INFERENCE IN SEMIPARAMETRIC BINARY RESPONSE MODELS WITH INTERVAL DATA*

YUANYUAN WAN† AND HAIQING XU‡

ABSTRACT. This paper studies the semiparametric binary response model with interval data investigated by Manski and Tamer (2002, MT). In this partially identified model, we propose a new estimator based on MT's modified maximum score (MMS) method by introducing density weights to the objective function, which allows us to develop asymptotic properties of the proposed set estimator for inference. We show that the density–weighted MMS estimator converges at a nearly cube–root–n rate. We propose an asymptotically valid inference procedure for the identified region based on subsampling. Monte Carlo experiments provide supports to our inference procedure.

Key words: Interval data, semiparametrc binary response model, density weights, $\mathscr{U}$–process

JEL: C12, C14 and C24

1. INTRODUCTION

Interval data is a common feature in empirical research. For example, as an explanatory variable, family income might be measured by a bracket with only upper and lower bounds reported to researchers. Models with interval data have been systematically investigated in

---

Date: Tuesday 16th September, 2014.

*We are grateful to two anonymous referees, an Associate Editor and the Co–Editor for their insightful comments, which helped generate a substantially improved paper. We have greatly benefited from interactions with Jason Abrevaya, Victor Aguirregabiria, Sung Jae Jun, Isabelle Perrigne, Joris Pinkse, and Quang Vuong, and the comments from Tim Armstrong, Jason Blevins, Federico Bugni, Qi Li, George Shoukry, Xun Tang and seminar participants at the 2013 Asia Meeting of Econometric Society, the 2013 China Meeting of Econometric Society and the 9th GNYEC. Wan gratefully acknowledges research support from SSHRC. All errors are ours.

† (corresponding author) Department of Economics, University of Toronto. 150 St. George Street, Max Gluskin House, Toronto, M5S3G7, ON, Canada. yuanyuan.wan@utoronto.ca.

‡Department of Economics, The University of Texas at Austin, h.xu@austin.utexas.edu.

a seminal paper by Manski and Tamer (2002, MT). For a semiparametric binary response model with interval data, MT propose a modified maximum score (MMS) set estimator and show its consistency. The convergence rate and other asymptotic properties of the MMS estimator, which are necessary for inference, however are not established. In this paper, we extend MT's method and propose a density–weighted MMS (set–valued) estimator, which allows us to establish the asymptotic properties. Further, we propose an asymptotically valid inference procedure for the identified set. Monte Carlo experiments are used to illustrate the finite sample performance of the proposed estimator and inference procedure.

When one explanatory variable ν is not observed but other variables x have been measured precisely, the conditional distribution $\mathbb{P}(y|x,\nu)$ is unknown in the population. MT suggest to characterize the identification region of model parameters based on $\mathbb{P}(y|x,\nu_0,\nu_1)$, where ν_0 and ν_1 are observed lower/upper bounds of ν in the interval data. Instead of modifying the original econometric model and objects of interests, e.g. replacing $\mathbb{P}(y|x,\nu)$ by $\mathbb{P}(y|x,d)$ where d is a discrete random variable indicating which bracket v belongs to, MT's approach treats the observability of data as a separate issue of modeling and data generating process. Although the observed bounds are less informative than ν, they still provide (partial) identification power for the object of interest. MT characterize the sharp identification region for the model parameters and show that their set estimators are consistent.[1] Following that direction, this paper focuses on the interval data issue in a semiparametric binary response model and provides an effective inference procedure for the partially identified parameters.

The issue of interval data also arises in estimating game theoretic models, where some equilibrium variables (e.g., equilibrium beliefs) are not observed but we could possibly derive their estimable upper/lower bounds from equilibrium conditions and model restrictions. For example, in a 2-by-2 game of incomplete information with correlated types, Wan and Xu (2012) show that each player's equilibrium strategy can be represented as a binary response

[1] Magnac and Maurin (2008) discuss the identification of the semiparametric binary response model with interval data when additional instrumental variables are available.

model, in which one of the explanatory variables, the equilibrium belief on the rival's choice, is unknown to researchers and bounded by some nonparametrically estimable functions.

In this paper, we extend MT's MMS method by introducing density weights to the objective function for their MMS estimation. The weighting does not change the identification region of parameters of interest, but allows us to obtain a sample objective function in a $\mathscr{U}$–process form. We further extend Kim and Pollard (1990)'s results on the asymptotic properties for maximum score point estimator to our setting and establish a set of conditions under which our density-weighted MMS estimator is nearly cube-root-n consistent.

Moreover, we follow Chernozhukov, Hong, and Tamer (2007) and construct confidence regions for the partially identified set as level sets of the sample objective function. Abrevaya and Huang (2005) show that the bootstrap for the asymptotic distribution of maximum score estimator is inconsistent. Their intuition carries through to our density–weighted MMS estimator in the partial identification scenario. Therefore, we propose to estimate the critical values by subsampling. Applying the results in Nolan and Pollard (1987, 1988), we show that the inferential statistic converges in distribution to a non–degenerate random variable, which ensures the validity of the subsampling procedure. In Section 4, we conduct Monte Carlo simulations under several choices of subsample sizes. The finite sample performance provide support to our inference procedure.

The key in our sample objective function is that it effectively controls the errors induced by the first stage nonparametric estimation in indicator functions. As in MT, our sample objective function also contains the term $\mathbf{1}\{\mathbb{E}(y|x, v_0, v_1) \geq 1 - \alpha\}$ for some $\alpha \in (0, 1)$, which demands a nonparametric plug–in estimator of the conditional expectation inside the indicator function. By choosing bandwidths and kernels properly, we show that first stage estimation errors are asymptotically negligible and will not distort the asymptotic behavior of the second stage estimator.

Our method is also related to the literature of using $\mathscr{U}$–process theory to derive asymptotic properties of estimators, e.g. Sherman (1994b) establishes the asymptotic properties of the $\mathscr{U}$–processes in the analysis of a generalized semiparametric regression model, which includes Ichimura (1993) and Klein and Spady (1993) as leading examples. The binary response model that we consider in this paper is different from Sherman (1994b) in two aspects. First, the parameters of interest are not point identified. Second, as a trade-off of the robustness from the conditional median assumption, our density–weighted MMS estimator has an "irregular" convergence rate which is slower than root–n. We do, however, discuss the extension of the density–weighting idea to a regular case — the parametric regression model with interval data. We propose a density–weighted modified minimum distance (MMD) method in a similar way to consistently estimate the identified set at a nearly parametric rate.

A line of literature on cube–root–n asymptotics has been developed for a variety of "irregular" estimators (see, e.g., Abrevaya, 2000). For the semiparametric binary response models, traditional maximum score type estimators have been reviewed, e.g. in Kim and Pollard (1990) and Horowitz (1998). The unusual cube–root–n convergence comes from the fact that maximum score sample criterion function is essentially a step function of parameters, which is "irregular" in the sense that it does not allow for a quadratic expansion.[2] Similar intuition carries through to the asymptotic analysis in our setting where the parameters of interest are partially identified: we show that the irregular set estimator converges to the identification region at a rate slightly slower than cube–root–n.[3] Blevins (2012) also studies the asymptotic problems of irregular set estimators, which is related to the present paper, but has a different focus.

The rest of the paper is organized as follows. Section 2 reviews the semiparametric response model with interval data and the MMS estimator proposed by MT. In Section 3, we introduce the density–weighed MMS estimator and provide the conditions for valid

[2] Under additional smoothness assumptions on the error term's density, Horowitz (1992) propose a smoothed MSE, which has a limiting normal distribution and a rate of convergence that is at least $n^{-2/5}$ and can be arbitrarily close to $n^{-1/2}$.

[3] On the other hand, a smoothed sample criterion function does not necessarily guarantee the corresponding estimator will converge at a parametric rate: in a simple setting of binary response models with a special regressor, Khan and Tamer (2010) show that the identification–at–infinity of parameters could also result in a convergence rate slower than the parametric rate. Chen, Khan, and Tang (2013) extend such a result.

inference. Section 4 reports Monte Carlo experiment results. We discuss some possible extensions in Section 5 and conclude the paper is Section 6.

2. Semiparametric Binary Response Model with Interval Data

Consider the following semiparametric binary response model studied in MT,

$$y = 1\left[x'\beta + \delta\nu + \epsilon > 0\right],$$

where $x \in \mathbb{R}^d$, $\nu \in \mathbb{R}$ and $\epsilon \in \mathbb{R}$. (y, x', ν_0, ν_1) are observed to researchers with $\nu_0 \le \nu \le \nu_1$. $\beta \in B \subset \mathbb{R}^d$ and $\delta \in \mathbb{R}$ are parameters of interest. The following assumption is made in MT, and throughout the present paper as well.

Assumption 2.1. *Let Semiparametric Binary Regression (SBR) assumptions hold.*

SBR-1 *For a specified* $\alpha \in (0,1)$, $q_\alpha(\epsilon|x,\nu) = 0$. $\mathbb{P}(\epsilon \le 0|x,\nu) = \alpha$.

SBR-2 $\mathbb{P}(\epsilon|x,\nu,\nu_0,\nu_1) = \mathbb{P}(\epsilon|x,\nu)$.

SBR-3 $\delta > 0$.

Assumption SBR-1 is the α–quantile–independence condition suggested by Manski (1975, 1985); SBR-2 asserts that observation of $[\nu_0, \nu_1]$ would not provide additional information for the distribution of ϵ if we know ν and x. SRB-2 holds if the bracket for each ν is generated at random, i.e., given x and ν, which bracket (with $\nu_0 \le \nu \le \nu_1$) to be reported has to be independent with ϵ (see Aucejo, Bugni, and Hotz, 2013). In practice, if the set of brackets are predetermined for reporting ν and forms a partition on the real line, then the conditional distribution of ν_0 and ν_1 given ν is degenerate and SBR-2 holds trivially. Assumption SBR-3 is strong but could be substituted with weaker model restrictions that identify the sign of δ. In addition, positive δ constitutes a normalization.

As pointed out by MT, the threshold–crossing condition is invariant to the scale of the parameters. Hence, we set $\delta = 1$ throughout as a scale normalization. Further, MT characterize the sharp identification region of β by

$$B^* = \{b \in \mathbb{R}^k : \mathbb{P}[T(b)] = 0\}, \tag{1}$$

where $T(b) = \{(x, v_0, v_1) : (xb + v_1 \leq 0 < x\beta + v_0) \cup (x\beta + v_1 \leq 0 < xb + v_0)\}$.

MT propose a consistent set estimator for B^*: the modified maximum score (MMS) estimator. Let $z = (x', v_0, v_1)'$ and $P(z) = \mathbb{P}(y = 1|z)$. Let further $\lambda(z) = 1[P(z) > 1 - \alpha]$ and $\text{sgn}(\cdot)$ be the conventional sign function.[4]

Assumption 2.2. $\mathbb{P}[P(z) = 1 - \alpha] = 0$.

Assumption 2.2 requires that there is no mass point at the $1 - \alpha$ quantile of the distribution of $P(z)$. This assumption excludes the identification power from those values of z's with zero scores, such that for identification of β, the maximum score type criterion function can exploit all the information provided by variations in z. The same assumption has also been made in MT.

Under Assumption 2.1, MT show that every $b \in B^*$ maximizes the following population criterion function,

$$S(b, \lambda) = \int [P(z) - (1 - \alpha)] \times \{\lambda(z) \cdot \text{sgn}(xb + v_1) + [1 - \lambda(z)] \cdot \text{sgn}(xb + v_0)\} dF(z). \quad (2)$$

If, in addition, Assumption 2.2 holds, then none of $b \notin B^*$ maximizes the above criterion function.

To estimate B^*, MT propose a modified maximum score (MMS) set estimator

$$B_n = [b \in B : S_n(b, \widehat{\lambda}) \geq \max_{c \in B} S_n(c, \widehat{\lambda}) - \gamma_n],$$

where $\gamma_n \downarrow 0$ a.s. at a specific rate and the sample criterion function is given by

$$S_n(b, \widehat{\lambda}) \equiv \frac{1}{n} \sum_{i=1}^{n} [y_i - (1 - \alpha)] \times \left\{\widehat{\lambda}(z_i) \cdot \text{sgn}(x_i b + v_{1i}) + \left[1 - \widehat{\lambda}(z_i)\right] \cdot \text{sgn}(x_i b + v_{0i})\right\}, \quad (3)$$

[4]We adopt the convention $sgn(0) = -1$.

in which $\widehat{\lambda}(z_i) = 1[P_n(z_i) > 1-\alpha]$ and P_n is a nonparamteric estimator of P. In particular, if $\nu_0 = \nu = \nu_1$, then the MMS estimator becomes the classical maximum score estimator.

Assumption 2.3. *B is compact. For any $b \in B$, $xb + \nu$ has a bounded probability density function with respect to the Lebesgue measure.*

Assumption 2.4. *$(y_i, z_i)_{i=1}^n$ is an i.i.d. random sample.*

Assumption 2.5. *$P_n(z) \xrightarrow{a.s.} P(z)$, for a.e. z.*

Under Assumptions 2.1 to 2.5, MT show that their MMS set estimator is consistent. Later, we will substitute Assumption 2.5 with primitive conditions on kernel functions and bandwidths, under which we will show that our density–weighted MMS estimator is nearly cube–root–n consistent.

3. Density–Weighted Maximum Score Method

We now introduce a density–weighted objective function, which is also maximized at the identification region B^*. We then define a set estimator B_n^f in a $\mathcal{U}$–process form as a level set of the sample criterion function. Further, we show that our estimator converges to B^* at a rate slightly slower than $n^{1/3}$ and propose an inference procedure for both β and B^*.

For notational simplicity and without loss of generality, we assume that z is continuously distributed and $f(\cdot)$ is the pdf, even though in empirical work discrete covariates are prevalent. Kernel estimation with discrete regressors can be accommodated.[5] Thus, we define the "density–weighted" population objective function as

$$L(b,\lambda) = \int [P(z) - (1-\alpha)] f(z) \times \{\lambda(z) \cdot \operatorname{sgn}(xb+\nu_1) + [1-\lambda(z)] \cdot \operatorname{sgn}(xb+\nu_0)\} \, dF(z). \quad (4)$$

[5] In practice, if z contains a discrete variable that is ordered and takes many values, we can treat it as a continuous variable or apply smoothing method to it without splitting the sample. See e.g. Racine and Li (2004). We thank a referee for pointing out this.

The objective function $L(b, \lambda)$ is weighted by density f as opposed to $S(b, \lambda)$, where the weights are always positive constants. As stated in Lemma 3.1, the set–maximizer of $L(\cdot, \lambda)$ is still B^*.

Lemma 3.1. *Suppose that Assumptions 2.1 and 2.2 holds. Let B^* be the set defined by equation (1). Then $b \in B^*$ if and only if b maximizes $L(b, \lambda)$.*

The proof is similar to that of MT's Lemma 1, and therefore omitted here. While any weighting function w, strictly positive on the support of z, provides the same identification region as B^*, here we choose $w = f$. The "density–weighted" idea can also be found in Powell, Stock, and Stoker (1989) and it permits applying the $\mathscr{U}$–process theory for asymptotic analysis.

Let P_n and f_n be kernel–based nonparametric estimators, defined by

$$P_n(z_i) = \sum_{j \neq i}^{n} y_j K\left(\frac{z_j - z_i}{h}\right) \Big/ \sum_{j \neq i}^{n} K\left(\frac{z_j - z_i}{h}\right), \quad f_n(z_i) = \frac{1}{(n-1)h^p} \sum_{j \neq i}^{n} K\left(\frac{z_j - z_i}{h}\right), \tag{5}$$

in which $K(\cdot)$ and $h \in \mathbb{R}^+$ are kernel function and bandwidth, respectively, and $p = d + 2$ is the dimension of z. Because $1\{P(z) > 1 - \alpha\} = 1\{[P(z) - (1-\alpha)] f(z) > 0\}$ almost surely, we propose to estimate $\lambda(z)$ by

$$\lambda_n(z) = 1\{[P_n(z) - (1-\alpha)] f_n(z) > 0\}.$$

Let $\vartheta(z, b, \lambda) = \lambda(z) \cdot sgn(xb + v_1) + [1 - \lambda(z)] \cdot sgn(xb + v_0)$. Now we define the sample analog of $L(\cdot, \lambda)$ by

$$L_n(b, \lambda_n) = \frac{1}{n} \sum_{i=1}^{n} [P_n(z_i) - (1-\alpha)] \times f_n(z_i) \times \vartheta(z_i, b, \lambda_n),$$

which can also be represented as a second order $\mathscr{U}$–process, i.e.

$$L_n(b, \lambda_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} g_n^*(z_i, z_j; b, \lambda_n), \tag{6}$$

where $g_n^*(z_i, z_j; b, \lambda_n) = \left[y_j - (1-\alpha)\right] \times \frac{1}{h^p} K\left(\frac{z_j - z_i}{h}\right) \times \vartheta(z_i, b, \lambda_n)$. By the $\mathscr{U}$–process representation of our sample criterion function, λ_n is essentially a nonparametric plug–in estimator of the first stage. Note that here we choose the same kernel and bandwidth for both λ_n and g_n^* in equation (6), which is not necessary in a more general analysis. Similar to MT, our second–stage set estimator is defined by

$$B_n^f = [b \in B : \sup_{c \in B} L_n(c, \lambda_n) - L_n(b, \lambda_n) \leq \gamma_n],$$

for some deterministic sequence $\gamma_n \downarrow 0$. Before exploring our set estimator in more detail, we make several assumptions for asymptotic analysis.

Assumption 3.1. *P and f are twice continuously differentiable at all values of z in the support.*

Assumption 3.2. *h is a deterministic sequence satisfying $nh^p \to \infty$ and $h \to 0$ as $n \to \infty$.*

Assumption 3.3. *The kernel $K : \mathbb{R}^p \to R$ is a symmetric function satisfying (1) $\int_{\mathbb{R}^p} K(u)du = 1$; (2) $\int_{\mathbb{R}^p} uK(u)du = 0$; (3) $\int_{\mathbb{R}^p} \|u\|^2 |K(u)|du < \infty$; (4) $\sup_u |K(u)| = \overline{K} < \infty$.*

Assumptions 3.1 to 3.3 are standard in the nonparametric estimation literature (see, e.g. Pagan and Ullah, 1999). Further, Assumption 3.1 implies that f is bounded above.

For any generic value of $a \in \mathbb{R}^d$ and any subset $A \subseteq B$, let $\rho(a, A) = \inf_{b \in A} \|a - b\|$, where $\|\cdot\|$ is the usual Euclidean norm. Theorem 1 below establishes the consistency of our set estimator, the proof of which is similar to that for Proposition 3 in MT.

Theorem 1. *Let Assumptions 2.1 to 2.4 and 3.1 to 3.3 hold. Then $\sup_{b \in B_n^f} \rho(b, B^*) \xrightarrow{p} 0$; if in addition $\sup_{b \in B} \left[L_n(b, \lambda) - L_n(\beta, \lambda)\right] / \gamma_n \xrightarrow{p} 0$ and $\sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)| / \gamma_n \xrightarrow{p} 0$, then $\sup_{b \in B^*} \rho(b, B_n^f) \xrightarrow{p} 0$.*

Proof. See Appendix A.1. □

Note that the condition $\sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)| / \gamma_n \xrightarrow{p} 0$ in Theorem 1 is also required by MT; the condition $\sup_{b \in B} \left[L_n(b, \lambda) - L_n(\beta, \lambda)\right] / \gamma_n \xrightarrow{p} 0$ is similar to Condition

C.1 (e) in Chernozhukov, Hong, and Tamer (2007). The conditions in Theorem 1 require us to specify a sequence γ_n converging to zero at a proper rate. In Sections 3.1 and 3.2, we will derive a uniform convergence rate for $L_n(\cdot, \lambda) - L_n(\beta, \lambda)$ and $L_n(\cdot, \lambda_n) - L_n(\cdot, \lambda)$ under additional conditions, which provides a guidance for choosing γ_n such that the resulting level set is a consistent estimator for B^*.

3.1. **Convergence rate.** In this section we derive the convergence rate for the density–weighted MMS estimator using the standard $\mathscr{U}$–process theory.

Let $g_n(z_i, z_j; b, \lambda) = (1/2)\left[g_n^*(z_i, z_j; b, \lambda) + g_n^*(z_j, z_i; b, \lambda)\right]$ be a symmetric function of z_i and z_j. By definition, $L_n(b, \lambda) = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} g_n(z_i, z_j; b, \lambda)$. Let further $\tilde{g}_n(z_i, z_j; b, \lambda) = g_n(z_i, z_j; b, \lambda) - g_n(z_i, z_j; \beta, \lambda)$ be the corresponding normalization such that $\tilde{g}_n(z_i, z_j; \beta, \lambda) = 0$. We define a $\mathscr{U}$–process sample criterion function by

$$U_n(b, \lambda) = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \tilde{g}_n(z_i, z_j; b, \lambda).$$

By definition, $U_n(b, \lambda) = L_n(b, \lambda) - L_n(\beta, \lambda)$. Equivalently, our set estimator is a level set of the $\mathscr{U}$–process sample criterion function,

$$B_n^f = [b \in B : \sup_{c \in B} U_n(c, \lambda_n) - U_n(b, \lambda_n) \le \gamma_n].$$

Let further $U(b, \lambda) = \mathbb{E} U_n(b, \lambda)$. Note that $U(b, \lambda) = \mathbb{E}\tilde{g}_n(z_1, z_2; b, \lambda)$, which depends on sample size n implicitly through bandwidth h.

To derive the convergence rate for our set estimator, the key step is to establish the uniform convergence rate of of $U_n(b, \lambda_n) - U(b, \lambda)$, which can be decomposed as

$$U_n(b, \lambda_n) - U(b, \lambda) = [U_n(b, \lambda) - U(b, \lambda)] + [U_n(b, \lambda_n) - U_n(b, \lambda)]. \qquad (7)$$

The first term of the right hand side of equation (7) is a demeaned $\mathscr{U}$–process. By extending Kim and Pollard (1990), Proposition 3.1 below provides an upper bound for the norm of this term. The second term on the right hand side of equation (7) is related to the asymptotic orthogonality condition in two–step semiparametric estimation.[6] In Proposition 3.2, we show that this term is asymptotically negligible (of the order $o_p(n^{-2/3})$), which exploits the fact that the plug–in estimator λ_n does not introduce much error to the sample criterion function unless the term $[P(z_i) - (1-\alpha)] \times f(z_i)$ is quite close to zero.

Sherman (1994b) studies the asymptotic properties of a generalized semiparametric regression estimator using a type of $\mathscr{U}$–process structure. The problem that we are looking at is different from that of Sherman's in the sense that our sample objective function is non–smooth and does not allow for a quadratic Taylor expansion. As a result, the convergence rate of our density–weighted MMS estimator is slower than the $\sqrt{n}$–rate in Sherman (1994b).

Proposition 3.1. *Let Assumptions 2.1 to 2.4 and 3.1 to 3.3 hold. Then for any $\epsilon > 0$, there exists a sequence of random variables $\{M_n\}$ of order $O_p(1)$, which does not depend on B^*, such that*

$$|U_n(b,\lambda) - U(b,\lambda)| \leq \epsilon\rho(b,B^*)^2 + n^{-\frac{2}{3}}M_n^2, \quad \textit{for all } b \in B.$$

Proof. See Appendix A.2. □

Note that Proposition 3.1 is not directly applicable to derive the convergence rate as it is stated under unknown function λ. To ensure the error introduced by the first stage estimation of λ is negligible, we strengthen Assumptions 3.1 to 3.3 as follows.

Assumption 3.4. *The kernel K satisfies all the conditions in Assumption 3.3 and for some $R \geq 1$, (i) $\int_{\mathbb{R}^p} u_1^{r_1}, ..., u_p^{r_p} K(u)du = 0$ if $1 \leq \sum_{k=1}^p r_k \leq R-1$; (ii) $\int_{\mathbb{R}^p} u_1^{r_1}, ..., u_p^{r_p} K(u)du = 1$ if $\sum_{k=1}^p r_k = R$, where $r_k \in \mathbb{N}^+$ for $k = 1, ..., p$.*

Let $f_{\epsilon|(x,v)}$ be the conditional density of ϵ given (x,v), similarly for $f_{v|z}$.

Assumption 3.5. *(1) f is everywhere R-continuously differentiable with bounded R-th partial derivatives. (2) $f_{\epsilon|(x,v)}$ is $(R-1)$-continuously differentiable with respect to v*

[6] For more discussion about the orthogonality condition in two step–semiparametric estimation, see Andrews (1994).

and has a bounded $(R-1)$–th partial derivatives in a neighbourhood of $x'\beta + v$ for all (x, v). $f_{v|z}$ is everywhere R-continuously differentiable with respect to (x, v_0, v_1) and has a bounded R-th partial derivative.

Assumption 3.5 imposes smoothness conditions for $f_{\epsilon|(x,v)}$, $f_{v|z}$ and the marginal density f of z. It ensures $P(\cdot)$ and $f(\cdot)$ are everywhere R–continuously differentiable with bounded R–th partial derivatives.

Assumption 3.6. *h is a deterministic sequence satisfying $n^{\frac{1}{3}+r}\left(nh^p\right)^{-\frac{1}{2}} \to 0$ and $n^{\frac{1}{3}+r}h^R \to 0$ for some $r > 0$.*

Assumption 3.6 requires that the high order kernel we use should satisfy $R > p$, a similar condition is also assumed in Powell, Stock, and Stoker (1989). Note that if we choose the optimal rate for bandwidth, i.e., $h \propto n^{-\frac{1}{2R+p}}$, then Assumption 3.6 requires $1/3 < R/(2R+p)$. It should also be noted that Assumptions 2.4 and 3.4 to 3.6 guarantee that $(f_n(z), P_n(z))$ converge to $(f(z), P(z))$ faster than $\sqrt[3]{n}$ for all $z \in \mathscr{Z}$.

Assumption 3.7. *There exists a neighborhood around zero, denoted as $\mathcal{N}_\delta$, and a constant $C_\delta > 0$, such that for any subset $S \subseteq \mathcal{N}_\delta$, there is*

$$\mathbb{P}(\xi \in S) \leq C_\delta \times \mu(S),$$

where $\xi \equiv [P(z) - (1-\alpha)]\, f(z)$ and μ is the Lebesgue measure.

By definition ξ is random variable. Assumption 3.7 is a technical condition that ensures that ξ is smoothly distributed in a small neighborhood of zero. It plays a similar role as Assumption 2.2.

Proposition 3.2. *Let Assumptions 2.1 to 2.4 and 3.4 to 3.7 hold. Then,*

$$\sup_{b \in B} |U_n(b, \lambda_n) - U_n(b, \lambda)| = o_p(n^{-2/3}).$$

Proof. See Appendix A.3. □

Proposition 3.2 is crucial because it allows us to focus on the infeasible sample analog $L_n(b, \lambda)$, or the $\mathscr{U}$–process $U_n(b, \lambda)$.

Assumption 3.8. *Let ∂B^* denote the boundary of the identified set B^* and β is in the interior of B^*. Then*

$$\sup_{b^* \in \partial B^*} \lim_{\eta \uparrow 0} \nabla^2_{\eta\eta} L\big(\eta\beta + (1-\eta)b^*, \lambda\big) < 0.$$

Assumption 3.8 is a partial identification condition. The primitive conditions can be derived using the method of derivatives as surface integrals (see Loomis and Sternberg, 1968). In particular, it holds when the structural error ϵ is independent of z, ϵ has a continuous density and the angular component of z also has a continuous density. In the point identification case, Assumption 3.8 is equivalent to the condition that the population objective function has a negative definite second deravtive matrix at β.

Assumption 3.9. $n^{2/3}\gamma_n \to \infty$ *and* $\gamma_n \to 0$.

Under Assumption 3.9, we can choose $\gamma_n \to 0$ at a rate slightly slower than $n^{-2/3}$. One possible choice is the iterated logrithm: $\gamma_n = n^{-2/3}\sqrt{2 \ln \ln n}$. Note that the statement in the second part of Theorem 1 imposes a restriction on the choice of γ_n, i.e. $\sup_{b \in B^*_\delta} |L_n(b, \lambda_n) - L(b, \lambda)| / \gamma_n \xrightarrow{p} 0$ for some fixed $\delta > 0$. By Propositions 3.1 and 3.2 and Assumption 3.8, it can be verified that $\gamma_n = n^{-2/3}\sqrt{2 \ln \ln n}$ satisfies the requirement.

For any two generic sets A and B, let $\rho_H(A, B)$ be the Hausdorff distance between them, i.e. $\rho_H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} \|a - b\|, \sup_{b \in B} \inf_{a \in A} \|a - b\|\}$.

Theorem 2. *Suppose Assumptions 2.1 to 2.4 and 3.4 to 3.9 are satisfied, then the Hausdorff distance* $\rho_H\left(B^f_n, B^*\right) = O_p\left(\sqrt{\gamma_n}\right)$.

Proof. See Appendix A.4. □

Assumptions 3.4 and 3.6 specify conditions on the bandwidth and kernel function for the first stage estimation. In Section 4, we examine the finite sample performance of our estimator under difference choices of bandwidth and kernel functions by simulations.

3.2. **Inference.** We now construct confidence regions for B^*. Our method follows the subsampling procedure proposed by Chernozhukov, Hong, and Tamer (2007), which is also used in Blevins (2012). To give a heuristic argument, let $1-\alpha \in (0,1)$ be the confidence level and C_α be a positive constant to be specified. Define a level set

$$\widehat{B}_\alpha = \{b : U_n(b,\lambda_n) \geq \sup_{b\in B} U_n(b,\lambda_n) - n^{-2/3}C_\alpha\}.$$

and a random variable

$$\mathcal{C}_n = n^{2/3}\sup_{b\in B} U_n(b,\lambda) - n^{2/3}\inf_{b\in B^*} U_n(b,\lambda).$$

Further, in Lemma B.13 it has been shown that $\mathcal{C}_n$ converges in distribution to a random variable $\mathcal{C}$. Suppose that $\mathcal{C}$ is continuous at its $1-\alpha$ quantile C_α. Then

$$\begin{aligned}
\mathbb{P}(B^* \subseteq \widehat{B}_\alpha) &= \mathbb{P}(\inf_{b\in B^*} U_n(b,\lambda_n) \geq \sup_{b\in B} U_n(b,\lambda_n) - n^{-2/3}C_\alpha) \\
&= \mathbb{P}(n^{2/3}\sup_{b\in B} U_n(b,\lambda) - n^{2/3}\inf_{b\in B^*} U_n(b,\lambda) \leq C_\alpha + o_p(1)) \\
&= \mathbb{P}(\mathcal{C}_n \leq C_\alpha + o_p(1)) = \mathbb{P}(\mathcal{C} \leq C_\alpha) + o(1) = 1-\alpha+o(1) \quad (8)
\end{aligned}$$

where in the second step we replace λ_n with λ in the sample analog and use Proposition 3.2. This gives a confidence region for B^* with asymptotic coverage probability $1-\alpha$. The critical value C_α can be estimated by following subsampling procedure.

Assumption 3.10. *Let $m = [n^a]$ for some $0 < a < 1$, where $[\cdot]$ denotes the integer part.*

Algorithm 1 (Subsampling)**.** *Our subsampling procedure consists the following steps.*

(1) *Obtain a consistent estimate B_n^f using the whole sample.*

(2) *Choose a subsample of size* m *such that* $m \to \infty$ *and* $m/n \to 0$. *Let* $T_n = \binom{n}{m}$ *be the number of subsamples.*

(3) *For the j–th subsample,* $j = 1, \cdots, T_n$, *compute* $\hat{\ell}_{j,m,n}$ *as*

$$\hat{\ell}_{j,m,n} = m^{2/3} \sup_{b \in B} U_{m,j}(b, \lambda_{m,j}) - m^{2/3} \inf_{b \in B_n^f} U_{m,j}(b, \lambda_{m,j}), \tag{9}$$

where $U_{m,j}$ *is the sample criterion function and* λ_m, j *is the nonparametric estimates of the j–th subsample, respectively.*[7]

(4) *Let* $\widehat{C}_\alpha$ *be the* $(1-\alpha)$ *empirical quantile of* $\{\hat{\ell}_{j,m,n}\}_{j=1,\cdots,T_n}$.

(5) *Calculate the confidence set of* B^* *at* $(1-\alpha) \times 100\%$ *level as*

$$\hat{B}_\alpha = \{b : U_n(b, \lambda_n) \geq \sup_{b \in B} U_n(b, \lambda_n) - n^{-2/3}\widehat{C}_\alpha\}.$$

Theorem 3. *Suppose that Assumptions 2.1 to 2.4 and 3.4 to 3.10 hold. Then*

$$\lim_{n \to \infty} \mathbb{P}\left(B^* \subseteq \widehat{B}_\alpha\right) \geq 1 - \alpha.$$

Proof. See Appendix A.5. □

The proof to Theorem 3 follows Chernozhukov, Hong, and Tamer (2007, Theorem 3.3). In particular, Lemmas B.13 and B.14 play similar roles as their conditions C.4 and C5, respectively.

4. Experiments

This section presents some simulation results on the finite sample performance of the density–weighted MMS estimator and the proposed subsampling inference procedure. We consider the following binary response model,

$$y = \mathbf{1}[\beta_0 + \beta_1 v + \beta_2 x - \epsilon \geq 0],$$

[7]Note that due to Proposition 3.2, we can also use the nonparametric estimate λ_n of the original sample to replace $\lambda_{m,j}$. As a matter of fact, we try both in the Monte Carlo experiments. Using λ_n yields a slightly better finite sample performance.

where $(\beta_0, \beta_1, \beta_2) = (1, 1, -1)$, (ν, x, ϵ) are mutually independent. $\epsilon \sim N(0, 1)$, $x \sim U[0, 5]$ and $\nu \sim U[-2, 3]$. We specify the bounds as $\nu_0 = \frac{1}{\kappa}\,\text{int}(\kappa\nu)$ and $\nu_1 = \frac{1}{\kappa}(\,\text{int}(\kappa\nu) + 1)$, where $\text{int}(\nu)$ denote the largest integer smaller than ν and $\kappa \in [1, \infty)$. The length of the interval is $1/\kappa$. When κ increases, the interval becomes smaller and more informative.[8]

Figure 1 plots the sample objective functions (fixing $\beta_0 = 1$) with sample size $n = 2 \times 10^6$. With such a large sample size, we can view them as good approximations to the population objective functions. For the purpose of comparison, all functions are rescaled such that the maximum values equal to 1. The solid and dashed lines are our density–weighted and MT's objective function, respectively. Note that across different values of κ, ours and MT's objective functions have the same maximum value region, which has been shown in Lemma 3.1. As κ increase, the maximum value region of the functions become shorter, reflecting the fact that the identification power of the interval becomes stronger.

Table 1 reports the estimation results for β_2 under different combinations of sample sizes and choices of slackness parameter γ_n. In particular, we consider $\gamma_n = cQ^* n^{-2/3}\sqrt{2 \ln \ln n}$, where $c \in \{0.1, 0.5, 1.0\}$, $Q^* = \sup_{b \in B} U_n(b, \lambda_n)$. We include Q^* in the expression of γ_n only for the purpose of normalization. We choose $\kappa \in \{1, 10\}$, representing two levels of informativeness of the interval. The set estimators are computed by a grid search with grid length equals to 0.001.[9] The numbers in the brackets are averages of bounds of the set estimators. The numbers in the parenthesis are the standard deviation of the estimated bounds. All results are based on 2000 replications.[10] Under appropriate choices of γ_n, both estimators are consistent. Our set estimates are slightly wider than MMSEs, but the standard deviations of the estimated bounds are smaller. Overall, it seems that neither estimator dominates the other in finite samples. See more discussions in Section 5.1.

[8]The numerical examples in Manski and Tamer (2002) set $\kappa = 1$.

[9]We use the fourth order Gaussian kernel and bandwidth $h = 1.08n^{-0.1}$. With $R = 4$ and $p = 3$, we can verify that the requirements in Assumption 3.6 are satisfied.

[10]Throughout this section, we fixed the value $\beta_0 = 1$. Based on a simulation with sample size 200,000, when $\kappa = 1$, the identified region for β_2 is $[-1.142, -0.852]$.

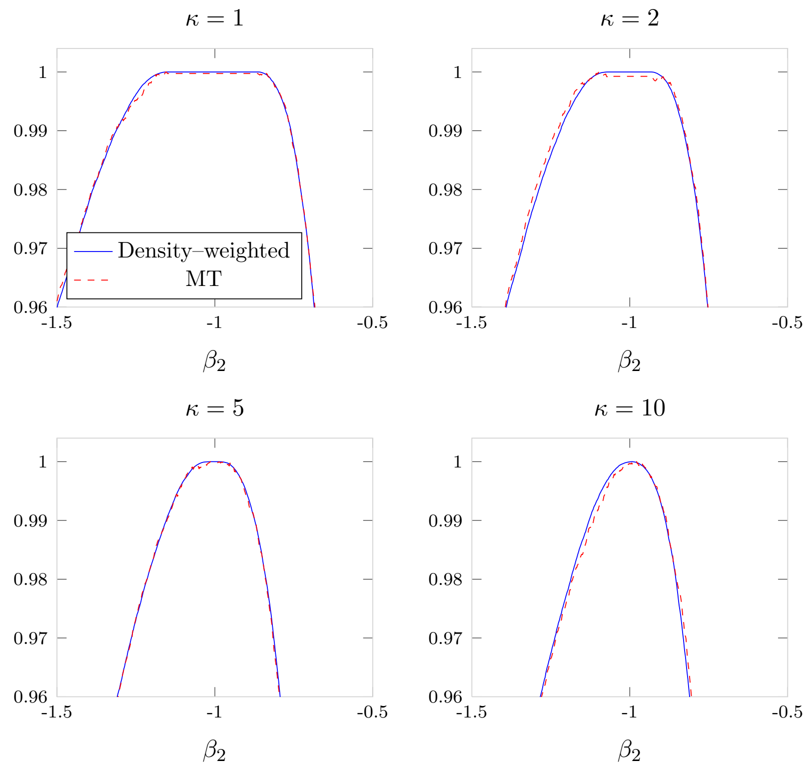

FIGURE 1. Objective Functions ($\beta_0 = 1, n = 2 \times 10^6$)

Figure 2 reports the root mean squared errors (RMSE), absolute value of the average bias and standard deviation of the left boundary estimator (based on $1,000$ replications) under different choices of bandwidths and kernel functions. In particular, we consider $h = c_h n^{-1/10}$ and investigate how the finite sample performance of the left boundary estimator changes as we change c_h. We also conduct simulations for other choices of rate which satisfy Assumption 3.6, all results are qualitatively similar and therefore omitted. In addition to the fourth order Gaussian kernel we used to obtain Table 1, we also consider fourth order Epanechnikov, biweight and triweight kernels (see Hansen, 2005, for detailed functional forms). The simulations is conducted with $\kappa = 1$, $n = 1600$ and $c = 0.1$.

Note first the magnitude of the RMSE and its shape as a function of bandwidth are similar across different kernel functions, especially among Epanechnikov, biweight and triweight

TABLE 1. Estimation, $\kappa \in \{1, 10\}$

n	γ_n	$\kappa = 1$ B_n^f	$\kappa = 1$ B_n^{MT}	$\kappa = 10$ B_n^f	$\kappa = 10$ B_n^{MT}
200	$c = 0.1$	$[-1.198, -0.812]$ $(0.140, 0.089)$	$[-1.088, -0.929]$ $(0.222, 0.197)$	$[-1.087, -0.946]$ $(0.117, 0.089)$	$[-1.047, -0.975]$ $(0.144, 0.129)$
1600		$[-1.193, -0.822]$ $(0.051, 0.033)$	$[-1.058, -0.982]$ $(0.188, 0.191)$	$[-1.052, -0.966]$ $(0.038, 0.031)$	$[-1.015, -0.996]$ $(0.069, 0.067)$
12800		$[-1.173, -0.835]$ $(0.020, 0.014)$	$[-1.053, -0.963]$ $(0.151, 0.166)$	$[-1.032, -0.977]$ $(0.013, 0.011)$	$[-1.007, -0.992]$ $(0.037, 0.036)$
200	$c = 0.5$	$[-1.401, -0.728]$ $(0.183, 0.073)$	$[-1.301, -0.785]$ $(0.208, 0.117)$	$[-1.125, -0.855]$ $(0.168, 0.078)$	$[-1.165, -0.891]$ $(0.172, 0.106)$
1600		$[-1.274, -0.778]$ $(0.056, 0.029)$	$[-1.222, -0.814]$ $(0.119, 0.091)$	$[-1.118, -0.916]$ $(0.042, 0.028)$	$[-1.070, -0.940]$ $(0.068, 0.056)$
12800		$[-1.212, -0.810]$ $(0.013, 0.011)$	$[-1.184, -0.837]$ $(0.178, 0.027)$	$[-1.064 - 0.949]$ $(0.013, 0.010)$	$[-1.040 - 0.960]$ $(0.034, 0.029)$
200	$c = 1.0$	$[-1.568, -0.670]$ $(0.254, 0.084)$	$[-1.456, -0.711]$ $(0.230, 0.083)$	$[-1.397, -0.798]$ $(0.219, 0.073)$	$[-1.284, -0.834]$ $(0.196, 0.088)$
1600		$[-1.335, -0.748]$ $(0.060, 0.027)$	$[-1.440, -0.721]$ $(0.083, 0.046)$	$[-1.294, -0.768]$ $(0.047, 0.027)$	$[-1.121, -0.902]$ $(0.069, 0.046)$
12800		$[-1.242, -0.792]$ $(0.013, 0.010)$	$[-1.227, -0.801]$ $(0.017, 0.002)$	$[-1.088, -0.930]$ $(0.014, 0.010)$	$[-1.066 - 0.940]$ $(0.031, 0.022)$

(1) Slackness parameter $\gamma_n = cn^{-2/3}\sqrt{2\ln\ln n}\sup_{b\in B} U_n(b, \lambda_n)$, $c \in \{0.1, 0.5, 1.0\}$
(2) Results are based on 2000 replications.

kernels. We can also observe the trade-off between the bias and variance. For each choice of kernel function, there is a wide range choices of bandwidth under which the RMSE of the second stage estimator varies very little.

Next we examine the finite sample performance of the proposed inference procedure. Figure 3 reports the coverage frequencies of the subsampling confidence region for the identified interval of β_2: B_2^*. We first compute set estimator B_n^f using slackness parameter

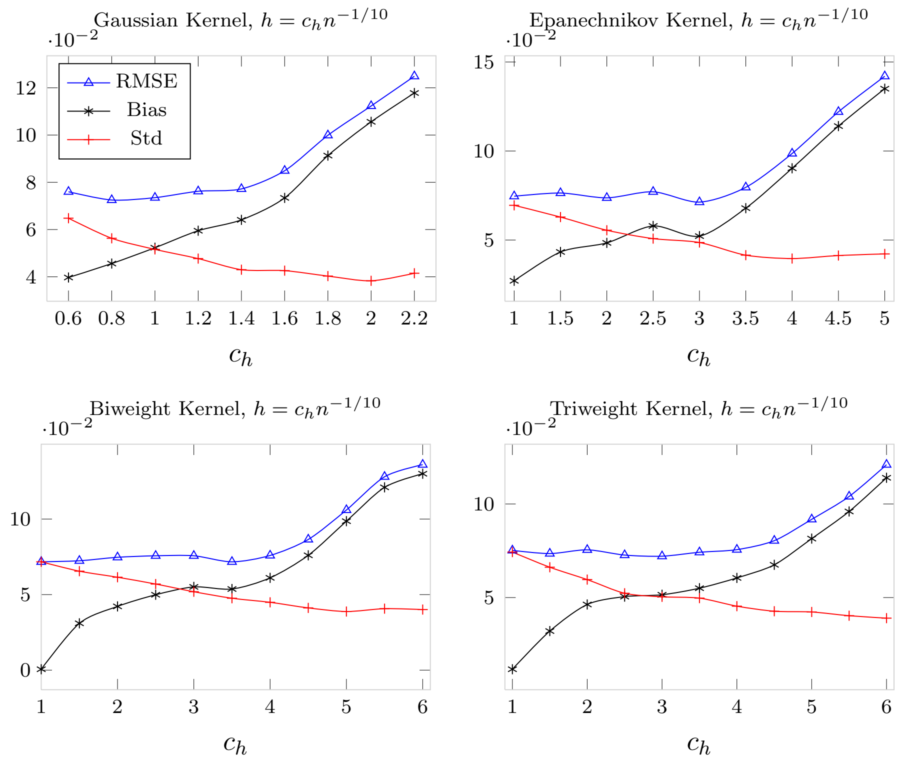

FIGURE 2. Performances under combinations of K and h

$\gamma_n = 0.1Q^* n^{-2/3}\sqrt{2\ln\ln n}$. We approximate the critical values using 4000 subsamples. For the j–th subsample, $j = 1, \cdots, 4000$, we draw m observations without replacement and compute the statistic $\hat{\ell}_{j,m,n}$ based on Equation (9). The critical value $\widehat{C}_\alpha$ is computed as the $100(1-\alpha)\%$ empirical quantile of $\{\hat{\ell}_{j,m,n}\}_{j=1}^{4000}$, with $\alpha = 0.01, 0.05$, and 0.1 respectively. Lastly, the confidence region for B_2^* is then chosen as

$$\widehat{B}_\alpha = \{b : U_n(b, \lambda_n) \geq \sup_{b \in B} U_n(b, \lambda_n) - n^{-2/3}\widehat{C}_\alpha\}.$$

All coverage frequencies are calculated based on 2000 replications.

To the best of our knowledge, there are no generic rules of choosing subsample size m. We report results for different choices as a robustness check. Figure 3 plots the coverage probabilities as a function of subsample sizes. It shows that the inference procedure in general works reasonably well, especially when sample size is large. For a wide range

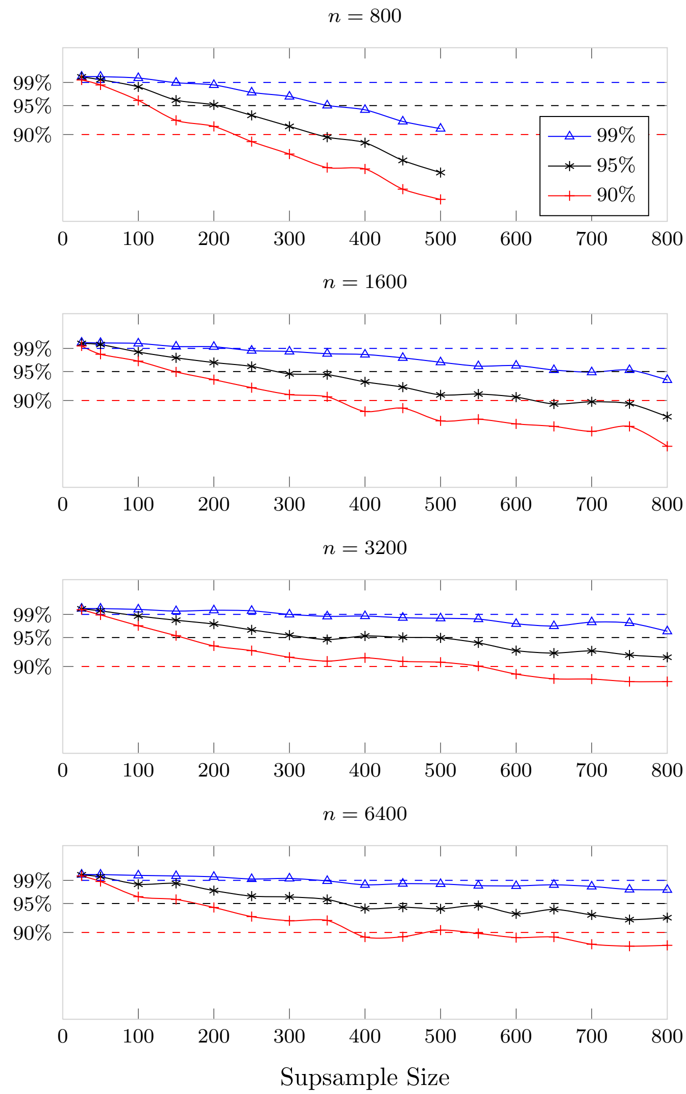

FIGURE 3. Coverage probabilities under combinations of (n, m).

subsample sizes, the coverage frequencies are close the targeted levels. For a given sample size, the coverage frequencies decrease with the subsample size. This is not surprising. When m is large, the sequence $\{\hat{\ell}_{j,m,n}\}_j$ in equation (9) tends to take smaller values (giving everything else equal), so does its empirical quantiles and $\widehat{C}_\alpha$. As a result, computed confidence sets are smaller. Based on the numerical results in Figure 3, the choice of $m \approx 10n^{0.45}$ seems to produce the best overall performance for this data generating process.

5. Discussions

5.1. **Efficiency.** The main focus of this paper is to provide a valid inference procedure in the MT's interval data framework. A related interesting question is to explore the efficiency of the proposed estimator. In particular, are there any efficiency gains for the proposed estimator over MT's MMS estimator?

In our Monte Carlo simulations, the finite sample behaviors of the density–weighted MMS estimator do not suggest obvious gains over MT's MMSE. Because the convergence rate and the limiting distribution of the latter are not available in the literature, it is unclear how to make a theoretical comparison between the two estimators in general cases. It is useful, however, to consider a speical case where the intervals are degenerate, i.e. $\nu_0 = \nu_1$ a.s.. The MMSE then becomes the traditional maximum–score–type point estimator, i.e.[11]

$$\beta_n = \text{argmax}_{b \in B} \frac{1}{n} \sum_{i=1}^{n} [y_i - (1-\alpha)]\, f(z_i)\, \text{sgn}(x_i' b + \nu_i).$$

Following Kim and Pollard (1990), $n^{1/3}(\beta_n - \beta)$ converges in distribution to the random vector that maximizes a Gaussian process with continuous sample paths. On the other hand, our density–weighted MMS estimator becomes

$$\beta_n^f = \text{argmax}_{b \in B} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} [y_j - (1-\alpha)]\, \frac{1}{h^p} K\left(\frac{z_j - z_i}{h}\right) \text{sgn}(x_i' b + \nu_i).$$

Under Assumptions 3.4 and 3.6, it can be shown that $n^{1/3}(\beta_n^f - \beta)$ converges in distribution to the random vector that maximizes the same Gaussian process. Thus β_n^f and β_n have the same limiting distribution. Above discussion suggests that we should not expect efficiency gains by just using the $\mathscr{U}$–process sample analog.

In the general non–degenerate case, another concern on the asymptotics is the the first stage estimation errors introduced by the plug–in estimator λ_n. Proposition 3.2 shows that

[11] Note that the density $f(z)$ is added to ensure that MT's MMS estimator has the same population objective function as our density weighted MMS estimator.

the errors are asymptotically negligible in our approach, which might not be the same case for the MMS estimator.

5.2. **Density Weighted Modified Minimum–distance Estimator.** In this subsection, we briefly describe an extension of our approach to the problem of inference on parametric regression models with interval data, i.e.

$$\mathbb{E}(y|x,v) = g(x,v;\theta)$$

in which g is a known function, $\theta \in \Theta \subset \mathbb{R}^p$ is the (finite dimensional) parameter of interest, and $v \in \mathbb{R}$ is the latent regressor with observed bounds (v_0, v_1), i.e. $v_0 \leq v \leq v_1$. Further, MT assume that $\mathbb{E}(y|x,v,v_0,v_1) = \mathbb{E}(y|x,v)$ and g is weakly monotone in v for each $x \in \mathbb{R}^d$ and $\theta \in \Theta$.

From above model restrictions, MT characterize the sharp identification region for the model parameter θ using moment inequalities: let Θ_I be a collection of $c \in \Theta$ satisfying the following conditions

$$g(x,v_0;c) \leq \mathbb{E}(y|x,v_0,v_1) \leq g(x,v_1;c), \;\; a.s.$$

MT further employ a minimum–distance type objective function for which Θ_I is a set–valued maximizer, the sample analog of which then leads to MT's modified minimum–distance (MMD) estimator. MMD is shown to be consistent under weak conditions.

Similarly, we can introduce density–weights to MT's objective function, for which we can establish asymptotic properties for the induced set estimator and provide an asymptotically valid inference procedure for the identified region in a similar way.

Let $z = (x', v_0, v_1)'$ and z is continuously distributed with a probability density function f. Let further $\lambda_b(z;c) = [\mathbb{E}(y|z) - g(x,v_b,c)] \times f(z)$ for $b = 0,1$ and $\lambda = (\lambda_0, \lambda_1)$. Then our density–weighted MMD objective function is defined as follows

$$Q(c,\lambda) \equiv \int \left\{ 1\left[\lambda_1(z;c) > 0\right] \lambda_1^2(z;c) + 1\left[\lambda_0(z;c) < 0\right] \lambda_0^2(z;c) \right\} dF(z).$$

Therefore, we define our sample objective function

$$Q_n(c,\hat{\lambda}) \equiv \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}\left\{\hat{\lambda}_1(z_i;c)>0\right\}\hat{\lambda}_1^2(z_i;c)+\mathbf{1}\left\{\hat{\lambda}_0(z_i;c)<0\right\}\hat{\lambda}_0^2(z_i;c)\Big],$$

where $\hat{\lambda}$ is the nonparametric estimator of λ: for $d=0,1$,

$$\hat{\lambda}_b(z_i,c)=\frac{1}{(n-1)h^{d+2}}\sum_{j\neq i}\left\{y_j-g(x_i,v_{bi};c)\right\}K\left(\frac{z_j-z_i}{h}\right).$$

Similarly to Proposition 3.2, it could be shown that $Q_n(c,\hat{\lambda})$ behaves asymptotically equivalent to the following infeasible sample criteria function

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}\left\{\lambda_1(z_i;c)>0\right\}\hat{\lambda}_1^2(z_i;c)+\mathbf{1}\left\{\lambda_0(z_i;c)<0\right\}\hat{\lambda}_0^2(z_i;c)\Big],$$

which indeed could also be written as a $\mathscr{U}$–process (after all cross–product terms have been left out).

Our set estimator $\widehat{\Theta}_I^f$ is defined as the set of parameter values which nearly minimize $Q_n(\cdot,\hat{\lambda})$,

$$\widehat{\Theta}_I^f=\{c\in\Theta: Q_n(c,\hat{\lambda})\leq \inf_{\tilde{c}\in\Theta}Q_n(\tilde{c},\hat{\lambda})+\gamma_n\},\quad \gamma_n\downarrow 0.$$

By choosing the deterministic sequence γ_n proportional to, e.g. $n^{-1}\sqrt{2\ln\ln n}$ and under a similar set of conditions, it could be shown that $\rho_H(\hat{\Theta}_I^f,\Theta_I)=O_p(n^{-1/2}\sqrt[4]{2\ln\ln n})$. We can also conduct inference on the identified set by taking appropriate level sets of Q_n, with the critical values chosen based on a similar subsampling procedure as in Section 3.2.

6. Conclusion

This paper studies the semiparametric binary response model with interval data investigated by Manski and Tamer (2002). We propose a density–weighted modified maximum score estimator and derive its asymptotic properties. Further, we propose to construct confidence sets for the identification region by subsampling. We also investigate the finite sample performance of the proposed estimator and inference procedure using Monte Carlo experiments.

Regarding potential empirical applications, here we suggest several possibilities for future studies. First, one can apply our method to conduct inferences on the wealth effect on smoking using Health and Retirement Study (HRS) data where wealth is reported by intervals, as in Manski and Tamer (2002). In particular, our approach provides a confidence interval for the wealth parameter. A second possible fruitful area to apply our method is the empirical game. For example, in the oligopoly entry model, firms' entry behaviors are explained by their "beliefs" on the rival's entry probabilities, as well as other exogenous profit shifters. Wan and Xu (2012) show that the equilibrium beliefs are partially identified by intervals that are nonparametrically estimable. An important question is then how the first–stage nonparametric estimation bias affects the performance of the second–stage density–weighted MMS estimator. We leave it for our future research.

REFERENCES

ABREVAYA, J. (2000): "Rank Estimation of a Generalized Fixed-effects Regression Model," *Journal of Econometrics*, 95(1), 1–23.

ABREVAYA, J., AND J. HUANG (2005): "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, 73(4), 1175–1204.

ANDREWS, D. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

AUCEJO, E. M., F. A. BUGNI, AND V. J. HOTZ (2013): "Identification and Inference on Regressions with Missing Covariate Data," Working paper.

BLEVINS, J. R. (2012): "Non-Standard Rates of Convergence of Criterion-Function-Based Set Estimators," Working paper.

CHEN, S., H. KHAN, AND X. TANG (2013): "Informational Content of Special Regressors in Heteroskedastic Binary Response Model," Working papers.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

DAVYDOV, Y. A., M. A. LIFSHITS, AND N. V. SMORODINA (1998): *Local Properties of Distributions of Stochastic Functionals*. Providence, RI: American Mathematical Society.

HANSEN, B. E. (2005): "Exact Mean Integrated Squared Error of Higher Order Kernel Estimators," *Econometric Theory*, 21(6), 1031–1057.

HOROWITZ, J. (1998): *Semiparametric Methods in Econometrics*. Springer.

HOROWITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-index Models," *Journal of Econometrics*, 58(1-2), 71–120.

KHAN, S., AND E. TAMER (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78(6), 2021–2042.

KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18(1), 191–219.

KLEIN, R. W., AND R. H. SPADY (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), 387–421.

KOSOROK, M. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

LOOMIS, L. H., AND S. STERNBERG (1968): *Advanced Calculus*. Addison–Wesley.

MAGNAC, T., AND E. MAURIN (2008): "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data," *Review of Economic Studies*, 75(3), 835–864.

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response : Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), 313 – 333.

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

NOLAN, D., AND D. POLLARD (1987): "U-Processes: Rates of Convergence," *The Annals of Statistics*, 15(2), 780–799.

NOLAN, D., AND D. POLLARD (1988): "Functional Limit Theorems for U-Process," *The Annals of Probabilities*, 16(3), 1291–1298.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

RACINE, J., AND Q. LI (2004): "Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data," *Journal of Econometrics*, 119(1), 99–130.

SHERMAN, R. (1994a): "Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators," *The Annals of Statistics*, 22(1), 439–459.

——— (1994b): "U-processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10(2), 372–395.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer Verlag.

WAN, Y., AND H. XU (2012): "Semiparametric Identification and Estimation of Binary Decision Games of Incomplete Information with Correlated Private Signals," Working paper.

APPENDIX A. PROOF OF MAIN RESULTS

A.1. **Proof of Theorem 1.**

Proof. Assumption 2.3 and Lemma 3.1 imply that the population objective function $L(\cdot,\lambda)$ is continuous on B and is maximized over B^*. Then for any given $\eta > 0$, let $\delta_\eta = \sup_{b\in B} L(b,\lambda) - \sup_{b\in B/B^*_\eta} L(b,\lambda) > 0$, where B^*_η is the η–expansion of B^*.

To show $\sup_{b\in B^f_n} \rho(b,B^*) \xrightarrow{p} 0$, it is equivalent to show that for any given $\eta > 0$, $B^f_n \subseteq B^*_\eta$ with probability approaching to one (w.p.a.1.), for which it suffices to show that $\inf_{b\in B^f_n} L(b,\lambda) > \sup_{b\in B/B^*_\eta} L(b,\lambda)$ w.p.a.1. Let $\Delta_n(b,\lambda) \equiv L_n(b,\lambda_n) - L(b,\lambda)$. By Lemmas B.1 and B.2,

$\sup_{b\in B}\left|\Delta_n(b,\lambda)\right| = o_p(1)$. Then we have

$$\begin{aligned}
\inf_{b\in B_n^f} L(b,\lambda) &\overset{(a)}{\geq} \inf_{b\in B_n^f} L_n(b,\lambda_n) - \sup_{b\in B_n^f} \Delta_n(b,\lambda) \geq \inf_{b\in B_n^f} L_n(b,\lambda_n) - \sup_{b\in B} |\Delta_n(b,\lambda)| \\
&\overset{(b)}{\geq} \sup_{b\in B} L_n(b,\lambda_n) - \gamma_n - \sup_{b\in B} |\Delta_n(b,\lambda)| \overset{(c)}{\geq} \sup_{b\in B} L(b,\lambda) - \gamma_n - 2\sup_{b\in B} |\Delta_n(b,\lambda)| \\
&\overset{(d)}{=} \sup_{b\in B/B_\eta^*} L(b,\lambda) + \delta_\eta - \gamma_n - 2\sup_{b\in B} |\Delta_n(b,\lambda)|,
\end{aligned}$$

where (a) is because of $\inf_x [g(x)+h(x)] \geq \inf_x g(x) + \inf_x h(x) = \inf_x g(x) - \sup_x[-h(x)]$; (b) is by the definition of B_n^f; (c) comes from the fact $\sup_x [g(x)+h(x)] \geq \sup_x g(x) - \sup_x[-h(x)] \geq \sup_x g(x) - \sup_x |h(x)|$; and (d) is by the definition of δ_η. Since $\gamma_n \downarrow 0$, $\sup_{b\in B} |\Delta_n(b,\lambda)| \xrightarrow{p} 0$, and $\delta_\eta > 0$, it follows that w.p.a.1., $\delta_\eta - \gamma_n - 2\sup_{b\in B} |\Delta_n(b,\lambda)| > 0$. Thus the result follows immediately.

Now we show the second part of Theorem 1. Suppose that $\sup_{b\in B} |L_n(b,\lambda_n) - L_n(\beta,\lambda_n)|/\gamma_n \xrightarrow{p} 0$ and $\sup_{b\in B} |L_n(b,\lambda_n) - L_n(b,\lambda)|/\gamma_n \xrightarrow{p} 0$. It suffices to show that $B^* \subseteq B_n^f$ w.p.a.1, which implies that $\sup_{b\in B^*} \rho(b, B_n^f) \xrightarrow{p} 0$. W.p.a.1,

$$\begin{aligned}
\inf_{b\in B^*} L_n(b,\lambda_n) &\geq \inf_{b\in B^*} L_n(b,\lambda) - \sup_{b\in B^*} |L_n(b,\lambda_n) - L_n(b,\lambda)| \\
&\overset{(e)}{\geq} \sup_{b\in B} L_n(b,\lambda) - \sup_{b\in B} |L_n(b,\lambda) - L_n(\beta,\lambda)| - \sup_{b\in B^*} |L_n(b,\lambda_n) - L_n(b,\lambda)| \\
&\geq \sup_{b\in B} L_n(b,\lambda_n) - \sup_{b\in B} [L_n(b,\lambda) - L_n(\beta,\lambda)] - 2\sup_{b\in B} |L_n(b,\lambda_n) - L_n(b,\lambda)| \overset{(f)}{\geq} \sup_{b\in B} L_n(b,\lambda_n) - \gamma_n,
\end{aligned}$$

where (e) is by the facts that (i) $\inf_{b\in B^*} L_n(b,\lambda) = L_n(\beta,\lambda)$ a.s., as implied by Assumption 2.2, $\vartheta(z,b,\lambda) = \vartheta(z,\beta,\lambda)$ a.s. for all $b \in B^*$; (ii) $L_n(\beta,\lambda) \geq \sup_{b\in B} L_n(b,\lambda) - \sup_{b\in B} |L_n(b,\lambda) - L_n(\beta,\lambda)|$. (f) is because of our assumption that $\sup_{b\in B} [L_n(b,\lambda) - L_n(\beta,\lambda)]/\gamma_n \xrightarrow{p} 0$ and $\sup_{b\in B} |L_n(b,\lambda_n) - L_n(b,\lambda)|/\gamma_n \xrightarrow{p} 0$; and also note that both $\sup_{b\in B} [L_n(b,\lambda_n) - L_n(\beta,\lambda_n)]$ and γ_n are non-negative. Thus, by the definition of B_n^f, it follows that $B^* \subseteq B_n^f$ w.p.a.1. □

A.2. **Proof of Proposition 3.1.**

Proof. Note that the point identification does not play a particular role in the proof of Kim and Pollard (1990, Lemma 4.1). We replace θ_0 in their proof with our identified set B^* (therefore $A(n,j) =$

$\{b : (j-1)n^{-1/3} \leq \rho(b, B^*) \leq jn^{-1/3}\}$) and using the fact that $U(b, \lambda) = U_n(b, \lambda) = 0$ for all $b \in B^*$ (by the definitions of B^*, U_n and U). Note also that the function class has an integrable envelope function (because of indicator functions). Lemma B.5 verifies the maximal inequality for the $\mathscr{U}$–process objective function.

To apply Kim and Pollard (1990, Lemma 4.1), it remains to verify that when n is large, $\mathbb{E}[\overline{G}^2] \leq CR_g$, where R_g is the constant defined in Equation (11) and C is any finite positive constant. Recall that

$$\bar{g}_n(z_i, b, \lambda) = \underbrace{\left| \frac{1}{2} \{ \mathbb{E}[g^*(z_i, z_j, b, \lambda)|z_i] - \mathbb{E}[g^*(z_i, z_j, \beta, \lambda)|z_i] \} \right|}_{\equiv \bar{g}_{n1}} + \underbrace{\left| \frac{1}{2} \{ \mathbb{E}[g^*(z_j, z_i, b, \lambda)|z_i] - \mathbb{E}[g^*(z_j, z_i, \beta, \lambda)|z_i] \} \right|}_{\equiv \bar{g}_{n2}}.$$

where $\bar{g}_{n1}$ and $\bar{g}_{n2}$ have their own point–wise upper bounds (with respect to index b), respectively. In particular, for $\bar{g}_{n1}$,

$$|\bar{g}_{n1}(z_i, b, \lambda)| \leq \frac{1}{2} \sup_z |\xi(z)| \int |K(u)| du \times \Big[\left| \mathrm{sgn}\{x_i' b + v_{1i}\} - \mathrm{sgn}\{x_i' \beta + v_{1i}\} \right| + \left| \mathrm{sgn}\{x_i' b + v_{0i}\} - \mathrm{sgn}\{x_i' \beta + v_{0i}\} \right| \Big] \equiv \overline{G}_1(z_i, b, \lambda);$$

and for the $\bar{g}_{n2}$ part, let $\tilde{\vartheta}(z_j, b, \lambda) = \vartheta(z_j, b, \lambda) - \vartheta(z_j, \beta, \lambda)$ and $\bar{f} = \sup_z f(z)$,

$$|\bar{g}_{n2}(z_i, b, \lambda)| = \left| (P(z_i) - 0.5) \mathbb{E} \left\{ \frac{1}{h^p} K\left(\frac{z_j - z_i}{h} \right) \tilde{\vartheta}(z_j, b, \lambda) | z_i \right\} \right|$$
$$= \left| (P(z_i) - 0.5) \int K(u) f(z_i + hu) \tilde{\vartheta}(z_i + hu, b, \lambda) du \right| \leq \bar{f} \left| \int K(u) \tilde{\vartheta}(z_i + hu, b, \lambda) du \right| \equiv \overline{G}_2(z_i, b, \lambda).$$

It is then sufficient to verify that $\mathbb{E}[\overline{G}_1^2]$ and $\mathbb{E}[\overline{G}_2^2]$ are bounded by $CR_g/2$. Similar to the analysis in (Kim and Pollard, 1990, Example 6.4), this is true for $\mathbb{E}[\overline{G}_1^2]$. For $\mathbb{E}[\overline{G}_2^2]$, let $\bar{K} = \sup_u K(u)$,

$$\mathbb{E}[\overline{G}_2^2] \leq \bar{f}^2 \int_z \left(\int_u K(u) \left| \tilde{\vartheta}(z_i + hu, b, \lambda) \right| du \right)^2 f(z)dz$$

$$\leq \bar{f}^2 \int_z \int_u K^2(u) \left| \tilde{\vartheta}(z_i + hu, b, \lambda) \right|^2 du f(z)dz = \bar{f}^2 \bar{K} \int_u \int_z \left| \tilde{\vartheta}(z_i + hu, b, \lambda) \right|^2 K(u) f(z) du dz$$

$$\leq \bar{f}^2 \bar{K} \int_u \int_z \left[\left| \mathrm{sgn}\{x'b + \ u_x'(hb) + \nu_1 + hu_{\nu_1}\} - \mathrm{sgn}\{x'\beta + \nu_1 + u_x'(h\beta) + \nu_1 + hu_{\nu_1}\} \right| \right.$$

$$\left. + \left| \mathrm{sgn}\{x'b + \ u_x'(hb) + \nu_0 + hu_{\nu_0}\} - \mathrm{sgn}\{x'\beta + \nu_0 + u_x'(h\beta) + \nu_0 + hu_{\nu_0}\} \right| \right] K(u) f(z) du dz$$

where $u = (u_x', u_{\nu_1}, u_{\nu_0})'$. Treating (z, u) as a $\mathbb{R}^{2p}$ dimensional random vector, and following the the same argument in (Kim and Pollard, 1990, Example 6.4), the right hand side integral is bounded by the order $\|b - \beta\| + \|hb - h\beta\|$, and therefore bounded by $CR_g/2$ when n is large. □

A.3. **Proof of Proposition 3.2.**

Proof. First note that $U_n(b, \lambda_n) - U_n(b, \lambda) = [L_n(b, \lambda_n) - L_n(b, \lambda)] - [L_n(\beta, \lambda_n) - L_n(\beta, \lambda)]$. Thus $\sup_{b \in B} |U_n(b, \lambda_n) - U_n(b, \lambda)| \leq 2 \sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)|$.

Moreover, for r satisfying Assumption 3.6, define $\varphi_n(z_i, r) = 1\left\{ |\xi_n(z_i)| > n^{-\frac{1}{3} - \frac{r}{2}} \right\}$, then

$$\sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)| = \sup_{b \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \xi_n(z_i) \times \phi(z_i, b) \times [\lambda_n(z_i) - \lambda(z_i)] \right\} \right|$$

$$\leq \frac{2}{n} \sum_{i=1}^{n} \left\{ |\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times \varphi_n(z_i, r) \right\}$$

$$+ \frac{2}{n} \sum_{i=1}^{n} \left\{ |\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times [1 - \varphi_n(z_i, r)] \right\}. \quad (10)$$

The conclusion follows from Lemmas B.8 and B.10. □

A.4. **Proof of Theorem 2.** First, by Theorem 1, $B_n^f \subseteq B_\delta^*$ w.p.a.1. for any $\delta > 0$, hence it is sufficient to consider a small neighborhood of B^*.

Part I. By Proposition 3.1, for any $\epsilon > 0$, there exists M_n of order $O_p(1)$ such that

$$U_n(b, \lambda) \leq U(b, \lambda) + \epsilon \rho (b, B^*)^2 + n^{-\frac{2}{3}} M_n^2, \quad \text{for all } b \in B.$$

Let ϵ_1, ϵ_2 and $\delta > 0$ be defined in Lemma B.11 and $\epsilon = \frac{\epsilon_2}{2} > 0$. Then for all $b \in B^*_\delta$ and n sufficiently large,

$$\frac{\epsilon_2}{2}\rho\left(b, B^*\right)^2 - \epsilon_1 h^R \rho\left(b, B^*\right) + U_n\left(b, \lambda\right) \leq n^{-\frac{2}{3}} M_n^2.$$

By definition of B_n^f, for any $\tilde{b} \in B_n^f$ we have $U_n\left(\tilde{b}, \lambda\right) \geq \sup_{b \in B} U_n\left(b, \lambda\right) - \gamma_n \geq -\gamma_n$. Then

$$\frac{\epsilon_2}{2}\rho\left(\tilde{b}, B^*\right)^2 - \epsilon_1 h^R \rho\left(\tilde{b}, B^*\right) \leq n^{-\frac{2}{3}} M_n^2 + \gamma_n,$$

which implies that

$$\frac{\epsilon_2}{2}\left[\rho\left(\tilde{b}, B^*\right) - \frac{\epsilon_1 h^R}{\epsilon_2}\right]^2 \leq n^{-\frac{2}{3}} M_n^2 + \gamma_n + \frac{\epsilon_1^2}{2\epsilon_2} h^{2R}.$$

By Assumptions 3.6 and 3.9, we have $\rho(\tilde{b}, B^*) = O_p(\sqrt{\gamma_n})$ for any $\tilde{b} \in B_n^f$.

Part II. By the proof for Theorem 1, it suffices to show $\sup_{b \in B} U_n(b, \lambda)/\gamma_n \xrightarrow{p} 0$ and $\sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)|/\gamma_n \xrightarrow{p} 0$, the latter of which is given by Proposition 3.2. For the first statement, $\mathbb{P}(B_n^f \subseteq B^*_\delta) \to 1$ for any fixed $\delta > 0$ implies that $\sup_{b \in B^*_\delta} U_n(b, \lambda_n) = \sup_{b \in B} U_n(b, \lambda_n)$ w.p.a.1.. Applying Proposition 3.2 to both sides, it follows that $\sup_{b \in B^*_\delta} U_n(b, \lambda) = \sup_{b \in B} U_n(b, \lambda)$ w.p.a.1.. By Lemma B.12, $\sup_{b \in B^*_\delta} U_n(b, \lambda)/\gamma_n \xrightarrow{p} 0$ for some $\delta > 0$. Therefore $\sup_{b \in B} U_n(b, \lambda)/\gamma_n \xrightarrow{p} 0$.

Combine Part I and II, we can conclude that $\rho_H\left(B_n^f, B^*\right) = O_p\left(\sqrt{\gamma_n}\right)$. □

A.5. **Proof of Theorem 3.** The argument follows the proof of Chernozhukov, Hong, and Tamer (2007, Theorem 3.3) and Blevins (2012, Theorem 4). Let $\eta_n = \sqrt{\gamma_n} = n^{-1/3}(2 \ln \ln n)^{1/4}$. Let $B^*_{\eta_n}$ be the η_n expansion of B^*. Under the rate condition of m, $\rho_H(B^*_{\eta_n}, B^*) = O(\eta_n) = o(m^{-1/3})$. Define

$$\hat{\ell}^-_{j,m,n} = m^{2/3} \sup_{b \in B} U_{m,j}(b, \lambda_{m,j}) - m^{2/3} \inf_{b \in B^*_{\eta_n}} U_{m,j}(b, \lambda_{m,j}),$$

and

$$\hat{\ell}^+_{j,m,n} = \sup_{\{K : \rho_H(K, B^*) \leq \eta_n\}} \left\{ m^{2/3} \sup_{b \in B} U_{m,j}(b, \lambda_{m,j}) - m^{2/3} \inf_{b \in K} U_{m,j}(b, \lambda_{m,j}) \right\}.$$

Recall that $\hat{\ell}_{j,m,n} = m^{2/3} \sup_{b \in B} U_{m,j}(b, \lambda_{m,j}) - m^{2/3} \inf_{b \in B_n} U_{m,j}(b, \lambda_{m,j})$, then $\hat{\ell}^-_{j,m,n} \leq \hat{\ell}_{j,m,n} \leq \hat{\ell}^+_{j,m,n}$ by construction, and

$$\underline{\hat{\mathcal{L}}} \equiv \frac{1}{T_n} \sum_{j=1}^{T_n} \mathbf{1}[\hat{\ell}^-_{j,m,n} \leq x] \leq \frac{1}{T_n} \sum_{j=1}^{T_n} \mathbf{1}[\hat{\ell}_{j,m,n} \leq x] \equiv \hat{\mathcal{L}} \leq \frac{1}{T_n} \sum_{j=1}^{T_n} \mathbf{1}[\hat{\ell}_{j,m,n} \leq x] \equiv \widehat{\overline{\mathcal{L}}}.$$

Consider $\widehat{\underline{\mathcal{L}}}$ first.

$$\widehat{\underline{\mathcal{L}}} \overset{(a)}{=} \mathbb{P}\left(m^{2/3} \sup_{b\in B} U_{m,j}(b,\lambda_{m,j}) - m^{2/3} \inf_{b\in B^*_{\eta n}} U_{m,j}(b,\lambda_{m,j}) \le x \right) + o_p(1)$$

$$\overset{(b)}{=} \mathbb{P}\left(m^{2/3} \sup_{b\in B} U_{m,j}(b,\lambda_{m,j}) - m^{2/3} \inf_{b\in B^*} U_{m,j}(b,\lambda_{m,j}) \le x \right) + o_p(1) \overset{(c)}{=} \mathbb{P}\left(\mathcal{C} \le x\right) + o_p(1).$$

Step (a) holds due to the fact that $\widehat{\underline{\mathcal{L}}}$ is an $\mathscr{U}$–statistic of order m and takes value from unit interval; (b) holds by Lemma B.14; (c) holds by Assumption 3.10 and Lemma B.13. By similar arguments, $\widehat{\overline{\mathcal{L}}} = \mathbb{P}\left(\mathcal{C} \le x\right) + o_p(1)$. If $\mathbb{P}(\mathcal{C} = 0) \ge 1-\alpha$, then $\lim_{n\to\infty} \mathbb{P}\left(B^* \subseteq \widehat{B}_\alpha\right) > 1-\alpha$; if $\mathbb{P}(\mathcal{C} = 0) < 1-\alpha$, then $\lim_{n\to\infty} \mathbb{P}\left(B^* \subseteq \widehat{B}_\alpha\right) = 1-\alpha$. □

Appendix B. Preliminary Lemmas

We begin with some preliminary lemmas. Lemmas B.1 and B.2 will be used to show the consistency of our estimator; Lemmas B.3 to B.12 are primarily for convergence rate; Lemmas B.13 and B.14 are for the validity of inference.

Let $\xi_n(z_i) = [P_n(z_i) - (1-\alpha)] f_n(z_i)$ and $\xi(z_i) = [P(z_i) - (1-\alpha)] f(z_i)$. Let further $\phi(z_i, b) = \text{sgn}(x_i b + v_{1i}) - \text{sgn}(x_i b + v_0)$. By definition, $|\phi(z,b)| \le 2$ for all b and z. Note that under Assumptions 2.4 and 3.1 to 3.3, it is standard in the nonparametric estimation literature that $\xi_n(z) \overset{L_2}{\to} \xi(z)$ for all $z \in \mathscr{Z}$, where L_r refers to the convergence in the r–th mean.

Lemma B.1. *Suppose that Assumptions 2.4 and 3.1 to 3.3 hold, then*

$$\sup_{b\in B} |L_n(b,\lambda) - L(b,\lambda)| = o_p(1).$$

Proof. Note that $L_n(b,\lambda)$ can be rewritten as $L_n(b,\lambda) \equiv \frac{1}{n}\sum_{i=1}^n [\xi_n(z_i) \times \vartheta(z_i,b,\lambda)]$. Let $\overline{L}_n(b,\lambda) = \frac{1}{n}\sum_{i=1}^n [\xi(z_i) \times \vartheta(z_i,b,\lambda)]$. Thus

$$\sup_{b\in B} \left|L_n(b,\lambda) - \overline{L}_n(b,\lambda)\right| \le \sup_{b\in B} \frac{1}{n}\sum_{i=1}^n \left|\xi_n(z_i) - \xi(z_i)\right| \times |\vartheta(z_i,b,\lambda)|.$$

By definition $|\vartheta(z_i, b, \lambda)| = 1$. It follows that

$$\mathbb{E} \sup_{b \in B} \left| L_n(b, \lambda) - \overline{L}_n(b, \lambda) \right| \leq \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} \left| \xi_n(z_i) - \xi(z_i) \right| \right]$$
$$= \mathbb{E} \left| \xi_n(z_1) - \xi(z_1) \right| \leq \left\{ \mathbb{E} \left[\xi_n(z_1) - \xi(z_1) \right]^2 \right\}^{1/2} \to 0,$$

where the last step comes from the fact that $\xi_n(z) \xrightarrow{L_2} \xi(z)$ for all $z \in \mathscr{Z}$.

By UWLLN, $\sup_{b \in B} \left| \overline{L}_n(b, \lambda) - L(b, \lambda) \right| = o_p(1)$. Hence, $\sup_{b \in B} \left| L_n(b, \lambda) - L(b, \lambda) \right| = o_p(1)$. □

Lemma B.2. *Suppose that Assumptions 2.4 and 3.1 to 3.3 hold, then*

$$\sup_{b \in B} \left| L_n(b, \lambda_n) - L_n(b, \lambda) \right| = o_p(1).$$

Proof. Because

$$|L_n(b, \lambda_n) - L_n(b, \lambda)| \leq \frac{1}{n} \sum_{i=1}^{n} |\xi_n(z_i)| \times |\vartheta(z_i, b, \lambda_n) - \vartheta(z_i, b, \lambda)|$$
$$\leq \frac{1}{n} \sum_{i=1}^{n} |\xi_n(z_i)| \times |\phi(z_i, b)| \times \left| 1\left\{ \xi_n(z_i) > 0 \right\} - 1\left\{ \xi(z_i) > 0 \right\} \right|$$
$$\leq \frac{2}{n} \sum_{i=1}^{n} |\xi_n(z_i)| \times \left| 1\left\{ \xi_n(z_i) > 0 \right\} - 1\left\{ \xi(z_i) > 0 \right\} \right|.$$

Note that the RHS does not depend on b, then

$$\mathbb{E} \left\{ \sup_{b \in B} |L_n(b, \lambda_n) - L_n(b, \lambda)| \right\} \leq \mathbb{E} \left\{ \frac{2}{n} \sum_{i=1}^{n} |\xi_n(z_i)| \times \left| 1\left\{ \xi_n(z_i) > 0 \right\} - 1\left\{ \xi(z_i) > 0 \right\} \right| \right\}$$
$$= \mathbb{E} \left\{ |\xi_n(z_i)| \times \left| 1\left\{ \xi_n(z_i) > 0 \right\} - 1\left\{ \xi(z_i) > 0 \right\} \right| \right\}.$$

Thus, it suffices to show for any $z \in \mathscr{Z}$, there is

$$\mathbb{E} \left\{ |\xi_n(z)| \times \left| 1\left\{ \xi_n(z) > 0 \right\} - 1\left\{ \xi(z) > 0 \right\} \right| \right\} \to 0.$$

Because

$$|\xi_n(z)| \times \left| 1\left\{ \xi_n(z) > 0 \right\} - 1\left\{ \xi(z) > 0 \right\} \right| \leq |\xi_n(z) - \xi(z)| \times \left| 1\left\{ \xi_n(z) > 0 \right\} - 1\left\{ \xi(z) > 0 \right\} \right|,$$

then

$$\mathbb{E}\left\{\sup_{b\in B}|L_n(b,\lambda_n)-L_n(b,\lambda)|\right\}\le \mathbb{E}\,|\xi_n(z)-\xi(z)|\le\left\{\mathbb{E}\left[\xi_n(z)-\xi(z)\right]^2\right\}^{1/2}\longrightarrow 0. \quad \square$$

Let $\bar{g}_n(z_i,b,\lambda)=\mathbb{E}\left[\tilde{g}_n(z_i,z_j,b,\lambda)|z_i\right]$. Given λ, and a constant $R_g>0$, for each n, define a class of functions mapping from $\mathscr{Z}$ to $\mathbb{R}$ indexed by $b\in B$, where $\mathscr{Z}$ is the suppport of z:

$$\overline{\mathscr{G}}_n\equiv\left\{\bar{g}_n(\cdot;b,\lambda):b\in B,\rho(b,B^*)\le R_g\right\}. \tag{11}$$

Let $\overline{G}$ be its envelop function. Recall that $g_n(z_i,z_j,b,\lambda)=\frac{1}{2}\{g^*(z_i,z_j,b,\lambda)+g^*(z_j,z_i,b,\lambda)\}$. Under Assumptions 3.1 and 3.3, $\bar{g}_n(z_i,b,\lambda)$ is the sum of a continuous function and indicator functions of finite dimensional parameter b and hence belongs to VC–class (Kosorok, 2008, Lemmas 9.6, 9.9 and 9.12).

To simplify the notation, we write $\bar{g}_i(b)$ and $\tilde{g}_{ij}(b)$ for $\bar{g}_n(z_i,b,\lambda)$ and $\tilde{g}_n(z_i,z_j,b,\lambda)$, respectively. Our proofs of Lemmas B.3 to B.5 is modified from the proof in Sherman (1994a).

Lemma B.3. *Suppose that Assumptions 2.1 to 2.4 and 3.1 to 3.3 hold. Then for some finite constant J,*

$$\sqrt{n}\mathbb{E}\left\{\sup_{b\in B}\left|\frac{1}{n}\sum_i\bar{g}_i(b)-\mathbb{E}\bar{g}_1(b)\right|\right\}\le J\sqrt{\mathbb{E}\overline{\mathrm{G}}^2}. \tag{12}$$

Proof. Since function class $\overline{\mathscr{G}}_n$ is a VC class, by maximal inequality 3.1 in Kim and Pollard (1990), it follows that there exists some universal constant J such that

$$\sqrt{n}\mathbb{E}\left\{\sup_{b\in B}\left|\frac{1}{n}\sum_i\bar{g}_i(b)-\mathbb{E}\bar{g}_1(b)\right|\right\}\le J\sqrt{\mathbb{E}\overline{\mathrm{G}}^2}. \qquad \square$$

Let $\tilde{\tilde{g}}_{ij}(b)=\tilde{g}_{ij}(b)-\bar{g}_i(b)-\bar{g}_j(b)+\mathbb{E}\tilde{g}_{ij}(b)$. By construction, $\mathbb{E}[\tilde{\tilde{g}}_{ij}(b)|z_i]=\mathbb{E}[\tilde{\tilde{g}}_{ij}(b)|z_j]=0$ for all b. Define the degenerate class $\tilde{\tilde{\mathscr{G}}}_n=\{\tilde{\tilde{g}}_{ij}:b\in B\}$. Then in Nolan and Pollard (1987)'s terminology, the process $\frac{1}{n(n-1)}\sum_{1\le j<i\le n}\tilde{\tilde{g}}_{ij}(\cdot)$ is $\mathbb{P}$–degenerate. Since $\tilde{\tilde{g}}_{ij}$ is a sum of functions of V–C class, by Kosorok (2008, Lemmas 9.6, 9.9), $\tilde{\tilde{\mathscr{G}}}_n$ is V-C class as well.

Lemma B.4. *Suppose that Assumptions 2.1 to 2.4 and 3.1 to 3.3 hold. Then for some finite constant J,*

$$\mathbb{E}\left\{\sup_{b\in B}\sqrt{n}\left|\frac{1}{n(n-1)}\sum_{1\le j<i\le n}\tilde{\tilde{g}}_{ij}(b)\right|\right\} = O\left(n^{-1/2}h^{-p/2}\right), \tag{13}$$

Proof. Since

$$|\tilde{\tilde{g}}_{ij}(b)| \le |\tilde{g}_{ij}(b)| + |\bar{g}_i(b)| + |\bar{g}_j(b)| + |\mathbb{E}\tilde{g}_{ij}(b)| \le |\tilde{g}_{ij}(b)| + 3\overline{\mathbb{G}},$$

It follows that the class $\tilde{\tilde{\mathscr{G}}}_n$ has an envelop function $\overline{\overline{\mathbb{G}}}_n \equiv |K(\frac{z_i-z_j}{h})/h^p| + 3\sup_z|\xi(z)f(z)| \times \int|K(u)|du$. Note that $\mathbb{E}\overline{\overline{\mathbb{G}}}_n^2 = O(h^{-p})$ under Assumptions 3.1 and 3.3. Then by Theorem 6 of Nolan and Pollard (1987), there exists a finite universal constant $J' > 0$

$$\sqrt{n}\mathbb{E}\left\{\sup_{b\in B}\left|\frac{1}{n(n-1)}\sum_{1\le j<i\le n}\tilde{\tilde{g}}_{ij}(b)\right|\right\} \le \frac{J'\sqrt{\mathbb{E}\overline{\overline{\mathbb{G}}}_n^2}}{\sqrt{n}} = O\left(n^{-1/2}h^{-p/2}\right). \qquad \square$$

The constant J' is finite because $\tilde{\tilde{\mathscr{G}}}_n$ is V–C class and Nolan and Pollard (1987, Lemma 16).

Lemma B.5. *Suppose that Assumptions 2.1 to 2.4 and 3.1 to 3.3 hold. Then for some finite constant J and large n,*

$$\sqrt{n}\mathbb{E}\left\{\sup_{b\in B}|U_n(b,\lambda) - \mathbb{E}U_n(b,\lambda)|\right\} \le J\sqrt{\mathbb{E}\overline{\mathbb{G}}^2}. \tag{14}$$

Proof. Note that

$$U_n(b,\lambda) - \mathbb{E}U_n(b,\lambda) = \frac{2}{n}\sum_i \bar{g}_i(b) - 2\mathbb{E}\bar{g}_1(b) + \frac{1}{n(n-1)}\sum_{1\le j<i\le n}\tilde{\tilde{g}}_{ij}(b).$$

The last term on the right hand side is negligible by Lemma B.4 and Assumption 3.2. The conclusion then follows from Lemma B.3. $\square$

Lemma B.6 (Bernstein's tail inequality)**.** *Let $X_1, \cdots, X_n$ be independent real-valued random variables with zero mean, such that $\forall\ i$, $|X_i| \le M$ a.s. Defining $\sigma^2 = n^{-1}\sum_{i=1}^n VarX_i$ and $S_n = \sum_{i=1}^n X_i$. We obtain for any $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}|S_n| > \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}M\epsilon}\right)$$

Proof. See Van der Vaart and Wellner (1996, lemma 2.2.9). $\square$

Lemma B.7. *Suppose that Assumptions 2.1 to 2.4 and 3.4 to 3.6 hold, then for any $q > 0$*

$$n^q \mathbb{P}\left\{|\xi_n(z_1) - \xi(z_1)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\} \to 0.$$

Proof. By Dominance Convergence theorem, it suffices to show $n^q \mathbb{P}\left\{|\xi_n(z) - \xi(z)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\} \to 0$ for any $z \in \mathscr{Z}$. Because

$$\mathbb{P}\left\{|\xi_n(z) - \xi(z)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\} \leq \mathbb{P}\left\{|\xi_n(z) - \mathbb{E}\xi_n(z)| + |\mathbb{E}\xi_n(z) - \xi(z)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\}$$
$$= \mathbb{P}\left\{|\xi_n(z) - \mathbb{E}\xi_n(z)| > n^{-\frac{1}{3}-\frac{r}{2}} - |\mathbb{E}\xi_n(z) - \xi(z)|\right\} = \mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(w_i - \mathbb{E}w_i)\right| > \tau_n\right\},$$

where $w_i = [y_i - (1-\alpha)] \times K\left(\frac{z_i - z}{h}\right)$ and $\tau_n = h^p\left[n^{-\frac{1}{3}-\frac{r}{2}} - |\mathbb{E}\xi_n(z) - \xi(z)|\right]$. Thus, by Bernstein's tail inequality (Lemma B.6),

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(w_i - \mathbb{E}w_i)\right| > \tau_n\right\} 2 \leq \exp\left(-\frac{n\tau_n^2}{2\mathrm{Var}(w_i) + \frac{2}{3}\overline{K}\tau_n}\right).$$

Note that $\mathbb{E}\xi_n(z) - \xi(z) = O_p(h^R)$ by Assumptions 3.4 and 3.7. Further, under Assumption 3.6, for sufficient large n there is $0.5h^p n^{-\frac{1}{3}-\frac{r}{2}} \leq \tau_n \leq h^p n^{-\frac{1}{3}-\frac{r}{2}}$. It should also be noted that

$$\mathrm{Var}(w_i) \leq \mathbb{E}w_i^2 \leq \mathbb{E}K^2\left(\frac{z_i - z}{h}\right) \leq h^p \times \overline{K}^2 \times \int_{\mathbb{R}^p} f(z + th)dt.$$

Since $\int_{\mathbb{R}^p} f(z+th)dz \to f(z)$, then for sufficient large n, there is

$$\mathrm{Var}(w_i) \leq C \times h^p,$$

for some constant $C > 0$. Hence, we have

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(w_i - \mathbb{E}w_i)\right| > \tau_n\right\} \leq 2\exp\left(-\frac{\frac{1}{4}nh^{2p}n^{-\frac{2}{3}-r}}{2Ch^p + \frac{2}{3}Mh^p n^{-\frac{1}{3}-\frac{r}{2}}}\right) = 2\exp\left(-\frac{\frac{1}{4}nh^p n^{-\frac{2}{3}-r}}{2C + \frac{2}{3}\overline{K}n^{-\frac{1}{3}-\frac{r}{2}}}\right).$$

Again, for sufficient large n, there is $\frac{2}{3}\overline{K}n^{-\frac{1}{3}-\frac{r}{2}} \leq 1$ and $\frac{1}{4}nh^p n^{-\frac{2}{3}-2r} \geq 1$ (by Assumption 3.6). Thus for n sufficiently large,

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(w_i - \mathbb{E}w_i)\right| > \tau_n\right\} \leq 2\exp\left(-\frac{0.25nh^p n^{-\frac{2}{3}-r}}{2C+1}\right) \leq 2\exp\left(-\frac{n^r}{2C+1}\right).$$

Therefore, for any $q > 0$

$$n^q \mathbb{P}\left\{|\xi_n(z) - \xi(z)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\} \leq 2n^q \exp\left(-\frac{n^r}{2C+1}\right) \to 0. \qquad \square$$

For r satisfying Assumption 3.6, define $\varphi_n(z_i, r) = 1\left\{|\xi_n(z_i)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\}$.

Lemma B.8. *Let Assumptions 2.1 to 2.4 and 3.4 to 3.6 hold, then*

$$\frac{2}{n}\sum_{i=1}^{n}\Big[|\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times \varphi_n(z_i, r)\Big] = o_p(n^{-\frac{2}{3}})$$

Proof. Because of Assumption 3.4, there is $|\xi_n(z_i)| \leq \overline{K}/h^p$ a.s.. Also note that $\lambda_n(z_i) = 1\{\xi_n(z_i) > 0\}$ and $\lambda(z_i) = 1\{\xi(z_i) > 0\}$, then

$$\begin{aligned}
&\mathbb{E}\left\{\frac{2}{n}\sum_{i=1}^{n}\Big[|\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times \varphi_n(z_i, r)\Big]\right\} \\
&\qquad \leq \frac{2\overline{K}}{h^p}\mathbb{E}1\left\{|\xi_n(z_1) - \xi(z_1)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\} = \frac{2\overline{K}}{h^p}\mathbb{P}\left\{|\xi_n(z_1) - \xi(z_1)| > n^{-\frac{1}{3}-\frac{r}{2}}\right\}.
\end{aligned}$$

Thus, the conclusion holds by Lemma B.7. $\square$

Lemma B.9. *Let Assumptions 2.1 to 2.4 and 3.4 to 3.7 hold, then*

$$\mathbb{E}1\left\{|\xi_n(z_1)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right\} = O_p(n^{-\frac{1}{3}-\frac{r}{2}}).$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}1\left\{|\xi_n(z_1)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right\} &= \mathbb{P}\left[|\xi_n(z_1)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right] \\
&\leq \mathbb{P}\left[|\xi(z_1)| \leq 2n^{-\frac{1}{3}-\frac{r}{2}}\right] + \mathbb{P}\left[|\xi_n(z_1) - \xi(z_1)| \geq n^{-\frac{1}{3}-\frac{r}{2}}\right].
\end{aligned}$$

The first term of RHS is $O(n^{-\frac{1}{3}-\frac{r}{2}})$ by Assumption 3.7; the second term is $o(n^{-q})$ for any $q > 0$ by Lemma B.7. $\square$

Lemma B.10. *Let Assumptions 2.1 to 2.4 and 3.4 to 3.7 hold, then*

$$\frac{2}{n}\sum_{i=1}^{n}\Big\{|\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times [1 - \varphi_n(z_i, r)]\Big\} = o_p(n^{-\frac{2}{3}}).$$

Proof. Note that

$$\frac{2}{n}\sum_{i=1}^{n}\left\{ |\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times [1-\varphi_n(z_i,r)] \right\}$$
$$\leq \frac{2}{n}\sum_{i=1}^{n}\left[|\xi_n(z_i)| \times [1-\varphi_n(z_i,r)] \right\} = \frac{2}{n}\sum_{i=1}^{n}\left\{ |\xi_n(z_i)| \times 1\left\{|\xi_n(z_i)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right\} \right]$$
$$\leq \frac{2}{n}\sum_{i=1}^{n}\left[n^{-\frac{1}{3}-\frac{r}{2}} \times 1\left\{|\xi_n(z_i)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right\}\right].$$

Thus

$$\mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n}\left\{ |\xi_n(z_i)| \times |\lambda_n(z_i) - \lambda(z_i)| \times [1-\varphi_n(z_i,r)] \right\}\right)$$
$$\leq 2n^{-\frac{1}{3}-\frac{r}{2}}\mathbb{E}1\left\{|\xi_n(z_1)| \leq n^{-\frac{1}{3}-\frac{r}{2}}\right\} = o(n^{-\frac{2}{3}}),$$

where the last equality follows from Lemma B.9. □

Lemma B.11. *Let Assumptions 2.1 to 2.4, 3.4 to 3.6 and 3.8 hold, then there exists $\epsilon_1, \epsilon_2 \in \mathbb{R}^+$, such that for some $\delta > 0$, $U(b,\lambda) \leq \epsilon_1 h^R \rho(b, B^*) - \epsilon_2 \rho^2(b, B^*)$, when $b \in B^*_\delta$ and n sufficiently large.*

Proof. Let b in the δ–neighborhood of B^* and $b \notin B^*$. Note that

$$U(b,\lambda) = \mathbb{E}L_n(b,\lambda) - \mathbb{E}L_n(\beta,\lambda)$$
$$= \mathbb{E}\left\{\int \xi(z_1+uh)K(u)du \times \vartheta(z_1,b,\lambda)\right\} - \mathbb{E}\left\{\int \xi(z_1+uh)K(u)du \times \vartheta(z_1,\beta,\lambda)\right\}.$$

Let $\Gamma_n(b) = \mathbb{E}\left\{\int \xi(z_1+uh)K(u)du \times \vartheta(z_1,b,\lambda)\right\}$. Note that $\Gamma_n(b) = \Gamma_n(\beta)$ for all $b \in B^*$. For each $b \in B$, let $b^* \in B^*$, which depends on b, such that $\|b^* - b\| = \rho(b, B^*)$. Thus $U(b,\lambda) = \Gamma_n(b) - \Gamma_n(b^*)$. By Taylor expansion

$$U(b,\lambda) = \nabla_b\Gamma_n(b^*) \times (b-b^*) + (b-b^*)' \times \nabla^2_{b'b}\Gamma_n(b^\eta) \times (b-b^*)$$

where $b^\eta = \eta b^* + (1-\eta)b$ for some $\eta \in (0,1)$.

To obtain the derivative $\nabla_b \Gamma_n(b^*)$, we first look at the derivative of $\Gamma_n(b) - \mathbb{E}\{\xi(z)\vartheta(z_1, b, \lambda)\}$. Let $\zeta_n(z) = \int \xi(z+uh)K(u)du - \xi(z)$. Then

$$\nabla_b \, \mathbb{E}\{\zeta_n(z) \times \vartheta(z,b,\lambda)\} = 2\nabla_b\, \partial \int \zeta_n(z) \times \mathbf{1}(\xi(z)>0) \times \mathbf{1}(xb+v_1>0) \times f(z)dz$$
$$+ 2\nabla_b\, \partial \int \zeta_n(z) \times \mathbf{1}(\xi(z)<0) \times \mathbf{1}(xb+v_0<0)) \times f(z)dz.$$

Note that

$$\nabla_b\, \partial \int \zeta_n(z)\mathbf{1}(\xi(z)>0)\mathbf{1}(xb^*+v_1>0)f(z)dz = \mathbb{E}[\xi_n(z)\mathbf{1}(\xi(z)>0)|xb^*+v_1=0] \times f_{xb^*+v_1}(0)$$

where $f_{xb^*+v_1}$ denotes the density function of xb^*+v_1 Because $\mathbb{E}\left(\zeta_n(z)\mathbf{1}(\xi>0)|x,v^1\right) \leq C_1 h^R$ for some constant $C_1>0$ and $f_{xb^*+v_1}$ is also bounded above (because f is bounded above), it follows that $\|\nabla_b \mathbb{E}\{\zeta_n(z_1) \times \vartheta(z_1,b^*,\lambda)\}\| \leq \epsilon_1 h^R$ for some constant $\epsilon_1>0$ that does not depend on the value of b^*. Moreover, note that the first order directional derivative of $\mathbb{E}\{\xi(z_1)\times\vartheta(z_1,b,\lambda)\}$ at $b^* \in B^*$ equals to zero by the definition of B^*. Hence $\sup_{b^*\in\partial B^*} \|\nabla_b \Gamma_n(b^*)\| \leq \epsilon_1 h^R$.

Moreover, note that

$$\nabla^2_{b'b}\Gamma_n(b^\eta) = \nabla^2_{\eta\eta}\Gamma_n(\eta b^* + (1-\eta)b)$$
$$= \nabla^2_{\eta\eta}L(\eta b^* + (1-\eta)b) + \nabla^2_{\eta\eta}\, \mathbb{E}\{\zeta_n(z)\times\vartheta(z,\eta b^*+(1-\eta)b,\lambda)\}.$$

Similarly, by Assumption 3.5, we have $\nabla^2_{\eta\eta}\, \mathbb{E}\{\zeta_n(z)\times\vartheta(z,\eta b^*+(1-\eta)b,\lambda)\} \leq C_2 h^R$ for some constant $C_2>0$. Further, by Assumption 3.8, for all b^η belongs to a neighborhood of B^*, there exists some positive constant ϵ_2 such that $(b-b^*)' \times \nabla^2_{b'b}\Gamma_n(b^\eta) \times (b-b^*) \leq -\epsilon_2\|b-b^*\|^2$ for n sufficiently large. □

Lemma B.12. *Let Assumptions 2.1 to 2.4, 3.4 to 3.6 and 3.8 hold, then* $\sup_{b\in B^*_\delta} U_n(b,\lambda) = O_p(n^{-2/3})$ *for some* $\delta>0$.

Proof. First, note that $\sup_{b\in B^*_\delta} U_n(b,\lambda) \geq 0$, since $U_n(\beta,\lambda)=0$. It suffices to show that $\sup_{b\in B^*_\delta} U_n(b,\lambda) \leq O_p(n^{-2/3})$ for some $\delta>0$. By Proposition 3.1, for any $\epsilon>0$, there exist random variables $\{M_n\}$ of order $O_p(1)$ independent with b such that

$$|U_n(b,\lambda) - U(b,\lambda)| \leq \epsilon\rho(b,B^*)^2 + n^{-\frac{2}{3}}M_n^2, \quad \text{for all } b\in B.$$

Then by setting $\epsilon = \epsilon_2/2 > 0$ and Lemma B.11, for all $b \in B^*_{\delta_0}$ there is $U_n(b,\lambda) \le \epsilon_1 h^R \rho(b, B^*) - \epsilon\rho(b, B^*)^2 + n^{-\frac{2}{3}} M_n^2$, which implies that $\sup_{b\in B^*_{\delta_0}} U_n(b,\lambda) \le \frac{\epsilon_1^2 h^{2R}}{4\epsilon} + n^{-\frac{2}{3}} M_n^2 = O_p(n^{-2/3})$. $\square$

Lemma B.13. *Suppose that Assumptions 2.1 to 2.4 and 3.4 to 3.9 are satisfied, then*

$$\mathcal{C}_n = n^{2/3} \sup_{b\in B} U_n(b, \lambda_n) \xrightarrow{d} \mathcal{C} \equiv \sup_{t\in\mathbb{R}^p, \tilde{b}\in\partial B^*} \left\{ S(\tilde{b}, t) - t' V(\tilde{b}) t \right\},$$

where for every $\tilde{b} \in \partial B^*$, $S(\tilde{b}, \cdot)$ is a mean zero Gaussian process , and $V(\tilde{b})$ is a positive definite matrix. Moreover, $\mathcal{C}$ is continuous on $(0, \infty)$.

Proof. By Proposition 3.2, $\mathcal{C}_n = n^{2/3} \sup_{b\in B} U_n(b,\lambda) + o_p(1)$. It it sufficient to consider $n^{2/3} \sup_{b\in B} U_n(b,\lambda)$.

Fixing a given $\tilde{b} \in \partial B^*$, we write $\tilde{q}_{ijn}(\tilde{b}, t) = n^{1/6} \tilde{g}_{ij}(\tilde{b} + t/\sqrt[3]{n}, \lambda)$. Note that $\tilde{q}_{ijn}(\tilde{b}, 0) = 0$ for all $\tilde{b} \in \partial B$ due to normalization. For every given $\tilde{b} \in \partial B$, we can rewrite $n^{2/3} U_n(\tilde{b}, \lambda)$ as a $\mathscr{U}$–process indexed by t:

$$S_n(\tilde{b}, t) = \frac{\sqrt{n}}{n(n-1)} \sum_{i\neq j} \left\{ \tilde{q}_{ijn}(\tilde{b}, t) - \mathbb{E}[\tilde{q}_{ijn}(\tilde{b}, t)] \right\}.$$

Note that $\tilde{g}_{ijn}$ is a product of continuous bounded function $\xi(z)$ and sign function $\tilde{\vartheta}$, and hence is Euclidean (Nolan and Pollard, 1987, lemma 16, 18, and 19). By Nolan and Pollard (1988, Theorem 5), $S_n(\tilde{b}, \cdot)$ converges in distribution to a mean zero Guassian process $S(\tilde{b}, \cdot)$ with a bounded and nonzero covariance kernel $H(\tilde{b}, \cdot, \cdot)$ defined by

$$H(\tilde{b}, t_1, t_2) = \lim_{n\to\infty} \mathbb{E} \left\{ \mathbb{E}[\tilde{q}_{ijn}(\tilde{b}, t_1)|z_i] \mathbb{E}[\tilde{q}_{ijn}(\tilde{b}, t_2)|z_i] \right\}. \tag{15}$$

By Assumption 3.8 and following similar argument in Lemma B.11, for any $\tilde{b} \in \partial B^*$ and $\tilde{b} + t/\sqrt[3]{n} \notin B^*$, there exist a positive semidefinite matrix $V(\tilde{b})$ such that $\sqrt{n}\mathbb{E}[\tilde{q}_{ijn}(\tilde{b}, t)] = -t'V(\tilde{b})t + o(1)$. $S_n(\tilde{b}, t)$ hence converges in distribution to $S(\tilde{b}, t) - t'V(\tilde{b})t$. Note that for each $(\tilde{b}, t)$, $S(\tilde{b}, t)$ is a Gaussian random variable, hence by Davydov, Lifshits, and Smorodina (1998, Theorem 11.1), $\mathcal{C} \equiv \sup_{(\tilde{b},t)} \{S(\tilde{b}, t) - t'V(\tilde{b})t\}$ is continuously distributed over $(0, +\infty)$. The proof completes by observing $\sup_{b\in B} U_n(b, \lambda) = \sup_{\tilde{b},t} S_n(\tilde{b}, t)$. $\square$

Lemma B.14. *Let Assumptions 2.1 to 2.4 and 3.4 to 3.9 be satisfied. Let $\tilde{B}_n$ be a sequence of subset such that $\rho_H(\tilde{B}_n, B^*) = o_p(n^{-1/3})$, then*

$$n^{2/3} \inf_{b \in B^*} U_n(b, \lambda_n) - n^{2/3} \inf_{b \in \tilde{B}_n} U_n(b, \lambda_n) = o_p(1).$$

Proof. By Proposition 3.2, and the fact that $U_n(b, \lambda) = 0$ for all $b \in B^*$, it is sufficient to show that $\inf_{b \in \tilde{B}_n} U_n(b, \lambda) = o_p(n^{-2/3})$.

Define $\tilde{A}(n, j) = \left\{b : (j-1)c_n n^{-1/3} \leq \rho(b, B^*) \leq j c_n n^{-1/3}\right\}$ for some $c_n \downarrow 0$. Then $\tilde{B}_n \subseteq \cup_{j=1}^{\infty} \tilde{A}_{j,j}$. We apply the same argument as in the proof of Kim and Pollard (1990, lemma 4.1) with $\tilde{A}(n, j)$ in the place of $A(n, j)$. Notice that $c_n \downarrow 0$, it follows that there exists $\tilde{M}_n = o_p(1)$ such that

$$|U_n(b, \lambda) - U(b, \lambda)| \leq \epsilon \rho(b, B^*)^2 + n^{-\frac{2}{3}} \tilde{M}_n^2, \quad \text{for all } b \in \tilde{B}_n.$$

By Lemma B.11, $\sup_{b \in \tilde{B}_n} U(b, \lambda) = o_p(n^{-2/3})$, then it follows that $\inf_{b \in \tilde{B}_n} U(b, \lambda) = o_p(n^{-2/3})$. □