REWEIGHTED NONPARAMETRIC LIKELIHOOD INFERENCE FOR LINEAR FUNCTIONALS

KARUN ADUSUMILLI, TAISUKE OTSU, AND CHEN QIU

Abstract. This paper is concerned with inference on finite dimensional parameters in semiparametric moment condition models, where the moment functionals are linear with respect to unknown nuisance functions. By exploiting this linearity, we reformulate the inference problem via the Riesz representer, and develop a general inference procedure based on nonparametric likelihood. For treatment effect or missing data analysis, the Riesz representer is typically associated with the inverse propensity score even though the scope of our framework is much wider. In particular, we propose a two-step procedure, where the first step computes the projection weights to approximate the Riesz representer, and the second step re-weights the moment conditions so that the likelihood increment admits an asymptotically pivotal chi-square calibration. Our re-weighting method is naturally extended to inference on treatment effects and data combination models, and other semiparametric problems. Simulation and empirical examples illustrate usefulness of the proposed method.

1. Introduction

There is a broad and important class of semiparametric models where finite dimensional parameters of interest are defined by moment conditions involving unknown nuisance functions, such as conditional mean functions. Examples include inference on missing at random observations, treatment effects, policy interventions, weighted average derivatives, and data combination models. Many such semiparametric models share a common feature: the moment functions for identifying the finite dimensional parameters are *linear* with respect to the unknown functions. By using the Riesz representation theorem, this linearity allows us to reformulate the original moment conditions as multiplicative or weighted moment conditions, where the weight function is given by the so called Riesz representer.

In the context of treatment effect analysis, this weight function is associated with balancing weights or the inverse propensity score. Recently, several methods that directly balance the distributional characteristics of covariates have been proposed (Hainmueller, 2012; Zubizarreta, 2015; Graham, Pinto and Egel, 2012, 2016; Athey, Imbens and Wager, 2018; Chan, Yam and Zhang, 2016). These methods, based on balancing weights, have been employed to obtain point estimates for the population parameters of interest, and the above authors have reported superior performance for the empirical balancing approach.

The purpose of this paper is to (i) develop a general framework of semiparametric inference for linear functional models (which covers the balancing weights as a special case), and (ii) construct a new weighting scheme that would be useful for conducting statistical inference (i.e., interval estimation and hypothesis testing) as opposed to point estimation. In our framework, the balancing weights are interpreted as the Riesz representer for the moment conditions of the linear functional models. Notably, since our framework allows the Riesz representer to take negative values, we are able to cover examples beyond the treatment effect or missing data analysis, such as weighted average derivatives, effects of policy interventions, data combination models, and bounds on consumer surplus. Our weighting scheme is composed of two steps. In the first step, we compute estimates of the Riesz representer for each observation using a projection argument. These estimates, called the projection weights, may take negative values, so they can be applied for general linear models. Then, in the second step, we *re-weigh* the moment conditions so that the resulting objective function admits an asymptotic chi-square calibration. In particular, in the second step, we capture the nonparametric likelihood increment in going from the baseline likelihood based on the projection weights to the one obtained by adding the identifying moment conditions for the parameters of interest. Since our likelihood ratio statistic is asymptotically pivotal, the resulting confidence set circumvents estimation of the asymptotic variance, which typically involves several nonparametric components (e.g., conditional means and variances and the Riesz representer). Also the confidence set is range preserving and transformation respecting, and its shape is determined by the data.

Our re-weighting method for constructing asymptotically pivotal statistics can be naturally extended for inference on treatment effect and data combination models. For treatment effects, we can employ empirical projection weights - which are similar to the empirical balancing weights of Hainmueller (2012) and Chan, Yam and Zhang (2016) - and reweigh the moment conditions in the second step to yield an asymptotically pivotal likelihood ratio statistic. Our approach is general enough to cover average and quantile treatment effects, among other quantities. For the data combination models, we consider the setup of Chen, Hong and Tarozzi (2008) and the Reisz representer is then related to projection weights that approximate the odds ratio of the propensity scores. Our simulation evidence and empirical example illustrate the usefulness of the proposed method.

We emphasize that treatment effect analysis is just one example in our framework. It can also be applied to settings such as effects of policy interventions and bounds on consumer surplus. More broadly, this paper contributes to the literature of estimation and inference on semiparametric models via the Riesz representer. For example, Newey and Robins (2018) first introduced a series-based estimator for the Riesz representer and considered estimation of finite dimensional parameters with fast decays for the remainders. Chernozhukov, Newey and Singh (2020) considered L_1-regularized Riesz representers for high dimensional scenarios. Hirshberg and Wager (2018) proposed to estimate the Riesz representer by applying the minimax approach. However, none of these papers focuses on likelihood-based inference on finite dimensional parameters by developing a asymptotically pivotal statistic.

This paper also contributes to the literature on empirical likelihood methods (see, Owen, 2001, for a survey). Qin and Zhang (2007) introduced the empirical likelihood approach for missing response problems with parametric propensity scores. Subsequently, Qin, Zhang and Leung (2009) proposed a unified empirical likelihood approach to missing data problems. We refer Qin (2017) for a comprehensive survey on the empirical likelihood methods, particularly in the context of missing and biased samples.

This paper is organized as follows. Section 2 introduces the basic setup and some examples. In Section 3, we develop the re-weighted nonparametric likelihood ratio statistic. Section 4 discusses extensions to inferences on treatment effects, data combination models, high-dimensional covariates by approximate balancing, and over-identified models. Sections 5 and 6 present simulation results and a real data example, respectively.

2. Setup and examples

Our dataset consists of a random sample of $(X,Y) \sim \mathbb{P}$ with support $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. Let $\mathbb{E}[\cdot]$ be expectation under $\mathbb{P}$ and $L_X^2 = \{f : \mathcal{X} \to \mathbb{R}, \mathbb{E}[f(X)^2] < \infty\}$. We consider inference on a finite dimensional vector of parameters $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ that can be identified by the moment conditions

$$\mathbb{E}[m(X, \gamma_{\theta_0}^{(l)}, \theta_0)] = 0 \qquad \text{for } l = 1, \ldots, d_\theta, \tag{1}$$

where for each $x \in \mathcal{X}$ and $\theta \in \Theta$, $\gamma \mapsto m(x,\gamma,\theta)$ is a known linear mapping such that $m(x,\gamma,\theta) - m(x,0,\theta)$ is linear in $\gamma \in L_X^2$, and $\gamma_{\theta_0}^{(l)}(\cdot) = \mathbb{E}[h^{(l)}(Y,X,\theta_0)|X = \cdot] \in L_X^2$ is the conditional expectation function for some known function $h^{(l)} : \mathcal{Y} \times \mathcal{X} \times \Theta \to \mathbb{R}$.

Based on this moment condition, we are interested in testing the parameter hypothesis

$$\mathrm{H}_0 : \theta_0 = c \quad \text{against} \quad \mathrm{H}_1 : \theta_0 \neq c,$$

for a given $c \in \mathbb{R}^{d_\theta}$. Assume that $\gamma \mapsto \mathbb{E}[m(X,\gamma,\theta_0)]$ is a continuous mapping on L_X^2. Then by the Riesz representation theorem, there exists a unique Riesz representer $\alpha_{\theta_0} : \mathcal{X} \to \mathbb{R}$ such that

$$\mathbb{E}[m(X,\gamma,\theta_0) - m(X,0,\theta_0)] = \mathbb{E}[\alpha_{\theta_0}(X)\gamma(X)] \qquad \text{for each } \gamma \in L_X^2. \tag{2}$$

By (2) and the law of iterated expectations, the moment condition (1) can be alternatively written as

$$\mathbb{E}[\alpha_{\theta_0}(X) h^{(l)}(Y,X,\theta_0) + m(X,0,\theta_0)] = 0 \qquad \text{for } l = 1, \ldots, d_\theta. \tag{3}$$

Note that (1) does not restrict how θ_0 enters to m or h. Therefore, γ_{θ_0} and/or α_{θ_0} may depend on θ_0 in possibly nonlinear manners.

This setup nests many well-known statistical inference problems. We give some examples below. Further examples are provided in Section 4.

Example 1 (Missing data model)**.** Consider a sequence of random variables $\{Y_{1i}, Z_i\}_{i=1}^N$, where Y_{1i} is observed only for a limited selection of individuals, and Z_i is a observable vector of covariates. In particular, we observe $Y_i = Y_{1i} D_i$ for all $i = 1, \ldots, N$, where D_i is the selection indicator (taking the value of one if Y_{1i} is observable, and zero otherwise). We wish to conduct statistical inference on the parameter θ_0, which is identified by the moment condition

$$\mathbb{E}[\psi(Y_1, Z, \theta_0)] = 0, \tag{4}$$

where ψ is possibly nonlinear in θ_0. Let $X = (D, Z)$, $h(Y,X,\theta_0) = \psi(Y,Z,\theta_0)$, and $\gamma_{\theta_0}(d,z) = \mathbb{E}[h(Y,Z,\theta_0)|D = d, Z = z]$. Further assume that ignorability (i.e., $Y_1 \perp D|X$) and overlap assumptions hold. Then it is easy to see that the identifying moment (4) can be rewritten as $\mathbb{E}[\gamma_{\theta_0}(1,Z)] = 0$. In this case, $m(X,\gamma_{\theta_0},\theta_0) = \gamma_{\theta_0}(1,Z)$, and the Riesz representer is written as

$\alpha_{\theta_0}(d,z) = \frac{d}{\mathbb{P}(D=1|Z=z)}$ so that

$$\mathbb{E}[\alpha_{\theta_0}(X)h(Y,X,\theta_0)] = 0. \tag{5}$$

Example 2 (Average effect after policy intervention). Let $\gamma_0(x) = \mathbb{E}[Y|X=x]$ be the conditional expectation and $\pi(\cdot)$ be a known policy function shifting the distribution of X to $\pi(X)$ after the policy intervention. The average policy effect is defined as $\mathbb{E}[\gamma_0(\pi(X))] - \mathbb{E}[\gamma_0(X)]$. The first term $\theta_0 = \mathbb{E}[\gamma_0(\pi(X))]$ can be analyzed by setting $m(x,\gamma_0,\theta_0) = \gamma_0(\pi(x)) - \theta_0$ and $h(y,x) = y$. Then the Riesz representer can be found by applying change of measure so that $\alpha_0(x) = \frac{dF_\pi}{dF}(x)$, where F_π is the cdf of $\pi(X)$, and $\mathbb{E}[\alpha_0(X)h(Y,X) - \theta_0] = 0$.

Example 3 (Bound on average equivalent variation and cosumer surplus). Let P_1 and P_2 denote the price of good 1 and a vector of prices of other goods in the consumption set, respectively. Also let V and Q be consumer's income and the quantity of good 1 bought by the consumer, respectively. We are interested in determining an upper bound θ_0 on the average equivalent variation for a price change from p_a to p_b of good 1, averaged over the other prices P_2 and income V. Let B denote a lower bound on the income effect for all individuals. Hausman and Newey (2017) showed that

$$\theta_0 = \mathbb{E}[\int l(p,V)\gamma_0(p,P_2,V)dp],$$

where $l(p,V) = w(V)\mathbb{I}\{p_a \leq p \leq p_b\}\exp(-B(p-p_a))$ with some known function w, and $\gamma_0(P_1,P_2,V) = \mathbb{E}[Q|P_1,P_2,V]$. It is easy to see that this setting can be subsumed into our framework by letting $X = (P_1,P_2,V)$, and $m(X,\gamma_0,\theta_0) = \int l(p,V)\gamma_0(p,P_2,V)dp - \theta_0$. The bound on the consumer surplus can be obtained similarly: we simply get rid of the conditioning variable P_2 in the above setup.

Example 4 (Average derivative). Consider the average partial derivative of the regression function $\gamma_0(x) = \mathbb{E}[Y|X=x]$:

$$\theta_0 = \mathbb{E}\left[w(X)\frac{\partial\gamma_0(X)}{\partial x}\right],$$

for some known weight function $w(\cdot)$. In this case, γ_{θ_0} and $h(Y,X,\theta_0)$ do not involve θ_0, and we set as $m(x,\gamma_0,\theta_0) = w(x)\frac{d\gamma_0(x)}{dx} - \theta_0$ and $h(y,x) = y$. Assuming $w(x) = 0$ at boundaries of $\mathcal{X}$, the Riesz representer is obtained as $\alpha_{\theta_0}(x) = -\frac{1}{f_X(x)}\frac{d[w(x)f_X(x)]}{dx}$, where f_X is the density of X.

3. Reweighted nonparametric likelihood inference

In this section we present our inference method for θ_0 defined by the linear functional model in (1), or alternatively (3). Let $\{q_j(\cdot)\}_{j=1}^\infty$ denote basis functions for the space L^2_X. Then the condition for the Riesz representer α_{θ_0} in (2) is equivalent to the infinite set of moment conditions

$$\mathbb{E}[m(X,q_j,\theta_0) - m(X,0,\theta_0) - \alpha_{\theta_0}(X)q_j(X)] = 0, \qquad \text{for all } j = 1,2,\ldots. \tag{6}$$

The equivalence between (2) and (6) exploits the fact that $\gamma \mapsto m(\cdot,\gamma,\cdot)$ is an linear functional, which ensures that for each $\gamma \in L^2_X$, there exists some sequence $\{\xi_j\}_{j=1}^\infty$ satisfying $\sum_{i=1}^\infty \xi_j^2 < \infty$ and $\mathbb{E}[m(X,\gamma,\theta_0) - m(X,0,\theta_0)] = \sum_{j=1}^\infty \xi_j\mathbb{E}[m(X,q_j,\theta_0) - \mathbb{E}[m(X,0,\theta_0)]$.

To approximate the Riesz representer α_{θ_0}, we employ a finite but growing number of the moment conditions from (6). Let $Q_K(\cdot) = (q_1(\cdot), \ldots, q_K(\cdot))'$ denote a vector of basis functions of dimension k, and $M_K(X_i, \theta_0)$ denote a K-dimensional vector whose k-th element is $m(X_i, q_k, \theta_0) - m(X_i, 0, \theta_0)$ for $k = 1, \ldots, K$. As in Newey and Robins (2018), we approximate $\{\alpha_{\theta_0}(X_i)\}_{i=1}^N$ by the projection weights $\{\hat{\alpha}_i\}_{i=1}^N$, which are obtained as a solution of

$$\min_{\alpha_1,\ldots,\alpha_N} \frac{1}{2}\sum_{i=1}^N \alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^N \{M_K(X_i,\theta_0) - \alpha_i Q_K(X_i)\} = 0,$$

i.e.,

$$\hat{\alpha}_i = Q_K(X_i)' \left[\frac{1}{N}\sum_{i=1}^N Q_K(X_i)Q_K(X_i)'\right]^{-1} \frac{1}{N}\sum_{i=1}^N M_K(X_i,\theta_0), \tag{7}$$

for $i = 1, \ldots, N$. Following Qiu (2020), one can interpret $\mathbb{E}[\alpha_{\theta_0}(X)^2]$ as the residual variance for estimation of θ_0, and $\sup_{\gamma \in L_X^2} \mathbb{E}[m(X,\gamma,\theta_0) - \alpha_{\theta_0}(X)\gamma]$ as the residual bias. Thus, (7) has an attractive interpretation of minimizing the empirical variance subject to a zero empirical bias constraint within the sieve space spanned by $Q_K(\cdot)$.

The construction of (7) is similar to, but distinct from, the empirical balancing weights that have been proposed in the literature on missing data, see e.g., Zubizarreta (2015), Chan, Yam and Zhang (2016), and Qiu and Otsu (2018). Recall that in Example 1 on missing data models, the Riesz representer is expressed as $\alpha_{\theta_0}(D_i, Z_i) = D_i/\mathbb{P}(D_i = 1|Z_i)$. The empirical balancing weights estimate the 'tilting function', $1/\mathbb{P}(D_i = 1|Z_i)$, instead of directly estimating $\alpha_{\theta_0}(D_i, Z_i)$. Although the estimates of $\alpha_{\theta_0}(D_i, Z_i)$ are only computed for observations without missing outcomes (but note that $\alpha_{\theta_0}(D_i, Z_i) = 0$ when $D_i = 0$ anyway), given the empirical balancing weights (say, $\hat{w}_i$), it is straightforward to obtain the estimates $\hat{\alpha}_i$ of α_i as $\hat{\alpha}_i = \hat{w}_i$ when $D_i = 1$, and $\hat{\alpha}_i = 0$ otherwise. A drawback of the empirical balancing weights, however, is that they are not applicable more generally, e.g., to average derivative estimation, where the Riesz representer may take negative values.

Based on the projection weights $\{\hat{\alpha}_i\}_{i=1}^N$ in (7), we now construct our nonparametric likelihood function. The basic idea is to 're-weigh' both the moment functions (2) and (3) after incorporating the the projection weights in (7). The 're-weighted' likelihood ratio then captures the likelihood increments associated with (3). Formally, let

$$\phi_\varsigma(\omega) = \frac{2}{\varsigma(\varsigma+1)}\{(N\omega)^\varsigma - 1\},$$

denote the Cressie and Read (1984) power divergence family if $\varsigma \neq -1, 0$, otherwise $\phi_{-1}(\omega) = -2\log(n\omega)$ and $\phi_0(\omega) = 2n\omega\log(n\omega)$. The cases of $\varsigma = -1$ and $\varsigma = 0$ are often called empirical likelihood and exponential tilting, respectively. Other popular choices for ς include the Neyman's modified χ^2 ($\varsigma = 1$), Hellinger or Freeman-Tukey ($\varsigma = -1/2$), and Pearson's χ^2 ($\varsigma = -2$). Based on this divergence, we consider the following re-weighting scheme:

$$\ell(\theta_0) = \min_{\omega_1,\ldots,\omega_N} \sum_{i=1}^N \phi_\varsigma(\omega_i),$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \omega_i \{M_K(X_i,\theta_0) - \hat{\alpha}_i Q_K(X_i)\} = 0, \qquad \sum_{i=1}^{N} \omega_i = 1, \qquad \omega_i \geq 0,$$

$$\sum_{i=1}^{N} \omega_i \{\hat{\alpha}_i h^{(l)}(Y_i, X_i, \theta_0) + m(X_i, 0, \theta_0)\} = 0 \quad \text{for } l = 1, \ldots, d_\theta, \tag{8}$$

where $\{\hat{\alpha}_i\}_{i=1}^N$ are the projection weights obtained in (7). Note that without the last condition in (8) (corresponding to the primary moment condition), the above maximization problem is solved by uniform weights $\omega_i = 1/N$ for all $i = 1, \ldots, N$ (because of (7)). Therefore, the above minimum $\ell(\theta_0)$ indeed corresponds to the likelihood increment by adding the last condition in (8). Let

$$g_i^K = \begin{pmatrix} M_K(X_i,\theta_0) - \hat{\alpha}_i Q_K(X_i) \\ \hat{\alpha}_i h^{(1)}(Y_i, X_i, \theta_0) + m(X_i, 0, \theta_0) \\ \vdots \\ \hat{\alpha}_i h^{(d_\theta)}(Y_i, X_i, \theta_0) + m(X_i, 0, \theta_0) \end{pmatrix}.$$

By applying the Lagrange multiplier method, the dual representation of the re-weighted non-parametric likelihood ratio statistic is

$$\ell(\theta_0) = \max_{\lambda} 2 \sum_{i=1}^{N} \{\rho_\varsigma(\lambda' g_i^K) - \rho_\varsigma(0)\}, \tag{9}$$

where

$$\rho_\varsigma(v) = -\frac{1}{\varsigma+1}(1+\varsigma v)^{(\varsigma+1)/\varsigma},$$

if $\varsigma \neq -1, 0$, otherwise $\rho_{-1}(v) = \log(1-v)$ and $\rho_0(v) = -e^v$. In practice, we employ this dual form to implement our inferential procedure.

To derive the limiting distribution of $\ell(\theta_0)$, we impose the following assumptions. Let $\zeta_K = \sup_{x\in\mathcal{X}} |Q_K(x)|$, $\varepsilon_K(x) = \alpha_{\theta_0}(x)Q_K(x) - M_K(x,\theta_0)$, $\zeta_{\varepsilon,K} = \sup_{x\in\mathcal{X}} |\varepsilon_K(x)|$, and $\tilde{m}(x,\cdot,\theta_0) = m(x,\cdot,\theta_0) - m(x,0,\theta_0)$.

Assumption.

(i): *$\{X_i, Y_i\}_{i=1}^N$ is iid. For each $x \in \mathcal{X}$, $\gamma \mapsto m(x,\gamma,\theta_0)$ is an linear mapping, and $\gamma \mapsto \mathbb{E}[m(x,\gamma,\theta_0)]$ is a continuous mapping from L_X^2 to $\mathbb{R}$.*

(ii): *All eigenvalues of $\mathbb{E}[Q_K(X)Q_K(X)']$ and $\mathbb{E}[\varepsilon_K(X)\varepsilon_K(X)']$ are bounded from above and away from zero for each $K \in \mathbb{N}$, $\frac{\zeta_K^2 \log K}{N} \to 0$, and $\frac{\zeta_{\varepsilon,K}^2 \log K}{N} \to 0$.*

(iii): *$\max_{1\leq i\leq N} |\hat{\alpha}_i - \alpha_{\theta_0}(X_i)| = O_p(\delta_{\alpha,N})$ for some $\delta_{\alpha,N} \to 0$. For each $l = 1, \ldots, d_\theta$ and $K \in \mathbb{N}$, there exists some $\beta_K^{(l)} \in \mathbb{R}^K$ such that $\sup_{x\in\mathcal{X}} |\gamma_{\theta_0}^{(l)}(x) - \beta_K^{(l)\prime} Q_K(x)| \lesssim \eta_K$ and $\mathbb{E}[\tilde{m}(X, \gamma_{\theta_0}^{(l)} - \beta_K^{(l)\prime} Q_K, \theta_0)^2] \lesssim \tilde{\eta}_K$ for some $\tilde{\eta}_K \to 0$.*

(iv): *For each $l = 1, \ldots, d_\theta$, it holds $\sup_{x\in\mathcal{X}} \mathbb{E}[\{h^{(l)}(X,Y,\theta_0) - \gamma_{\theta_0}^{(l)}(X)\}^2 | X = x] \lesssim 1$, $\mathbb{E}[\{\alpha_0(X)\gamma_{\theta_0}^{(l)}(X) - \tilde{m}(X, \gamma_{\theta_0}^{(l)}, \theta_0)\}^2] \lesssim 1$, $\mathbb{E}[m(X, \gamma_{\theta_0}^{(l)}, \theta_0)^2] \lesssim 1$, $\sup_{x\in\mathcal{X}} |\alpha_{\theta_0}(x)| \lesssim 1$, and there exists some $\kappa > 2$ such that $\mathbb{E}[|h^{(l)}(X,Y,\theta_0)|^\kappa] \lesssim 1$ and $N^{1/\kappa}\delta_{\alpha,N} \to 0$.*

(v): *For each* $j, k = 1, \ldots, K$,

$$\begin{aligned} \mathbb{E}[\tilde{m}(X, q_j(\cdot)\tilde{m}(\cdot, q_k, \theta_0), \theta_0)] &= \mathbb{E}[\tilde{m}(X, q_j, \theta_0)\tilde{m}(X, q_k, \theta_0)], \\ \mathbb{E}[\tilde{m}(X, m(\cdot, 0, \theta_0)q_j(\cdot), \theta_0)] &= \mathbb{E}[m(X, 0, \theta_0)\tilde{m}(X, q_j, \theta_0)]. \end{aligned} \tag{10}$$

Assumption (i) is reasonable for all the examples listed in this paper. An extension to dependent data is left for future research. Assumption (ii) is on the vector of basis functions $Q_K(\cdot)$ and the approximation error $\varepsilon_K(\cdot)$. The first condition in Assumption (iii) imposes a basic approximation quality of $\hat{\alpha}_i$ in terms of sup norm. This could be verified by more primitive conditions; see Newey and Robins (2018). The second condition in Assumption (iii) is on the approximation error for the conditional mean function $\gamma^{(l)}_{\theta_0}(\cdot)$ by the vector of basis functions $Q_K(\cdot)$. Assumption (iv) is a set of regularity conditions. Notably, we require the Riesz representer α_{θ_0} to be bounded, and existence of higher moments for h.

Assumption (v) is a key requirement that needs to be checked for each application. It can be thought of as placing some constraints on the form of $m(\cdot)$. All our examples satisfy this assumption except for the average derivative example. The assumption is trivially satisfied if the moment function is multiplicative in γ (in addition to being linear). For instance, in the missing data example, $m(X, 0, \theta_0) = 0$ and $\tilde{m}(X, \gamma_{\theta_0}, \theta_0) = \gamma_{\theta_0}$, and it is easy to see that (10) is satisfied. A similar reasoning also applies to the average effect after policy intervention. For the average equivalent variation and consumer surplus example, $m(X, \gamma_0, \theta_0) = \int l(p, V)\gamma_0(p, P_2, V)dp - \theta_0$, and so $m(x, 0, \theta_0) = -\theta_0$ and $\tilde{m}(x, \gamma_0, \theta_0) = \int l(p, V)\gamma_0(p, P_2, V)dp$. The second equation of (10) is satisfied trivially. As for the first equation:

$$\begin{aligned} \mathbb{E}[\tilde{m}(X, q_j(\cdot)\tilde{m}(\cdot, q_k, \theta_0), \theta_0)] &= \mathbb{E}\left[\int l(p, V)q_j(p, P_2, V) \underbrace{\int l(p, V)q_k(p, P_2, V)dp}\, dp\right] \\ &= \mathbb{E}\left[\underbrace{\int l(p, V)q_k(p, P_2, V)dp} \int l(p, V)q_j(p, P_2, V)dp\right] \\ &= \mathbb{E}[\tilde{m}(X, q_j, \theta_0)\tilde{m}(X, q_k, \theta_0)], \end{aligned}$$

where we can take $\underbrace{\cdot}$ out since it is not a function of p.

For the average derivative example, equation (10) is not generally satisfied. One exception is the setting where $w(x) = 1$ and $d^2 q_k(x)/dx^2 = 0$ for all k, i.e., the basis functions all have zero second derivatives. This implies that the assumption is valid if one employs basis functions of indicator or linear form, such as linear regression splines or Strömberg wavelets of order 0 (assuming compact support for X and that its density $f_X(\cdot)$ is bounded).

Based on these assumptions, the asymptotic distribution of the likelihood ratio statistic $\ell(\theta_0)$ is obtained as follows.

Theorem. *Suppose Assumption (i)-(iv) hold true. In addition,* $\delta_{\alpha,N}\zeta_K\zeta_{\varepsilon,K} \to 0$, $\sqrt{K}\eta_K \to 0$, $\sqrt{N}\delta_{\alpha,N}\eta_K \to 0$. *Then*

$$\ell(\theta_0) \xrightarrow{d} \chi^2_{d_\theta}, \quad \text{as } N \to \infty.$$

This theorem says that our likelihood ratio statistic is asymptotically pivotal and converges to the chi-squared distribution under the null hypothesis. Based on this result, the $100(1-\alpha)\%$

asymptotic confidence set for θ_0 can be given by $\{\theta : \ell(\theta) \leq \chi^2_{d_\theta,\alpha}\}$, where $\chi^2_{d_\theta,\alpha}$ is the $(1-\alpha)$-th quantile of the $\chi^2_{d_\theta}$ distribution. Furthermore, it is straightforward to extend this theorem for testing the null $\mathrm{H}_0 : r(\theta) = r_0$ for a possibly nonlinear function $r : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_r}$ with $d_r \leq d_\theta$. In this case, the likelihood ratio statistic is obtained by $\min_{\theta : r(\theta) = r_0} \ell(\theta)$, which can be shown to converge to the $\chi^2_{d_r}$ distribution.

If Assumption (v) is violated, the re-weighted statistic $\ell(\theta_0)$ loses its asymptotic pivotalness and converges to a weighted χ^2 distribution, where the weights involve unknown components. In particular, suppose the condition (5) does not hold but $\delta_{\alpha,N}\zeta_K\zeta^3_{\varepsilon,K} \to 0$ and $\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(m(\gamma_0)|\varepsilon_K)\mathcal{P}_i(m(\gamma_0)|\varepsilon_K) \xrightarrow{p} V^*$ for some $d_\theta \times d_\theta$ matrix V^*, where $\mathcal{P}_i(m(\gamma_0)|\varepsilon_K)$ is the empirical projection of $m(\gamma_0) = (m(X_1,\gamma_0,\theta_0),\ldots,m(X_N,\gamma_0,\theta_0))'$ with $m(X_i,\gamma_0,\theta_0) = (m(X_i,\gamma_0^{(1)},\theta_0),\ldots,m(X_i,\gamma_0^{(d_\theta)},\theta_0))'$ on $\varepsilon_K = (\varepsilon_K(X_1),\ldots,\varepsilon_K(X_N))'$. Then by inspection the proof of this theorem, we obtain

$$\ell(\theta_0) \xrightarrow{d} U'(V - V^*)U, \quad \text{as } N \to \infty,$$

where $U \sim N(0,V)$ and V is the variance matrix of the d_θ-dimensional random vector whose l-th element is $m(X,\gamma_0^{(l)},\theta_0) + \alpha_0(X)\{h^{(l)}(X,Y,\theta_0) - \gamma^{(l)}_{\theta_0}(X)\}$. One may conduct inference based on this limiting distribution by estimating the variance components V and V^*, or by bootstrapping (see, Section 2.3 of Hjort, McKeague and Van Keilegom, 2009). However, given the asymptotic pivotalness in our theorem, we recommend employing basis functions $Q_K(\cdot)$ that satisfy the condition in (3).

The proof of this theorem indicates that under our assumptions, our likelihood ratio statistic has the same local power function as the Wald or t-test based on the globally semiparametric efficient estimator. However, in contrast to the Wald test, we circumvent estimation of the asymptotic variance which can be quite involved.

4. Extensions

4.1. **Treatment effects.** It is straightforward to extend our likelihood ratio construction to conduct inference on various measures of treatment effects under unconfoundedness. Let Y_1 and Y_0 be potential outcomes associated with a binary treatment variable $D \in \{0,1\}$. The observed outcome is $Y = DY_1 + (1-D)Y_0$. Let Z be a vector of covariates. Suppose we want to conduct inference on the parameter θ_0 identified by the moment condition

$$\mathbb{E}[\psi_1(Y_1,Z,\theta_0) - \psi_0(Y_0,Z,\theta_0)] = 0, \tag{11}$$

where ψ_1 and ψ_0 have the same dimension as θ_0. This setup accommodates many popular inferential problems as special cases. For example, if we set $\psi_1(Y_1,X,\theta_0) = Y_1 - \theta_0$ and $\psi_0(Y_0,X,\theta_0) = Y_0$, then θ_0 is the average treatment effect.

To see how this fits into our setup, let us denote $X = (D,Z)$, $\gamma^{(1)}_{\theta_0}(D,Z) = \mathbb{E}[\psi_1(Y,Z,\theta_0)|D,Z]$ and $\gamma^{(0)}_{\theta_0}(D,Z) = \mathbb{E}[\psi_0(Y,Z,\theta_0)|D,Z]$. Under conditional independence assumption, (11) can be rewritten as

$$\mathbb{E}[\gamma^{(1)}_{\theta_0}(1,Z) - \gamma^{(0)}_{\theta_0}(0,Z)] = 0. \tag{12}$$

Further, under the overlap condition, (11) gives rise to the two Riesz representers $\alpha_{\theta_0}^{(1)}(x) = \frac{d}{\mathbb{P}(D=1|Z=z)}$ and $\alpha_{\theta_0}^{(0)}(x) = \frac{1-d}{1-\mathbb{P}(D=1|Z=z)}$ for $x = (d, z)$ so that

$$\mathbb{E}[\alpha_{\theta_0}^{(1)}(X)\gamma(X)] = \mathbb{E}[\gamma(X)], \quad \text{for each } \gamma \in L_X^2,$$
$$\mathbb{E}[\alpha_{\theta_0}^{(0)}(X)\gamma(X)] = \mathbb{E}[\gamma(X)], \quad \text{for each } \gamma \in L_X^2.$$

Again, we consider the testing problem $\mathrm{H}_0 : \theta_0 = c$ against $\mathrm{H}_1 : \theta_0 = c$. Note that the identifying moment for θ_0 is

$$\mathbb{E}[\alpha_{\theta_0}^{(1)}(X)\psi_1(Y, Z, \theta_0) - \alpha_{\theta_0}^{(0)}(X)\psi_0(Y, Z, \theta_0)] = 0, \tag{13}$$

Applying our methodology, in the first step we calibrate two sets of projection weights $\{\hat{\alpha}_i^{(1)}\}_{i=1}^N$ and $\{\hat{\alpha}_i^{(0)}\}_{i=1}^N$ according to

$$\min_{\alpha_1^{(1)},\ldots,\alpha_N^{(1)}} \frac{1}{2}\sum_{i=1}^N D\alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^N \{Q_K(X_i) - \alpha_i^{(1)} D_i Q_K(X_i)\} = 0;$$
$$\min_{\alpha_1^{(0)},\ldots,\alpha_N^{(0)}} \frac{1}{2}\sum_{i=1}^N (1-D)\alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^N \{Q_K(X_i) - \alpha_i^{(0)} (1-D_i) Q_K(X_i)\} = 0;$$

Based on the approximated Riesz representers $\{D\hat{\alpha}_i^{(1)}\}_{i=1}^N$ and $\{(1-D_i)\hat{\alpha}_i^{(0)}\}_{i=1}^N$, our reweighted likelihood ratio statistic can be constructed as

$$\bar{\ell}(\theta_0) = \min_{\omega_1,\ldots,\omega_N} \sum_{i=1}^N \phi_\varsigma(\omega_i),$$

$$\text{s.t.} \quad \sum_{i=1}^N \omega_i Q_K(X_i)(\hat{\alpha}_i^{(1)} D_i - 1) = 0, \quad \sum_{i=1}^N \omega_i Q_K(X_i)(\hat{\alpha}_i^{(0)}(1-D_i) - 1) = 0, \quad \sum_{i=1}^N \omega_i = 1, \quad \omega_i \geq 0,$$
$$\sum_{i=1}^N \omega_i \{\hat{\alpha}_i^{(1)} D_i \psi_1(Y_i, X_i, \theta_0) - \hat{\alpha}_i^{(0)}(1-D_i)\psi_0(Y_i, X_i, \theta_0)\} = 0.$$

The dual form of $\bar{\ell}(\theta_0)$ is obtained in the same manner as $\ell(\theta_0)$. Also, under analogous conditions to the ones in the Theorem, it can be shown that $-2\bar{\ell}(\theta_0) \xrightarrow{d} \chi^2_{d_\theta}$ under H_0, where d_θ is the dimension of θ_0. Again, our likelihood ratio statistic is asymptotically pivotal, and is free from variance estimation. If we are interested in some p_1-dimensional function $r(\beta)$ (e.g., quantile treatment effects), the likelihood ratio statistic for $\mathrm{H}_0 : r(\beta) = r_0$ can be modified as $\min_{\beta: r(\beta)=r_0} \bar{\ell}(\beta) \xrightarrow{d} \chi^2_{p_1}$.

4.2. **Data combination models.** Data combination models are another important class of missing data models. Let $W = (Y_1, Y_0, Z')'$ denote a vector of random variables from a study population. We are interested in conducting inference for the d_θ-dimensional vector of parameters, θ_0, which is just-identified by the moment condition

$$\mathbb{E}_s[\psi(W, \theta_0)] = 0,$$

where $\mathbb{E}_s[\cdot]$ denotes the expectation under the study sample. However we do not observe the entire vector W. Rather, we only observe N_s measurements of $(Y_1, Z')'$ from the study sample, but we have access to N_a measurements of $(Y_0, Z')'$ drawn from an auxiliary sample. Thus the variables Z are common to the both samples.

We shall assume that the conditional distribution of Y_0 given Z is the same in the both samples (however the marginal distributions of Z may differ). Also, we assume that the support of Z in the auxiliary sample is at least as large as the study sample. Under these conditions, Chen, Hong and Tarozzi (2008) showed that the parameter vector θ_0 is identified as long as $\psi(\cdot)$ is separable in Y_1 and Y_0 in the sense that

$$\psi(Y_1, Y_0, Z, \theta_0) = \psi_s(Y_1, Z, \theta_0) - \psi_a(Y_0, Z, \theta_0),$$

for some $\psi_s(\cdot)$ and $\psi_a(\cdot)$. This framework covers many important statistical problems including estimation of the average treatment effect on the treated (ATT), two-sample instrumental variables (Angrist and Krueger, 1992), counterfactual distributions (Dinardo, Fortin and Lemieux, 1996), semiparametric differences-in-differences (Abadie, 2005), and models with mismeasured regressors in the presence of validation samples (Carroll and Wand, 1991).

Following Graham, Pinto and Egel (2016), we employ a multinomial sampling framework by assuming that a unit is drawn at random from the distribution of the study sample with probability π, and from that of the auxiliary sample with probability $1 - \pi$. Let D denote a binary random variable that takes value 1 when the observation is in the study sample and 0 when it is in the auxiliary sample. Under this framework we can treat the 'merged' realization $(D_i, Z_i, D_i Y_{1i}, (1 - D_i)Y_{0i})$ as a random draw from a synthetic 'merged' population (Graham, Pinto and Egel, 2016). Let $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability and expectation, respectively, in this merged population. Finally, let $Y = DY_1 + (1 - D)Y_0$ denote the observed 'outcome' variable.

This set of models also fits into our current setup. To observe this, let us denote $X = (D, Z)$, $\gamma_\theta^{(1)}(D, Z) = \mathbb{E}[\psi_s(Y, Z, \theta_0)|D, Z]$ and $\gamma_\theta^{(2)}(D, Z) = \mathbb{E}[\psi_u(Y, Z, \theta_0)|D, Z]$. Then the identifying moment condition can be rewritten as

$$\int \{\gamma_\theta^{(1)}(z, 1) - \gamma_\theta^{(2)}(z, 0)\} dF_s(z) = \int \{\gamma_\theta^{(1)}(z, 1) - \gamma_\theta^{(2)}(z, 0)\} \frac{dF_s(z)}{dF(z)} dF(z) = 0.$$

The support condition above assures existence of some $\kappa > 0$ such that $\kappa \leq \mathbb{P}(D = 1|Z = z) \leq 1$ for all $z \in \mathbb{R}^{dz}$. Importantly, we do not place any functional form assumptions on the propensity score, apart from some smoothness assumptions. As before, we consider the testing problem $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$, which is equivalent to test the identifying moment

$$\mathbb{E}\left[D\psi_s(W, X, \theta_0) - (1 - D)\frac{\mathbb{P}(D = 1|Z)}{1 - \mathbb{P}(D = 1|Z)}\psi_a(W, X, \theta_0)\right] = 0,$$

subject to the auxiliary moment conditions identifying the Riesz representer as $\alpha_{\theta_0}(x) = (1 - d)\frac{\mathbb{P}(D=1|Z=z)}{1-\mathbb{P}(D=1|Z=z)}$ since

$$\mathbb{E}[\alpha_{\theta_0}(X)\gamma(X)] = \mathbb{E}[D\gamma(X)], \quad \text{for each } \gamma \in L_X^2,$$

(see, Graham, 2011). Let $N = N_a + N_s$. We shall order the observations such that the first N_a terms correspond to the auxiliary sample (i.e., $D_i = 0$ for $i = 1, \ldots, N_a$ and 1 for $i = N_a + 1, \ldots, N$). The projection weights $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{N_a})$ for data combination models are obtained as the solution of

$$\min_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^{N} \frac{1}{2}(1-D)\alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^{N} D_i Q_K(X_i) = \frac{1}{N}\sum_{i=1}^{N} \tilde{\alpha}_i (1-D_i) Q_K(X_i).$$

In this case, our likelihood ratio statistic is obtained as

$$\tilde{\ell}(\theta_0) = \max_{\omega_1,\ldots,\omega_N} \sum_{i=1}^{N} \phi_\varsigma(\omega_i),$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{N} \omega_i \{D_i - \tilde{\alpha}_i(1-D_i)\} Q_K(X_i) = 0, \quad \sum_{i=1}^{N} \omega_i = 1, \quad \omega_i \geq 0 \\
& \sum_{i=1}^{N} \omega_i \{D_i \psi_s(W_i, X_i, \theta_0) - (1-D_i)\tilde{\alpha}_i \psi_a(W_i, X_i, \theta_0)\} = 0.
\end{aligned}$$

The dual form of $\tilde{\ell}(\theta_0)$ is obtained in the same manner as $\ell(\theta_0)$. Also under analogous conditions to the ones in Theorem, it can be shown that $\tilde{\ell}(\theta_0) \xrightarrow{d} \chi^2_{d_\theta}$ under H_0, where d_θ is the dimension of θ_0.

4.3. **Over-identified models.** Thus far we have considered inference under just-identification. In some applications however, the parameters β could be over-identified (e.g., moment conditions with side information, and two-sample instrumental variable models with more instruments than regressors). While our testing procedure still controls size in such contexts, it is no longer first-order efficient. In this section we show how it can be modified to recover efficiency.

Consider the missing data setup in Section 2. Suppose now that the dimension p_1 of the moment function $\psi(\cdot)$ is greater than p, the dimension of β. Then we can construct a likelihood ratio test by considering the discrepancy in the log-likelihoods evaluated at the estimated and hypothesized values of β. In particular, based on the likelihood ratio statistic in (8), the likelihood ratio test statistic for testing $\mathrm{H}_0 : \beta = \beta_0$ against $\mathrm{H}_1 : \beta \neq \beta_0$ is given by

$$\ell^R(\beta_0) = \ell(\beta_0) - \min_{\beta} \ell(\beta).$$

Under analogous conditions to the Theorem in Section 3, it can be shown that $\ell^R(\beta_0) \xrightarrow{d} \chi^2_p$ under H_0. Note that the degree of freedom of the limiting distribution is p, the dimension of β. On the other hand, the statistic $\ell(\beta_0)$ converges to the chi-square distribution with degree of freedom p_1, the dimension of ψ.

5. Simulation

In this section, we study the finite sample performance of the proposed likelihood ratio test. We consider three different data generating processes (DGPs). The first DGP (DGP1) is taken from Abadie and Imbens (2016, Supplementary material), adapted for the case of missing data. We generate a two dimensional vector (Z_1, Z_2) of covariates by drawing both variables from a uniform $[-1/2, 1/2]$ distribution independently of each other. The 'true' outcome variable is generated as $Y_1 = 5 + 2Z_1 + 4Z_2 + U$, where U is a standard normal random variable. The

TABLE 1. Rejection frequencies under the null for inference on β_a

		$N = 100$		$N = 200$		$N = 500$		$N = 5000$	
		$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$
DGP1	LR	0.067	0.077	0.056	0.062	0.057	0.063	0.048	0.054
	Wald	0.064	0.074	0.057	0.062	0.058	0.063	0.047	0.054
DGP2	LR	0.092	0.119	0.074	0.082	0.054	0.066	0.059	0.056
	Wald	0.098	0.118	0.076	0.085	0.056	0.066	0.058	0.056
DGP3	LR	0.105	0.147	0.083	0.102	0.079	0.065	0.126	0.057
	Wald	0.143	0.226	0.102	0.122	0.091	0.071	0.129	0.058

propensity score is given by the logistic function

$$\mathbb{P}(D = 1|Z) = \frac{\exp(Z_1 + tZ_2)}{1 + \exp(Z_1 + tZ_2)}. \tag{14}$$

The treatment D is generated by this probability, and the observed outcome variable is generated by $Y = DY_1$. For DGP1, we choose $t = 2$. The second DGP (DGP2) differs from DGP1 only in the choice of t: in particular, we set $t = 4$. The effect of increasing t is to reduce the amount of overlap in the propensity score. For the third DGP (DGP3), we generate a two-dimensional vector (Z_1, Z_2) of covariates by drawing both variables (independently of each other) as $Z_a \sim 2V - 1$ for $a = 1, 2$, with $V \sim \text{beta}(2, 4)$. The unobserved outcome variable is generated as $Y_1 = 5 + Z_1^2 + Z_2^2 + U$. The remainder of DGP3 follows the same construction as DGP1, i.e., we obtain the propensity scores by setting $t = 2$ in (14). Compared to DGP1, the distribution of Y is more asymmetric due to the use of an asymmetric beta distribution for the covariates. Additionally, the underlying model for Y_1 is more non-linear.

We first consider inference on the average outcome $\beta_a = \mathbb{E}[Y_1]$. Table 1 reports the performance of the likelihood ratio (LR) procedure for inference on β_a under the null, along with inference based on the Wald statistic using the variance estimate proposed by Chan, Yam and Zhang (2016). The nominal significance level is 0.05. For all DGPs, we report results with $K = 3$, corresponding to $q^K(X) = (1, Z_1, Z_2)'$, and $K = 5$, corresponding to $q^K(X) = (1, Z_1, Z_2, Z_1^2, Z_2^2)$. All the simulation results are based on 2,500 Monte Carlo repetitions.

From Table 1, we can observe that when the DGP is linear and the overlap is good (e.g., DGP1 and DGP2), both the LR and Wald procedures behave very similarly. However, when there is non-linearity in the covariates and the underlying distribution is asymmetric, the LR procedure provides more accurate inference, as seen in the simulation results for DGP3.

Next we consider inference on the median outcome $\beta_m = \text{median}(Y_1)$. Here the 'identifying' moment condition for β_m is given by $\mathbb{E}[\mathbb{I}\{Y < \beta_m\} - 0.5] = 0$. Table 2 reports the performance of the LR procedure for inference on β_m under the null for all the DGPs. Again the nominal significance level is 0.05. The LR procedure provides excellent size control for all DGPs, with the proviso that one employs a proper choice of K for DGP3. Note also that the Wald statistic is difficult to obtain here due to the complicated nature of the variance estimate for quantile

TABLE 2. Rejection frequencies under the null for inference on β_m

	$N = 100$		$N = 200$		$N = 500$		$N = 5000$	
	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$
DGP1	0.056	0.067	0.058	0.052	0.048	0.047	0.056	0.053
DGP2	0.072	0.098	0.054	0.076	0.058	0.055	0.091	0.063
DGP3	0.128	0.206	0.097	0.101	0.069	0.080	0.088	0.049

estimators; indeed we are not aware of any variance estimate that has been proposed for this context.

6. Real data example

We illustrate our inferential procedure by applying it on data taken from the influential study of Card and Krueger (1994). These authors were interested in studying the effect of the raise, in 1993, of New Jersey's state minimum wage on employment. To this end, they collected data on employment in fast food restaurants in New Jersey and neighboring Pennsylvania, following the minimum wage hike. The restaurants in Pennsylvania, which did not witness a change in the minimum wage, form the control group. While the original study was based on a differences-in-differences design, later authors including Rosenbaum (2002) and Imbens and Rubin (2015) re-analyzed the data as if it arose from an unconfoundedness assumption, i.e., conditional on covariates, the probability of being treated (i.e., being from New Jersey as opposed to Pennsylvania) does not depend on the potential outcomes. Subsequently, our results in this section are based on the latter assumption.

The data consist of 273 restaurants from New Jersey (treated units), and 67 from Pennsylvania (control units). The covariate data consist of the following pre-treatment variables: number of employed in each restaurant prior to minimum wage hike (empft), starting wages (wage_st), average duration for the first raise (inctime), and indicators for the identity of the chain: (burger king, kfc, roys, wendys). The outcome (Y) is the number of employed in each restaurant after the increase in minimum wage (part time employees are weighted by 0.5). Our parameter of interest, β_0, is the average treatment effect on employment levels due to the minimum wage hike.

To provide inference on β_0, we consider two empirical balancing schemes: one where we only balance a single covariate, empft, i.e., $q^K(X) = (1, \mathsf{empft})$, corresponding to $K = 2$; and the other where we balance all the covariates Z, i.e., $q^K(X) = Z$, corresponding to $K = 7$. The first scheme in particular is based on the analysis of Imbens and Rubin (2015) who found that empft was the only variable selected by their iterative balance checking algorithm for inclusion in the propensity score. Table 3 presents 90 and 95% confidence regions for β_0 based on our inferential procedure, along with the Wald confidence regions. We also report the estimates, $\hat{\beta}$, of β_0 under both $K = 2$ and 7. Both values are very close to the estimate of $\hat{\beta}_m = 0.84$ obtained by Imbens and Rubin (2015) using matching.

Table 3. Confidence regions for β_0 using Likelihood Ratio and Wald procedures

	$K=2$		$K=7$	
Estimate	$\hat{\beta}=0.840$		$\hat{\beta}=0.873$	
	90% CI	95% CI	90% CI	95% CI
LR	$[-0.782,\ 2.382]$	$[-1.110,\ 2.682]$	$[-0.608,\ 2.262]$	$[-0.909,\ 2.527]$
Wald	$[-0.766,\ 2.445]$	$[-1.073,\ 2.753]$	$[-0.590,\ 2.335]$	$[-0.870,\ 2.615]$

Appendix A. Mathematical appendix

Notation: Hereafter we use the following notation: Let $|A|$ mean the Euclidean norm for a vector A and the spectral norm for a matrix A, "wpa1" mean "with probability approaching one", and

$$
\begin{aligned}
h_i &= h(Y_i, X_i, \theta_0), \qquad \alpha_{0i} = \alpha_{\theta_0}(X_i), \qquad \gamma_{0i} = \gamma_{\theta_0}(X_i), \\
m_i(\gamma_0) &= m(X_i, \gamma_0, \theta_0), \qquad m_i(0) = m(X_i, 0, \theta_0), \qquad \tilde{m}_i(\gamma) = m_i(\gamma) - m_i(0), \\
Q_{Ki} &= Q_K(X_i), \quad M_{Ki} = M_K(X_i, \theta_0), \\
\hat{\varepsilon}_{Ki} &= \hat{\alpha}_i Q_{Ki} - M_{Ki}, \qquad \varepsilon_{Ki} = \alpha_{0i} Q_{Ki} - M_{Ki}, \\
\hat{\varepsilon}_{hi} &= \hat{\alpha}_i h_i - \tilde{m}_i(\gamma_0), \qquad \varepsilon_{hi} = \alpha_{0i} h_i - \tilde{m}_i(\gamma_0), \\
\hat{\varepsilon}_{\gamma_0 i} &= \hat{\alpha}_i \gamma_{0i} - \tilde{m}_i(\gamma_0), \qquad \varepsilon_{\gamma_0 i} = \alpha_{0i}\gamma_{0i} - \tilde{m}_i(\gamma_0).
\end{aligned}
$$

Also recall $\zeta_K = \sup_{x\in\mathcal{X}} |Q_K(x)|$ and $\zeta_{\varepsilon,K} = \sup_{x\in\mathcal{X}} |\varepsilon_K(x)|$. Let $\mathcal{P}_i(a_i|\hat{\varepsilon}_{Ki}) = \hat{\varepsilon}'_{Ki}(\hat{\varepsilon}'\hat{\varepsilon})^{-1}\hat{\varepsilon}'a$ for $i = 1, \ldots, N$ be the empirical projection of a vector $a = (a_1, \ldots, a_N)'$ to $\hat{\varepsilon} = (\hat{\varepsilon}_{K1}, \ldots, \hat{\varepsilon}_{KN})'$.

A.1. **Proof of Theorem.** By Lemma 1 (iv), $\ell(\theta_0)$ exists uniquely wpa1, and we can establish a quadratic expansion of the dual form in (9) as

$$
\ell(\theta_0) = \left(\frac{1}{\sqrt{N}}\sum_{i=1}^N g_i^K\right)\left(\frac{1}{N}\sum_{i=1}^N g_i^K g_i^{K\prime}\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N g_i^K\right) + R_N, \tag{15}
$$

where R_N is the remainder term. Based on Lemma 3 (v) and $\max_{1\le i\le N}|g_i^K| \le \max_{1\le i\le N}|D_{1i}| + \max_{1\le i\le N}|D_{2i}| = o_p(\sqrt{N})$ (by Lemma 1 (i)-(ii)), a similar argument as that used in Hjort, McKeague and Van Keilegom (2009, proof of Theorem 2.1, p. 1105) yields $R_N \xrightarrow{p} 0$. Since $\sum_{i=1}^N (M_{Ki} - \hat{\alpha}_i Q_{Ki}) = 0$ (due to (7)), the definition of g_i^K and inversion formula for partitioned matrices imply that the first term on the right hand side of (15) can be written as

$$
\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\}\right)' [\hat{V}_0 - \hat{V}_1]^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\}\right),
$$

where $\hat{V}_0 = \frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\}^2$ and

$$
\hat{V}_1 = \left(\frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\}\hat{\varepsilon}_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\}\hat{\varepsilon}_{Ki}\right).
$$

Now, Lemma 2 implies $\frac{1}{\sqrt{N}}\sum_{i=1}^N\{\hat{\alpha}_i h_i + m_i(0)\} \xrightarrow{d} N(0,V)$, and Lemma 4 implies $\hat{V}_0 - \hat{V}_1 \xrightarrow{p} V_0 - V_1$. Since the condition in (10) guarantees $V = V_0 - V_1$, the conclusion follows.

On the other hand, if (10) does not hold, the conclusion follows by Lemma 6.

A.2. **Lemmas.**

Lemma 1. *Let $D_{1i} = (\varepsilon_{Ki}', \alpha_{0i}h_i + m_i(0))'$ and $D_{2i} = ((\hat{\alpha}_i - \alpha_{0i})Q_{Ki}', (\hat{\alpha}_i - \alpha_{0i})h_i)'$. Under Assumptions 1-3, the following statements hold true.*

- **(i):** $\max_{1\le i\le N}|D_{1i}| = o_p(\sqrt{N})$.
- **(ii):** $\max_{1\le i\le N}|D_{2i}| = o_p(1)$.
- **(iii):** *all eigenvalues of $\mathbb{E}[D_{1i}D_{1i}']$ are bounded away from zero for all $K \in \mathbb{N}$.*
- **(iv):** $\mathbb{P}\{0 \in \mathcal{C}_n\} \to 1$*, where $\mathcal{C}_n$ is the interior of the convex hull of $\{g_i^{K+1}, i = 1,\dots,N\}$.*

Proof of (i): The triangle inequality implies $\max_{1\le i\le N}|D_{1i}| \le D_{11} + D_{12}$, where

$$D_{11} = \max_{1\le i\le N}|\varepsilon_{Ki}|, \qquad D_{12} = \max_{1\le i\le N}|\alpha_{0i}h_i + m_i(0)|.$$

Note that $D_{11} \le \zeta_{\varepsilon,K} = o(\sqrt{N})$ by the definition of $\zeta_{\varepsilon,K}$ and Assumption 2 (ii). Also, since $\mathbb{E}[\alpha_{0i}h_i + m_i(0)]^2 < \infty$ by assumption, Owen (2001, Lemma 11.2) implies that $D_{12} = o_p(\sqrt{N})$. Thus, we obtain the conclusion.

Proof of (ii): Note that for each $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}\left\{\max_{1\le i\le N}|h_i| > n^{1/\kappa}C_\epsilon\right\} \le \sum_{i=1}^N \mathbb{P}\{|h_i| > n^{1/\kappa}C_\epsilon\} \le \frac{\mathbb{E}[|h_i|^\kappa]}{C^\kappa} \le \varepsilon,$$

where the first inequality follows from the union bound, the second inequality follows from Markov's inequality, and the last inequality follows from Assumption 3 (iii). Therefore, by Assumption 3 (iii),

$$\max_{1\le i\le N}|D_{2i}| \le \delta_{\alpha,N}\left(\max_{1\le i\le N}|\ Q_{Ki}\ | + \max_{1\le i\le N}|h_i|\right) = \delta_{\alpha,N}(\zeta_K + O_p(n^{1/\kappa})) = o_p(1). \tag{16}$$

Proof of (iii): Note that $\lambda_{\min}\{\mathbb{E}[D_{1i}D_{1i}']\} = \min\{\mathring{V}, \lambda_{\min}\{\mathbb{E}[\varepsilon_{Ki}\varepsilon_{Ki}]\}$, where

$$\mathring{V} = \mathbb{E}[\alpha_{0i}h_i + m_i(0)]^2 - \mathbb{E}[(\alpha_{0i}h_i + m_i(0))\varepsilon_{Ki}]'\left(\mathbb{E}[\varepsilon_{Ki}\varepsilon_{Ki}]\right)^{-1}\mathbb{E}[(\alpha_{0i}h_i + m_i(0))\varepsilon_{Ki}].$$

Similar to Lemma 4 (ii), we can show

$$\begin{aligned}
&\mathbb{E}[(\alpha_0 h + m(0))\varepsilon^{Q_K}]'\left(\mathbb{E}[\varepsilon^{Q_K}\varepsilon^{Q_K\prime}]\right)^{-1}\mathbb{E}[(\alpha_0 h + m(0))\varepsilon^{Q_K}] \\
\to\ &\mathbb{E}[\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}^2] + 2\mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}
\end{aligned}$$

as well. Hence $\mathring{V} \to \mathbb{E}[m_i(\gamma_0) + \alpha_{0i}(h_i - \gamma_{0i})]^2 > 0$ as well. And by assumption $\lambda_{\min}(\mathbb{E}[\varepsilon^{Q_K}\varepsilon^{Q_K\prime}])$ has all eigenvalues bounded away from zero by assumption. Conclusion follows.

Proof of (iv): Denote $\hat{H}_n(a) = \min_{1\le i\le N}(a'g_i^{K+1})$. It suffices to show

$$\mathbb{P}\left\{\max_{a\in\mathbb{S}^K}\hat{H}_n(a) < 0\right\} \to 1, \tag{17}$$

as $n \to \infty$, where $\mathbb{S}^K = \{a \in \mathbb{R}^{K+1} : |a| = 1\}$. To this end, let $H_n(a) = \min_{1\le i\le N}(a'D_{1i})$. Observe that $|\hat{H}_n(a) - H_n(a)| \le \max_{1\le i\le N}|a'D_{2i}| \le \max_{1\le i\le N}|D_{2i}| = o_p(1)$ for all $a \in \mathbb{S}^K$, where the last inequality follows from Lemma 1 (i). Similarly, we have $|H_n(a) - H_n(b)| \leq |a-b| \max_{1\leq i\leq N} |D_{1i}| \leq |a-b| o_p(\sqrt{N})$ for all $a, b \in \mathbb{S}^K$, where the last inequality follows from Lemma 1 (ii). Let $\mathbb{U}_{N,K}$ be the union of a finite number $C_{K,N}$ of rectangles with side length δ_N, where $C_{K,N}\delta_N^{K-1} \geq 2\pi^{K/2}/\Gamma(K/2)$ for the gamma function $\Gamma(\cdot)$ [note: $2\pi^{K/2}/\Gamma(K/2)$ is the the surface area of of $\mathbb{S}^K$]. It follows

$$\max_{a\in\mathbb{S}^K} \hat{H}_n(a) \leq \max_{a\in\mathbb{U}_{N,K}} H_n(a) + \delta_n \max_{1\leq i\leq N} |D_{1i}| + \max_{1\leq i\leq N} |D_{2i}|.$$

For (17), it is sufficient to show show that for each $\epsilon > 0$,

$$\mathbb{P}\left\{\max_{1\leq i\leq N} |D_{2i}| \leq \frac{\epsilon}{2}\right\} \rightarrow 1, \tag{18}$$

$$\mathbb{P}\left\{\delta_N \max_{1\leq i\leq N} |D_{1i}| \leq \frac{\epsilon}{2}\right\} \rightarrow 1, \tag{19}$$

$$\mathbb{P}\left\{\max_{a\in\mathbb{U}_{N,K}} H_n(a) < -\epsilon\right\} \rightarrow 1. \tag{20}$$

The convergence in (18) is guaranteed by $\max_{1\leq i\leq N} |D_{2i}| = o_p(1)$. The convergence in (19) is guaranteed by setting $\delta_N = \frac{\epsilon}{C\sqrt{N}}$ for some $C > 0$ since $\max_{1\leq i\leq N} |D_{1i}| = o_p(\sqrt{N})$. By Hjort, McKeague and Van Keilegom (2009, Lemma 4.2), the convergence in (20) is guaranteed if $\frac{K\log N}{N} \rightarrow 0$ and $\mathbb{E}[D_{1i}D_{1i}']$ has eigenvalues bounded away from zero, which follow by Assumption 2 and Lemma 1 (iii).

Lemma 2. *Under the assumptions of Theorem, it holds* $\frac{1}{\sqrt{N}}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\} \xrightarrow{d} N(0, V)$, *where* $V = \mathbb{E}[\{m_i(\gamma_0) + \alpha_{0i}(h_i - \gamma_{0i})\}^2]$.

Proof: Decompose $\frac{1}{\sqrt{N}}\sum_{i=1}^N \{\hat{\alpha}_i h_i + m_i(0)\} = \frac{1}{\sqrt{N}}\sum_{i=1}^N \phi_i + E_1 + E_2$, where

$$\begin{aligned}
\phi_i &= m_i(\gamma_0) + \alpha_{0i}(h_i - \gamma_{0i}), \\
E_1 &= \frac{1}{\sqrt{N}}\sum_{i=1}^N (\hat{\alpha}_i - \alpha_{0i})(h_i - \gamma_{0i}), \qquad E_2 = \frac{1}{\sqrt{N}}\sum_{i=1}^N (\hat{\alpha}_i\gamma_{0i} - \tilde{m}_i(\gamma_0)).
\end{aligned}$$

Since $\frac{1}{\sqrt{N}}\sum_{i=1}^N \phi_i \xrightarrow{d} N(0, V)$ by the central limit theorem, it is sufficient for the conclusion to show that $E_1 \xrightarrow{p} 0$ and $E_2 \xrightarrow{p} 0$.

Since $\mathbb{E}[h_i - \gamma_{0i}|X_i = x] = 0$ (by the definition of γ_0), the law of iterated expectations yields

$$\mathbb{E}[E_1] = \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{E}\left[(\hat{\alpha}_i - \alpha_{0i})\mathbb{E}[h_i - \gamma_{0i}|X_1, \cdots X_N]\right] = 0.$$

Also as $\sup_{x\in\mathcal{X}} \mathbb{E}[(h_i - \gamma_{0i})^2|X_i = x] \lesssim 1$, the same argument in Qiu (2020, Lemma S4) implies

$$\begin{aligned}
&Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N (\hat{\alpha}_i - \alpha_{0i})(h_i - \gamma_{0i})\right) \lesssim \mathbb{E}\left[Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N (\hat{\alpha}_i - \alpha_{0i})(h_i - \gamma_{0i})\middle| X_1, \cdots, X_N\right)\right] \\
= \; & \frac{1}{N}\sum_{i=1}^N \mathbb{E}[(\hat{\alpha}_i - \alpha_{0i})^2 Var(h_i - \gamma_{0i}|X_1, \cdots, X_N)] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[(\hat{\alpha}_i - \alpha_{0i})^2 Var(h_i - \gamma_{0i}|X_i)] \\
\lesssim \; & \frac{1}{N}\sum_{i=1}^N \mathbb{E}(\hat{\alpha}_i - \alpha_{0i})^2 \leq \sup_{x\in\mathcal{X}} |\hat{\alpha}(x) - \alpha_0(x)|^2.
\end{aligned}$$

Thus, Markov's inequality and Assumption 1 (iv) imply $E_1 \xrightarrow{p} 0$.

We now show $E_2 \xrightarrow{p} 0$. By linearity of $\tilde{m}$ and $\gamma_{0i} = \beta_K' Q_{Ki} + r_{Ki}$, we have $E_2 = E_{21} + E_{22} + E_{23}$, where

$$\begin{aligned} E_{21} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \beta_K' (\hat{\alpha}_i Q_{Ki} - M_{Ki}), \qquad E_{22} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\alpha_{0i} r_{Ki} - \tilde{m}(r_{Ki})), \\ E_{23} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_{0i}) r_{Ki}. \end{aligned}$$

Note that $E_{21} = 0$ by the construction of $\hat{\alpha}_i$ in (7). For E_{22}, note that $\mathbb{E}[\alpha_{0i} r_{Ki} - \tilde{m}(r_{Ki})] = 0$ and

$$\mathbb{E}[E_{22}^2] \lesssim \mathbb{E}[\alpha_{0i}^2 r_{Ki}^2] + \mathbb{E}[\tilde{m}(r_{Ki})^2] \lesssim \eta_K^2,$$

where the last inequality follows from Assumption 2 (iii). So, Markov's inequality implies $E_{22} \xrightarrow{p} 0$. Finally, Assumptions 1 (iv) and 2 (iii) and the condition $\sqrt{N}\delta_{\alpha,N}\eta_K \to 0$ guarantee $|E_{23}| \leq \sqrt{N}\delta_{\alpha,N}\eta_K = o_p(1)$. Combining these results, we obtain $E_2 \xrightarrow{p} 0$, and the conclusion follows.

Lemma 3. *Under Assumptions 1-3, the following statements hold true.*

- **(i):** $\left| \frac{1}{N} \sum_{i=1}^{N} Q_{Ki} Q_{Ki}' - \mathbb{E}[Q_{Ki} Q_{Ki}'] \right| \xrightarrow{p} 0$, *and* $\lambda_{\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} Q_{Ki} Q_{Ki}' \right\}$ *is bounded away from zero wpa1.*
- **(ii):** $\left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{Ki} \varepsilon_{Ki}' - \mathbb{E}[\varepsilon_{Ki} \varepsilon_{Ki}'] \right| \xrightarrow{p} 0$, *and* $\lambda_{\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{Ki} \varepsilon_{Ki}' \right\}$ *is bounded away from zero wpa1.*
- **(iii):** $\left| \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{Ki} \hat{\varepsilon}_{Ki}' \right| = O_p(1)$.
- **(iv):** *Under* $\delta_{\alpha,N} \zeta_K \zeta_{\varepsilon,K} \to 0$, *it holds* $\left| \left(\frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{Ki} \hat{\varepsilon}_{Ki}' \right)^{-1} \right| = O_p(1)$.
- **(v):** $\lambda_{\max} \left\{ \frac{1}{N} \sum_{i=1}^{N} g_i^K g_i^{K\prime} \right\} = O_p(1)$ *and* $\lambda_{\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} g_i^K g_i^{K\prime} \right\} = O_p(1)$.

Proof of (i): The proof is similar to that of Part (ii).

Proof of (ii): It follows from Belloni *et al.* (2015, Lemma 6.2) for the first statement, and Tropp (2015, Theorem 5.1.1) for the first statement.

Proof of (iii): By the triangle inequality, stated assumptions and Lemma 3 (i)-(ii), we have

$$\begin{aligned} \left| \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{Ki} \hat{\varepsilon}_{Ki}' \right| &= \max_{s \in \mathbb{S}^{K-1}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (s' \hat{\varepsilon}_{Ki})^2 \right\} \\ &\lesssim \max_{s \in \mathbb{S}^{K-1}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (s' \varepsilon_{Ki})^2 \right\} + \max_{s \in \mathbb{S}^{K-1}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (s' (\hat{\alpha}_i - \alpha_{0i}) Q_{Ki})^2 \right\} \\ &\leq \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{Ki} \varepsilon_{Ki}' \right| + \sup_{x \in \mathcal{X}} |\hat{\alpha}(x) - \alpha_0(x)| \left| \frac{1}{N} \sum_{i=1}^{N} Q_{Ki} Q_{Ki}' \right| \\ &= O_p(1) + O_p(\delta_{\alpha,N}) O_p(1) = O_p(1). \end{aligned}$$

Proof of (iv): Since $(a+b)^2 \geq a^2+b^2-2|ab|$ for $a,b \in \mathbb{R}$, we have

$$\begin{aligned}\lambda_{\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right\} &= \min_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}(s'\hat{\varepsilon}_{Ki})^2\right\} \\ &\geq \min_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}(s'(\hat{\alpha}_i-\alpha_{0i})Q_{Ki})^2\right\}+\min_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}(s'\varepsilon_{Ki})^2\right\} \\ &\quad -2\max_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}|(\hat{\alpha}_i-\alpha_{0i})(s'Q_{Ki})(s'\varepsilon_{Ki})|\right\}.\end{aligned}$$

Note that

$$\max_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}|(\hat{\alpha}_i-\alpha_{0i})(s'Q_{Ki})(s'\varepsilon_{Ki})|\right\}=O_p(\delta_{\alpha,N}\zeta_K\zeta_{\varepsilon,K})=o_p(1).$$

Also $\min_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}(s'(\hat{\alpha}_i-\alpha_{0i})Q_{Ki})^2\right\}=o_p(1)$, and $\min_{s\in\mathbb{S}^{K-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}(s'\varepsilon_{Ki})^2\right\}$ is bounded away from zero wpa1 by Lemma 3 (i)-(ii). Thus, $\lambda_{\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right\}$ is also bounded away from zero wpa1, and the conclusion follows.

Proof of (v): By the definitions of eigenvalue and determinant for partitioned matrix, we have

$$\begin{aligned}\lambda_{\max}\left\{\frac{1}{N}\sum_{i=1}^{N}g_i^K g_i^{K\prime}\right\} &= \max\left\{\hat{V},\lambda_{\max}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}\right\}\right\}. \\ \lambda_{\min}\left\{\frac{1}{N}\sum_{i=1}^{N}g_i^K g_i^{K\prime}\right\} &= \min\left\{\hat{V},\lambda_{\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}\right\}\right\}.\end{aligned}$$

Thus, the conclusion follows from Lemmas 4 and 3 (iii), and $0<\mathbb{E}[m_i(\gamma_0)+\alpha_{0i}(h_i-\gamma_{0i})]^2<\infty$.

Lemma 4. *Under Assumptions 1-3, the following statements hold true.*

(i): $\hat{V}_0=\frac{1}{N}\sum_{i=1}^{N}\{\hat{\alpha}_i h_i+m_i(0)\}^2 \xrightarrow{p} V_0$, *where*

$$V_0=\mathbb{E}[\{\alpha_{0i}h_i+m_i(0)\}^2]=\mathbb{E}[\alpha_{0i}^2\{h_i-\gamma_{0i}\}^2]+\mathbb{E}[\{\alpha_{0i}\gamma_{0i}+m_i(0)\}^2].$$

(ii): $\hat{V}_1 \xrightarrow{p} V_1$, *where*

$$V_1=\mathbb{E}[\{\alpha_{0i}\gamma_{0i}+m_i(0)-m_i(\gamma_0)\}^2]+2\mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i}+m_i(0)-m_i(\gamma_0)\}]$$

Proof of (i): Note that $\hat{V}_0-V_0=\hat{V}_{01}+\hat{V}_{02}$, where

$$\begin{aligned}\hat{V}_{01} &= \frac{1}{N}\sum_{i=1}^{N}\{\hat{\alpha}_i h_i+m_i(0)\}^2-\frac{1}{N}\sum_{i=1}^{N}\{\alpha_{0i}h_i+m_i(0)\}^2. \\ \hat{V}_{02} &= \frac{1}{N}\sum_{i=1}^{N}\{\alpha_{0i}h_i+m_i(0)\}^2-\mathbb{E}[\{\alpha_{0i}h_i+m_i(0)\}^2].\end{aligned}$$

Since the weak law of large numbers implies $\hat{V}_{02}\xrightarrow{p}0$, it suffices to show $\hat{V}_{01}\xrightarrow{p}0$. By using $a^2-b^2=2b(a-b)+(a-b)^2$ for $a,b\in\mathbb{R}$, and the triangle inequality,

$$|\hat{V}_{01}|\leq 2\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})h_i\{\alpha_{0i}h_i+m_i(0)\}\right|+\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})^2h_i^2\right|.$$

The weak law of large numbers implies $\frac{1}{N}\sum_{i=1}^{N}|h_i\{\alpha_{0i}h_i+m_i(0)\}| \xrightarrow{p} \mathbb{E}|h_i\{\alpha_{0i}h_i+m_i(0)\}| \leq \sqrt{\mathbb{E}[h_i^2]}\sqrt{\mathbb{E}[\{\alpha_{0i}h_i+m_i(0)\}^2]}$ and $\frac{1}{N}\sum_{i=1}^{N}h_i^2 \xrightarrow{p} \mathbb{E}[h_i^2]$. Thus, $\hat{V}_{01} \xrightarrow{p} 0$ follows from Assumption 1(iv).

Proof of (ii): Recall $\hat{\varepsilon}_{hi} = \hat{\alpha}_i h_i + m_i(0) - m_i(\gamma_0)$. Decompose

$$\begin{aligned}\hat{V}_1 &= \left(\frac{1}{N}\sum_{i=1}^{N}\{\hat{\varepsilon}_{hi}+m_i(\gamma_0)\}\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\{\hat{\varepsilon}_{hi}+m_i(\gamma_0)\}\hat{\varepsilon}_{Ki}\right)\\ &= A_N + 2B_N + C_N,\end{aligned}$$

where

$$\begin{aligned}A_N &= \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right),\\ B_N &= \left(\frac{1}{N}\sum_{i=1}^{N}m_i(\gamma_0)\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right),\\ C_N &= \left(\frac{1}{N}\sum_{i=1}^{N}m_i(\gamma_0)\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}m_i(\gamma_0)\hat{\varepsilon}_{Ki}\right).\end{aligned}$$

Thus, it is sufficient for the conclusion to show that

$$A_N \xrightarrow{p} \mathbb{E}[\{\alpha_{0i}\gamma_{0i}+m_i(0)-m_i(\gamma_0)\}^2], \tag{21}$$

$$B_N \xrightarrow{p} \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i}+m_i(0)-m_i(\gamma_0)\}], \tag{22}$$

$$C_N \xrightarrow{p} 0. \tag{23}$$

Proof of (21). Let $e_{hi} = h_i - \gamma_{0i}$. Observe that A_N can be decomposed as $A_N = A_{N1} + A_{N2} + 2A_{N3}$, where

$$\begin{aligned}A_{N1} &= \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right),\\ A_{N2} &= \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right),\\ A_{N3} &= \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right),\\ \hat{\varepsilon}_{\gamma_0 i} &= \hat{\alpha}_i\gamma_{0i} - \tilde{m}_i(\gamma_0).\end{aligned}$$

First, we show $A_{N1} \xrightarrow{p} \mathbb{E}[\{\alpha_{0i}\gamma_{0i}+m_i(0)-m_i(\gamma_0)\}^2]$. Observe that

$$A_{N1} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{P}_i^2(\hat{\varepsilon}_{\gamma_0}|\hat{\varepsilon}_K) = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{\gamma_0 i}^2 - \frac{1}{N}\sum_{i=1}^{N}\hat{\mathcal{E}}_{\gamma_0 i}^2,$$

where $\hat{\mathcal{E}}^2_{\gamma_0 i}$ is the projection error of the empirical projection of $\hat{\varepsilon}_{\gamma_0}$ onto $\hat{\varepsilon}_K$. Recall $\varepsilon_{\gamma_0 i} = \alpha_{0i}\gamma_{0i} - \tilde{m}_i(\gamma_0)$. For the first term in A_{N1}, note triangle inequality implies

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}^2_{\gamma_0 i} - \mathbb{E}[\varepsilon^2_{\gamma_0 i}]\right| &\leq 2\left|\frac{1}{N}\sum_{i=1}^N (\hat{\alpha}_i - \alpha_{0i})\gamma_{0i}\varepsilon_{\gamma_0 i}]\right| + \frac{1}{N}\sum_{i=1}^N (\hat{\alpha}_i - \alpha_{0i})^2\gamma^2_{0i} \\ &\quad + \left|\frac{1}{N}\sum_{i=1}^N \varepsilon^2_{\gamma_0 i} - \mathbb{E}[\varepsilon^2_{\gamma_0 i}]\right|.\end{aligned}$$

Thus, Assumptions 1 (iv) and 2 (ii) and the weak law of large numbers imply

$$\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}^2_{\gamma_0 i} \xrightarrow{p} \mathbb{E}[\varepsilon^2_{\gamma_0 i}]. \tag{24}$$

For the second term in A_{N1}, note that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{E}}^2_{\gamma_0 i} &\leq \frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki}) - \tilde{m}_i(\gamma_0 - \beta_K' Q_K)\}^2 \\ &\leq 2\frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki})\}^2 + 2\frac{1}{N}\sum_{i=1}^N \tilde{m}_i(\gamma_0 - \beta_K' Q_K)^2 \\ &= 2A_{N11} + 2A_{N12},\end{aligned}$$

where the first inequality follows from the fact that $\hat{\mathcal{E}}_{\gamma_0 i}$ is the empirical projection error.

For A_{N11}, Assumption 2 (iii) and the triangle inequality imply $A_{N11} \leq \eta^2_{K,N}\left\{\frac{1}{N}\sum_{i=1}^N(\hat{\alpha}_i - \alpha_{0i})^2 + \frac{1}{N}\sum_{i=1}^N \alpha^2_{0i}\right\}$, and the weak law of large numbers and Assumptions 1 (iv) and 2 (iii) yield $A_{N11} \xrightarrow{p} 0$.

For A_{N12}, by Assumption 3 (i), $A_{N12} = \frac{1}{N}\sum_{i=1}^N \tilde{m}_i(\gamma_0 - \beta_K' Q_K)^2 \lesssim \frac{1}{N}\sum_{i=1}^N(\gamma_{0i} - \beta_K' Q_{Ki})^2$ with probability approaching one. Thus $A_{N12} \xrightarrow{p} 0$ by Assumption 2 (iii). Combining these results, we have$A_{N1} \xrightarrow{p} \mathbb{E}[\alpha_{0i}\gamma_{0i} - \tilde{m}_i(\gamma_0)]^2$.

Next, we show $A_{N2} \xrightarrow{p} 0$. Observe by linearity of empirical projection and triangle inequality

$$A_{N2} = \frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i(\hat{\alpha}e_h|\hat{\varepsilon}_K) \leq 2\frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i((\hat{\alpha} - \alpha_0)e_h|\hat{\varepsilon}_K) + 2\frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i(\alpha_0 e_h|\hat{\varepsilon}_K).$$

By definition of empirical projection

$$\frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i((\hat{\alpha} - \alpha_0)e_h|\hat{\varepsilon}_K) \leq \frac{1}{N}\sum_{i=1}^N(\hat{\alpha}_i - \alpha_{0i})^2 e^2_{hi} \leq \left(\sup_{x\in\mathcal{X}}|\hat{\alpha}(x) - \alpha_0(x)|\right)^2 \frac{1}{N}\sum_{i=1}^N e^2_{hi} = o_p(1),$$

where the last inequality follows by law of large numbers under Assumption 3 (ii) and $\sup_{x\in\mathcal{X}}|\hat{\alpha}(x) - \alpha_0(x)| \xrightarrow{p} 0$ by Assumption 1 (iv). Next we show $\frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i(\alpha_0 e_h|\hat{\varepsilon}_K) = o_p(1)$ as well. Since

$$\frac{1}{N}\sum_{i=1}^N \mathcal{P}^2_i(\alpha_0 e_h|\hat{\varepsilon}_K) = \left(\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\hat{\varepsilon}_{Ki}\right)'\left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\hat{\varepsilon}_{Ki}\right),$$

and $\left|\left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right)^{-1}\right| = O_p(1)$ by Lemma 3 (ii), it suffices to show $\left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\hat{\varepsilon}_{Ki}\right| = o_p(1)$. Note

$$\left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\hat{\varepsilon}_{Ki}\right| \leq \left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}(\hat{\alpha}_{0i}-\alpha_{0i})Q_{Ki}\right| + \left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\varepsilon_{Ki}\right|,$$

where the first term is bounded as

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}(\hat{\alpha}_{0i}-\alpha_{0i})Q_{Ki}\right| &\leq \sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\zeta_K\frac{1}{N}\sum_{i=1}^N|\alpha_{0i}e_{hi}| \\ &= O_p(\delta_{\alpha,N}\zeta_K) = o_p(1),\end{aligned}$$

by Assumption 1 (iv) and law of large numbers by Assumption 3 (ii). For the second term, by definition of e_{hi} and iid assumption

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\varepsilon_{Ki}\right|^2 = \frac{1}{N}\mathbb{E}\alpha_{0i}^2e_{hi}^2\varepsilon_{Ki}'\varepsilon_{Ki} \lesssim \frac{\zeta_{\varepsilon,K}^2}{N} \to 0.$$

It follows by Markov inequality that $\frac{1}{N}\sum_{i=1}^N \alpha_{0i}e_{hi}\varepsilon_{Ki} = O_p(\frac{\zeta_{\varepsilon,K}^2}{N}) = o_p(1)$ as well.

Finally, we show $A_{N3} \xrightarrow{p} 0$. Observe that by Cauchy-Schwarz inequality,

$$|A_{N3}| = \left|\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(\hat{\alpha}e_h|\hat{\varepsilon}_K)\hat{\varepsilon}_{\gamma_0}\right| \leq \sqrt{\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i^2(\hat{\alpha}e_h|\hat{\varepsilon}_K)}\sqrt{\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{\gamma_0}^2} = o_p(1),$$

by $A_{N2} \xrightarrow{0} 0$ and (24).

Proof of (22). Recall $\hat{\varepsilon}_{hi} = \hat{\alpha}_i h_i - \tilde{m}_i(\gamma_0)$. By using the empirical projections, decompose

$$\begin{aligned}B_N &= \left(\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\hat{\varepsilon}_{Ki}\right)'\left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right) \\ &= \frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\mathcal{P}_i(\hat{\varepsilon}_h|\hat{\varepsilon}_K) \\ &= \frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\mathcal{P}_i(\hat{\alpha}\gamma_0 + m(0) - m(\gamma_0)|\hat{\varepsilon}_K) + \frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\mathcal{P}_i(\hat{\alpha}e_h|\hat{\varepsilon}_K) \\ &= B_{N1} + B_{N2}.\end{aligned}$$

Let $\hat{\mathcal{E}}_i$ be the empirical projection error of $\hat{\alpha}\gamma_0 + m(0) - m(\gamma_0)$ onto $\hat{\varepsilon}_K$. By the definition of the empirical projection

$$B_{N1} = \frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\{\hat{\alpha}_i\gamma_{0i} + m_i(0) - m_i(\gamma_0)\} - \frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\hat{\mathcal{E}}_i.$$

For the first term of B_{N1}, the triangle inequality implies

$$\begin{aligned}
& \left| \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)\{\hat{\alpha}_i \gamma_{0i} + m_i(0) - m_i(\gamma_0)\} - \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}] \right| \\
\leq \; & \frac{1}{N} \sum_{i=1}^{N} |(\hat{\alpha}_i - \alpha_{0i}) m_i(\gamma_0)\gamma_{0i}| \\
& + \left| \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\} - \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}] \right| \\
\overset{p}{\to} \; & 0,
\end{aligned}$$

where the convergence follows from the weak law of large numbers and Assumption 1 (iv).

For the second term of B_{N1}, the definition of the empirical projection and assumption imply

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{E}}_i^2 \; & \leq \; \frac{1}{N} \sum_{i=1}^{N} \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki}) + \tilde{m}_i(\beta_K' Q_K - \gamma_0)\}^2 \\
& \lesssim \; \frac{1}{N} \sum_{i=1}^{N} \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki})\}^2 + \frac{1}{N} \sum_{i=1}^{N} \tilde{m}_i(\beta_K' Q_K - \gamma_0)^2. \qquad (25)
\end{aligned}$$

For the first term of (25), it holds $\frac{1}{N}\sum_{i=1}^{N}\{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki})\}^2 \lesssim \eta_K^2 \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_i^2 = o_p(1)$ by Assumptions 1 (iv) and 2 (iii). For the second term of (25), the weak law of large numbers and Assumption 2 (iii) yield $\frac{1}{N}\sum_{i=1}^{N} \tilde{m}_i(\beta_K' Q_K - \gamma_0)^2 = o_p(1)$. Thus, we have $\frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{E}}_i^2 \overset{p}{\to} 0$. By this and the Cauchy Schwarz inequality, we obtain

$$\left| \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)\hat{\mathcal{E}}_i \right| \leq \sqrt{\frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)^2} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{E}}_i^2} = o_p(1).$$

Therefore, B_{N1} satisfies $B_{N1} \overset{p}{\to} \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}]$.

For the term B_{N2}, by Cauchy-Schwarz inequality

$$|B_{N2}| \; = \; \left| \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)\mathcal{P}_i(\hat{\alpha} e_h | \hat{\varepsilon}_K) \right| \leq \sqrt{\frac{1}{N} \sum_{i=1}^{N} m_i^2(\gamma_0)} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \mathcal{P}_i^2(\hat{\alpha} e_h | \hat{\varepsilon}_K)}.$$

By law of large numbers and Assumption 3 (ii), $\frac{1}{N}\sum_{i=1}^{N} m_i^2(\gamma_0) = O_p(1)$, and by the proof of $A_{N2} \overset{p}{\to} 0$, $\frac{1}{N}\sum_{i=1}^{N} \mathcal{P}_i^2(\hat{\alpha} e_h | \hat{\varepsilon}_K) = o_p(1)$. Thus, $B_{N2} = o_p(1)$. Conclusion follows by combining the probability limits of B_{N1} and B_{N2}.

Proof of (23). Recall $\varepsilon_{Ki} = \alpha_{0i} Q_{Ki} - M_{Ki}$. Decompose

$$\begin{aligned}
C_N \; & = \; \left(\frac{1}{N} \sum_{i=1}^{N} \{m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i}) Q_{Ki} + m_i(\gamma_0)\varepsilon_{Ki}\} \right)' \left[\frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{Ki} \hat{\varepsilon}_{Ki}' \right]^{-1} \\
& \quad \times \left(\frac{1}{N} \sum_{i=1}^{N} \{m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i}) Q_{Ki} + m_i(\gamma_0)\varepsilon_{Ki}\} \right) \\
& = \; C_{N1} + 2C_{N2} + C_{N3},
\end{aligned}$$

where

$$
\begin{aligned}
C_{N1} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right), \\
C_{N2} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right), \\
C_{N3} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right).
\end{aligned}
$$

For C_{N2}, we further decompose $C_{N2} = C_{N21} + C_{N22}$, where

$$
\begin{aligned}
C_{N21} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right), \\
C_{N22} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}].
\end{aligned}
$$

Also, $C_{N3} = C_{N31} + 2C_{N32} + C_{N33}$, where

$$
\begin{aligned}
C_{N31} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right), \\
C_{N32} &= \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right)' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}], \\
C_{N33} &= \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]' \left[\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}].
\end{aligned}
$$

Note that

$$
\left|\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right| \le \zeta_K \sup_{x\in\mathcal{X}}|\hat{\alpha}(x) - \alpha_0(x)| \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\right) = O_p(\zeta_K\delta_{\alpha,N}),
$$

and

$$
\left|\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right| = O_p(\zeta_{\varepsilon,K}/\sqrt{N}),
$$

where the last equality follows from Markov inequality combined with

$$
\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]\right|^2\right] = \frac{1}{N}\mathbb{E}[m_i(\gamma_0)^2|\varepsilon_{Ki}|^2] \le \frac{\zeta^2_{\varepsilon,K}}{N}.
$$

By these results, Lemma 5, and Assumptions 1 (iv) and 2 (iii), we have

$$
\begin{aligned}
|C_{N1}| &= O_p(\zeta_K^2\delta_{\alpha,N}^2) = o_p(1), \\
|C_{N21}| &= O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}/\sqrt{N}) = o_p(1), \qquad |C_{N22}| = O_p(\zeta_K\delta_{\alpha,N}\sqrt{K}\eta_K) = o_p(1), \\
|C_{N31}| &= O_p(\zeta^2_{\varepsilon,K}/N) = o_p(1), \qquad |C_{N32}| = O_p(\zeta_{\varepsilon,K}\eta_K\sqrt{K/N}) = o_p(1), \\
|C_{N33}| &= O_p(K\eta_K^2) = o_p(1),
\end{aligned}
$$

and the conclusion follows.

Lemma 5. *Under Assumption 1-3, it holds* $|\mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}]| = O(\sqrt{K}\eta_K)$.

Proof: Note that $\mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}] = \mathbb{E}[m_i(0)\varepsilon_{Ki}] + \mathbb{E}[\tilde{m}_i(\gamma_0)\varepsilon_{Ki}]$. Let $r_K = \gamma_0 - \beta_K' Q_K$. It follows

$$\begin{aligned}\mathbb{E}[\tilde{m}_i(\gamma_0)\varepsilon_{Ki}] &= \mathbb{E}[\tilde{m}_i(\gamma_0)\alpha_{0i}Q_{Ki}] - \mathbb{E}[\tilde{m}_i(\gamma_0)M_{Ki}] \\ &= \mathbb{E}[\{\beta_K' M_{Ki} + \tilde{m}_i(r_K)\}\alpha_{0i}Q_{Ki}] - \mathbb{E}[\{\beta_K' M_{Ki} + \tilde{m}_i(r_K)\}M_{Ki}] \\ &= \Xi_1 + \Xi_2,\end{aligned}$$

where $\Xi_1 = \mathbb{E}[\alpha_{0i}Q_{Ki}M_{Ki}']\beta_K - \mathbb{E}[M_{Ki}M_{Ki}']\beta_K$ and $\Xi_2 = \mathbb{E}[(\alpha_{0i}Q_{Ki} - M_{Ki})\tilde{m}_i(r_K)]$. Note that (5) implies $\Xi_1 = 0$. By Cauchy and Schwarz inequality, we have

$$\begin{aligned}|\Xi_2|^2 &\leq \mathbb{E}[|\alpha_{0i}Q_{Ki} - M_{Ki}|^2]\mathbb{E}[\tilde{m}_i(r_K)^2] \\ &\lesssim \text{trace}(\mathbb{E}[\varepsilon_{Ki}\varepsilon_{Ki}'])\mathbb{E}[r_K^2] \lesssim K\eta_K^2 \to 0.\end{aligned}$$

Also, by (5),

$$\begin{aligned}\mathbb{E}[m_i(0)\varepsilon_{Ki}] &= \mathbb{E}[m_i(0)\alpha_{0i}Q_{Ki}] - \mathbb{E}[m_i(0)M_{Ki}] \\ &= \mathbb{E}\begin{bmatrix}\tilde{m}_i(m(0)q_1) \\ \vdots \\ \tilde{m}_i(m(0)q_K)\end{bmatrix} - \mathbb{E}\begin{bmatrix}m_i(0)\tilde{m}_i(q_1) \\ \vdots \\ m_i(0)\tilde{m}_i(q_K)\end{bmatrix} = 0.\end{aligned}$$

Combining these results, the conclusion follows.

Lemma 6. *Suppose assumptions of Theorem 1 hold true except display (10). In addition, (1) if* $\zeta_{\varepsilon,K}^3\zeta_K\delta_{\alpha,N} \to 0$, *and* plim $\left[\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(m(\gamma_0)|\varepsilon_K)^2\right] = \mathcal{V}^*$, *then* $C_N \xrightarrow{p} \mathcal{V}^*$ *; (2) otherwise, if* $\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(m(\gamma_0)|\hat{\varepsilon}_K)^2 \xrightarrow{p} \mathcal{V}^{**}$, *then* $C_N \xrightarrow{p} \mathcal{V}^{**}$.

Proof: Statement (2) is straightforward. We only show statement (1). Note the following decomposition of C_N still holds:

$$C_N = C_{N1} + 2C_{N2} + C_{N3},$$

where C_{N1}, C_{N2}, C_{N3} are defined in the proof of for display (23). Specifically, it still holds $|C_{N1}| = O_p(\zeta_K^2\delta_{\alpha,N}^2) = o_p(1)$. It remains to bound C_{N2} and C_{N3}. Note $\left|\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\varepsilon_{Ki}\right| \leq \zeta_{\varepsilon,K}\frac{1}{N}\sum_{i=1}^N |m_i(\gamma_0)| = O_p(\zeta_{\varepsilon,K})$ since $\frac{1}{N}\sum_{i=1}^N |m_i(\gamma_0)| = O_p(1)$ by the law of large numbers. Recall $\left|\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right| = O_p(\zeta_K\delta_{\alpha,N})$. Hence by Lemma 3(iii),

$$\begin{aligned}|C_{N2}| &\leq \left|\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right| \left|\left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right)^{-1}\right| \left|\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\varepsilon_{Ki}\right| \\ &= O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}) = o_p(1).\end{aligned}$$

For C_{N3}, notice

$$\begin{aligned}C_{N3} &= \left(\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\varepsilon_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)\varepsilon_{Ki}\right) \\ &= \tilde{C}_{N31} + \tilde{C}_{N32},\end{aligned}$$

where

$$\tilde{C}_{N31} = \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right)' \left\{\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} - \left[\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{Ki}\varepsilon'_{Ki}\right]^{-1}\right\} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right),$$

$$\tilde{C}_{N32} = \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{Ki}\varepsilon'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right).$$

Let $\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki} = \hat{\Sigma}_N$, $\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{Ki}\varepsilon'_{Ki} = \Sigma_N$. Then

$$\begin{aligned} |\Sigma_N - \hat{\Sigma}_N| &= \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\varepsilon}_{Ki}-\varepsilon_{Ki})(\hat{\varepsilon}_{Ki}-\varepsilon_{Ki})' + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{Ki}(\hat{\varepsilon}_{Ki}-\varepsilon_{Ki})' + \frac{1}{N}\sum_{i=1}^{N}(\hat{\varepsilon}_{Ki}-\varepsilon_{Ki})\varepsilon'_{Ki}\right| \\ &\le \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})^2 Q_{Ki}Q'_{Ki}\right| + \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})\varepsilon_{Ki}Q'_{Ki}\right| + \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})Q_{Ki}\varepsilon'_{Ki}\right|. \end{aligned}$$

Note $\left|\frac{1}{N}\sum_{i=1}^{N}Q_{Ki}Q'_{Ki}\right| = O_p(1)$ by Lemma 3 (i), so

$$\begin{aligned} \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})^2 Q_{Ki}Q'_{Ki}\right| &= \max_{a\in\mathbb{S}^{K-1}}\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})^2(a'Q_{Ki})^2 \\ &\le \left(\sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\right)^2 \max_{a\in\mathbb{S}^{K-1}}\frac{1}{N}\sum_{i=1}^{N}(a'Q_{Ki})^2 \\ &= \left(\sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\right)^2 \left|\frac{1}{N}\sum_{i=1}^{N}Q_{Ki}Q'_{Ki}\right| = O_p(\delta^2_{\alpha,N}). \end{aligned}$$

Also $\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})\varepsilon_{Ki}Q'_{Ki}\right| \le \sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\sup_{x\in\mathcal{X}}|\varepsilon_K(x)|\sup_{x\in\mathcal{X}}|Q_K(x)| = O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K})$. Similarly $\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})Q_{Ki}\varepsilon'_{Ki}\right| = O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K})$ as well. So $|\Sigma_N - \hat{\Sigma}_N| = O_p(\delta^2_{\alpha,N} + \zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}) = O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K})$. It follows

$$\begin{aligned} |\tilde{C}_{N31}| &= \left|\left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right)' \hat{\Sigma}_N^{-1}\{\Sigma_N - \hat{\Sigma}_N\}\Sigma_N^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki}\right)\right| \\ &= O_p(\zeta^3_{\varepsilon,K}\zeta_K\delta_{\alpha,N}) = o_p(1) \end{aligned}$$

by assumption and Lemmas 3 (i) and (iv). Finally $\tilde{C}_{N32} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{P}_i(m(\gamma_0)|\varepsilon_K)^2 \xrightarrow{p} V^*$ by assumption. The conclusion follows.

References

[1] Abadie, A. (2005) Semiparametric difference-in-differences, *Review of Economic Studies*, 72, 1-19.

[2] Abadie, A. and G. W. Imbens (2016) Matching on the estimated propensity score, *Econometrica*, 84, 781-807.

[3] Angrist, J. D. and A. B. Krueger (1992) The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples, *Journal of the American Statistical Association*, 87, 328-336.

[4] Athey, S., Imbens, G. W. and S. Wager (2018) Approximate residual balancing: de-biased inference of average treatment effects in high dimensions, *Journal of the Royal Statistical Society*, B 80, 597-623.

[5] Belloni, A., Chernozhukov, V., Chetverikov, D. and K. Kato (2015) Some new asymptotic theory for least squares series: pointwise and uniform results, *Journal of Econometrics*, 186, 345-366.

[6] Card, D. and A. B. Krueger (1994) Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review*, 84(4), 772-793.

[7] Carroll, R. J. and M. P. Wand (1991) Semiparametric estimation in logistic measurement error models, *Journal of the Royal Statistical Society*, B 53, 573-585

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

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

[10] Cressie, N. and R. C. Read (1984) Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society*, B 46, 440-464.

[11] Chernozhukov, V., Newey, W. K. and R. Singh (2018) Learning L2 continuous regression functionals via regularized Riesz representers, Working paper.

[12] Dinardo, J., Fortin, N. M. and T. Lemieux (1996) Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach, *Econometrica*, 64, 1001-1044.

[13] Graham, B. S. (2011) Efficiency bounds for missing data models with semiparametric restrictions, *Econometrica*, 79, 437-452.

[14] Graham, B., Pinto, C. and D. Egel (2012) Inverse probability tilting for moment condition models with missing data, *Review of Economic Studies*, 79, 1053-1079.

[15] Graham, B. S., Pinto, C. and D. Egel (2016) Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST), *Journal of Business and Economic Statistics*, 34, 288-301.

[16] Hainmueller, J. (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis*, 20, 25-46.

[17] Hausman, J. A. and W. K. Newey (2017) Nonparametric welfare analysis, *Annual Review of Economics*, 9, 521-546.

[18] Hirshberg, D. A. and S. Wager (2018) Augmented minimax linear estimation, Working paper.

[19] Hjort, N. L., I. W. McKeague and I. Van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[20] Imbens, G. W. and D. B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

[21] Newey, W. K. and J. M. Robins (2018) Cross-fitting and fast remainder rates for semiparametric estimation, Working paper.

[22] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC.

[23] Qin, J. (2017) *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer.

[24] Qin, J. and B. Zhang (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies, *Journal of the Royal Statistical Society*, B 69, 101-122.

[25] Qin, J., Zhang, B. and D. H. Y. Leung (2009) Empirical likelihood in missing data problems, *Journal of the American Statistical Association*, 104, 1492-1503.

[26] Qiu, C. (2020) Near optimal estimation of average regression functionals, Working paper.

[27] Qiu, C. and T. Otsu (2018) Information theoretic approach to high dimensional multiplicative models: stochastic discount factor and treatment effect, Working paper.

[28] Rosenbaum, P. R. (2002) *Observational studies*, Springer, New York.

[29] Tropp, J. A. (2015) An introduction to matrix concentration inequalities, *Foundations and Trends in Machine Learning*, 8, 1-230.

[30] Wang, Y. and J. R. Zubizaretta (2017) Approximate balancing weights: characterizations from a shrinkage estimation perspective, Working paper.

[31] Zubizaretta, J. R. (2015) Stable weights that balance covariates for estimation with incomplete outcome data, *Journal of the American Statistical Association*, 110, 910-922.

Department of Economics, University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104, USA.

Email address: akarun@sas.upenn.edu

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

Email address: t.otsu@lse.ac.uk

The Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK.

Email address: c.qiu@lse.ac.uk