Electronic Journal of Statistics
Vol. 17 (2023) 2810–2848
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2168

# Reweighted nonparametric likelihood inference for linear functionals

**Karun Adusumilli**

*Department of Economics, University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104, USA*
*e-mail:* akarun@sas.upenn.edu

**Taisuke Otsu**

*Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK*
*e-mail:* t.otsu@lse.ac.uk

**Chen Qiu**

*Department of Economics, Cornell University, Uris Hall, Ithaca, NY14853, USA*
*e-mail:* cq62@cornell.edu

**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 reweights the moment conditions so that the likelihood increment admits an asymptotically pivotal chi-square calibration. Our reweighting method is naturally extended to inference on missing data, treatment effects, and data combination models, and other semiparametric problems. Simulation and real data examples illustrate usefulness of the proposed method. We note that our reweighting method and theoretical results are limited to linear functionals.

**MSC2020 subject classifications:** Primary 62G20, 62G10; secondary 62P20.
**Keywords and phrases:** Nonparametric likelihood, linear functional, Riesz representer.

Received May 2022.

## 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 functionals 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 the balancing weights or the inverse propensity score. Recently, several methods that directly balance the distributional characteristics of covariates have been proposed [19, 35, 17, 4, 10]. These methods, based on balancing weights, have been employed to obtain point estimates for the population parameters of interest, and the above authors reported superior performance for the empirical balancing approach.

This paper extends the balancing approach to more general contexts and proposes nonparametric likelihood inference of finite dimensional parameters defined via moment conditions or estimating equations containing the unknown Riesz representer. To this end, one possibility is to construct an empirical likelihood (EL) statistic based on the estimated Riesz representer plugged in the moment conditions. As shown by [23], however, the resulting EL statistic is not asymptotically pivotal in general, invalidating calibration by the chi-squared distribution. [6] restored the asymptotic pivotalness of the EL statistic by working with a debiased version of the original moment condition. In contrast to these works, we show that the plug-in approach can, in fact, achieve asymptotic pivotalness, once all the relevant information of the model has been properly accounted for. However, it should be noted that our approach is limited to linear functionals unlike the general results by [23] and [6] allowing nonlinear functionals. A new insight from our paper is that in order for the plug-in approach to restore the limiting chi-squared distribution, it is crucial to rewrite the original moment condition as a system of *growing* moment conditions, where the additional growing moment conditions are derived from the definition of the Riesz representer.

Our approach is general enough to cover many linear functional models including balancing weights for the average treatment effect as a special case. Specifically, our approach is composed of two steps that involve two different weighting schemes. In the first step, we compute estimates of the Riesz representer using a projection argument. Inspired by the balancing literature for estimating average treatment effects, we call these estimates *projection weights*. Importantly, our projection weights may take negative values, so they can be applied for general linear functional models. In the second step, we *reweigh* the original moment conditions: we capture the nonparametric likelihood increment in going from the baseline likelihood based on the projection weights to the one obtained by adding the original moment conditions for the parameters of interest. The second reweighting step is the key to restoring asymptotic pivotalness.

Our new weighting scheme is useful for conducting statistical inference (i.e., interval estimation and hypothesis testing). 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 reweighting 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 [19] and [10] – 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 [11] and the Reisz representer is then related to projection weights that approximate the odds ratio of the propensity scores. Our simulation evidence and real data example illustrate the usefulness of the proposed method.

Moreover, 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. More broadly, this paper contributes to the literature of estimation and inference on semiparametric models via the Riesz representer. For example, [13] and [25] introduced a series-based estimator for the Riesz representer and considered estimation of finite dimensional parameters with fast decays of the remainders. In a high dimensional framework, [14] considered Wald-type inference with $L_1$-regularized Riesz representers. [22] proposed to estimate the Riesz representer by applying the minimax approach. However, none of these papers considers likelihood-based inference on finite dimensional parameters by developing an asymptotically pivotal statistic.

This paper also contributes to the literature on EL methods (see, [27], for a survey). [29] introduced the EL approach for missing response problems with parametric propensity scores. Subsequently, [30] proposed a unified EL approach to missing data problems. We refer [28] for a comprehensive survey on the EL methods, particularly in the context of missing and biased samples. Our paper shares a similar motivation with the recent work of [6], but our approach is inherently different. Working in a more general framework, [6] proposed asymptotically pivotal nonparametric EL inference by using a bias corrected moment equation. They modify the original moment by adding a correction term that accounts for the impact of the nuisance parameters on the identifying moment. This term, which is nonparametric in nature, is estimated using kernel methods. Our paper complements their work by showing that in a large class of moment models, the bias corrected moment in [6] is not required for asymp-

totical pivotalness (although their correction terms can be easily computed for all the examples considered in our paper), and it is possible to work with the original moment instead. We rely on a reweighting scheme that effectively captures the same information as the bias corrected moment without estimating the additional nuisance parameters or functions. Furthermore, our first step in the reweighting procedure involves computing balancing weights, which practitioners would compute anyway (at least in the case of missing data models) to obtain point estimates. Our confidence intervals also have a useful property that they always include these point estimates. We thus believe our methods may be more appealing to users of balancing weight methods that have gained in popularity in recent years. On the other hand, it should be acknowledged that our reweighting approach and theoretical analysis are confined to linear functionals and that the bias corrected moment approach by [6] is broadly applicable for nonlinear functionals as well.

This paper is organized as follows. Section 2 introduces the basic setup and some examples. In Section 3, we develop the reweighted nonparametric likelihood ratio statistic. Section 4 discusses extensions to inferences on treatment effects, data combination models, 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 $\gamma_{\theta_0}^{(l)}(\cdot) = \mathbb{E}[h^{(l)}(Y, X, \theta_0) \mid X = \cdot] \in L_X^2$ is the conditional expectation function for some known function $h^{(l)} : \mathcal{Y} \times \mathcal{X} \times \Theta \to \mathbb{R}$. We emphasize that our methodology and asymptotic theory are limited to linear functionals unlike other methods, such as [6].

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 $\theta_0$ enters into $m$ or $h$. Therefore, $\gamma_{\theta_0}$ and/or $\alpha_{\theta_0}$ may depend on $\theta_0$ in a possibly nonlinear manner.

This setup covers many well-known statistical inference problems. We give some examples below. In particular, our setup can deal with quantile models in missing data problems and other less investigated problems, such as consumer welfare analysis and data combination models, which are further discussed in Section 4. We note that for these examples with linear functionals, the bias corrected moments in [6] are easy to compute.

**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 $\theta_0$, which is identified by the moment condition

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

where $\psi$ is possibly nonlinear in $\theta_0$. 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 $F_\pi$ 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 consumer 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 $\theta_0$ 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. [20] showed that

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

where $l(p, V) = w(V)\mathbb{I}\{p_a \le p \le p_b\}\exp(-B(p - p_a))$ 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, $\gamma_{\theta_0}$ and $h(Y,X,\theta_0)$ do not involve $\theta_0$, 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 $\theta_0$ defined by the linear functional model in (1), or alternatively (3). Let $\{q_j(\cdot)\}_{j=1}^{\infty}$ denote basis functions for the space $L_X^2$. Then the condition for the Riesz representer $\alpha_{\theta_0}$ in (2) is equivalent to the infinite set of moment conditions

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

The equivalence between (2) and (6) exploits the fact that $\gamma\mapsto m(\cdot,\gamma,\cdot)$ is a linear functional, which ensures that for each $\gamma\in L_X^2$, 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 $\alpha_{\theta_0}$, 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, 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 [25], 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

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

i.e.,

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

for $i=1,\ldots,N$. Following [31], one can interpret $\mathbb{E}[\alpha_{\theta_0}(X)^2]$ as the residual variance for estimation of $\theta_0$, 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., [35, 10, 32]. 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 $\alpha_i$ 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 'reweigh' both the moment functions (2) and (3) after incorporating the the projection weights in (7). The 'reweighted' likelihood ratio then captures the likelihood increments associated with (3). Formally, let

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

denote the Cressie and Read [12] 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 EL and exponential tilting, respectively. Other popular choices for $\varsigma$ include the Neyman's modified $\chi^2$ ($\varsigma = 1$), Hellinger or Freeman-Tukey ($\varsigma = -1/2$), and Pearson's $\chi^2$ ($\varsigma = -2$). Based on this divergence, we consider the following reweighted likelihood:

$$\begin{aligned}
&\ell(\theta_0) = \min_{\omega_1,\ldots,\omega_N} \sum_{i=1}^N \phi_\varsigma(\omega_i), \\
\text{s.t.} \quad &\sum_{i=1}^N \omega_i\{M_K(X_i,\theta_0) - \hat{\alpha}_i Q_K(X_i)\} = 0, \qquad \sum_{i=1}^N \omega_i = 1, \qquad \omega_i \geq 0, \\
&\sum_{i=1}^N \omega_i\{\hat{\alpha}_i h^{(l)}(Y_i, X_i, \theta_0) + m(X_i, 0, \theta_0)\} = 0 \quad \text{for } l = 1, \ldots, d_\theta, \qquad (8)
\end{aligned}$$

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 reweighted nonparametric 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 inference procedure.

**Remark 1.** Our approach is reminiscent of the generalized empirical likelihood (GEL) inference for moment condition models [26]. In fact, if the Riesz representer is known up to $\theta_0$ in (3), then the GEL methodology directly applies for inference on $\theta_0$. In our setup, however, the Riesz representer is unknown. Therefore, our setup is more involved than the standard GEL framework and needs to consider the impact of not knowing the Riesz representer in our asymptotic analysis. The growing set of moment constraints in (8) exactly captures the price we need to pay to restore the asymptotic pivotalness of the likelihood ratio statistic.

**Remark 2.** On the other hand, our inference approach has some connection to extended versions of the GMM/GEL approach since we may interpret our inference problem using just-identified but growing number of moment conditions. To elaborate on this connection, suppose $\alpha_{\theta_0}(\cdot)$ can be approximated by $a_0' Q_K(\cdot)$ with the $K$-dimensional vector of basis functions $Q_K(\cdot)$. Then our setup in (1) and (2) can be alternatively characterized as

$$\mathbb{E}[a_0' Q_K(X) h(Y, X, \theta_0) + m(X, 0, \theta_0)] = 0, \tag{10}$$

$$\mathbb{E}[m(X, q_j, \theta_0) - m(X, 0, \theta_0) - q_j(X) Q_K'(x) a_0] = 0, \tag{11}$$

for $j = 1, \ldots, K$, where (10) is the moment condition to identify $\theta_0$, and (11) is the auxiliary moment conditions to estimate $\alpha_{\theta_0}(\cdot)$. A similar argument as in [16] indicates that these moment conditions lead to the semiparametric efficiency bound for $\theta_0$ as $K \to \infty$, i.e., (10) and (11) contain all relevant information for $\theta_0$. Furthermore, the GEL ratio statistic (say, $GELR_K$) for $\mathrm{H}_0 : \theta_0 = c$ may be constructed based on (10) and (11), and we expect that $GELR_K$ will converge to the $\chi^2_{d_\theta}$ distribution.

Compared to $GELR_K$, our approach offers new insight for the popular plug-in approach of EL inference. As opposed to the well-known result in [23], we show that in fact, plug-in approach can still achieve asymptotic pivotalness, once the information of the model has been properly accounted for. Indeed, if the Reisz representer is actually known, and we use them in place of $\hat{\alpha}_i$ in (8), $\ell(\theta_0)$ will converge to $\chi^2_{K+d_\theta}$. Surprisingly, however, we show that plugging-in the estimates from balancing leads to a $\chi^2_{d_\theta}$ distribution. The construction of balancing weights and the definition of $\ell(\theta_0)$ implies that the latter does behave

like a likelihood ratio test statistic, and is in sharp contrast of the result in [23]. In terms of the treatment of the unknown Riesz representer, our approach relies on the first step balancing in (7), while $GELR_K$ requires a sieve approximation. Moreover, $\ell(\theta_0)$ also has computational advantage over $GELR_K$ particularly when we compute confidence sets for $\theta_0$. To construct the confidence set by inverting $GELR_K$, we need to profile out the nuisance parameters $a_0$ for each hypothetical value of $\theta_0$. On the other hand, the confidence interval obtained by inverting our statistic $\ell(\theta_0)$ does not involve such profiling out.

**Remark 3.** Note that when $\varsigma = 1$, the criterion $\rho_\varsigma(v)$ becomes the least square likelihood (as in [7]) or the Euclidean likelihood (as in [26]), where $\rho_1(v) = -\frac{1}{2}(1+v)^2$. In this case, the minimization problem in (8) admits an explicit solution that depends only on $\hat{\alpha}_i$ in (7). Thus, this likelihood ratio statistic is computationally attractive.

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 a 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)** *$\sup_{x\in\mathcal{X}}|\hat{\alpha}(x) - \alpha_{\theta_0}(x)| = O_p(\delta_{\alpha,N})$ for some $\delta_{\alpha,N} \to 0$. For each $l = 1,\dots,d_\theta$ and $K \in \mathbb{N}$, there exists some $\beta_K^{(l)} \in \mathbb{R}^K$ such that*

$$\sup_{x\in\mathcal{X}}|\gamma_{\theta_0}^{(l)}(x) - \beta_K^{(l)\prime}Q_K(x)| \lesssim \eta_K \quad \text{and} \quad \mathbb{E}[\tilde{m}(X,\gamma_{\theta_0}^{(l)} - \beta_K^{(l)\prime}Q_K,\theta_0)^2] \lesssim \eta_K$$

*for some $\eta_K \to 0$.*

**(iv)** *For each $l = 1,\dots,d_\theta$, 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*

$$\mathbb{E}[|h^{(l)}(X,Y,\theta_0)|^\kappa] \lesssim 1 \quad \text{and} \quad N^{1/\kappa}\delta_{\alpha,N} \to 0.$$

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

$$\begin{aligned}\mathbb{E}[\tilde{m}(X,q_j(\cdot)\tilde{m}(\cdot,q_k,\theta_0),\theta_0)] &= \mathbb{E}[\tilde{m}(X,q_j,\theta_0)\tilde{m}(X,q_k,\theta_0)],\\ \mathbb{E}[\tilde{m}(X,m(\cdot,0,\theta_0)q_j(\cdot),\theta_0)] &= \mathbb{E}[m(X,0,\theta_0)\tilde{m}(X,q_j,\theta_0)].\end{aligned} \tag{12}$$

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 basic approximation quality of $\hat{\alpha}_i$ in terms of sup norm. This could be verified by more primitive conditions; see [26]. The second condition in Assumption (iii) is on the approximation error for the conditional mean function $\gamma_{\theta_0}^{(l)}(\cdot)$ by the vector of basis functions $Q_K(\cdot)$. Assumption (iv) is a set of regularity conditions. Notably, we require the Riesz representer $\alpha_{\theta_0}$ 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 $\gamma$ (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 (12) 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 (12) 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, the condition in (12) is not generally satisfied. One exception is the setting where $w(x) = 1$ and $d^2q_k(x)/dx^2 = 0$ 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 Ströberg 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 that Assumptions (i)–(v) hold true, and*

$$\delta_{\alpha,N}\zeta_K\zeta_{\varepsilon,K} \to 0, \sqrt{K}\eta_K \to 0, \text{ and } \sqrt{N}\delta_{\alpha,N}\eta_K \to 0.$$

*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 $\theta_0$ 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 reweighted statistic $\ell(\theta_0)$ loses its asymptotic pivotalness and converges to a weighted $\chi^2$ 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 $d_\theta$-dimensional random vector whose $l$-th element is $m(X,\gamma_0^{(l)},\theta_0) + \alpha_0(X)\{h^{(l)}(X,Y,\theta_0) - \gamma_{\theta_0}^{(l)}(X)\}$. 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 [23]). 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 $\theta_0$ identified by the moment condition

$$\mathbb{E}[\psi_1(Y_1,Z,\theta_0) - \psi_0(Y_0,Z,\theta_0)] = 0, \tag{13}$$

where $\psi_1$ and $\psi_0$ have the same dimension as $\theta_0$. 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 $\theta_0$ 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, (13) can be rewritten as

$$\mathbb{E}[\gamma^{(1)}_{\theta_0}(1, Z) - \gamma^{(0)}_{\theta_0}(0, Z)] = 0. \tag{14}$$

Further, under the overlap condition, (13) gives rise to the two Riesz representers $\alpha^{(1)}_{\theta_0}(x) = \frac{d}{\mathbb{P}(D=1|Z=z)}$ and $\alpha^{(0)}_{\theta_0}(x) = \frac{1-d}{1-\mathbb{P}(D=1|Z=z)}$ for $x = (d, z)$ so that

$$\begin{aligned}
\mathbb{E}[\alpha^{(1)}_{\theta_0}(X)\gamma(X)] &= \mathbb{E}[\gamma(X)], \quad \text{for each } \gamma \in L^2_X, \\
\mathbb{E}[\alpha^{(0)}_{\theta_0}(X)\gamma(X)] &= \mathbb{E}[\gamma(X)], \quad \text{for each } \gamma \in L^2_X.
\end{aligned}$$

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 $\theta_0$ is

$$\mathbb{E}[\alpha^{(1)}_{\theta_0}(X)\psi_1(Y, Z, \theta_0) - \alpha^{(0)}_{\theta_0}(X)\psi_0(Y, Z, \theta_0)] = 0, \tag{15}$$

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

$$\min_{\alpha^{(1)}_1, \ldots, \alpha^{(1)}_N} \frac{1}{2}\sum_{i=1}^N D_i\alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^N \{Q_K(X_i) - \alpha^{(1)}_i D_i Q_K(X_i)\} = 0;$$

$$\min_{\alpha^{(0)}_1, \ldots, \alpha^{(0)}_N} \frac{1}{2}\sum_{i=1}^N (1 - D_i)\alpha_i^2 \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^N \{Q_K(X_i) - \alpha^{(0)}_i (1 - D_i) Q_K(X_i)\} = 0;$$

Based on the approximated Riesz representers

$$\{D_i\hat{\alpha}^{(1)}_i\}_{i=1}^N \quad \text{and} \quad \{(1 - D_i)\hat{\alpha}^{(0)}_i\}_{i=1}^N,$$

our reweighted likelihood ratio statistic can be constructed as

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

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

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 $\mathrm{H}_0$, where $d_\theta$ is the dimension of $\theta_0$. 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 $d_\theta$-dimensional vector of parameters, $\theta_0$, 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, [11] showed that the parameter vector $\theta_0$ 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 [3], counterfactual distributions [15], semiparametric differences-in-differences [1], and models with mismeasured regressors in the presence of validation samples [9].

Following [17], we employ a multinomial sampling framework by assuming that a unit is drawn at random from the distribution of the study sample with probability $\pi$, 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 ([18]). 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 $X = (D, Z)$, $\gamma^{(1)}_\theta(D, Z) = \mathbb{E}[\psi_s(Y, Z, \theta_0)|D, Z]$ and $\gamma^{(2)}_\theta(D, Z) =$

$\mathbb{E}[\psi_u(Y,Z,\theta_0)|D,Z]$. 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 $z\in\mathbb{R}^{d_Z}$. 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, [16]). 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

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

In this case, our likelihood ratio statistic is obtained as

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

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

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 $\mathrm{H}_0$, where $d_\theta$ is the dimension of $\theta_0$.

### *4.3. Over-identified models*

Thus far we have considered inference under just-identification. In some applications however, the parameters $\beta$ 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 $\beta$. Then we can construct a likelihood ratio test by considering the discrepancy in the log-likelihoods evaluated at the estimated and hypothesized values of $\beta$. 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 $\mathrm{H}_0$. Note that the degree of freedom of the limiting distribution is $p$, the dimension of $\beta$. On the other hand, the statistic $\ell(\beta_0)$ converges to the chi-square distribution with degree of freedom $p_1$, the dimension of $\psi$.

## 5. Simulation

In this section, we study the finite sample performance of the proposed likelihood ratio test in a missing data setup. Our main findings are: when the data generating process (DGP) is relatively simple and when the overlap is good, all methods perform well for testing the mean. However, when the overlap worsens, the performance of the Wald statistic is less satisfactory. The debiased likelihood ratio (DLR) test proposed in [6] performs better than the Wald statistic but appears to perform worse than our method particularly when the support of outcome variables is unbounded. When it comes to testing the median, the Wald statistic using bootstrapped variance performs erratically while the both likelihood ratio tests perform much better. The DLR test performs slightly better than our method for testing the median.[1]

We consider two different DGPs. The first DGP (DGP1) is taken from [2] (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 the 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

[1] For example, the conventional Wald test using the inverse probability weighting requires the propensity score to be correctly specified and highly smooth (i.e., continuously differentiable of order larger than $7d_x$ as in [21]). On the other hand, our test basically requires $\sqrt{n}\delta_{\alpha,N}\eta_K \to 0$, where $\delta_{\alpha,N}$ and $\eta_K$ are the sup-norm convergence rates for estimating the Riesz representer and the conditional expectation function, respectively. This condition may be significantly weaker than the one for the Wald test. Compared with [6] approach, their test also needs estimation of the propensity score, which enters the moment condition nonlinearly. Our method avoids estimating the propensity score, and instead estimates the Riesz representer, which enters the moment condition linearly. These features could be reasons to explain robustness of our test for different degrees of overlaps compared to the other tests.

random variable. The propensity score is given by the logistic function

$$\mathbb{P}(D=1|Z)=\frac{\exp(Z_1+tZ_2)}{1+\exp(Z_1+tZ_2)}, \tag{16}$$

for $t \in [1,6]$. The effect of increasing $t$ is to reduce amount of overlaps in the propensity score. The treatment $D$ is generated by this probability, and the observed outcome variable is generated by $Y = DY_1$.

The second DGP (DGP2) is a more challenging case, where the potential outcome has an unbounded support. We generate $Y_1 = 1 + Z_1 + 2Z_2 + U$, where $(Z_1, Z_2)$ follows the bivariate standard normal, and $U$ is a standard normal random variable independent of $(Z_1, Z_2)$, and the true propensity score is set as

$$\mathbb{P}(D=1|Z)=\frac{\exp(tZ_2)}{1+\exp(tZ_2)},$$

for $t \in [1,6]$. Again the larger value of $t$ implies reduced overlaps in the propensity score.

We compare inference on the average outcome $\beta_a = \mathbb{E}[Y_1]$ using four methods: (1) reweighted likelihood ratio (RLR) test proposed in our paper, with $\varsigma = -1$, corresponding to EL; (2) Wald statistic (Wald 1) using inverse propensity score weighting proposed in [21]; (3) Wald statistic (Wald 2) using balancing and the variance estimate proposed by [10]; (4) debiased likelihood ratio (DLR) test proposed in [6] with EL.

Figures 1 and 2 plot the rejection frequencies under the null $H_0 : \beta_a = 5$ for different values of $t$ under DGP1 and DGP2, respectively. The nominal significance level is 0.05. We report the results with the sample size $N = 100$, 200, and 500, and with the number of basis functions $K = 3$ (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 simulation results are based on 5,000 Monte Carlo repetitions.

For DGP1 displayed in Figure 1, when the overlap is good (say, $t \leq 2$), all four methods perform reasonably well. However, the Wald statistics are highly sensitive to overlap conditions and their performance deteriorate quickly as $t$ increases. For example, when $t = 6$, $K = 3$, and $N = 100, 200$, the null rejection frequencies of Wald 1 are around 20% for the nominal 5% level. When $t = 6$, $K = 5$, and $N = 100$, the null rejection frequency of Wald 2 is around 15%. Although RLR and DLR also exhibit over-rejections as $t$ increases particularly when $n$ is small, they are less sensitive to the Wald statistics, and achieve valid size controls when $N = 500$. For DGP1, the performances of RLR and DLR are comparable.

DGP2 displayed in Figure 2 is a clear case that our proposed method (RLR) outperforms other methods. First of all, the over-rejections by the Wald statistics become much severer; both may over-rejects around 50% or more in some cases. Second, when $K = 5$, DLR severely over-rejects as well; around 30% for $N = 100$, and 20% even for $N = 500$ for the nominal 5% level. Although such over-rejections for DLR are relatively less severe when $K = 3$ (but still larger than 10%), these results indicate sensitivity of DLR for the number of basis

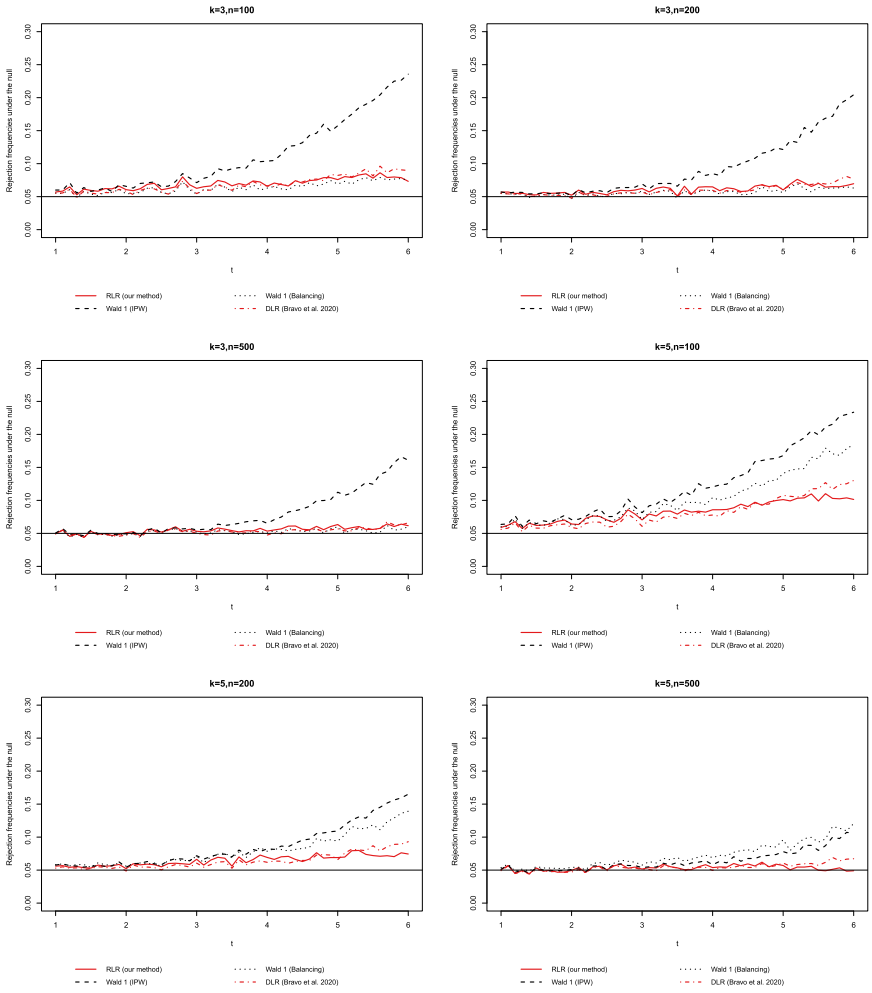


FIG 1. *Rejection frequencies under the null for inference on $\beta_a$, DGP1.*

functions $K$. Finally, the proposed RLR exhibits robust size properties across all the cases.

Table 1 reports the size adjusted power of the proposed method as well as the other methods for testing alternative hypotheses $H_1 : \beta_a = b$, where $b \in \{4, 4.1, \ldots, 5.9, 6\}$. Since the results for other specifications are similar, we only report the results for $K = 3$ and $N = 100$. We find that after size adjustment, the power of the four methods are similar. However, we note that our procedure, RLR, has slightly better power than DLR for most alternative hypotheses considered in Table 1.

Finally, we explore the performance of the tests on the median outcome $\beta_m = \text{median}(Y_1)$. Here the moment condition for $\beta_m$ is given by $\mathbb{E}[\mathbb{I}\{Y <$

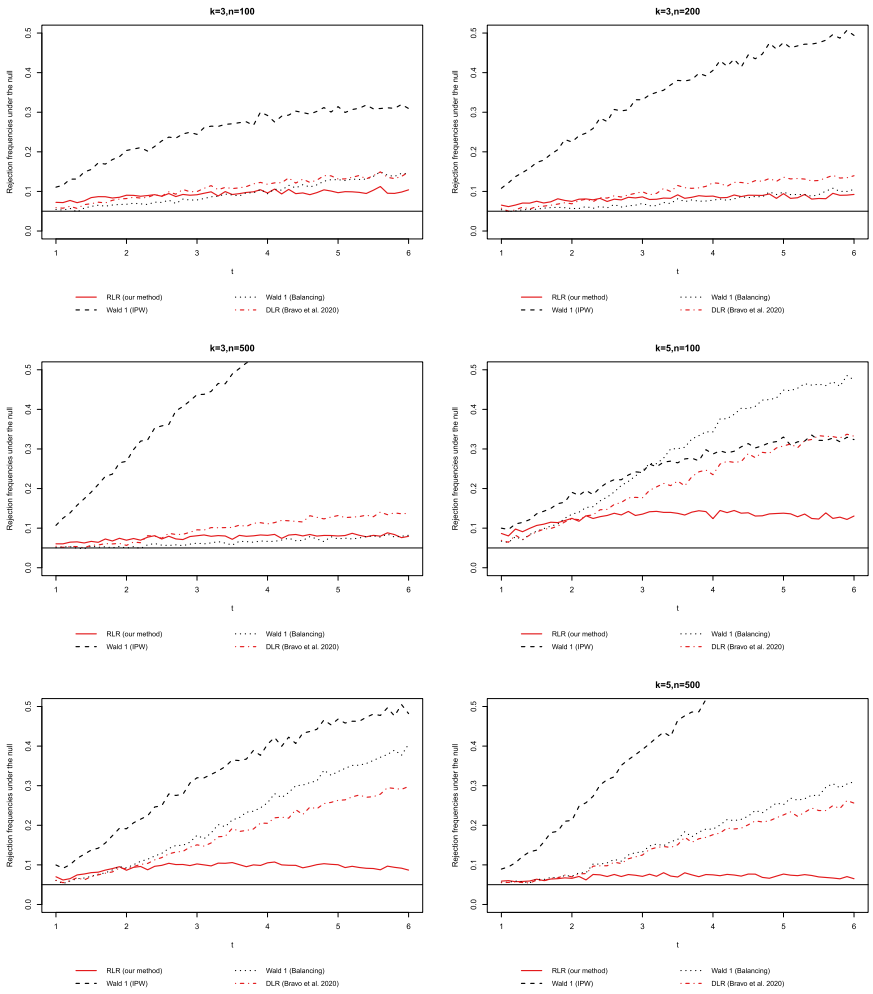


FIG 2. *Rejection frequencies under the null for inference on $\beta_a$, DGP2.*

$\beta_m\} - 0.5] = 0$. Thus the parameter of interest $\beta_m$ enters the moment condition in a nonlinear way. The data generating process follows that of DGP1. We first compare the rejection frequencies of the tests under the null $H_0 : \beta_m = 5$ and with nominal size 0.05. For Wald statistic, the point estimator is the IPW-GMM estimator, which is similar to the one proposed in [11]. Note in this case the Wald statistic is difficult to obtain due to the complicated nature of the variance estimate for quantile estimators. We instead use a bootstrapped variance estimator with 500 bootstrap replications. In this simulation, the number of Monte Carlo replications is set as 2500.

Figure 3 plots the rejection frequencies under the null. Notice that the Wald statistic does not have the correct coverage in all scenarios and under-reject in

TABLE 1
*Size adjusted power under alternatives for inference on $\beta_a$, DGP1.*

| | DGP1, $t = 2$, $K = 3$, $N = 100$ | | | |
|---|---|---|---|---|
| $H_1$ | RLR (our method) | Wald 1 (IPW) | Wald 2 (Balancing) | DLR (Bravo et al.) |
| 4.0 | 0.9988 | 0.997 | 0.9984 | 0.9932 |
| 4.1 | 0.9928 | 0.9884 | 0.9936 | 0.9836 |
| 4.2 | 0.9728 | 0.9664 | 0.9724 | 0.9612 |
| 4.3 | 0.918 | 0.9124 | 0.9196 | 0.9032 |
| 4.4 | 0.814 | 0.8036 | 0.8172 | 0.7936 |
| 4.5 | 0.6824 | 0.6772 | 0.688 | 0.6576 |
| 4.6 | 0.4836 | 0.4816 | 0.4916 | 0.4704 |
| 4.7 | 0.312 | 0.31 | 0.3188 | 0.3076 |
| 4.8 | 0.158 | 0.1664 | 0.164 | 0.1572 |
| 4.9 | 0.0828 | 0.0876 | 0.0796 | 0.0812 |
| 5.0 | 0.05 | 0.05 | 0.05 | 0.05 |
| 5.1 | 0.0848 | 0.0828 | 0.0748 | 0.0828 |
| 5.2 | 0.1636 | 0.1516 | 0.1524 | 0.1644 |
| 5.3 | 0.3228 | 0.2976 | 0.3032 | 0.3176 |
| 5.4 | 0.5208 | 0.4916 | 0.5028 | 0.5116 |
| 5.5 | 0.7216 | 0.6884 | 0.7036 | 0.7188 |
| 5.6 | 0.8712 | 0.8544 | 0.8552 | 0.8628 |
| 5.7 | 0.9448 | 0.9256 | 0.9352 | 0.9352 |
| 5.8 | 0.9828 | 0.9752 | 0.98 | 0.9796 |
| 5.9 | 0.994 | 0.9912 | 0.9924 | 0.9924 |
| 6.0 | 0.9984 | 0.998 | 0.9984 | 0.998 |

most cases. We speculate that this under-rejection is due to the optimization error involved in the bootstrap repetitions when it comes to estimating the nonlinear parameter $\beta_m$. On the other hand, the both likelihood ratio statistics perform much better since their optimization is carried out only under the null. However, DLR perform slightly better than our RLR in most cases. Since the performance of the Wald statistic is dominated by the LR tests under the null, we compares the size-adjusted power of the two competing LR tests (RLR and DLR) under $H_1 : \beta_m = b$ for $b \in \{4, 4.1, \ldots, 5.9, 6\}$ in Table 2. Since the results are similar, we only report the results for $K = 3$ and $N = 100$ with 5000 Monte Carlo replications. The power properties are overall comparable for RLR and DLR.

## 6. Real data example

We illustrate our inference procedure by applying it on data taken from the influential study of [8]. 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 [33] and [24] 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

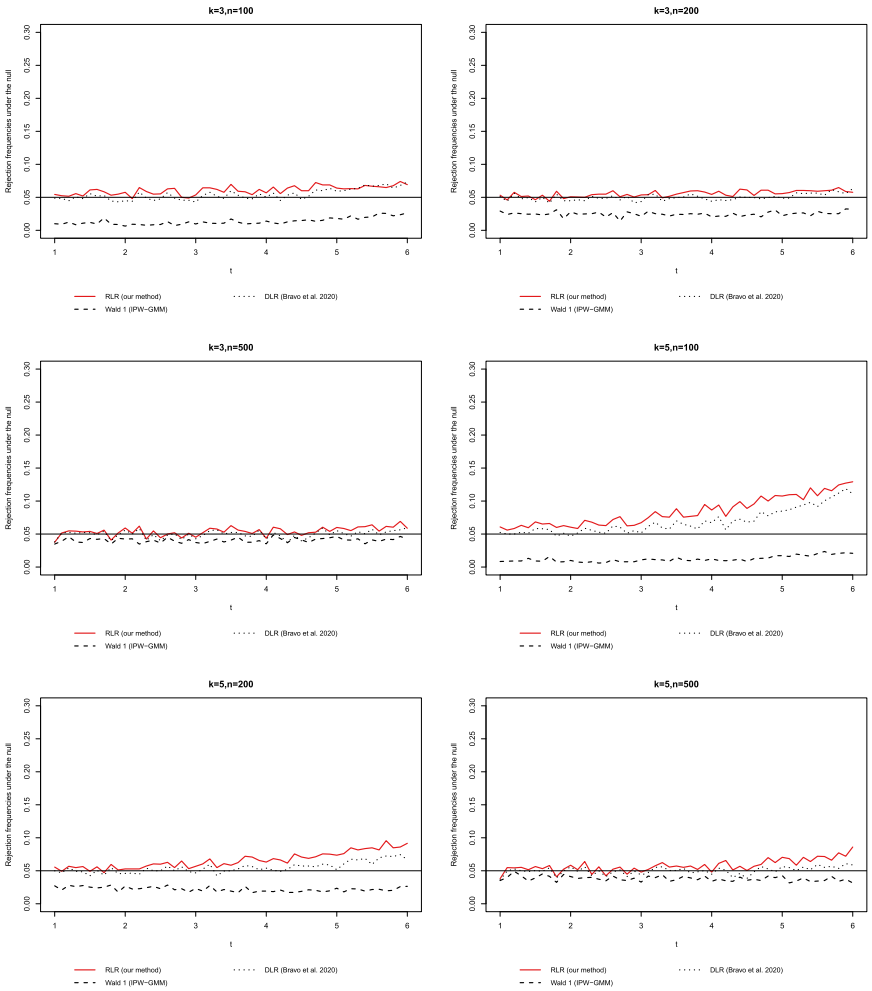


FIG 3. *Rejection frequencies under the null for inference on $\beta_m$, DGP1.*

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:

$$(\text{burger king}, \text{kfc}, \text{roys}, \text{wendys}).$$

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

TABLE 2
*Size adjusted power under alternatives for inference on $\beta_m$, DGP1.*

| | DGP1, $t = 2$, $K = 3$, $N = 100$ | |
|---|---|---|
| $H_1$ | RLR (our method) | DLR (Bravo et al.) |
| 4.0 | 0.9244 | 0.8984 |
| 4.1 | 0.8676 | 0.8284 |
| 4.2 | 0.78 | 0.7508 |
| 4.3 | 0.6836 | 0.6504 |
| 4.4 | 0.5476 | 0.5164 |
| 4.5 | 0.4164 | 0.3932 |
| 4.6 | 0.288 | 0.2748 |
| 4.7 | 0.1828 | 0.1772 |
| 4.8 | 0.112 | 0.1068 |
| 4.9 | 0.0568 | 0.06 |
| 5.0 | 0.05 | 0.05 |
| 5.1 | 0.0576 | 0.0624 |
| 5.2 | 0.0948 | 0.1036 |
| 5.3 | 0.188 | 0.2008 |
| 5.4 | 0.2944 | 0.2984 |
| 5.5 | 0.4408 | 0.448 |
| 5.6 | 0.6112 | 0.6244 |
| 5.7 | 0.7196 | 0.724 |
| 5.8 | 0.8376 | 0.8344 |
| 5.9 | 0.908 | 0.9032 |
| 6.0 | 0.9564 | 0.954 |

TABLE 3
*Confidence regions for $\beta_0$ 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 | $[-0.766,\ 2.445]$ | $[-1.073,\ 2.753]$ | $[-0.590,\ 2.335]$ | $[-0.870,\ 2.615]$ |

of interest, $\beta_0$, is the average treatment effect on employment levels due to the minimum wage hike.

To provide inference on $\beta_0$, we consider two empirical balancing schemes: one where we only balance a single covariate, empft, i.e., $q^K(X) = (1, \text{empft})$, 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 [24] 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 $\beta_0$ based on our inferential procedure, along with the Wald confidence regions. We also report the estimates, $\hat{\beta}$, of $\beta_0$ under both $K = 2$ and 7. Both values are very close to the estimate of $\hat{\beta}_m = 0.84$ obtained by [24] using matching.

## 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

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

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

$$\ell(\theta_0) = \left(\frac{1}{\sqrt{N}}\sum_{i=1}^N g_i^K\right)\left(\frac{1}{N}\sum_{i=1}^N g_i^K g_i^{K\prime}\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N g_i^K\right) + R_N, \tag{17}$$

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 [23] (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 (17) 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

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

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 (12) guarantees $V = V_0 - V_1$, the conclusion follows.

On the other hand, if (12) 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 (i)–(v), 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 (ii). Also, since $\mathbb{E}[\alpha_{0i}h_i + m_i(0)]^2 < \infty$ by assumption, [27] (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 (iv). Therefore, by Assumption (iv),

$$\begin{aligned}\max_{1\le i\le N}|D_{2i}| &\le \delta_{\alpha,N}\left(\max_{1\le i\le N}|Q_{Ki}| + \max_{1\le i\le N}|h_i|\right)\\ &= \delta_{\alpha,N}(\zeta_K + O_p(n^{1/\kappa})) = o_p(1). \qquad (18)\end{aligned}$$

□

*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

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

Similar to Lemma 4 (ii), we can show

$$\mathbb{E}[(\alpha_0 h + m(0))\varepsilon^{Q_K}]'\left(\mathbb{E}[\varepsilon^{Q_K}\varepsilon^{Q_K\prime}]\right)^{-1}\mathbb{E}[(\alpha_0 h + m(0))\varepsilon^{Q_K}]$$

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

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

$$\mathbb{P}\left\{\max_{a\in\mathbb{S}^K} \hat{H}_n(a) < 0\right\} \to 1, \tag{19}$$

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 $|H_n(a) - H_n(b)| \le |a - b| \max_{1\le i\le N} |D_{1i}| \le |a - b| o_p(\sqrt{N})$ 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 $\delta_N$, where $C_{K,N}\delta_N^{K-1} \ge 2\pi^{K/2}/\Gamma(K/2)$ 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

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

For (19), it is sufficient to show show that for each $\epsilon > 0$,

$$\mathbb{P}\left\{\max_{1\le i\le N} |D_{2i}| \le \frac{\epsilon}{2}\right\} \to 1, \tag{20}$$

$$\mathbb{P}\left\{\delta_N \max_{1\le i\le N} |D_{1i}| \le \frac{\epsilon}{2}\right\} \to 1, \tag{21}$$

$$\mathbb{P}\left\{\max_{a\in\mathbb{U}_{N,K}} H_n(a) < -\epsilon\right\} \to 1. \tag{22}$$

The convergence in (20) is guaranteed by $\max_{1\le i\le N} |D_{2i}| = o_p(1)$. The convergence in (21) is guaranteed by setting $\delta_N = \frac{\epsilon}{C\sqrt{N}}$ for some $C > 0$ since $\max_{1\le i\le N} |D_{1i}| = o_p(\sqrt{N})$. By [23] (2009, Lemma 4.2), the convergence in (22) is guaranteed if $\frac{K \log N}{N} \to 0$ and $\mathbb{E}[D_{1i} D_{1i}']$ has eigenvalues bounded away from zero, which follow by Assumption (ii) 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

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

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 $\gamma_0$), the law of iterated expectations yields

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

Also as $\sup_{x\in\mathcal{X}}\mathbb{E}[(h_i - \gamma_{0i})^2|X_i = x] \lesssim 1$, the same argument in [31] (Lemma S4) implies

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

Thus, Markov's inequality and Assumption (iii) 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

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

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 (iii). So, Markov's inequality implies $E_{22} \xrightarrow{p} 0$. Finally, Assumptions (iii)–(iv) 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 (i)–(v), 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 [5] (Lemma 6.2) for the first statement, and [34] (Theorem 5.1.1) for the second statement. □

*Proof of (iii).* By the triangle inequality, stated assumptions and Lemma 3 (i)–(ii), we have

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

□

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

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

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 (i)–(v), 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

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

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

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,

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

The weak law of large numbers implies

$$\frac{1}{N}\sum_{i=1}^{N}|h_i\{\alpha_{0i}h_i + m_i(0)\}| \xrightarrow{p} \mathbb{E}|h_i\{\alpha_{0i}h_i + m_i(0)\}|$$
$$\le \sqrt{\mathbb{E}[h_i^2]}\sqrt{\mathbb{E}[\{\alpha_{0i}h_i + m_i(0)\}^2]},$$

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 (iii). □

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

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

where

$$A_N = \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{Ki}\hat{\varepsilon}_{Ki}'\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{hi}\hat{\varepsilon}_{Ki}\right),$$

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

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

Thus, it is sufficient for the conclusion to show that

$$A_N \xrightarrow{p} \mathbb{E}[\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}^2], \tag{23}$$

$$B_N \xrightarrow{p} \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}], \tag{24}$$

$$C_N \xrightarrow{p} 0. \tag{25}$$

*Proof of (23).* Let $e_{hi} = h_i - \gamma_{0i}$. Observe that $A_N$ can be decomposed as $A_N = A_{N1} + A_{N2} + 2A_{N3}$, where

$$A_{N1} = \left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right),$$

$$A_{N2} = \left(\frac{1}{N}\sum_{i=1}^N \hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^N \hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right),$$

$$A_{N3} = \left(\frac{1}{N}\sum_{i=1}^N \hat{\alpha}_i e_{hi}\hat{\varepsilon}_{Ki}\right)' \left[\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right]^{-1} \left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{\gamma_0 i}\hat{\varepsilon}_{Ki}\right),$$

$$\hat{\varepsilon}_{\gamma_0 i} = \hat{\alpha}_i\gamma_{0i} - \tilde{m}_i(\gamma_0).$$

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 $\hat{\mathcal{E}}_{\gamma_0 i}^2$ 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

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

Thus, Assumptions (iii) and (ii) and the weak law of large numbers imply

$$\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{\gamma_0 i}^2 \xrightarrow{p} \mathbb{E}[\varepsilon_{\gamma_0 i}^2]. \tag{26}$$

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

$$\frac{1}{N}\sum_{i=1}^N \hat{\mathcal{E}}_{\gamma_0 i}^2 \le \frac{1}{N}\sum_{i=1}^N \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki}) - \tilde{m}_i(\gamma_0 - \beta_K' Q_K)\}^2$$

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

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

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

For $A_{N12}$, note $A_{N12}\xrightarrow{p}0$ by Assumption (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

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

By definition of empirical projection

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

where the last inequality follows by law of large numbers under Assumption (iv) and $\sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\xrightarrow{p}0$ by Assumption (iii). Next, we show

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{P}_i^2(\alpha_0 e_h|\hat{\varepsilon}_K)=o_p(1)$$

as well. Since

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

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 (iii) and law of large numbers by Assumption (iv). 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 (26).

*Proof of (24).* 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

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

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

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

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{E}}_i^2 &\leq \frac{1}{N}\sum_{i=1}^{N} \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki}) + \tilde{m}_i(\beta_K' Q_K - \gamma_0)\}^2 \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N} \{\hat{\alpha}_i(\gamma_{0i} - \beta_K' Q_{Ki})\}^2 \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \tilde{m}_i(\beta_K' Q_K - \gamma_0)^2. \qquad (27)
\end{aligned}$$

For the first term of (27), 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 Assumption (iii). For the second term of (27), the weak law of large numbers and Assumption (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 $\frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{E}}_i^2 \xrightarrow{p} 0$. 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 $B_{N1} \xrightarrow{p} \mathbb{E}[m_i(\gamma_0)\{\alpha_{0i}\gamma_{0i} + m_i(0) - m_i(\gamma_0)\}]$.

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

*Proof of (25).* 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

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

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

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

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

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

$$\times \left( \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0)\varepsilon_{Ki} - \mathbb{E}[m_i(\gamma_0)\varepsilon_{Ki}] \right),$$

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

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

Note that

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

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

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

By these results, Lemma 5, and Assumption (iii), we have

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

and the conclusion follows. □

**Lemma 5.** *Under Assumptions (i)–(v), 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

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

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 the Theorem hold true except display (12). 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 (25). 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),

$$|C_{N2}| \leq \left|\frac{1}{N}\sum_{i=1}^N m_i(\gamma_0)(\hat{\alpha}_i - \alpha_{0i})Q_{Ki}\right| \left|\left(\frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_{Ki}\hat{\varepsilon}'_{Ki}\right)^{-1}\right|$$

$$\times \left| \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0) \varepsilon_{Ki} \right|$$
$$= O_p(\zeta_K \delta_{\alpha,N} \zeta_{\varepsilon,K}) = o_p(1).$$

For $C_{N3}$, notice

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

where

$$\tilde{C}_{N31} = \left( \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0) \varepsilon_{Ki} \right)' \left\{ \left[ \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{Ki} \hat{\varepsilon}'_{Ki} \right]^{-1} - \left[ \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{Ki} \varepsilon'_{Ki} \right]^{-1} \right\}$$
$$\times \left( \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0) \varepsilon_{Ki} \right),$$
$$\tilde{C}_{N32} = \left( \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0) \varepsilon_{Ki} \right)' \left[ \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{Ki} \varepsilon'_{Ki} \right]^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} m_i(\gamma_0) \varepsilon_{Ki} \right).$$

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

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

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

$$\left| \frac{1}{N} \sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_{0i})^2 Q_{Ki} Q'_{Ki} \right|$$
$$= \max_{a \in \mathbb{S}^{K-1}} \frac{1}{N} \sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_{0i})^2 (a' Q_{Ki})^2$$

$$\leq \left(\sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\right)^2 \max_{a\in\mathbb{S}^{K-1}} \frac{1}{N}\sum_{i=1}^{N}(a'Q_{Ki})^2$$
$$= \left(\sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\right)^2 \left|\frac{1}{N}\sum_{i=1}^{N} Q_{Ki}Q_{Ki}'\right| = O_p(\delta_{\alpha,N}^2).$$

Also,

$$\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i-\alpha_{0i})\varepsilon_{Ki}Q_{Ki}'\right| \leq \sup_{x\in\mathcal{X}}|\hat{\alpha}(x)-\alpha_0(x)|\sup_{x\in\mathcal{X}}|\varepsilon_K(x)|\sup_{x\in\mathcal{X}}|Q_K(x)|$$
$$= O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}).$$

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,

$$|\Sigma_N-\hat{\Sigma}_N| = O_p(\delta_{\alpha,N}^2+\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}) = O_p(\zeta_K\delta_{\alpha,N}\zeta_{\varepsilon,K}).$$

It follows

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

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. □

## Funding

This research was conducted with support from the Cornell University Center for Advanced Computing.

## References

[1] Abadie, A. (2005) Semiparametric difference-in-differences, *Review of Economic Studies*, 72, 1–19. MR2116973

[2] Abadie, A. and G. W. Imbens (2016) Matching on the estimated propensity score, *Econometrica*, 84, 781–807. MR3481379

[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. MR3849336

[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. MR3343791

[6] Bravo, F., Escanciano, J. C. and I. Van Keilegom (2020) Two-step semiparametric empirical likelihood inference, *Annals of Statistics*, 48, 1–26. MR4065150

[7] Brown, B. M., and Chen, S. X. (1998). Combined and least squares empirical likelihood. *Annals of the Institute of Statistical Mathematics*, 50, 697–714. MR1671990

[8] 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, 772–793.

[9] 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. MR1125715

[10] 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. MR3506798

[11] Chen, X., Hong, H. and A. Tarozzi (2008) Semiparametric efficiency in GMM models with auxiliary data, *Annals of Statistics*, 36, 808–843. MR2396816

[12] Cressie, N. and R. C. Read (1984) Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society B*, 46, 440–464. MR0790631

[13] Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and J. M. Robins (2016) Locally robust semiparametric estimation, *arXiv preprint arXiv:1608.00033*. MR4467437

[14] Chernozhukov, V., Newey, W. K. and R. Singh (2020) Automatic debiased machine learning of causal and structural effects, *arXiv preprint arXiv: 1809.05224*. MR4436042

[15] 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.

[16] Graham, B. S. (2011) Efficiency bounds for missing data models with semiparametric restrictions, *Econometrica*, 79, 437–452. MR2809376

[17] 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. MR2986390

[18] 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. MR3475879

[19] Hainmueller, J. (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies,

*Political Analysis*, 20, 25–46.
[20] Hausman, J. A. and W. K. Newey (2017) Nonparametric welfare analysis, *Annual Review of Economics*, 9, 521–546.
[21] Hirano, K., Imbens, G. W. and G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161–1189. MR1995826
[22] Hirshberg, D. A. and S. Wager (2018) Augmented minimax linear estimation, Working paper. MR4352528
[23] Hjort, N. L., I. W. McKeague and I. Van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079–1111. MR2509068
[24] Imbens, G. W. and D. B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press. MR3309951
[25] Newey, W. K. and J. M. Robins (2018) Cross-fitting and fast remainder rates for semiparametric estimation, Working paper.
[26] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1), 219–255. MR2031017
[27] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC.
[28] Qin, J. (2017) *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer. MR3675467
[29] 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. MR2301502
[30] 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. MR2750574
[31] Qiu, C. (2020) Near optimal estimation of average regression functionals, Working paper.
[32] Qiu, C. and T. Otsu (2022) Information theoretic approach to high dimensional multiplicative models: stochastic discount factor and treatment effect, *Quantitative Economics*, 13, 63–94. MR4399603
[33] Rosenbaum, P. R. (2002) *Observational Studies*, Springer, New York. MR1899138
[34] Tropp, J. A. (2015) An introduction to matrix concentration inequalities, *Foundations and Trends in Machine Learning*, 8, 1–230. MR4408832
[35] Zubizarreta, J. R. (2015) Stable weights that balance covariates for estimation with incomplete outcome data, *Journal of the American Statistical Association*, 110, 910–922. MR3420672