CROSS-FITTED EMPIRICAL LIKELIHOOD ON HIGH DIMENSIONAL SEMIPARAMETRIC MODELS

CHEN QIU

ABSTRACT. We consider empirical likelihood ratio for low dimensional parameters in the presence of infinite dimensional nuisance parameters. When nuisance parameters are estimated by modern high dimensional machine learning methods, Donsker Theorem can be rather restrictive. Instead, by using locally robust estimating equations and a cross fitting procedure, we establish a Wilks type theorem that validates empirical likelihood inference in high dimensional models. We construct easy-to-verify low level conditions and present how our results can be applied to many econometric models: partially linear model, treatment effect analysis and partially log linear model. Two pieces of simulation exercises show our method keeps up well with Wald statistics in linear case while outperforms its counterpart when the moment condition becomes non linear in parameters.

1. INTRODUCTION

Estimation and inference of a class of semiparametric models have re-drawn a lot of discussion in recent literature since the exciting development of modern machine learning methods. In many economic problems researchers are interested in some low dimensional, or causal parameters in the presence of some infinite dimensional nuisance parameters. A meaningful inference exercise for these models usually requires the $\sqrt{n}$ estimability of the low dimensional object. This can often be achieved when the infinite dimensional parameter is relatively simple: we model the dimension of their covariates to grow slowly enough compared to sample size (For example, Bickel, 1982; Robinson, 1988; Andrews, 1994; Newey, 1994; Newey and McFadden, 1994; Ai and Chen, 2003, 2012, etc.). However, in a data-rich environment, such default assumption becomes inadequate, especially when nuisance parameters are highly complex. Our paper develops a general theory of nonparametric likelihood inference under such situations. A Wilks type theorem is established to complement the relatively well-developed Wald type inference by work of Belloni et al. (2017); Chernozhukov et al. (2017), etc. The common strength of empirical likelihood (Owen, 1990, 1991, 2001) under low dimensions can be carried forward. For instance, our method does not require estimation of variance and it is less sensitive to nonlinearity. Since our likelihood ratio works under the null, it is less dependent on the quality of nuisance estimate and therefore more robust.

To fix ideas, we are interested in the empirical likelihood inference for θ_0, a finite dimensional object defined in the following moment condition with a known function g:

$$\mathbb{E}[g(Z, \theta_0, \beta_0)] = 0, \tag{1}$$

where Z is an observable vector valued random variable, and β_0 is an infinite dimensional nuisance parameter. To achieve $\sqrt{n}$ estimability of the Euclidean part, the convergence rate of the infinite dimensional parameter should be sufficiently fast, usually at least $o_p(n^{-1/4})$ (Chen, 2007, page 5587). More importantly, nuisance parameters should not be too complicated so that Donsker theorems can be used. In this paper we focus on the interesting case when β_0 is estimated by possibly high dimensional machine learning methods and therefore Donsker properties, which are underlying assumptions in traditional semiparametric inference, usually break down.

We circumvent Donsker Theorem by first choosing score function g (or its corresponding estimating equation) that has a locally robust property and then carrying out a cross fitting procedure. These methods have been proven powerful in removing hard-to-control bias in high dimensional models (eg., Belloni et al., 2017; Chernozhukov et al., 2017).

A locally robust g automatically implements a debiasing procedure since it minimizes the impact of nuisance parameter estimate on θ_0. In the context of empirical likelihood, it also aligns the variance estimate and empirical process part in log likelihood ratio. Thus, it is known to construct empirical likelihood ratio based on locally robust g instead of original ones (See for example, Hjort et al., 2009; Bravo et al., 2015; Matsushita and Otsu, 2017). The formal definition of a locally robust g is deferred to Theorem 1, but it is usually the influence function derived from pathwise derivative using von Mises calculus (eg., Newey, 1990, 1994; Ichimura and Newey, 2015), and possesses a doubly robust or Neyman orthogonal property (eg., Rothe and Firpo, 2016; Chernozhukov et al., 2016b).

On the other hand, cross fitting procedure removes further bias by splitting a random sample into several subgroups where nuisance parameters are estimated only using data not from its own group. Independence among sub samples ensures a conditional Markov inequality can be invoked. Such idea stemmed from Angrist and Krueger (1995); Bickel (1982); Van der Vaart (1998) and was brought to high dimensional or semiparametric models by work such as Belloni et al. (2011, 2012); Robins et al. (2013); Newey et al. (2017). Sample splitting method in empirical likelihood is relatively new. As far as we know, only Bertail (2006) briefly mentioned sample splitting as a possible direction to relax entropy conditions in empirical likelihood but it did not provide a theoretic proof and its focus was inherently different from ours.

To summarize, the first main contribution of this paper is that we fill the gap between estimation and inference in high dimensional semiparametric models. Our paper extends existing results in both empirical likelihood and high dimensional econometrics literature.

A second contribution of this paper is that we establish low level conditions that are easy to verify for most high dimensional semiparametric models. We show that convergence of estimated score function $\hat{g}$ in mean square and some "maximum" norm fashion is needed, which in turn requires sufficient convergence of estimated nuisance parameters in some suitable norm. Our results are much simpler in terms of asymptotics because do not need to differentiate between a linear and a nonlinear score—we work under the null. This implies that, compared to Wald statistics, our requirements on the functional class of g as well as the quality of nuisance estimate are weaker. As a result, our method is more flexible and robust. Simulation results confirm this. Lastly, our method might also be practically easier to extend to an over-identified case.

Hereafter our paper is organized as follows: we present our main theoretic results in Section 2, followed by discussions of its applications to some important economic problems in Section 3. Monte Carlo simulation results are collected in section 4. Section 5 concludes. All proofs, lemmas, tables and figures are deferred to Appendix.

Notations. $\mathbb{P}$ stands for the objective probability operator. Let $\mathbb{E}[\cdot] = \mathbb{E}_{\mathbb{P}}[\cdot]$ be expectation under $\mathbb{P}$, $\mathbb{E}_n[\cdot]$ be the empirical average, $\mathbb{I}\{A\}$ be the indicator function for an event A, $\|B\| = \sqrt{\text{trace}(B'B)}$ be the Euclidean norm for a scalar, vector, or matrix B, and $a \vee b = \max\{a, b\}$. For a function f, define its L_2 norm as $\|f\| = \|f\|_{\mathbb{P},2} = \left[\int |f(x)|^2 \, d\mathbb{P}(x)\right]^{1/2}$, and its L_∞ norm as $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.

2. Main Results

Suppose the low dimensional parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$ satisfies the following extended moment condition

$$\mathbb{E}[g(Z, \theta_0, \eta_0)] = 0, \tag{2}$$

where $g(\cdot) : \mathcal{Z} \times \Theta \times \mathcal{H} \to \mathbb{R}^d$ is a vector of known functions. $Z \in \mathcal{Z} \subset \mathbb{R}^{dz}$ is a vector valued random variable defined over a measurable objective probability space $(\Omega, \mathcal{S}, \mathbb{P})$. $\eta_0 \in \mathcal{H}$ is the infinite dimensional nuisance parameter, where $\mathcal{H}$ is a convex subset of some normed space with its norm denoted as $\|\cdot\|_H$. We observe an independently and identically distributed (hereafter iid) array of sample $\{Z_i\}_{i=1}^n$ from Z. As we can see, (2) just-identifies θ_0. It would be relatively easy to extend our model to an over-identified case (see remark 6).

To proceed, we introduce the notion of differentiability for a function in the presence of infinite dimensional parameters (Chapter 20.2, Van der Vaart, 1998). Consider a map $\eta \to f(\eta)$ from $\mathcal{H}_\eta$ to $\mathbb{R}^d$, where $\mathcal{H}_\eta$ is a subset of $\mathcal{H}$ containing η. We say $f(\eta)$ is Hadamard differentiable at η if there exists a continuous and linear map $f_\eta^{(1)}[h] : \mathcal{H} \to \mathbb{R}^d$

such that:

$$\left\| \frac{f(\eta + th_t) - f(\eta)}{t} - f_\eta^{(1)}[h] \right\| \to 0 \qquad \text{as } t \downarrow 0, \text{ every } h_t \to h.$$

We say $f(\eta)$ is continuously Hadamard differentiable if $f(\eta)$ is differentiable in a neighborhood of $\eta-$values. The second order Hadamard derivative of $f(\eta)$ is defined similarly and denoted as $f_\eta^{(2)}[h]$.

We next explain how cross-fitting, as a sample splitting method, works in the context of empirical likelihood inference: For an observation index $[n] = \{1, \cdots n\}$, we randomly partition it into approximately equal sized K subsamples $(I_k)_{k=1}^K$, where $K \geq 2$. Therefore each subsample I_k is of size $m = \frac{n}{K}$ and without loss of generality we assume m to be an integer. For each subsample I_k, define $I_k^c = \{1, \cdots n\} \setminus I_k$, i,e., I_k^c only includes observations *not* from subsample I_k. For each I_k we estimate η_0 by $\hat{\eta}_k = \hat{\eta}((Z_j)_{j \in I_k^c})$. That is, $\hat{\eta}_k$ is estimated only using data not from I_k . Finally, each observation i in subsample I_k will be fitted by the estimated functional form $\hat{\eta}_k$. For each θ, define empirical likelihood ratio as follows:

$$R_n(\theta) = \sup_{\{p_i\}_{i=1}^n} \left\{ \Pi_{i=1}^n (np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g(Z_i, \theta, \hat{\eta}_k) = 0 \right\} \tag{3}$$

We impose the following assumptions:

Assumption A.

In model (2),

(A0) θ_0 is the unique solution of model (2); $\mathbb{E}[g(Z, \theta_0, \eta)]$ is twice continuously Hadamard differentiable with respect to η on $\mathcal{H}$; $\mathbb{E}\{g(Z, \theta_0, \eta_0) g'(Z, \theta_0, \eta_0)\}$ exists and is non-singular; There exist some $\mathcal{H}_n \subset \mathcal{H}$ around η_0 such that $\hat{\eta}_k$ lies in $\mathcal{H}_n$ with probability at least $1 - \varepsilon_n$, for all random subsamples $I_k, k = 1 \cdots K$;

And the following rate conditions are satisfied:

(A1) $\sup_{\eta \in \mathcal{H}_n} \max_{i \in I_k} \|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\| = \tilde{r}_n = o(1)$ *for all* $I_k, k = 1 \cdots K$;

(A2) $\sup_{\eta \in \mathcal{H}_n} \left\{ \mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\|^2 \right] \right\}^{\frac{1}{2}} = r_n = o(1)$;

(A3) $\sup_{\eta \in \mathcal{H}_n} \left\| [\mathbb{E} g(Z, \theta_0, \eta)]_{\eta_0}^{(1)} [\eta - \eta_0] \right\| = a_n = o(n^{-\frac{1}{2}})$ *(locally robustness);*

(A4) $\sup_{\eta \in \mathcal{H}_n, 0 \leq \tilde{t} \leq 1} \left\| [\mathbb{E} g(Z, \theta_0, \eta)]_{\eta_0 + \tilde{t}(\eta - \eta_0)}^{(2)} [\eta - \eta_0] \right\| = b_n = o(n^{-\frac{1}{2}})$ *(regularity).*

Theorem 1. *Under Assumption A, we have:*

$$-2 \log R_n(\theta_0) \xrightarrow{p} \chi_d^2,$$

and for any critical level α

$$\mathbb{P}\left\{ -2 \log R_n(\theta_0) \leq \chi_{d,1-\alpha}^2 \right\} \to 1 - \alpha.$$

Remark 1. (A0) are common identification and regularity conditions and are fairly weak. Thus most effort shall be spent checking rate conditions listed in (A1)-(A4). (A1) and (A2) are concerned with structural properties of estimating equations $g(Z, \theta_0, \eta)$. For most time they can be easily verified by the convergence of $\hat{\eta}_k$ in some norm (eg., sup or L_2 norm). (A1) and (A2) imply the convergence rate of $\hat{\eta}_k$ should be sufficiently fast such that we have mean square convergence for estimating equations (A2) and something a bit more (A1). (A1) is required so that $\log R_n(\theta_0)$ exists nontrivially with a large probability (see Lemma 1). This seems distinct compared to existing high dimensional inference results, for example, Belloni et al. (2017); Chernozhukov et al. (2017) only imposed (A2). With this respect, our conditions are a bit stronger than its Wald counterpart. But here we do not need to differentiate between linear or nonlinear scores, which simplifies proofs considerably and imposes much less technical conditions when the score is nonlinear (See for example, Theorem 3.3 in Chernozhukov et al. 2016a). Notice, (A1) and (A2) can be jointly satisfied if we have:

(A5) $\sup_{\eta \in \mathcal{H}_n} \sup_{Z \in \mathcal{Z}} \|g(Z, \theta_0, \eta) - g(Z, \theta_0, \eta_0)\| = r_n' = o(1)$

However, (A5) might not be applicable as in many models of interest sup norm of $g(\cdot)$ might not even exist (See our applications in section 3). Therefore (A1) and (A2) are weaker than similar uniform convergence condition in Bravo et al. (2015). It turns out that although in some cases sup norm does not exist, the "maximum" condition in (A1) could still stand if we impose stronger moment existence conditions.

Remark 2. In traditional empirical likelihood literature, we need to assume that the function class $\mathcal{F} := \{g(Z, \theta_0, \eta) : \eta \in \mathcal{H}_n\}$ is $\mathbb{P}-$Donsker with probability 1. This requires that η_0 should not be too complicated in terms of its entropy. This is too stringent in a data-rich environment, as often we need to model covariates dimension k grows together with sample size n. Cross-fitting procedure significantly reduces our reliance on Donsker Theorem. This makes Theorem 1 attractive because essentially what a researcher needs to do is to ensure that we have a good quality estimate for nuisance parameters. For instance, if the nuisance parameter has an approximate sparse structure, then it usually requires that the dimension of non-sparse regressors should not grow too fast compared to sample size. However, notice that our result does not rely on sparsity assumptions per se.

Remark 3. Condition (A3) is the locally robustness condition. It guarantees that the impact of nuisance parameters is negligibly small. In many cases we can achieve $a_n = 0$, which corresponds to the exact orthogonal or perfectly debiased situation. If not, (A3) is the minimal requirement we need. To find a locally robust score, we can often take an influence function approach or "partialling-out" approach from original moment condition. There is a vast body of literature with detailed treatment on this, for instance, Pfanzagl

(1982); Bickel et al. (1993); Newey (1994); Tsiatis (2007); Ichimura and Newey (2015); Chernozhukov et al. (2016b), etc. Here we give a little flavor by an example. Suppose we are interested in estimating expected conditional variance (Newey et al., 2017):

$$\theta_0 = \mathbb{E}[Cov(W, Y \mid X)] = \mathbb{E}\left\{W[Y - \mathbb{E}(Y \mid X)]\right\}, \tag{4}$$

where X, Y, W are all scalar valued random variables. (4) can be viewed as a special case of (1) where the score $g(Z, \theta_0, \beta_0)$ admits a simple linear structure θ:

$$\theta_0 = \mathbb{E}[\psi(Z, \beta_0)], \tag{5}$$

where $Z = (X, Y, W)'$, $\beta_0 = \mathbb{E}(Y \mid X)$ and $\psi(Z, \beta_0) = W[Y - \beta_0]$. The locally robust g for (1) is straightforward:

$$g(Z, \theta_0, \eta_0) = [W - \gamma_0)][Y - \beta_0] - \theta_0, \tag{6}$$

where $\gamma_0 = \mathbb{E}(W \mid X)$. Notice (6) is also the influence function for each observation and gives formula for calculating asymptotic variance. It implies we have to additionally account for another infinite dimensional object γ_0.

Remark 4. To see how our conditions translate to raw rate requirement on $\hat{\eta}_k$, let $\|\hat{\eta} - \eta_0\|_\infty = O_p(\epsilon_n)$. First, we require $a_n = o(\frac{1}{\sqrt{n}})$, although in many cases $a_n = 0$. If $g(Z, \theta_0, \eta)$ shows some quadratic behavior, we can often bound b_n and r_n such that: $b_n = O(\epsilon_n^2), r_n^2 = O(\epsilon_n^2)$. If in addition we are willing to assume certain residuals with fourth moment existence, it is possible to bound $\tilde{r}_n$ as $\tilde{r}_n = o(n^{\frac{1}{4}}\epsilon_n)$. So all we need is $\epsilon_n = o(n^{-1/4})$. This means nuisance parameters under high dimensions should be estimated at least at rate $o_p(n^{-1/4})$.

Remark 5. This paper is not about finite sample properties. Importantly, the choice of K does not affect our asymptotic results while of course it might have significant impact in finite samples. In our simulations we find a larger sample size can usually bear with a larger K. Setting $K = 2$ or 4 is often a good start for practitioners. There are also some ways to mitigate finite sample problems. For example, consider the following procedure: Repeat the process in Theorem 1 for T times, and each time we get a different statistic $-2 \log R_{n,t}(\theta_0)(t = 1 \cdots T)$. The following statistic

$$\mathcal{R}_{n,T} = \frac{1}{T} \sum_{n=1}^{T} [-2 \log R_{n,t}(\theta_0)],$$

might make result less sensitive to uncertainty induced by random partitioning in finite samples.

Remark 6. It would be interesting to extend our result to an over-identified case. Suppose the dimension of estimating equations d is larger than $\dim(\theta_0)$, then our result in Theorem 1 still stands; However, it is now a joint test of over-identifying moment conditions as well

as parameters. To test for parameters only, similar to the procedure in Kitamura (2006) we could propose the following statistic:

$$\tilde{R}_n = -2\left[\log R_n(\theta_0) - \sup_{\theta\in\Theta} \log R_n(\theta)\right],$$

which we suspect should follow a chi-square distribution of degrees of freedom $\dim(\theta_0)$. This is computationally simpler than a Wald procedure in Belloni et al. (2017), which requires additional estimation of infinite dimensional parameters.

3. Applications

3.1. **Partially Linear Model.** Consider a partially linear projection where Y_i is projected on the the space of functions that has the form $\beta' X + h_0(W_i)$. Y_i is a scalar valued random variable, X_i, W_i are vectors of random variables, and h_0 is an unknown function such that $\|h_0\|_{\mathbb{P},2} < \infty$. Since $(X_i - \mathbb{E}[X_i \mid W_i])$ is orthogonal to $h(W_i) + \mathbb{E}[X_i \mid W_i]$, a naive moment condition to identify β_0 could be written as

$$\mathbb{E}\left\{\left[Y_i - \beta_0'\left(X_i - \mathbb{E}[X_i \mid W_i]\right)\right]\left[\left(X_i - \mathbb{E}[X_i \mid W_i]\right)\right]\right\} = 0, \tag{7}$$

where we project Y_i on the space of $(X_i - \mathbb{E}[X_i \mid W_i])$. This model can be alternatively interpreted as a partially linear regression

$$Y_i = \beta_0' X_i + h_0(W_i) + u_i, \mathbb{E}\left\{u_i \mid X_i, W_i\right\} = 0, \tag{8}$$

with X_i being an instrument

$$\mathbb{E}\left\{\left[Y_i - \beta_0' X_i - h_0(W_i)\right] X_i\right\} = 0. \tag{9}$$

This has been well studied in a low dimensional case (Robinson, 1988; Donald and Newey, 1994). Neither (7) or (9) satisfies the locally robust moment condition. The locally robust moment condition corresponding to (7) can be derived through the famous "partialling out" method (Robinson, 1988):

$$\mathbb{E}g(Z, \beta_0, \eta_0) = \mathbb{E}\left\{\left[Y - m_0^Y(W) - \beta_0'\left(X - m_0^X(W)\right)\right]\left(X - m_0^X(W)\right)\right\} = 0, \tag{10}$$

where $Z = (Y, X', W')'$, $\eta_0 = (m_0^Y(W), m_0^X(W)')'$, and $m_0^Y(W) = \mathbb{E}[Y \mid W]$, $m_0^X(W) = \mathbb{E}[X \mid W]$. And the locally robust moment condition for (9) is

$$\mathbb{E}\left\{g(Z, \beta_0, \eta_0)\right\} = \mathbb{E}\left[Y - \beta_0' X - h_0(W)\right]\left[X - m_0^X(W)\right] = 0, \tag{11}$$

where $Z = (Y, X', W')'$, $\eta_0 = (h_0(W), m_0^X(W)')'$. Notice both (10) and (11) expand the number of nuisance parameters to be estimated compared to original (7) and (9). Consider the cross-fitted empirical likelihood ratio defined as (3) where for (10)

$$g(Z_i, \beta_0, \hat{\eta}_k) = \left[Y_i - \hat{m}_k^Y(W_i) - \beta_0'\left(X_i - \hat{m}_k^X(W_i)\right)\right]\left[X_i - \hat{m}_k^X(W_i)\right],$$

and for (11)

$$g(Z_i, \beta_0, \hat{\eta}_k) = \left[Y_i - \beta_0' X_i - \hat{h}_k(W_i)\right]\left[x_i - \hat{m}_k^X(W_i)\right],$$

where $\hat{m}_k^Y(W_i)$, $\hat{m}_k^X(W_i)$ and $\hat{h}_k(W_i)$ are cross-fitted estimators for $m_0^Y(W_i)$, $m_0^X(W_i)$ and $h_0(W_i)$, respectively.

For any $\eta = (\eta_i)_{i=1}^J$, where each η_i is an infinite dimensional parameter from $\mathcal{Z}$ to $\mathbb{R}$, define $\|\eta\|_\infty = \max_{1\le i\le J} \|\eta_i\|_\infty$. Moreover, define $u^X = X - m_0^X(W); u^Y = Y - \beta_0' X - h_0(W)$. We impose the following assumptions.

Assumption B.

(1) β_0 is the unique solution of model (10) or (11).

(2) The error terms u^Y and u^X are such that

$$\mathbb{E}\left[\left(u^Y\right)^4\right] < \infty, \qquad \mathbb{E}\left\{\left[u^X\left(u^X\right)'\right]^2\right\} < \infty.$$

(3) $\mathbb{E}\left\{\left(u^Y\right)^2 u^X\left(u^X\right)'\right\}$ is non-singular.

(4) For all random subsamples $I_k, k = 1\cdots K$, where $K \ge 2$ and $\frac{n}{K}$ is an integer, $\hat{\eta}_k$ is such that with probability at least $1 - \varepsilon_n$,

$$\|\hat{\eta}_k(W_i) - \eta_0(W_i)\|_\infty = o(n^{-\frac{1}{4}}).$$

That is: For model (10):

$$\left\|\hat{m}_k^Y(W_i) - m_0^Y(W_i)\right\|_\infty = o(n^{-\frac{1}{4}}); \left\|\hat{m}_k^X(W_i) - m_0^X(W_i)\right\|_\infty = o(n^{-\frac{1}{4}});$$

For model (11):

$$\left\|\hat{m}_k^X(W_i) - m_0^X(W_i)\right\|_\infty = o(n^{-\frac{1}{4}}); \left\|\hat{h}_k(W_i) - h_0(W_i)\right\|_\infty = o(n^{-\frac{1}{4}}).$$

Theorem 2. *Under Assumption B, the empirical likelihood ratio constructed in (3) based on model (10) or (11) satisfies:*

$$-2\log R_n(\beta_0) \xrightarrow{p} \chi_d^2,$$

and

$$\mathbb{P}\left\{-2\log R_n(\beta_0) \le \chi_{d,1-\alpha}^2\right\} \to 1 - \alpha.$$

3.2. **Treatment Effects.** Consider inference on treatment effect in the model studied in Rubin (1974); Rosenbaum and Rubin (1983). For a random sample of size N from population, we observe a vector of variables $\{D, Y, X'\}'$, where D denotes a binary treatment variable $D \in \{0, 1\}$, Y is the scalar-valued outcome variable and X represents a vector of covariates. For each individual $i$there exist two potential outcomes, denoted $Y_i(1)$ and

$Y_i(0)$. $Y_i(1)$ represents the potential outcome for i when its treatment status is $D_i = 1$ (being treated) and $Y_i(0)$ represents the potential outcome when $D_i = 0$ (under control). Average treatment effect θ_0 is defined as:

$$\theta_0 = E\left[Y_i(1) - Y_i(0)\right].$$

Since for each i we only either observe $Y_i(1)$ or $Y_i(0)$, denote the observed outcome $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot(1-D_i)$. Under unconfoundedness and overlap assumptions, θ_0 can be identified as:

$$\theta_0 = \mathbb{E}\left[\frac{D_i Y_i}{\pi_0(X_i)} - \frac{(1-D_i)Y_i}{1-\pi_0(X_i)}\right], \tag{12}$$

where π is the propensity score and defined as

$$\pi_0 = \mathbb{P}(D = 1 \mid X).$$

12 implies to carry out inference on θ_0 we need to have a first stage estimator for nuisance parameter π_0. Usually the functional form of $\pi_0(X)$ might be very complicated and the the dimension of covariates X might be very large compared to sample size. Under such scenarios, to construct an empirical likelihood confidence region for θ_0 we need to first find the locally robust moment condition, which is well-known from Robins et al. (1995); Hahn (1998):

$$\mathbb{E}\left\{\frac{DY}{\pi_0} - \frac{(1-D)Y}{1-\pi_0} - (D-\pi_0)\left[\frac{\mathbb{E}[Y(1)|X]}{\pi_0} + \frac{\mathbb{E}[Y(0)|X]}{1-\pi_0}\right]\right\} = \theta_0. \tag{13}$$

(13) says in addition to original nuisance parameter—propensity score π_0, we still need to estimate two more nuisance parameters, $\mathbb{E}[Y(1)|X]$ and $\mathbb{E}[Y(0)|X]$. Notice under unconfoundedness assumption, both $\mathbb{E}[Y(1)|X]$ and $\mathbb{E}[Y(0)|X]$ can be estimated by observables Y, D and X. Therefore, our method is flexible enough to accommodate any modern machine learning estimators for any of the nuisance parameters in (13).

Another approach to estimating treatment effect is through information theoretics (Qiu and Otsu, 2018) or similarly, covariate-balancing (Zubizarreta, 2015; Chan et al., 2016). Our method can also be applied to this branch of growing literature. Denoting $\omega_1(X) = \pi_0(X)^{-1}$, $\omega_0(X) = \{1-\pi_0(X)\}^{-1}$, θ_0 can be identified as

$$\theta_0 = \mathbb{E}[\omega_1(X_i)D_iY_i] - \mathbb{E}[\omega_0(X_i)(1-D_i)Y_i], \tag{14}$$

which admits a simple linearly multiplicative structure. Notice the locally robust estimating equation under this case can be written in a structure simpler than (13):

$$\mathbb{E}\left\{\omega_1(X_i)D_i\left(Y_i - \mathbb{E}[Y(1)|X]\right) - \left[\omega_0(X_i)(1-D_i)\left(Y_i - \mathbb{E}[Y(0)|X]\right)\right] + \mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X]\right\} = \theta_0. \tag{15}$$

(15) is easier to analyze than (13) because of its linearly multiplicative structure, although now the number of nuisance parameters to be estimated increases from 3 to 4.

Let $Z_i = (D_i, Y_i, X_i')'$. For model (13), let $\eta_0 = (\pi_0, m_0, l_0)$ where $m_0 = \mathbb{E}[Y(1)|X]$ and $l_0 = \mathbb{E}[Y(0)|X]$; For model (15), $\eta_0 = (\omega_1, \omega_0, m_0, l_0)$ where $\omega_1 = \omega_1(X), \omega_0 = \omega_0(X)$. Define $\|\eta\|_\infty = \max_i \{\|\eta_i\|_\infty\}$. Denote the cross-fitted machine learning estimator $\hat{\eta}_k = (\hat{\pi}_k(X_i), \hat{m}_k(X_i), \hat{l}_k(X_i))$ or $(\hat{\omega}_{1k}(X_i), \hat{\omega}_{0k}(X_i), \hat{m}_k(X_i), \hat{l}_k(X_i))$. Then let:

$$g(Z_i, \theta_0, \hat{\eta}_k) = \frac{D_i Y_i}{\hat{\pi}_k(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{\pi}_k(X_i)} - (D_i - \hat{\pi}_k(X_i)) \left[\frac{\hat{m}_k(X_i)}{\hat{\pi}_k(X_i)} + \frac{\hat{l}_k(X_i)}{1-\hat{\pi}_k(X_i)} \right] - \theta_0$$

for model (13) and

$$g(Z_i, \theta_0, \hat{\eta}_k) = \left\{ \hat{\omega}_{1k} D_i \left(Y_i - \hat{m}_k(X_i)\right) - \left[\hat{\omega}_{0k}(X_i)(1-D_i) \left(Y_i - \hat{l}_k(X_i) \right) \right] + \hat{m}_k(X_i) - \hat{l}_k(X_i) \right\} - \theta_0$$

for model (15). The empirical likelihood ratio is defined the same way as in (3). We further need the following conditions:

Assumption C.

(1) θ_0 is the unique solution of model (13) or (15) ; $D \perp (Y(1), Y(0))|X$ (unconfoundedness); there exists some strictly positive constant ε such that $\varepsilon \leq \pi_0 \leq 1-\varepsilon$ (overlap).

(2) $\mathbb{E}[Y^4(0)] < \infty, \mathbb{E}[Y^4(1)] < \infty$.

(3) $\mathbb{E}\left\{ \{\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X] - \theta\}^2 + \frac{\mathrm{Var}(Y(1)|X)}{\pi(X)} + \frac{\mathrm{Var}(Y(0)|X)}{1-\pi(X)} \right\} \neq 0$.

(4) For all random subsamples $I_k, k = 1 \cdots K$, where $K \geq 2$ and $\frac{n}{K}$ is an integer, $\hat{\eta}_k$ is such that $\|\hat{\eta}_k\|_\infty < \infty$ and satisfies the following rate conditions: with probability at least $1 - \varepsilon_n$, for model (13):

$$\|\hat{\pi}_k(X_i) - \pi_0(X_i)\|_\infty = o(n^{-\frac{1}{4}}); \qquad \|\hat{m}_k(X_i) - m_0(X_i)\|_\infty = o(n^{-\frac{1}{4}});$$
$$\left\|\hat{l}_k(X_i) - l_0(X_i)\right\|_\infty = o(n^{-\frac{1}{4}}); \qquad \varepsilon \leq \|\hat{\pi}_k(X_i)\|_\infty \leq 1-\varepsilon.$$

For model (11):

$$\|\hat{\omega}_{1k}(X_i) - \omega_1(X_i)\|_\infty = o(n^{-\frac{1}{4}}); \qquad \|\hat{\omega}_{0k}(X_i) - \omega_0(X_i)\|_\infty = o(n^{-\frac{1}{4}});$$
$$\|\hat{m}_k(X_i) - m_0(X_i)\|_\infty = o(n^{-\frac{1}{4}}); \qquad \left\|\hat{l}_k(X_i) - l_0(X_i)\right\|_\infty = o(n^{-\frac{1}{4}});$$
$$\varepsilon \leq \|\hat{\omega}_{1k}(X_i)\|_\infty \leq 1-\varepsilon; \qquad \varepsilon \leq \|\hat{\omega}_{0k}(X_i)\|_\infty \leq 1-\varepsilon.$$

Theorem 3. *Under Assumption C, the empirical likelihood ratio constructed according to (3) for model (13) or (11) satisfies:*

$$-2 \log R_n(\theta_0) \xrightarrow{p} \chi_1^2,$$

and

$$\mathbb{P}\left\{-2\log R_n(\theta_0) \leq \chi^2_{1,1-\alpha}\right\} \to 1-\alpha.$$

3.3. **Partially log-linear model.** We revisit the model discussed in Robins et al. (1992). It is similar to the partially linear model in subsection 3.1 but with an exponential link

$$Y_i = \exp\left[\beta_0' X_i + \xi_0(W_i)\right] + u_i; E(u_i \mid X_i, W_i) = 0. \tag{16}$$

We might use model (16) instead of (8) if we suspect Y displays a Poisson flavored distribution and is often encountered in biostatistic or epidemiological investigations. Model (16) can also be viewed as a special case of the general partially linear model with a known link Φ, which was studied in Horowitz et al. (2004); Ai and Chen (2003). The parameter of interest is still the Euclidean parameter β_0, which can be interpreted as the causal effect of X_i after controlling for confounding factors W_i. Based on work from Robins et al. (2013); Vermeulen and Vansteelandt (2015) who studied doubly-robust estimators for β_0, it is easy to propose the following locally robust moment condition that can identify β_0:

$$\mathbb{E}\left\{g(Z,\beta_0,\eta_0)\right\} = \mathbb{E}\left\{\left[\exp(-\beta_0' X)Y - h_0(W)\right](X - \pi_0(W))\right\} = 0, \tag{17}$$

where $Z = (Y, X', W')'$, $\eta_0(h_0(W), \pi_0(W)')'$, and $h_0(W) = \exp(\xi_0(W)) = \mathbb{E}[Y \mid X = 0, W]$, $\pi_0(W) = \mathbb{E}[X \mid W]$. When direct estimation of h_0 in (17) is difficult, the following alternative locally robust moment condition could be applied in the spirit of "partialling out"

$$\mathbb{E}\left\{g(Z,\beta_0,\eta_0)\right\} = \mathbb{E}\left\{\left[\exp(-\beta_0' X)Y - \frac{m_0(W)}{l_0(W)}\right](X - \pi_0(W))\right\} = 0, \tag{18}$$

where $m_0(W) = E(Y \mid W)$ and $l_0(W) = \mathbb{E}\left[\exp(\beta_0' X) \mid W\right]$. Notice moment condition (18) is slightly different from the functional form covered by our general theory in section 2, since now $l_0(W)$ is in fact also a function of unknown β_0. Nevertheless, Theorem 1 can still be utilized here since under H_0, β_0 is known and $l_0(W)$ can be easily estimated under the the null hypothesis as a conditional expectation. Consider the cross-fitted log empirical likelihood ratio constructed as in (3) where for model (17)

$$g(Z_i,\beta_0,\hat{\eta}_k) = \left[\exp(-\beta_0' X_i)Y_i - \hat{h}_k(W_i)\right](X_i - \hat{\pi}_k(W_i)),$$

where $\hat{h}_k(W_i)$ and $\hat{\pi}_k(W_i)$ are some machine learning estimators of $h_0(W)$ and $\pi_0(W)$, respectively, based on subsample I_k^c; and for model (18)

$$g(Z_i,\beta_0,\hat{\eta}_k) = \left[\exp(-\beta_0' X_i)Y_i - \frac{\hat{m}_k(W_i)}{\hat{l}_k(W_i)}\right](X_i - \hat{\pi}_k(W_i)),$$

where $\hat{m}_k(W_i)$ and $\hat{l}_k(W_i)$ are the cross-fitting machine learning estimators for their theoretic counterparts. Same as previous examples, $\|\eta\|_\infty = \max_i \{\|\eta_i\|_\infty\}$, and define

$u^Y = Y - \exp(\beta_0' X) h_0(W)$, and $u^X = X - \pi_0(W)$. We impose the following set of assumptions:

Assumption D.

(1) β_0 is the unique solution of model (17) or (18); $X \in \mathcal{X}$, a compact set in $\mathbb{R}^d$.

(2) The error terms, u^Y and u^X are such that

$$\mathbb{E}\left[\left(u^Y\right)^4\right] < \infty, \qquad \mathbb{E}\left\{\left[u^X \left(u^X\right)'\right]^2\right\} < \infty;$$

(3) $\mathbb{E}\left\{\exp(-2\beta_0' X)\left(u^Y\right)^2 u^X \left(u^X\right)'\right\}$ is non-singular.

(4) For all random subsamples $I_k, k = 1 \cdots K$, where $K \geq 2$ and $\frac{n}{K}$ is an integer, $\hat{\eta}_k$ satisfies the following rate conditions: with probability at least $1 - \varepsilon_n$, for model (17):

$$\|\hat{\pi}_k(W_i) - \pi_0(W_i)\|_\infty = o(n^{-\frac{1}{4}}); \left\|\hat{h}_k(W_i) - h_0(W_i)\right\|_\infty = o(n^{-\frac{1}{4}}).$$

For model (18):

$$\|\hat{\pi}_k(W_i) - \pi_0(W_i)\|_\infty = o(n^{-\frac{3}{8}}); \|\hat{m}_k(W_i) - m_0(W_i)\|_\infty = o(n^{-\frac{3}{8}});$$
$$\left\|\hat{l}_k(W_i) - l_0(W_i)\right\|_\infty = o(n^{-\frac{3}{8}}).$$

And there exists some strictly positive constant $\underline{\delta}$ and $\bar{\delta}$ such that for model (18):

$$\|h_0(W_i)\|_\infty < \bar{\delta}; \qquad \|m_0(W_i)\|_\infty < \bar{\delta};$$
$$\|l_0(W_i)\|_\infty > \underline{\delta}; \qquad \left\|\hat{l}_k(W_i)\right\|_\infty > \underline{\delta}.$$

Theorem 4. *Under Assumption D, the empirical likelihood ratio constructed in (3) based on model (17) or (14) satisfies:*

$$-2 \log R_n(\beta_0) \xrightarrow{p} \chi^2_d$$

where $d = \dim(\beta_0)$, and

$$\mathbb{P}\left\{-2 \log R_n(\beta_0) \leq \chi^2_{d,1-\alpha}\right\} \to 1 - \alpha.$$

4. Simulation

4.1. **Missing data.** We consider the problem of evaluating population mean in the presence of missing data. This can be viewed as a simplified version of the treatment effect model discussed in Subsection 3.2. The response variable $Y(1)$ is generated as:

$$Y_i(1) = \lambda(1)' X + e_1,$$

where X is a p dimensional vector of covariates with p possibly large compared to sample size n. X includes an intercept, with the remainder drawn from $N(0, \Sigma)$, where $\Sigma_{[j_1,j_2]} = 2^{-|j_1-j_2|}$ for $2 \leq j_1, j_2 \leq p$. Let e_1 be independent of X and standard normal. The propensity score is designed to have a logistic fashion:

$$\pi_0 = \mathbb{P}(D = 1 \mid X) = \exp(\gamma' X)/[1 + \exp(\gamma' X)].$$

Similar to Farrell (2015), we add more structure for coefficients $\lambda(1), \gamma$ such that the model necessarily admits approximate sparsity. Let

$$\lambda(1) = \rho_1(-1, -1, 2^{-\alpha_1}, -3^{-\alpha_1}, \cdots j^{-\alpha_1}, \cdots p^{-\alpha_1})$$
$$\gamma = \rho_\gamma(1, 1, -2^{-\alpha_\gamma}, 3^{-\alpha_\gamma}, \cdots j^{-\alpha_\gamma}, \cdots - p^{-\alpha_\gamma}),$$

where $(\rho_1, \rho_\gamma, \alpha_1, \alpha_\gamma)$ are parameters set by researchers: ρ_1, ρ_γ control the strength of signal compared to noise while α_1, α_γ controls degree of approximate sparsity. As usual we are interested in the population parameter

$$\vartheta_0 = \mathbb{E}\left[Y_i(1)\right],$$

where we always observe $Y_i = Y_i(1) \cdot D_i$. In simulation we let $p = 501$, and $n = 500$. Monte Carlo replications are set as 1000. We consider a variety of combinations of $(\rho_1, \rho_\gamma, \alpha_1, \alpha_\gamma)$ to evaluate the performance of our method compared to Wald-type inference. To showcase the flexibility of our methods, we consider locally robust score function to be

$$g(Z, \vartheta_0, \eta_0) = \frac{DY}{\pi_0} - \frac{\mathbb{E}[Y(1)|X]}{\pi_0}(D - \pi_0) - \vartheta_0, \tag{19}$$

or the information theoretic type:

$$\tilde{g}(Z, \vartheta_0, \eta_0) = \omega_1 D\left(Y - \mathbb{E}[Y(1)|X]\right) + \mathbb{E}[Y(1)|X] - \vartheta_0. \tag{20}$$

Sample-splitting index K is set to be 2, 4 or 5. For such high dimensional models, many estimation methods for nuisance parameters exist. For simplicity we only use a Lasso method with 10 fold cross validation. We evaluate the empirical coverage probability under a nominal rejection rate of 5%. Results are reported in Table 1 and Table 2.

The strength of the signal of propensity score ρ_γ seems important in determining empirical performance in our setup. See Figure 1. When $\rho_\gamma = 1$ (strong signal), cross fitted empirical likelihood ratio keeps up very well with Wald counterpart. There is no significant evidence showing which one is better. This is expected since the true model is approximately sparse as well as linear, they should perform in a very similar way. We also see that setting $K = 4$ removes most bias. Empirical performance when $K = 5$ is not necessarily better than when $K = 4$. It might be the reason that finite sample problem

might be dominant in the former case. On the other hand, when $\rho_\gamma = -1$ (weak signal), the performance of both EL and Wald drops considerably, which is also expected. Again, no evidence showing which one is the winner. However, we do notice that the empirical likelihood based on (20) is the most stable one, even when the signal of the propensity score is weak. This partly shows the merit of using information-theoretic-based methods.

4.2. **Partially log linear model.** In this subsection we examine the performance of our method when the score is defined in a nonlinear way. We design a partly log linear model in Subsection 3.3 with a flavor similar to Kang et al. (2007). Let $W = (W_1, W_2, W_3, W_4)$ be generated from standard multivariate normal distribution. The response variable Y is defined as

$$Y = \exp(\beta_0 X + 0.5W_1 + 1.8W_2 - 1.8W_3 + 1.8W_4) + e,$$

with $e \sim N(0,1)$, and X is generated as $X = \Phi[0.7W_1 - 1.5W_2 + 1.5W_3 - 0.3W_4] + \tilde{e}$ where $\tilde{e} \sim N(0,1)$ and $\Phi(\cdot)$ is the CDF of a standard normal. We carry out hypothesis testing exercises for β_0, by means of the locally robust score defined in (17). In the first step the two nuisance parameters, $h_0(W) = \exp(\xi_0(W)) = \mathbb{E}[Y \mid X = 0, W]$ and $\pi_0(W) = \mathbb{E}[X \mid W]$, are estimated using straightforward series methodology since they are conditional expectations. Notice $h_0(W)$ is estimated by first evaluating $\mathbb{E}[Y \mid X, W]$ followed by a plug-in when $X = 0$. We consider the following sets of regressors: (1) W; (2) A polynomial of order two based on W. In the latter case researchers might worry about a small sample problem thus decide to use some regularization. For simplicity, we only consider a combination of l_1 and l_2 penalty methods to estimate $h_0(W)$ and $\pi_0(W)$. For comparison we also report results when one of the nuisance parameters, $\pi_0(W)$ is known as well as when both nuisance parameters are known. We set sample size to be 100. Sample splitting indexes are either 2 or 4. Experiments are repeated 1000 times. Table 3 presents our empirical coverage probabilities when nominal rejection is 5%.

Results show that our EL procedure dominantly performs better and more robustly than a Wald statistic. For most cases EL ratio maintains coverage probabilities much better. These results show the robustness of our procedure compared to its Wald counterpart. This is also expected, because: (1) When the score is highly nonlinear, estimate based on method of moment procedure might be subject to more apparent finite sample problems; variance also needs to be re-estimated and while in a linear case estimation of variance is much straightforward. (2) A nonlinear score usually requires a more stringent condition on functional classes of g and a better quality estimate of nuisance parameters. This problem is avoided in our test since we work under the null. Therefore, Wald statistic

seems more sensitive to the functional form of the moment condition as well as the method used to estimate nuisance parameters.

5. Concluding Remarks

We established simple and easy-to-verify conditions under which empirical likelihood ratio test is pivotal, for a class of semiparametric moment condition models. In addition to some usual second mean type convergence, a "maximum" norm condition is proposed that will guarantee the nontrivial existence of EL ratio. Most of our requirements can be translated to nuisance parameters being estimated consistently at rate at least $o_p(n^{-1/4})$, if the estimating function shows some quadratic behavior. We use three examples to showcase how our main theorem can be applied. Notice our rate conditions are sufficient and not necessarily the weakest possible. But they are much straightforward and easy to verify. Simulation exercises confirm that our method works well, even when the locally robust score becomes highly nonlinear. For future research, it might be interesting to study the effect of K on finite sample performance of our method, and to extend our model to an over-identified case. Moreover, research is also needed in our context that would allow nuisance estimate to converge slower than $o_p(n^{-1/4})$.

Appendix A. Lemmas

Under Assumption A, we have the following lemmas.

Lemma 1. $\mathbb{P}\{\mathbf{0} \in \mathcal{C}_n\} \to 1$, *where* $\mathbf{0}$ *is the zero vector in* $\mathbb{R}^d$ *and* $\mathcal{C}_n$ *is the interior of the convex hull of* $\{g(Z_i, \theta_0, \hat{\eta}_k), i = 1 \cdots n, k = 1 \cdots K\}$.

Lemma 2. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [g(Z_i, \theta_0, \hat{\eta}_k)] \xrightarrow{d} N(0, \Omega)$, *where* $\Omega = \mathbb{E}[g(Z, \theta_0, \eta_0) g'(Z, \theta_0, \eta_0)]$.

Lemma 3. $\frac{1}{n} \sum_{i=1}^{n} [g(Z_i, \theta_0, \hat{\eta}_k) g'(Z_i, \theta_0, \hat{\eta}_k)] \xrightarrow{p} \Omega$.

Lemma 4. $\max_{1 \le i \le n} \|g(Z_i, \theta_0, \hat{\eta}_k)\| = o_p(\sqrt{n})$.

Appendix B. Proofs

Notation. To ease notational burden in the proofs denote:

$$g(Z_i, \theta_0, \eta_0) = g_{0i}, g(Z_i, \theta, \hat{\eta}_k) = \hat{g}_{ki}.$$

Theorem 1.

Proof. We verify Conditions (A0)-(A3) in Hjort et al. (2009). Lemma 1 corresponds to (A0), Lemma 2 verifies (A1), Lemma 3 verifies (A2) and Lemma 4 confirms (A3). The conclusion follows directly by Theorem 2.1 in Hjort et al. (2009). □

Lemma 1.

Proof. Since K is fixed, we only need to show for each $k = 1 \cdots K$:

$$\mathbb{P}\{\mathbf{0} \in \mathcal{C}_{n,k}\} \to 1,$$

where $\mathcal{C}_{n,k}$ is the interior of the convex hull of $\{g(Z_i, \theta_0, \hat{\eta}_k), i = 1 \cdots n\}$. Since also the dimension of $\hat{g}_{ki}$ is fixed, it suffices to show $\mathbb{P}\left\{\max_{i \in I_k} (\hat{g}'_{ki} e) > 0\right\} \to 1$ for any e in the space of unit vectors in $\mathbb{R}^d$. Fix e, since

$$\max_{i \in I_k} \hat{g}'_{ki} e \ge \max_{i \in I_k} g'_{0i} e - \max_{i \in I_k} |(g_{0i} - \hat{g}_{ki})' e|,$$

it suffices to show

$$\mathbb{P}\left\{\max_{i \in I_k} g'_{0i} e - \max_{i \in I_k} |(g_{0i} - \hat{g}_{ki})' e| > 0\right\} \to 1.$$

As $\mathbb{E}[g'_{0i} e] = 0$ and $\mathbb{E}(g'_{0i} e)^2 < \infty$ by assumption (A0), we have $\mathbb{P}\{g(Z, \theta_0, \eta_0)' e > 0\} > 0$ by Owen (1990, Lemma 2). Then, similar to argument (2.7) in Owen (1990, page 100), with probability approaching 1, we can find always find some positive constant $\varepsilon_{n,k}$ such that

$$\max_{i \in I_k} g'_{0i} e > \varepsilon_{n,k}. \tag{21}$$

Let $\Xi_{1,n,k}$ denote event (21), and $\mathscr{E}_{n,k}$ be the event that $\widehat{\eta}_k \in \mathcal{H}_n$. Conditional on $\Xi_{1,n,k}$ and $\mathcal{E}_n$, by Cauchy-Schwarz inequality and (A1):

$$\begin{aligned}
&\mathbb{P}\left\{\max_{i\in I_k}|(g_{0i}-\hat{g}_{ki})'e| \le \varepsilon_{n,k}\right\} \\
\ge &\mathbb{P}\left\{\sup_{\eta\in\mathcal{H}_n}\max_{i\in I_k}\|g(Z_i,\theta_0,\eta)-g_{0i}\| \le \varepsilon_{n,k}\right\} \to 1. \qquad (22)
\end{aligned}$$

Denote the event in (22) $\Xi_{2,n,k}$. Further let $\Xi_{3,n,k}$ be the event that all $\Xi_{1,n,k}$, $\mathscr{E}_{n,k}$ and $\Xi_{2,n,k}$ stand. Conditional on $\Xi_{3,n,k}$, we have:

$$\begin{aligned}
\max_{i\in I_k} g_{0i}'e - \max_{i\in I_k}|(g_{0i}-\hat{g}_{ki})'e| &= \max_{i\in I_k} g_{0i}'e - \varepsilon_{n,k} \\
&\quad + \varepsilon_{n,k} - \max_{i\in I_k}|(g_{0i}-\hat{g}_{ki})'e| \\
&> 0
\end{aligned}$$

by our construction. Notice also $\mathbb{P}(\Xi_{3,n,k}) \to 1$, we have

$$\mathbb{P}\left\{\max_{i\in I_k} g_{0i}'e - \max_{i\in I_k}|(g_{0i}-\hat{g}_{ki})'e|\right\} \to 1,$$

which completes our proof. □

Lemma 2.

Proof. First, denote $\mathcal{E}_n$ as the event that $\widehat{\eta}_k \in \mathcal{H}_n$, for all $k = 1\cdots K$. By Assumption (A0) and Boole's inequality, $\mathbb{P}\{\mathcal{E}_n\} \ge 1 - K\varepsilon_n \to 1$ since $\varepsilon_n = o(1)$. Second, we demonstrate that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{ki} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{0i} + o_p(1).$$

Notice:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{ki} &= \sqrt{n}\frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{m}\sum_{i\in I_k}\hat{g}_{ki}\right\} \\
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{0i} &= \sqrt{n}\frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{m}\sum_{i\in I_k} g_{0i}\right\},
\end{aligned}$$

since K is fixed. So it suffices to show:

$$\frac{1}{\sqrt{m}}\sum_{i\in I_k}(\hat{g}_{ki} - g_{0i}) = o_p(1), \forall k = 1\cdots K.$$

Now for each $k \in \{1 \cdots K\}$:

$$\left\| \frac{1}{\sqrt{m}} \sum_{i \in I_k} (\hat{g}_{ki} - g_{0i}) \right\| \leq A_1 + A_2$$

$$A_1 = \left\| \frac{1}{\sqrt{m}} \sum_{i \in I_k} \left\{ \hat{g}_{ki} - \mathbb{E}\left[\hat{g}_{ki} \mid (Z_j)_{j \in I_k^c} \right] \right\} - \frac{1}{\sqrt{m}} \sum_{i \in I_k} \left\{ g_{0i} - \mathbb{E} g_{0i} \right\} \right\|$$

$$A_2 = \sqrt{m} \left\| \mathbb{E}\left[\hat{g}_{ki} \mid (Z_j)_{j \in I_k^c} \right] - \mathbb{E} g_{0i} \right\|$$

Also notice for any a vector and x:

$$\begin{aligned} \mathbb{E}[\|a - \mathbb{E}(a \mid x)\|^2 \mid x] &= \mathbb{E}\left\{ [a - \mathbb{E}(a \mid x)]' [a - \mathbb{E}(a \mid x)] \mid x \right\} \\ &= \mathbb{E}\left\{ [a'a - 2a'\mathbb{E}(a \mid x) + \mathbb{E}'(a \mid x)\mathbb{E}(a \mid x)] \mid x \right\} \\ &= \mathbb{E}\left(\|a\|^2 \mid x \right) - \|\mathbb{E}(a \mid x)\|^2 \\ &\leq \mathbb{E}\left(\|a\|^2 \mid x \right) \end{aligned} \tag{23}$$

As a result, on event $\mathcal{E}_n$:

$$\begin{aligned} \mathbb{E}\left[A_1^2 \mid (Z_j)_{j \in I_k^c} \right] &= \mathbb{E}\left[\frac{1}{m} \left\| \sum_{i \in I_k} \left\{ \hat{g}_{ki} - \mathbb{E}\left[\hat{g}_{ki} \mid (Z_j)_{j \in I_k^c} \right] - g_{0i} + \mathbb{E} g_{0i} \right\}_{i=1}^m \right\|^2 \mid (Z_j)_{j \in I_k^c} \right] \\ &\leq \mathbb{E}\left[\left\| \hat{g}_{ki} - g_{0i} - \mathbb{E}\left[\hat{g}_{ki} \mid (Z_j)_{j \in I_k^c} \right] + \mathbb{E} g_{0i} \right\|^2 \mid (Z_j)_{j \in I_k^c} \right] \\ &\leq \mathbb{E}\left[\|\hat{g}_{ki} - g_{0i}\|^2 \mid (Z_j)_{j \in I_k^c} \right] \end{aligned}$$

where the first inequality is because conditional on $(Z_j)_{j \in I_k^c}$, $\hat{\eta}_k$ is fixed, and the second inequality is due to $\mathbb{E} g_{0i} = \mathbb{E}(g_{0i} \mid (Z_j)_{j \in I_k^c})$ and(23). Therefore by (A1):

$$\begin{aligned} \mathbb{E}\left[A_1^2 \mid (Z_j)_{j \in I_k^c} \right] &\leq \mathbb{E}\left[\|\hat{g}_{ki} - g_{0i}\|^2 \mid (Z_j)_{j \in I_k^c} \right] \\ &\leq \sup_{\eta \in \mathcal{H}_n} \mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g_{0i}\|^2 \mid (Z_j)_{j \in I_k^c} \right] \\ &\leq \sup_{\eta \in \mathcal{H}_n} \mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g_{0i}\|^2 \right] \\ &= r_n^2 = o(1). \end{aligned}$$

Thus by a conditional version of Chebyshev's inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left\{ A_1 > \epsilon \mid (Z_j)_{j \in I_k^c} \right\} \to 0.$$

Since $\mathbb{P}\left\{ A_1 > \epsilon \mid (Z_j)_{j \in I_k^c} \right\}$ can be straightforwardly verified as uniformly integrable, by Theorem 25.12 of Billingsley (2008),

$$\mathbb{P}\left\{ A_1 > \epsilon \right\} = \mathbb{E}\left\{ \mathbb{P}\left\{ A_1 > \epsilon \mid (Z_j)_{j \in I_k^c} \right\} \right\} \to 0.$$

This confirms

$$A_1 = o_p(1). \tag{24}$$

Now we bound A_2. Let $f(t) = \mathbb{E}\left[g[Z_i, \theta_0, \eta_0 + t(\hat{\eta}_k - \eta_0)] \mid (Z_j)_{j\in I_k^c}\right]$. By assumption (A0), $\mathbb{E}[g(Z, \theta_0, \eta)]$ is twice continuously Hadamard differentiable, so Taylor expansion yields:

$$\mathbb{E}\left[\hat{g}_{ki} \mid (Z_j)_{j\in I_k^c}\right] = f(1) = f(0) + f^{(1)}(0) + \frac{1}{2}f^{(2)}(\tilde{t})$$

where $\tilde{t}$ lies on the line joining 0 and 1. Notice $f(0) = \mathbb{E}\left[g[Z_i, \theta_0, \eta_0] \mid (Z_j)_{j\in I_k^c}\right] = \mathbb{E}\left[g(Z_i, \theta_0, \eta_0)\right]$. Therefore[1],

$$\begin{aligned}\left\|\mathbb{E}\left[g(z_i, \theta_0, \hat{\eta}_k) \mid (Z_j)_{j\in I_k^c}\right] - \mathbb{E}\left[g(z_i, \theta_0, \eta_0)\right]\right\| &= \left\|f^{(1)}(0) + \frac{1}{2}f^{(2)}(\tilde{t})\right\| \\ &\leq \|f'(0)\| + \frac{1}{2}\sup_{0\leq\tilde{t}\leq 1}\left\|f^{(2)}(\tilde{t})\right\|\end{aligned}$$

Notice also $f^{(1)}(\cdot), f^{(2)}(\cdot)$ are readily the first and second Hadamard derivatives, respectively:

$$f^{(1)}(0) = \left\{\mathbb{E}\left[g(Z_i, \theta_0, \eta) \mid (Z_j)_{j\in I_k^c}\right]\right\}^{(1)}_{\eta_0}[\hat{\eta}_k - \eta_0] : \mathcal{H}_n \to \mathbb{R}^d$$

and

$$f^{(2)}(\tilde{t}) = \left\{\mathbb{E}\left[g(Z_i, \theta_0, \eta) \mid (Z_j)_{j\in I_k^c}\right]\right\}^{(2)}_{\eta_0+\tilde{t}(\hat{\eta}_k-\eta_0)}[\hat{\eta}_k - \eta_0] : \mathcal{H}_n \to \mathbb{R}^d.$$

On event $\mathcal{E}_n$, by (A2) and (A3):

$$\begin{aligned}A_2 &\leq \sqrt{m}\sup_{\eta\in\mathcal{H}_n}\left\|[\mathbb{E}g(Z_i, \theta_0, \eta)]^{(1)}_{\eta_0}[\eta - \eta_0]\right\| \\ &\quad + \frac{\sqrt{m}}{2}\sup_{\eta\in\mathcal{H}_n, 0\leq\tilde{t}\leq 1}\left\|[\mathbb{E}g(Z_i, \theta_0, \eta)]^{(2)}_{\eta_0+\tilde{t}(\eta-\eta_0)}[\eta - \eta_0]\right\| \\ &= O(\sqrt{m}a_n) + O(\sqrt{m}b_n) \\ &= O(\sqrt{n}a_n) + O(\sqrt{n}b_n) \\ &= o(1),\end{aligned} \tag{25}$$

since $m = \frac{n}{K}$ and K is fixed. (24) and (25) imply

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{ki} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_{0i} + o_p(1).$$

The conclusion follows by invoking Cramer-Wold device and Lindeberg-Lévy CLT. □

[1] Or by Proposition 5.3.13 in Atkinson and Han (2005), we also have: $\|f(1) - f(0)\| \leq \left\|f^{(1)}(0)\right\| + \frac{1}{2}\sup_{0\leq\tilde{t}\leq 1}\left\|f^{(2)}(\tilde{t})\right\|$.

Lemma 3.

Proof. Since:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{ki}\hat{g}'_{ki}) = \frac{1}{K}\sum_{k=1}^{K}(\frac{1}{m}\sum_{i\in I_k}\hat{g}_{ki}\hat{g}'_{ki})$$

and K is fixed, it suffices to show

$$\frac{1}{m}\sum_{i\in I_k}\hat{g}_{ki}\hat{g}_{ki} \xrightarrow{p} \Omega, \forall 1\cdots K.$$

Next we bound

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i\in I_k}(\hat{g}_{ki}\hat{g}'_{ki}) - \mathbb{E}\left(g_{0i}g'_{0i}\right)\right\| &\leq \left\|\frac{1}{m}\sum_{i\in I_k}(\hat{g}_{ki}\hat{g}'_{ki}) - \frac{1}{m}\sum_{i\in I_k}(g_{0i}g'_{0i})\right\| \\ &\quad + \left\|\frac{1}{m}\sum_{i\in I_k}(g_{0i}g'_{0i}) - \mathbb{E}\left(g_{0i}g'_{0i}\right)\right\| \\ &= B_1 + B_2. \end{aligned} \tag{26}$$

We first bound B_2. Weak law of large numbers directly yields:

$$\left\|\frac{1}{m}\sum_{i\in I_k}(g_{0i}g'_{0i}) - \mathbb{E}\left(g_{0i}g'_{0i}\right)\right\| = o_p(1). \tag{27}$$

To bound B_1, let $\Delta_{ki} = \hat{g}_{ki} - g_{0i}$. By triangle and Cauchy–Schwarz inequality:

$$\begin{aligned} B_2 &= \left\|\frac{1}{m}\sum_{i\in I_k}(\hat{g}_{ki}\hat{g}'_{ki}) - \frac{1}{m}\sum_{i\in I_k}(g_{0i}g'_{0i})\right\| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\|\Delta_{ki}\|\,\|2g_{0i}+\Delta_i\| \\ &\leq \left\{\frac{1}{m}\sum_{i=1}^{m}\|\Delta_{ki}\|^2\right\}^{1/2}\left\{\frac{1}{m}\sum_{i=1}^{m}\|2g_{0i}+\Delta_i\|^2\right\}^{1/2} \\ &\leq \left\{\frac{1}{m}\sum_{i=1}^{m}\|\Delta_{ki}\|^2\right\}^{1/2}\left\{\frac{8}{m}\sum_{i=1}^{m}\|g_{0i}\|^2 + \frac{2}{m}\sum_{i=1}^{m}\|\Delta_i\|^2\right\}^{1/2} \\ &= B_{21}^{1/2}\cdot[8B_{22}+2B_{21}]^{1/2}, \end{aligned}$$

where $B_{21} = \frac{1}{m}\sum_{i=1}^{m}\|\Delta_i\|^2$ and $B_{22} = \frac{1}{m}\sum_{i=1}^{m}\|g_{0i}\|^2$.

On event $\mathcal{E}_n$:

$$\begin{aligned} B_{21} &= \frac{1}{m}\sum_{i=1}^{m} \|\hat{g}_{ki} - g_{0i}\|^2 \\ &\leq \left\{ \max_{i\in I_k} \|\hat{g}_{ki} - g_{0i}\| \right\}^2 \\ &\leq \left\{ \sup_{\eta\in\mathcal{H}_n} \max_{i\in I_k} \|g(Z_i,\theta_0,\eta) - g_{0i}\| \right\}^2 \\ &= \tilde{r}_n^2. \end{aligned}$$

On the other hand, by weak law of large numbers,

$$\begin{aligned} B_{22} &= \frac{1}{m}\sum_{i=1}^{m} \|g_{0i}\|^2 \\ &\xrightarrow{p} \mathbb{E}\,\|g_{0i}\|^2 \\ &= O_p(1). \end{aligned}$$

Therefore,

$$\begin{aligned} B_2 &= \left\{\tilde{r}_n^2\right\}^{1/2} \left[O_p(1) + \tilde{r}_n^2\right]^{1/2} \\ &= O_p(\tilde{r}_n) \\ &= o_p(1). \end{aligned} \tag{28}$$

Combining (26), (27) and (28) completes the proof. □

Lemma 4.

Proof. Since K is fixed, it suffices to show:

$$\max_{i\in I_k} \|\hat{g}_{ki}\| = o_p(\sqrt{m}), \forall k = 1\cdots K.$$

Pick any $\delta > 0$:

$$\begin{aligned} \mathbb{P}\left\{ \max_{i\in I_k} \|\hat{g}_{ki}\| > 2\delta\sqrt{m} \right\} &= \mathbb{P}\left\{ \max_{i\in I_k} \|\hat{g}_{ki}\| > 2\delta\sqrt{m}; \max_{i\in I_k} \|\hat{g}_{ki}\| - \max_{i\in I_k} \|g_{0i}\| \geq \delta\sqrt{m} \right\} \\ &+ \mathbb{P}\left\{ \max_{i\in I_k} \|\hat{g}_{ki}\| > 2\delta\sqrt{m}; \max_{i\in I_k} \|\hat{g}_{ki}\| - \max_{i\in I_k} \|g_{0i}\| < \delta\sqrt{m} \right\} \\ &\leq \mathbb{P}\left\{ \max_{i\in I_k} \|\hat{g}_{ki} - g_{0i}\| \geq \delta\sqrt{m} \right\} \\ &+ \mathbb{P}\left\{ \max_{i\in I_k} \|g_{0i}\| > \delta\sqrt{m} \right\} \\ &= P_1 + P_2. \end{aligned}$$

Since $\mathbb{E}\{g_{0i}g_{0i}'\} < \infty$, Owen (Lemma 11.2, 2001) leads to $P_2 \to 0$ as $m \to 0$.

To bound P_1 , first notice by a conditional Markov inequality, on event $\mathcal{E}_n$:

$$\begin{aligned}\sum_{i\in I_k}\mathbb{P}\left[\|\hat{g}_{ki}-g_{0i}\|\geq\delta\sqrt{m}\mid(Z_j)_{j\in I_k^c}\right] &\leq \frac{\mathbb{E}\left[\|\hat{g}_{ki}-g_{0i}\|^2\mid(Z_j)_{j\in I_k^c}\right]}{\delta}\\ &\leq \frac{\mathbb{E}\left[\|\hat{g}_{ki}-g_{0i}\|^2\mid(Z_j)_{j\in I_k^c}\right]}{m\delta}\\ &\leq \frac{\sup_{\eta\in\mathcal{H}_n}\mathbb{E}\left(\|g(Z_i,\theta_0,\eta)-g_{0i}\|^2\right)}{\delta}\\ &= \frac{r_n^2}{\delta}\to 0.\end{aligned}$$

Since $\mathbb{P}\left[\|\hat{g}_{ki}-g_{0i}\|\geq\delta\sqrt{m}\mid(Z_j)_{j\in I_k^c}\right]$ is uniformly integrable, Theorem 25.12 of Billingsley (2008) implies as $m\to 0$

$$\begin{aligned}P_1 &\leq \mathbb{P}\left\{\max_{i\in I_k}\|\hat{g}_{ki}-g_{0i}\|\geq\delta\sqrt{m}\right\}\\ &\leq \sum_{i\in I_k}\mathbb{P}\left\{\|\hat{g}_{ki}-g_{0i}\|\geq\delta\sqrt{m}\right\}\\ &= \mathbb{E}\left\{\sum_{i\in I_k}\mathbb{P}\left[\|\hat{g}_{ki}-g_{0i}\|\geq\delta\sqrt{m}\mid(Z_j)_{j\in I_k^c}\right]\right\}\to 0.\end{aligned}$$

□

Theorem 2.

Proof. For ease of notation here we only consider the case when $d=1$, i.e., when X is scalar valued random variable. The case when $d\geq 2$ can be proved similarly. It suffices to verify conditions in Theorem 1.

Model (10):

Choose $\mathcal{H}_n$ to be the set of $\eta=(m^Y(W),m^X(W))'$ such that $\left\|m^Y(W_i)-m_0^Y(W_i)\right\|_\infty = o(n^{-\frac{1}{4}})$; $\left\|m^X(W_i)-m_0^X(W_i)\right\|_\infty = o(n^{-\frac{1}{4}})$. Define $\triangle^Y = m^Y(W)-m_0^Y(W)$; $\triangle^X = m^X(W)-m_0^X(W)$; $\triangle^h = h(W)-h_0(W)$. By assumption, $\hat{\eta}_k\in\mathcal{H}_n$ with probability at least $1-\varepsilon_n$ for all $k=1\cdots K$.

(A0): Based on Assumption B (1)-(3), A(0) will be satisfied if we can verify the existence of $\mathbb{E}\left[g^2(Z,\beta_0,\eta_0)\right]$. Notice $Y-m_0^Y=\beta_0\left[X-\mathbb{E}\left(X\mid W\right)\right]+u^Y=\beta_0u^X+u^Y$. So, by Cauchy–Schwarz inequality and Assumption B (2):

$$\begin{aligned}\mathbb{E}\left\{\left\{\left[Y-m_0^Y-\beta_0\left(X-m_0^X\right)\right]\left(X-m_0^X\right)\right\}^2\right\} &= \mathbb{E}\left\{\left(u^Y\right)^2\left(u^X\right)^2\right\}\\ &\leq \left\{\mathbb{E}\left[\left(u^Y\right)^4\right]\right\}^{\frac{1}{2}}\mathbb{E}\left[\left(u^X\right)^4\right]^{\frac{1}{2}}<\infty.\end{aligned}$$

A(1) and A(2): First notice by plug-in,

$$Y - m^Y - \beta_0 \left(X - m^X\right) = u^Y - \triangle^Y + \beta_0 \triangle^X.$$

Then for any $\eta \in \mathcal{H}_n$, we have for any $Z_i \in \mathcal{Z}$,

$$\begin{aligned}
|g(Z_i, \beta_0, \eta) - g(Z_i, \beta_0, \eta_0)| &= \left|\left(u_i^Y - \triangle_i^Y + \beta_0 \triangle_i^X\right)\left(u_i^X - \triangle_i^X\right) - u_i^Y u_i^X\right| \\
&= \left|-u_i^Y \triangle_i^X + \left(\beta_0 \triangle_i^X - \triangle_i^Y\right) u_i^X + \triangle_i^X \triangle_i^Y - \beta_0 \left(\triangle_i^X\right)^2\right| \\
&\leq \left|u_i^Y\right| \left|\triangle_i^X\right| + |\beta_0| \left(\left|\triangle_i^X\right| + \left|\triangle_i^Y\right|\right) \left|u_i^X\right| + \left|\triangle_i^X\right| \left|\triangle_i^Y\right| + |\beta_0| \left|\triangle_i^X\right|^2.
\end{aligned}$$

By an argument similar to Owen (Lemma 3, 1990), $\max_{i \in I_k} \left|u_i^Y\right| = o(n^{\frac{1}{4}})$, $\max_{i \in I_k} \left|u_i^X\right| = o(n^{\frac{1}{4}})$ by Assumption B (2). Therefore:

$$\begin{aligned}
\max_{i \in I_k} \|g(Z_i, \beta_0, \eta) - g(Z_i, \beta_0, \eta_0)\| &\leq \left\|\triangle_i^X\right\|_\infty \max_{i \in I_k} \left|u_i^Y\right| + |\beta_0| \left(\left\|\triangle_i^X\right\|_\infty + \left\|\triangle_i^Y\right\|_\infty\right) \max_{i \in I_k} \left|u_i^X\right| \\
&\quad + \left\|\triangle_i^X\right\|_\infty \left\|\triangle_i^Y\right\|_\infty + |\beta_0| \left\|\triangle_i^X\right\|_\infty^2 \\
&\leq o(n^{\frac{1}{4}}) o(n^{-\frac{1}{4}}) + |\beta_0| \, o(n^{\frac{1}{4}}) \left[o(n^{-\frac{1}{4}}) + o(n^{-\frac{1}{4}})\right] + |\beta_0| \, o(n^{-\frac{1}{4}})^2 \\
&= o(1).
\end{aligned}$$

Also, there exists some $C_1 > 0$ such that:

$$\begin{aligned}
\mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\|^2\right] &\leq C_1 \{\mathbb{E}[\left|u_i^Y\right|^2 \left|\triangle_i^X\right|^2] + \mathbb{E}[(\left|\triangle_i^X\right| + \left|\triangle_i^Y\right|)^2 \left|u_i^X\right|^2] \\
&\quad + \mathbb{E}[\left|\triangle_i^X\right|^2 \left|\triangle_i^Y\right|^2] + \mathbb{E}[\left|\triangle_i^X\right|^4]\} \\
&\leq C_1 \{\left\|\triangle_i^X\right\|_\infty^2 \mathbb{E}[(u_i^Y)^2] + (\left\|\triangle_i^X\right\|_\infty + \left\|\triangle_i^Y\right\|_\infty)^2 \mathbb{E}[(u_i^X)^2] \\
&\quad + \left\|\triangle_i^X\right\|_\infty^2 \left\|\triangle_i^Y\right\|_\infty^2 + \left\|\triangle_i^X\right\|_\infty^4\} \\
&= o(n^{-\frac{1}{2}}),
\end{aligned}$$

which leads to $\left\{\mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\|^2\right]\right\}^{\frac{1}{2}} = o(n^{-\frac{1}{4}}) = o(1)$.

A(3): Notice

$$\begin{aligned}
\left[\mathbb{E}g(Z, \beta_0, \eta)\right]_{\eta_0}^{(1)} \left[\eta - \eta_0\right] &= \frac{\partial \mathbb{E}g(Z, \beta_0, \eta_0 + t(\eta - \eta_0))}{\partial t} \Big|_{t=0} \\
&= -\mathbb{E}\left\{\left(m^Y - m_0^Y\right)\left(X - m_0^X\right)\right\} \\
&\quad + 2\beta_0 \mathbb{E}\left\{(X - m_0^X)(m^X - m_0^X)\right\} \\
&\quad - \mathbb{E}\left\{\left(m^X - m_0^X\right)\left(Y - m_0^Y\right)\right\}.
\end{aligned}$$

By law of iterated expectations and by definition of $m_0^X(W)$,

$$\mathbb{E}\left\{\left(m^Y - m_0^Y\right)\left(X - m_0^X\right)\right\} = \mathbb{E}\left\{\left[m^Y(W) - m_0^Y(W)\right] \mathbb{E}\left(X - m_0^X(W) \mid W\right)\right\} = 0.$$

Similarly,

$$\mathbb{E}\left\{(X - m_0^X)(m^X - m_0^X)\right\} = \mathbb{E}\left\{\mathbb{E}\left[(X - m_0^X(W)) \mid W\right](m^X(W) - m_0^X(W))\right\} = 0,$$

$$\mathbb{E}\left\{(m^X - m_0^X)\left(Y - m_0^Y\right)\right\} = \mathbb{E}\left\{(m^X - m_0^X)\mathbb{E}\left(Y - m_0^Y(W) \mid W\right)\right\} = 0.$$

Therefore, $[\mathbb{E}g(Z, \theta_0, \eta)]^{(1)}_{\eta_0}[\eta - \eta_0] = 0$.
(A4): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and $0 < \tilde{t} < 1$.

$$\begin{aligned}[\mathbb{E}g(Z, \theta_0, \eta)]^{(1)}_{\eta_0 + \tilde{t}(\eta - \eta_0)}[\eta - \eta_0] = &-\mathbb{E}\left\{\left(m^Y - m_0^Y\right)\left[X - \left[m_0^X + \tilde{t}\left(m^X - m_0^X\right)\right]\right]\right\} \\ &+ 2\beta_0\mathbb{E}\left\{(X - \left[m_0^X + \tilde{t}\left(m^X - m_0^X\right)\right])(m^X - m_0^X)\right\} \\ &- \mathbb{E}\left\{(m^X - m_0^X)\left(Y - \left[m_0^Y + \tilde{t}\left(m^Y - m_0^Y\right)\right]\right)\right\}.\end{aligned}$$

So:

$$\begin{aligned}[\mathbb{E}g(Z, \theta_0, \eta)]^{(2)}_{\eta_0 + \tilde{t}(\eta - \eta_0)}[\eta - \eta_0] = &2\mathbb{E}\left\{\left(m^Y - m_0^Y\right)\left(m^X - m_0^X\right)\right\} \\ &+ 2\beta_0\mathbb{E}\left\{\left(m^X - m_0^X\right)^2\right\}.\end{aligned}$$

Therefore:

$$\begin{aligned}\left|[\mathbb{E}g(Z, \theta_0, \eta)]^{(2)}_{\eta_0 + \tilde{t}(\eta - \eta_0)}[\eta - \eta_0]\right| &\leq 2\mathbb{E}\left\{\left|\left(m^Y - m_0^Y\right)\left(m^X - m_0^X\right)\right|\right\} \\ &+ 2\left|\beta_0\right|\mathbb{E}\left\{\left|\left(m^X - m_0^X\right)\right|^2\right\} \\ &\leq 2\left\|m^Y(W) - m_0^y(W)\right\|_\infty \left\|m^X(W) - m_0^X(W)\right\|_\infty \\ &+ 2\left|\beta_0\right|\left\|m^X(W) - m_0^X(W)\right\|_\infty^2 \\ &= o^2(n^{-\frac{1}{4}}) \\ &= o(n^{-\frac{1}{2}}).\end{aligned}$$

Thus we conclude verifying all conditions for model (10).
Model (11):
(A0): Existence of second moments:

$$\begin{aligned}\mathbb{E}\left\{\left[Y - \beta_0 X - h_0(W)\right]^2\left[X - m_0^X(W)\right]^2\right\} &= \mathbb{E}\left\{\left(u^Y\right)^2\left(u^X\right)^2\right\} \\ &\leq \left\{\mathbb{E}\left[\left(u^Y\right)^4\right]\right\}^{\frac{1}{2}}\mathbb{E}\left[\left(u^X\right)^4\right]^{\frac{1}{2}} \\ &< \infty.\end{aligned}$$

(A1) and (A2): For any $\eta \in \mathcal{H}_n$, we have for any $Z_i \in \mathcal{Z}$,

$$\begin{aligned}
|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)| &= \left|\left(u_i^Y - \triangle_i^h\right)\left(u_i^X - \triangle_i^X\right) - u_i^Y u_i^X\right| \\
&= \left|-\triangle_i^X u_i^Y - \triangle_i^h u_i^X + \triangle_i^h \triangle_i^X\right| \\
&\leq \left|\triangle_i^X\right|\left|u_i^Y\right| + \left|\triangle_i^h\right|\left|u_i^X\right| + \left|\triangle_i^h\right|\left|\triangle_i^X\right|.
\end{aligned}$$

Therefore, by the same argument used for model (10):

$$\begin{aligned}
\max_{i \in I_k} \|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\| &\leq \left\|\triangle_i^X\right\|_\infty \max_{i \in I_k}\left|u_i^Y\right| + \left\|\triangle_i^h\right\|_\infty \max_{i \in I_k}\left|u_i^X\right| + \left\|\triangle_i^X\right\|_\infty \left\|\triangle_i^Y\right\|_\infty \\
&= o(n^{\frac{1}{4}})o(n^{-\frac{1}{4}}) = o(1).
\end{aligned}$$

There also exists some $C_2 > 0$ such that:

$$\begin{aligned}
\mathbb{E}\left[\|g(Z_i, \theta_0, \eta) - g(Z_i, \theta_0, \eta_0)\|^2\right] &\leq C_2\{\left\|\triangle_i^X\right\|_\infty^2 \mathbb{E}[(u_i^Y)^2] + \left\|\triangle_i^h\right\|_\infty^2 \mathbb{E}[(u_i^X)^2] + \left\|\triangle_i^h\right\|_\infty \left\|\triangle_i^X\right\|_\infty\} \\
&= o(n^{-\frac{1}{2}}).
\end{aligned}$$

(A3): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and some $t > 0$ and $t \downarrow 0$. Then:

$$\begin{aligned}
[\mathbb{E}g(z, \beta_0, \eta)]_{\eta_0}^{(1)}[\eta - \eta_0] &= \frac{\partial \mathbb{E}g(Z, \beta_0, \eta_0 + t(\eta - \eta_0))}{\partial t}\Big|_{t=0} \\
&= -\mathbb{E}\left\{(h - h_0)\left(X - m_0^X\right)\right\} \\
&\quad - \mathbb{E}\left\{(Y - \beta_0 X - h_0)(m^X - m_0^X)\right\}.
\end{aligned}$$

The result follows by by using law of iterated expectations for the above two terms and definition $\mathbb{E}[Y \mid W] = \beta_0 \mathbb{E}(X \mid W) + h_0(W)$.

(A4): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and $0 < \tilde{t} < 1$.

$$\begin{aligned}
[\mathbb{E}g(Z, \beta_0, \eta)]_{\eta_0 + \tilde{t}(\eta - \eta_0)}^{(1)}[\eta - \eta_0] &= -\mathbb{E}\left\{(h - h_0)\left(X - \left(m_0^X + \tilde{t}\left(m^X - m_0^X\right)\right)\right)\right\} \\
&\quad - \mathbb{E}\left\{\left(Y - \beta_0 X - \left(h_0 + \tilde{t}\left(h - h_0\right)\right)\right)(m^X - m_0^X)\right\}.
\end{aligned}$$

So:

$$[\mathbb{E}g(Z, \theta_0, \eta)]_{\eta_0 + \tilde{t}(\eta - \eta_0)}^{(2)}[\eta - \eta_0] = 2\mathbb{E}\left\{(h - h_0)\left(m^X - m_0^X\right)\right\}.$$

Therefore:

$$\begin{aligned}
\left|[\mathbb{E}g(Z, \theta_0, \eta)]_{\eta_0 + \tilde{t}(\eta - \eta_0)}^{(2)}[\eta - \eta_0]\right| &\leq 2\|\eta(W_i) - \eta_0(W_i)\|_\infty \|m^x(W_i) - m_0^x(W_i)\|_\infty \\
&= o^2(n^{-\frac{1}{4}}) = o(n^{-\frac{1}{2}}).
\end{aligned}$$

□

Theorem 3.

Proof. It suffices to verify conditions in Theorem 1. Verification for model (15) is much more straightforward than model (13) and thus is omitted. Here we focus on (13). Assumption C (1) corresponds to regularity conditions in (A0). Assumption 3 (2) and (3) implies second moment of g exists and is non-singular. It remains to (A1)-(A4).

Define $\triangle^\pi = \pi - \pi_0; \triangle^m = m - m_0; \triangle^l = l - l_0$. Choose $\mathcal{H}_n$ to be the set of $\eta = (\pi, m, l)$ such that $\|\pi(X_i) - \pi_0(X_i)\|_\infty = O(r_n); \|m(X_i) - m_0(X_i)\|_\infty = O(r_n); \|l(X_i) - l_0(X_i)\|_\infty = O(r_n)$ and $\sqrt{n} r_n^2 \to 0$. By definition, $\hat{\eta}_k \in \mathcal{H}_n$ with probability at least $1 - \varepsilon_n$ for all $k = 1 \cdots K$.

(A1) and (A2): Rearrange terms we get:

$$\begin{aligned} g(Z, \theta_0, \eta) &= \frac{DY}{\pi} - \frac{(1-D)Y}{1-\pi} - (D - \pi)\left[\frac{m}{\pi} + \frac{l}{1-\pi}\right] - \theta_0 \\ &= m - l + \frac{D}{\pi}(Y - m) - \frac{(1-D)(Y-l)}{1-\pi} - \theta_0. \end{aligned}$$

Then

$$|g(Z, \theta_0, \eta) - g(Z, \theta_0, \eta_0)| \le T_1 + T_2 + T_3.$$

where

$$\begin{aligned} T_1 &= \left| \frac{D}{\pi}(Y - m) - \frac{D}{\pi_0}(Y - m_0) \right| \\ T_2 &= \left| \frac{(1-D)(Y-l)}{1-\pi} - \frac{(1-D)(Y-l_0)}{1-\pi_0} \right| \\ T_3 &= |\triangle^m| + \left|\triangle^l\right|. \end{aligned}$$

Apparently

$$\max_{i \in I_k} T_3 = o(n^{-\frac{1}{4}}), \tag{29}$$

$$\mathbb{E}(T_3^2) \le o^2(n^{-\frac{1}{4}}) = o(n^{-\frac{1}{2}}). \tag{30}$$

It remains to bound T_1 and T_2. Since:

$$\begin{aligned} T_1 &\le D\,|Y| \left| \frac{1}{\pi} - \frac{1}{\pi_0} \right| + D \left| \frac{m_0}{\pi_0} - \frac{m}{\pi} \right| \\ &\le D\,|Y| \frac{|\triangle^\pi|}{\pi \pi_0} + D \frac{|\triangle^\pi|\,|m_0|}{\pi_0 \pi} + D \frac{|\triangle^m|}{\pi}, \end{aligned}$$

we have for any $\eta \in \mathcal{H}_n$ and any $Z_i \in \mathcal{Z}$, by Assumption C (1), (2)and (4)

$$\begin{aligned} \max_{i \in I_k} T_1 &\le \frac{1}{\varepsilon^2} \|\triangle^\pi\|_\infty \max_{i \in I_k} |Y_i| + \frac{1}{\varepsilon^2} \|\triangle^\pi\|_\infty \|m_0\|_\infty + \frac{1}{\varepsilon} \|\triangle^m\|_\infty \\ &= o(n^{-\frac{1}{4}}) o(n^{\frac{1}{4}}) + o(n^{-\frac{1}{4}}) + o(n^{-\frac{1}{4}}) \\ &= o(1). \end{aligned} \tag{31}$$

The above illustration also shows that there exists some $C_3 > 0$ such that

$$\begin{aligned}\mathbb{E}(T_1^2) &\leq C_3\left(\frac{\mathbb{E}\left|Y\right|^2\left|\triangle^\pi\right|^2}{\varepsilon^4} + \frac{\mathbb{E}\left|\triangle^\pi\right|^2\left|m_0\right|^2}{\varepsilon^4} + \frac{\mathbb{E}\left|\triangle^m\right|^2}{\varepsilon^2}\right) \\ &\leq C_3\left(\frac{\left\|\triangle^\pi\right\|_\infty^2\mathbb{E}\left|Y\right|^2}{\varepsilon^4} + \frac{\left\|\triangle^\pi\right\|_\infty^2\mathbb{E}\left|m_0\right|^2}{\varepsilon^4} + \frac{\left\|\triangle^m\right\|_\infty^2}{\varepsilon^2}\right) \\ &\leq o(n^{-\frac{1}{2}}) = o(1). \end{aligned} \tag{32}$$

Similarly, for T_2:

$$T_2 \leq (1-D)\left|Y\right|\frac{\left|\triangle^\pi\right|}{(1-\pi)(1-\pi_0)} + (1-D)\frac{\left|\triangle^l\right|}{(1-\pi)} + (1-D)\frac{\left|l_0\right|\left|\triangle^\pi\right|}{(1-\pi)(1-\pi_0)}.$$

By Assumption C (1) (2) and (4), for any $\eta \in \mathcal{H}_n$ and any $Z_i \in \mathcal{Z}$:

$$\begin{aligned}\max_{i\in I_k} T_2 &\leq \frac{\max_{i\in I_k}\left|Y\right|}{\varepsilon^2}\left\|\triangle^\pi\right\|_\infty + \frac{\left\|\triangle^l\right\|_\infty}{\varepsilon} + \frac{\left\|l_0\right\|_\infty\left\|\triangle^\pi\right\|_\infty}{\varepsilon^2} \\ &= o(n^{\frac{1}{4}})o(n^{-\frac{1}{4}}) + o(n^{-\frac{1}{4}}) + o(n^{-\frac{1}{4}}) \\ &= o(1), \end{aligned} \tag{33}$$

and for some $C_4 > 0$

$$\begin{aligned}\mathbb{E}(T_2^2) &\leq C_4\left(\frac{\mathbb{E}\left|Y\right|^2\left|\triangle^\pi\right|^2}{\varepsilon^4} + \frac{\mathbb{E}\left|\triangle^\pi\right|^2\left|l_0\right|^2}{\varepsilon^4} + \frac{\mathbb{E}\left|\triangle^l\right|^2}{\varepsilon^2}\right) \\ &\leq o(n^{-\frac{1}{2}}) = o(1). \end{aligned} \tag{34}$$

Summing up, (29), (31), (33) confirms (A1); (30), (32), and (34) confirms (A2).

(A3): Recall $\eta_0 = (\pi_0, m_0, l_0)$, so for any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and some $t > 0$ and $t \downarrow 0$. Then:

$$\begin{aligned}\left[\mathbb{E}g(Z,\theta_0,\eta)\right]_{\eta_0}^{(1)}\left[\eta-\eta_0\right] &= \frac{\partial \mathbb{E}g(Z,\theta_0,\eta_0 + t(\eta-\eta_0))}{\partial t}\Big|_{t=0} \\ &= \mathbb{E}\left[\left(\frac{Dm_0}{\pi_0^2} - \frac{DY}{\pi_0^2}\right)(\pi-\pi_0)\right] \\ &+ \mathbb{E}\left\{\left[\frac{(1-D)\,l_0}{(1-\pi_0)^2} - \frac{(1-D)\,Y}{(1-\pi_0)^2}\right](\pi-\pi_0)\right\} \\ &- \mathbb{E}\left[\left(\frac{D-\pi_0}{\pi_0}\right)(m-m_0)\right] - \mathbb{E}\left[\left(\frac{D-\pi_0}{1-\pi_0}\right)(l-l_0)\right]. \end{aligned}$$

Notice $Y = Y(1)\cdot D + Y(0)\cdot(1-D)$, $D^2 = D$, $(1-D)D = 0$:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{Dm_0}{\pi_0^2} - \frac{DY}{\pi_0^2}\right)(\pi-\pi_0)\right] &= \mathbb{E}\left\{\left[\frac{D\left(m_0 - Y(1)\right)}{\pi_0^2}\right](\pi-\pi_0)\right\} \\ &= 0, \end{aligned} \tag{35}$$

where (35) is due to unconfoundedness, law of iterated expectations and$m_0 = \mathbb{E}[Y(1) \mid X]$. By the same argument,

$$\mathbb{E}\left\{\left[\frac{(1-D)\,l_0}{(1-\pi_0)^2} - \frac{(1-D)\,Y}{(1-\pi_0)^2}\right](\pi-\pi_0)\right\} = \mathbb{E}\left\{\left[\frac{(1-D)\,(l_0 - Y(0))}{(1-\pi_0)^2}\right](\pi-\pi_0)\right\}$$
$$= 0,$$

since $l_0 = \mathbb{E}[Y(0) \mid X]$. Similarly, $\mathbb{E}\left[\left(\frac{D-\pi_0}{\pi_0}\right)(m-m_0)\right] = 0$ and $\mathbb{E}\left[\left(\frac{D-\pi_0}{1-\pi_0}\right)(l-l_0)\right]$ because $\pi_0 = \mathbb{E}[D \mid X]$.

(A4): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and $0 < \tilde{t} < 1$.

$$\begin{aligned}
[\mathbb{E}g(z,\theta_0,\eta)]^{(1)}_{\eta_0+\tilde{t}(\eta-\eta_0)}[\eta-\eta_0] &= \frac{\partial \mathbb{E}g(Z,\theta_0,\eta_0+t\,(\eta-\eta_0))}{\partial t}\,|_{t=0} \\
&= \mathbb{E}\left[\left(\frac{D\left[m_0+\tilde{t}(m-m_0)\right]}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^2} - \frac{DY}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^2}\right)(\pi-\pi_0)\right] \\
&+ \mathbb{E}\left\{\left[\frac{(1-D)\left[l_0+\tilde{t}(l-l_0)\right]}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^2} - \frac{(1-D)\,Y}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0))\right]^2}\right](\pi-\pi_0)\right\} \\
&- \mathbb{E}\left[\left(\frac{D-\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]}{\pi_0+\tilde{t}(\pi-\pi_0)}\right)(m-m_0)\right] - \mathbb{E}\left[\left(\frac{D-\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]}{1-\pi_0-\tilde{t}(\pi-\pi_0)}\right)(l-l_0)\right].
\end{aligned}$$

So:

$$[\mathbb{E}g(z,\theta_0,\eta)]^{(2)}_{\eta_0+\tilde{t}(\eta-\eta_0)}[\eta-\eta_0] = T_4 + T_5 + T_6,$$

where

$$\begin{aligned}
T_4 &= \mathbb{E}\left\{\frac{DY(\pi-\pi_0)^2}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^3} - \frac{2D\left[m_0+\tilde{t}(m-m_0)\right](\pi-\pi_0)^2}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^3}\right\} \\
T_5 &= \mathbb{E}\left\{\frac{2\,(1-D)\left[l_0+\tilde{t}(l-l_0)\right](\pi-\pi_0)^2}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^3} - \frac{(1-D)\,Y(\pi-\pi_0)^2}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^3}\right\} \\
T_6 &= 2\mathbb{E}\left\{\frac{D(\pi-\pi_0)\,(m-m_0)}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^2}\right\} + 2\mathbb{E}\left\{\frac{(1-D)\,(\pi-\pi_0)\,(l-l_0)}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^2}\right\}.
\end{aligned}$$

Since $\varepsilon \le \|\pi_0\|_\infty \le 1-\varepsilon$, $\varepsilon \le \|\pi\|_\infty \le 1-\varepsilon$ by assumption, we have

$$\varepsilon \le \pi_0 + \tilde{t}(\pi-\pi_0) = \tilde{t}\pi + (1-\tilde{t})\pi_0 \le 1-\varepsilon.$$

Also notice

$$\begin{aligned}
\left\|m_0+\tilde{t}(m-m_0)\right\|_\infty &\le \|m_0\|_\infty + \|m\|_\infty \\
\left\|l_0+\tilde{t}(l-l_0)\right\|_\infty &\le \|l_0\|_\infty + \|l\|_\infty.
\end{aligned}$$

Therefore:

$$
\begin{aligned}
|T_4| &\leq \mathbb{E}\left|\frac{Y(\pi-\pi_0)^2}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^3}\right| + 2\mathbb{E}\left|\frac{\left[m_0+\tilde{t}(m-m_0)\right](\pi-\pi_0)^2}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^3}\right| \\
&\leq \varepsilon^3 \left\|\triangle^\pi\right\|_\infty^2 \mathbb{E}\left|Y\right| + 2\varepsilon^3 \left\|\triangle^\pi\right\|_\infty^2 \left(\left\|m_0\right\|_\infty + \left\|m\right\|_\infty\right) \\
&= o^2(n^{-\frac{1}{4}}) = o(n^{-\frac{1}{2}}).
\end{aligned}
$$

$$
\begin{aligned}
|T_5| &\leq 2\mathbb{E}\left|\frac{\left[l_0+\tilde{t}(l-l_0)\right](\pi-\pi_0)^2}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^3}\right| + \mathbb{E}\left|\frac{Y(\pi-\pi_0)^2}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^3}\right| \\
&\leq 2\varepsilon^3 \left\|\triangle^\pi\right\|_\infty^2 \left(\left\|l_0\right\|_\infty + \left\|l\right\|_\infty\right) + \varepsilon^3\mathbb{E}\left|Y\right| \left\|\triangle^\pi\right\|_\infty^2 \\
&= o^2(n^{-\frac{1}{4}}) = o(n^{-\frac{1}{2}}).
\end{aligned}
$$

$$
\begin{aligned}
|T_6| &= 2\mathbb{E}\left|\frac{(\pi-\pi_0)(m-m_0)}{\left[\pi_0+\tilde{t}(\pi-\pi_0)\right]^2}\right| + 2\mathbb{E}\left|\frac{(\pi-\pi_0)(l-l_0)}{\left[1-\pi_0-\tilde{t}(\pi-\pi_0)\right]^2}\right| \\
&\leq 2\varepsilon^2 \left\|\triangle^\pi\right\|_\infty \left\|\triangle^m\right\|_\infty + 2\varepsilon^2 \left\|\triangle^\pi\right\|_\infty \left\|\triangle^l\right\|_\infty \\
&= o^2(n^{-\frac{1}{4}}) = o(n^{-\frac{1}{2}}).
\end{aligned}
$$

□

Theorem 4.

Proof. We verify conditions listed in Theorem 1. For simplicity we only consider the case when $\dim(\beta_0) = 1$. Since $\frac{m_0(W)}{l_0(W)} = h_0(W)$, under Assumption D (1)-(3), Assumption (A0) are apparently satisfied for both (17) and (18). We focus on (A1)-(A4).

Model (17):

Define $\triangle^h = h - h_0; \triangle^\pi = \pi - \pi_0$. We choose $\mathcal{H}_n$ to be the set of $\eta = (h(W), \pi(W))'$ such that $\left\|h(W) - h_0(W)\right\|_\infty = o(n^{-\frac{1}{4}}); \left\|\pi(W) - \pi_0(W)\right\|_\infty = o(n^{-\frac{1}{4}})$. By definition, $\hat{\eta}_k \in \mathcal{H}_n$ with probability at least $1 - \varepsilon_n$ for all $k = 1 \cdots K$.

(A1) and (A2): Notice by definition and plug-in:

$$
\begin{aligned}
g(Z, \beta_0, \eta_0) &= \exp(-\beta_0 X) u^Y u^X \\
g(z, \beta_0, \eta) &= \left[\exp(-\beta_0 X) u^Y - \triangle^h\right]\left(u^X - \triangle^\pi\right),
\end{aligned}
$$

which leads to

$$|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)| = \left|\exp(-\beta_0 X_i)u_i^Y\triangle_i^\pi+\triangle_i^h u_i^X-\triangle_i^h\triangle_i^\pi\right|$$
$$\leq \exp(-\beta_0 X_i)\left|u_i^Y\right|\left|\triangle_i^\pi\right|+\left|\triangle_i^h\right|\left|u_i^X\right|+\left|\triangle_i^h\right|\left|\triangle_i^\pi\right|.$$

By Assumption D (1), (2) and (4):

$$\begin{aligned}\max_{i\in I_k}|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)| &\leq \sup_{x\in\mathcal{X}}\left[\exp(-\beta_0 x)\right]\left\|\triangle_i^\pi\right\|_\infty \max_{i\in I_k}\left|u_i^Y\right|+\left\|\triangle_i^h\right\|_\infty\max_{i\in I_k}\left|u_i^X\right|\\ &+\left\|\triangle_i^\pi\right\|_\infty\left\|\triangle_i^h\right\|_\infty\\ &\leq O(1)o(n^{-\frac{1}{4}})o(n^{\frac{1}{4}})+o(n^{-\frac{1}{4}})o(n^{\frac{1}{4}})+o^2(n^{-\frac{1}{4}})\\ &= o(1).\end{aligned}$$

And there exists some $C_5>0$ such that

$$\begin{aligned}\mathbb{E}\left[|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)|^2\right] &\leq C_5\{\mathbb{E}\left[\exp(-2\beta_0 X_i)\left|u_i^Y\right|^2\left|\triangle_i^\pi\right|^2\right]+\mathbb{E}\left[\left|\triangle_i^h\right|^2\left|u_i^X\right|^2\right]\\ &+\mathbb{E}\left[\left|\triangle_i^h\right|^2\left|\triangle_i^\pi\right|^2\right]\}\\ &\leq C_5\{\sup_{x\in\mathcal{X}}\left[\exp(-2\beta_0 x)\right]\left\|\triangle_i^\pi\right\|_\infty^2\mathbb{E}\left(\left|u_i^Y\right|^2\right)+\left\|\triangle_i^h\right\|_\infty^2\mathbb{E}\left(\left|u_i^X\right|^2\right)\\ &+\left\|\triangle_i^\pi\right\|_\infty^2\left\|\triangle_i^h\right\|_\infty^2\}\\ &\leq O(1)o^2(n^{-\frac{1}{4}})O(1)+o^2(n^{-\frac{1}{4}})O(1)+o^4(n^{-\frac{1}{4}})\\ &= o(n^{-\frac{1}{2}})=o(1).\end{aligned}$$

(A3):

$$\begin{aligned}\left[\mathbb{E}g(Z,\beta_0,\eta)\right]^{(1)}_{\eta_0}\left[\eta-\eta_0\right] &= \frac{\partial\mathbb{E}g(Z,\beta_0,\eta_0+t(\eta-\eta_0))}{\partial t}\mid_{t=0}\\ &= -\mathbb{E}\left[(h-h_0)(X-\pi_0)\right]-\mathbb{E}\left\{\left[\exp(-\beta_0'X)Y-h_0\right](\pi-\pi_0)\right\}.\end{aligned}$$

By law of iterated expectations and definition of h_0 and π_0:

$$\mathbb{E}\left[(X-\pi_0)(h-h_0)\right]=\mathbb{E}\left[\mathbb{E}\left[(X-\pi_0)\mid W\right](h(W)-h_0(W))\right]=0,$$

$$\mathbb{E}\left\{\left[\exp(-\beta_0'X)Y-h_0(W)\right](\pi-\pi_0)\right\}=\mathbb{E}\left\{\left[\exp(-\beta_0'X)\mathbb{E}(Y\mid W,X)-h_0(W)\right]\left[\pi(W)-\pi_0(W)\right]\right\}=0.$$

(A4): For any $\eta\in\mathcal{H}_n$, pick direction $(\eta-\eta_0)$ and $0<\tilde{t}<1$:

$$\begin{aligned}\left[\mathbb{E}g(Z,\beta_0,\eta)\right]^{(1)}_{\eta_0+\tilde{t}(\eta-\eta_0)}\left[\eta-\eta_0\right] &= -\mathbb{E}\left[(h-h_0)\left(X-\left(\pi_0+\tilde{t}(\pi-\pi_0)\right)\right)\right]\\ &-\mathbb{E}\left\{\left[\exp(-\beta_0'X)Y-\left(h_0+\tilde{t}(h-h_0)\right)\right](\pi-\pi_0)\right\}\end{aligned}$$

$$[\mathbb{E}g(Z,\beta_0,\eta)]^{(2)}_{\eta_0+\tilde{t}(\eta-\eta_0)}[\eta-\eta_0] = 2\mathbb{E}\left[(h-h_0)(\pi-\pi_0)\right].$$

Therefore:

$$\begin{aligned}\left|[\mathbb{E}g(Z,\beta_0,\eta)]^{(2)}_{\eta_0+\tilde{t}(\eta-\eta_0)}[\eta-\eta_0]\right| &\le 2\mathbb{E}\left\{|(h-h_0)(\pi-\pi_0)|\right\}\\ &\le 2\left\|h(W_i)-h_0(W_i)\right\|_\infty \left\|\pi(W_i)-\pi_0(W_i)\right\|_\infty\\ &= o(n^{-\frac{1}{2}}).\end{aligned}$$

Model (18): Additionally define $\triangle^m = m - m_0$, $\triangle^l = l - l_0$.
(A1) and (A2): Notice by plug-in:

$$\begin{aligned}g(Z,\beta_0,\eta) &= \left[\exp(-\beta_0 X)Y - \frac{m}{l}\right](x-\pi)\\ &= \left[\exp(-\beta_0 X)u^Y + \frac{h_0\triangle^l}{l} - \frac{\triangle^m}{l}\right]\left(u^X - \triangle^\pi\right).\end{aligned}$$

So:

$$|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)| \le \exp(-\beta_0 X_i)\left|u_i^Y\triangle_i^\pi\right| + \left|\frac{h_0 u_i^X\triangle_i^l}{l}\right| + \left|\frac{\triangle_i^m u_i^X}{l}\right| + \left|\frac{h_0\triangle_i^l\triangle_i^\pi}{l}\right| + \left|\frac{\triangle_i^m\triangle_i^\pi}{l}\right|.$$

By Assumption D (1) (2) (4) we have:

$$\begin{aligned}\max_{i\in I_k}|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)| &\le \sup_{x\in\mathcal{X}}\left[\exp(-\beta_0 x)\right]o(n^{-\frac{3}{8}})o(n^{\frac{1}{4}}) + \frac{\bar{\delta}}{\underline{\delta}}o(n^{-\frac{3}{8}})o(n^{\frac{1}{4}})\\ &\quad + \frac{1}{c}o(n^{-\frac{3}{8}})o(n^{\frac{1}{4}}) + \frac{\bar{\delta}}{\underline{\delta}}o^2(n^{-\frac{3}{8}}) + \frac{o^2(n^{-\frac{3}{8}})}{\underline{\delta}}\\ &= o(n^{-\frac{1}{8}}).\end{aligned}$$

And similarly there exists some $C_6 > 0$ such that

$$\begin{aligned}\mathbb{E}\left[|g(Z_i,\beta_0,\eta)-g(Z_i,\beta_0,\eta_0)|^2\right] &\le C_6\{\mathbb{E}\left[\exp(-2\beta_0 X_i)\left|u_i^Y\right|^2\left|\triangle_i^\pi\right|^2\right] + \mathbb{E}\left|\frac{h_0^2\left(u_i^X\right)^2\left(\triangle_i^l\right)^2}{l^2}\right|\\ &\quad + \mathbb{E}\left|\frac{\left(\triangle_i^m\right)^2\left(u_i^X\right)^2}{l^2}\right| + \mathbb{E}\left|\frac{h_0^2\left(\triangle_i^l\right)^2\left(\triangle_i^\pi\right)^2}{l^2}\right| + \mathbb{E}\left|\frac{\left(\triangle_i^m\right)^2\left(\triangle_i^\pi\right)^2}{l^2}\right|\}\\ &\le O(1)o^2(n^{-\frac{3}{8}}) + o^2(n^{-\frac{3}{8}})O(1) + o^2(n^{-\frac{3}{8}})O(1) + O(1)o^4(n^{-\frac{3}{8}})\\ &= o(n^{-\frac{3}{4}}).\end{aligned}$$

(A3): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and some $t > 0$ and $t \downarrow 0$:

$$\begin{aligned}
[\mathbb{E}g(Z, \beta_0, \eta)]^{(1)}_{\eta_0} [\eta - \eta_0] &= \frac{\partial \mathbb{E}g(z, \beta_0, \eta_0 + t(\eta - \eta_0))}{\partial t} \mid_{t=0} \\
&= -\mathbb{E}\left[\frac{m - m_0}{l_0}(X - \pi_0)\right] \\
&+ \mathbb{E}\left[\frac{m_0}{l_0^2}(l - l_0)(X - \pi_0)\right] \\
&- \mathbb{E}\left\{\left[\exp(-\beta_0 X)Y - \frac{m_0}{l_0}\right](\pi - \pi_0)\right\}.
\end{aligned}$$

Therefore (A3) is satisfied by by invoking law of iterated expectations.
(A4): For any $\eta \in \mathcal{H}_n$, pick direction $(\eta - \eta_0)$ and $0 < \tilde{t} < 1$.

$$\begin{aligned}
[\mathbb{E}g(Z, \beta_0, \eta)]^{(1)}_{\eta_0 + \tilde{t}(\eta - \eta_0)} [\eta - \eta_0] &= \mathbb{E}\left[-\frac{m - m_0}{l_0 + \tilde{t}(l - l_0)}\left[X - \left(\pi_0 + \tilde{t}(\pi - \pi_0)\right)\right]\right] \\
&+ \mathbb{E}\left[\frac{m_0 + \tilde{t}(m - m_0)}{\left[l_0 + \tilde{t}(l - l_0)\right]^2}\left[X - \left(\pi_0 + \tilde{t}(\pi - \pi_0)\right)\right](l - l_0)\right] \\
&\mathbb{E}\left\{-\left[\exp(-\beta_0 X)Y - \frac{m_0 + \tilde{t}(m - m_0)}{l_0 + \tilde{t}(l - l_0)}\right](\pi - \pi_0)\right\}.
\end{aligned}$$

So:

$$[\mathbb{E}g(Z, \beta_0, \eta)]^{(2)}_{\eta_0 + \tilde{t}(\eta - \eta_0)} [\eta - \eta_0] = \nabla_1 + \nabla_2 + \nabla_3,$$

where

$$\begin{aligned}
\nabla_1 &= \mathbb{E}\frac{\triangle^m \triangle^\pi}{\left[l_0 + \tilde{t}\triangle^l\right]} + \mathbb{E}\frac{\triangle^m \triangle^l \left(u^X - \tilde{t}\triangle^\pi\right)}{\left[l_0 + \tilde{t}\triangle^l\right]^2} \\
\nabla_2 &= \mathbb{E}\left[\frac{\triangle^l \triangle^m \left[u^X - \tilde{t}\triangle^\pi\right]}{\left[l_0 + \tilde{t}\triangle^l\right]^2} - \frac{\triangle^l \triangle^\pi \left[m_0 + \tilde{t}\triangle^m\right]}{\left[l_0 + \tilde{t}\triangle^l\right]^2}\right] \\
&- 2\mathbb{E}\left[\frac{\left(m_0 + \tilde{t}\triangle^m\right)\left[u^X - \tilde{t}\triangle^\pi\right](\triangle^l)^2}{\left[l_0 + \tilde{t}\triangle^l\right]^3}\right] \\
\nabla_3 &= \mathbb{E}\left[\frac{\triangle^\pi \triangle^m}{l_0 + \tilde{t}\triangle^l} - \frac{\triangle^\pi \triangle^l \left(m_0 + \tilde{t}\triangle^m\right)}{\left(l_0 + \tilde{t}\triangle^l\right)^2}\right].
\end{aligned}$$

After rearranging terms we get:

$$[\mathbb{E}g(Z, \beta_0, \eta)]^{(2)}_{\eta_0 + \tilde{t}(\eta - \eta_0)} [\eta - \eta_0] = 2(\nabla_4 + \nabla_5 + \nabla_6 + \nabla_7),$$

where $\nabla_4 = \mathbb{E}\frac{\triangle^m \triangle^\pi}{\left[l_0 + \tilde{t}\triangle^l\right]}$, $\nabla_5 = \mathbb{E}\frac{\triangle^m \triangle^l \left(u^X - \tilde{t}\triangle^\pi\right)}{\left[l_0 + \tilde{t}\triangle^l\right]^2}$, $\nabla_6 = -\mathbb{E}\frac{\triangle^l \triangle^\pi \left[m_0 + \tilde{t}\triangle^m\right]}{\left[l_0 + \tilde{t}\triangle^l\right]^2}$, and $\nabla_7 = -\mathbb{E}\left[\frac{\left(m_0 + \tilde{t}\triangle^m\right)\left[u^x - \tilde{t}\triangle^\pi\right](\triangle^l)^2}{\left[l_0 + \tilde{t}\triangle^l\right]^3}\right]$.

Notice by Assumption D (4):$\left\|l_0+\tilde{t}\triangle^l\right\|_\infty=\left\|(1-\tilde{t})l_0+\tilde{t}l\right\|_\infty\geq\underline{\delta}$. Thus Assumption D (2) and (4) yield:

$$|\nabla_4|\leq\frac{1}{\underline{\delta}}o^2(n^{-\frac{3}{8}})=o(n^{-\frac{3}{4}}); \tag{36}$$

$$\begin{aligned}|\nabla_5|&=\frac{1}{\underline{\delta}^2}o^2(n^{-\frac{3}{8}})(o(n^{\frac{1}{4}})+o(n^{-\frac{3}{8}}))\\&=o(n^{-\frac{1}{2}});\end{aligned} \tag{37}$$

$$|\nabla_6|=\frac{1}{\underline{\delta}^2}o^2(n^{-\frac{3}{8}})O(1)=o(n^{-\frac{3}{4}}); \tag{38}$$

$$\begin{aligned}|\nabla_7|&=\frac{1}{\underline{\delta}^3}O(1)\left[o(n^{\frac{1}{4}})+o(n^{-\frac{3}{8}})\right]o^2(n^{-\frac{3}{8}})\\&=o(n^{-\frac{1}{2}}).\end{aligned} \tag{39}$$

Combining (36) (37) (38) and (39) concludes the proof. □

Appendix C. Tables and Figures

Figure 1. Kernel density of propensity score under strong and weak signals

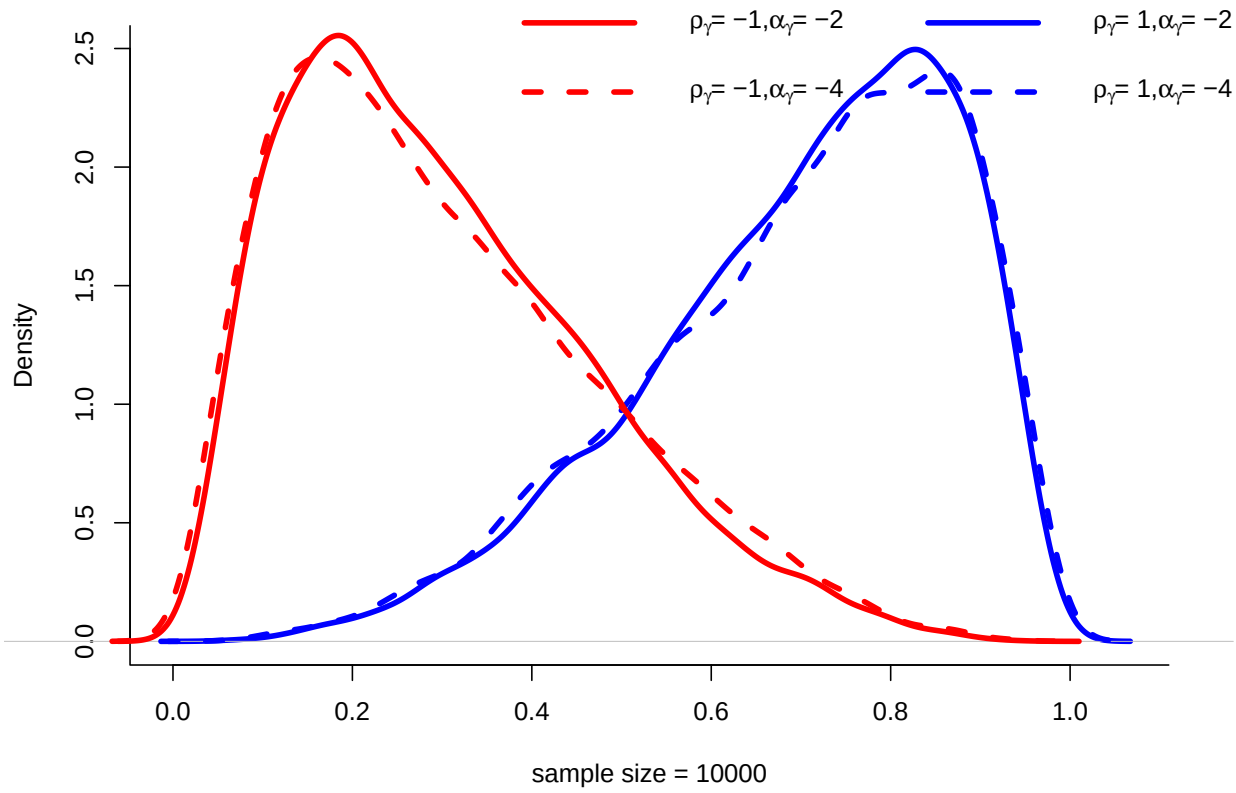

TABLE 1. Coverage probability for population mean when propensity score signal is strong

ρ_γ						1						
ρ_1						1						
α_γ				-2						-4		
α_1		-2			-4			-2			-4	
	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald
$K = 2$	0.913	0.923	0.919	0.924	0.925	0.932	0.928	0.932	0.921	0.921	0.931	0.926
$K = 4$	0.937	0.943	0.920	0.920	0.944	0.931	0.933	0.928	0.949	0.944	0.938	0.940
$K = 5$	0.923	0.930	0.924	0.941	0.931	0.935	0.962	0.928	0.931	0.949	0.940	0.933
Baseline		0.956	0.957		0.944	0.943		0.953	0.948		0.953	0.939
ρ_1						3						
α_γ				-2						-4		
α_1		-2			-4			-2			-4	
	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald
$K = 2$	0.930	0.939	0.938	0.944	0.944	0.948	0.935	0.940	0.935	0.949	0.934	0.947
$K = 4$	0.949	0.954	0.943	0.948	0.954	0.958	0.950	0.944	0.955	0.947	0.944	0.961
$K = 5$	0.936	0.941	0.953	0.947	0.944	0.927	0.947	0.943	0.936	0.947	0.929	0.946
Baseline		0.946	0.946		0.946	0.949		0.930	0.941		0.952	0.945

Note: This table reports empirical coverage probability when nominal rejection rate is set at 5% and when propensity score signal is strong ($\rho_\gamma = 1$). The parameter of interest is ϑ_0, a population mean in the presence of missing data. EL1 and Wald use the locally robust score based on (19) and EL 2 uses (20). EL 1 and EL 2 correspond to the empirical likelihood ratio procedure and Wald refers to a Wald ratio procedure. All nuisance parameters are constructed using Lasso with penalty level chosen by 10 fold cross-validation. Sample size is 500 and Monte Carlo experiment is repeated 1000 times. For comparison, the baseline case is provided when both nuisance parameters are known. Under this case EL1 and EL2 are trivially the same model.

TABLE 2. Coverage probability when propensity score signal is weak

ρ_γ	-1											
ρ_1	1											
α_γ	-2						-4					
α_1	-2			-4			-2			-4		
	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald
$K = 2$	0.767	0.865	0.760	0.849	0.869	0.810	0.830	0.862	0.805	0.840	0.881	0.796
$K = 4$	0.866	0.900	0.826	0.864	0.899	0.880	0.859	0.912	0.855	0.879	0.899	0.878
$K = 5$	0.843	0.909	0.837	0.899	0.907	0.866	0.881	0.903	0.861	0.868	0.913	0.876
Baseline	0.949		0.955	0.942		0.959	0.947		0.944	0.941		0.942
ρ_1	3											
α_γ	-2						-4					
α_1	-2			-4			-2			-4		
	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald	EL 1	EL 2	Wald
$K = 2$	0.821	0.851	0.796	0.901	0.915	0.866	0.874	0.905	0.872	0.893	0.906	0.890
$K = 4$	0.862	0.901	0.860	0.911	0.909	0.900	0.911	0.927	0.907	0.908	0.918	0.925
$K = 5$	0.880	0.905	0.867	0.920	0.927	0.906	0.905	0.915	0.916	0.915	0.929	0.912
Baseline	0.951		0.935	0.945		0.947	0.941		0.952	0.942		0.953

Note: This table reports empirical coverage probability when nominal rejection rate is set at 5% and when propensity score signal is weak ($\rho_\gamma = -1$). The parameter of interest is ϑ_0, a population mean in the presence of missing data. EL1 and Wald use locally robust score based on (19) and EL 2 uses (20). EL 1 and EL 2 correspond to the empirical likelihood ratio procedure and Wald refers to a Wald ratio procedure. All nuisance parameters are constructed using Lasso with penalty level chosen by 10 fold cross-validation. Sample size is 500 and Monte Carlo experiment is repeated 1000 times. For comparison, the baseline case is provided when both nuisance parameters are known. Under this case EL1 and EL2 are trivially the same model.

TABLE 3. Empirical coverage probability for β_0 in partly log linear model

β_0	1		0.5		-0.5		-1	
	EL	Wald	EL	Wald	EL	Wald	EL	Wald
Panel A: original Z								
$K=2$	0.785	0.671	0.752	0.654	0.741	0.654	0.748	0.668
$K=4$	0.899	0.785	0.868	0.733	0.837	0.715	0.862	0.733
Panel B: original Z, with π_0 known								
$K=2$	0.888	0.747	0.886	0.783	0.899	0.762	0.885	0.771
$K=4$	0.885	0.750	0.877	0.770	0.891	0.764	0.891	0.780
Panel C: Z polynomial of order 2								
$K=2$,$\alpha=0$	0.749	0.668	0.729	0.649	0.755	0.704	0.807	0.763
$K=4$, $\alpha=0$	0.814	0.720	0.794	0.678	0.773	0.727	0.835	0.763
$K=2$, $\alpha=0.3$	0.781	0.674	0.764	0.683	0.761	0.700	0.767	0.762
$K=4$, $\alpha=0.3$	0.833	0.730	0.827	0.699	0.794	0.712	0.845	0.748
$K=2$, $\alpha=0.7$	0.780	0.676	0.773	0.685	0.749	0.721	0.782	0.749
$K=4$, $\alpha=0.7$	0.842	0.756	0.837	0.723	0.812	0.728	0.824	0.784
$K=2$, $\alpha=1$	0.791	0.689	0.785	0.671	0.766	0.704	0.786	0.749
$K=4$, $\alpha=1$	0.858	0.715	0.821	0.720	0.820	0.711	0.824	0.736
Panel D: Z polynomial of order 2, with π_0 known								
$K=2$,$\alpha=0$	0.834	0.750	0.840	0.736	0.821	0.761	0.840	0.756
$K=4$, $\alpha=0$	0.851	0.747	0.819	0.731	0.837	0.758	0.851	0.770
$K=2$, $\alpha=0.3$	0.836	0.742	0.843	0.747	0.837	0.750	0.841	0.789
$K=4$, $\alpha=0.3$	0.850	0.746	0.844	0.735	0.840	0.761	0.841	0.757
$K=2$, $\alpha=0.7$	0.875	0.765	0.822	0.749	0.808	0.743	0.844	0.755
$K=4$, $\alpha=0.7$	0.843	0.744	0.854	0.765	0.838	0.710	0.828	0.776
$K=2$, $\alpha=1$	0.835	0.754	0.855	0.736	0.839	0.742	0.840	0.806
$K=4$, $\alpha=1$	0.862	0.753	0.853	0.752	0.848	0.736	0.860	0.767
Panel E: both π_0 and h_0 known								
	0.866	0.972	0.911	0.984	0.911	0.976	0.867	0.981

Note: This table reports empirical coverage probability when nominal rejection rate is set at 5% for β_0 in partially log linear model. The true value of β_0 is set at 1, 0.5, -0.5, or -1. In Panel A and B nuisance parameters are estimated based on covariates W; In Panel C and D they are estimated using a polynomial of order 2 based on W. For comparison Panel B and D assume that π_0 is known, and Panel E assumes both nuisance parameters are known. The locally robust score moment equation is evaluated through (17). EL ratio is constructed according to (3). For Wald statistic, the point estimate as well as variance estimate is based on a trivial GMM procedure setting weight equal 1. All nuisance parameters are estimated in a linear way plus a linear combination of l_1 and l_2 penalty. In Panel C and D α denotes the weight to l_1 penalty. In Panel A and B no penalty is imposed. 10 fold cross validation is used to choose the final data-driven penalty level. Sample size is 100 and Monte Carlo experiment is repeated 1000 times.

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Atkinson, K. and Han, W. (2005). *Theoretical numerical analysis*, volume 39. Springer.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Lasso methods for gaussian instrumental variables models.

Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli*, 12(2):299–331.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 10(3):647–671.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Bravo, F., Escanciano, J. C., and Van Keilegom, I. (2015). Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference. 16(2015).

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, a. W. (2016a). Double Machine Learning for Treatment and Causal Parameters. *ArXiv*, page 30.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017). Double/debiased machine learning for treatment and causal parameters. Technical report, arXiv. org.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. K. (2016b). Locally Robust Semiparametric Estimation. pages 1–42.

Donald, S. and Newey, W. (1994). Series Estimation of Semilinear Models.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, pages 1079–1111.

Horowitz, J. L., Mammen, E., et al. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32(6):2412–2443.

Ichimura, H. and Newey, W. K. (2015). The Influence Function of Semiparametric Estimators.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice.

Matsushita, Y. and Otsu, T. (2017). Likelihood inference on semiparametric models: average derivative and treatment effect. pages 1–21.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K., Robins, J. M., Newey, W. K., and Robins, J. M. (2017). Cross-fitting and fast remainder rates for semiparametric estimation Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *Working Paper*.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18(1):90–120.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747.

Owen, A. B. (2001). *Empirical likelihood*. Wiley Online Library.

Pfanzagl, J. (1982). Lecture notes in statistics. *Contributions to a general asymptotic statistical theory*, 13.

Qiu, C. and Otsu, T. (2018). Information theoretic approach to high dimensional multiplicative models: Stochastic discount factor and treatment effect.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Robins, J. M., Zhang, P., Ayyagari, R., Logan, R., Tchetgen, E. T., Li, L., Lumley, T., and van der Vaart, A. (2013). New statistical approaches to semiparametric regression with application to air pollution research. *Health Effects Institute*, 175(175):3–129.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothe, C. and Firpo, S. (2016). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. Technical report, Working paper.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: c.qiu@lse.ac.uk