Maximum score estimation of preference parameters for a binary choice model under uncertainty

Le-Yu Chen
Sokbae Lee
Myung Jae Sung

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP14/13

Maximum Score Estimation of Preference Parameters for a Binary Choice Model under Uncertainty *

Le-Yu Chen
Institute of Economics, Academia Sinica

Sokbae Lee
Department of Economics, Seoul National University

Myung Jae Sung†
School of Economics, Hongik University

18 April 2013

Abstract

This paper develops maximum score estimation of preference parameters in the binary choice model under uncertainty in which the decision rule is affected by conditional expectations. The preference parameters are estimated in two stages: we estimate conditional expectations nonparametrically in the first stage and then the preference parameters in the second stage based on Manski (1975, 1985)'s maximum score estimator using the choice data and first stage estimates. The paper establishes consistency and derives the rate of convergence of the corresponding two-stage estimator, which is of independent interest for maximum score estimation with generated regressors. The paper also provides results of some Monte Carlo experiments.

Keywords: *discrete choice, maximum score estimation, generated regressor, preference parameters, M-estimation, cube root asymptotics*
JEL Codes: C12, C13, C14.

*This work was in part supported by European Research Council (ERC-2009-StG-240910-ROMETA) and by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A5A8023573). We would like to thank a co-editor and two anonymous referees for helpful comments.

†Corresponding author. Address: School of Economics, Hongik University, 94 Wausan-ro, Mapo-Gu, Seoul, South Korea 121-791. E-mail: mjaesung@hongik.ac.kr.

1 Introduction

This paper develops a semiparametric two-stage estimator of preference parameters in the binary choice model where the agent's decision rule is affected by conditional expectations of outcomes which are uncertain at the choice making stage and the preference shocks are nonparametrically distributed with unknown form of heteroskedasticity. The pioneering papers of Manski (1991, 1993) establish nonparametric identification of agents' expectations in the discrete choice model under uncertainty when the expectations are fulfilled and conditioned only on observable variables. Utilizing this result, Ahn and Manski (1993) proposed a two-stage estimator for a binary choice model under uncertainty where agent's utility was linear in parameter and the unobserved preference shock had a known distribution. Specifically, Ahn and Manski (1993) estimated the agent's expectations nonparametrically in the first stage and then the preference parameters in the second stage by maximum likelihood estimation using the choice data and the expectation estimates. Ahn (1995, 1997) extended the two-step approach further. On one hand, Ahn (1995) considered nonparametric estimation of conditional choice probabilities in the second stage. On the other hand, Ahn (1997) retained the linear index structure of the Ahn-Manski model but estimated the preference parameters in the second stage using average derivative method hence allowing for unknown distribution of the unobservable. In principle, alternative approaches accounting for nonparametric unobserved preference shock can also be applied in the second step estimation of this framework. Well known methods include Cosslett (1983), Powell et al. (1989), Ichimura (1993), Klein and Spady (1993), and Coppejans (2001), among many others.

The aforementioned papers allow for nonparametric setting of the distribution of the preference shock. But the unobserved shock is assumed either to be independent of or to have specific dependence structure with the covariates. By contrast, Manski (1975, 1985) considered a binary choice model under the conditional median restriction and thus allowed for general form of heteroskedasticity for the unobserved shock. It is particularly important, as shown in Brown and Walker (1989), to account for heteroskedasticity in random utility models. Therefore, this paper develops the semiparametric two-stage estimation method for the Ahn-Manski model where the second stage is based on Manski (1975, 1985)'s maximum score estimator and thus can accommodate nonparametric preference shock with unknown form of heteroskedasticity.

From a methodological perspective, this paper also contributes to the literature of two-stage M-estimation method with non-smooth criterion functions. When the true parameter value can be formulated as the unique root of certain population moment equations, the problem of M-estimation can be reduced to that of Z-estimation. Chen et al. (2003) considered semiparametric non-smooth Z-estimation problem with estimated nuisance parameter, while allowing for over-identifying restrictions. Chen and Pouzo (2009, 2012) developed general estimation methods for semiparametric and nonparametric conditional moment models with possibly nonsmooth generalized residuals. For the general M-estimation problem, Ichimura and Lee (2010) assumed some degree of second-order expansion of the underlying objective function and established conditions under which one can obtain a $\sqrt{N}$-consistent estimator of the finite dimensional parameter where N is the sample size when the nuisance parameter at the first stage is estimated at a slower rate. For more recent papers on two-step semiparametric estimation, see Ackerberg et al. (2012), Chen et al. (2013), Escanciano et al. (2012, 2013), Hahn and Ridder (2013), and Mammen et al. (2013), among others. None of the aforementioned papers include the maximum score estimation in the second stage estimation.

For this paper, the second stage maximum score estimation problem cannot be reformulated as a Z-estimation problem. Furthermore, even in absence of nuisance parameter, Kim and Pollard (1990) demonstrated that the maximum score estimator can only have the cube root rate of convergence and its asymptotic distribution is non-standard. The most closely related paper is Lee and Pun (2006) who showed that m out of n bootstrapping can be used to consistently estimate sampling distributions of nonstandard M-estimators with nuisance parameters. Their general framework includes the maximum score estimator as a special case, but allowing for only parametric nuisance parameters. Therefore, established results in the two-stage estimation literature are not immediately applicable and the asymptotic theory developed in this paper may also be of independent interest for non-smooth M-estimation with nonparametrically generated covariates.

The rest of the paper is organized as follows. Section 2 sets up the binary choice model under uncertainty and presents the two-stage maximum score estimation procedure of the preference parameters. Section 3 states regularity assumptions and derives consistency and rate of convergence of the estimator. Section 4 presents Monte Carlo studies assessing finite sample performance of the estimator. Section 5 concludes this paper. Proofs of technical results along with some preliminary lemmas are given in the Appendices.

2 Maximum Score Estimation of a Binary Choice Model under Uncertainty

Suppose an agent must choose between two actions denoted by 0 and 1. The utility from choosing action $j \in \{0,1\}$ is

$$U_j = {z_j}'\beta_1 + y'\beta_2 + \varepsilon_j.$$

Realization of the random vector $(z_j, \varepsilon_j) \in R^k \times R$ is known to the agent before the action is chosen and the random vector $y \in R^p$ is realized only after the action is chosen. Random vectors (z_1, ε_1) and (z_0, ε_0) are not necessarily identical. Distribution of y depends on the chosen action and realization of a random vector $x \in R^q$. Let $E^s(\cdot|\cdot)$ denote the agent's subjective conditional expectation. Given the realization of (z_j, ε_j), the agent chooses the action d that maximizes the expected utility:

$$z_j'\beta_1 + E^s(y|x, d = j)'\beta_2 + \varepsilon_j, j \in \{0,1\}.$$

Thus the decision rule has the form

$$d = 1\left\{z'\beta_1 + [E^s(y|x, d = 1) - E^s(y|x, d = 0)]'\beta_2 > \varepsilon\right\}, \tag{2.1}$$

where $z \equiv z_1 - z_0$, $\varepsilon \equiv \varepsilon_0 - \varepsilon_1$, and $1\{\cdot\}$ is an indicator function whose value is one if the argument is true and zero otherwise.

As in Ahn and Manski (1993), suppose that expectations are fulfilled:

$$E^s(y|x, d = j) = E(y|x, d = j).$$

We assume that the researcher does not observe realization of ε and $E(y|x, d = j)$, but that of (z, x, d, y).

Let $G(x) \equiv E(y|x, d = 1) - E(y|x, d = 0)$ and let $w \equiv (z, G(x)) \in \mathcal{W} \subset R^{k+p}$, where $\mathcal{W}$ denotes the support of the distribution of w. Then, equation (2.1) can be written as

$$d = 1\{w'\beta > \varepsilon\}, \tag{2.2}$$

where $\beta \equiv (\beta_1, \beta_2)$ is a vector of unknown preference parameters. The set of assumptions leading to the binary choice model in (2.2) is equivalent to that of Ahn and Manski (1993, equations (1)-(3)).

However, in this paper we consider an important deviation from Ahn and Manski (1993)'s setup where the unobserved preference shock ε is independent of (z, x) with a known distribution function. Instead, we consider inference under a flexible specification of the unobserved model component. Following Manski (1985), we impose the restriction:

$$\text{Med}(\varepsilon|z, x) = 0. \tag{2.3}$$

The conditional median independence assumption in (2.2) allows for heteroskedasticity of unknown form, and hence, is substantially weaker than the assumption imposed in Ahn and Manski (1993). Given (2.3), the model (2.1) then satisfies

$$\text{Med}(d|z, x) = 1\{w'\beta > 0\}. \tag{2.4}$$

Let Θ denote the space of preference parameters, and let Λ_j, $j \in \{1, ..., p\}$, denote the function space of difference of conditional expectations $E(y_j|x, d = 1) - E(y_j|x, d = 0)$. Moreover, let $b \equiv (b_1, b_2)$ and $\gamma_j(x)$, $j \in \{1, ..., p\}$, denote generic elements of Θ and Λ_j, respectively. Let $\gamma(x) \equiv (\gamma_1(x), ..., \gamma_p(x))$ and $\Lambda \equiv \prod_{j=1}^{p} \Lambda_j$ be the space of γ. We refer to $\beta \equiv (\beta_1, \beta_2)$ and $G(x)$ as the true finite-dimensional and infinite-dimensional parameters.

Suppose that data consist of random samples $(z_i, x_i, d_i, y_i), i = 1, \cdots, N$. We estimate in the first stage the conditional expectations which are not observed. Let $\widehat{G}(x_i)$ denote an estimate of the difference in conditional expectations. Using the estimate $\widehat{G}$, we estimate the preference parameters β in the second stage by the method of maximum score estimation of Manski (1975, 1985). For any b and γ, define the sample score function

$$S_N(b, \gamma) \equiv \frac{1}{N} \sum_{i=1}^{N} \tau_i (2d_i - 1) 1\{z_i' b_1 + \gamma(x_i)' b_2 > 0\}, \tag{2.5}$$

where $\tau_i \equiv \tau(x_i)$ is a predetermined weight function to avoid unduly influences from estimated $G(x_i)$ at the boundaries of the support of x_i. The two-stage estimator of β is now defined as

$$\widehat{\beta} = \arg\max_{b \in \Theta} \mathrm{S}_N(b, \widehat{G}). \tag{2.6}$$

3 Consistency and Rate of Convergence of $\widehat{\beta}$

Let $F(t; b)$ and $f(t; b)$, respectively, denote the distribution and density of $w'b$. To simplify the analysis, we consider fixed trimming such that $\tau(x) = 1(x \in \mathcal{X})$, where $\mathcal{X} \subset \mathcal{R}^q$ is a predetermined, compact, and convex subset of the support of x. For any real vector b, let $\|b\|_E$ denote the Euclidean norm of b. For any p-dimensional vector of functions $h(x)$, let $\|h\|_\infty \equiv \left\| \left(\|h_1\|_{\sup}, ..., \|h_p\|_{\sup} \right) \right\|_E$ where $\|h_j\|_{\sup} \equiv \sup\{|h_j(x)| : x \in \mathcal{X}\}$ and $h_j(x)$ denote the jth component of h. Let $\widetilde{z}$ be the subvector of z excluding the component z_1. Write $b_1 = (b_{1,1}, \widetilde{b}_1)$ and $\beta_1 = (\beta_{1,1}, \widetilde{\beta}_1)$. We assume the following regularity conditions.

Assumption 1. *Assume that:*

C1. *$\Theta = \{-1, 1\} \times \Upsilon$, where Υ is a compact subspace of R^{k+p-1} and $\left(\widetilde{\beta}_1, \beta_2\right)$ is an interior point of Υ.*

C2. *(a) The support of the distribution of w is not contained in any proper linear subspace of R^{k+p}. (b) $0 < P(d = 1|w) < 1$ for almost every w. (c) For almost every $(\widetilde{z}, x)$, the distribution of z_1 conditional on $(\widetilde{z}, x)$ has everywhere positive density with respect to Lebesgue measure.*

C3. *$Med(\varepsilon|z, x) = 0$ for almost every (z, x).*

C4. *There is a positive constant $L < \infty$ such that $|F(t_1; b) - F(t_2; b)| \leq L|t_1 - t_2|$ for all $(t_1, t_2) \in R^2$ uniformly over $b \in \Theta$.*

C5. $\left\|\widehat{G} - G\right\|_\infty = o_p(1)$.

Because the scale of β for the model characterized by (2.4) cannot be identified, Assumption C1 imposes scale normalization by requiring that the absolute value of the first coefficient is unity. Assumption C2 implies that $F(t; b)$ is absolutely continuous and has density $f(t; b)$ for each $b \in \{-1, 1\} \times \Upsilon$. Assumptions C1 - C3 are standard in the maximum score estimation literature (see e.g., Manski (1985), Horowitz (1992), and Florios and Skouras (2008)). Assumption C4 is a mild condition on the distribution of the index variable $w'b$. Assumption C5 requires uniform consistency of first stage estimation. This assumption can be easily verified for standard nonparametric estimators such as series estimators (Newey (1997, Theorem 1)) and the kernel regression estimator (Bierens (1983, Theorem 1), Bierens (1987, Theorem 2.3.1) and Andrews (1995, Theorem 1)).

Given these regularity conditions, we have the following result.

Theorem 1 (Consistency)**.** *Let Assumption 1 (**C1** - **C5**) hold. Then the two-stage estimator given by (2.6) converges to β in probability as $N \longrightarrow \infty$.*

In addition to consistency, we also study rate of convergence of the estimator $\widehat{\beta}$. Let $\widetilde{w} \equiv (\widetilde{z}, G(x))$, $\widetilde{b} \equiv (\widetilde{b}_1, b_2)$ and $\widetilde{\beta} \equiv (\widetilde{\beta}_1, \beta_2)$. Let $F_\varepsilon(\cdot|z, x)$ denote the distribution function of ε conditional on (z, x) and $g_1(z_1|\widetilde{z}, x)$ denote the density function of z_1 conditional on $(\widetilde{z}, x)$. Let $p_1(\cdot, \widetilde{z}, x)$ denote the partial derivative of $P(d = 1|z, x)$ with respect to z_1. Define the following matrix

$$V \equiv \beta_{1,1} E\left[\tau p_1(-\widetilde{w}'\widetilde{\beta}/\beta_{1,1}, \widetilde{z}, x) g_1(-\widetilde{w}'\widetilde{\beta}/\beta_{1,1}|\widetilde{z}, x)\widetilde{w}\widetilde{w}'\right].$$

Since the objective function of (2.5) is non-smooth, we require the nonparametric parameter of the estimation problem should possess certain degree of smoothness to facilitate derivation of the rate of convergence result. In particular, we consider the following well known class of smooth functions (see, e.g., van der Vaart and Wellner (1996, Section 2.7.1)) : For $0 < \alpha < \infty$, let C_M^α denote the class of functions $f: \mathcal{X} \longmapsto \mathcal{R}$ with $\|f\|_\alpha \leq M$ where for any q dimensional vector of non-negative integers $k = (k_1, ..., k_q)$,

$$\|f\|_\alpha \equiv \max_{\sigma(k) \leq \underline{\alpha}} \left\|D^k f\right\|_{\sup} + \max_{\sigma(k) \leq \underline{\alpha}} \sup_{x \neq x'} \frac{\left|D^k f(x) - D^k f(x')\right|}{\|x - x'\|_E^{\alpha - \underline{\alpha}}}$$

where $\sigma(k) \equiv \sum_{j=1}^q k_q$, $\underline{\alpha}$ denotes the greatest integer smaller than α, and D^k is the differential operator

$$D^k \equiv \frac{\partial^{\sigma(k)}}{\partial x_1^{k_1} \cdots \partial x_q^{k_q}}.$$

Given the norm $\|\cdot\|_\alpha$, for any p-dimensional vector of functions $h(x)$, let $\|h\|_{\alpha,p} \equiv \left\|\left(\|h_1\|_\alpha, ..., \|h_p\|_\alpha\right)\right\|_E$ where $h_j(x)$ denote the jth component of h. Note that $\|\cdot\|_{\alpha,p}$ is a stronger norm than $\|\cdot\|_\infty$ used in condition C5 for the uniform consistency of the first stage estimator.

The regularity conditions imposed for the convergence rate result are stated as follows.

Assumption 2. *Assume that:*

C6. *The support of $\widetilde{z}$ is bounded.*

C7. *There is a positive constant $\overline{B} < \infty$ such that (i) for every z_1 and for almost* *every* $(\widetilde{z}, x)$,

$$g_1(z_1|\widetilde{z}, x) < \overline{B}, |\partial g_1(z_1|\widetilde{z}, x)/\partial z_1| < \overline{B}, \text{ and } \left|\partial^2 g_1(z_1|\widetilde{z}, x)/\partial z_1^2\right| < \overline{B},$$

and (ii) for non-negative integers i *and* j *satisfying* $i + j \leq 2$,

$$\left|\partial^{i+j} F_\varepsilon(t|z, x)/\partial t^i \partial z_1^j\right| < \overline{B}$$

for every t *and* z_1 *and for almost every* $(\widetilde{z}, x)$.

C8. *All elements of the vector* $\widetilde{w}$ *have finite third absolute moments.*

C9. *The matrix* V *is positive definite.*

C10. *For each* $j \in \{1, ..., p\}$, $\Lambda_j = C_M^\alpha$ *for some* $\alpha \geq 2q$ *and* $M < \infty$.

C11. $\left\|\widehat{G} - G\right\|_{\alpha,p} = O_p(\varepsilon_N)$ *where* ε_N *is a non-stochastic positive real sequence such that* $N^{1/3}\varepsilon_N \leq 1$ *for each* N.

Assumption C6 is standard in deriving asymptotic properties of Manski's maximum score estimator (see, e.g. Kim and Pollard (1990), pp. 213 - 216). Assumption C7 requires some smoothness of the density $g_1(z_1|\widetilde{z}, x)$ and the distribution $F_\varepsilon(t|z, x)$. Assumption C8 is mild. Since $-V$ corresponds to the second order derivative of $E[S_N(b, \gamma)]$ with respect to $\widetilde{b}$ evaluated at true parameter values, Assumption C9 is analogous to the classic condition of Hessian matrix being non-singular in the M-estimation framework. Assumption C10 imposes smoothness for the nonparametric parameter γ and hence helps to control complexity of the space Λ.

Assumption C11 requires that the first stage estimator should converge under the norm $\|\cdot\|_{\alpha,p}$ at a rate no slower than $N^{-1/3}$. Note that convergence of $\widehat{G}$ to G in the norm $\|\cdot\|_{\alpha,p}$ also implies uniform convergence of derivatives of $\widehat{G}$ to those of G. For integer-valued $\alpha > 0$, Assumption C11 is fulfilled provided that for vector of non-negative integers $k = (k_1, ..., k_q)$ that satisfies $\sigma(k) \leq \alpha$,

$$\left\|D^k \widehat{G}_{t,j} - D^k G_{t,j}\right\|_{\sup} = O_p(\varepsilon_N) \tag{3.1}$$

where $\widehat{G}_{t,j}(x)$ denotes the estimate of $G_{t,j}(x) \equiv E(y_j|x, d = t)$ for $(t, j) \in \{0, 1\} \times \{1, ..., p\}$. The condition (3.1) can also be verified for series estimators (Newey (1997, Theorem 1)) and the kernel regression estimator (Andrews (1995, Theorem 1)).

Theorem 2 (Rate of Convergence)**.** *In addition to Assumption 1 (**C1** - **C5**), let Assumption 2 (**C6** - **C11**) also hold. Then* $\left\|\widehat{\beta} - \beta\right\|_E = O_p(N^{-1/3})$.

Note that if G were priorly known to the researcher, the preference parameters β could be estimated using covariates w and the resulting maximum score estimator would have the cube root rate of convergence (Kim and Pollard (1990)). In the case of unknown G, Theorem 2 implies that the two-stage estimator $\widehat{\beta}$ retains the same convergence rate as its corresponding infeasible estimator.

We conclude this section by making some remarks on the asymptotic distribution of the two-stage estimator $\widehat{\beta}$. Without the first stage estimation, Kim and Pollard (1990) obtained the limiting distribution of the maximum score estimator. In view of this, we conjecture the limiting distribution of our proposed estimator of β might be the same as that of Kim and Pollard (1990), as long as the first stage estimator converges uniformly in probability at a sufficiently faster rate than $N^{-1/3}$ with other regularity conditions. Once we show this, the inference on β can be carried out by subsampling (Delgado et al. (2001)) since the standard bootstrap cannot be used to estimate the distribution of the maximum score estimator consistently (Abrevaya and Huang (2005)). There does not seem to exist a known result on nonstandard M-estimation with nonparametrically generated regressors. It is thus a future research topic to establish the limiting distribution of our estimator and more generally to develop a general approach for nonstandard M-estimation with nonparametrically generated nuisance parameters.

4 Monte Carlo Simulations

We adopt the following DGP in simulation study of the two-stage maximum score estimator:

$$d = 1\{\beta_0 + z\beta_1 + G(x)\beta_2 > \varepsilon\},$$

where $x = (x_1, x_2)$, $G(x) = E(y|x, d = 1) - E(y|x, d = 0)$, $z \sim Logistic$, $x_1 \sim U(-1, 1)$, $x_2 \sim Beta(2, 2)$ and $\varepsilon|(z, x) \sim N(0, 1 + z^2 + x_1^2 + x_2^2)$. The scalar random variable y is generated according to

$$y = d(\gamma_{01} + \gamma_{11}x_1 + \gamma_{21}x_2 + u_1) + (1 - d)(\gamma_{00} + \gamma_{10}x_1 + \gamma_{20}x_2 + u_0), \tag{4.1}$$

where (u_1, u_0) are independent of (x, z, ε) and are jointly normally distributed with $E(u_1) = E(u_0) = 0$, $Var(u_1) = Var(u_0) = 1$, and $Cov(u_1, u_0) = \rho$. Given (4.1),

$$G(x) = \gamma_{01} - \gamma_{00} + (\gamma_{11} - \gamma_{10})x_1 + (\gamma_{21} - \gamma_{20})x_2.$$

The true parameter values are specified in Table 1.

We compare infeasible single-stage estimator using $(z, G(x))$ as regressors and also the feasible two-stage estimator using $(z, \widehat{G}(x))$ as regressors. We consider both parametric and nonparametric first stage estimators. For the former, we estimate $E(y|x, d = j)$ by running OLS of y on x using $d = j$ subsamples. For the latter, we implement Nadaraya-Watson kernel regression estimators. The nonparametric estimators of $E(y|x, d = j)$, $j \in \{0, 1\}$ are thus constructed as

$$\frac{\sum_{i=1}^{N} y_i K(\widehat{\Omega}_j^{-1/2} h_N^{-1} (x - x_i)) 1\{d_i = j\}}{\sum_{i=1}^{N} K(\widehat{\Omega}_j^{-1/2} h_N^{-1} (x - x_i)) 1\{d_i = j\}} \tag{4.2}$$

where $x_i = (x_{1,i}, x_{2,i})$, $\widehat{\Omega}_j$ is the diagonal matrix whose kth diagonal element is the estimated variance of $x_{k,i}$ conditional on $d_i = j$, and h_N is a deterministic bandwidth sequence. Here, $K(.)$ is a multivariate kernel function of the 12th order (see, e.g., Bierens (1987, p. 112) and Andrews (1995, p. 567)) such that

$$K(x) \equiv \sum_{m=1}^{6} a_m b_m^{-2} \exp\left[-x'x/(2b_m^2)\right],$$

where the constants (a_m, b_m), $m \in \{1, ..., 6\}$ satisfy

$$\sum_{m=1}^{6} a_m = 1 \text{ and } \sum_{m=1}^{6} a_m b_m^{2l} = 0 \text{ for } l \in \{1, ..., 5\}. \tag{4.3}$$

We specify $b_m = m^{-1/2}$ and then solve a_m as solution of the system of linear equations (4.3). The bandwidth h_N is set to be $cN^{-1/36}$ with $c \in \{3, 3.5, 4, 4.5, 5, 5.5, 6\}$. As noted by Bierens (1987, p. 113), the choice of the constants (a_m, b_m) for the kernel function is less crucial since its effect on the asymptotic variance of the conditional mean estimator can be captured via the scale constant c associated with the bandwidth h_N. By Theorem 1(b) of Andrews (1995), the resulting first stage estimator (4.2) has the convergence property required in (3.1) with $\sigma(k) \leq 4$ and $\varepsilon_N = N^{-1/3}$, thus fulfilling regularity conditions C5 and C11 of Section 3.

To implement the second-stage estimator using nonparametric first stage estimators, we trim the data by setting $\tau_i = 1\{|x_{1i}| \leq 1-\varepsilon, \varepsilon \leq x_{2i} \leq 1-\varepsilon\}$ where τ_i is the weight introduced in (2.5) and ε is set to be 0.01. The estimates of β_0, β_1 and β_2 are obtained using grid search method. Since the model (2.2) allows for identification of preference parameters only up to scale normalization, we report simulation results of the estimated ratio $\widehat{\lambda} \equiv \widehat{\beta}_2/\widehat{\beta}_1$.

Let $\widehat{\lambda}_{Single}$, $\widehat{\lambda}_{OLS}$ and $\widehat{\lambda}_{Kernel}$ respectively denote the estimators $\widehat{\lambda}$ that are constructed based on the infeasible single-stage, two-stage (OLS first stage) and two-stage (kernel regression first stage) preference parameter estimators. We compute bias, median, root mean squared error (RMSE), mean absolute deviation (mean AD) and median absolute deviation (median AD) of these estimators based on 1000 simulation repetitions. Table 2 presents simulation results of $\widehat{\lambda}_{Single}$ and $\widehat{\lambda}_{OLS}$ and Table 3 gives those of $\widehat{\lambda}_{Kernel}$ for various values of the bandwidth parameter c. We find that there seems to be systematically downward bias among all the simulation configurations including the infeasible single-stage estimation cases. However, magnitude of the bias diminishes as sample size increases. The precision in terms of RMSE, mean AD and median AD of the two-stage estimators $\widehat{\lambda}_{OLS}$ and $\widehat{\lambda}_{Kernel}$ is quite close to that of the infeasible single-stage estimator $\widehat{\lambda}_{Single}$. We notice that simulation results of $\widehat{\lambda}_{Kernel}$ do not appear to be very sensitive with respect to choice of the bandwidth parameter c though setting c to be around 4.5 tends to yield better overall finite sample performance. In short, our proposed two-stage maximum score estimator seems to work well in the simulations.

5 Conclusions

This paper has developed maximum score estimation of preference parameters in the binary choice model under uncertainty in which the decision rule is affected by conditional expectations. The estimation procedure is implemented in two stages: we estimate conditional expectations nonparametrically in the first stage and obtain the maximum score estimate of the preference parameters in the second stage using the choice data and first stage estimates. The paper has established consistency and the rate of convergence of the corresponding two-stage estimator, which is of independent interest for non-smooth M-estimation with generated regressors.

It would be an alternative approach to develop the second stage estimator using Horowitz (1992)'s smoothed maximum score estimator or using a Laplace estimator proposed in Jun, Pinkse, and Wan (2013). These alternative methods would produce faster convergence rates but require extra tuning parameters. Alternatively, we might build the second stage estimator based on Lewbel (2000), who introduced the idea of a special regressor satisfying certain conditional independence restriction. These are interesting future research topics.

A Proof of Consistency

Recall that $w = (z, G(x))$ and $S_N(b, \gamma)$ is the sample score function defined by (2.5). We first state and prove a preliminary lemma that will be invoked in proving Theorem 1 of the paper.

Lemma 1. *Under Assumptions C1, C4 and C5,*

$$\sup_{b\in\Theta}\left|S_N(b,\widehat{G})-S_N(b,G)\right| \xrightarrow{p} 0. \tag{A.1}$$

Proof of Lemma 1. Note that

$$\left|S_N(b,\widehat{G})-S_N(b,G)\right| \leq \frac{1}{N}\sum_{i=1}^{N}\tau_i 1\left\{\left|(\widehat{G}(x_i)-G(x_i))'b_2\right| \geq |w_i'b|\right\}. \tag{A.2}$$

By Assumption C1, $\|b_2\|_E < B_2$ for some finite positive constant B_2. Therefore, the right-hand side of the inequality (A.2) is bounded above by

$$\tilde{\Gamma}_N \equiv P_N\left(\tau=1, B_2\left\|\widehat{G}-G\right\|_\infty \geq |w'b|\right), \tag{A.3}$$

where P_N denotes the empirical probability. Note that the term (A.3) is further bounded above by

$$\Gamma_N \equiv P_N\left(B_2\left\|\widehat{G}-G\right\|_\infty \geq |w'b|\right). \tag{A.4}$$

Let E_η denote the event $\left\|\widehat{G}-G\right\|_\infty < \eta$ for some $\eta > 0$. Then given $\epsilon > 0$,

$$\begin{aligned} & P(\sup_{b\in\Theta}\Gamma_N > \epsilon) \leq P(\sup_{b\in\Theta}\Gamma_N > \epsilon, E_\eta) + P(E_\eta^c) \\ \leq\ & P\left[\sup_{b\in\Theta} P_N\left(B_2\eta \geq |w'b|\right) > \epsilon\right] + P(E_\eta^c). \end{aligned}$$

By Assumption C5, $P(E_\eta^c) \longrightarrow 0$ as $N \longrightarrow \infty$. Hence, to show (A.1), it remains to establish that as $N \longrightarrow \infty$,

$$P\left[\sup_{b\in\Theta} P_N\left(B_2\eta \geq |w'b|\right) > \epsilon\right] \longrightarrow 0. \tag{A.5}$$

Note that by Assumption C4, $P\left(B_2\eta \geq |w'b|\right) \leq 2LB_2\eta$. Therefore, we have that

$$\begin{aligned} & P\left[\sup_{b\in\Theta} P_N\left(B_2\eta \geq |w'b|\right) > \epsilon\right] \\ \leq\ & P\left[\sup_{b\in\Theta}\left|P_N\left(B_2\eta \geq |w'b|\right) - P\left(B_2\eta \geq |w'b|\right)\right| > \epsilon - 2LB_2\eta\right], \end{aligned} \tag{A.6}$$

where η is taken to be sufficiently small such that $\epsilon - 2LB_2\eta > 0$ for the given ϵ. By Lemma 9.6, 9.7 (ii) and 9.12 (i) of Kosorok (2008), the family of sets $\{B_2\eta \geq |w'b|\}$ for $b \in \Theta$ forms a Vapnik-Červonenkis class. Therefore, by Glivenko-Cantelli Theorem (see, e.g. Theorem 2.4.3 of van der Vaart and Wellner (1996)), the right-hand side of (A.6) tends to zero as $N \longrightarrow \infty$. Hence, the convergence result in (A.5) holds and Lemma 1 thus follows. ∎

We now prove Theorem 1 for consistency of $\widehat{\beta}$.

Proof of Theorem 1. For any (b, γ), define

$$S(b, \gamma) \equiv E\left[\tau(2d-1)1\{z'b_1 + \gamma(x)'b_2 > 0\}\right].$$

Given Assumptions C1 - C3 and by Manski (1985, Lemma 3, p. 321), β uniquely satisfies $\beta = \arg\max_{b \in \Theta} S(b, G)$. We now look at the difference

$$\left|S_N(b, \widehat{G}) - S(b, G)\right| \leq \left|S_N(b, \widehat{G}) - S_N(b, G)\right| + \left|S_N(b, G) - S(b, G)\right|, \tag{A.7}$$

where by Lemma 1, the first term of the right-hand side of (A.7) converges to zero in probability uniformly over $b \in \Theta$, whilst by Manski (1985, Lemma 4, p. 321), the second term converges to zero almost surely uniformly over $b \in \Theta$. Therefore, we have that

$$\sup_{b \in \Theta} \left|S_N(b, \widehat{G}) - S(b, G)\right| \xrightarrow{p} 0.$$

By Lemma 5 of Manski (1985, p. 322), $S(b, G)$ is continuous in b. Given these results, Theorem 1 thus follows by application of the consistency theorem in Newey and McFadden (1994, Theorem 2.1). ∎

B Lemma on the Rates of Convergence of a Two-Stage M-Estimator with a Non-smooth Criterion Function

We first present and prove a general lemma establishing the rates of convergence of a general two-stage M-estimator under high level assumptions. In next section, we prove Theorem 2 by verifying these assumptions for the particular estimator given by (2.6) under the regularity conditions of C1 - C11.

To present a general result, let $s \mapsto m_{\theta,h}(s)$ be measurable functions indexed by parameters (θ, h). Let Θ and H be the space of parameters θ and h, respectively. Let (θ^*, h^*) denote the true parameter value. We assume $(\theta^*, h^*) \in \Theta \times H$. Let $S_N(\theta, h) \equiv \sum_{i=1}^N m_{\theta,h}(s_i)/N$ be the empirical criterion of the M-estimation problem where $(s_i)_{i=1}^N$ are i.i.d. random vectors. Suppressing the individual index, let $S(\theta, h) \equiv E[m_{\theta,h}(s)]$ be the population criterion. For a given first stage estimate $\widehat{h}$, let the estimator $\widehat{\theta}$ be constructed as

$$\widehat{\theta} = \arg\sup_{\theta \in \Theta} S_N\left(\theta, \widehat{h}\right). \tag{B.1}$$

Let $d_\Theta(\theta, \theta^*)$ and $d_H(h, h^*)$ be non-negative functions measuring discrepancies between θ and θ^*, and h and h^*, respectively. Note that d_Θ and d_H are usually related to but not necessarily the same as the metrics specified for the spaces Θ and H. Given a non-stochastic positive real sequence ε_N, define $H_N(C) \equiv \{h \in H : d_H(h, h^*) \leq C\varepsilon_N\}$. To simplify the presentation, we use the notation $\lesssim$ to denote being bounded above up to a universal constant. Define the recentered criterion

$$\widetilde{S}_N(\theta, h) \equiv (S_N(\theta, h) - S_N(\theta^*, h)) - (S(\theta, h) - S(\theta^*, h)). \tag{B.2}$$

The following lemma modifies the rate of convergence results developed by van der Vaart (1998, Theorem 5.55) and provides sufficient conditions ensuring that $\widehat{\theta}$ retains the same convergence rate as it would have if h^* were known.

Lemma 2 (Rate of convergence for a general two-stage M-estimator)**.** *For any fixed and sufficiently large $C > 0$, assume that for all sufficiently large N,*

$$\sup_{h \in H_N(C)} |S(\theta^*, h) - S(\theta^*, h^*)| \lesssim (C\varepsilon_N)^2 \tag{B.3}$$

and there is a sequence of non-stochastic functions $e_N : \Theta \times H_N(C) \longmapsto R$ such that for all sufficiently small $\delta > 0$ and for every $(\theta, h) \in \Theta \times H_N(C)$ satisfying $d_\Theta(\theta, \theta^*) \leq \delta$,

$$S(\theta, h) - S(\theta^*, h^*) + e_N(\theta, h) \lesssim -d_\Theta^2(\theta, \theta^*) + d_H^2(h, h^*), \tag{B.4}$$

$$\sup_{d_\Theta(\theta,\theta^*) \leq \delta, (\theta,h) \in \Theta \times H_N(C)} |e_N(\theta, h)| \lesssim C\delta\varepsilon_N, \tag{B.5}$$

and

$$E\left[\sup_{d_\Theta(\theta,\theta^*) \leq \delta, (\theta,h) \in \Theta \times H_N(C)} \left|\widetilde{S}_N(\theta, h)\right|\right] \lesssim \frac{\phi_N(\delta)}{\sqrt{N}}, \tag{B.6}$$

where $\phi_N(\delta)$ is a sequence of functions defined on $(0, \infty)$ and satisfies that $\phi_N(\delta)\delta^{-\alpha}$ is decreasing for some $\alpha < 2$. Suppose $d_H(\widehat{h}, h^*) = O_p(\varepsilon_N)$, $d_\Theta(\widehat{\theta}, \theta^*) = o_p(1)$ and there is a non-stochastic positive real sequence δ_N which tends to zero as $N \longrightarrow \infty$ and satisfies that $\varepsilon_N \leq \delta_N$ and $\phi_N(\delta_N) \leq \sqrt{N}\delta_N^2$ for every N. Then $d_\Theta(\widehat{\theta}, \theta^*) = O_p(\delta_N)$.

Proof. Based on the peeling technique of van der Vaart (1998, Theorem 5.55), for each natural number N, integer j and positive real M, construct the set

$$A_{N,j,M}(C) \equiv \left\{(\theta, h) \in \Theta \times H_N(C) : 2^{j-1}\delta_N < d_\Theta(\theta, \theta^*) \leq 2^j\delta_N,\ d_H(h, h^*) \leq 2^{-M} d_\Theta(\theta, \theta^*)\right\}.$$

Then we have that for any $\epsilon > 0$,

$$\begin{aligned}
& P\left(d_\Theta(\widehat{\theta},\theta^*) \geq 2^M \left(\delta_N + d_H(\widehat{h},h^*)\right), \widehat{h} \in H_N(C)\right) \\
\leq \; & P(2d_\Theta(\widehat{\theta},\theta^*) > \epsilon) + P\left((\widehat{\theta},\widehat{h}) \in \bigcup\nolimits_{j\geq M, 2^j\delta_N \leq \epsilon} A_{N,j,M}(C)\right) \\
\leq \; & P(2d_\Theta(\widehat{\theta},\theta^*) > \epsilon) + \\
& \sum\nolimits_{j\geq M, 2^j\delta_N\leq\epsilon} P\left(\sup_{(\theta,h)\in A_{N,j,M}(C)} [S_N(\theta,h) - S_N(\theta^*,h)] \geq 0\right) \qquad \text{(B.7)}
\end{aligned}$$

where the last inequality follows from the definition of $\widehat{\theta}$ given by (B.1). Since $d_\Theta(\widehat{\theta},\theta^*) = o_p(1)$, the term $P(2d_\Theta(\widehat{\theta},\theta^*) > \epsilon)$ tends to zero as $N \longrightarrow \infty$. Hence the remaining part of the proof is to bound the terms in the sum (B.7).

Let N be large enough such that (B.3) holds and choose ϵ to be small enough such that assumptions (B.4), (B.5) and (B.6) hold for every $\delta \leq \epsilon$. Note that for every sufficiently large M, if $(\theta,h) \in A_{N,j,M}(C)$, then $d_H^2(h,h^*) - d_\Theta^2(\theta,\theta^*) \lesssim -\delta_N^2 2^{2j}$ so that by (B.4),

$$S(\theta,h) - S(\theta^*,h^*) + e_N(\theta,h) \lesssim -\delta_N^2 2^{2j} \qquad \text{(B.8)}$$

and thus

$$S_N(\theta,h) - S_N(\theta^*,h) \lesssim \left[\widetilde{S}_N(\theta,h) + S(\theta^*,h^*) - S(\theta^*,h) - e_N(\theta,h)\right] - \delta_N^2 2^{2j}.$$

Therefore, by Markov inequality, each term in the sum (B.7) can be bounded above by

$$\delta_N^{-2} 2^{-2j} E\left[\sup_{(\theta,h)\in A_{N,j,M}(C)} \left|\widetilde{S}_N(\theta,h) + S(\theta^*,h^*) - S(\theta^*,h) - e_N(\theta,h)\right|\right]. \qquad \text{(B.9)}$$

By (B.3), (B.5), (B.6) and applying triangular inequality, the term (B.9) is bounded above by

$$\delta_N^{-2} 2^{-2j} \left[N^{-1/2}\phi_N(2^j\delta_N) + 2^j C\delta_N\varepsilon_N + (C\varepsilon_N)^2\right]. \qquad \text{(B.10)}$$

By the monotonicity property of the mapping $\delta \mapsto \phi_N(\delta)\delta^{-\alpha}$, we have that $\phi_N(2^j\delta_N) \leq$

$2^{j\alpha}\phi_N(\delta_N)$. Furthermore, since $\phi_N(\delta_N) \leq \sqrt{N}\delta_N^2$, the first term in the bracket of (B.10) can thus be bounded by $2^{j\alpha}\delta_N^2$. Given that $\varepsilon_N \leq \delta_N$, the term (B.10) can be further bounded above by $2^{j(\alpha-2)} + C2^{-j} + C^2 2^{-2j}$. Using this fact and the condition $\alpha < 2$, it follows that the sum (B.7) tends to zero as $M \longrightarrow \infty$.

Since $d_H(\widehat{h}, h^*) = O_p(\varepsilon_N)$, $P(\widehat{h} \in H_N(C))$ can be made arbitrarily close to 1 by choosing a sufficiently large value of C for every sufficiently large N. Therefore, Lemma 2 follows by putting together all these results and noting that $\delta_N + d_H(\widehat{h}, h^*) = O_p(\delta_N)$. ∎

C Proof of the Rate of convergence for $\widehat{\beta}$

To establish the convergence rate of $\widehat{\beta}$, we apply Lemma 2 by setting $(\theta, h) = (b, \gamma)$, $(\theta^*, h^*) = (\beta, G)$, $\Theta = \{-1, 1\} \times \Upsilon$, $H = \Lambda$, $s = (\tau, d, z, x)$ and

$$m_{b,\gamma}(s) \equiv \tau(2d-1)1\{z'b_1 + \gamma(x)'b_2 > 0\}.$$

Assumptions (B.3), (B.4), (B.5) and (B.6) of Lemma 2 are non-trivial and will be verified using primitive condition C1 - C11 of the model. Assumption (B.4) is concerned with the quadratic expansion of $S(b, \gamma)$ around (β, G) by which we obtain the functional form of $e_N(b, \gamma)$. Recall that $w = (z, G(x))$, $z = (z_1, \widetilde{z})$, $\widetilde{w} = (\widetilde{z}, G(x))$, $b_1 = (b_{1,1}, \widetilde{b}_1)$, $\beta_1 = (\beta_{1,1}, \widetilde{\beta}_1)$, $\widetilde{b} = (\widetilde{b}_1, b_2)$ and $\widetilde{\beta} = (\widetilde{\beta}_1, \beta_2)$. The following lemma will be used to establish expansion of the population criterion $S(b, \gamma)$.

Lemma 3. *Under conditions C3 and C7, the sign of* $p_1(-\widetilde{w}'\widetilde{\beta}/\beta_{1,1}, \widetilde{z}, x)$ *is the same as that of* $\beta_{1,1}$ *for almost every* $(\widetilde{z}, x)$.

Proof. Note that the model (2.2) implies that

$$P(d = 1|z, x) = F_\varepsilon(w'\beta|z, x).$$

Thus, by C7(ii), $P(d=1|z,x)$ is differentiable with respect to z_1 and

$$\frac{\partial}{\partial z_1}P(d=1|z,x) = \beta_{1,1}\frac{\partial}{\partial t}F_\varepsilon(t|z,x)\Big|_{t=w'\beta} + \frac{\partial}{\partial z_1}F_\varepsilon(t|z,x)\Big|_{t=w'\beta}.$$

Consider the mapping $z_1 \mapsto h(z_1) \equiv \frac{\partial}{\partial z_1}F_\varepsilon(t|z,x)\Big|_{t=z_1\beta_{1,1}+\widetilde{w}'\widetilde{\beta}}$. By C3, $h(-\widetilde{w}'\widetilde{\beta}/\beta_{1,1}) = 0$ for almost every $(\widetilde{z},x)$. Therefore, Lemma 3 follows from this fact and the monotonicity of $F_\varepsilon(t|z,x)$ in the argument t. ∎

By assumption C1, the space of the coefficient $b_{1,1}$ is $\{-1,1\}$ and thus $b_{1,1}=\beta_{1,1}$ when $\|b-\beta\|_E < \delta$ for δ small enough. Let $p(z,x) \equiv P(d=1|z,x)$ and

$$S_1(\widetilde{b},\gamma) \equiv E\left[\tau(2p(z,x)-1)1\{z_1\beta_{1,1}+\widetilde{z}'\widetilde{b}_1+\gamma(x)'b_2>0\}\right]. \tag{C.1}$$

We now derive the quadratic expansion of $S_1(\widetilde{b},\gamma)$ around $(\widetilde{\beta},G)$.

Lemma 4. *For sufficiently small* $\left\|\widetilde{b}-\widetilde{\beta}\right\|_E$ *and* $\|\gamma-G\|_\infty$ *and under conditions C3, C7, C8 and C9, we have that*

$$\left|S_1(\widetilde{\beta},\gamma)-S_1(\widetilde{\beta},G)\right| \lesssim \|\gamma-G\|_\infty^2$$

and there are constants $c_1>0$ *and* $c_2\geq 0$ *such that*

$$S_1(\widetilde{b},\gamma)-S_1(\widetilde{\beta},G)+e(\widetilde{b},\gamma) \leq -c_1\left\|\widetilde{b}-\widetilde{\beta}\right\|_E^2 + c_2\|\gamma-G\|_\infty^2$$

for some function $e(\widetilde{b},\gamma)$ *that satisfies*

$$\left|e(\widetilde{b},\gamma)\right| \lesssim \left\|\widetilde{b}-\widetilde{\beta}\right\|_E \|\gamma-G\|_\infty.$$

Proof. We prove Lemma 4 explicitly for the case $\beta_{1,1}=1$. Proof for the case $\beta_{1,1}=-1$ can be done by similar arguments.

Suppose now $\beta_{1,1} = 1$. Then

$$
\begin{aligned}
& S_1(\widetilde{b}, \gamma) - S_1\left(\widetilde{\beta}, G\right) \\
= \; & E\left(\tau(2p(z,x)-1)\left[1\{z_1 + \widetilde{z}'\widetilde{\beta}_1 + G(x)'\beta_2 \le 0\} - 1\{z_1 + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2 \le 0\}\right]\right).
\end{aligned}
$$

Let

$$
\begin{aligned}
\lambda(t) & \equiv \widetilde{z}'\left(\widetilde{\beta}_1 + t\left(\widetilde{b}_1 - \widetilde{\beta}_1\right)\right) + \left(G(x) + t\left(\gamma(x) - G(x)\right)\right)'\left(\beta_2 + t\left(b_2 - \beta_2\right)\right), \\
\Psi(t) & \equiv -E(\tau(2p(z,x)-1)1\{z_1 + \lambda(t) \le 0\}).
\end{aligned}
$$

The first-order and second-order derivatives of $\Psi(t)$ are derived as follows:

$$
\begin{aligned}
\Psi'(t) &= E\left(\tau\lambda'(t)\left(2p(-\lambda(t), \widetilde{z}, x) - 1\right) g_1(-\lambda(t)|\widetilde{z}, x)\right), \\
\Psi''(t) &= -E\left\{\tau\left(\lambda'(t)\right)^2\left[2p_1\left(-\lambda(t), \widetilde{z}, x\right) g_1(-\lambda(t)|\widetilde{z}, x)\right.\right. \\
& \quad \left.\left. + \left(2p(-\lambda(t), \widetilde{z}, x) - 1\right)\frac{\partial}{\partial z_1} g_1(-\lambda\left(t\right)|\widetilde{z}, x)\right]\right\} \\
& \quad + E\left(2\tau\left[(2p(-\lambda(t), \widetilde{z}, x) - 1)\right] g_1(-\lambda(t)|\widetilde{z}, x)(\gamma(x) - G(x))'(b_2 - \beta_2)\right).
\end{aligned}
$$

Then the second order expansion of $S_1(\widetilde{b}, \gamma) - S_1\left(\widetilde{\beta}, G\right)$ takes the form

$$
\Psi'(0) + \Psi''(0)/2 + o\left(\left(\max\left\{\left\|\widetilde{b} - \widetilde{\beta}\right\|_E, \|\gamma - G\|_\infty\right\}\right)^2\right)
$$

where by C7 and C8, the remainder term has the stated order uniformly over $\widetilde{b}$ and γ. Given assumption C3, it follows that $p(-\widetilde{w}'\widetilde{\beta}, \widetilde{z}, x) = 1/2$ for almost every $(\widetilde{z}, x)$. Let

$$
\kappa(\widetilde{z}, x) = 2p_1\left(-\widetilde{w}'\widetilde{\beta}, \widetilde{z}, x\right) g_1(-\widetilde{w}'\widetilde{\beta}|\widetilde{z}, x).
$$

Then we have that

$$
\begin{aligned}
\Psi'(0) + \Psi''(0)/2 &= -E\left(\tau\kappa(\widetilde{z}, x)\left(\widetilde{w}'(\widetilde{b} - \widetilde{\beta}) + (\gamma(x) - G(x))'\beta_2\right)^2\right) \\
&= -\left(A_1 + A_2 + e(\widetilde{b}, \gamma)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &\equiv (\widetilde{b}-\widetilde{\beta})' E(\tau\kappa(\widetilde{z},x)\widetilde{w}\widetilde{w}')(\widetilde{b}-\widetilde{\beta}), \\
A_2 &\equiv E\left(\tau\kappa(\widetilde{z},x)\left(\gamma(x)-G(x)\right)'\beta_2\beta_2'\left(\gamma(x)-G(x)\right)\right), \\
e(\widetilde{b},\gamma) &\equiv 2(\widetilde{b}-\widetilde{\beta})' E\left(\tau\kappa(\widetilde{z},x)\widetilde{w}\beta_2'\left(\gamma(x)-G(x)\right)\right).
\end{aligned}
$$

Under condition C9, $E(\tau\kappa(\widetilde{z},x)\widetilde{w}\widetilde{w}')$ is positive definite, so that $A_1 \geq c_1 \left\|\widetilde{b}-\widetilde{\beta}\right\|_E^2$ for some positive real constant c_1. By Lemma 3, $p_1\left(-\widetilde{w}'\widetilde{\beta},\widetilde{z},x\right) \geq 0$ and thus $\kappa(\widetilde{z},x) \geq 0$. By Cauchy-Schwarz inequality, $0 \leq A_2 \leq c_2 \|\gamma - G\|_\infty^2$, where $c_2 \equiv E(\tau\kappa(\widetilde{z},x))\|\beta_2\|_E^2 \geq 0$, and the function $e(\widetilde{b},\gamma)$ satisfies that

$$
\begin{aligned}
\left|e(\widetilde{b},\gamma)\right| &\leq 2E\left(\tau\kappa(\widetilde{z},x)\left|(\widetilde{b}-\widetilde{\beta})'\widetilde{w}\beta_2'\left(\gamma(x)-G(x)\right)\right|\right) \\
&\leq 2E\left(\tau\kappa(\widetilde{z},x)\|\widetilde{w}\|_E\right)\|\beta_2\|_E\left\|\widetilde{b}-\widetilde{\beta}\right\|_E\|\gamma-G\|_\infty.
\end{aligned}
$$

Hence Lemma 4 follows by noting that when $\left\|\widetilde{b}-\widetilde{\beta}\right\|_E$ and $\|\gamma - G\|_\infty$ are sufficiently small,

$$
\left|S_1(\widetilde{\beta},\gamma)-S_1(\widetilde{\beta},G)\right| = \left|A_2 + o\left(\|\gamma-G\|_\infty^2\right)\right| \leq c_2\|\gamma-G\|_\infty^2
$$

and

$$
\begin{aligned}
S_1(\widetilde{b},\gamma)-S_1\left(\widetilde{\beta},G\right)+e(\widetilde{b},\gamma) &\leq -A_1 + A_2 \\
&\leq -c_1\left\|\widetilde{b}-\widetilde{\beta}\right\|_E^2 + c_2\|\gamma-G\|_\infty^2.
\end{aligned}
$$

∎

We now verify assumption (B.6) of Lemma 2. Note that for δ sufficiently small, assumption C1 implies that $b_{1,1} = \beta_{1,1}$ when $\|b-\beta\|_E \leq \delta$. Therefore we can focus on analyzing (B.6) for the case of $b_{1,1} = \beta_{1,1}$ and $\left\|\widetilde{b}-\widetilde{\beta}\right\|_E \leq \delta$. For any $s = (\tau, d, z, x)$, consider the following recentered function

$$\widetilde{m}_{\widetilde{b},\gamma}(s) \equiv \tau(2d-1)\left[1\{z_1\beta_{1,1} + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2 > 0\} - 1\{z_1\beta_{1,1} + \widetilde{z}'\widetilde{\beta}_1 + \gamma(x)'\beta_2 > 0\}\right] \tag{C.2}$$

and the class of functions

$$\mathcal{F}_{\delta,\varepsilon} \equiv \left\{\widetilde{m}_{\widetilde{b},\gamma} : \left\|\widetilde{b} - \widetilde{\beta}\right\|_E \le \delta, \|\gamma - G\|_{\alpha,p} \le \varepsilon\right\}. \tag{C.3}$$

Let $\|\cdot\|_{L_r(P)}$ denote the $L_r(P)$ norm such that $\|f\|_{L_r(P)} \equiv [E(|f(\tau,d,z,x)|^r)]^{1/r}$ for any measurable function f. For any $\epsilon > 0$, let $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ denote the $L_r(P)$ - bracketing number for a given function space $\mathcal{F}$. Namely, $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of $L_r(P)$ - brackets of length ϵ required to cover $\mathcal{F}$ (see e.g., van der Vaart (1998, p. 270)). The logarithm of bracketing number for $\mathcal{F}$ is referred to as the bracketing entropy for $\mathcal{F}$. Assumption (B.6) is a stochastic equicontinuity condition concerning the complexity of the function space $\mathcal{F}_{\delta,\varepsilon}$ in terms of its envelope function and bracketing entropy. Let $M_{\delta,\varepsilon}$ denote an envelope for $\mathcal{F}_{\delta,\varepsilon}$ such that $\left|\widetilde{m}_{\widetilde{b},\gamma}(s)\right| \le \left|M_{\delta,\varepsilon}(s)\right|$ for all s and for all $\widetilde{m}_{\widetilde{b},\gamma} \in \mathcal{F}_{\delta,\varepsilon}$. The next lemma derives the envelope function $M_{\delta,\varepsilon}$.

Lemma 5. *Let δ and ε be sufficiently small. Then under conditions C1, C4 ,C6 and C10, for some real constants $a_1 > 0$ and $a_2 > 0$, we can take*

$$M_{\delta,\varepsilon} = 1\{a_1 \max\{\delta, \varepsilon\} \ge |w'\beta|\}$$

and furthermore,

$$\left\|M_{\delta,\varepsilon}\right\|_{L_2(P)} \le a_2\sqrt{\max\{\delta,\varepsilon\}}. \tag{C.4}$$

Proof. Note that

$$\begin{aligned}
&\left|\widetilde{m}_{\widetilde{b},\gamma}(\tau,d,z,x)\right| \\
&\le 1\{z_1\beta_{1,1} + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2 > 0 \ge z'\beta_1 + \gamma(x)'\beta_2 \quad \text{or} \\
&\quad z'\beta_1 + \gamma(x)'\beta_2 > 0 \ge z_1\beta_{1,1} + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2\}.
\end{aligned}$$

Under condition C6, there is a positive real constant B such that $\|\widetilde{z}\|_E < B$ with probability 1. Hence if $\left\|\widetilde{b} - \beta\right\|_E \leq \delta$ and $\|\gamma - G\|_{\alpha,p} \leq \varepsilon$, then we have that

$$\begin{aligned}
& z_1\beta_{1,1} + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2 > 0 \geq z'\beta_1 + \gamma(x)'\beta_2 \\
\iff \quad & \widetilde{z}'(\widetilde{b}_1 - \widetilde{\beta}_1) + \gamma(x)'(b_2 - \beta_2) > -\left[z'\beta_1 + \gamma(x)'\beta_2\right] \geq 0 \\
\implies \quad & \delta\left[\|\widetilde{z}\|_E + \|\gamma\|_\infty\right] \geq -\left[z'\beta_1 + \gamma(x)'\beta_2\right] \text{ and } 0 \geq w'\beta + (\gamma(x) - G(x))'\beta_2 \\
\implies \quad & w'\beta + (\gamma(x) - G(x))'\beta_2 \geq -\delta\left[\|\widetilde{z}\|_E + \varepsilon + \|G\|_\infty\right] \text{ and } \varepsilon\|\beta_2\|_E \geq w'\beta \\
\implies \quad & \delta\left[B + \varepsilon + \|G\|_\infty\right] + \varepsilon\|\beta_2\|_E \geq w'\beta \geq -\delta\left[B + \varepsilon + \|G\|_\infty\right] - \varepsilon\|\beta_2\|_E
\end{aligned}$$

Based on similar arguments, it also follows that

$$\begin{aligned}
& z'\beta_1 + \gamma(x)'\beta_2 > 0 \geq z_1\beta_{1,1} + \widetilde{z}'\widetilde{b}_1 + \gamma(x)'b_2 \\
\implies \quad & \delta\left[B + \varepsilon + \|G\|_\infty\right] + \varepsilon\|\beta_2\|_E \geq w'\beta \geq -\delta\left[B + \varepsilon + \|G\|_\infty\right] - \varepsilon\|\beta_2\|_E
\end{aligned}$$

Therefore, Lemma 5 follows by noting that for ε sufficiently small (e.g., $\varepsilon < 1$), we can take

$$M_{\delta,\varepsilon} = 1\{a_1 \max\{\delta, \varepsilon\} \geq |w'\beta|\}$$

where $a_1 \equiv 2\max\{(B + 1 + \|G\|_\infty), \|\beta_2\|_E\}$. By C1 and C10, $0 < a_1 < \infty$ and hence by C4, $\left\|M_{\delta,\varepsilon}\right\|_{L_2(P)} \leq a_2\sqrt{\max\{\delta, \varepsilon\}}$ with $a_2 \equiv \sqrt{2a_1L}$ where L is the positive constant stated in condition C4. ∎

The following lemma establishes the bound for the bracketing entropy for $\mathcal{F}_{\delta,\varepsilon}$.

Lemma 6. *Given conditions C1, C4, C6, C7, C8 and C10, we have that*

$$\log N_{[]}(\epsilon, \mathcal{F}_{\delta,\varepsilon}, L_2(P)) \lesssim \sqrt{\max\{\delta, \varepsilon\}}/\epsilon$$

for sufficiently small δ and ε and for $\epsilon \leq a_2\sqrt{\max\{\delta, \varepsilon\}}$ where a_2 is the constant stated in (C.4).

Proof. For $j \in \{1, ..., p\}$, let $\widetilde{\Lambda}_j(\varepsilon)$ and $\widetilde{\Lambda}_j\mathbf{B}_j(\delta, \varepsilon)$ be classes of functions defined as

$$\begin{aligned} \widetilde{\Lambda}_j(\varepsilon) &\equiv \left\{(\gamma_j - G_j)/\varepsilon : \|\gamma_j - G_j\|_\alpha \leq \varepsilon\right\}, \\ \widetilde{\Lambda}_j\mathbf{B}_j(\delta, \varepsilon) &\equiv \left\{(\gamma_j(x) - G_j(x))(b_{2,j} - \beta_{2,j})/\left(\varepsilon\delta\right) : \|\gamma_j - G_j\|_\alpha \leq \varepsilon, |b_{2,j} - \beta_{2,j}| \leq \delta\right\}. \end{aligned}$$

Assumption C10 implies that both $\widetilde{\Lambda}_j(\varepsilon)$ and $\widetilde{\Lambda}_j\mathbf{B}_j(\delta, \varepsilon)$ are C_1^α with $\alpha \geq 2q$. By Corollary 2.7.2 of van der Vaart and Wellner (1996, p. 157), we have that for $j \in \{1, ..., p\}$,

$$\log N_{[]}(\epsilon^2, \widetilde{\Lambda}_j(\varepsilon), L_1(P)) \lesssim \epsilon^{-2q/\alpha} \text{ and } \log N_{[]}(\epsilon^2, \widetilde{\Lambda}_j\mathbf{B}_j(\delta, \varepsilon), L_1(P)) \lesssim \epsilon^{-2q/\alpha}. \quad \text{(C.5)}$$

Note that for $s = (\tau, d, z, x)$, $\widetilde{m}_{\widetilde{b},\gamma}(s)$ defined by (C.2) can be rewritten as

$$\widetilde{m}_{\widetilde{b},\gamma}(s) = \tau d \left[1\{h(s; \widetilde{b}) > 0\} - 1\{h(s; \widetilde{\beta}) > 0\}\right] + \tau(1-d)\left[1\{h(s; \widetilde{b}) \leq 0\} - 1\{h(s; \widetilde{\beta}) \leq 0\}\right]$$

where

$$\begin{aligned} h(s; \widetilde{b}) &\equiv w'\beta + \widetilde{w}'(\widetilde{b} - \widetilde{\beta}) + (\gamma(x) - G(x))'(b_2 - \beta_2) + (\gamma(x) - G(x))'\beta_2, \\ h(s; \widetilde{\beta}) &\equiv w'\beta + (\gamma(x) - G(x))'\beta_2. \end{aligned}$$

Consider the following spaces:

$$\begin{aligned} \Theta_1 &\equiv \{\widetilde{w}'(\widetilde{b} - \widetilde{\beta}) : \left\|\widetilde{b} - \widetilde{\beta}\right\|_E \leq \delta\}, \\ \Theta_{2,j} &\equiv \left\{(\gamma_j(x) - G_j(x))(b_{2,j} - \beta_{2,j}) : \|\gamma_j - G_j\|_\alpha \leq \varepsilon, |b_{2,j} - \beta_{2,j}| \leq \delta\right\} \text{ for } j \in \{1, ..., p\}, \\ \Theta_2 &\equiv \left\{(\gamma(x) - G(x))'(b_2 - \beta_2) : \|\gamma - G\|_{\alpha,p} \leq \varepsilon, \|b_2 - \beta_2\|_E \leq \delta\right\}, \\ \Theta_{3,j} &\equiv \{(\gamma_j(x) - G_j(x))\beta_{2,j} : \|\gamma_j - G_j\|_\alpha \leq \varepsilon\} \text{ for } j \in \{1, ..., p\}, \\ \Theta_3 &\equiv \{(\gamma(x) - G(x))'\beta_2 : \|\gamma - G\|_{\alpha,p} \leq \varepsilon\}, \\ \Theta_4 &\equiv \{h(\tau, d, z, x; \widetilde{b}) - w'\beta : \|\gamma - G\|_{\alpha,p} \leq \varepsilon, \left\|\widetilde{b} - \widetilde{\beta}\right\|_E \leq \delta\}. \end{aligned}$$

Let $n_i(\epsilon) \equiv \log N_{[]}(\epsilon, \Theta_i, L_1(P))$ for $i \in \{1, 2, 3, 4\}$ and $n_{k,j}(\epsilon) \equiv \log N_{[]}(\epsilon, \Theta_{k,j}, L_1(P))$ for $(k, j) \in \{2, 3\} \times \{1, ..., p\}$. Since Θ_1 is a pointwise Lipschitz class of functions with envelope $\|\widetilde{w}\|_E \delta$. By condition C8, $E(\|\widetilde{w}\|_E)$ is finite. Thus applying Theorem 2.7.11 of van der Vaart and Wellner (1996, p. 164), we have that

$$n_1(\epsilon^2) \lesssim \log(\delta/\epsilon^2) \lesssim \sqrt{\delta}/\epsilon \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon. \tag{C.6}$$

Note that for any norm $\|\cdot\|$, any fixed real valued c, any class of functions $\mathcal{F}$, it is straightforward to verify that

$$N_{[]}(\epsilon, c\mathcal{F}, \|\cdot\|) = 1 \text{ for } c = 0$$
$$N_{[]}(\epsilon, c\mathcal{F}, \|\cdot\|) \leq N_{[]}(\epsilon/|c|, \mathcal{F}, \|\cdot\|) \text{ for } c \neq 0$$

where $c\mathcal{F} \equiv \{cf : f \in \mathcal{F}\}$.

Using this fact, we have that $n_{2,j}(\epsilon^2) = \log N_{[]}(\epsilon^2/(\varepsilon\delta), \widetilde{\Lambda}_j \mathbf{B}_j(\delta,\varepsilon), L_1(P))$ and $n_{3,j}(\epsilon^2) = 0$ for $\beta_{2,j} = 0$ and $n_{3,j}(\epsilon^2) \leq \log N_{[]}(\epsilon^2/(\varepsilon|\beta_{2,j}|), \widetilde{\Lambda}_j(\varepsilon), L_1(P))$ for $\beta_{2,j} \neq 0$. Hence for sufficiently small δ and ε (e.g., $\delta < 1$ and $\varepsilon < 1$) and by (C.5), it follows that

$$n_{2,j}(\epsilon^2) \leq \log N_{[]}((\epsilon/\sqrt{\max\{\delta,\varepsilon\}})^2, \widetilde{\Lambda}_j \mathbf{B}_j(\delta,\varepsilon), L_1(P)) \lesssim (a_2\sqrt{\max\{\delta,\varepsilon\}}/\epsilon)^{2q/\alpha}.$$

Since $\alpha \geq 2q$, we have that $n_{2,j}(\epsilon^2) \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon$ for $\epsilon \leq a_2\sqrt{\max\{\delta,\varepsilon\}}$. Using similar arguments, we can also deduce that $n_{3,j}(\epsilon^2) \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon$ for $\epsilon \leq a_2\sqrt{\max\{\delta,\varepsilon\}}$.

By preservation of bracketing metric entropy (see, e.g., Lemma 9.25 of Kosorok (2008, p. 169)), we have that for $i \in \{2,3\}$,

$$n_i(\epsilon) \leq n_{i,p}(\epsilon 2^{1-p}) + \sum\nolimits_{j=1}^{p-1} n_{i,j}(\epsilon 2^{-j}).$$

and $n_4(\epsilon) \leq n_1(\epsilon/2) + n_2(\epsilon/4) + n_3(\epsilon/4)$. Therefore by the bounds derived above, it follows that $n_2(\epsilon^2) \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon$, $n_3(\epsilon^2) \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon$ and also $n_4(\epsilon^2) \lesssim \sqrt{\max\{\delta,\varepsilon\}}/\epsilon$.

Now let $f_1^L \leq f_1^U, \ldots, f^L_{N_{[]}(\epsilon^2,\Theta_3,L_1(P))} \leq f^U_{N_{[]}(\epsilon^2,\Theta_3,L_1(P))}$ and $g_1^L \leq g_1^U, \ldots, g^L_{N_{[]}(\epsilon^2,\Theta_4,L_1(P))} \leq$

$g^U_{N_{[]}(\epsilon^2,\Theta_4,L_1(P))}$ be the ϵ^2-brackets with bracket length defined by $L_1(P)$ for the spaces Θ_3 and Θ_4, respectively. For $1 \le k \le N_{[]}(\epsilon^2, \Theta_3, L_1(P))$ and $1 \le j \le N_{[]}(\epsilon^2, \Theta_4, L_1(P))$, define

$$
\begin{aligned}
m^L_{jk}(\tau, d, z, x) &\equiv \tau d \left[1\{w'\beta + g^L_j(z,x) > 0\} - 1\{w'\beta + f^U_k(z,x) > 0\}\right] \\
&\quad + \tau(1-d)\left[1\{w'\beta + g^U_j(z,x) \le 0\} - 1\{w'\beta + f^L_k(z,x) \le 0\}\right], \\
m^U_{jk}(\tau, d, z, x) &\equiv \tau d \left[1\{w'\beta + g^U_j(z,x) > 0\} - 1\{w'\beta + f^L_k(z,x) > 0\}\right] \\
&\quad + \tau(1-d)\left[1\{w'\beta + g^L_j(z,x) \le 0\} - 1\{w'\beta + f^U_k(z,x) \le 0\}\right].
\end{aligned}
$$

Note that

$$
0 \le m^U_{jk} - m^L_{jk} \le 2\left[1\{g^L_j \le -w'\beta < g^U_j\} + 1\{f^L_k \le -w'\beta < f^U_k\}\right].
$$

Thus

$$
E\left(m^U_{jk} - m^L_{jk}\right)^2 \le 12P(g^L_j \le -w'\beta < g^U_j) + 4P(f^L_k \le -w'\beta < f^U_k). \tag{C.7}
$$

By condition C1 and given $(\widetilde{z}, x)$, the mapping $z_1 \longmapsto w'\beta$ is one-to-one. Hence by condition C7, the density of $w'\beta$ conditional on $(\widetilde{z}, x)$ is bounded and by (C.7), it then follows that $\left\|m^U_{jk} - m^L_{jk}\right\|_{L_2(P)} \lesssim \epsilon$. Moreover for each $\widetilde{m}_{\widetilde{b},\gamma} \in \digamma_{\delta,\varepsilon_N}$, there is a bracket $\left[m^L_{jk}, m^U_{jk}\right]$ in which it lies. Therefore,

$$
\log N_{[]}(\epsilon, \digamma_{\delta,\varepsilon_N}, L_2(P)) \lesssim n_3(\epsilon^2) + n_4(\epsilon^2) \lesssim \sqrt{\max\{\delta, \varepsilon\}}/\epsilon.
$$

∎

Replacing (θ, h) and θ^* with $((\beta_{1,1}, \widetilde{b}), \gamma)$ and $(\beta_{1,1}, \widetilde{\beta})$, respectively in the definition of $\widetilde{S}_N$ given by (B.2), we now verify assumption (B.6) in the next lemma.

Lemma 7. *For sufficiently small δ and ε, under conditions C1, C4, C6, C7, C8 and C10,*

$$
E\left[\sup_{\left\|\widetilde{b}-\widetilde{\beta}\right\|_E \le \delta, \|\gamma - G\|_{\alpha,p} \le \varepsilon} \left|\widetilde{S}_N(\widetilde{b}, \gamma)\right|\right] \lesssim \frac{\sqrt{\max\{\delta, \varepsilon\}}}{\sqrt{N}}.
$$

Proof. By Lemmas 5 and 6, we have that

$$\begin{aligned}
&\int_0^{\|M_{\delta,\varepsilon}\|_{L_2(P)}} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}_{\delta,\varepsilon}, L_2(P))} d\epsilon \\
\leq\ &\int_0^{a_2\sqrt{\max\{\delta,\varepsilon\}}} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}_{\delta,\varepsilon}, L_2(P))} d\epsilon \\
\lesssim\ &\sqrt{\max\{\delta,\varepsilon\}}.
\end{aligned}$$

Lemma 7 hence follows by applying Corollary 19.35 of van der Vaart (1998, p. 288). ∎

We now prove Theorem 2.

Proof of Theorem 2. We take $\delta_N = N^{-1/3}$, $d_\Theta(b,\beta) = \sqrt{c_1}\,\|b-\beta\|_E$ and $d_H(\gamma, G) = \sqrt{c_2}\,\|\gamma - G\|_{\alpha,p}$ in the application of Lemma 2, where c_1 and c_2 are real constants stated in Lemma 4.

Since $c_1 > 0$, the norm by the metric $d_\Theta(\cdot,\cdot)$ is equivalent to the Euclidean norm and thus by Theorem 1, $d_\Theta(\widehat{\beta}, \beta) = o_p(1)$. Moreover since $c_2 \geq 0$, assumption C11 implies that $d_H(\widehat{G}, G) = O_p(\varepsilon_N)$. Given assumption C1, for sufficiently small δ, we have that $b_{1,1} = \beta_{1,1}$ when $d_\Theta(b,\beta) \leq \delta$. Hence for sufficiently small δ and ε_N, by Lemma 4 and noting that $\|\cdot\|_{\alpha,p}$ is stronger than $\|\cdot\|_\infty$, assumptions (B.3), (B.4) and (B.5) hold.

By Lemma 7 and by taking C sufficiently large in the definition of $H_N(C)$ of Lemma 2, assumption (B.6) also holds with $\phi_N(\delta) = \sqrt{\max\{\delta, \varepsilon_N\}}$. Clearly, $\phi_N(\delta)\delta^{-\alpha}$ is decreasing for some $\alpha < 2$. By assumption C11, $\varepsilon_N \leq \delta_N$ and thus $\phi_N(\delta_N) \leq \sqrt{N}\delta_N^2$ for every N. Therefore, all conditions stated in Lemma 2 are fulfilled and the result of Theorem 2 hence follows. ∎

References

[1] Abrevaya, J., and Huang, J. (2005), "On the bootstrap of the maximum score estimator", *Econometrica*, 73 (4), 1175-1204.

[2] Ackerberg, D., X. Chen, and J. Hahn (2012): A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators, *Review of Economics and Statistics* 94, 481-498.

[3] Ahn, H., (1995), "nonparametric Two-Stage Estimation of Conditional Choice Probabilities in a Binary Choice Model under Uncertainty", *Journal of Econometrics*, 67, 337-378.

[4] Ahn, H., (1997), "Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors", *Econometric Theory*, 13, 3-31.

[5] Ahn, H. and C. F. Manski (1993), "Distribution Theory for the Analysis of Binary Choice under Uncertainty with Nonparametric Estimation of Expectations", *Journal of Econometrics*, 56, 291-321.

[6] Andrews, D. W. K. (1995), "Nonparametric Kernel Estimation for Semiparametric Models", *Econometric Theory*, 11, 560-596.

[7] Bierens, H. J. (1983), "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions", *Journal of the American Statistical Association*, 78, 699-707.

[8] Bierens, H. J. (1987), "Kernel Estimators of Regression Functions", in Truman F.Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, Vol. 1, Cambridge University Press, 99-144.

[9] Brown, Bryan W., Walker, Mary Beth (1989), "The random utility hypothesis and inference in demand systems", *Econometrica*, 57, 815-829.

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

[11] Chen, X. and Pouzo, D. (2009), "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals", *Journal of Econometrics*, 152, 46-60.

[12] Chen, X. and Pouzo, D. (2012), “Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals”, *Econometrica*, 80, 277-321.

[13] Chen, X., Hahn, J., Liao, Z. and Ridder G. (2013), ”Nonparametric Two-Step Sieve M Estimation and Inference”, Working paper.

[14] Coppejans, M. (2001), ”Estimation of the Binary Response Model Using a Mixture of Distributions Estimator (MOD)”, *Journal of Econometrics*, 102, 231-261.

[15] Cosslett, S. R. (1983), ”Distribution-free Maximum Likelihood Estimator of the Binary Choice Model”, *Econometrica*, 51, 765-782.

[16] Delgado, M.A., Rodriguez-Poo, J.M.,and Wolf, M. (2001), “Subsampling inference in cube root asymptotics with an application to Manski’s maximum score estimator”, *Economics Letters*, 73 (2), 241–250.

[17] Escanciano, J., D. Jacho-Chávez, and A. Lewbel (2012): Identification and Estimation of Semiparametric Two Step Models, Working Paper.

[18] Escanciano, J., D. Jacho-Chávez, and A. Lewbel (2013): Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing, Working Paper.

[19] Florios, K. and Skouras, S. (2008), ”Exact computation of max weighted score estimators”, *Journal of Econometrics*, 146, 86-91.

[20] Hahn, J., and G. Ridder (2013): The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors, *Econometrica*, 81, 315-340.

[21] Horowitz, J. (1992), ”A Maximum Score Estimator for the Binary Response Model”, *Econometrica*, 60, 505-531.

[22] Ichimura, H. (1993), ”Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models”, *Journal of Econometrics*, 58, 71-120.

[23] Ichimura, H. and Lee, S. (2010), "Characterization of the asymptotic distribution of semiparametric M-estimators", *Journal of Econometrics*, 159, 252-266.

[24] Jun, S. J., Pinkse, J. and Wan, Y. (2013), Classical Laplace Estimation for $\sqrt[3]{n}$-Consistent Estimators: Improved Convergence Rates and Rate-Adaptive Inference, Working paper, Department of Economics, Pennsylvania State University.

[25] Kim, J. and Pollard, D. (1990), "Cube root asymptotics", *Annals of Statistics*, 18, 191-219.

[26] Klein, R. and Spady, R. (1993), "An Efficient Semiparametric Estimator for Binary Response Models", *Econometrica*, 61, 387-421.

[27] Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer, New York.

[28] Lee, S. M. S., and Pun, M.C. (2006), "On m out of n bootstrapping for nonstandard M-estimation", *Journal of the American Statistical Association*, 101, 1185-1197.

[29] Lewbel, A. (2000), "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables", *Journal of Econometrics*, 97, 145-177.

[30] Manski, C. F. (1975), "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205-228.

[31] Manski, C. F. (1985), "Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator", *Journal of Econometrics*, 27, 313-333.

[32] Manski, C. F. (1991), "Nonparametric Estimation of Expectations in the Analysis of Discrete Choice under Uncertainty" in W. Barnett, J. Powell, and G. Tauchen, (editors), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge: Cambridge Press.

[33] Manski, C. F. (1993), "Dynamic choice in social settings : Learning from the experiences of others", *Journal of Econometrics*, 58, 121-136.

[34] Mammen, E., C. Rothe, and M. Schienle (2013). Semiparametric Estimation with Generated Covariates, Working Paper.

[35] Newey, W.K. and McFadden, D.L. (1994), "Large Sample Estimation and Hypothesis Testing", In: Engle R. F. and McFadden, D. (eds), Handbook of Econometrics, Vol. 4, 2111-2245.

[36] Newey, W.K. (1997), "Convergence rates and asymptotic normality for series estimators", *Journal of Econometrics*, 79, 147-168.

[37] Powell, J. L., Stock, J. H., and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

[38] Van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, New York.

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

Table 1 : Parameter configuration in the DGP

Parameter	β_0	β_1	β_2	γ_{01}	γ_{11}	γ_{21}	γ_{00}	γ_{10}	γ_{20}	ρ
Value	0	1	1	0.2	0.15	0.1	0.1	0.08	0.4	-0.8

Table 2 : Simulation Results for $\widehat{\lambda}_{Single}$ and $\widehat{\lambda}_{OLS}$

N	Bias	RMSE	Median	mean AD	median AD
		Single-stage estimation			
300	-0.058	0.199	0.890	0.184	0.225
500	-0.048	0.190	0.928	0.172	0.202
1000	-0.040	0.184	0.942	0.166	0.187
		Two-stage estimation : OLS first stage			
300	-0.084	0.199	0.839	0.183	0.223
500	-0.070	0.191	0.876	0.174	0.205
1000	-0.055	0.187	0.901	0.170	0.194

Table 3 : Simulation Results for $\widehat{\lambda}_{Kernel}$

c	Bias	RMSE	Median	mean AD	median AD
Two-stage estimation : kernel first stage ($N = 300$)					
3	-0.112	0.195	0.818	0.178	0.206
3.5	-0.097	0.193	0.842	0.175	0.207
4	-0.087	0.194	0.845	0.177	0.211
4.5	-0.078	0.195	0.856	0.178	0.216
5	-0.071	0.198	0.862	0.180	0.220
5.5	-0.071	0.202	0.848	0.187	0.230
6	-0.085	0.202	0.835	0.186	0.234
Two-stage estimation : kernel first stage ($N = 500$)					
3	-0.100	0.190	0.841	0.173	0.200
3.5	-0.088	0.192	0.849	0.175	0.207
4	-0.074	0.189	0.873	0.170	0.198
4.5	-0.062	0.190	0.893	0.172	0.196
5	-0.065	0.197	0.872	0.180	0.218
5.5	-0.058	0.193	0.901	0.175	0.209
6	-0.078	0.198	0.858	0.182	0.223
Two-stage estimation : kernel first stage ($N = 1000$)					
3	-0.081	0.186	0.862	0.168	0.195
3.5	-0.077	0.186	0.867	0.168	0.191
4	-0.056	0.189	0.903	0.171	0.197
4.5	-0.055	0.183	0.908	0.165	0.183
5	-0.058	0.188	0.899	0.171	0.194
5.5	-0.058	0.187	0.902	0.169	0.195
6	-0.062	0.190	0.897	0.172	0.199