Bootstrap Confidence Sets under Semiparametric Conditional Moment Restrictions with Single-Index Components

Kyungchul Song[1]
University of Pennsylvania

Very Preliminary Draft

October 24, 2008

Abstract

This paper investigates models of semiparametric conditional moment restrictions where the restrictions contain a nonparametric function of a single-index as a nuisance parameter. It is assumed that this nonparametric function and the single-index are identified and estimated as a first step prior to the estimation of the parameter of interest under conditional moment restrictions. This paper finds that the estimated parameter of interest is robust to the quality of the estimated single-index component. More specifically, based on symmetrized nearest neighborhood estimation of this nonparametric function, this paper shows that the influence of the estimated single-index is asymptotically negligible even when the estimated single-index follows cube-root asymptotics. Using this finding, this paper proposes a method to construct bootstrap confidence sets that have three characteristics. First, the confidence sets are asymptotically valid in the presence of $n^{1/3}$-converging single-index components. Second, the confidence sets accommodate conditional heteroskedasticity. Third, the method is computationally easy as it does not require re-estimation of the single-index for each bootstrap sample. Some results from Monte Carlo simulations are presented and discussed.

Key words and Phrases: Conditional Moment Restrictions; Single-Index Restrictions; Boostrap; Confidence Sets.

JEL Classifications: C12, C14, C52.

[1]Department of Economics, University of Pennsylvania, 528 McNeil Building, 3718 Locust Walk, Philadelphia, Pennsylvania 19104-6297. Email: kysong@sas.upenn.edu

1 Introduction

In empirical researches of labor economics or development economics, a number of variables of demographic characteristics are used to alleviate various endogeneity problems. However, using too many covariates in nonparametric estimation tends to worsen the quality of the empirical results significantly. A promising approach to deal with this situation would be to introduce a single-index restriction. With a single-index restriction, one can retain flexibility in the specification of the regression function while avoiding the curse of dimensionality. Not surprisingly, the single-index restriction has long been actively investigated in the literature of statistics and econometrics for this reason. For example, Klein and Spady (1993) and Ichimura (1993) proposed M-estimation approaches to estimate the single-index coefficients. An alternative approach is the approach of average derivatives pioneered by Stoker (1986) and Powell, Stock and Stoker (1989). See also Härdle and Tsybakov (1993), Horowitz and Härdle (1996), and Hristache, Juditsky and Spokoiny (2001). A single-index restriction can be tested using various methods. To name but a few, see Fan and Li (1996), Stute and Zhu (2005) and Escanciano and Song (2008) and references therein.

Most literatures deal with a single-index model as an isolated object, whereas empirical researchers often face the necessity to use the single-index specification in the context of estimating a larger model. A prototypical example is a structural model in labor economics that requires a prior estimation of components such as wage equations. When single-index components are plugged in a larger context of extremum estimation, the introduction of single-index restrictions do not improve the rate of convergence of an extremum estimator which already achieves the parametric rate of $\sqrt{n}$. However, the use of single-index restrictions in such a situation have several own merits. The use of a single-index restriction requires weaker assumptions on the nonparametric function and on the kernel function. When the nonparametric function is defined on a space of a large dimension, stronger conditions on the nonparametric function and higher-order kernels are usually required. (See Hristache, Juditsky and Spokoiny (2001) for more details.)

This paper focuses on semiparametric conditional moment restrictions where the restrictions contain nonparametric functions of single-indices that are identified and estimated prior to the estimation of the parameter of interest. Surprisingly, this paper finds that the influence of the estimated single-indices is asymptotically negligible, and that this is true even when the estimated single-indices follow cube-root asymptotics. In other words, the quality of the estimated parameter of interest is robust to the quality of the first step single-index estimators. To investigate this phenomenon closely in the light of Newey (1994), this paper considers functionals that involve conditional expectations where the conditioning variable

involves an unknown parameter. In this situation, we show that the first order Fréchet derivative of the functional with respect to the unknown parameter is zero. This means that the influence of the parameter is negligible as long as the estimator has a convergence rate $o(n^{-1/4})$. Therefore, the phenomenon has a generic nature, although this paper uses a specific nonparametric estimation method for concreteness.

Utilizing this finding, this paper proposes a bootstrap procedure that has three characteristics. First, the bootstrap procedure is valid even when the single-index component follows cube-root asymptotics. This is interesting in the light of the fact that bootstrap confidence sets of semiparametric estimators that follow cube-root asymptotics are invalid. (Abrevaya and Huang (2005)). Nevertheless, this paper's proposal affirms that this is no longer a problem when the $n^{1/3}$-converging single-index estimator enters as a plug-in first step estimator. Second, the bootstrap method accommodates conditional heteroskedasticity. Note that conditional heteroskedascity is natural for models under conditional moment restrictions. Third, the bootstrap method does not require re-estimation of the single-index component or the nonparametric function for each bootstrap sample. Hence it is computationally attractive when the dimension of the single-index coefficient vector is large and its estimation involves numerical optimization. This is indeed the case when the single-index is estimated through maximum score estimation and the number of covariates is large.

While the asymptotic negligibility of the $n^{1/3}$-converging, estimated single-index has a generic nature, for the sake of concreteness, this paper's result is built on a uniform Bahadur representation of symmetrized nearest neighborhood estimators over function spaces that is established in the appendix. A Bahadur representation of such type was originally found by Stute and Zhu (2005) who established a non-uniform result in a restrictive context. This paper puts their finding in a broader perspective and shows that the phenomenon arises even when the single-index component has a convergence rate slower than $n^{-1/2}$. The representation is also useful for many other purposes, for example, for analyzing various semiparametric specification tests.

The paper is organized as follows. In the next section, we introduce models of conditional moment restrictions that have nonparametric functions of single-index components. Section 3 introduces an estimation method and establishes the asymptotic distribution of the estimator. Section 4 is devoted to the bootstrap method that this paper proposes. Section 5 presents and discusses Monte Carlo simulation results. Section 6 concludes. Technical proofs are relegated to the Appendix. In the Appendix, a general uniform Bahadur representation of symmetrized nearest neighborhood estimators is presented.

2 Semiparametric Conditional Moment Restrictions

In this section, we define the scope of this paper by introducing models under semiparametric conditional moment restrictions. Let $\{S_i\}_{i=1}^n$ be an observable random sample from a distribution P of a random vector $S = (S_1, W_1)$ taking values in $\mathbf{R}^{d_{S_1}+d_{W_1}}$. Let S_1 be constituted by (possibly overlapping) subvectors Y, X, and V, which take values in $\mathbf{R}^{d_Y}$, $\mathbf{R}^{d_X}$, and $\mathbf{R}^{d_V}$ respectively. Define a semiparametric $\mathbf{R}^{d_Y}$-valued function:

$$\mu(X;\theta_0) \equiv \mathbf{E}\left[Y|\lambda(X;\theta_0)\right], \text{ for some } \theta_0 \in \mathbf{R}^{d_\Theta} \tag{1}$$

where $\lambda(\cdot;\theta_0)$ is a real function known up to $\theta_0 \in \mathbf{R}^{d_\Theta}$. A prototypical example is a linear index, $\lambda(X;\theta_0) = X^\top\theta_0$. In this paper, we assume that the distribution of $\lambda(X;\theta_0)$ is absolutely continuous. Then, this paper focuses on the following type of conditional moment restrictions:

$$\mathbf{E}\left[\rho(V, \mu(X;\theta_0);\beta_0)|W\right] = 0,$$

where $W = (W_1, \lambda(X;\theta_0))$, W_1 being some observable random vector in $\mathbf{R}^{d_{W_1}}$, and $\rho(\cdot,\cdot;\beta_0) : \mathbf{R}^{d_V+d_Y} \to \mathbf{R}$ is known up to an unknown finite-dimensional parameter $\beta \in \mathcal{A} \subset \mathbf{R}^{d_\beta}$. The function ρ_β is often called a *generalized residual function,* reminiscent of a residual in the regression.

In some situations, the parameter θ_0 in the single-index component is estimated jointly with β in the GMM estimation. (See e.g. Ichimura (1993).) Such a situation is not of our interest, because the main interest of this paper, i.e., the effect of the estimation error in $\hat{\theta}$ upon the asymptotic covariance matrix of $\hat{\beta}$, becomes irrelevant in this situation. Hence this paper maintains the framework in which the object of interest is β_0 and the single-index component $\lambda(\cdot;\theta)$ is a nuisance parameter that is identified and estimated prior to the GMM estimation step.

This paper's framework is different from the literatures of conditional moment restrictions with endogeneity on the nonparametric components. (See Ai and Chen (2003), and Newey and Powell (2003) and references therein.) This paper assumes that the semiparametric regression function μ and the single-index coefficient θ_0 are identified and estimated as a first step. Hence the framework does not encounter an ill-posed inverse problem while allowing for endogeneity of $\mu(X;\theta_0)$. As we illustrate by examples below, this set-up is empirically relevant in many situations.

One further extension that this paper's set-up contains as compared to the existing literature is that the instrumental variable W is only partially observed as it is allowed to depend on $\lambda(X;\theta_0)$. In some cases, it is more relevant to assume what we call *single-index*

exogeneity, where exogeneity is required of some unknown combination of the components in X, rather than all the components. In many practical situations of empirical researches, confirming exogeneity can be a delicate matter both in terms of theoretical reasoning and in terms of empirical testing. In this situation, single-index exogeneity can be a more reasonable assumption than exogeneity on the whole vector X. This paper's framework accommodates such a situation. In this paper, we assume that the single-index $\lambda(X;\theta_0)$ in W is precisely the same single-index that constitutes the conditioning variable in μ. This specificity is motivated by some examples that we consider below and by the consideration that concreteness of the set-up is preferable to an abstract generality given the space constraint of this paper. The paper's framework can be applied to a situation where $W = (W_1, \tilde{\lambda}(W_2;\gamma_0))$, as long as $\tilde{\lambda}$ is a known function up to an unknown finite dimensional parameter γ_0 and this parameter γ_0 is identified and estimated as a first step.

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

$$\begin{aligned} Y^* &= \beta_0^\top Z + v \text{ and} \\ D &= 1\{X^\top\theta_0 \geq \varepsilon\}, \end{aligned}$$

where the first equation is an outcome equation with Y^* denoting the latent outcome variable and Z a vector of covariates that affect the outcome. The binary variable D represents the selection of the variable into the observed data set, (Y, Z) is observed only when $D = 1$. The incidence of selection is governed by a single index $X^\top\theta_0$ of another set of covariates. The variables v and ε represent unobserved heterogeneity in the individual observation. We assume that ε can be correlated with X but maintains that

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

In this model, we also assume that Z is independent of (v, ε) conditional on the observable component $X^\top\theta_0$ in the selection mechanism. Therefore, an individual component of X can be correlated with v. Note that in this case,

$$\mathbf{E}\left[v|D = 1, Z, X^\top\theta_0\right] = \frac{\mathbf{E}\left[v1\{X^\top\theta_0 \geq \varepsilon\}|X^\top\theta_0\right]}{P\{X^\top\theta_0 \geq \varepsilon|X^\top\theta_0\}} = \tau(X^\top\theta_0), \text{ say.}$$

Hence conditional on the event that the sample is selected, we write the following model for observed data set $(D_i = 1)$,

$$Y = \beta_0^\top Z + \tau(X^\top\theta_0) + u,$$

where u satisfies that $\mathbf{E}[u|D=1,Z,X^\top\theta_0]=0$ and τ is an unknown nonparametric function. This model can be estimated by using the method of Robinson (1988). First observe that

$$Y_i-\mu_Y(X_i^\top\theta_0)=\beta_0^\top\{Z_i-\mu_Z(X_i^\top\theta_0)\}+\eta_i \tag{2}$$

where $\mathbf{E}[\eta|D=1,Z,X^\top\theta_0]=0$, $\mu_Y(\cdot)=\mathbf{E}[Y|D=1,X^\top\theta_0=\cdot]$, and $\mu_Z(\cdot)=\mathbf{E}[Z|D=1,X^\top\theta_0=\cdot]$. Note that we do not impose a single-index restriction on nonparametric functions $\mathbf{E}[Y|D=1,X=\cdot]$ and $\mathbf{E}[Z|D=1,X=\cdot]$. The single-index restictions in this situation naturally stem from the selection mechanism D and the assumption that Z is independent of (v,ε) conditional on $X^\top\theta_0$. Then, the identifying restriction of β_0 can be written as the following conditional moment restriction:

$$\mathbf{E}\left[\{Y_i-\mu_Y(X_i^\top\theta_0)\}-\beta_0^\top\{Z_i-\mu_Z(X_i^\top\theta_0)\}|D_i=1,Z_i,X_i^\top\theta_0\right]=0.$$

This model falls into the framework of this paper with $W_i=(X_i^\top\theta_0,Z_i)$.

In this situation, one may consider estimating θ_0 first using the maximum score estimation and μ_Y and μ_Z and then plugging these in the moment restrictions to estimate β_0. The nuisance parameter estimator $\hat{\theta}$ in the first step follows the cube-root asymptotics and it is this paper's main concern how the estimator affects the estimator of β_0. ■

Example 2: Endogenous Binary Regressors with a Median Restriction: Consider the following model:

$$\begin{aligned} Y &= Z^\top\beta_0+D\gamma+\varepsilon, \text{ and} \\ D &= 1\{X^\top\theta_0\geq\eta\}, \end{aligned}$$

where ε satisfies that $\mathbf{E}[\varepsilon|X^\top\theta_0]=0$ and $Med(\eta|X)=0$. Therefore, the index $X^\top\theta_0$ plays the role of the instrumental variable (IV). However, the IV exogeneity condition is weaker than the assumptions used in the literature. First, 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. The researcher does not have to know which combination of components in X will be exogenous. It is not assumed either that η and X are independent. It only assumes a weaker condition that $Med(\eta|X)=0$. The model allows that the distribution of η depends on X in a certain way. The model also allows conditional heteroskedasticity for ε. In other words, we assume that $\mathbf{E}[\varepsilon^2|X]$ is a function of X that is an unknown form. In this case, we can write

$$\mathbf{E}\left[Y-Z^\top\beta_0-P\{D=1|X^\top\theta_0\}\gamma|X^\top\theta_0\right]=0.$$

Again, this conditional moment restriction is a special case of this paper's framework with $W = X^\top \theta_0$. ■

The main thesis of this paper can also be applied to a situation where the estimator is explicitly defined as an estimated functional of data that involves a nonparametric function. Examples include average derivative estimators (Stoke (1986)) and treatment effect estimators (Hirano, Imbens, and Ridder (2003)). Since the analysis in this case is simpler than the case of this paper, we omit the development in this direction for brevity.

3 Estimation and Construction of Confidence Sets

3.1 Estimation

In this section, we consider estimation of $\mu(X;\theta_0)$. First we estimate the single index $\lambda(X;\theta_0)$ to obtain $\lambda(X;\hat{\theta})$. The estimation of the coefficient in the single-index has long been investigated in many researches. There are researches that employed the M-estimation approach in estimating θ_0. For example, see Klein and Spady (1993) and Ichimura (1993). An alternative approach is the approach of average derivatives pioneered by Stoker (1986) and Powell, Stock and Stoker (1989). See also Härdle and Tsybakov (1993), Horowitz and Härdle (1996), and Hristache, Juditsky and Spokoiny (2001).

In this paper, following Stute and Zhu (2005), we reparametrize the nonparametric function μ by using the probability integral transform of the single-index. More specifically, let F_θ be the distribution function of $\lambda(X;\theta)$ and define $U_\theta = F_\theta(\lambda(X;\theta))$ and $U = F_{\theta_0}(\lambda(X;\theta_0))$. Based on this reparametrization, we consider a symmetrized nearest neighborhood estimator of μ. More specifically, define

$$\hat{U}_{n,j} = \frac{1}{n}\sum_{i=1}^{n} 1\{\lambda(X_i;\hat{\theta}) \le \lambda(X_j;\hat{\theta})\}.$$

Then, we consider using the following estimator:

$$\hat{\mu}(X_j;\hat{\theta}) = \frac{\sum_{i=1}^{n} Y_i K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}{\sum_{i=1}^{n} K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}, \tag{3}$$

where $K_h(u) = K(u/h)/h$ and $K : [0,1] \to \mathbf{R}$ is a kernel function. This is a symmetrized nearest neighborhood estimator proposed by Yang (1981). Since the probability integral transform of $\lambda(X;\theta_0)$ turns its density into a uniform density on $[0,1]$, we could use constant

1 in the denominator of the estimator $\hat{\mu}(X_j;\hat{\theta})$ in (3). However, this paper proposes using the kernel density estimator in place of the true density in (3). This approach ensures the uniform convergence of nonparametric regression estimator of $\mathbf{E}[Y|U=\cdot]$ including the boundary points of $[0,1]$. (See Song (2008b)).

The use of probability integral transform has several merits. First, it simplifies the problem of choosing an appropriate kernel function. Note that the choice of a kernel function eventually requires certain prior information about the density function of $\lambda(X;\theta_0)$ which depends on the unknown parameter θ_0. Second, the probability integral transform obviates the need to introduce a trimming sequence. The trimming sequence is often required for theoretical reasons (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 elliminates such a nuisance altogether.

Suppose that we have a prior estimator $\hat{\mu}(X;\hat{\theta})$ of $\mu(X;\theta)$ using the data set $\{Y_i, X_i\}_{i=1}^n$. Then, following Domínguez and Lobato (2004), we can estimate β as follows:

$$\hat{\beta} = \underset{\beta\in B}{\operatorname{argmin}} \sum_{j=1}^{n} \left\{ \sum_{i=1}^{n} \rho(V_i, \hat{\mu}(X_i;\hat{\theta});\beta) 1\{\hat{W}_i \le \hat{W}_j\} \right\}^2 ,$$

where $\hat{W}_j = (W_{1j}, \hat{U}_{n,j})$. In the following, we present the results of asymptotic properties of the estimator $\hat{\beta}$. We introduce the following assumptions:

Assumption 1: (i) The sample $\{S_i\}_{i=1}^n$ is a random sample.
(ii) $\mathbf{E}[\rho(V,\mu(X;\theta_0);\beta)|W]=0$ a.s. if and only if $\beta=\beta_0$ where $\beta_0 \in \text{int}(B)$, B a compact set in $\mathbf{R}^{d_\Theta}$.
(iii) $\rho(v,\mu;\beta)$ as a function of $(\mu,\beta)\in \mathbf{R}^{d_Y}\times\mathbf{R}^{d_\Theta}$ is second order continuously differentiable in (μ,β) with derivatives $\rho_\beta(v,\mu;\beta)$, $\rho_\mu(v,\mu;\beta)$, $\rho_{\beta\beta}(v,\mu;\beta)$, $\rho_{\beta\mu}(v,\mu;\beta)$ and $\rho_{\mu\mu}(v,\mu;\beta)$, such that for some $\delta>0$, $\mathbf{E}[\sup_{\beta\in B}||\tilde{\rho}(V,\mu(X;\theta_0);\beta)||^p]<\infty$, $p>4$, for all $\tilde{\rho}\in\{\rho_\beta,\rho_\mu,\rho_{\beta\beta},\rho_{\beta\mu},\rho_{\mu\mu}\}$.

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

Assumption 3: (A) There exist $\delta>0$ and $C>0$ such that the following three conditions hold.
(i) For each $\theta\in\Theta(\delta)\equiv\{\theta\in\mathbf{R}^{d_\Theta}: ||\theta-\theta_0||<\delta\}$, $\lambda(X;\theta)$ is a continuous random variable, and

$$\sup_{\theta\in\Theta(\delta)}|F_\theta(\lambda_1)-F_\theta(\lambda_2)| \le C|\lambda_1-\lambda_2|, \text{ for all } \lambda_1,\lambda_2\in\mathbf{R}.$$

(ii)(a) The conditional density $f(s|u)$ of S given $U=u$ with respect to a σ-finite measure is L times continuously differentiable in u, $L>8$, with derivatives $f^{(j)}(s|u)$ such that

$\sup_{(s,u)\in\mathcal{S}\times[0,1]}|f^{(j)}(s|u)/f(s)| < C$ where $f(s)$ is the density of S with respect to a σ-finite measure and $\mathcal{S}$ is the support of S.
(b) For each $\theta \in \Theta(\delta)$, the conditional density $f_\theta(s|u_1, u_2)$ of S given $(U_\theta, U) = (u_1, u_2)$ with respect to a σ-finite measure satisfies that for all $\nu > 0$

$$\sup_{(s,u_2)\in\mathcal{S}\times[0,1]}|f_\theta(s|u_1 - \nu, u_2) - f_\theta(s|u_1 + \nu, u_2)| \leq C\eta_\theta(s),$$

where $\eta_\theta : \mathbf{R}^{d_S} \to \mathbf{R}_+$ is such that $\sup_{(s,\theta)\in\mathcal{S}\times\Theta(\delta)}|\eta_\theta(s)/f(s)| < C$.
(B) $\mathbf{E}|Y|^p < \infty$, $p > 4$, and $\mathbf{E}[Y|U = \cdot]$ is bounded and twice continuously differentiable with bounded derivatives.
(C) For some strictly bounded map $G : \mathbf{R} \to [0, 1]$ and a constant $C > 0$,

$$|(G \circ \lambda)(x; \theta_1) - (G \circ \lambda)(x; \theta_2)| \leq C||\theta_1 - \theta_2||$$

for all $\theta_1, \theta_2 \in \Theta(\delta)$.

Assumption 1 is standard in models of conditional moment restrictions. Assumption 2 requires that the estimator $\hat{\theta}$ has the convergence rate of either $n^{-1/2}$ or $n^{-1/3}$. The moment conditions Assumption 3A(ii) requires more explanation as it does not appear in the literature often. This assumption is introduced to control the behavior of conditional expectations given $\lambda_\theta(X)$ when θ is perturbed around θ_0. This assumtion does not require that the distribution of (V, W, Y) be absolutely continuous. When the vector is discrete, we may view $f_\theta(w_1, y, y|u)$ as a conditional probability mass function. Assumption 3(C) is a regularity condition for the single-index function $\lambda(\cdot; \theta)$. When $\lambda(X; \theta) = X^\top\theta$, we can choose G to be a normal cdf function to fulfill (C). Introduction of the map G is made to emphasize the flexibility of the specifications of $\lambda(\cdot; \theta)$. The map G is not used for actual inference.

Assumption 4: (i) $K(\cdot)$ is bounded, symmetric, compact supported, infinite times continuously differentiable with bounded derivatives and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^4 + n^{-1/2}h^{-2} \to 0$.

The condition for the kernel in Assumption 4(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 4(ii) does not require undersmoothing. The bandwidth condition is satisfied for any $h = n^{-s}$ with $1/8 < s < 1/4$.

Theorem 1: *Suppose that* Assumptions 1-4 *hold. Then,*

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

where $\dot{H}(w) = \mathbf{E}[\rho_\beta(V_i; \mu(X_i; \theta_0); \beta_0) 1\{W_i \leq w\}]$ *and* B *is a centered Gaussian process on* $\mathbf{R}^{d_W}$ *that has a covariance kernel given by*

$$C(w_1, w_2) = \mathbf{E}\left[\xi_i(w_1)\xi_i(w_2)\right]$$

and

$$\begin{aligned} \xi_i(w) &= \rho(V_i; \mu(X_i; \theta_0); \beta_0) 1\{W_i \leq w\} \\ &\quad -\mathbf{E}\left[\rho_\mu(V_i; \mu(X_i; \theta_0); \beta_0)^\top 1\{W_i \leq w\} | U_i\right] (Y_i - \mu(X_i; \theta_0)). \end{aligned} \tag{4}$$

Compared with the asymptotic covariance matrix of Domínguez and Lobato (2004), the asymptotic covariance matrix contains an additional term involving $Y_i - \mu(X_i; \theta_0)$ in the covariance kernel (4). This is due to the nonparametric estimation error in $\hat{\mu}$. It is important to note at this point that the asymptotic covariance matrix remains the same regardless of whether we use the estimated single index $\lambda(X_i; \hat{\theta})$ or the true single-index $\lambda(X_i; \theta_0)$. This is true even if $\hat{\theta}$ converges at the rate of $n^{-1/3}$.

3.2 A Heuristic Analysis

The result of Theorem 1 shows that the influence of the estimator $\hat{\theta}$ can be ignored in this situation even if $\hat{\theta}$ converges at the rate of $n^{-1/3}$. This phenonomenon appears unexpected because the estimator $\hat{\theta}$ does not even have a usual asymptotic linear representation in this situation. In this section, we attempt to put this phenomenon in perspective in the light of Newey (1994) who systematically exposited how the first-step estimators affects the asymptotic covariance matrix of the second step estimators. In analyzing the effect, it is crucial to investigate the behavior of the parameter in response to the perturbation of the nuisance parameter. To put the result of Theorem 1 in perspective, we introduce a generic set-up. Let $l_\infty(\mathbf{R}^{d_X})$ be a Banach space of bounded functions equipped with the sup norm $||\cdot||_\infty : ||f||_\infty = \sup_{x \in \mathbf{R}^{d_X}} |f(x)|$. Then, let $\Lambda \subset l_\infty(\mathbf{R}^{d_X})$ be a class of real functions on $\mathbf{R}^{d_X}$ and define $\mu(X; \lambda) = \mathbf{E}[Y|\lambda(X)]$ for $\lambda \in \Lambda$. Given a random vector ξ, suppose that the parameter of focus takes the form of

$$\Gamma(\lambda) = \mathbf{E}\left[\mu(X; \lambda)^\top \xi\right].$$

Then, we show that under certain generic conditions, the parameter $\Gamma(\lambda)$ has the first order Fréchet derivative (in λ) equal to zero.

Theorem 2: *Let* λ_0 *be a fixed function such that* $\lambda_0(X)$ *is continuous and let* Λ_0 *be the*

collection of paths $\lambda_t,\ t \in [0,1]$ *passing through* λ_0 *that satisfy the following property*:

(RC) *There exists* $C > 0$ *such that*

$$\sup_{y,\bar{\xi} \in \mathbf{R}^{d_Y} \times \mathbf{R}^d} \left| f_t(y, \bar{\xi} | \bar{\lambda}_1 + \delta, \bar{\lambda}_2) - f_t(y, \bar{\xi} | \bar{\lambda}_1 - \delta, \bar{\lambda}_2) \right| < C\delta,$$

where $f_t(y, \bar{\xi} | \bar{\lambda}_1, \bar{\lambda}_2)$ *denotes the conditional pdf of* (Y, ξ) *given* $(\lambda_t(X), \lambda_0(X)) = (\bar{\lambda}_1, \bar{\lambda}_2)$ *with respect to a* σ*-finite measure and* C *is a constant independent of* $\lambda_t,\ \bar{\lambda}_1$ *and* $\bar{\lambda}_2$.

Then, the first order Fréchet derivative of $\Gamma(\lambda)$ *in* λ *in* Λ_0 *is zero.*

The regularity condition (RC) does not require that (Y, ξ) be a continuous random vector. Neither does the condition concerns the behavior of the conditional density function f_t at the perturbation of t at 0. The condition is merely an equicontinuity condition: for each t, the function should be Lipschitz continuous in the conditioning variable in a manner uniform over t. The result of Theorem 2 shows that the parameter of this form has Fréchet derivative equal to zero. This fact is solely due to the property of conditional expectations (except for the condition RC that is used in the proof.)

The implication of Theorem 2 is that in the case of i.i.d. series where we consider $t = c/\sqrt{n} \to 0$, the influence of λ upon the estimator of $\Gamma(\lambda_0)$ is negligible as long as $||\lambda - \lambda_0||_\infty = o(n^{-1/4})$. To see this clearly in the context of conditional moment restriction, observe that the influence of the nonparametric estimation $\hat{\mu}$ and $\hat{\theta}$ is summarized in the following difference: for λ such that $||\lambda - \lambda_0||_\infty = o_P(n^{-1/4})$,

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho \left\{ (V_i, \hat{\mu}(X; \lambda); \beta_0) - \rho(V_i, \mu(X; \lambda_0); \beta_0) \right\} \qquad (5) \\
\approx\ & \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho_\mu(V_i, \mu(X_i; \lambda_0); \beta_0) \left\{ \hat{\mu}(X_i; \lambda) - \mu(X_i; \lambda_0) \right\} \\
=\ & \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho_\mu(V_i, \mu(X_i; \lambda_0); \beta_0) \left\{ \hat{\mu}(X_i; \lambda) - \mu(X_i; \lambda) \right\} \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho_\mu(V_i, \mu(X_i; \lambda_0); \beta_0) \left\{ \mu(X_i; \lambda) - \mu(X_i; \lambda_0) \right\}
\end{aligned}$$

By using the Hoeffding's decomposition and the usual arguments of U-process theory, the

second to the last sum is asymptotically equivalent to

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{E}\left[\rho_\mu(V_i, \mu(X_i;\lambda_0);\beta_0)|\lambda(X)\right]\{Y_i - \mu(X_i;\lambda)\} \\ \approx \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{E}\left[\rho_\mu(V_i, \mu(X_i;\lambda_0);\beta_0)|\lambda_0(X)\right]\{Y_i - \mu(X_i;\lambda_0)\}, \end{aligned}$$

using arguments of stochastic equicontinuity. The last sum above which was originated from the estimation error of $\hat{\mu}$ contributes to the asymptotic covariance matrix of $\hat{\beta}$. After subtracting its mean, the last sum in (5) becomes asymptotically negligible, so that we are left with

$$\sqrt{n}\mathbf{E}\left[\rho_\mu(V_i, \mu(X_i;\lambda_0);\beta_0)\{\mu(X_i;\lambda) - \mu(X_i;\lambda_0)\}\right].$$

The Fréchet derivative of the expectation with respect to λ is zero under regularity conditions by Theorem 2. In fact, under these conditions, one can show that the expectation above is $O(||\lambda - \lambda_0||_\infty^2)$. Therefore, whenever $||\lambda - \lambda_0||_\infty = o(n^{-1/4})$. The last sum in (5) becomes asymptotically negligible. The rate condition for λ includes the cube-root rate $n^{-1/3}$.

3.3 Bootstrap Confidence Sets

The asymptotic distribution of $\hat{\beta}$ is complicated, and naturally, one might consider using a bootstrap method to construct confidence sets. The finding of Theorem 1 suggests that there may be a bootstrap method that is valid even when the single-index estimator $\hat{\theta}$ follows cube-root asymptotics. However, as far as the author is concerned, it is not clear how one can analyze the asymptotic refinements of a bootstrap method in this situation. Leaving this to a future research, this paper rather 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).

Suppose that $\hat{\mu}(X_i;\hat{\theta})$ is a first step estimator defined before and introduce the following quantities:

$$\begin{aligned} \hat{\rho}_{jk}(\beta) &= 1\{W_j \leq W_k\} \times \rho(V_j, \hat{\mu}(X_j;\hat{\theta});\beta) \text{ and} \\ \hat{\rho}_{\mu,ik} &= 1\{W_i \leq W_k\} \times \rho_\mu(V_i, \hat{\mu}(X_j;\hat{\theta});\hat{\beta}). \end{aligned}$$

Then, we construct the following symmetrized nearest neighborhood estimator:

$$\hat{r}_{jk} = \frac{\sum_{i=1}^n \hat{\rho}_{\mu,ik} \times K_h(\hat{U}_{n,i} - \hat{U}_{n,j})}{\sum_{i=1}^n K_h(\hat{U}_{n,i} - \hat{U}_{n,j})}.$$

The bootstrap procedure that this paper suggests is a wild bootstrap in the following form:

Step 1: For each $b = 1, \cdots, B$, draw i.i.d. random variables $\{\omega_{i,b}\}_{i=1}^n$ 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

$$\hat{\beta}_b^* = \underset{\beta \in B}{\operatorname{argmin}} \sum_{k=1}^{n} \left\{ \sum_{j=1}^{n} \left[\hat{\rho}_{jk}(\hat{\beta}) - \hat{\rho}_{jk}(\beta) + \omega_{j,b} \left\{ \hat{\rho}_{jk}(\hat{\beta}) + \hat{r}_{jk}^{\top} \times (Y_j - \hat{\mu}(X_j; \hat{\theta})) \right\} \right] \right\}^2$$

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

The bootstrap procedure is very simple. In particular, one does not need to estimate the nonparametric function μ nor the single-index coefficient θ_0 using the bootstrap sample. The estimator $\hat{\mu}(X_i; \hat{\theta})$ is stored once and repeatedly used for each bootstrap sample. This computational merit is prominent in particular when the dimension of the parameter θ_0 is large and one has to resort to a numerical optimization algorithm for its estimation. The bootstrap procedure has an additional term ${\hat{r}_{jk}}^{\top} \times (Y_j - \hat{\mu}(X_j; \hat{\theta}))$ as compared to typical wild bootstrap. This term is introduced to account for the first order effect of the estimation error in $\hat{\mu}$ upon the asymptotic distribution of the estimator. In the following, we establish the asymptotic validity of the bootstrap confidence sets.

Theorem 3: *Suppose that* Assumptions 1-4 *hold. Then, conditional on almost every sequence* $\{S_j\}_{j=1}^n$,

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

where $\dot{H}(w)$ *and* B *are as in* Theorem 1.

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

$$\begin{aligned} Z_i &\sim U_{1i} - \eta_{1i}/2 \text{ and} \\ X_i &\sim U_{2i} - \eta_i/2 \end{aligned}$$

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

DGP A1: $T_i \sim t$ distribution with degree of freedom 1.
DGP A2: $T_i \sim$ log-normal distribution with median zero.

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

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

where $v_i \sim \zeta_i + e_i$, 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 formula due to the least squares problem.

Table 1 shows the performance of the estimators. There are four combinations, according to whether it is assumed that θ_0 is known (TR) or unknown and estimated through maximum score estimation (ES) and according to whether a symmetrized nearest neighborhood estimation was used (NN) or usual kernel estimation was used (KN). For the latter case, we

used standard normal pdf as a kernel. The bandwidth choice was made using a least-squares cross-validation method. The current version used a crude 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	KN-TR	NN-ES	KN-ES
$n = 200$	DGP A1	3	MAE	0.3534	0.3535	0.3413	0.3673
			RMSE	0.2010	0.2005	0.1969	0.2266
		6	MAE	0.3709	0.3815	0.3654	0.3970
			RMSE	0.2050	0.2234	0.2108	0.2513
	DGP A2	3	MAE	0.3147	0.3134	0.3104	0.3424
			RMSE	0.1510	0.1520	0.1518	0.2000
		6	MAE	0.3018	0.3015	0.3030	0.3275
			RMSE	0.1433	0.1430	0.1462	0.1725
$n = 500$	DGP A1	3	MAE	0.2268	0.2249	0.2282	0.2545
			RMSE	0.0801	0.0791	0.0817	0.1056
		6	MAE	0.2143	0.2164	0.2200	0.2455
			RMSE	0.0728	0.0740	0.0765	0.0959
	DGP A2	3	MAE	0.1947	0.1913	0.1941	0.2354
			RMSE	0.0586	0.0580	0.0592	0.0863
		6	MAE	0.1816	0.1807	0.1827	0.2112
			RMSE	0.0516	0.0512	0.0521	0.0690

The results show that the performance of the estimators does not change significantly as we increase the number of covariates from 3 to 6. Some simulations unreported here showed a similar result when we increase the number of covariates to 9. This indirectly 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 that uses θ_0 and the estimator that uses $\hat{\theta}$. The performance does not show much difference between these two estimators. The performance of the estimator that does not use probability integral transform appears to perform slightly better than that does use probability integral transform. 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 appears conspicuous.

4.2 The Performance of the Bootstrap Procedure

In this subsection, we investigate the bootstrap procedure that this paper proposes. In the case of the bootstrap performance we consider the following drawing of T_i that constitutes the error term in the selection equation:

DGP B1: $T_i \sim N(0,1)$
DGP B2: $T_i \sim t$ distribution with degree of freedom 1.

Hence the selection mechanism has errors that are conditionally heteroskedastic and, in the case of DGP B2, are 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 $e_i \sim N(0,1)$. Hence the errors in the outcome equation are conditionally heteroskedastic. We set θ_0 to be the vector of 2's and $\beta_0 = 2$ as before.

To illustrate this paper's proposal, we explain the procedure of estimation and the bootstrap procedure in detail. First, define

$$r_{jk} = \hat{w}_{jk} - \frac{\sum_{i=1}^n \hat{w}_{ik} \times K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}{\sum_{i=1}^n K_h\left(\hat{U}_{n,i} - \hat{U}_{n,j}\right)}.$$

Then, for each draw of $\{\omega_{j,b}\}_{j=1}^n$, we estimate

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

Again, the bootstrap estimator is also a solution to the least squares problem, adding to the computational expediency. We construct a confidence interval using the empirical distribution of $\{\hat{\beta}_b^*\}_{b=1}^B$.

Table 2: The Performance of the Proposed Bootstrap Methods

		k	Nom. Cov. Prob.	NN-TR	KN-TR	NN-ES	KN-ES
			99%	0.9320	0.9460	0.9420	0.9500
		3	95%	0.8840	0.9100	0.8740	0.8960
	DGP B1		90%	0.8380	0.8520	0.8120	0.8520
			99%	0.9220	0.9520	0.9180	0.9680
		6	95%	0.8680	0.9180	0.8700	0.9140
$n = 100$			90%	0.8000	0.8600	0.8060	0.8680
			99%	0.9180	0.9400	0.9280	0.9660
		3	95%	0.8780	0.9000	0.8720	0.9220
	DGP B2		90%	0.8280	0.8620	0.8200	0.8780
			99%	0.9280	0.9420	0.9520	0.9780
		6	95%	0.8880	0.9120	0.8960	0.9500
			90%	0.8300	0.8720	0.8360	0.9020
			99%	0.9740	0.9780	0.9600	0.9780
		3	95%	0.9240	0.9260	0.9140	0.9320
	DGP B1		90%	0.8780	0.8600	0.8640	0.8780
			99%	0.9880	0.9880	0.9760	0.9840
		6	95%	0.9340	0.9340	0.9340	0.9480
$n = 300$			90%	0.8840	0.8880	0.8840	0.9000
			99%	0.9680	0.9760	0.9620	0.9780
		3	95%	0.9360	0.9300	0.9220	0.9280
	DGP B2		90%	0.8740	0.8820	0.8520	0.8840
			99%	0.9760	0.9800	0.9820	0.9820
		6	95%	0.9340	0.9360	0.9320	0.9360
			90%	0.8900	0.8940	0.8880	0.8920

Table 2 contains finite sample coverage probabilities for a variety of estimators. When the sample size was 100, the bootstrap coverage probability is smaller than the nominal ones. When the sample size was 300, 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 $\hat{\theta}$. In the similar way, the performance is also similar across different numbers of covariates 3 and 6. However, overall performance of the confidence set using the kernel estimator (KN)

appears to perform slightly better than the nearest neighborhood estimator (NN). Finally, the bootstrap performance does not make much difference with regard to the heavy tailedness of the error distribution in the selection equation.

5 Conclusion

This paper finds that the first step estimator of a single-index component of a nonparametric estimator does not affect the quality of the second step estimator in models of semiparametric conditional moment restrictions. In particular, the influence of the first step estimator converging even at the rate of $n^{-1/3}$ is shown to be asymptotically negligible. An heuristic analysis was performed in terms of Fréchet derivative 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 and computationally attractive. Therefore, while the usual bootstrap procedure is known to fail for $n^{1/3}$-converging estimators, we can still use bootstrap when such an estimator is a first step plug-in estimator in a larger model of conditional moment restrictions. The simulation studies reported in this paper are small scales. An extended simulation study is in progress now.

6 Appendix: Mathematical Proofs

6.1 The Proofs of the Main Results

Proof of Theorem 1: Write $\mu(x) = \mu(x;\theta)$, $\mu_0(x) = \mu(x;\theta_0)$ and $\hat{\mu}(x) = \hat{\mu}(x;\hat{\theta})$. Then define

$$\begin{aligned} Q_n(\beta,\mu) &= \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(V_i,\mu(X_i);\beta)1\{\hat{W}_i \le \hat{W}_j\}\right\}^2 \text{ and} \\ Q(\beta,\mu) &= \int \left\{\mathbf{E}\left[\rho(V_i,\mu(X_i);\beta)1\{W_i \le w\}\right]\right\}^2 dP_W(w). \end{aligned}$$

and define $q_n(t;\mu) = Q_n(\beta_0+t,\mu) - Q_n(\beta_0,\mu)$ and $q(t) = Q(\beta_0+t,\mu) - Q(\beta_0,\mu)$. Lastly, we also let

$$\begin{aligned} \xi_n(\mu) = \frac{2}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(V_i,\mu(X_i);\beta_0)1\{\hat{W}_i \le \hat{W}_j\}\right\} \times \frac{1}{n}\sum_{i=1}^{n}\rho_\beta(V_i,\mu(X_i);\beta_0)1\{\hat{W}_i \le \hat{W}_j\} \\ - \int 2\mathbf{E}\left[\rho(V_i,\mu(X_i);\beta_0)1\{W_i \le w\}\right]\mathbf{E}\left[\rho_\beta(V_i,\mu(X_i);\beta_0)1\{W_i \le w\}\right] dP_W(w). \end{aligned}$$

We will show the following three claims later.

Claim 1 : $q_n(t_n, \hat{\mu}) - q(t_n) - \xi_n(\hat{\mu})^\top t_n = o_P(||t_n||^2)$ for all $t_n \to 0$.
Claim 2 : $q(t_n) = t_n^\top \int \dot{H}(w)\dot{H}(w)^\top dP_W(w) t_n + o(||t_n||^2)$, for all $t_n \to 0$.
Claim 3 : $\sqrt{n}\xi_n(\hat{\mu}) = \sqrt{n}\tilde{\xi}_n + o_P(1)$, where $\tilde{\xi}_n = \int \dot{H}(w)\frac{1}{n}\sum_{i=1,j\neq i}^n \xi_i(w) dP_W(w)$.

Then, as in the proof of Theorem 3.2.16 of van dar Vaart and Wellner (1996), we can write $\sqrt{n}(\hat{\theta}-\theta_0) = \Omega^{-1}\sqrt{n}\tilde{\xi}_n + o_P(1)$. (This theorem is originated from Pollard (1985, 1991).) Observe that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i(\cdot) \Longrightarrow B.$$

This can be shown using the fact that indicator functions are VC classes. Hence the continuous mapping theorem gives the wanted result.

As for Claim 1, we first show that

$$q_n(t_n, \hat{\mu}) - \xi_n(\hat{\mu})^\top t_n = q_n(t_n, \mu) - \xi_n(\mu)^\top t_n + o_P(||t_n||^2) \tag{6}$$

By the Taylor expansion, $q_n(t_n, \mu) = \xi_n(\mu)^\top t_n + t_n^\top \xi_n^\Delta(\mu) t_n + o_P(||t_n||^2)$, where $\xi_n^\Delta(\mu)$ denotes the second order derivative of $q_n(t, \mu)$ in t at $t \in [0, t_n]$. Therefore,

$$\begin{aligned} & \{q_n(t_n, \hat{\mu}) - q_n(t_n, \mu_0)\} - (\xi_n(\hat{\mu}) - \xi_n(\mu_0))^\top t_n \\ = \;& t_n^\top \{\xi_n^\Delta(\hat{\mu}) - \xi_n^\Delta(\mu_0)\} t_n/2 + o_P(||t_n||^2) = o_P(||t_n||^2). \end{aligned}$$

The last equality follows by using the fact that $||\xi_n^\Delta(\hat{\mu}) - \xi_n^\Delta(\mu_0)|| = O_P(||\hat{\mu} - \mu_0||_\infty)$. This term is $o_P(1)$ because $||\hat{\mu} - \mu_0||_\infty = o_P(1)$ by Lemma A4 of Song (2008b). Hence we have established (6). Therefore, for Claim 1, it suffices to show that

$$q_n(t_n, \hat{\mu}) - q(t_n) - \xi_n(\hat{\mu})^\top t_n = o_P(||t_n||^2). \tag{7}$$

We expand $q_n(t_n, \hat{\mu}) - q(t_n)$ up to the second order term which then becomes $t_n^\top \xi_n^{\Delta\Delta} t_n/2$ where $\xi_n^{\Delta\Delta}$ is the second order derivative of $q_n(t, \hat{\mu}) - q(t)$ in t at $t = 0$, We can easily check that $\xi_n^{\Delta\Delta} = O_P(n^{-1/2})$. Since the first order derivative of $q_n(t, \hat{\mu}) - q(t)$ in t at $t = 0$ is equal to $\xi_n(\hat{\mu})$, the remainder term in the expansion in (7) is equal to $O_P(||t_n||^2/\sqrt{n}) + o(||t_n||^2) = o(||t_n||^2)$. Hence Claim 1 follows.

As for Claim 2, the expansion of $q(t)$ in t up to the second order delivers the wanted

result. It remains to show Claim 3. For this, first we write

$$\begin{aligned}\xi_n(\hat{\mu}) &= \xi_n(\mu_0) + \xi_n(\hat{\mu}) - \xi_n(\mu_0) \\ &= \xi_n(\mu_0) + \frac{2}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1,j\neq i}^{n} 1\{\hat{W}_i \leq \hat{W}_j\}\rho_{\mu,i}^{\top}(\hat{\mu}(X_i) - \mu_0(X_i))\right\} \\ &\quad \times \left\{\frac{1}{n}\sum_{k=1,k\neq i,k\neq j}^{n} 1\{\hat{W}_k \leq \hat{W}_j\}\rho_{\beta,k}\right\} + o_P(1),\end{aligned}$$

where $\rho_{\mu,i} = \rho_\mu(V_i, \mu_0(X_i); \beta_0)$ and $\rho_{\beta,i} = \rho_\beta(V_i, \mu_0(X_i); \beta_0)$. By applying the uniform Bahadur representation in Lemma A1 below,

$$\begin{aligned}\frac{1}{n}\sum_{i=1,j\neq i}^{n} 1\{\hat{W}_i &\leq \hat{W}_j\}\rho_{\mu,i}^{\top}(\hat{\mu}(X_i) - \mu_0(X_i)) \\ &= \frac{1}{n}\sum_{i=1,j\neq i}^{n} \mathbf{E}\left[1\{W_i \leq W_j\}\rho_{\mu,i}^{\top}|U_i\right](Y_i - \mu_0(X_i)) + o_P(n^{-1/2}).\end{aligned}$$

(Note that the bracketing entropy condition for the space of functions of the form $\psi_w(u) = 1\{F_{n,\theta,i}(\lambda(x;\theta)) \leq u\}$ can be established using Lemma A1 of Song (2008b).) Using the fact that the last sum is $O_P(n^{-1/2})$, we obtain that

$$\begin{aligned}\xi_n(\hat{\mu}) &= \xi_n(\mu_0) + \int \mathbf{E}\left[\rho_{\beta,k}1\{W_i \leq w\}\right] \\ &\quad \times \sum_{i=1,j\neq i}^{n} \mathbf{E}\left[1\{W_i \leq w\}\rho_{\mu,i}^{\top}|U_i\right](Y_i - \mu_0(X_i))dP_W(w) + o_P(n^{-1/2}).\end{aligned}$$

Now we turn to $\xi_n(\mu_0)$, which we write as

$$\xi_n(\mu_0) = \frac{2}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(V_i, \mu_0(X_i); \beta_0)1\{\hat{W}_i \leq \hat{W}_j\}\right\} \times \frac{1}{n}\sum_{i=1}^{n}\rho_{\beta,i}1\{\hat{W}_i \leq \hat{W}_j\},$$

because $\mathbf{E}\left[\rho(V_i, \mu_0(X_i); \beta_0)1\{W_i \leq w\}\right] = 0$. Since $\frac{1}{n}\sum_{i=1}^{n}\rho(V_i, \mu_0(X_i); \beta_0)1\{W_i \leq W_j\}$ is the sample mean of mean-zero random variables,

$$\begin{aligned}\xi_n(\mu_0) &= \frac{2}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(V_i, \mu_0(X_i); \beta_0)1\{W_i \leq W_j\}\right\} \times \dot{H}(W_j) + o_P(n^{-1/2}) \\ &= \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\rho(V_i, \mu_0(X_i); \beta_0)1\{W_i \leq W_j\} \times \dot{H}(W_j) + o_P(n^{-1/2}).\end{aligned}$$

Note that $\mathbf{E}[\rho(V_i, \mu_0(X_i); \beta_0) 1\{W_i \leq W_j\} \times \dot{H}(W_j)|W_j] = 0$. Therefore, the Hoeffding's decomposition renders the leading term above as

$$\int \frac{2}{n} \sum_{i=1}^{n} \rho(V_i, \mu_0(X_i); \beta_0) 1\{W_i \leq w\} \dot{H}(w) dP_W(w)$$

plus a degenerate U-process. This degenerate U-process can be shown to be $o_P(n^{-1/2})$ using the maximal inequality in Sherman (1994) or Turki-Moalla (1998). Hence Claim 3 is obtained. ■

Proof of Theorem 2: Simply write $\mu_\lambda(x) = \mu(x; \lambda)$ and $\mu_0(x) = \mu(x; \lambda_0)$. First write

$$\begin{aligned}
& \mathbf{E}\left[\xi \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] = \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda(X_i), \lambda_0(X_i)\right] \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] \\
= & \mathbf{E}\left[\left(\mathbf{E}\left[\xi|\lambda(X_i), \lambda_0(X_i)\right] - \mathbf{E}\left[\xi|\lambda_0(X_i)\right]\right) \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] \\
& + \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] \\
= & \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] + O(||\lambda - \lambda_0||_\infty^2)
\end{aligned}$$

by applying Lemma A2(ii) of Song (2008a). The last expectation is equal to

$$\begin{aligned}
& \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] \\
= & \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mathbf{E}\left[Y_i|\lambda(X_i)\right] - \mathbf{E}\left[Y_i|\lambda(X_i), \lambda_0(X_i)\right]\right\}\right] \\
& + \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mathbf{E}\left[Y_i|\lambda(X_i), \lambda_0(X_i)\right] - \mathbf{E}\left[Y_i|\lambda_0(X_i)\right]\right\}\right] \\
= & \mathbf{E}\left[\mathbf{E}\left[\xi|\lambda_0(X_i)\right] \left\{\mathbf{E}\left[Y_i|\lambda(X_i)\right] - \mathbf{E}\left[Y_i|\lambda(X_i), \lambda_0(X_i)\right]\right\}\right] \\
= & \mathbf{E}\left[\left\{\mathbf{E}\left[\xi|\lambda_0(X_i)\right] - \mathbf{E}\left[\xi|\lambda(X_i)\right]\right\} \left\{\mathbf{E}\left[Y_i|\lambda(X_i)\right] - \mathbf{E}\left[Y_i|\lambda(X_i), \lambda_0(X_i)\right]\right\}\right].
\end{aligned}$$

The last equality follows because $\mathbf{E}\left[\mathbf{E}\left[\xi|\lambda(X_i)\right] \left\{\mathbf{E}\left[Y_i|\lambda(X_i)\right] - \mathbf{E}\left[Y_i|\lambda(X_i), \lambda_0(X_i)\right]\right\}\right] = 0$. Applying Lemma A2(ii) of Song (2008a) again, the last expectation is equal to $O(||\lambda - \lambda_0||_\infty^2)$. Hence we conclude that

$$\mathbf{E}\left[\xi \left\{\mu_\lambda(X_i) - \mu_0(X_i)\right\}\right] = O(||\lambda - \lambda_0||_\infty^2),$$

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

Proof of Theorem 3: Let

$$Q_{n,b}(\beta) = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\rho}_{jk}(\hat{\beta}) - \hat{\rho}_{jk}(\beta) + \omega_{j,b}\left\{\hat{\rho}_{jk}(\hat{\beta}) + \hat{r}_{jk}\times\{Y_j - \hat{\mu}(X_j;\hat{\theta})\}\right\}\right]\right\}^2 \text{ and}$$

$$Q_{n}(\beta) = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\rho}_{jk}(\hat{\beta}) - \hat{\rho}_{jk}(\beta)\right]\right\}^2$$

and define $q_{n,b}(t) = Q_{n,b}(\hat{\beta}+t) - Q_{n,b}(\hat{\beta})$ and $q_n(t) = Q_n(\hat{\beta}+t) - Q_n(\hat{\beta})$. Let

$$\xi_n = \frac{2}{n}\sum_{k=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\omega_{j,b}\left[\hat{\rho}_{jk}(\hat{\beta}) + \hat{r}_{jk}\times\{Y_j - \hat{\mu}(X_j;\hat{\theta})\}\right]\right\}\times\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\beta,jk}(\hat{\beta}).$$

Similarly as in the proof of Theorem 1, we can show the following:

Claim 1 **:** $q_n(t_n) - q(t_n) - {\xi_n}^{\top}t_n = o_P(||t_n||)$ for all $t_n \to 0$.
Claim 2 **:** $q(t_n) = t_n^{\top}\int \dot{H}(w)\dot{H}(w)^{\top}dP_W(w)t_n + o(||t_n||^2)$, for all $t_n \to 0$.

Then, as in the proof of Theorem 3.2.16 of van dar Vaart and Wellner (1996) again, we can write $\sqrt{n}(\hat{\theta}_b^* - \hat{\theta}) = \Omega^{-1}\sqrt{n}\xi_n + o_P^*(1)$. Therefore, it remains to analyze the sum ξ_n. Observe that by using the usual arguments of stochastic equicontinuity,

$$\begin{aligned}\sqrt{n}\xi_n &= \frac{2}{n}\sum_{k=1}^{n}\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\omega_{j,b}\left[\hat{\rho}_{jk}(\hat{\beta}) + \hat{r}_{jk}\times\{Y_j - \hat{\mu}(X_j;\hat{\theta})\}\right]\right\}\times\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\beta,jk}(\hat{\beta}) \\ &= \frac{2}{n}\sum_{k=1}^{n}\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\omega_{j,b}\left[\rho_j(W_k;\beta_0) + r_j(W_k)\times\{Y_j - \mu(X_j;\theta_0)\}\right]\right\}\times\dot{H}(W_k) + o_P(1),\end{aligned}$$

where $\rho_j(w;\beta_0) = 1\{W_j \le w\}\rho(V_j,\mu_0(X_j);\beta_0)$ and $r_j(w) = \mathbf{E}[\rho_{\mu,j}1\{W_j \le w\}|U_j]$. (This is possible because $\omega_{j,b}$ is mean zero, bounded and independent of other random components. The nonparametric estimator $\hat{\mu}$ can be handled by using Lemma A1 below.) 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

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

by the strong law of large numbers. Let

$$F_n(w; \{S_j, W_{1j}\}_{j=1}^n) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \omega_{j,b} \left[\rho_j(w; \beta_0) + r_j(w) \times \{Y_j - \mu(X_j; \theta_0)\}\right] \times \dot{H}(w).$$

Now, by the conditional multiplier central limit theorem of Ledoux and Talagrand (1988), conditional on almost every sequence $\{S_j, W_{1j}\}_{j=1}^\infty$,

$$F_n(\cdot; \{S_j, W_{1j}\}_{j=1}^n) \Longrightarrow B.$$

Therefore, by the almost sure representation theorem (e.g. Theorem 6.7 of Billingsley), there is a sequence $\tilde{F}_n(\cdot)$ such that $\tilde{F}_n(\cdot)$ is distributionally equivalent to $F_n(\cdot)$ and $\tilde{F}_n(\cdot) \to_{a.s.} B$. Then, by the previous arguments, condiitonal on almost every sequence $\{S_j, W_{1j}\}_{j=1}^n$, we have

$$\Gamma_n(\tilde{F}_n(\cdot; \{S_j, W_{1j}\}_{j=1}^n) \to_d \int B(w)\dot{H}(w)P_W(w),$$

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

6.2 Uniform Bahadur Representation of Symmetrized Nearest Neighborhood Estimators

In this section, we present a general asymptotic representation of sums of symmetrized nearest neighborhood estimators that is uniform over certain function spaces. Stute and Zhu (2005) obtained a non-unifom result in a different format. 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 in the proof. Suppose that we are given a random sample $\{(W_i, X_i, Y_i)\}_{i=1}^n$ drawn from the distribution of a random vector $S = (W, X, Y) \in \mathbf{R}^{d_W + d_X + 1}$. Let Λ_n and $\Lambda_0 \subset \Lambda_n$ be classes of real functions on $\mathbf{R}^{d_X}$ with generic elements denoted respectively by λ and λ_0. We also let Φ and Ψ be classes of functions on $\mathbf{R}^{d_Y}$ and $\mathbf{R}^{d_W}$with generic elements φ and ψ. The nonparametric regression function of interest in this situation is $\mathbf{E}[\varphi(Y)|\lambda_0(X)]$. When $\lambda_0(X)$ is continuous, we focus instead on $g_\varphi(u) = \mathbf{E}[\varphi(Y)|U = u]$, where $U = F_{\lambda_0}(\lambda_0(X))$ and $F_{\lambda_0}(\cdot)$ is the distribution function of $\lambda_0(X)$. Similarly, we define $g_\psi(u) = \mathbf{E}[\psi(W)|U = u]$.

Let F_λ be the distribution function of $\lambda(X_i)$ and

$$F_{n,\lambda,i}(\cdot) = (n-1)^{-1}\Sigma_{j=1,j\neq i}^n 1\{\lambda(X_j) \leq \cdot\},$$

For brevity, we write $U_{n,\lambda,i} = F_{n,\lambda,i}(\lambda(X_i))$. Then we define

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

and $\hat{f}_{\lambda,i}(u) = (n-1)^{-1} \sum_{j=1,j\neq i}^{n} K_h(U_{n,\lambda,j} - u)$. We also let .

For each $\lambda_0 \in \Lambda_n$, we define $\Lambda_n(\lambda_0) = \{\lambda \in \Lambda : ||\lambda - \lambda_0||_\infty \leq n^{-b}\}$ for $b \in (1/4, 1/2]$. Hence $\Lambda_n(\lambda_0)$ can be regarded as a shrinking neighborhood of a nonparametric or parametric function λ_0. The semiparametric process of focus takes the following form:

$$\tilde{\nu}_n(\lambda, \varphi, \psi) = \frac{1}{n} \sum_{i=1}^{n} \psi(W_i) \left\{\hat{g}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_\varphi(U_i)\right\},$$

with $(\lambda, \varphi, \psi) \in \Lambda_n(\lambda_0) \times \Phi_n \times \Psi_n$. We introduce the following assumptions:

Assumption P1 : (i) Classes Φ and Ψ for some $C > 0,\ p > 4$, and $b_\Psi,\ b_\Phi \in (0, 2)$,

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

and envelopes $\tilde{\varphi}$ and $\tilde{\psi}$ satisfy that $\mathbf{E}[|\tilde{\varphi}(Y)|^p|X] < \infty$ and $\mathbf{E}[|\tilde{\psi}(W)|^p|X] < \infty$, a.s.
(ii) There exist a strictly increasing, bounded map G, $b_\Lambda \in (0, 1)$, and $C > 0$ such that the class $G \circ \Lambda_n = \{G \circ \lambda : \lambda \in \Lambda_n\}$ satisfies the following:

$$\log N_{[]}(\varepsilon, G \circ \Lambda_n, ||\cdot||_\infty) \leq C\varepsilon^{-b_\Lambda}, \text{ for each } \varepsilon > 0.$$

Assumption P2 : (i) For each $\lambda \in \Lambda_n$, the variable $\lambda(X)$ is continuous.
(ii) There exists $C > 0$ such that

$$|F_\lambda(\lambda_1) - F_\lambda(\lambda_2)| \leq C|\lambda_1 - \lambda_2|, \text{ for all } \lambda_1, \lambda_2 \in \mathbf{R}.$$

(iii) (a) The conditional density $f(s|u)$ of S given $U = u$ with respect to a σ-finite measure is L times continuously differentiable in $u,\ L > 8$, with bounded derivatives $f^{(j)}(s|u)$ such that $\sup_{(s,u)\in\mathcal{S}\times[0,1]}|f^{(j)}(s|u)/f(s)| < C,\ j = 1, \cdots, L$, where $f(s)$ is the density of S with respect to a σ-finite measure and $\mathcal{S}$ is the support of S.
(b) For each $\lambda \in \Lambda_n$, the conditional density $f_\lambda(s|u)$ of S given $(U_\lambda, U) = (u_1, u_2)$ with respect to a σ-finite measure satisfies that for all $\nu > 0$

$$\sup_{(s,u_2)\in\mathcal{S}\times[0,1]}|f_\lambda(s|u_1 - \nu, u_2) - f_\lambda(s|u_1 + \nu, u_2)| \leq C\eta_\lambda(s),$$

where $\eta_\lambda : \mathcal{S} \to \mathbf{R}_+$ is such that $\sup_{(s,\lambda)\in\mathcal{S}\times\Lambda_n}|\eta_\lambda(s)/f(s)| < C$.
(c) $\sup_{\varphi\in\Phi}||g_\varphi||_\infty < \infty$ and $g_\varphi(\cdot)$ is twice continuously differentiable with bounded derivatives.

Assumption P3 : (i) $K(\cdot)$ is bounded above, symmetric, compact supported, infinite times differentiable with bounded derivatives, and $\int K(t)dt = 1$.
(ii) $n^{1/2}h^4 + n^{1/2-(L+1)b}h^{-(L+2)} \to 0$.

The following theorem offers the uniform Bahadur representation of $\hat{g}_{\varphi,\lambda,i}$.

Lemma A1 : *Suppose that* Assumptions P1-P3 *hold. Then,*

$$\sup_{(\lambda,\varphi,\psi)\in\Lambda_n(\lambda_0)\times\Phi\times\Psi}\left|\sqrt{n}\tilde{\nu}_n(\lambda,\varphi,\psi) - \tilde{\zeta}_n(\varphi,\psi)\right| = o_P(1),$$

where $\tilde{\zeta}_n(\varphi,\psi) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g_\psi(U_i)\{\varphi(Y_i) - g_\varphi(U_i)\}$.

It is worth noting that the representation holds regardless of whether the estimator $\hat{\lambda}$ has a parametric rate of $n^{-1/2}$ or cube-root rate $n^{-1/3}$ as in the maximum score estimator, or a nonparametric rate n^{-b} with $b \in (1/4, 1/2]$.

Proof of Lemma A1 : Without loss of generality, assume that the support of K is contained in $[-1,1]$. Also observe that the difference $\hat{g}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_\varphi(U_i)$ remains the same when we replace $\lambda \in \Lambda$ by $G \circ \lambda$. Hence we write Λ_n for $G \circ \Lambda_n$ for simplicity. Throughout the proofs, the notation $\mathbf{E}_{Z_i}$ indicates the conditional expectation given Z_i. Define $\hat{\rho}_{\varphi,\lambda,i}(t) = (n-1)^{-1}\sum_{j=1,j\neq i}^n K_h(U_{n,\lambda,j} - t)\varphi(Y_j)$. Then we write $\hat{g}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_{\varphi,\lambda}(U_{n,\lambda,i})$ as

$$\begin{aligned} R_{1i}(\lambda,\varphi) &= \frac{\hat{\rho}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_\varphi(U_i)\hat{f}_{\lambda,i}(U_{n,\lambda,i})}{f_{\lambda_0}(U_i)} \\ &\quad + \frac{[\hat{\rho}_{\varphi,\lambda,i}(U_{n,\lambda,i}) - g_\varphi(U_i)\hat{f}_{\lambda,i}(U_{n,\lambda,i})](f_{\lambda_0}(U_i) - \hat{f}_{\lambda,i}(U_{n,\lambda,i}))}{\hat{f}_{\lambda,i}(U_{n,\lambda,i})f_{\lambda_0}(U_i)} \\ &= R_{1i}^A(\lambda,\varphi) + R_{1i}^B(\lambda,\varphi), \text{ say,} \end{aligned}$$

where $f_{\lambda_0}(u) = 1\{u \in [0,1]\}$. We simply put $\pi = (\lambda,\varphi,\psi)$ and $\Pi_n = \Lambda_n(\lambda_0)\times\Phi\times\Psi$, and write

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

Note that as for the second term,

$$r_{1n}^B(\pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(W_i)\{\hat{g}_{\varphi,\lambda}(U_{n,\lambda,i}) - g_\varphi(U_i)\}(f_{\lambda_0}(U_i) - \hat{f}_{\lambda,i}(U_{n,\lambda,i})) = O_P(\sqrt{n}w_n^2),$$

where $w_n = n^{-(2p+2)/(4p-3)}h^{-1} + h^2$, by Lemma A4 of Song (2008b). Since $p > 4$,

$$n^{(1/2)-(4p+4)/(4p-3)}h^{-2} + n^{1/2}h^4 = o(n^{-1}h^{-2} + n^{1/2}h^4) = o(1)$$

by Assumption P3. Hence it suffices to show that

$$\sup_{\pi \in \Pi_n} |r_{1n}^A(\pi) - \xi_n(\varphi, \psi)| = o_P(1).$$

We write $r_{1n}^A(\pi)$ as

$$\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \psi_i \Delta_{\varphi,ij} K_{h,ij} + \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \psi_i \Delta_{\varphi,ij} \left\{K_{h,ij}^\lambda - K_{h,ij}\right\}, \qquad (8)$$

where $\psi_i = \psi(W_i)$, $\Delta_{\varphi,ij} = \varphi(Y_j) - g_\varphi(U_i)$, $K_{h,ij}^\lambda = K_h(U_{n,\lambda,j} - U_{n,\lambda,i})$ and $K_{h,ij} = K_h(U_j - U_i)$. We consider the second sum first, which we write as

$$\begin{aligned}
&\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \psi_i \Delta_{\varphi,ij} K'_{h,ij} \left\{U_{n,\lambda,j} - U_{n,\lambda,i} - (U_j - U_i)\right\} \\
&+ \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{l=2}^L \frac{1}{l!} \psi_i \Delta_{\varphi,ij} K_{h,ij}^{(l)} \left\{U_{n,\lambda,j} - U_{n,\lambda,i} - (U_j - U_i)\right\}^l \\
&+ \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \frac{1}{(L+1)!} \psi_i \Delta_{\varphi,ij} K_{h,ij}^{(L+1)} \left\{U_{n,\lambda,j} - U_{n,\lambda,i} - (U_j - U_i)\right\}^{L+1} \\
= \;& A_{1n} + A_{2n} + A_{3n}, \text{ say,}
\end{aligned}$$

where $K_{h,ij}^{(l)} = h^{-(l+1)} \partial K(t)/\partial t|_{t=(U_i - U_j)/h}$ and

$$K_{h,ij}^{(L+1)} = h^{-(L+2)} \partial^{L+1} K(t)/\partial t^{L+1}|_{t=\{(1-a_{ij})(U_i - U_j) + a_{ij}(U_{n,\lambda,i} - U_{n,\lambda,j})\}/h},$$

for some $a_{ij} \in [0, 1]$. First, note that

$$\sup_{\lambda \in \Lambda_n} \sup_{x \in \mathbf{R}^{d_X}} |F_{n,\lambda,i}(\lambda(x)) - F_{\lambda_0}(\lambda_0(x))| = O_P(n^{-b}). \qquad (9)$$

We can show this following the proof of Lemma A3 of Song (2008b). It is easy to show that

$$A_{3n} = O_P(n^{1/2-(L+1)b}h^{-(L+2)}) = o_P(1).$$

In Lemma A2 below, it is shown that $A_{2n} = o_P(1)$. We turn to A_{1n}. Recall the notation $\delta_i^\lambda = U_{n,\lambda,i} - U_i$ and write

$$A_{1n} = \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \psi_i \Delta_{\varphi,ij} K'_{h,ij} \delta_j^\lambda - \frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \psi_i \Delta_{\varphi,ij} K'_{h,ij} \delta_i^\lambda. \quad (10)$$

We consider the leading term which we write as (up to $O(n^{-1})$)

$$\frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\psi_i\Delta_{\varphi,ij}K'_{h,ij} - \mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|W_j\right]\right\}\right]\delta_j^\lambda + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|W_j\right]\delta_j^\lambda. \quad (11)$$

We can easily show that the normalized sum in the bracket is $O_P(1)$ uniformly over $j = 1, \cdots, n$ and over $(\psi, \varphi) \in \Psi \times \Phi_n$, by using the maximal inequality and the bracketing entropy conditions in Assumption P1. Hence the first term is $o_P(1)$ by the fact that $\max_{1\leq j\leq n} ||\delta_j^\lambda|| = O_P(n^{-b}) = o_P(1)$ from (9). We deduce a similar result for the last term in (10) so that we write

$$\begin{aligned} A_{1n} &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_j\right](U_{n,j} - U_j) \quad (12)\\ &\quad -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_i\right](U_{n,i} - U_i)\\ &\quad +\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|W_j\right](U_{n,\lambda,j} - U_{n,j})\\ &\quad -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|W_i\right](U_{n,\lambda,i} - U_{n,i}). \end{aligned}$$

We show that the first two sums cancel out asymptotically. As for the first term, observe that

$$\begin{aligned} \mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_j = u_1\right] &= \frac{1}{h^2}\int_0^1 K'\left(\frac{u-u_1}{h}\right) g_\psi(u)\{g_\varphi(u_1) - g_\varphi(u)\}du\\ &= \frac{1}{h}\int_{(-u_1/h)\vee 1}^{((1-u_1)/h)\wedge 1} K'(u)\, g_\psi(u_1+uh)\{g_\varphi(u_1) - g_\varphi(u_1+uh)\}du \end{aligned}$$

and similarly,

$$\begin{aligned}\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_i=u_1\right] &= \frac{1}{h^2}\int_0^1 K'\left(\frac{u_1-u}{h}\right)g_\psi(u_1)\{g_\varphi(u)-g_\varphi(u_1)\}du \qquad (13)\\ &= \frac{1}{h}\int_{(-u_1/h)\vee 1}^{((1-u_1)/h)\wedge 1} K'(u)\,g_\psi(u_1)\{g_\varphi(u_1)-g_\varphi(u_1+uh)\}du\end{aligned}$$

Therefore, by using Hoeffding's decomposition and taking care of the degenerate U-process,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^n \mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_j\right](U_{n,j}-U_j)\\ =\;& \frac{1}{\sqrt{n}}\sum_{k=1}^n\int_0^1\int_0^1 K'(u)\,g_\psi(u_1)\frac{g_\varphi(u_1)-g_\varphi(u_1+uh)}{h}du\,(1\{U_k\le u_1\}-u_1)\,du_1+O_P(h)\\ =\;& \frac{1}{\sqrt{n}}\sum_{k=1}^n\int_0^1 g_\psi(u_1)g'_\varphi(u_1)\,(1\{U_k\le u_1\}-u_1)\,dudu_1+O_P(h).\end{aligned}$$

The second to the last equality follows because the inner integral is not zero only when $u_1>1-h$ or $u_1<h$ and the Lebesgue measure of this set is $O(h)$. The last equality follows because $\int_{-1}^1 K'(u)\,du=-1$. Similarly, using (13), we deduce that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^n \mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_i\right](U_{n,i}-U_i)\\ =\;& \frac{1}{\sqrt{n}}\sum_{k=1}^n\int_0^1 g_\psi(u_1)g'_\varphi(u_1)\,(1\{U_k\le u_1\}-u_1)\,dudu_1+o_P(1).\end{aligned}$$

Therefore, the first two sums in (12) cancel out. We focus on the last two sums. We show that the second to the last sum is $o_P(1)$. The last sum can be shown to be $o_P(1)$ in a similar manner. Applying Lemma UA of Escanciano and Song (2008), we can write the second to the last sum in (12) as

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^n \mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|W_j\right](U_{n,\lambda,j}-U_{n,j})\\ =\;& \sqrt{n}\mathbf{E}\left[\left\{\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_{\lambda,j},U_j\right]-\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K'_{h,ij}|U_j\right]\right\}(U_{\lambda,j}-U_j)\right]+O_P(n^{1/2-2b}).\end{aligned}$$

By applying Lemma A2(ii) of Song (2008a) combined with Assumption P2(iii) and by using Assumption P2(ii), we find that the leading expectation above is $O(n^{-2b})$. Hence we obtain that the above terms are $O_P(n^{1/2-2b})=o_P(1)$. We conclude that $A_{1n}=o_P(1)$. Therefore the second sum in (8) is $o_P(1)$.

We turn to the first sum in (8). We define $q^{\pi}_{n,ij} \equiv q^{\pi}_{n}(Z_i, Z_j) \equiv \psi_i \Delta_{\varphi,ij} K_{h,ij}$ and write the sum as

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

Similarly as before, let $\rho^{\pi}_{n,ij} \equiv \rho^{\pi}_{n}(Z_i, Z_j) \equiv q^{\pi}_{n,ij} - \mathbf{E}_{Z_i}[q^{\pi}_{n,ij}] - \mathbf{E}_{Z_j}[q^{\pi}_{n,ij}] + \mathbf{E}[q^{\pi}_{n,ij}]$ and define

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

Now write $\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} q^{\pi}_{n,ij} - \sqrt{n}\mathbf{E}[q^{\pi}_{n,ij}]$ as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \mathbf{E}_{Z_i}[q^{\pi}_{n,ij}] - \mathbf{E}_{Z_j}[q^{\pi}_{n,ij}] - 2\mathbf{E}[q^{\pi}_{n,ij}] \right\} + u_{2n}(\pi). \tag{14}$$

We will later show that $\sup_{\pi} |u_n(\pi)| = o_P(1)$.

First we note that through some tedious computations,

$$\mathbf{E}\left[\sup_{\pi \in \Pi_n} \left| \mathbf{E}_{Z_i}[q^{\pi}_{n,ij}] \right|^2 \right] \leq \int_0^1 \sup_{\varphi \in \Phi} g^2_{\tilde{\psi}}(t_1) \left[\int_0^1 \{g_{\varphi}(t_2) - g_{\varphi}(t_1)\} K_h(t_2 - t_1) dt_2 \right]^2 dt_1. \tag{15}$$

By change of variables, the integral inside the bracket becomes

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

When $h \leq t_1 \leq 1-h$, the integrand is of $O(h^2)$ because $\int_{-1}^{1} t_2 K(t_2) dt_2 = 0$. When $h > t_2$ or $t_2 > 1-h$, the integrand is of the order $O(h)$. Since the Lebesgue measure for this set of t_2's is $O(h)$, the integral above is equal to $O(h^2)$. Hence the expectation on the left-hand side in (15) is $O(h^4)$. Using this result, take

$$\tilde{\mathcal{J}}_n = \{\mathbf{E}[q^{\pi}_{n,ij} | Z_i = \cdot] : \pi \in \Pi_n\}$$

with an envelope J such that $||J||_2 \leq Ch^2$. Using the maximal inequality and the bracketing

entropy condition in (20) below,

$$\begin{aligned} & \mathbf{E}\left[\sup_{\pi\in\Pi_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbf{E}_{Z_i}[q^{\pi}_{n,ij}]-\mathbf{E}[q^{\pi}_{n,ij}]\right\}\right|\right] \qquad (16)\\ \leq\ & \int_0^{Ch^2}\sqrt{1+\log N_{[]}(\varepsilon,\tilde{\mathcal{J}}_n,||\cdot||_2)}d\varepsilon \\ =\ & O(h^{1-(b_\Phi\wedge b_\Psi)/2})=o(1), \end{aligned}$$

because $b_\Phi \wedge b_\Psi < 2$. We conclude from (14) that

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

Now, we consider the following:

$$\begin{aligned} & \mathbf{E}\left[\sup_{\pi\in\Pi_n}\left\{\mathbf{E}_{Z_j}[q^{\pi}_{n,ij}]-g_\psi(U_j)\{\varphi(Y_j)-g_\varphi(U_j)\}\right]^2\right] \qquad (17)\\ =\ & \int\sup_{\pi\in\Pi_n}\left\{\int_0^1 A_{n,\psi}(t_1,t_2,w)dt_1\right\}^2 dF_{\lambda_0}(w,t_2), \end{aligned}$$

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

$$A_{n,\psi}(t_1,t_2,w)=g_\psi(t_1)\{\varphi(w)-g_\varphi(t_1)\}K_h(t_1-t_2)-g_\psi(t_2)\{\varphi(w)-g_\varphi(t_2)\}.$$

We consider the term in (17). From similar arguments after (15), this term becomes $O(h^4)$ and we conclude that

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\mathbf{E}_{Z_j}[q^{\pi}_{n,ij}]-g_\psi(U_j)\{\varphi(Y_j)-g_\varphi(U_j)\}\right)=o_P(1). \qquad (18)$$

The arguments here, as those after (15) involves the bracketing entropy bound for the class $\{q^{\pi}_n(\cdot,\cdot)-g_\psi(\cdot)(\varphi(\cdot)-g_\varphi(\cdot):\pi\in\Pi_n\}$ and its vanishing sequence of envelopes. The procedures are very similar to those used before, and hence we omit the details. Combined with (16), the asymptotic representation in (18) yields the wanted result of Claim 1.

Now, it remains to deal with the denerate U-process u_n and show that $\sup_{\pi\in\Pi_n}|u_n(\pi)| = o_P(1)$. For this, let us define

$$\mathcal{J}_n=\{q^{\pi}_n(\cdot,\cdot):\pi\in\Pi_n\} \qquad (19)$$

and write $q_n^\pi(z_1, z_2; \pi) = \psi(s_1)\{\varphi(w_2) - g_\varphi(u_1)\} \times h^{-2}(\partial/\partial t)K_h(t)|_{t=(u_1-u_2)/h}$. Hence

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

Therefore,

$$\log N_{[]}(\varepsilon, \mathcal{J}_n, ||\cdot||_p) \leq C\varepsilon^{-(b_\Phi \vee b_\Psi)}. \tag{20}$$

We take arbitrary $\varepsilon > 0$ such that $(b_\Phi \vee b_\Psi)(1/2 + \varepsilon) < 1$. Then, note that

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

By Theorem 1 of Turki-Moalla (1998), p.878, for any small $\Delta > 0$,

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

Hence the proof is complete.

Lemma A2: *Under the assumptions of* Lemma A1, $A_{3n} = o_P(1)$ *uniformly in* $(\psi, \varphi, \lambda) \in \Psi \times \Phi \times \Lambda_n$.

Proof of Lemma A2: Define

$$\mathcal{T}_n = \left\{ \frac{1}{n-1} \sum_{j=2}^n 1\{\lambda(x_j) \leq \lambda(\cdot)\} : (\lambda, \{x_j\}) \in \Lambda_n \times \mathbf{R}^{(n-1)d_X} \right\}. \tag{21}$$

Then, the bracketing entropy bound for $\mathcal{T}_n$ can be computed from Lemma A1 of Song (2008b):

$$\log N_{[]}(\varepsilon, \mathcal{T}_n, ||\cdot||_p) \leq \log N_{[]}(\varepsilon^p, \Lambda_n, ||\cdot||_\infty) + C/\varepsilon \leq C\varepsilon^{-pb_\Lambda} + C/\varepsilon.$$

Let $\tau_0(x) = F_0(\lambda_0(x))$. It suffices to show that the sum below is $o_P(1)$ uniformly over $(\psi, \varphi, \tau) \in \Psi \times \Phi \times \mathcal{T}_n$:

$$\frac{1}{(n-1)\sqrt{n}} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \psi_i \Delta_{\varphi,ij} K_{h,ij}^{(l)} \delta_{\tau,ij}^l,$$

where $\delta_{\tau,j} = \tau(X_j) - \tau_0(X_j)$, $\delta_{\tau,ij} = \delta_{\tau,j} - \delta_{\tau,i}$. We write the sum as

$$\frac{1}{n} \sum_{j=1}^n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \psi_i \Delta_{\varphi,ij} K_{h,ij}^{(l)} \delta_{\tau,i}^l - \mathbf{E}\left[\psi_i \Delta_{\varphi,ij} K_{h,ij}^{(l)} \delta_{\tau,i}^l\right] \right\} \right] + \sqrt{n}\mathbf{E}\left[\psi_i \Delta_{\varphi,ij} K_{h,ij}^{(l)} \delta_{\tau,i}^l\right]. \tag{22}$$

Write the last term as $\sqrt{n}\mathbf{E}\left[G_{\psi,\varphi,\tau}(U_i,U_j)K_{h,ij}^{(l)}\right]$, where $G_{\psi,\varphi,\tau}(U_i,U_j)=\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}\delta_{\tau,i}^l|U_i,U_j\right]$. By change of variables and integration by parts, the last term is written as

$$\begin{aligned}\frac{\sqrt{n}}{h^l}\int\int G_{\psi,\varphi,\tau}(u_2+hu_1,u_2)K^{(l)}(u_1)\,du_1du_2 &= \sqrt{n}\int\int G_{\psi,\varphi,\tau}^{(l)}(u_2+hu_1,u_2)K(u_1)\,du_1du_2\\ &= O_P(\sqrt{n}n^{-2b})=o_P(1),\end{aligned}$$

because $l\geq 2$. We turn to the leading term in (22). Note that

$$\psi_i\Delta_{\varphi,ij}K_{h,ij}^{(l)}\delta_{\tau,i}^l\leq Cn^{-bl}\tilde{\psi}(W_i)\{\tilde{\varphi}(Y_j)+g_{\tilde{\varphi}}(U_i)\}|K_{h,ij}^{(l)}|.$$

Furthermore, using Hoeffding's inequality, we are left with the terms such as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K_{h,ij}^{(l)}\delta_{\tau,i}^l|U_i=u\right]-\mathbf{E}\left[\psi_i\Delta_{\varphi,ij}K_{h,ij}^{(l)}\delta_{\tau,i}^l\right]\right\}.$$

The envelope of the class of functions that index the above sum has the L_2-norm as follows:

$$\begin{aligned}&\sqrt{\int_{-1}^1\left\{\mathbf{E}\left[\tilde{\psi}(W_i)\{\tilde{\varphi}(Y_j)+g_{\tilde{\varphi}}(U_i)\}|K_{h,ij}^{(l)}||U_i=u\right]\right\}^2du}\\ =\;&\frac{1}{h^{l+1}}\sqrt{\int_{-1}^1\left\{\int_{-1}^1\mathbf{E}\left[\tilde{\psi}(W_i)\{\tilde{\varphi}(Y_j)+g_{\tilde{\varphi}}(U_i)\}|U_i=u_1\right]\left|K^{(l)}\left(\frac{u-u_1}{h}\right)\right|du_1\right\}^2du}\\ =\;&\frac{1}{h^l}\sqrt{\int_{-1}^1\left\{\int_{\{(u-1)/h\}\vee(-1)}^{\{u/h\}\wedge 1}\mathbf{E}\left[\tilde{\psi}(W_i)\{\tilde{\varphi}(Y_j)+g_{\tilde{\varphi}}(U_i)\}|U_i=u_2\right]|K^{(l)}(u_2)|\,du_2\right\}^2du}\\ =\;&O(h^{-l}),\end{aligned}$$

Hence, by using the maximal inequality, we deduce that the leading term in (22) is $O_P(n^{-bl}h^{-l})=O_P(n^{-2b}h^{-2})=o_P(1)$. Therefore, we obtain the wanted result. ■

References

[1] Abrevaya, J. and J. Huang (2005), "On the bootstrap of the maximum score estimator," *Econometrica*, 73, 1175-1204.

[2] Ai, C. and X. Chen (2003), "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795-1843.

[3] 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.

[4] Billingsley, P (1999), *Convergence of Probability Measures*, Second Edition, John Wiley & Sons, New York.

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

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

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

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

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

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

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

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

[13] 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.

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

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

[16] Ledoux, M. and M. Talagrand (1988), "Un critére sur les petite boules dans le théorème limite central," *Probability Theory and Related Fields* 77, 29-47.

[17] 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.

[18] Newey, W. and J. Powell (2003), "Instrumental variable estimation of nonparametric function," *Econometrica*, 1565-1578.

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

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

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

[22] Sherman, R. (1994), "Maximal inequalities for degenerate U-processes with application to optimization estimators," *Annals of Statistics*, 22, 439-459.

[23] Song, K. (2008a), "Uniform convergence of series estimators over function spaces," forthcoming in *Econometric Theory*.

[24] Song, K. (2008b), "Testing conditional independence using Rosenblatt transforms," Working paper, University of Pennsylvania.

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

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

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

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

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

[30] 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.