{(J+1)}$ is as defined prior to (6) using $\{\hat{\lambda}_{J+1}(X_i)\}_{i=1}^n$. The estimation method is similar to the proposal by Domínguez and Lobato (2004). Let $\Theta(\delta) \equiv \{\theta \in \mathbf{R}^{d_\Theta} : ||\theta - \theta_0|| < \delta\}$.

Assumption 1 : (i) The sample $\{(V_i, X_i, Y_i, W_{1i})\}_{i=1}^n$ is a random sample.
(ii)(a) $\mathbf{E}[\rho(V, \mu_0(X; \lambda_0); \beta)|W] = 0$ a.s. iff $\beta = \beta_0$. (b) $\beta_0 \in \text{int}(B)$ with B compact.
(iii) $\rho(v, \mu; \beta)$ as a function of $(\mu, \beta) \in \mathbf{R}^J \times B$ is twice continuously differentiable with the first order derivatives ρ_β and ρ_μ and the second order derivatives $\rho_{\beta\beta}$, $\rho_{\beta\mu}$ and $\rho_{\mu\mu}$ such that $\mathbf{E}[\sup_{\beta \in B} ||\tilde{\rho}(V, \mu_0(X; \lambda_0); \beta)||^p] < \infty$, $p > 2$, for all $\tilde{\rho} \in \{\rho, \rho_\beta, \rho_\mu, \rho_{\beta\beta}, \rho_{\beta\mu}\}$.
(iv) For some $M > 0$ and $p > 8$, $\mathbf{E}[||Y_i||^p] < M$, $\mathbf{E}[||S_i||^p] < M$, and

$$\mathbf{E}[\sup_{(\beta, \bar{\mu}) \in B \times [-M, M]} ||\rho_{\mu\mu}(V_i, \bar{\mu}; \beta)||^q] < \infty, \ q > 4, \tag{9}$$

where $S_i = \rho_\mu(V_i, \mu_0(X_i; \lambda_0); \beta_0)$.

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

Assumptions 1(i)-(iii) are standard in models of conditional moment restrictions. The condition $\mathbf{E}[||S_i||^p] < M$ and (9) in Assumption 1(iv) are trivially satisfied when $\rho(v, \mu; \beta)$ is linear in μ as in Examples 1 and 2. Assumption 2 allows $\hat{\theta}$ to converge at the rate of $n^{-1/3}$.

Let $S_i^{(j)}$ be the j-th entry of S_i defined in Assumption 1(iv) and let $Z_i^{(j)} = (S_i^{(j)}, W_{1i}, U_{0,i}^{(J+1)})$ if $U_{0,i}^{(J+1)} \neq U_{0,i}^{(j)}$ and $Z_i^{(j)} = (S_i^{(j)}, W_{1i})$ if $U_{0,i}^{(J+1)} = U_{0,i}^{(j)}$. We set $\tilde{\psi}(Z_i^{(j)}) = |S_i^{(j)}|$. Define $f_\theta^{(j)}(y|u_0, u_1)$ to be the conditional density of $Y_i^{(j)}$ given $(U_{0,i}^{(j)}, U_{\theta,i}^{(j)}) = (u_0, u_1)$ with respect to a σ-finite measure, where $U_{0,i}^{(j)} = F_{0,j}(\lambda_{0,j}(X_i))$ and $U_{\theta,i}^{(j)} = F_{\theta,j}(\lambda_j(X_i; \theta))$ and $F_{0,j}$ and

$F_{\theta,j}$ are the cdfs of $\lambda_{0,j}(X)$ and $\lambda_j(X;\theta)$. Similarly define $h_\theta^{(j)}(z|u_0,u_1)$ to be the conditional density of $Z_i^{(j)}$ given $(U_{0,i}^{(j)}, U_{\theta,i}^{(j)}) = (u_0,u_1)$ with respect to a σ-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 3 : For each $j = 1,\cdots,J+1$, there exist $\delta_j > 0$ and $C_j > 0$ such that
(i) for each $j = 1,\cdots,J+1$,

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

(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) $\sup_{u\in[0,1]}\mathbf{E}[|Y_i^{(j)}||U_{0,i}^{(j)} = u] < \infty$, and (b) $\mathbf{E}[Y_i^{(j)}|U_{0,i}^{(j)} = \cdot]$ is twice continuously differentiable with bounded derivatives.

Assumption 3(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 1. The regularity of $\mathcal{P}_{Z,j}(\delta_j)$ in Assumption 3(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, $J = 2$, $U_{0,i}^{(1)} = U_{0,i}^{(2)} = U_{0,i}^{(3)}$, and in the case of the model with the single-index instrument in Example 2, $J = 1$, $U_{0,i}^{(1)} = U_{0,i}^{(2)}$. In both cases, S_i is a constant vector of -1's. Hence it suffices for the regularity of $\mathcal{P}_{Z,j}(\delta_j)$ that the conditional density function of W_{1i} given $(U_{0,i}^{(1)}, U_{\theta,i}^{(1)}) = (u_0,u_1)$ is continuously differentiable in (u_0,u_1) with a derivative uniformly bounded over $\theta \in \Theta(\delta_j)$ and W_{1i} has a bounded support. The requirement that W_{1i} have a bounded support can always be made to be fulfilled by using a strictly increasing, continuous and bounded map $G : \mathbf{R}^{d_{W_1}} \to [0,1]^{d_{W_1}}$ and substituting $W_{1i}^G = G(W_{1i})$ for W_{1i}.

Theorem 2 : *Suppose that* Assumptions 1-3 *hold. Furthermore, K satisfies* Assumption G3(i) *and h satisfies that $n^{1/2}h^{3-1/q} + n^{-1/2}h^{-2}(-\log h) \to 0$. Then,*

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d \left(\int \dot{H}\dot{H}^\top dP_W\right)^{-1} \int \dot{H}\zeta dP_W$$

where $\dot{H}(w) = \mathbf{E}[\rho_\beta(V_i;\mu_0(X_i;\lambda_0);\beta_0)1\{W_i \le w\}]$, P_W is the distribution of W, and ζ is a centered Gaussian process on $\mathbf{R}^{d_W}$ that has a covariance kernel given by $C(w_1,w_2) =$

$\mathbf{E}\left[\xi_i(w_1)\xi_i(w_2)\right]$ *with*

$$\begin{aligned}\xi_i(w) &= \rho(V_i;\mu_0(X_i;\lambda_0);\beta_0)1\{W_i \le w\} \\ &\quad -\Sigma_{j=1}^{J}\mathbf{E}[S_i^{(j)}1\{W_i \le w\}|U_i^{(j)}](Y_i^{(j)} - \mu_{0,j}(X_i;\lambda_{0,j})).\end{aligned} \tag{10}$$

The bandwidth condition is slightly stronger than Assumption G3(ii). This condition is used to ensure Condition B1 as well as (7) in this context. Still the bandwidth condition does not require undersmoothing. Compared with the asymptotic covariance matrix of Domínguez and Lobato (2004), the asymptotic covariance matrix contains additional terms involving $Y_i^{(j)} - \mu_{0,j}(X_i;\lambda_{0,j})$ in (10). 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 $\lambda_j(X_i;\hat{\theta})$ or the true indices $\lambda_j(X_i;\theta_0)$. This is true even if $\hat{\theta}$ follows cube root asymptotics.

While one can construct confidence sets for β_0 based on the asymptotic theory, the estimation of the asymptotic covariance matrix appears complicated requiring a choice of multiple bandwidths. Alternatively, one might consider bootstrap. Theorem 2 suggests that there may be a bootstrap method that is valid even when $\hat{\theta}$ follows cube root asymptotics. As far as the author is concerned, it is not clear how one can analyze the asymptotic refinement properties of a bootstrap method in this situation. Leaving this to a future research, this paper chooses to develop 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).)

Suppose that $\hat{\mu}(X_i;\hat{\lambda})$ is a first step estimator defined in (6) and let

$$\begin{aligned}\hat{\rho}_{lk}(\beta) &= 1\{\hat{W}_l \le \hat{W}_k\}\rho(V_l,\hat{\mu}(X_l;\hat{\lambda});\beta) \text{ and} \\ \hat{\rho}_{\mu,ik} &= 1\{\hat{W}_i \le \hat{W}_k\}\rho_\mu(V_i,\hat{\mu}(X_i;\hat{\lambda});\hat{\beta}).\end{aligned}$$

Then, let $\hat{r}_{lk} = [\hat{r}_{lk}^{(1)},\cdots,\hat{r}_{lk}^{(J)}]^\top$ where

$$\hat{r}_{lk}^{(j)} = \frac{\sum_{i=1}^n \hat{\rho}_{\mu,ik}^{(j)} K_h\left(\hat{U}_{n,i}^{(j)} - \hat{U}_{n,l}^{(j)}\right)}{\sum_{i=1}^n K_h\left(\hat{U}_{n,i}^{(j)} - \hat{U}_{n,l}^{(j)}\right)},$$

and $\hat{\rho}_{\mu,ik}^{(j)}$ is the j-th component of $\hat{\rho}_{\mu,ik}$. 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

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

and use its empirical distribution to construct the confidence set for β_0.

The bootstrap procedure is very simple. In particular, one does not need to estimate μ_0 or θ_0 using the bootstrap sample. The estimator $\hat{\mu}(X_i; \hat{\lambda})$ is stored once and repeatedly used for each bootstrap sample. This computational merit is prominent when the dimension of the parameter θ_0 is large and one has to resort to a numerical optimization algorithm for its estimation as in the case of maximum score estimation.

Theorem 3 : *Suppose that* Assumptions 1-3 *hold. Then,*

$$\sqrt{n}(\hat{\beta}_b^* - \hat{\beta}) \to_d \left(\int \dot{H} \dot{H}^{\top} dP_W \right)^{-1} \int \dot{H} \zeta dP_W, \text{ conditional on } \{(V_i, X_i, Y_i, W_{1i})\}_{l=1}^n, \text{ in } P$$

where $\dot{H}$ *and* ζ *are as in* Theorem 2.

Theorem 3 shows that the bootstrap procedure is asymptotically valid. Therefore, even when $\hat{\theta}$ follows cube root asymptotics, we can still bootstrap $\hat{\beta}$ in this situation.

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. $U[0,1]$ random variable, U_{2i} and η_i are random vectors in $\mathbf{R}^k$ with entries equal to i.i.d random variables of $U[0,1]$. The dimension k is chosen from $\{3, 6\}$. The random variable η_{1i} is the first component of η_i. Then, the selection mechanism is defined as

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

where ε_i follows the distribution of $2T_i \times \frac{1}{d_X} \sum_{k=1}^{d_X} \Phi\left(X_{ik}^2 + |X_{ik}|\right) + \zeta_i$, $\zeta_i \sim N(0,1)$, Φ 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 $v_i \sim (\zeta_i + e_i) \times \Phi\left(Z_i^2 + |Z_i|\right)$ with $e_i \sim N(0,1)$. We set θ_0 to be the vector of 2's and $\beta_0 = 2$.

We first estimate θ_0 by using the maximum score estimation to obtain $\hat{\theta}$. Using this $\hat{\theta}$, we construct $\hat{U}_{n,i}$ and

$$\begin{aligned}
\hat{\mu}_{Y,j} &= \frac{\sum_{i=1,i\neq j}^{n} Y_i \times K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}{\sum_{i=1,i\neq j}^{n} K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)} \text{ and} \\
\hat{\mu}_{Z,j} &= \frac{\sum_{i=1,i\neq j}^{n} Z_i \times K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}{\sum_{i=1,i\neq j}^{n} K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}.
\end{aligned}$$

Then, we estimate β from the following optimization:

$$\hat{\beta} = \underset{\beta\in B}{\operatorname{argmin}} \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{ij}\left(Y_i - \hat{\mu}_{Y,i} - \{Z_i - \hat{\mu}_{Z,i}\}\beta\right)\right\}^2,$$

where $\hat{w}_{ij} = 1\{Z_i \leq Z_j\}1\{\hat{U}_{n,i} \leq \hat{U}_{n,j}\}$. Note that we do not resort to numerical optimization, as $\hat{\beta}$ has an explicit form from the least squares problem. The sample sizes were chosen from $\{200, 500, 800\}$ and the Monte Carlo simulation number was 2000.

Table 1 shows the performance of the estimators. There are four combinations, according to whether θ_0 is assumed to be known (TR) or estimated through maximum score estimation (ES) and according to 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. The bandwidth choice was made using a least-squares cross-validation method, selecting among ten equal-spaced points between 0 and 1.

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

		k		NN-TR	NN-ES	KN-TR	KN-ES
$n = 200$	DGP A1	3	MAE	0.4243	0.4234	0.4276	0.4417
			RMSE	0.2892	0.2881	0.2942	0.3088
		6	MAE	0.4089	0.4131	0.4105	0.4202
			RMSE	0.2616	0.2653	0.2652	0.2727
	DGP A2	3	MAE	0.4276	0.4297	0.4298	0.4386
			RMSE	0.2881	0.2890	0.2924	0.2991
		6	MAE	0.4334	0.4314	0.4331	0.4402
			RMSE	0.2909	0.2874	0.2868	0.3002
$n = 500$	DGP A1	3	MAE	0.2688	0.2696	0.2742	0.2783
			RMSE	0.1144	0.1157	0.1193	0.1221
		6	MAE	0.2620	0.2624	0.2616	0.2670
			RMSE	0.1093	0.1097	0.1090	0.1130
	DGP A2	3	MAE	0.2827	0.2820	0.2870	0.2894
			RMSE	0.1231	0.1237	0.1270	0.1290
		6	MAE	0.2641	0.2630	0.2636	0.2670
			RMSE	0.1100	0.1089	0.1095	0.1114
$n = 800$	DGP A1	3	MAE	0.2123	0.2124	0.2171	0.2188
			RMSE	0.0709	0.0708	0.0737	0.0746
		6	MAE	0.2067	0.2066	0.2072	0.2097
			RMSE	0.0670	0.0671	0.0672	0.0691
	DGP A2	3	MAE	0.2204	0.2214	0.2226	0.2268
			RMSE	0.0777	0.0781	0.0795	0.0818
		6	MAE	0.2112	0.2119	0.2124	0.2147
			RMSE	0.0697	0.0706	0.0706	0.0726

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 θ_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 to perform 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.

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 θ_0 (TR) and the case with the estimated parameter θ_0 (ES) is almost negligible. This again affirms the robustness of the bootstrap procedure to the quality of the first step estimator $\hat{\theta}$. 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 Conclusion

This paper finds that the influence of the first step index estimators in nonparametric functions is asymptotically negligible. A heuristic analysis was performed in terms of the Fréchet derivatives of a relevant class of functionals. Hence this phenomenon appears to have a generic nature. Then this paper proposes 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.

6 Appendix: Mathematical Proofs

Throughout the proofs, the notation C denotes a positive constant that may assume different values in different contexts.

6.1 The Proofs of the Main Results

Proof of Lemma 1 : 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.

Table 2: 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

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 $\mathbf{E}_{0,\varphi}$ denotes 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 $\mathbf{E}_{0,\varphi}[A_1] = 1$ as in the proof of Lemma A5 of Song (2009). Let $\tilde{\mu}_\varphi(\bar{\lambda}_j, \bar{\lambda}_0) \equiv \mathbf{E}_{0,\varphi}[\varphi(Y)A_j] / \mathbf{E}_{0,\varphi}[A_j] = \mathbf{E}_{0,\varphi}[\varphi(Y)A_j]$, $j = 0, 1$. Then,

$$\begin{aligned} \left\|\mu_\varphi(x; \lambda_1, \lambda_0) - \mu_\varphi(x; \lambda_0)\right\| &\leq \left\|\mu_\varphi(\bar{\lambda}_1, \bar{\lambda}_0) - \tilde{\mu}_\varphi(\bar{\lambda}_1, \bar{\lambda}_0)\right\| + \left\|\tilde{\mu}_\varphi(\bar{\lambda}_1, \bar{\lambda}_0) - \mu_\varphi(\bar{\lambda}_0)\right\| \\ &= (I) + (II), \text{ say.} \end{aligned}$$

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 $\lambda_1(X)$ given $\lambda_0(X) = \bar{\lambda}_0$. 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

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

Let us turn to (II) which we write as

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

where $V \equiv \varphi(Y) - \mathbf{E}_{0,\varphi}[\varphi(Y)]$ because $\mathbf{E}_{0,\varphi}[A_1] = 1$. The term (II) is bounded by the absolute value of

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

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

Proof of Lemma 2 : 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

$$\begin{aligned}
& \mathbf{E}\left[\psi(Z)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] = \mathbf{E}\left[\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_0(X)\right\}\right] \\
= \; & \mathbf{E}\left[\left(\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]-\mu_{\psi,0}(X)\right)^\top\left(\mu_{\varphi,\lambda}(X)-\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]\right)\right] \\
& +\mathbf{E}\left[\left(\mathbf{E}\left[\psi(Z)|\lambda(X),\lambda_0(X)\right]-\mu_{\psi,0}(X)\right)^\top\left(\mathbf{E}\left[\varphi(Y)|\lambda(X),\lambda_0(X)\right]-\mu_{\varphi,0}(X)\right)\right] \\
& +\mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right] \\
= \; & \mathbf{E}\left[\mu_{\psi,0}(X)^\top\left\{\mu_{\varphi,\lambda}(X)-\mu_{\varphi,0}(X)\right\}\right]+O(||\lambda-\lambda_0||_\infty^2)
\end{aligned}$$

by applying Lemma 1 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 1 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. ■

Proof of Theorem 1 : From the proof of Theorem 3.2.16 in van der Vaart and Wellner (1996), it can be shown that

$$\sqrt{n}(\hat{\beta}-\beta_0)=\Omega^{-1}\sqrt{n}\xi_n(\hat{\mu})+o_P(1).$$

Hence it suffces to show (7). Observe that for any $\lambda_j\in\Lambda_j$ and $\Delta_j=||\lambda_j-\lambda_{0,j}||_\infty$,

$$\begin{aligned}
& |F_{\lambda,j}(\lambda_j(x))-F_{0,j}(\lambda_{0,j}(x))| \\
\leq \; & P\left\{\lambda_{0,j}(X)\leq\lambda_{0,j}(x)+2\Delta_j\right\}-P\left\{\lambda_{0,j}(X)\leq\lambda_{0,j}(x)-2\Delta_j\right\}\leq C\Delta_j
\end{aligned}$$

by Assumption G1(i)(b). Hence by Assumption G1(i)(a), $\Lambda_j(\delta_n)$ with $\delta_n=n^{-b'}$, $b'\in(1/4,b)$, contains $\hat{\lambda}_j$ with probability approaching one. The result of the Bahadur representation in Lemma A1 below yields (7). ■

Proof of Theorem 2 : Write $\mu_0(x) = \mu(x;\lambda_0)$ and $\hat{\mu}(x) = \hat{\mu}(x;\hat{\lambda})$. Put briefly, $\hat{1}_{il} = 1\{\hat{W}_i \leq \hat{W}_l\}$ and $1_{il} = 1\{W_i \leq W_l\}$ and

$$\begin{aligned}\rho_i(\beta) &= \rho(V_i, \mu_0(X_i);\beta),\ \rho_{\mu,i}(\beta) = \rho_\mu(V_i, \mu_0(X_i);\beta),\\ \hat{\rho}_i(\beta) &= \rho(V_i, \hat{\mu}(X_i);\beta), \text{ and } \hat{\rho}_{\beta,i}(\beta) = \rho_\beta(V_i, \hat{\mu}(X_i);\beta).\end{aligned}$$

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

$$\hat{Q}(\beta) = \frac{1}{n}\sum_{l=1}^n \left\{\frac{1}{n}\sum_{i=1}^n \hat{\rho}_i(\beta)\hat{1}_{il}\right\}^2 \text{ and } \tilde{Q}(\beta) = \frac{1}{n}\sum_{l=1}^n \left\{\frac{1}{n}\sum_{i=1}^n \rho_i(\beta)1_{il}\right\}^2.$$

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 $\hat{g}_j(u) = \sum_{i=1}^n Y_{ji}K_h(\hat{U}_{n,i}^{(j)} - u)/\sum_{i=1}^n K_h(\hat{U}_{n,i}^{(j)} - u)$ and $g_j(u) = \mathbf{E}[Y^{(j)}|F_{0,j}(\lambda_{0,j}(X)) = u]$. Note that $||\hat{\mu} - \mu_0||_\infty$ is bounded by

$$\sup_{u\in[0,1]}||\hat{g}_j(u) - g_j(u)|| + \sup_{x\in\mathcal{X}}||g_j(F_{n,\hat{\theta},j}(\lambda_j(x;\hat{\theta}))) - g_j(F_{0,j}(\lambda_j(x;\theta_0)))||. \tag{11}$$

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, $||\hat{\mu}-\mu_0||_\infty = o_P(1)$. Now, for $\bar{\mu}(X_i)$ lying between $\hat{\mu}(X_i)$ and $\mu_0(X_i)$,

$$\begin{aligned}&\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^n \{\hat{\rho}_i(\beta) - \rho_i(\beta)\}\hat{1}_{il}\right| \leq \frac{||\hat{\mu}-\mu_0||_\infty}{n}\sum_{i=1}^n \sup_{\beta\in B}||\rho_\mu(V_i,\mu_0(X_i);\beta)|| \qquad (12)\\ &+\frac{||\hat{\mu}-\mu_0||_\infty^2}{2n}\sum_{i=1}^n \sup_{(\beta,\bar{\mu})\in B\times[-M,M]}||\rho_{\mu\mu}(V_i,\bar{\mu};\beta)||\\ =\ & o_P(1),\end{aligned}$$

by Assumption 1(iii)(iv).

Note also that from large n on,

$$\begin{aligned}&\mathbf{E}\left(\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^n \rho_i(\beta)\left(\hat{1}_{il} - 1_{il}\right)\right|\right) \qquad (13)\\ \leq\ & \frac{1}{n}\sum_{i=1,i\neq l}^n \left\{\mathbf{E}\left[\sup_{\beta\in B}|\rho_i(\beta)|^2\right]\right\}^{1/2}\sqrt{P\{U_{0,l}^{(J+1)} - \Delta_n < U_{0,i}^{(J+1)} \leq U_{0,l}^{(J+1)} + \Delta_n\}},\end{aligned}$$

where $\Delta_n = \max_{1\leq i\leq n}\sup_{\theta\in B(\theta_0,\delta_n)}||U_{n,\theta,i}^{(J+1)} - U_{0,i}^{(J+1)}||$, $\delta_n = n^{-1/3+\varepsilon}$, with small $\varepsilon > 0$, and $U_{n,\theta,i}^{(J+1)} = \frac{1}{n}\sum_{j=1,j\neq i}^n 1\{\lambda_{J+1}(X_j;\theta) \leq \lambda_{J+1}(X_i;\theta)\}$. Similarly as in the proof of Lemma A3 of

Song (2009), $\Delta_n = O_P(\delta_n)$, so that the last term in (13) is $o(1)$. From (12) and (13),

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

Since $\rho(v, \mu_0(x); \beta)$ is Lipschitz in β 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 β_0,

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

Using consistency of $\hat{\beta}$ and following similar steps in (12) and (13), we can show that $G_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\})$ is equal to

$$G_n(\beta_0, \mu_0, \{W_l\}) + o_P(1) = \int \dot{H}(w) \dot{H}(w)^\top dP_W(w) + o_P(1),$$

by the law of large numbers. We turn to the analysis of $\sqrt{n} \xi_n(\hat{\beta}, \hat{\mu}, \{\hat{W}_l\})$. Let $\mu_\theta(X_i) = \mathbf{E}[Y_i | \lambda(X_i; \theta)]$ and write

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\rho}_i(\beta_0) \hat{1}_{il} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \rho(V_i, \hat{\mu}(X_i); \beta_0) - \rho(V_i, \mu_{\hat{\theta}}(X_i); \beta_0) \right\} \hat{1}_{il} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \rho(V_i, \mu_{\hat{\theta}}(X_i); \beta_0) - \rho(V_i, \mu_0(X_i); \beta_0) \right\} \hat{1}_{il} \\ &= A_{1n} + A_{2n}, \text{ say.} \end{aligned}$$

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

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

where $\bar{\mu}(X_i)$ lies between $\hat{\mu}(X_i)$ and $\mu_{\hat{\theta}}(X_i)$. We deal with B_{2n} first. By Hölder inequality,

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

Note that $\mathbf{E}[\sup_{\bar{\mu}\in[-M,M]}||\rho_{\mu\mu}(V_i,\bar{\mu};\beta_0)||^q]<\infty$ and

$$\begin{aligned}
&\int_{\mathcal{S}_X}\left|\left(\hat{\mu}_r(x)-\mu_{r,\hat{\theta}}(x)\right)\left(\hat{\mu}_s(x)-\mu_{s,\hat{\theta}}(x)\right)\right|^{\frac{q}{q-1}}dP_X(x) \qquad (14)\\
=\;&\int_{\mathcal{D}_{1n}}\left|\left(\hat{\mu}_r(x)-\mu_{r,\hat{\theta}}(x)\right)\left(\hat{\mu}_s(x)-\mu_{s,\hat{\theta}}(x)\right)\right|^{\frac{q}{q-1}}dP_X(x)\\
&+\int_{\mathcal{D}_{2n}}\left|\left(\hat{\mu}_r(x)-\mu_{r,\hat{\theta}}(x)\right)\left(\hat{\mu}_s(x)-\mu_{s,\hat{\theta}}(x)\right)\right|^{\frac{q}{q-1}}dP_X(x),
\end{aligned}$$

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 (11) and in the proof of Lemma A4 of Song (2009), the first term is bounded by

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

where $w_n=n^{-1/2}h^{-1}\sqrt{-\log h}+h^2$ and $g_{r,\theta}(u)=\mathbf{E}[Y^{(r)}|F_{\theta,r}(\lambda_r(X;\theta))=u]$. Similarly, the last term in (14) is bounded by $C\int_{u\in[0,1]:|u-1|\leq h/2}\hat{D}(u)du$, where $\hat{D}(u)$ is equal to

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

When $|u-1|\leq 2h$, $\left|\left(\hat{g}_r(u)-g_{r,\hat{\theta}}(u)\right)\left(\hat{g}_s(u)-g_{s,\hat{\theta}}(u)\right)\right|^{\frac{q}{q-1}}=O_P(h^{\frac{2q}{q-1}})$ 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 (14) 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 $\lambda_{J+1}(X_i;\hat{\theta}) \leq \lambda_{J+1}(X_l;\hat{\theta})$. Then,

$$\begin{aligned} U^{(J+1)}_{n,\hat{\theta},i} &\leq \frac{1}{n-1}\sum_{k=1,k\neq i}^{n} 1\left\{\lambda_{J+1}(X_k;\hat{\theta}) \leq \lambda_{J+1}(X_i;\hat{\theta})\right\} \\ &= \frac{1}{n-1}\sum_{k=1,k\neq l}^{n} 1\left\{\lambda_{J+1}(X_k;\hat{\theta}) \leq \lambda_{J+1}(X_l;\hat{\theta})\right\} = U^{(J+1)}_{n,\hat{\theta},l}. \end{aligned}$$

Exchanging the roles of i and l, we find that if $\lambda_{J+1}(X_i;\hat{\theta}) \geq \lambda_{J+1}(X_l;\hat{\theta})$, $U^{(J+1)}_{n,\hat{\theta},i} \geq U^{(J+1)}_{n,\hat{\theta},l}$. Therefore, letting $\tilde{W}_{\theta,i} = (W_{1i}, U^{(J+1)}_{\theta,i})$ and $\tilde{1}_{i,\theta}(w) = 1\{\tilde{W}_{\theta,i} \leq w\}$, we write

$$1\{\hat{W}_i \leq \hat{W}_l\} = \tilde{1}_{i,\hat{\theta}}(\tilde{W}_{\hat{\theta},l}).$$

Using this, we write

$$B_{1n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \rho_\mu(V_i, \mu_{\hat{\theta}}(X_i);\beta_0)^\top \tilde{1}_{i,\hat{\theta}}(\tilde{W}_{\hat{\theta},l}) \left(\hat{\mu}(X_i) - \mu_{\hat{\theta}}(X_i)\right).$$

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

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

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

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

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

$$\log N_{[]}(\varepsilon,\mathcal{H}_n,||\cdot||_q) \leq C\log\varepsilon + C/\varepsilon, \text{ for } \varepsilon > 0. \tag{15}$$

By Lemma 1 and Assumption 3(i),

$$\begin{aligned} &\left\|\rho_\mu(v,\mu_{\theta_1}(x);\beta_0) - \rho_\mu(v,\mu_{\theta_2}(x);\beta_0)\right\| \\ \leq\ & C\sup_{\bar{\mu}\in[-M,M]}\left\|\rho_{\mu\mu}(v,\bar{\mu};\beta_0)\right\| \times \|\theta_1 - \theta_2\|. \end{aligned}$$

Therefore, using this, (9) and (15), we conclude that for $\Psi = \{\psi_{\theta,x,w} : (\theta, x, w) \in B(\theta_0, \delta_n) \times \mathcal{S}_X \times \mathcal{S}_{W_1}\}$,

$$\log N_{[]}(\varepsilon, \Psi, ||\cdot||_q) \leq C \log \varepsilon + C/\varepsilon, \text{ for } \varepsilon > 0. \tag{16}$$

By applying (Step 1) in the proof of Lemma A1 below, we find that $\tilde{\nu}_n(\theta, x, w)$ is equal to

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{l=1}^{J} \mathbf{E}\left[\psi^{(l)}_{\theta,x,w}(V_i, X_i, W_{1i}) | U^{(l)}_{\theta,i}\right] \left(Y^{(l)}_i - \mu_{\theta,l}(X_i)\right) + o_P(1) \\ = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{l=1}^{J} \mathbf{E}\left[\psi^{(l)}_{0,x,w}(V_i, X_i, W_{1i}) | U^{(l)}_{0,i}\right] \left(Y^{(l)}_i - \mu_{0,l}(X_i)\right) + o_P(1), \end{aligned}$$

uniformly over $(\theta, u) \in B(\theta_0, \delta_n) \times [0, 1]$, where $\psi^{(l)}_{\theta,x,w}$ denotes the l-th component of $\psi_{\theta,x,w}$ and $\psi^{(l)}_{0,x,w} = \psi^{(l)}_{\theta_0,x,w}$. The equality above follows from (Step 2) in the proof of Lemma A1 below. Therefore, we conclude that

$$A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{l=1}^{J} \mathbf{E}\left[\psi^{(l)}_{0,x,w}(V_i, X_i, W_{1i}) | U^{(l)}_{0,i}\right]_{w=\tilde{W}_{0,l}} \left(Y^{(l)}_i - \mu_{0,l}(X_i)\right) + o_P(1).$$

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

$$A_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\hat{\theta},X_l,W_{1l}}(V_i, X_i, W_{1i})^\top \left(\mu_{\hat{\theta}}(X_i) - \mu_0(X_i)\right).$$

Using previous arguments yielding (16), we can establish a similar bracketing entropy bound for $\mathcal{F}_n = \{\psi_{\theta,\bar{x},\bar{w}}(\cdot,\cdot,\cdot)\left(\mu_\theta(\cdot) - \mu_0(\cdot)\right) : (\theta, \bar{x}, \bar{w}) \in B(\theta_0, \delta_n) \times \mathcal{S}_X \times \mathcal{S}_{W_1}\}$. Following the usual stochastic equicontinuity arguments and using Lemma 1, Lemma 2 and Assumption 3(i), we deduce that

$$\begin{aligned} |A_{2n}| &\leq \sup_{(\theta,\bar{x},\bar{w})} \left|\sqrt{n}\mathbf{E}\left[\psi_{\theta,\bar{x},\bar{w}}(V_i, X_i, W_{1i})\left(\mu_\theta(X_i) - \mu_0(X_i)\right)\right]\right| + o_P(1) \\ &\leq \sqrt{n}\sup_{(\theta,\bar{x},\bar{w})} \left|\mathbf{E}\left[\psi_{0,\bar{x},\bar{w}}(V_i, X_i, W_{1i})\left\{\mu_\theta(X_i) - \mu_0(X_i)\right\}\right]\right| \\ &\quad + O(\sqrt{n}\delta_n^2) + o_P(1) = O(\sqrt{n}\delta_n^2) + o_P(1) = o_P(1), \end{aligned}$$

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

$$\begin{aligned} z_n(w) \equiv & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \rho_i(\beta_0) 1\{W_i \leq w\} \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{l=1}^{J} \mathbf{E}\left[\rho^{(l)}_{\mu,i}(\beta_0) 1\{W_i \leq w\} | U^{(l)}_{0,i}\right]^\top \left(Y^{(l)}_i - \mu_{0,l}(X_i)\right), \end{aligned}$$

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

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

Since $\sup_{w\in\mathbf{R}^{d_W}}|z_n(w)| = O_P(1)$, using (12) and (13) again, we conclude that

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

The wanted result now follows by applying the weak convergence of z_n to ζ and the continuous mapping theorem. ■

Proof of Theorem 3 : First, define $m(\beta;w) \equiv \mathbf{E}\left[\rho_l(\beta)1\{W_l \le w\}\right]$,

$$\begin{aligned}
\hat{m}_b(\beta;\hat{W}_k) &\equiv \frac{1}{n}\sum_{l=1}^{n}\left[\left\{\hat{\rho}_l(\hat{\beta}) - \hat{\rho}_l(\beta)\right\}\hat{1}_{lk} + \omega_{l,b}\left\{\rho_l(\hat{\beta})\hat{1}_{lk} + \hat{r}_{lk}^{\top}\{Y_l - \hat{\mu}(X_l)\}\right\}\right], \text{ and}\\
\tilde{m}_b(\beta;W_k) &\equiv \frac{1}{n}\sum_{l=1}^{n}\left[\{\rho_l(\beta_0) - \rho_l(\beta)\}1_{lk} + \omega_{l,b}\left\{\rho_l(\beta_0)1_{lk} + r_l^{\top}(W_k)\{Y_l - \mu_0(X_l)\}\right\}\right],
\end{aligned}$$

where $r_l(w) \equiv [r_l^{(1)}(w),\cdots,r_l^{(J)}(w)]^{\top}$ and $r_l^{(j)}(w) \equiv \mathbf{E}[\rho_l(\beta_0)1\{W_l \le w\}|U_l^{(j)}]$. Then, we introduce

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

and $\tilde{Q}(\beta) \equiv \mathbf{E}\left[m(\beta;W_k)^2\right]$. We first show that the bootstrap estimator is consistent conditional on $\mathcal{G}_n \equiv \{(V_i,Y_i,X_i,W_{1i})\}_{i=1}^n$ in probability. (Following the conventions, we use notations O_{P^*} and o_{P^*} that indicate conditional stochastic convergences given $\mathcal{G}_n$.) For this, it suffices to show that

$$\sup_{\beta\in B}|\hat{Q}_b^*(\beta) - \tilde{Q}(\beta)| = o_{P^*}(1) \text{ in } P. \tag{17}$$

For this, we first show that

$$\sup_{\beta\in B}|\hat{Q}_b^*(\beta) - \tilde{Q}(\beta)| = \sup_{\beta\in B}|\tilde{Q}_b^*(\beta) - \tilde{Q}(\beta)| + o_{P^*}(1) \text{ in } P. \tag{18}$$

Then the multiplier CLT of Ledoux and Talagrand (1988) (e.g. Theorem 2.9.7 of van der Vaart and Wellner (1996)) applied to $\{\tilde{m}_b(\beta;w) : (\beta,w) \in B\times\mathbf{R}^{d_W}\}$ yields that

$\sup_{\beta\in B}|\tilde{Q}_b^*(\beta)-\tilde{Q}(\beta)|=o_{P^*}(1)$ in P, affirming (17). We turn to (18). We write

$$\begin{aligned} & \hat{m}_b(\beta;\hat{W}_k)-\tilde{m}_b(\beta;W_k) \\ = \; & \frac{1}{n}\sum_{l=1}^{n}\left[\left\{\hat{\rho}_l(\hat{\beta})-\hat{\rho}_l(\beta)\right\}\hat{1}_{lk}-\{\rho_l(\beta_0)-\rho_l(\beta)\}\,1_{lk}\right]+\eta_n, \end{aligned} \tag{19}$$

where η_n is equal to

$$\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}\left[\rho_l(\hat{\beta})\hat{1}_{lk}-\rho_l(\beta_0)1_{lk}\right]+\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}\left[\hat{r}_{lk}^{\top}\{Y_l-\hat{\mu}(X_l)\}-r_l^{\top}(W_k)\{Y_l-\mu_0(X_l)\}\right].$$

It is not hard to show that the first sum in (19) is $o_P(1)$ uniformly in $(\beta,k)\in B\times\{1,\cdots,n\}$ using the similar arguments in the proof of Theorem 2. We show that $\eta_n=o_P(1)$. For a future use, we show a stronger statement:

$$\eta_n=o_P(n^{-1/2}). \tag{20}$$

Using the fact that $\omega_{l,b}$ is a bounded, mean-zero random variables independent of the data, we can follow the steps in the proof of Theorem 2 to show that the leading sum in the definition of η_n is $o_{P^*}(n^{-1/2})$ in P. We focus on the last sum in the definition of η_n which we write as

$$\begin{aligned} & \frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}Y_l^{\top}(\hat{r}_{lk}-r_l(W_k))-\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}\left[\hat{r}_{lk}^{\top}\hat{\mu}(X_l)-r_l^{\top}(W_k)\mu_0(X_l)\right] \\ = \; & \frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}\{Y_l-\mu_0(X_l)\}^{\top}(\hat{r}_{lk}-r_l(W_k))-\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}r_l^{\top}(W_k)\{\hat{\mu}(X_l)-\mu_0(X_l)\} \\ & -\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}(\hat{r}_{lk}-r_l(W_k))^{\top}\{\hat{\mu}(X_l)-\mu_0(X_l)\}. \end{aligned}$$

Using arguments used to deal with B_{2n} in the proof of Theorem 2, we can show that the last sum vanishes at the rate $o_{P^*}(n^{-1/2})$ in P. As for the first sum,

$$\begin{aligned}
& \mathbf{E}\left[\left|\frac{1}{n}\sum_{l=1}^{n}\omega_{l,b}\{Y_l-\mu_0(X_l)\}^\top(\hat{r}_{lk}-r_l(W_k))\right||\mathcal{G}_n\right] \\
= \; & \mathbf{E}\left[\left|\frac{1}{n}\sum_{l=1}^{n}\{\omega_{l,b}-\mathbf{E}\left[\omega_{l,b}|\mathcal{G}_n\right]\}\{Y_l-\mu_0(X_l)\}^\top(\hat{r}_{lk}-r_l(W_k))\right||\mathcal{G}_n\right] \\
\leq \; & o_P(n^{-1/2})\times\sqrt{\frac{1}{n}\sum_{l=1}^{n}\mathbf{E}\left[\{\omega_{l,b}-\mathbf{E}\left[\omega_{l,b}|\mathcal{G}_n\right]\}^2|\mathcal{G}_n\right]||Y_l-\mu_0(X_l)||^2} \\
= \; & o_P(n^{-1/2}).
\end{aligned}$$

Similarly, we can deduce that the second sum vanishes at the rate $o(n^{-1/2})$ conditional on $\mathcal{G}_n$ in P. Therefore, we obtain (20). This yields that

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

From this, we deduce (18) 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 2, 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

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

where $\bar{\beta}_b^*$ lies between $\hat{\beta}_b^*$ and $\hat{\beta}$. Again, similarly as in the proof of Theorem 2, we can show that

$$\begin{aligned}
G_n^*(\hat{\beta},\hat{\mu},\{\hat{W}_l\}) &= G_n(\beta_0,\mu_0,\{W_l\})+o_{P^*}(1)\text{ in }P \\
&= \int\dot{H}(w)\dot{H}(w)^\top dP(w)+o_P(1)+o_{P^*}(1)\text{ in }P.
\end{aligned}$$

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 β_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 (20),

$$\begin{aligned}\sqrt{n}\xi_n^*(\hat{\beta}_b^*, \hat{\mu}, \{\hat{W}_l\}) &= \frac{1}{\sqrt{n}}\sum_{k=1}^n \left\{ \frac{1}{n}\sum_{i=1}^n \rho_\beta(V_i, \hat{\mu}(X_i); \hat{\beta}_b^*)\hat{1}_{ik} \right\} \\ &\quad \times \left\{ \frac{1}{n}\sum_{i=1}^n \omega_{i,b}\left\{\rho_i(\beta_0)1_{ik} + r_i^\top(W_k)\{Y_i - \mu_0(X_i)\}\right\} \right\} + o_{P^*}(1) \text{ in } P.\end{aligned}$$

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

$$\frac{1}{n}\sum_{k=1}^n \dot{H}(W_k)\left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^n \omega_{i,b}\left\{\rho_i(\beta_0)1_{ik} + r_i^\top(W_k)\{Y_i - \mu_0(X_i)\}\right\} \right\} + o_{P^*}(1) \text{ in } P.$$

Let $\Gamma_n(f) = \int f(w)d\mathbb{P}_n(w)$ and $\Gamma(f) = \int f(w)dP_W(w)$, where $\mathbb{P}_n$ is the empirical measure of $\{W_k\}_{k=1}^n$. Then, choose any sequence f_n. Then, for a subsequence $f_{n'}$ such that $||f_{n'} - f||_\infty \to 0$, for some f, we have

$$\begin{aligned}\Gamma_{n'}(f) - \Gamma(f) &= \int f_{n'}(w)d\mathbb{P}_{n'}(w) - \int f(w)dP_W(w) \\ &= \int (f_{n'}(w) - f(w))\, d\mathbb{P}_{n'}(w) + \int f(w)d(\mathbb{P}_{n'}(w) - P_W(w)) \\ &= o(1) + o_{a.s.}(1),\end{aligned}$$

by the strong law of large numbers. Let

$$F_n(w; \mathcal{G}_n) = \frac{1}{\sqrt{n}}\sum_{l=1}^n \omega_{l,b}\left[\rho_l(\beta_0)1\{W_l \le w\} + r_l(w)^\top\{Y_l - \mu(X_l;\theta_0)\}\right] \times \dot{H}(w).$$

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

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

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

$$\Gamma_n(\tilde{F}_n(\cdot;\mathcal{G}_n)) \to_d \int \zeta(w)\dot{H}(w)dP_W(w),$$

by the continuous mapping theorem (e.g. Theorem 18.11 of van der Vaart (1998)). ■

6.2 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 Λ be a class of $\mathbf{R}$-valued functions on $\mathbf{R}^{d_X}$ with generic elements denoted by λ. We also let Φ and Ψ be classes of real functions on $\mathbf{R}^J$ and $\mathbf{R}^{d_Z}$ with generic elements φ and ψ. We fix $\lambda_0 \in \Lambda$ such that $\lambda_0(X)$ is continuous. 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 σ-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 A1 : (i) Classes Φ and Ψ 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 A2 : (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 A3 : (i) $K(\cdot)$ is symmetric, compact supported, twice continuously differentiable with bounded derivatives, and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^3 + n^{-1/2}h^{-2}(-\log h) \to 0$.

The following theorem offers the uniform representation of ν_n.

Lemma A1 : *Suppose that* Assumptions A1-A3 *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 representation remains the same when we replace $\nu_n(\lambda, \varphi, \psi)$ by $\nu_n(\lambda_0, \varphi, \psi)$.

Proof of Lemma A1 : 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 λ and λ_0 with $F_\lambda\circ\lambda$ there and using Assumption A1(ii)), it follows that

$$\sup_{\lambda\in\Lambda_n}\sup_{x\in\mathbf{R}^{d_X}}|F_{n,\lambda,i}(\lambda(x))-F_\lambda(\lambda(x))|=O_P(n^{-1/2}),\tag{21}$$

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 (21) and employing similar arguments around (14) 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

$$\sup_{\pi\in\Pi_n}|R_{2n}(\pi)|\to_P 0.\tag{22}$$

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

$$\begin{aligned} & \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} \delta_j^\lambda \\ & - \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \psi_i \Delta_{\varphi,\lambda,ij} K'_{h,ij} \delta_i^\lambda \\ = \;& B_{1n}(\pi) + B_{2n}(\pi), \text{ say.} \end{aligned} \tag{23}$$

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

$$\begin{aligned} A_{1n}(\pi) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,j}\right](U_{n,\lambda,j}-U_{\lambda,j}) \\ &\quad -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,\lambda,ij}K'_{h,ij}|U_{\lambda,i}\right](U_{n,\lambda,i}-U_{\lambda,i})+o_P(1) \\ &= D_{1n}(\pi)-D_{2n}(\pi)+o_P(1), \text{ say.} \end{aligned} \tag{24}$$

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

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

Similarly, using the first order differentiability of $g_{\psi,\lambda}(\cdot)$, we observe that

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

It is not hard to show that the sum above is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^1\int_0^1 g_{\psi,\lambda}(u_2)\left\{g_{\varphi,\lambda}(u_2)-g_{\varphi,\lambda}(u_1)\right\}K'\left(\frac{u_2-u_1}{h}\right)du_2\,(1\{U_{\lambda,j}\le u_1\}-u_1)\,du_1.$$

Therefore, $D_{1n}(\pi)=D_{2n}(\pi)+o_P(1)$ uniformly over $\pi\in\Pi_n$. We conclude that $\sup_{\pi\in\Pi_n}|A_{1n}(\pi)|=o_P(1)$, which completes the proof of (22).

It suffices for (Step 1) to show that

$$\sup_{\pi\in\Pi_n}\left|R_{1n}(\pi)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{\psi,\lambda}(U_{\lambda,i})\{\varphi(Y_i)-g_{\varphi,\lambda}(U_{\lambda,i})\}\right|=o_P(1). \tag{25}$$

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

$$\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}q^{\pi}_{n,ij}. \tag{26}$$

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 (26) as

$$\frac{1}{(n-1)\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left\{\mathbf{E}_{S_i}[q^{\pi}_{n,ij}]-\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]-\mathbf{E}[q^{\pi}_{n,ij}]\right\}+u_n(\pi). \tag{27}$$

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 $\max_{1\le i,j\le n}\sup_{\lambda\in\Lambda_n}||d^2_{\lambda,ji}||=O_P(n^{-1})$ by (21). 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

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

where $1_n = 1\left\{|U_{\lambda_0,i} - U_{\lambda_0,j}| \leq h + Cn^{-b}\right\}$. We bound the last term again by

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

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. The second term is equal to $O(n^{-1/2}h^{-2}) = o(1)$.

Proof of C2 : Note that $K'(\cdot/h)$ is unformly bounded and bounded variation. Let $\mathcal{K}_{1,\Lambda} = \{K'(\sigma(\cdot)/h) : (\sigma,h) \in \mathcal{I}_n \times (0,\infty)\}$, where $\sigma_{\lambda,u}(x) = (F_\lambda \circ \lambda)(x_1) - u$ and $\mathcal{I}_n = \{\sigma_{\lambda,u} : (\lambda,u) \in \Lambda_n \times [0,1]\}$. By Lemma A1 of Song (2009) and Assumption A1(ii),

$$\log N_{[]}(\varepsilon, \mathcal{K}_{1,\Lambda}, ||\cdot||_p) \leq \log N(\varepsilon, \mathcal{I}_n, ||\cdot||_\infty) + C/\varepsilon \leq C\varepsilon^{-b_\Lambda}. \tag{28}$$

Using (28) and following standard arguments, we can easily 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 $\max_{1\leq j\leq n}||\delta_j^\lambda|| = O_P(n^{-1/2})$, the wanted result follows because $O_P(n^{-1/2}h^{-2}) = o_P(1)$.

Proof of C3 : First we note that

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\pi\in\Pi_n}\left|\mathbf{E}_{S_i}[q^\pi_{n,ij}]\right|^2\right] \qquad (29)\\
\leq\ &\int_0^1\left\{g^2_{\tilde{\psi},\lambda_0}(t_1) + Cn^{-2b}\right\}\sup_{(\varphi,\lambda)\in\Phi\times\Lambda_n}\left[\int_0^1\{g_{\varphi,\lambda}(t_2) - g_{\varphi,\lambda}(t_1)\}K_h(t_2-t_1)dt_2\right]^2 dt_1.
\end{aligned}$$

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 (29) is $O(h^3)$.

Let $\mathcal{J}_n = \{h\mathbf{E}[q^\pi_{n,ij}|S_i = \cdot] : \pi \in \Pi_n\}$ with an envelope J 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 unformly bounded and bounded

variation. Let $\mathcal{K}_\Lambda = \{K(\sigma(\cdot)/h) : (\sigma, h) \in \mathcal{I}_n \times (0, \infty)\}$. Then by Lemma A1 of Song (2009),

$$\log N_{[]}(\varepsilon, \mathcal{K}_\Lambda, ||\cdot||_p) \leq \log N(\varepsilon, \mathcal{I}_n, ||\cdot||_\infty) + C/\varepsilon \leq C\varepsilon^{-b_\Lambda}. \tag{30}$$

Let us define $\tilde{\mathcal{J}}_n = \{hq_n^\pi(\cdot,\cdot) : \pi \in \Pi_n\}$. Observe that for any $\lambda_1, \lambda_2 \in \Lambda_n$,

$$\begin{aligned} \|g_{\varphi,\lambda_1}(F_{\lambda_1}(\lambda_1(\cdot))) - g_{\varphi,\lambda_2}(F_{\lambda_2}(\lambda_2(\cdot)))\|_\infty &\leq C||(F_{\lambda_1}\circ\lambda_1) - (F_{\lambda_2}\circ\lambda_2)||_\infty \text{ and} \\ \|g_{\psi,\lambda_1}(F_{\lambda_1}(\lambda_1(\cdot))) - g_{\psi,\lambda_2}(F_{\lambda_2}(\lambda_2(\cdot)))\|_\infty &\leq C||(F_{\lambda_1}\circ\lambda_1) - (F_{\lambda_2}\circ\lambda_2)||_\infty, \end{aligned} \tag{31}$$

by Lemma 1. From this, it is easy to show that

$$\log N_{[]}(\varepsilon, \tilde{\mathcal{J}}_n, ||\cdot||_{p/2}) \leq \log N_{[]}(\varepsilon/C, \Phi, ||\cdot||_p) + \log N_{[]}(\varepsilon/C, \Psi, ||\cdot||_p) + C\varepsilon^{-b_\Lambda}. \tag{32}$$

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)),

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

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 arbitrarily small $\Delta > 0$ and take $\eta = 1/4 + \Delta$ such that $\eta + 1/2 < 1 - 2/p$ 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, $\sup_{\pi\in\Pi_n}|u_{1n}(\pi)| = o_P(n^{-\eta+\Delta/2}h^{-1}) = o_P(n^{-1/4-\Delta/2}h^{-1}) = o_P(1)$. Hence the proof is complete.

Proof of C5 : We consider the following:

$$\begin{aligned} & \mathbf{E}\left[\sup_{\pi\in\Pi_n}\left\{\mathbf{E}_{S_j}[q^{\pi}_{n,ij}]-g_{\psi,\lambda}(U_{\lambda,j})\{\varphi(Y_j)-g_{\varphi,\lambda}(U_{\lambda,j})\}\right\}^2\right] \qquad (33)\\ = & \int \sup_{\pi\in\Pi_n}\left\{\int_0^1 A_{n,\psi}(t_1,t_2,w)dt_1\right\}^2 dF_{\lambda_0}(w,t_2), \end{aligned}$$

where $\int \cdot dF_{\lambda_0}$ denotes the integration with respect to the joint distribution of $(Y_i, U_{\lambda,i})$ and

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

After some tedious algebra, we can show that the last term in (33) 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 $\varepsilon>0$. Using (31) and standard arguments, we find that

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

Since $b_\Phi\vee b_\Psi\vee b_\Lambda<2$, the wanted bracketing integral entropy condition follows. We take an envelope which we choose as

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

for some large M. Clearly, this function F_M is $L_{2+\varepsilon}(P)$-bounded by Assumption A1. 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 Λ_n is a shrinking neighborhood of λ_0 and $\mathbf{E}[\Delta_i^{\varphi,\psi}(\lambda)-\Delta_i^{\varphi,\psi}(\lambda_0)]=0$, 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 method in econometrics," in *The Handbook of Econometrics*, Vol. IV, ed. by R. F. Engle and D. L. McFadden, Amsterdam: North-Holland.

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

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

[7] Chen, S. and S. Khan (2003), "Semiparametric estimation of a heteroskedastic sample selection model," *Econometric Theory* 19, 1040-1064.

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

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

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

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

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

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

[14] Härdle, W., P. and Tsybacov (1993), "How senstitive are average derivatives," *Journal of Econometrics*, 58, 31-48.

[15] Heckman, J. J. (1990), "Varieties of selection bias," *American Economic Review*, 80, 313-328.

[16] 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.

[17] Heckman, J. J., Ichimura, H. and P. Todd (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261-294.

[18] Hirano, K., G. W. Imbens, and G. Ridder (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71, 1161-1189.

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

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

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

[22] Liu, R. Y. (1988), "Bootstrap procedures under some non i.i.d. models," *Annals of Statistics*, 16, 1696-1708.

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

[24] Newey, W., Powell, J. and J. Walker (1990), "Semiparametric estimation of selection models: some empirical results," *American Economic Review,* 80:324-8.

[25] Powell, J. (1989), "Semiparametric estimation of bivariate latent variable models," Manuscript, University of Wisconsin Madison.

[26] Powell, J., Stock, J. and T. Stoker (1989), "Semiparametric estimation of index coefficients," *Econometrica*, 57, 6, 1403-1430.

[27] Robinson, P. (1988), "Root-N consistent nonparametric regression," *Econometrica,* 56, 931-954.

[28] Song, K. (2008), "Uniform convergence of series estimators over function spaces," *Econometric Theory*, 24, 1463-1499.

[29] Song, K. (2009), "Testing conditional independence using Rosenblatt transforms," Working paper, University of Pennsylvania.

[30] Stoker, T. (1986), "Consistent estimation of scaled coefficients," *Econometrica*, 54, 1461-1481.

[31] Stute, W. (1984), "Asymptotic normality of nearest neighbor regression function estimates," *Annals of Statistics*, 12, 917-926.

[32] Stute, W. and L. Zhu (2005): "Nonparametric checks for single-index models," *Annals of Statistics,* 33, 1048-1083.

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

[34] van der Vaart, A. W. (1996), "New Donsker classes," *Annals of Probability,* 24, 2128-2140.

[35] van der Vaart, A. W. (1998), *Asymptotic Statistics,* Cambridge University Press, New York.

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

[37] Wu, C. F. J. (1986), "Jacknife, bootstrap and other resampling methods in regression analysis," *Annals of Statistics*, 1261-1295.

[38] 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.