Electronic Journal of Statistics
Vol. 19 (2025) 87–130
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2338

Pseudo-empirical likelihood methods for causal inference

Jingyue Huang

Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania, Philadelphia, PA, United States,
e-mail: jingyue.huang@pennmedicine.upenn.edu

Changbao Wu

Department of Statistics and Actuarial Science
University of Waterloo, Waterloo, ON N2L 3G1, Canada,
e-mail: cbwu@uwaterloo.ca

and

Leilei Zeng

Department of Statistics and Actuarial Science
University of Waterloo, Waterloo, ON N2L 3G1, Canada,
e-mail: lzeng@uwaterloo.ca

Abstract: Causal inference remains a key area of research for assessing treatment effects across various real-world settings. This paper introduces two inferential procedures for the average treatment effect (ATE) using a two-sample pseudo-empirical likelihood (PEL) approach. The first procedure uses estimated propensity scores for the formulation of the PEL function, yielding a maximum PEL estimator equivalent to the inverse probability weighted estimator. Our focus in this scenario is on the PEL ratio statistic and establishing its theoretical properties. The second procedure incorporates outcome regression models for doubly robust PEL inference. We establish the asymptotic distribution of the PEL ratio statistic and also propose a bootstrap method for constructing confidence intervals to bypass a scaling constant involved in the asymptotic distribution of the PEL ratio statistic that is very difficult to calculate. Simulation studies and an real data application illustrate the performance of our proposed methods with comparisons to existing ones.

MSC2020 subject classifications: Primary 62G05, 62G10; Secondary 62G20.
Keywords and phrases: Average treatment effect, propensity scores, pseudo-empirical likelihood ratio statistic, model calibration, double robustness, bootstrap method.

Received October 2023.

Contents

1. Introduction

It is straightforward to establish causal results when treatment assignments are completely randomized. It is known to the research community, however, that causal inferences with observational studies are a challenging task due to the non-randomized treatment assignments that are influenced by the covariates associated with outcomes. In such cases, the difference in outcomes between the treatment and the control groups is due to not only the different treatment exposures but also the different characteristics of units in the two groups as reflected by the covariates' imbalances. These covariates, often referred to as "confounders", make the naive mean causal effect estimators invalid.

The propensity score, defined as the probability of being in the treatment group given the covariates, plays an essential role in causal inference for balancing covariates and valid statistical procedures, as highlighted in the seminal paper of Rosenbaum and Rubin (1983). Many statistical methods are proposed based on propensity scores to obtain consistent estimators for the average treatment effect (ATE) under the so-called strongly ignorable assumption, such as matching (Rosenbaum and Rubin, 1985; Abadie and Imbens, 2006), post-stratification (Rosenbaum and Rubin, 1984; Rosenbaum, 1987), and weighting (Robins et al., 2000; Hirano et al., 2003).

A practical issue of the aforementioned methods is that the propensity scores are usually unknown and must be estimated. Misspecification of the propensity score model leads to invalid inverse probability weighted (IPW) estimators. One way to mitigate this issue is to construct estimators that combine inverse probability weighting with outcome regression modelling, known as the augmented inverse probability weighted (AIPW) estimators (Robins et al., 1994). The AIPW estimators are doubly robust in the sense that the consistency of such estimators only requires the correct specification of one of the two sets of models: the set of propensity score model and the set of outcome regression models (Scharfstein et al., 1999). The asymptotic variances of the AIPW estimators can be obtained based on their influence functions (Tsiatis, 2006).

Propensity score-based weighting methods can balance covariates in a large sample but may fail to do so in finite samples. Additionally, applied researchers often engage in a cyclical process of propensity score modelling and covariate balance checking until they achieve satisfactory results, which has been criticized as the "propensity score tautology" (Imai et al., 2008). To address these issues, Hainmueller (2012) introduced entropy balancing to estimate the average treatment effect of the treated. Entropy balancing exactly balances the sample moments of the covariates between the treatment and the control groups by maximizing the entropy of the weights subject to some calibration constraints that ensure the equivalence of the moments from the two groups. Zhao and Percival (2017) demonstrated that entropy balancing is doubly robust with respect to linear outcome regression and logistic propensity score regression, though there is no modelling in its original form. However, this approach may require a considerable number of constraints.

In this paper, we propose two procedures for the estimation and inference of the ATE through a two-sample pseudo-empirical likelihood (PEL) approach. Two point estimators are constructed. One corresponds to the IPW estimator through the explicit use of estimated propensity scores in forming the PEL function, and the other achieves double robustness through the inclusion of additional model-calibration constraints based on the outcome regression models. Moreover, for each procedure, we establish the asymptotic properties of the PEL ratio statistic, enabling the construction of confidence intervals and tests of hypotheses. The methods we developed have attractive features shared by general empirical likelihood based approaches, which include (1) range-respecting and transformation-invariant properties of the PEL ratio confidence intervals; and (2) problem formulations through a constrained optimization procedure, which enables incorporations of suitable auxiliary information through additional constraints for more efficient or robust estimators (Hall and Scala, 1990; Owen, 2001). These features of our proposed methods are demonstrated through simulation studies reported in Section 5 and a real data analysis in Section 6, with additional remarks provided in Section 7.

The remainder of this paper is structured as follows. Section 2 introduces fundamental concepts of causal inference, alongside commonly used causal inference methods. In Section 3, we present the formulation of the two-sample PEL approach to causal inference and develop the point estimator and the associated PEL ratio confidence intervals for the ATE. In Section 4, we examine doubly robust inference under the proposed PEL framework through the inclusion of the model-calibration constraints. Results from simulation studies on finite sample performances of proposed methods with comparisons to existing ones are reported in Section 5. We also conducted a real data analysis to investigate the ATE of maternal smoking during pregnancy on birth weights in Section 6. Some additional remarks are given in Section 7. Regularity conditions, proofs and technical details are presented in Section 8.

2. Causal inference

2.1. Basic setting and propensity scores

We follow the commonly used setup of causal inference with the potential outcome framework. Let T be a binary variable denoting the treatment assignment, with $T = 1$ indicating treatment and $T = 0$ for control. The potential outcome variable under the treatment or the control is represented by Y_1 or Y_0, respectively. The parameter of interest is the average treatment effect (ATE), $\theta = \mu_1 - \mu_0$, where $\mu_1 = \mathrm{E}(Y_1)$ and $\mu_0 = \mathrm{E}(Y_0)$ are the expectations of the potential outcome variables. As each individual can only be assigned to one of the two groups, only one of the potential outcomes can be observed for each individual, and the observed outcome is denoted as $Y = TY_1 + (1 - T)Y_0$.

Let T_j, Y_{1j} and Y_{0j} be respectively the values of T, Y_1 and Y_0 associated with subject j. Let $Y_j = T_jY_{1j} + (1 - T_j)Y_{0j}$. Consider a random sample $\mathcal{S}$ of size n from an infinite target population. The sample data are represented by $\{(\boldsymbol{x}_j, Y_j, T_j), j \in \mathcal{S}\}$, where $\boldsymbol{x}_j$ denotes the value of the vector of auxiliary variables for subject j, $j = 1, \ldots, n$. We define $\mathcal{S}_1 = \{j \mid T_j = 1 \text{ and } j \in \mathcal{S}\}$ and $\mathcal{S}_0 = \{j \mid T_j = 0 \text{ and } j \in \mathcal{S}\}$ as the sub-samples of subjects receiving treatment and control, respectively. The available sample data can be partitioned into two subsets: $\{(\boldsymbol{x}_j, Y_{1j}, T_j = 1), j \in \mathcal{S}_1\}$ and $\{(\boldsymbol{x}_j, Y_{0j}, T_j = 0), j \in \mathcal{S}_0\}$. Denote the sizes of $\mathcal{S}_1$ and $\mathcal{S}_0$ respectively by n_1 and n_0. We have $n = n_1 + n_0$.

Balancing the distributions of confounders, covariates that are simultaneously associated with the treatment and potential response variables, in the two treatment groups is essential for investigating causal effects in observational studies. The propensity scores are a widely used tool for this purpose. The propensity score (PS) is defined as the conditional probability of receiving the treatment given the covariates, denoted as $P(T = 1 \mid \boldsymbol{x})$. The properties and the essential role of propensity scores in causal inference are elaborated in the seminal paper by Rosenbaum and Rubin (1983). However, it is important to note that certain critical assumptions must be met in order to obtain consistent estimators of causal effects through propensity score adjustments. Two key assumptions are given below.

A1. *Strongly Ignorable Treatment Assignment (SITA).* The treatment indicator (T) and the potential response variables (Y_1, Y_0) are independent given the set of covariates ($\boldsymbol{x}$).

A2. *Positivity Assumption.* The propensity scores are strictly in the range of $(0, 1)$, i.e., $0 < P(T = 1 \mid \boldsymbol{x}) < 1$ for all possible values of $\boldsymbol{x}$.

The SITA assumption **A1** implies that there are no unmeasured confounders and treatment assignments depend only on the observed covariates. The positivity assumption **A2** guarantees that there is no restriction preventing individuals from being assigned to either the treatment or the control group.

Propensity scores are typically unknown and require to be estimated in practice. Under a parametric model, we have the propensity score $\tau_j = P(T_j = 1 \mid \boldsymbol{x}_j) = \tau(\boldsymbol{x}_j; \boldsymbol{\alpha})$, where $\tau(\cdot;\cdot)$ has a known form and $\boldsymbol{\alpha}$ is the vector of unknown parameters. The maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ can be obtained by maximizing the likelihood function $L(\boldsymbol{\alpha}) = \prod_{j=1}^n \{\tau(\boldsymbol{x}_j; \boldsymbol{\alpha})\}^{T_j}\{1 - \tau(\boldsymbol{x}_j; \boldsymbol{\alpha})\}^{1-T_j}$ using the observed dataset $\{(\boldsymbol{x}_j, T_j), j \in \mathcal{S}\}$, which yields estimated propensity scores $\hat{\tau}_j = \tau(\boldsymbol{x}_j; \hat{\boldsymbol{\alpha}})$, $j = 1, \ldots, n$. In practice, logistic regression models are commonly employed for the propensity scores, leading to the following closed-form expressions for the estimated propensity scores,

$$\hat{\tau}_j = \tau(\boldsymbol{x}_j; \hat{\boldsymbol{\alpha}}) = \text{expit}\left(\tilde{\boldsymbol{x}}_j^\top \hat{\boldsymbol{\alpha}}\right) = \frac{\exp\left(\tilde{\boldsymbol{x}}_j^\top \hat{\boldsymbol{\alpha}}\right)}{1 + \exp\left(\tilde{\boldsymbol{x}}_j^\top \hat{\boldsymbol{\alpha}}\right)},$$

where $\tilde{\boldsymbol{x}}_j = (1, \boldsymbol{x}_j^\top)^\top$. The discussions in the rest of this paper assume a logistic regression model for propensity scores.

2.2. Inverse probability weighted estimators

The method of inverse probability weighting was originally introduced by Horvitz and Thompson (1952) for estimating a finite population total using a probability survey sample, where the weighting is done through the inverses of the known sample inclusion probabilities. The resulting weighted estimator is called the Horvitz-Thompson (HT) estimator, which is one of the backbones for design-based inference in survey sampling. The concept was successfully adapted to causal inference and missing data analysis based on estimated propensity scores.

The inverse probability weighted (IPW) estimators, also referred to as the propensity score-adjusted estimators, play a crucial role in causal inference. The IPW estimators of μ_1 and μ_0 are given by

$$\hat{\mu}_{1\text{IPW1}} = \frac{1}{n}\sum_{j \in \mathcal{S}} \frac{T_j Y_j}{\hat{\tau}_j} = \frac{1}{n}\sum_{j \in \mathcal{S}_1} \frac{Y_{1j}}{\hat{\tau}_j} \quad \text{and} \quad \hat{\mu}_{0\text{IPW1}} = \frac{1}{n}\sum_{j \in \mathcal{S}} \frac{(1 - T_j)\, Y_j}{1 - \hat{\tau}_j} = \frac{1}{n}\sum_{j \in \mathcal{S}_0} \frac{Y_{0j}}{1 - \hat{\tau}_j}.$$

The corresponding IPW estimator of the ATE is defined as $\hat{\theta}_{\text{IPW1}} = \hat{\mu}_{1\text{IPW1}} - \hat{\mu}_{0\text{IPW1}}$.

Note that $\mathrm{E}(\sum_{l\in\mathcal{S}_1}\tau_l^{-1}) = \mathrm{E}(\sum_{l\in\mathcal{S}} T_l\tau_l^{-1}) = n$. Replacing n by its consistent estimator $\sum_{l\in\mathcal{S}_1}\hat{\tau}_l^{-1}$ in $\hat{\mu}_{1\text{IPW1}}$ yields the Hájek-type (Hájek, 1971) IPW estimator for μ_1,

$$\hat{\mu}_{1\text{IPW2}} = \sum_{j\in\mathcal{S}_1}\frac{\hat{\tau}_j^{-1}}{\sum_{l\in\mathcal{S}_1}\hat{\tau}_l^{-1}}Y_{1j} = \sum_{j\in\mathcal{S}_1}\tilde{a}_{1j}Y_{1j}\,, \tag{2.1}$$

where $\tilde{a}_{1j} = \hat{\tau}_j^{-1}/\sum_{l\in\mathcal{S}_1}\hat{\tau}_l^{-1}$ and $\sum_{j\in\mathcal{S}_1}\tilde{a}_{1j} = 1$. Similarly, the Hájek-type IPW estimator for μ_0 takes the form

$$\hat{\mu}_{0\text{IPW2}} = \sum_{j\in\mathcal{S}_0}\frac{(1-\hat{\tau}_j)^{-1}}{\sum_{l\in\mathcal{S}_0}(1-\hat{\tau}_l)^{-1}}Y_{0j} = \sum_{j\in\mathcal{S}_0}\tilde{a}_{0j}Y_{0j}\,, \tag{2.2}$$

where $\tilde{a}_{0j} = (1-\hat{\tau}_j)^{-1}/\sum_{l\in\mathcal{S}_0}(1-\hat{\tau}_l)^{-1}$. The resulting estimator for the ATE is then given by $\hat{\theta}_{\text{IPW2}} = \hat{\mu}_{1\text{IPW2}} - \hat{\mu}_{0\text{IPW2}}$.

The IPW estimators $\hat{\theta}_{\text{IPW1}}$ and $\hat{\theta}_{\text{IPW2}}$ of the ATE are consistent under the assumed propensity score model and certain regularity conditions. This can be shown by recognizing that they belong to the so-called m-estimators and following the technical arguments presented in Section 3.2 of Tsiatis (2006). Furthermore, the variance of the IPW estimators can be estimated by the sandwich variance estimator. Details are presented in Section 8.2. The Hájek-type IPW estimator $\hat{\theta}_{\text{IPW2}}$ is typically more efficient than the IPW estimator $\hat{\theta}_{\text{IPW1}}$ for finite samples (Särndal et al., 1992).

2.3. Augmented IPW estimators

The IPW estimators can be biased when the propensity score model is misspecified. The biases can sometimes be mitigated through the incorporation of an outcome regression model. Robins et al. (1994) proposed a class of augmented IPW estimators under the two-model framework, which usually exhibits greater efficiency than the IPW estimators when both models are correct. Scharfstein et al. (1999) later found that the consistency of these estimators only requires one of the two models to be correctly specified. Hence, such estimators are also termed as doubly robust.

Under the potential outcome framework, consider two outcome regression models: $\mathrm{E}(Y_1 \mid \boldsymbol{x}) = m_1(\boldsymbol{x};\boldsymbol{\beta}_1)$ and $\mathrm{E}(Y_0 \mid \boldsymbol{x}) = m_0(\boldsymbol{x};\boldsymbol{\beta}_0)$. Given the SITA assumption **A1**, i.e., $T \perp\!\!\!\perp \{Y_1, Y_0\} \mid \boldsymbol{x}$, we have

$$\mathrm{E}(Y_1 \mid \boldsymbol{x}) = \mathrm{E}(Y_1 \mid T=1,\boldsymbol{x}) \quad \text{and} \quad \mathrm{E}(Y_0 \mid \boldsymbol{x}) = \mathrm{E}(Y_0 \mid T=0,\boldsymbol{x})\,.$$

It follows that the model parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ can be estimated by fitting the corresponding model using observed datasets $\{(\boldsymbol{x}_j, Y_{1j}), j\in\mathcal{S}_1\}$ and $\{(\boldsymbol{x}_j, Y_{0j}), j\in\mathcal{S}_0\}$, respectively. Let $\hat{m}_{1j} = m_1(\boldsymbol{x}_j;\hat{\boldsymbol{\beta}}_1)$ and $\hat{m}_{0j} = m_0(\boldsymbol{x}_j;\hat{\boldsymbol{\beta}}_0)$, for all $j\in\mathcal{S}$, where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ are the corresponding estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$.

Following Robins et al. (1994), the augmented IPW estimators for μ_1 and μ_0 are constructed as

$$\hat{\mu}_{1\text{AIPW1}} = \frac{1}{n}\sum_{j\in\mathcal{S}_1}\frac{Y_{1j}-\hat{m}_{1j}}{\hat{\tau}_j}+\frac{1}{n}\sum_{j=1}^{n}\hat{m}_{1j}\,,$$

$$\hat{\mu}_{0\text{AIPW1}} = \frac{1}{n}\sum_{j\in\mathcal{S}_0}\frac{Y_{0j}-\hat{m}_{0j}}{1-\hat{\tau}_j}+\frac{1}{n}\sum_{j=1}^{n}\hat{m}_{0j}\,,$$

and the augmented IPW estimator for the ATE is given by $\hat{\theta}_{\text{AIPW1}} = \hat{\mu}_{1\text{AIPW1}} - \hat{\mu}_{0\text{AIPW1}}$.

Each of the two augmented IPW estimators, $\hat{\mu}_{1\text{AIPW1}}$ and $\hat{\mu}_{0\text{AIPW1}}$, contains an IPW estimator based on the "residuals" $Y_{1j}-\hat{m}_{1j}$ or $Y_{0j}-\hat{m}_{0j}$, and each IPW estimator can be replaced by a Hájek-type IPW estimator, leading to

$$\hat{\mu}_{1\text{AIPW2}} = \sum_{j\in\mathcal{S}_1}\tilde{a}_{1j}\left(Y_{1j}-\hat{m}_{1j}\right)+\bar{\hat{m}}_1 \quad\text{and}\quad \hat{\mu}_{0\text{AIPW2}} = \sum_{j\in\mathcal{S}_0}\tilde{a}_{0j}\left(Y_{0j}-\hat{m}_{0j}\right)+\bar{\hat{m}}_0\,,$$

where $\bar{\hat{m}}_1 = \sum_{j=1}^{n}\hat{m}_{1j}/n$, $\bar{\hat{m}}_0 = \sum_{j=1}^{n}\hat{m}_{0j}/n$, $\tilde{a}_{1j}$ and $\tilde{a}_{0j}$ are defined in (2.1) and (2.2). The corresponding estimator for the ATE is given by $\hat{\theta}_{\text{AIPW2}} = \hat{\mu}_{1\text{AIPW2}} - \hat{\mu}_{0\text{AIPW2}}$.

Remark 1. *The estimators $\hat{\theta}_{AIPW1}$ and $\hat{\theta}_{AIPW2}$ are doubly robust in the sense that they are consistent estimators of θ if one of the two sets of models, the propensity score model $\tau(\boldsymbol{x};\boldsymbol{\alpha})$ or the set of outcome regression models $m_1(\boldsymbol{x};\boldsymbol{\beta}_1)$ and $m_0(\boldsymbol{x};\boldsymbol{\beta}_0)$, is correctly specified. This is clearly an attractive property that offers some protection against misspecification of one set of models.*

Doubly robust variance estimation based on a doubly robust point estimator is an active research topic pursued by several authors. In the context of missing data, Cao et al. (2009) briefly mentioned that the variance of some augmented IPW estimators for the population mean can be estimated by the usual empirical sandwich technique (Stefanski and Boos, 2002) if the point estimators are obtained by solving a set of m-estimating equations jointly. The resulting variance estimator is robust against misspecifications of one or both of the propensity score and outcome regression models. In the current setting, the sandwich variance estimators for the two augmented IPW estimators, $\hat{\theta}_{\text{AIPW1}}$ and $\hat{\theta}_{\text{AIPW2}}$, are consistent for the true variance even if one or both of the two sets of models are incorrectly specified. The sandwich variance estimators for the augmented IPW estimators are constructed in the same manner as those for the IPW estimators, where the additional estimating equations for the outcome regression models are included. We formally state this result in the following proposition. Proof of the result is presented in Section 8.3, where we also present the detailed formulation of the sandwich variance estimator for the augmented IPW estimator.

Proposition 1. *Under certain regularity conditions, the variance of the augmented IPW estimator $\hat{\theta}_{AIPW1}$ or $\hat{\theta}_{AIPW2}$ can be estimated by the sandwich variance estimator regardless of the correctness of the propensity score model or the outcome regression models.*

The required regularity conditions for Proposition 1 as well as other major results presented in the paper are given in Section 8.1. When both sets of models are misspecified, the point estimators become invalid, and the corresponding variance estimators are not practically useful. We mainly focus on the doubly robust property of the sandwich variance estimators for the augmented IPW estimators. We also demonstrate that a version of bootstrap variance estimators is doubly robust and can provide valid standard errors when either the outcome regression models or the propensity score model is misspecified, through simulation studies in Section 5. Let $\hat{\theta}_{\text{DR}}$ denote $\hat{\theta}_{\text{AIPW1}}$ or $\hat{\theta}_{\text{AIPW2}}$. The algorithm for obtaining the bootstrap variance estimator is described as follows:

Step 1. Select a bootstrap sample $\mathcal{S}^{[b]}$ of size n from the original sample $\mathcal{S}$ using simple random sampling with replacement.

Step 2. Calculate the bootstrap version of the doubly robust estimator, denoted as $\hat{\theta}_{\text{DR}}^{[b]}$, based on the bootstrap sample $\mathcal{S}^{[b]}$.

Step 3. Repeat Steps 1 and 2 a large number B times, independently, to obtain $\hat{\theta}_{\text{DR}}^{[b]}$, $b = 1, 2, \cdots, B$.

Step 4. Compute the bootstrap variance estimator of $\hat{\theta}_{\text{DR}}$ as

$$\text{var}\left(\hat{\theta}_{\text{DR}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{\text{DR}}^{[b]} - \bar{\hat{\theta}}_{\text{DR}}\right)^2,$$

where $\bar{\hat{\theta}}_{\text{DR}} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_{\text{DR}}^{[b]}$.

3. The PEL approach to causal inference

3.1. Pseudo-empirical likelihood

Empirical likelihood methods, first proposed by Owen (1988), are a powerful nonparametric tool for statistical inference under the likelihood principle analogous to parametric likelihood methods. Maximum empirical likelihood estimators are obtained through constrained maximization of the empirical likelihood function, and confidence intervals and hypothesis tests can be constructed through the empirical likelihood ratio function. Owen (1988) first established the empirical likelihood theorem, a nonparametric version of the Wilks' Theorem (Wilks, 1938), and generalized the result to multivariate functionals (Owen, 1990) and to linear regression models (Owen, 1991). Chen and Qin (1993) applied the empirical likelihood methods to finite population parameters under simple random sampling for more efficient estimation through the inclusion of additional constraints using auxiliary information. Qin and Lawless (1994) provided a general framework by combining empirical likelihood methods with unbiased estimating equations.

Chen and Sitter (1999) proposed the pseudo-empirical likelihood (PEL) approach to complex survey data, with the survey weights incorporated into the formulation of the empirical likelihood, to provide design-consistent point estimators for finite population parameters. Wu and Rao (2006) developed an alternative form of the PEL function for complex survey data to construct PEL ratio confidence intervals that retain all the attractive features of the empirical likelihood methods. The PEL function of Wu and Rao (2006) serves as the starting point of our proposed methods for causal inference. In this section, we focus on PEL methods for the ATE under an assumed propensity score model. A doubly robust PEL approach to causal inference is discussed in Section 4.

3.2. PEL estimation of the ATE

The pseudo-empirical likelihood function presented in Wu and Rao (2006) in the context of survey sampling can be adapted for causal inference problems. We first replace the sampling weights by suitable replacements using the estimated propensity scores for subjects in $\mathcal{S}_1$ and $\mathcal{S}_0$, and then construct the PEL function using a two-sample empirical likelihood formulation as in Wu and Yan (2012). Our proposed joint PEL function for the two samples, $\mathcal{S}_1$ and $\mathcal{S}_0$, is given by

$$\ell_{\text{PEL}}\left(\boldsymbol{p}_1, \boldsymbol{p}_0\right) = n\left\{ w_1 \sum_{j\in\mathcal{S}_1} \tilde{a}_{1j} \log\left(p_{1j}\right) + w_0 \sum_{j\in\mathcal{S}_0} \tilde{a}_{0j} \log\left(p_{0j}\right) \right\}, \tag{3.1}$$

where $\boldsymbol{p}_i = (p_{i1}, \cdots, p_{in_i})^\top$ is a set of discrete probability measure imposed over $\mathcal{S}_i$ for $i = 0, 1$, $w_1 = w_0 = 1/2$, and the normalized weights $\tilde{a}_{1j}$ and $\tilde{a}_{0j}$ are defined in Section 2.2 using the estimated propensity scores. The use of $w_1 = w_0 = 1/2$ in the PEL function $\ell_{\text{PEL}}\left(\boldsymbol{p}_1, \boldsymbol{p}_0\right)$ is to mimic the PEL function of Wu and Rao (2006) for stratified sampling where the two potential outcome variables Y_1 and Y_0 are defined over the same target population and hence with equal "stratum weights" which are also used by Wu and Yan (2012) for two-sample empirical likelihood. The formulation leads to the adaptation of simple computational procedures as described in Wu (2004, 2005). Our theoretical development, however, can be carried out with any choices of $w_1 > 0$ and $w_0 > 0$ such that $w_1 + w_0 = 1$. It becomes an interesting question for future research on whether other choices of (w_1, w_0) can lead to more efficient inference. Maximizing $\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ in (3.1) subject to the normalization constraints

$$\sum_{j\in\mathcal{S}_1} p_{1j} = 1 \quad \text{and} \quad \sum_{j\in\mathcal{S}_0} p_{0j} = 1 \tag{C1}$$

yields $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{in_i})^\top$, where $\hat{p}_{ij} = \tilde{a}_{ij}$ for $j \in \mathcal{S}_i$, $i = 1, 0$. The maximum PEL estimator of μ_i is computed as

$$\hat{\mu}_{i\text{PEL}} = \sum_{j\in\mathcal{S}_i} \hat{p}_{ij} Y_{ij} = \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij} Y_{ij}$$

for $i = 1$ and 0, which is identical to the Hájek-type IPW estimators of μ_i given in (2.1) and (2.2). The maximum PEL estimator of the ATE is given by $\hat{\theta}_{\text{PEL}} = \hat{\mu}_{1\text{PEL}} - \hat{\mu}_{0\text{PEL}}$, which is also the same as the Hájek-type IPW estimator $\hat{\theta}_{\text{IPW2}}$.

3.3. PEL ratio confidence intervals

One major advantage of the PEL approach is that we can construct the PEL ratio statistic for the parameter of interest, the ATE, as well as other parameters. Studying the asymptotic properties of the PEL ratio statistic enables the construction of confidence intervals and hypothesis tests. The "global" maximum of the PEL function $\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ under the normalization constraints (C1) is achieved at $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{in_i})^\top$, where $\hat{p}_{ij} = \tilde{a}_{ij}$ for $j \in \mathcal{S}_i$, $i = 1, 0$, as shown in Section 3.2. Let $\hat{\boldsymbol{p}}_1(\theta) = (\hat{p}_{11}(\theta), \cdots, \hat{p}_{1n_1}(\theta))^\top$ and $\hat{\boldsymbol{p}}_0(\theta) = (\hat{p}_{01}(\theta), \cdots, \hat{p}_{0n_0}(\theta))^\top$ be the "restricted" maximizer of $\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ under the normalization constraints (C1) and the following constraint (C2) induced by the parameter of interest, $\theta = \mu_1 - \mu_0$,

$$\sum_{j \in \mathcal{S}_1} p_{1j} Y_{1j} - \sum_{j \in \mathcal{S}_0} p_{0j} Y_{0j} = \theta \tag{C2}$$

for a given θ. The profile PEL function for θ is then given by

$$\ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta)) = n \left\{ w_1 \sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} \log(\hat{p}_{1j}(\theta)) + w_0 \sum_{j \in \mathcal{S}_0} \tilde{a}_{0j} \log(\hat{p}_{0j}(\theta)) \right\}. \tag{3.2}$$

The maximum PEL estimator of the ATE θ can be alternatively defined as the maximizer of the profile PEL function $\ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta))$ with respect to θ. This alternative definition is equivalent to $\hat{\theta}_{\text{PEL}} = \hat{\mu}_{1\text{PEL}} - \hat{\mu}_{0\text{PEL}}$ presented in Section 3.2.

Lemma 1. *The maximum PEL estimator $\hat{\theta}_{PEL} = \hat{\mu}_{1PEL} - \hat{\mu}_{0PEL}$ maximizes the profile PEL function $\ell_{PEL}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta))$ given in (3.2) with respect to θ.*

We now present our first major result on the PEL ratio statistic for the ATE, θ. For clarity of presentation, we let μ_1^0 and μ_0^0 denote the true values of the population means μ_1 and μ_0, and $\theta^0 = \mu_1^0 - \mu_0^0$ be the true value of the ATE. The PEL ratio function for θ is defined as

$$r_{\text{PEL}}(\theta) = \ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta)) - \ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_0)\,, \tag{3.3}$$

where $(\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_0)$ and $(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta))$ are the "global" maximizer and the "restricted" maximizer of the PEL function $\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ for a given θ, respectively. Let χ_1^2 denote a chi-squared distribution with one degree of freedom. The following theorem states the limiting distribution of the PEL ratio statistic $r_{\text{PEL}}(\theta)$ at $\theta = \theta^0$.

Theorem 1. *Under certain regularity conditions and a correctly specified propensity score model, the scaled PEL ratio function $-2r_{PEL}(\theta)/\hat{c}$ at $\theta = \theta^0$ converges in distribution to a χ_1^2 distributed random variable as $n \to \infty$, where the adjusting factor $\hat{c}$ is given in (3.4).*

The scaling constant can be theoretically defined at the population level as c. In applications, it suffices to use a consistent estimator $\hat{c}$ of c, which is given by

$$\hat{c} = n \left\{ 2 \left[\sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} Y_{1j}^2 + \sum_{j \in \mathcal{S}_0} \tilde{a}_{0j} Y_{0j}^2 - \hat{\mu}_{1\text{IPW2}}^2 - \hat{\mu}_{0\text{IPW2}}^2 \right] \right\}^{-1} \text{var}\left(\hat{\theta}_{\text{IPW2}} \right), \quad (3.4)$$

where $\text{var}(\hat{\theta}_{\text{IPW2}})$ is the variance estimator of the Hàjek-type IPW estimator $\hat{\theta}_{\text{IPW2}}$. The explicit expression of $\text{var}(\hat{\theta}_{\text{IPW2}})$ is given in Section 8.2. Proof of Theorem 1 is presented in Section 8.4.

Using the result of Theorem 1, we can construct a $100(1-\alpha)\%$-level PEL ratio confidence interval for θ in the form of

$$\{\theta \mid -2 r_{\text{PEL}}(\theta)/\hat{c} \le \chi_1^2(\alpha)\} ,$$

where $\chi_1^2(\alpha)$ is the $100(1-\alpha)$th quantile of the χ_1^2 distribution for a given $\alpha \in (0,1)$. Wu (2005) contains algorithmic details on how to find the PEL ratio confidence interval using a simple bi-section search method.

4. The PEL approach to doubly robust estimation

The PEL approach proposed in Section 3 to causal inference can be further extended to achieve doubly robust estimation through the inclusion of model-calibration constraints. Calibration methods were first developed in survey sampling (Deville and Särndal, 1992), where calibration constraints were imposed directly over individual auxiliary variables with known population controls. Wu and Sitter (2001) proposed a model-calibration technique based on a working linear or nonlinear model, where a single constraint is formed based on fitted values. The model-calibration estimators were shown to be optimal among a class of calibration estimators (Wu, 2003). In this section, we demonstrate that outcome regression models can be used to form model-calibration constraints to achieve doubly robust estimation for the ATE.

4.1. Model-calibrated maximum PEL estimators

We consider two model-calibration constraints using the two outcome regression models. Note that $\hat{m}_{1j} = m_1(\boldsymbol{x}_j; \hat{\boldsymbol{\beta}}_1)$, $\hat{m}_{0j} = m_0(\boldsymbol{x}_j; \hat{\boldsymbol{\beta}}_0)$, $\bar{\hat{m}}_1 = \sum_{j=1}^n \hat{m}_{1j}/n$, and $\bar{\hat{m}}_0 = \sum_{j=1}^n \hat{m}_{0j}/n$. The two constraints on the empirical probabilities $\boldsymbol{p}_i = (p_{i1}, \cdots, p_{in_i})^\top$ are formed as

$$\sum_{j \in \mathcal{S}_1} p_{1j} \hat{m}_{1j} = \bar{\hat{m}}_1 \quad \text{and} \quad \sum_{j \in \mathcal{S}_0} p_{0j} \hat{m}_{0j} = \bar{\hat{m}}_0 \,. \quad \text{(C3)}$$

The model-calibrated maximum PEL estimator of μ_i is computed as $\hat{\mu}_{i\text{MCP}} = \sum_{j \in \mathcal{S}_i} \hat{p}_{ij} Y_{ij}$, where $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \cdots, \hat{p}_{in_i})^\top$, $i = 0, 1$ maximizes the PEL function $\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ subject to the normalization constraints (C1) and the model-calibration constraints (C3). It can be shown by using the Lagrange multiplier method that the constrained maximizers are given by $\hat{p}_{ij} = \tilde{a}_{ij}/(1 + \hat{\lambda}_i \hat{u}_{ij})$, where $\hat{\lambda}_i$ is the solution to

$$\sum_{j \in \mathcal{S}_i} \frac{\tilde{a}_{ij} \hat{u}_{ij}}{1 + \lambda_i \hat{u}_{ij}} = 0 , \quad i = 1, 0 ,$$

and $\hat{u}_{ij} = \hat{m}_{ij} - \bar{\hat{m}}_i$. The asymptotic properties of $\hat{\mu}_{1\text{MCP}}$ are summarized in the theorem below. Parallel results can be stated regarding $\hat{\mu}_{0\text{MCP}}$.

Theorem 2. *Under suitable regularity conditions, the model-calibrated maximum PEL estimator $\hat{\mu}_{1MCP}$ of μ_1 is doubly robust with respect to the propensity score model and the outcome regression model. Moreover, if the propensity score model is correctly specified, the estimator $\hat{\mu}_{1MCP}$ admits the asymptotic expansion*

$$\hat{\mu}_{1MCP} = \sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} Y_{1j} + \hat{B} \left\{ \bar{\hat{m}}_1 - \sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} \hat{m}_{1j} \right\} + o_p \left(n^{-1/2} \right) , \tag{4.1}$$

where

$$\hat{B} = \left\{ \sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} \hat{u}_{1j} Y_{1j} \right\} \left\{ \sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} \hat{u}_{1j}^2 \right\}^{-1} . \tag{4.2}$$

The inclusion of the model-calibration constraints leads to a doubly robust estimator for the ATE as $\hat{\theta}_{\text{MCP}} = \hat{\mu}_{1\text{MCP}} - \hat{\mu}_{0\text{MCP}}$. This estimator is more efficient than the IPW estimator when both sets of models are correctly specified. Specifically, if the outcome regression model is correctly specified, the $\hat{B}$ given in (4.2) converges to 1, and the newly constructed estimator $\hat{\mu}_{1\text{MCP}}$ is asymptotically equivalent to the augmented IPW estimator $\hat{\mu}_{1\text{AIPW2}}$ and is particularly efficient. The $\hat{B}$ does not converge to 1 otherwise. The expression (4.1) also implies that $\hat{\mu}_{1\text{MCP}} = \hat{\mu}_{1\text{IPW2}} + O_p(n^{-1/2})$, which further implies $\hat{\mu}_{1\text{MCP}} - \mu_1^0 = O_p(n^{-1/2})$ when the propensity score model is correctly specified. Proof of Theorem 2 is given in Section 8.5.

An important feature of our proposed PEL approach to causal inference is the flexibility of including additional constraints under the constrained maximization framework. Useful information can be incorporated through suitable constraints, which often lead to more robust and more efficient estimators (Chan and Yam, 2014). This is related to a topic of recent research interests, the so-called multiply robust estimators; see Han and Wang (2013), Han (2014b), Han (2014a), Han (2016), Chen and Haziza (2017), and Duan and Yin (2017), among others, for further discussion.

4.2. Model-calibrated PEL ratio confidence intervals

We now establish the limiting distribution of the PEL ratio statistic with the inclusion of model-calibration constraints. Let $\hat{\boldsymbol{p}}_1(\theta) = (\hat{p}_{11}(\theta), \cdots, \hat{p}_{1n_1}(\theta))^\top$ and $\hat{\boldsymbol{p}}_0(\theta) = (\hat{p}_{01}(\theta), \cdots, \hat{p}_{0n_0}(\theta))^\top$ be the maximizer of the PEL function

$\ell_{\text{PEL}}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ given in (3.1) under the normalization constraints (C1), the model-calibration constraints (C3), and the parameter constraint (C2) for a fixed θ. Let $\ell_{\text{PEL}}(\theta) = \ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta))$ be the profile PEL function of θ, which is the "restricted" maximum of the PEL function under the constraints (C1), (C3) and (C2) for the given θ.

Lemma 2. *The model-calibrated maximum PEL estimator $\hat{\theta}_{MCP}$ maximizes the profile PEL function $\ell_{PEL}(\theta)$ with respect to θ.*

The profile PEL function $\ell_{\text{PEL}}(\theta)$ with a given θ can be computed by existing algorithms for empirical likelihood through reformulations of the constraints. Under the normalization constraints (C1), the model calibration constraints (C3) can be rewritten as $\sum_{j\in\mathcal{S}_i} p_{ij}\hat{u}_{ij} = 0$ for $i = 1, 0$, and the parameter constraint (C2) is equivalent to $\sum_{j\in\mathcal{S}_1} p_{1j}r_{1j} - \sum_{j\in\mathcal{S}_0} p_{0j}r_{0j} = 0$, where $r_{1j} = Y_{1j} - \theta/2$ and $r_{0j} = Y_{0j} + \theta/2$. Note that $w_1 = w_0 = 1/2$. The constraints (C1)-(C3) altogether can be thus reformulated as

$$\begin{aligned} \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} p_{ij} &= 1 \\ \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} p_{ij}\boldsymbol{g}_{ij}(\theta) &= \boldsymbol{0}\ , \end{aligned} \tag{C4}$$

where $\boldsymbol{g}_{1j}(\theta) = (1-w_1, \hat{u}_{1j}, \hat{u}_{1j}, r_{1j}/w_1)^\top$ and $\boldsymbol{g}_{0j}(\theta) = (-w_1, \hat{u}_{0j}, -\hat{u}_{0j}, -r_{0j}/w_0)^\top$.

The major motivation behind this reformulation is to use computational procedures for stratified sampling as described in Wu and Rao (2006) and Wu (2005). Maximizing (3.1) subject to the constraints in (C4) for a fixed θ yields the solution

$$\hat{p}_{ij}(\theta) = \frac{\tilde{a}_{ij}}{1 + \hat{\boldsymbol{\lambda}}^\top \boldsymbol{g}_{ij}(\theta)}\ ,$$

where the Lagrange multiplier $\hat{\boldsymbol{\lambda}}$ is obtained by solving

$$\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \frac{\tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)} = \boldsymbol{0}\ .$$

Let $\boldsymbol{W}$ be the limit of $\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta^0)\boldsymbol{g}_{ij}(\theta^0)^\top$, $\boldsymbol{\Gamma} = (0,0,0,-1)^\top$, and $\sigma = (\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}\boldsymbol{\Gamma})^{-1}$. We denote the asymptotic variance-covariance matrix of $\sqrt{n}\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta^0)$ by $\boldsymbol{\Omega}$. In what follows, we present the limiting distribution of the maximum PEL estimator $\hat{\theta}_{\text{MCP}}$ in Theorem 3 and the limiting distribution of the PEL ratio statistic $r_{\text{PEL}}(\theta) = \ell_{\text{PEL}}(\theta) - \ell_{\text{PEL}}(\hat{\theta}_{\text{MCP}})$ at $\theta = \theta^0$ in Theorem 4. Proofs of Theorem 3 and Theorem 4 are given in Section 8.6 and Section 8.7, respectively.

Theorem 3. *Under certain regularity conditions and a correctly specified propensity score model, we have $\sqrt{n}\left(\hat{\theta}_{MCP} - \theta^0\right) \xrightarrow{d} \mathrm{N}(0, V_1)$ as $n \to \infty$, where $V_1 = \sigma^2\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}\boldsymbol{\Omega}\boldsymbol{W}^{-1}\boldsymbol{\Gamma}$ and $\xrightarrow{d}$ denotes convergence in distribution.*

In practice, we use $\hat{\boldsymbol{W}} = \sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \tilde{a}_{ij} \boldsymbol{g}_{ij}(\hat{\theta}_{\text{MCP}}) \boldsymbol{g}_{ij}(\hat{\theta}_{\text{MCP}})^\top$ as an estimate for the limiting matrix $\boldsymbol{W}$, which also leads to $\hat{\sigma} = (\boldsymbol{\Gamma}^\top \hat{\boldsymbol{W}}^{-1} \boldsymbol{\Gamma})^{-1}$. The estimated variance-covariance matrix ($\boldsymbol{\Omega}$) is $\hat{\boldsymbol{\Omega}} = n^{-1} \sum_{j \in \mathcal{S}} (\hat{\boldsymbol{h}}_j - \bar{\hat{\boldsymbol{h}}})(\hat{\boldsymbol{h}}_j - \bar{\hat{\boldsymbol{h}}})^\top$, where $\hat{\boldsymbol{h}}_j$ is the plug-in estimator of

$$\boldsymbol{h}_j = \begin{pmatrix} 0 \\ \frac{1}{2}\left\{\left(\frac{T_j}{\tau_j^0} - 1\right)\left(m_{1j}^* - \mathrm{E}\left(m_{1j}^*\right)\right) + \left(\frac{1-T_j}{1-\tau_j^0} - 1\right)\left(m_{0j}^* - \mathrm{E}\left(m_{0j}^*\right)\right) - (\boldsymbol{A} + \boldsymbol{E})\,\boldsymbol{C}^{-1}\tilde{\boldsymbol{x}}_j\left(T_j - \tau_j^0\right)\right\} \\ \frac{1}{2}\left\{\left(\frac{T_j}{\tau_j^0} - 1\right)\left(m_{1j}^* - \mathrm{E}\left(m_{1j}^*\right)\right) - \left(\frac{1-T_j}{1-\tau_j^0} - 1\right)\left(m_{0j}^* - \mathrm{E}\left(m_{0j}^*\right)\right) - (\boldsymbol{A} - \boldsymbol{E})\,\boldsymbol{C}^{-1}\tilde{\boldsymbol{x}}_j\left(T_j - \tau_j^0\right)\right\} \\ \frac{T_j}{\tau_j^0}\left(Y_{1j} - \mu_1^0\right) - \frac{1-T_j}{1-\tau_j^0}\left(Y_{0j} - \mu_0\right) - (\boldsymbol{J} - \boldsymbol{G})\,\boldsymbol{C}^{-1}\tilde{\boldsymbol{x}}_j\left(T_j - \tau_j^0\right) \end{pmatrix}$$

and $\bar{\hat{\boldsymbol{h}}} = n^{-1} \sum_{j \in \mathcal{S}} \hat{\boldsymbol{h}}_j$. In the above expression of $\boldsymbol{h}_j$, $\tau_j^0 = \tau(\boldsymbol{x}_j; \boldsymbol{\alpha}^0)$ is the propensity score under the true propensity score model with $\boldsymbol{\alpha}^0$ representing the true value of $\boldsymbol{\alpha}$, $m_{ij}^* = m_i\left(\boldsymbol{x}_j, \boldsymbol{\beta}_i^*\right)$ for $i = 0, 1$, where $\boldsymbol{\beta}_i^*$ is the probability limit of $\hat{\boldsymbol{\beta}}_i$ under the assumed outcome regression model, and

$$\begin{aligned} \boldsymbol{A} &= -\mathrm{E}\left[\left\{m_{1j}^* - \mathrm{E}(m_{1j}^*)\right\}\left(1 - \tau_j^0\right)\tilde{\boldsymbol{x}}_j^\top\right], \quad \boldsymbol{E} = \mathrm{E}\left[\left\{m_{0j}^* - \mathrm{E}(m_{0j}^*)\right\}\tau_j^0 \tilde{\boldsymbol{x}}_j^\top\right], \\ \boldsymbol{J} &= -\mathrm{E}\left[T_j\left(Y_{1j} - \mu_1^0\right)\left(1 - \tau_j^0\right)\tilde{\boldsymbol{x}}_j^\top / \tau_j^0\right], \quad \boldsymbol{G} = \mathrm{E}\left[\left(1 - T_j\right)\left(Y_{0j} - \mu_0^0\right)\left(1 - \tau_j^0\right)^{-1}\tau_j^0 \tilde{\boldsymbol{x}}_j^\top\right], \\ \boldsymbol{C} &= -\mathrm{E}\left[\tau_j^0\left(1 - \tau_j^0\right)\tilde{\boldsymbol{x}}_j \tilde{\boldsymbol{x}}_j^\top\right]. \end{aligned}$$

Note that the full expression of the model-calibrated PEL ratio function $r_{\text{PEL}}(\theta) = \ell_{\text{PEL}}(\theta) - \ell_{\text{PEL}}(\hat{\theta}_{\text{MCP}})$ is given by

$$r_{\text{PEL}}(\theta) = \ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta)) - \ell_{\text{PEL}}\left(\hat{\boldsymbol{p}}_1\left(\hat{\theta}_{\text{MCP}}\right), \hat{\boldsymbol{p}}_0\left(\hat{\theta}_{\text{MCP}}\right)\right).$$

The asymptotic distribution of $r_{\text{PEL}}(\theta)$ at $\theta = \theta^0$ is given in the theorem below. Note that the rank of the matrix $\boldsymbol{M} = \sigma \boldsymbol{\Omega}^{1/2} \boldsymbol{W}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1} \boldsymbol{\Omega}^{1/2}$ is one.

Theorem 4. *Under certain regularity conditions and a correctly specified propensity score model, we have* $-2r_{\text{PEL}}(\theta) \xrightarrow{d} \delta \chi_1^2$ *when* $\theta = \theta^0$*, where* δ *is the unique non-zero eigenvalue of the matrix* $\boldsymbol{M}$*.*

The above result can be used to construct a $100(1 - \alpha)\%$ model-calibrated PEL ratio confidence interval for θ as

$$\left\{\theta \mid -2r_{\text{PEL}}(\theta)/\hat{\delta} \le \chi_1^2(\alpha)\right\},$$

where $\hat{\delta}$ is the non-zero eigenvalue of $\hat{\boldsymbol{M}} = \hat{\sigma} \hat{\boldsymbol{\Omega}}^{1/2} \hat{\boldsymbol{W}}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Gamma}^\top \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{\Omega}}^{1/2}$. Computational details and R codes pertaining to the reformulated constrained maximization problem and the method for constructing the model-calibrated PEL ratio confidence interval can be found in Wu (2005).

4.3. Bootstrap PEL ratio confidence intervals

The bootstrap-calibrated empirical likelihood method can provide an improved approximation to the sampling distribution of the empirical likelihood ratio function, especially when the sample size is small (Owen, 2001). Wu and Rao (2010) proposed a bootstrap procedure for approximating the sampling distribution of the pseudo-empirical likelihood ratio function for complex survey data under certain sampling designs. Chen et al. (2018) discussed bootstrap-calibration procedures for approximating the asymptotic distribution of the PEL ratio function for the nonrandomized pretest-posttest study designs.

The asymptotic distributions of the proposed PEL ratio functions in Sections 3 and 4 involve scaling constants that need to be estimated. Estimation of the scaling constants requires variance estimation, which involves heavy analytic expressions. We propose a bootstrap approach to constructing model-calibrated PEL ratio confidence intervals for the ATE as follows. This procedure bypasses the need for estimating the scaling constant required for the χ^2 approximation.

Step 1 Select the b-th bootstrap sample $\mathcal{S}^{[b]}$ of n units from the initial sample $\mathcal{S}$ by simple random sampling with replacement. Let $\mathcal{S}_1^{[b]} = \{j | j \in \mathcal{S}^{[b]} \text{ and } T_j = 1\}$ and $\mathcal{S}_0^{[b]} = \{j | j \in \mathcal{S}^{[b]} \text{ and } T_j = 0\}$.

Step 2 Fit the propensity score model and the outcome regression models to the b-th bootstrap sample, and obtain estimates $\hat{\boldsymbol{\alpha}}^{[b]}$, $\hat{\boldsymbol{\beta}}_1^{[b]}$ and $\hat{\boldsymbol{\beta}}_0^{[b]}$. Calculate the estimated propensity scores by $\hat{\tau}_j^{[b]} = \tau(\boldsymbol{x}_j; \hat{\boldsymbol{\alpha}}^{[b]})$ and the fitted values $\hat{m}_{1j}^{[b]} = m_1(\boldsymbol{x}_j; \hat{\boldsymbol{\beta}}_1^{[b]})$ and $\hat{m}_{0j}^{[b]} = m_0(\boldsymbol{x}_j; \hat{\boldsymbol{\beta}}_0^{[b]})$ for all the units $j \in \mathcal{S}^{[b]}$.

Step 3 Construct the bootstrap version of the PEL function as

$$\ell_{\text{PEL}}^{[b]}(\boldsymbol{p}_1, \boldsymbol{p}_0) = n \left\{ w_1 \sum_{j \in \mathcal{S}_1^{[b]}} \tilde{a}_{1j}^{[b]} \log(p_{1j}) + w_0 \sum_{j \in \mathcal{S}_0^{[b]}} \tilde{a}_{0j}^{[b]} \log(p_{0j}) \right\}, \quad (4.3)$$

where $\tilde{a}_{1j}^{[b]} = (\hat{\tau}_j^{[b]})^{-1} / \sum_{l \in \mathcal{S}_1^{[b]}} (\hat{\tau}_l^{[b]})^{-1}$ for $j \in \mathcal{S}_1^{[b]}$ and $\tilde{a}_{0j}^{[b]} = (1 - \hat{\tau}_j^{[b]})^{-1} / \sum_{l \in \mathcal{S}_0^{[b]}} (1 - \hat{\tau}_l^{[b]})^{-1}$ for $j \in \mathcal{S}_0^{[b]}$. Specify the bootstrap versions of the constraints of (C1), (C2) and (C3) as

$$\begin{aligned} &\sum_{j \in \mathcal{S}_1^{[b]}} p_{1j} = 1 \quad \text{and} \quad \sum_{j \in \mathcal{S}_0^{[b]}} p_{0j} = 1; \\ &\sum_{j \in \mathcal{S}_1^{[b]}} p_{1j} Y_{1j} - \sum_{j \in \mathcal{S}_0^{[b]}} p_{0j} Y_{0j} = \theta; \\ &\sum_{j \in \mathcal{S}_1^{[b]}} p_{1j} \hat{m}_{1j}^{[b]} = \frac{1}{n} \sum_{j \in \mathcal{S}^{[b]}} \hat{m}_{1j}^{[b]} \quad \text{and} \quad \sum_{j \in \mathcal{S}_0^{[b]}} p_{0j} \hat{m}_{0j}^{[b]} = \frac{1}{n} \sum_{j \in \mathcal{S}^{[b]}} \hat{m}_{0j}^{[b]}. \end{aligned} \quad (4.4)$$

Compute the PEL ratio function $r_{\text{PEL}}^{[b]}(\theta)$ using $\ell_{\text{PEL}}^{[b]}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ with constraints (4.4) and $\theta = \hat{\theta}_{\text{MCP}}$.

Step 4 Repeat Steps 1-3 a large number B times, independently, to obtain values of $r_{\text{PEL}}^{[b]}(\theta)$ for $b = 1, \ldots, B$, all at $\theta = \hat{\theta}_{\text{MCP}}$.

The number B is typically set to be $1,000$ but can be larger if more computing resources are available. Let α^B denote the lower 100α-th quantile of the sequence $r_{\mathrm{PEL}}^{[b]}(\hat{\theta}_{\mathrm{MCP}})$, $b = 1, \ldots, B$. The bootstrap-calibrated $(1-\alpha)$-level PEL ratio confidence interval for θ is computed as

$$\left\{\theta \mid r_{\mathrm{PEL}}(\theta) > \alpha^B\right\} .$$

The validity of the bootstrap approach under a correctly specified propensity score model is justified in Section 8.8. When the propensity score model is misspecified, the arguments used in the justification no longer hold. However, our simulation results in Section 5 show that the bootstrap procedure works well when the outcome regression models are correctly specified.

5. Simulation studies

5.1. Point estimators and confidence intervals

We conduct limited simulation studies to investigate the performance of our proposed maximum PEL point estimators and PEL ratio confidence intervals for the average treatment effect based on $n_{sim} = 1,000$ simulation samples. The results are compared with the commonly used IPW estimator and the AIPW estimator. We use $B = 1,000$ for the bootstrap methods with each simulated sample. The performance measurement metrics and the candidate methods are as follows.

(i) For point estimators, we compute the percentage relative bias ($\%RB$) and the mean squared error (MSE) for an estimator $\hat{\theta}$ of the parameter θ as

$$\%RB = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} \frac{\hat{\theta}^{(s)} - \theta^0}{\theta^0} \times 100 \quad \text{and} \quad MSE = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} \left(\hat{\theta}^{(s)} - \theta^0\right)^2 ,$$

where $\hat{\theta}^{(s)}$ is the estimator $\hat{\theta}$ computed from the sth simulation sample. Results of $\%RB$ and MSE for estimators $\hat{\theta}_{\mathrm{IPW2}}$, $\hat{\theta}_{\mathrm{PEL}}$, $\hat{\theta}_{\mathrm{AIPW2}}$, and $\hat{\theta}_{\mathrm{MCP}}$ are presented.

(ii) For confidence intervals, we compute the percentage coverage probability ($\%CP$) and the average length (AL) for the confidence interval $\mathcal{I}$ as

$$\%CP = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} I\left(\theta^0 \in \mathcal{I}^{(s)}\right) \times 100 \;\; \text{and} \;\; AL = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} \left(UB^{(s)} - LB^{(s)}\right),$$

where $I(\cdot)$ is the indicator function and $\mathcal{I}^{(s)}$ is the interval $\mathcal{I}$ computed from the sth simulation sample with $UB^{(s)}$ as the upper bound and $LB^{(s)}$ as the lower bound, i.e., $\mathcal{I}^{(s)} = (LB^{(s)}, UB^{(s)})$. We include results of $\%CP$ and AL for the Wald-type confidence interval using $\hat{\theta}_{\mathrm{IPW2}}$ and the sandwich variance estimator ($\mathcal{I}_{\mathrm{IPW2}}$), the PEL ratio confidence interval based on $\hat{\theta}_{\mathrm{PEL}}$ ($\mathcal{I}_{\mathrm{PELR}}$), the Wald-type confidence interval using $\hat{\theta}_{\mathrm{AIPW2}}$ and the sandwich variance estimator ($\mathcal{I}_{\mathrm{AIPW2}}$) or the Bootstrap variance estimator ($\mathcal{I}_{\mathrm{AIPW2B}}$), the model-calibrated PEL ratio confidence interval based on $\hat{\theta}_{\mathrm{MCP}}$ and the scaled chi-squared distribution ($\mathcal{I}_{\mathrm{MCP}}$) or the bootstrap method ($\mathcal{I}_{\mathrm{MCPB}}$). The nominal value of the coverage probability is 95%.

We consider an independent sample $\mathcal{S}$ of size n from an infinite population $(Y_1, Y_0, T, \boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, x_3)^\top$. We consider $n = 100$, 200 and 400 in the simulation. For each subject j, $j = 1, \ldots, n$, $\boldsymbol{x}_j = (x_{j1}, x_{j2}, x_{j3})^\top$, and the three covariates are generated by $x_{j1} = v_{j1}$, $x_{j2} = v_{j2} + 0.2x_{j1}$ and $x_{j3} = v_{j3} + 0.3(x_{j1} + x_{j2})$, with $v_{j1} \sim \mathrm{N}(0, 1)$, $v_{j2} \sim$ Bernoulli(0.6) and $v_{j3} \sim$ Exponential(1). The true propensity scores $\tau_j^0 = P(T_j = 1 \mid \boldsymbol{x}_j)$ for treatment assignments are given by

$$\tau : \quad \tau_j^0 = \mathrm{expit}(\alpha_0 + 0.2x_{j1} + 0.2x_{j2} - 0.5x_{j3})\,, \quad j = 1, \ldots, n,$$

where the value of α_0 controls the expected proportion of treatment subjects denoted by t. We consider three scenarios with $t = 0.3$, 0.5, or 0.7. Under the scenario $t = 0.3$, for instance, there are approximately 30% of the subjects in the sample $\mathcal{S}$ belonging to the treatment group. The treatment assignment for subject j is decided based on the treatment indicator $T_j \sim \mathrm{Bernoulli}(\tau_j^0)$. The outcome regression models are specified as

$$m_1 : \quad Y_{1j} = 4.5 + x_{j1} - 2x_{j2} + 3x_{j3} + a_1\epsilon_j$$

and

$$m_0 : \quad Y_{0j} = 1 + x_{j1} + x_{j2} + 2x_{j3} + a_0\epsilon_j\ ,$$

where $\epsilon_j \sim \mathrm{N}(0, 1)$, $j = 1, \ldots, n$. The values of a_1 and a_0 are chosen to control the correlation coefficient ρ between the linear predictor of $\boldsymbol{x}_j$ and the potential outcomes Y_{1j} and Y_{0j}. We consider three scenarios with $\rho = 0.3$, 0.5, or 0.7, representing weak, mild and strong prediction power of the covariates. For all three scenarios, the true ATE is $\theta^0 = 2.88$. The sample dataset is given by $\{(\boldsymbol{x}_j, T_j, Y_j), j \in \mathcal{S}\}$, where $Y_j = T_jY_{1j} + (1 - T_j)Y_{0j}$.

We consider three scenarios for the misspecification of the propensity score model τ and the outcome regression models m_1 and m_0.

(i) TT: Both the propensity score model and the outcome regression models are correctly specified.
(ii) TF: The propensity score model is correctly specified, but the two outcome regression models are misspecified by excluding x_3 from the model.
(iii) FT: The outcome regression models are correctly specified, but the propensity score model is misspecified by omitting x_3 in the model.

Our simulation studies are conducted for each of the combinations of ρ, t, n and working models, resulting in a total of $3 \times 3 \times 3 \times 3 = 81$ simulation settings. Tables 1, 2 and 3 present the percentage relative biases ($\%RB$'s) and the mean squared errors (MSE's) $\times 100$ for the point estimators under different combinations of the sample size n, the correlation ρ, and the model specification scenarios, when the expected treatment proportion t is 0.3, 0.5 and 0.7. Since the misspecification of the outcome regression models has no impact on $\hat{\theta}_{\text{IPW2}}$ and $\hat{\theta}_{\text{PEL}}$, their simulation results under the "TF" scenario are identical to the ones under the "TT" scenario, and hence are not shown.

From the simulation results, we can see that (1) $\hat{\theta}_{\text{IPW2}}$ and $\hat{\theta}_{\text{PEL}}$ are exactly the same, and they perform well when the propensity score model is correctly

Table 1
%RB and MSE(×100) for Point Estimators when $t = 0.3$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
100	TT	$\hat{\theta}_{\text{IPW2}}$	-6.9	635.6	-6.2	208.3	-5.8	90.5
		$\hat{\theta}_{\text{PEL}}$	-6.9	635.6	-6.2	208.3	-5.8	90.5
		$\hat{\theta}_{\text{AIPW2}}$	-1.4	676.0	-0.8	202.2	-0.5	71.8
		$\hat{\theta}_{\text{MCP}}$	-1.6	686.4	-1.0	204.8	-0.6	72.6
	TF	$\hat{\theta}_{\text{AIPW2}}$	-8.4	624.5	-7.8	203.0	-7.5	87.0
		$\hat{\theta}_{\text{MCP}}$	-9.6	609.6	-8.8	195.7	-8.4	82.4
	FT	$\hat{\theta}_{\text{IPW2}}$	-38.2	680.4	-37.4	299.6	-37.0	193.5
		$\hat{\theta}_{\text{PEL}}$	-38.2	680.4	-37.4	299.6	-37.0	193.5
		$\hat{\theta}_{\text{AIPW2}}$	-1.3	665.3	-0.7	199.1	-0.4	70.8
		$\hat{\theta}_{\text{MCP}}$	-1.1	694.9	-0.8	208.6	-0.5	74.1
200	TT	$\hat{\theta}_{\text{IPW2}}$	-5.0	334.4	-4.0	118.7	-3.6	61.7
		$\hat{\theta}_{\text{PEL}}$	-5.0	334.4	-4.0	118.7	-3.6	61.7
		$\hat{\theta}_{\text{AIPW2}}$	-2.1	327.1	-1.2	98.6	-0.8	35.4
		$\hat{\theta}_{\text{MCP}}$	-1.9	327.8	-1.1	98.8	-0.8	35.4
	TF	$\hat{\theta}_{\text{IPW2}}$	-5.8	327.8	-4.8	114.6	-4.3	59.4
		$\hat{\theta}_{\text{MCP}}$	-6.5	323.1	-5.7	108.1	-5.0	53.1
	FT	$\hat{\theta}_{\text{IPW2}}$	-38.0	410.1	-37.3	210.3	-37.0	154.3
		$\hat{\theta}_{\text{PEL}}$	-38.0	410.1	-37.3	210.3	-37.0	154.3
		$\hat{\theta}_{\text{AIPW2}}$	-2.1	324.5	-1.3	97.8	-0.8	35.1
		$\hat{\theta}_{\text{MCP}}$	-2.4	329.8	-1.5	100.2	-1.0	36.0
400	TT	$\hat{\theta}_{\text{IPW2}}$	-3.8	168.8	-2.7	58.1	-2.2	27.7
		$\hat{\theta}_{\text{PEL}}$	-3.8	168.8	-2.7	58.1	-2.2	27.7
		$\hat{\theta}_{\text{AIPW2}}$	-2.3	158.0	-1.3	47.3	-0.7	16.8
		$\hat{\theta}_{\text{MCP}}$	-2.3	156.8	-1.2	46.9	-0.7	16.7
	TF	$\hat{\theta}_{\text{AIPW2}}$	-4.1	166.3	-3.0	56.6	-2.5	26.6
		$\hat{\theta}_{\text{MCP}}$	-4.4	164.8	-3.3	55.7	-2.8	25.7
	FT	$\hat{\theta}_{\text{IPW2}}$	-38.2	262.1	-37.4	161.5	-36.9	132.7
		$\hat{\theta}_{\text{PEL}}$	-38.2	262.1	-37.4	161.5	-36.9	132.7
		$\hat{\theta}_{\text{AIPW2}}$	-1.9	153.4	-1.1	45.9	-0.6	16.3
		$\hat{\theta}_{\text{MCP}}$	-1.9	157.9	-1.1	47.3	-0.6	16.8

specified ("TT"); although the approximately -6% RB is slightly large in the case of $(n,t) = (100, 0.3)$, this might be due to the small sample size in the treatment group; there are noticeable biases when the propensity score model is not correctly specified ("FT"), revealing the failure of the two estimators in this scenario; (2) the performances of $\hat{\theta}_{\text{AIPW2}}$ and $\hat{\theta}_{\text{MCP}}$ are similar as expected, and they perform well in all of the three model specification scenarios, showing that they are doubly robust; although the absolute values of the RB's of $\hat{\theta}_{\text{AIPW2}}$ and $\hat{\theta}_{\text{MCP}}$ are over 5% for some "TF" cases, these values are still reasonable because those of $\hat{\theta}_{\text{IPW2}}$ and $\hat{\theta}_{\text{PEL}}$ are close to or over 5% in the corresponding "TT" cases (for example, $(n,t) = (100, 0.3)$) and the consistency of $\hat{\theta}_{\text{AIPW2}}$ or $\hat{\theta}_{\text{MCP}}$ relies on the propensity score model in "TF" cases; (3) when both models are correctly specified ("TT"), the mean squared errors of $\hat{\theta}_{\text{AIPW2}}$ and $\hat{\theta}_{\text{MCP}}$ seem to be smaller than those of $\hat{\theta}_{\text{IPW2}}$ and $\hat{\theta}_{\text{PEL}}$; and (4) the MSE of each estimator decreases as

Table 2
%RB and MSE(×100) for Point Estimators when $t = 0.5$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
100	TT	$\hat{\theta}_{\text{IPW2}}$	0.1	483.9	-1.1	164.5	-1.7	75.0
		$\hat{\theta}_{\text{PEL}}$	0.1	483.9	-1.1	164.5	-1.7	75.0
		$\hat{\theta}_{\text{AIPW2}}$	2.2	459.0	1.2	139.2	0.7	50.7
		$\hat{\theta}_{\text{MCP}}$	2.2	453.7	1.2	137.6	0.7	50.2
	TF	$\hat{\theta}_{\text{AIPW2}}$	-0.8	474.5	-2.0	156.3	-2.6	67.2
		$\hat{\theta}_{\text{MCP}}$	-1.3	461.8	-2.5	147.9	-3.0	60.4
	FT	$\hat{\theta}_{\text{IPW2}}$	-34.7	513.7	-36.2	245.6	-37.0	174.2
		$\hat{\theta}_{\text{PEL}}$	-34.7	513.7	-36.2	245.6	-37.0	174.2
		$\hat{\theta}_{\text{AIPW2}}$	2.7	451.3	1.5	136.7	0.8	49.8
		$\hat{\theta}_{\text{MCP}}$	2.2	451.0	1.2	136.5	0.7	49.6
200	TT	$\hat{\theta}_{\text{IPW2}}$	0.6	226.4	-0.7	72.0	-1.3	29.6
		$\hat{\theta}_{\text{PEL}}$	0.6	226.4	-0.7	72.0	-1.3	29.6
		$\hat{\theta}_{\text{AIPW2}}$	2.7	224.2	1.4	67.6	0.7	24.4
		$\hat{\theta}_{\text{MCP}}$	2.7	224.1	1.3	67.5	0.7	24.4
	TF	$\hat{\theta}_{\text{AIPW2}}$	0.2	224.9	-1.1	71.2	-1.7	29.2
		$\hat{\theta}_{\text{MCP}}$	-0.1	224.1	-1.3	70.7	-1.9	28.7
	FT	$\hat{\theta}_{\text{IPW2}}$	-35.8	326.1	-37.2	188.2	-37.9	152.2
		$\hat{\theta}_{\text{PEL}}$	-35.8	326.1	-37.2	188.2	-37.9	152.2
		$\hat{\theta}_{\text{AIPW2}}$	2.6	221.9	1.3	66.9	0.7	24.1
		$\hat{\theta}_{\text{MCP}}$	2.5	222.8	1.2	67.1	0.6	24.2
400	TT	$\hat{\theta}_{\text{IPW2}}$	-1.3	122.5	-1.1	38.3	-1.0	15.3
		$\hat{\theta}_{\text{PEL}}$	-1.3	122.5	-1.1	38.3	-1.0	15.3
		$\hat{\theta}_{\text{AIPW2}}$	-0.4	122.2	-0.2	37.0	-0.1	13.4
		$\hat{\theta}_{\text{MCP}}$	-0.4	121.5	-0.2	36.8	-0.1	13.3
	TF	$\hat{\theta}_{\text{AIPW2}}$	-1.6	122.6	-1.4	38.4	-1.3	15.3
		$\hat{\theta}_{\text{MCP}}$	-1.7	122.4	-1.5	38.3	-1.4	15.2
	FT	$\hat{\theta}_{\text{IPW2}}$	-37.8	234.8	-37.6	155.9	-37.5	133.9
		$\hat{\theta}_{\text{PEL}}$	-37.8	234.8	-37.6	155.9	-37.5	133.9
		$\hat{\theta}_{\text{AIPW2}}$	-0.5	119.4	-0.3	36.2	-0.2	13.1
		$\hat{\theta}_{\text{MCP}}$	-0.5	119.6	-0.3	36.3	-0.2	13.2

n increases.

The percentage coverage probabilities ($\%CP$'s) and average lengths (AL's) $\times 100$ of the 95% confidence intervals under different settings are reported in Tables 4, 5 and 6 for $t = 0.3$, $t = 0.5$ and $t = 0.7$, respectively. Since $\mathcal{I}_{\text{IPW2}}$ and $\mathcal{I}_{\text{PELR}}$ do not rely on the outcome regression models, their simulation results in "TT" and "TF" are identical and are only presented for the scenario "TT". The results show that (1) when both models are correctly specified ("TT"), none of these methods fails; (2) all methods work well if the outcome regression models are misspecified but the propensity score model is correctly specified ("TF"); (3) the two confidence intervals $\mathcal{I}_{\text{IPW2}}$ and $\mathcal{I}_{\text{PELR}}$ are not reliable when the propensity score model is misspecified ("FT"), as the corresponding point estimators are not valid in this scenario; (4) confidence intervals $\mathcal{I}_{\text{AIPW2}}$, $\mathcal{I}_{\text{AIPW2B}}$, $\mathcal{I}_{\text{MCP}}$ and $\mathcal{I}_{\text{MCPB}}$ seem to perform well in all three model specification scenarios, meaning

Table 3
%RB and MSE(×100) for Point Estimators when $t = 0.7$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
100	TT	$\hat{\theta}_{\text{IPW2}}$	-1.4	544.9	-1.3	168.0	-1.3	64.2
		$\hat{\theta}_{\text{PEL}}$	-1.4	544.9	-1.3	168.0	-1.3	64.2
		$\hat{\theta}_{\text{AIPW2}}$	-0.6	545.9	-0.3	164.7	-0.2	59.4
		$\hat{\theta}_{\text{MCP}}$	-0.6	546.3	-0.3	164.6	-0.2	59.3
	TF	$\hat{\theta}_{\text{AIPW2}}$	-2.6	543.1	-2.5	165.7	-2.4	61.9
		$\hat{\theta}_{\text{MCP}}$	-2.7	543.0	-2.6	165.7	-2.6	61.9
	FT	$\hat{\theta}_{\text{IPW2}}$	-40.9	648.6	-40.8	312.2	-40.8	219.5
		$\hat{\theta}_{\text{PEL}}$	-40.9	648.6	-40.8	312.2	-40.8	219.5
		$\hat{\theta}_{\text{AIPW2}}$	-0.8	531.5	-0.4	160.4	-0.3	57.9
		$\hat{\theta}_{\text{MCP}}$	-0.7	540.2	-0.2	162.2	-0.1	58.3
200	TT	$\hat{\theta}_{\text{IPW2}}$	0.7	259.3	-0.2	78.5	-0.6	29.1
		$\hat{\theta}_{\text{PEL}}$	0.7	259.3	-0.2	78.5	-0.6	29.1
		$\hat{\theta}_{\text{AIPW2}}$	1.6	256.9	0.8	76.5	0.3	27.1
		$\hat{\theta}_{\text{MCP}}$	1.7	256.5	0.8	76.4	0.4	27.0
	TF	$\hat{\theta}_{\text{AIPW2}}$	0.0	257.8	-0.9	77.9	-1.3	28.6
		$\hat{\theta}_{\text{MCP}}$	0.0	257.5	-0.9	77.8	-1.3	28.5
	FT	$\hat{\theta}_{\text{IPW2}}$	-39.7	383.8	-40.7	226.0	-41.2	183.0
		$\hat{\theta}_{\text{PEL}}$	-39.7	383.8	-40.7	226.0	-41.2	183.0
		$\hat{\theta}_{\text{AIPW2}}$	1.5	251.3	0.7	74.8	0.3	26.5
		$\hat{\theta}_{\text{MCP}}$	1.4	251.7	0.7	74.9	0.3	26.5
400	TT	$\hat{\theta}_{\text{IPW2}}$	2.3	129.2	1.2	40.3	0.6	15.8
		$\hat{\theta}_{\text{PEL}}$	2.3	129.2	1.2	40.3	0.6	15.8
		$\hat{\theta}_{\text{AIPW2}}$	2.6	126.1	1.4	38.0	0.8	13.7
		$\hat{\theta}_{\text{MCP}}$	2.6	125.8	1.4	37.9	0.8	13.6
	TF	$\hat{\theta}_{\text{AIPW2}}$	2.1	128.5	0.9	39.9	0.3	15.4
		$\hat{\theta}_{\text{MCP}}$	2.0	128.2	0.9	39.7	0.3	15.3
	FT	$\hat{\theta}_{\text{IPW2}}$	-38.5	245.3	-39.6	171.6	-40.1	152.9
		$\hat{\theta}_{\text{PEL}}$	-38.5	245.3	-39.6	171.6	-40.1	152.9
		$\hat{\theta}_{\text{AIPW2}}$	2.5	122.8	1.4	37.0	0.8	13.3
		$\hat{\theta}_{\text{MCP}}$	2.5	124.0	1.3	37.3	0.8	13.4

that they are doubly robust to the misspecification of either the propensity score model or the outcome regression models; (5) the intervals $\mathcal{I}_{\text{AIPW2B}}$ and $\mathcal{I}_{\text{MCPB}}$ are usually wider than the other two with higher coverage probabilities, which is an advantage of the bootstrap-calibrated methods in some settings (such as $(t, n) = (0.3, 200)$) where the coverage probabilities of the other two are close to 92%, slightly low compared to the nominal level of 95%; (6) between the two bootstrap-calibrated methods, $\mathcal{I}_{\text{MCPB}}$ is usually wider than $\mathcal{I}_{\text{AIPW2B}}$ with higher coverage probabilities; (7) when the sample size is small, for scenario "TT", as ρ increases, the coverage probability of $\mathcal{I}_{\text{MCP}}$ reaches to the nominal level of 95% faster than $\mathcal{I}_{\text{AIPW2}}$; for example, when $\rho = 0.7$ and $(n, t) = (100, 0.3)$, the coverage probability of $\mathcal{I}_{\text{MCP}}$ is 95.2%, while that of $\mathcal{I}_{\text{AIPW2}}$ is 92.7%; however, in other scenarios ("TF" and "FT"), this advantage disappears; (8) all confidence intervals based on doubly robust methods are similar to each other and perform well when n is large and get narrower as n increases.

5.2. Hypothesis tests

The basic hypothesis test problem in causal inference is to test whether there is a treatment effect, i.e., to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The power of a test (at the 0.05 level) can be computed through the rejection rate of the corresponding 95% confidence interval not containing zero over repeated samples. Specifically, for a test method corresponding to a confidence interval $\mathcal{I}$, we have

$$\text{Rejection Rate} = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} I\left(0 \notin \mathcal{I}^{(s)}\right) .$$

We consider four test methods based on the confidence intervals $\mathcal{I}_{\text{AIPW2}}$, $\mathcal{I}_{\text{AIPW2B}}$, $\mathcal{I}_{\text{MCP}}$ and $\mathcal{I}_{\text{MCPB}}$, respectively. When the true value of θ equals zero, the rejection rate is an approximation to the type I error probability; otherwise, the rejection rate is an approximation to the power of the test for a given value of the true ATE, θ^0.

The sample data are generated following the same setup in Section 5.1, except that the two outcome regression models are modified as

$$m_1 : \quad Y_{1j} = \theta^0 + 4.5 + x_{j1} - 2x_{j2} + 3x_{j3} + a_1\epsilon_j$$

and

$$m_0 : \quad Y_{0j} = 3.88 + x_{j1} + x_{j2} + 2x_{j3} + a_0\epsilon_j ,$$

where θ^0 is the assigned value of the true ATE. The setting allows θ^0 to vary from 0 to 3 to show the pattern of the power function of the test.

Figure 1 depicts the power functions of tests for different sample sizes and model specifications with $t = 0.5$ and $\rho = 0.5$. For $n = 100$, it is clear that the two curves for the bootstrap-calibrated confidence intervals are below the other two, implying that they have smaller type I error probabilities and test powers, though there is not much difference between all four curves. If the sample size increases, the four curves get closer and become indistinguishable when $n = 400$. The test powers of the four methods all go up as θ^0 departs from 0 for all the cases, meaning that they are working well; especially when $n = 400$, the test powers increase to 1 rapidly. The test using $\mathcal{I}_{\text{AIPW2}}$ seems to have larger powers than others in many cases. However, a further examination reveals that it also has inflated type I error probabilities in some cases, as shown in Figure 2.

5.3. Multiply robust estimators

One of the advantages of the constrained optimization formulation is the flexibility of including additional constraints. For general doubly robust estimation, a working model is required for propensity scores and for the outcome regression. When a set of feasible working models can be considered for the propensity

Table 4
%CP and AL(×100) for 95% Confidence Intervals when $t = 0.3$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			%CP	AL	%CP	AL	%CP	AL
100	TT	$\mathcal{I}_{\text{IPW2}}$	93.4	916.6	93.6	518.2	92.4	331.2
		$\mathcal{I}_{\text{PELR}}$	93.9	926.0	93.7	522.6	92.7	333.3
		$\mathcal{I}_{\text{AIPW2}}$	92.1	938.2	92.4	514.0	92.7	307.5
		$\mathcal{I}_{\text{AIPW2B}}$	95.5	1109.7	96.1	607.0	96.2	361.5
		$\mathcal{I}_{\text{MCP}}$	93.0	974.7	93.5	570.8	95.2	386.5
		$\mathcal{I}_{\text{MCPB}}$	97.0	1241.0	96.9	677.4	97.2	404.0
	TF	$\mathcal{I}_{\text{AIPW2}}$	93.7	915.8	93.0	515.2	92.1	325.8
		$\mathcal{I}_{\text{AIPW2B}}$	95.8	1005.1	95.8	568.5	94.7	365.0
		$\mathcal{I}_{\text{MCP}}$	93.6	933.0	92.6	529.1	91.4	340.0
		$\mathcal{I}_{\text{MCPB}}$	96.4	1066.8	96.2	607.2	95.3	396.4
	FT	$\mathcal{I}_{\text{IPW2}}$	91.1	902.0	85.8	516.3	75.5	340.3
		$\mathcal{I}_{\text{PELR}}$	91.5	910.7	86.1	521.3	75.2	343.8
		$\mathcal{I}_{\text{AIPW2}}$	92.4	950.3	93.0	520.6	92.9	311.2
		$\mathcal{I}_{\text{AIPW2B}}$	95.5	1098.2	95.5	600.8	95.5	357.9
		$\mathcal{I}_{\text{MCP}}$	90.9	895.6	90.5	493.7	92.2	301.1
		$\mathcal{I}_{\text{MCPB}}$	96.9	1249.6	96.8	683.2	97.0	407.3
200	TT	$\mathcal{I}_{\text{IPW2}}$	91.6	664.2	92.0	378.1	90.9	243.9
		$\mathcal{I}_{\text{PELR}}$	91.6	668.3	92.1	380.3	91.1	245.2
		$\mathcal{I}_{\text{AIPW2}}$	91.9	667.2	92.0	365.5	91.9	218.5
		$\mathcal{I}_{\text{AIPW2B}}$	93.8	714.8	93.8	391.3	94.0	233.5
		$\mathcal{I}_{\text{MCP}}$	91.6	693.8	92.4	402.8	94.6	269.4
		$\mathcal{I}_{\text{MCPB}}$	94.7	750.1	94.4	411.0	94.7	244.9
	TF	$\mathcal{I}_{\text{AIPW2}}$	91.6	662.5	91.7	376.0	89.5	241.1
		$\mathcal{I}_{\text{AIPW2B}}$	93.0	697.9	92.8	400.6	91.9	262.9
		$\mathcal{I}_{\text{MCP}}$	91.8	680.4	91.5	387.0	89.3	250.2
		$\mathcal{I}_{\text{MCPB}}$	93.9	713.9	93.3	410.1	92.9	273.0
	FT	$\mathcal{I}_{\text{IPW2}}$	89.2	647.2	78.3	370.1	60.4	243.6
		$\mathcal{I}_{\text{PELR}}$	89.3	650.9	78.4	372.2	60.7	245.0
		$\mathcal{I}_{\text{AIPW2}}$	92.2	670.4	92.4	367.2	92.5	219.5
		$\mathcal{I}_{\text{AIPW2B}}$	93.7	707.5	93.4	387.3	94.0	231.2
		$\mathcal{I}_{\text{MCP}}$	91.2	644.7	91.3	355.2	91.8	214.9
		$\mathcal{I}_{\text{MCPB}}$	94.7	755.2	95.1	414.3	95.2	246.8
400	TT	$\mathcal{I}_{\text{IPW2}}$	94.0	478.8	93.9	272.3	93.3	176.1
		$\mathcal{I}_{\text{PELR}}$	94.1	480.4	93.9	273.2	93.3	176.7
		$\mathcal{I}_{\text{AIPW2}}$	94.0	473.2	94.3	259.2	94.4	155.0
		$\mathcal{I}_{\text{AIPW2B}}$	95.3	488.9	95.3	267.7	95.2	159.9
		$\mathcal{I}_{\text{MCP}}$	94.2	481.6	94.8	276.3	95.9	180.7
		$\mathcal{I}_{\text{MCPB}}$	95.3	499.0	96.2	273.1	95.7	163.1
	TF	$\mathcal{I}_{\text{AIPW2}}$	94.0	477.3	94.3	270.5	93.3	174.0
		$\mathcal{I}_{\text{AIPW2B}}$	94.5	491.6	95.0	281.9	94.5	185.4
		$\mathcal{I}_{\text{MCP}}$	94.2	479.5	94.1	272.9	93.0	176.5
		$\mathcal{I}_{\text{MCPB}}$	95.4	498.0	95.3	285.7	94.7	188.5
	FT	$\mathcal{I}_{\text{IPW2}}$	83.8	457.4	62.6	261.7	31.4	172.3
		$\mathcal{I}_{\text{PELR}}$	84.0	458.7	62.6	262.4	31.7	172.9
		$\mathcal{I}_{\text{AIPW2}}$	94.1	473.4	93.8	259.3	94.3	155.0
		$\mathcal{I}_{\text{AIPW2B}}$	95.0	483.4	94.9	264.8	94.8	158.2
		$\mathcal{I}_{\text{MCP}}$	92.8	457.0	93.4	251.6	94.0	151.8
		$\mathcal{I}_{\text{MCPB}}$	95.2	502.6	95.3	275.2	95.3	164.1

Table 5
%CP and AL(×100) for 95% Confidence Intervals when $t = 0.5$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			%CP	AL	%CP	AL	%CP	AL
100	TT	$\mathcal{I}_{\text{IPW2}}$	94.0	819.8	93.5	459.9	93.6	288.5
		$\mathcal{I}_{\text{PELR}}$	94.2	827.1	93.7	463.2	93.6	289.6
		$\mathcal{I}_{\text{AIPW2}}$	93.6	822.0	93.7	451.0	93.4	270.7
		$\mathcal{I}_{\text{AIPW2B}}$	95.5	873.4	95.0	478.8	95.2	286.8
		$\mathcal{I}_{\text{MCP}}$	94.4	843.2	95.0	479.4	95.8	311.8
		$\mathcal{I}_{\text{MCPB}}$	96.9	927.8	96.7	508.7	96.3	304.9
	TF	$\mathcal{I}_{\text{AIPW2}}$	94.2	818.8	93.5	458.0	93.4	285.6
		$\mathcal{I}_{\text{AIPW2B}}$	94.7	860.6	95.0	484.4	95.0	307.3
		$\mathcal{I}_{\text{MCP}}$	94.4	828.7	93.7	463.6	93.4	290.2
		$\mathcal{I}_{\text{MCPB}}$	96.2	894.6	96.4	504.0	95.9	321.3
	FT	$\mathcal{I}_{\text{IPW2}}$	92.9	811.2	86.5	468.1	73.4	313.4
		$\mathcal{I}_{\text{PELR}}$	92.9	818.0	86.8	471.9	73.6	316.1
		$\mathcal{I}_{\text{AIPW2}}$	94.5	821.7	94.2	450.8	94.0	270.6
		$\mathcal{I}_{\text{AIPW2B}}$	95.7	862.9	95.3	473.1	95.5	283.5
		$\mathcal{I}_{\text{MCP}}$	93.4	804.9	93.8	442.4	93.5	266.8
		$\mathcal{I}_{\text{MCPB}}$	97.0	924.7	96.6	508.0	96.6	304.5
200	TT	$\mathcal{I}_{\text{IPW2}}$	94.8	583.4	94.8	326.2	94.2	204.0
		$\mathcal{I}_{\text{PELR}}$	95.2	586.1	95.2	327.6	94.4	204.9
		$\mathcal{I}_{\text{AIPW2}}$	95.2	581.4	95.5	319.1	95.5	191.7
		$\mathcal{I}_{\text{AIPW2B}}$	95.9	598.8	96.1	328.5	96.0	197.2
		$\mathcal{I}_{\text{MCP}}$	95.5	587.6	95.6	331.4	96.2	211.0
		$\mathcal{I}_{\text{MCPB}}$	96.6	614.5	97.0	337.2	96.4	202.2
	TF	$\mathcal{I}_{\text{AIPW2}}$	94.6	582.3	94.7	325.2	94.1	203.0
		$\mathcal{I}_{\text{AIPW2B}}$	95.4	599.7	95.3	338.4	95.0	215.6
		$\mathcal{I}_{\text{MCP}}$	94.8	584.7	94.5	326.7	93.8	204.3
		$\mathcal{I}_{\text{MCPB}}$	95.8	610.4	95.9	344.0	95.5	219.3
	FT	$\mathcal{I}_{\text{IPW2}}$	88.8	575.6	76.6	332.2	51.4	222.4
		$\mathcal{I}_{\text{PELR}}$	88.7	578.2	76.8	333.7	52.0	223.5
		$\mathcal{I}_{\text{AIPW2}}$	95.9	581.6	95.6	319.2	95.7	191.7
		$\mathcal{I}_{\text{AIPW2B}}$	96.4	593.3	96.8	325.5	96.4	195.4
		$\mathcal{I}_{\text{MCP}}$	95.1	571.7	94.8	314.4	95.2	189.3
		$\mathcal{I}_{\text{MCPB}}$	96.8	614.8	96.8	337.3	96.5	202.2
400	TT	$\mathcal{I}_{\text{IPW2}}$	92.9	413.0	92.5	230.8	92.5	144.6
		$\mathcal{I}_{\text{PELR}}$	93.0	414.0	92.7	231.3	92.5	144.9
		$\mathcal{I}_{\text{AIPW2}}$	92.7	410.9	92.7	225.5	92.7	135.4
		$\mathcal{I}_{\text{AIPW2B}}$	93.4	416.3	93.0	228.4	93.0	137.1
		$\mathcal{I}_{\text{MCP}}$	92.9	412.4	93.2	229.4	93.5	142.2
		$\mathcal{I}_{\text{MCPB}}$	93.8	422.5	93.7	232.0	93.2	139.2
	TF	$\mathcal{I}_{\text{AIPW2}}$	93.1	412.9	92.7	230.5	92.6	144.1
		$\mathcal{I}_{\text{AIPW2B}}$	93.4	418.7	93.1	235.5	93.2	149.4
		$\mathcal{I}_{\text{MCP}}$	92.8	413.5	92.6	230.9	92.4	144.5
		$\mathcal{I}_{\text{MCPB}}$	93.6	423.7	93.5	238.2	93.3	151.2
	FT	$\mathcal{I}_{\text{IPW2}}$	81.5	407.5	54.9	235.3	23.9	157.8
		$\mathcal{I}_{\text{PELR}}$	81.6	408.5	54.9	235.9	23.9	158.2
		$\mathcal{I}_{\text{AIPW2}}$	93.1	410.3	93.4	225.2	93.1	135.2
		$\mathcal{I}_{\text{AIPW2B}}$	93.5	413.7	93.5	227.0	93.2	136.3
		$\mathcal{I}_{\text{MCP}}$	92.4	404.5	92.5	222.4	92.5	133.8
		$\mathcal{I}_{\text{MCPB}}$	93.9	422.7	93.7	231.9	93.3	139.2

Table 6
%CP and AL(×100) for 95% Confidence Intervals when $t = 0.7$

n	Scenario	Estimator	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
			$\%CP$	AL	$\%CP$	AL	$\%CP$	AL
100	TT	$\mathcal{I}_{\text{IPW2}}$	93.2	852.5	93.7	474.9	93.4	294.8
		$\mathcal{I}_{\text{PELR}}$	93.6	860.7	93.9	479.1	93.4	297.2
		$\mathcal{I}_{\text{AIPW2}}$	92.9	847.8	92.8	465.1	93.3	279.0
		$\mathcal{I}_{\text{AIPW2B}}$	94.8	917.3	94.9	502.8	94.9	300.9
		$\mathcal{I}_{\text{MCP}}$	92.8	864.9	93.6	485.8	94.7	307.1
		$\mathcal{I}_{\text{MCPB}}$	96.1	979.7	96.1	536.7	96.1	321.0
	TF	$\mathcal{I}_{\text{AIPW2}}$	93.0	849.1	93.6	471.0	93.6	289.7
		$\mathcal{I}_{\text{AIPW2B}}$	94.8	913.1	95.1	510.4	95.5	319.6
		$\mathcal{I}_{\text{MCP}}$	93.2	858.7	93.6	478.4	93.7	296.8
		$\mathcal{I}_{\text{MCPB}}$	96.2	959.4	95.4	535.8	96.2	335.9
	FT	$\mathcal{I}_{\text{IPW2}}$	90.1	851.4	83.0	498.7	71.9	342.7
		$\mathcal{I}_{\text{PELR}}$	90.3	859.5	83.6	503.5	72.5	346.3
		$\mathcal{I}_{\text{AIPW2}}$	93.7	851.5	93.7	467.1	93.9	280.2
		$\mathcal{I}_{\text{AIPW2B}}$	94.7	907.6	95.0	497.5	95.3	297.8
		$\mathcal{I}_{\text{MCP}}$	93.2	838.1	93.5	464.7	94.3	285.6
		$\mathcal{I}_{\text{MCPB}}$	95.8	977.9	96.0	535.8	96.2	320.5
200	TT	$\mathcal{I}_{\text{IPW2}}$	94.0	610.9	95.4	338.3	94.7	207.5
		$\mathcal{I}_{\text{PELR}}$	94.2	614.1	95.4	339.9	94.6	208.5
		$\mathcal{I}_{\text{AIPW2}}$	93.7	608.7	94.1	333.8	94.2	200.2
		$\mathcal{I}_{\text{AIPW2B}}$	94.7	626.6	94.9	343.5	95.0	205.8
		$\mathcal{I}_{\text{MCP}}$	93.7	613.3	94.5	339.8	95.1	208.7
		$\mathcal{I}_{\text{MCPB}}$	95.3	645.3	95.3	353.8	95.6	212.0
	TF	$\mathcal{I}_{\text{AIPW2}}$	93.8	609.4	94.9	336.4	94.8	204.9
		$\mathcal{I}_{\text{AIPW2B}}$	95.2	627.5	95.6	348.4	95.5	214.9
		$\mathcal{I}_{\text{MCP}}$	94.5	612.7	95.0	338.6	94.8	207.0
		$\mathcal{I}_{\text{MCPB}}$	95.5	642.5	96.1	356.3	96.1	219.7
	FT	$\mathcal{I}_{\text{IPW2}}$	88.7	608.9	75.3	356.7	52.9	245.1
		$\mathcal{I}_{\text{PELR}}$	88.9	612.0	75.4	358.6	53.3	246.6
		$\mathcal{I}_{\text{AIPW2}}$	94.3	608.0	94.6	333.4	94.9	199.9
		$\mathcal{I}_{\text{AIPW2B}}$	95.3	622.1	95.1	341.1	95.4	204.4
		$\mathcal{I}_{\text{MCP}}$	94.0	599.7	94.3	331.3	95.2	201.7
		$\mathcal{I}_{\text{MCPB}}$	95.4	644.3	95.6	353.1	96.0	211.5
400	TT	$\mathcal{I}_{\text{IPW2}}$	94.5	434.7	94.8	241.0	94.4	148.0
		$\mathcal{I}_{\text{PELR}}$	94.5	435.9	94.8	241.6	94.4	148.4
		$\mathcal{I}_{\text{AIPW2}}$	94.2	432.3	94.4	237.1	94.1	142.1
		$\mathcal{I}_{\text{AIPW2B}}$	94.7	437.6	94.6	239.9	94.5	143.8
		$\mathcal{I}_{\text{MCP}}$	94.4	436.2	94.7	241.8	94.6	148.1
		$\mathcal{I}_{\text{MCPB}}$	95.0	444.7	95.1	243.9	95.0	146.2
	TF	$\mathcal{I}_{\text{AIPW2}}$	94.4	434.2	94.5	240.3	94.4	146.9
		$\mathcal{I}_{\text{AIPW2B}}$	94.9	440.1	94.9	244.7	94.6	151.0
		$\mathcal{I}_{\text{MCP}}$	94.5	435.6	94.7	241.2	94.3	147.7
		$\mathcal{I}_{\text{MCPB}}$	95.0	446.2	95.2	247.7	95.6	152.8
	FT	$\mathcal{I}_{\text{IPW2}}$	82.8	431.1	55.9	252.4	24.7	173.4
		$\mathcal{I}_{\text{PELR}}$	82.9	432.3	56.0	253.1	25.2	173.9
		$\mathcal{I}_{\text{AIPW2}}$	94.6	430.9	94.7	236.3	94.5	141.7
		$\mathcal{I}_{\text{AIPW2B}}$	94.6	434.7	94.7	238.3	94.9	142.8
		$\mathcal{I}_{\text{MCP}}$	94.2	425.7	94.5	234.7	94.7	142.3
		$\mathcal{I}_{\text{MCPB}}$	94.9	444.3	94.9	243.3	95.3	145.8

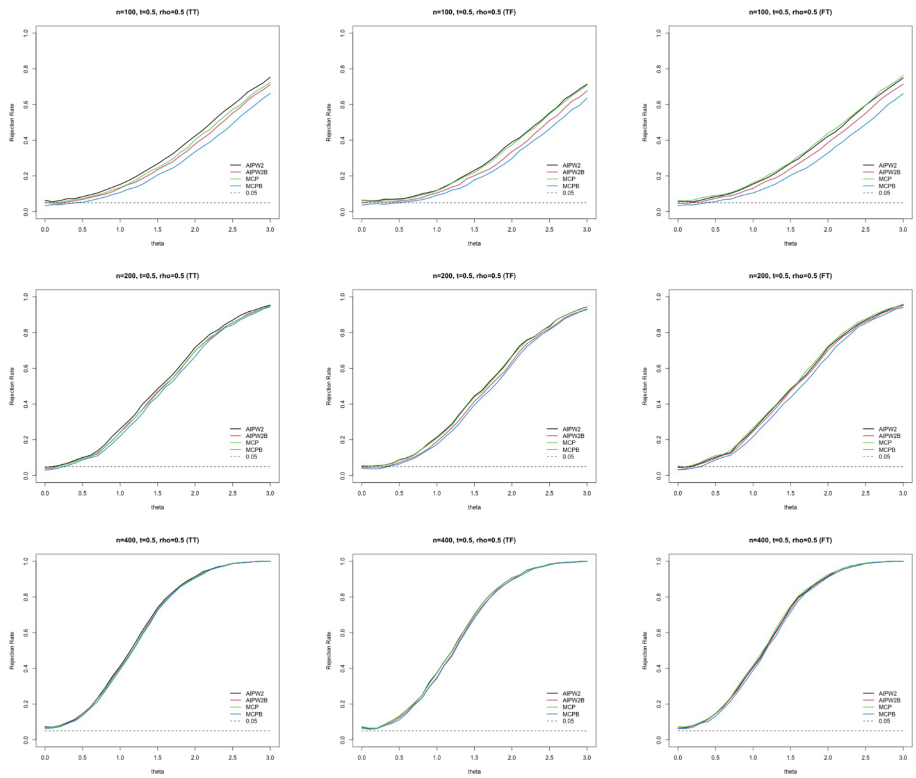

Fig 1. *Power functions of tests when* $t = 0.5$ *and* $\rho = 0.5$

scores and sets of models are available for the outcome regression, the so-called multiply robust estimators can be constructed by including additional model-calibration constraints. See Han and Wang (2013) and Zhang et al. (2023) for further discussions.

Our proposed pseudo-EL approach to causal inference allows the use of multiple working models for the outcome regression. We demonstrate the procedure through a simulation study. The sample data are generated following the same mechanisms described in Section 5.1. Table 7 gives the scenarios for model specifications.

Table 7
Model Specification Scenarios.

Model	TTF	TFF	FTF	FFF
PS	T	T	F	F
REGRESSION	TF	FF	TF	FF

A single working model is used for propensity scores (PS), with the true model (T) including all three x variables and the misspecified model (F) without having x_3. A set of two working models is used for the outcome regression $Y_1 \mid \boldsymbol{x}$ and

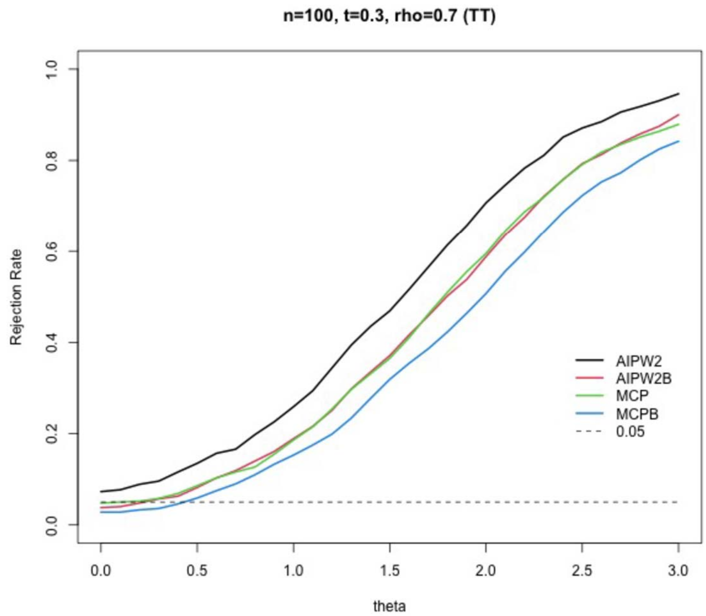

Fig 2. *Power functions of tests with* $(n, t, \rho) = (100, 0.3, 0.7)$ *and "TT".*

a set of two working models is used for the outcome regression $Y_0 \mid \boldsymbol{x}$. The case for which each set contains one correct model with all three x variables and one misspecified model by omitting x_3 is denoted as TF for "REGRESSION", and the case for which each set contains two misspecified regression models, one by omitting x_2 and the other by omitting x_3, is represented by FF. It leads to four scenarios of model specifications for PS and REGRESSION, denoted as TTF, TFF, FTF, and FFF. The performance of the model-calibrated maximum PEL estimator is examined in terms of the percent relative bias and mean square errors and the results are reported in table 8. It can be seen that the biases are negligible when either the PS model is correctly specified or one of working regression models is correctly specified. When all three sets of models are misspecified, the biases become bigger but are still quite acceptable in most cases.

6. A real data example

We now apply the methods discussed in this paper to examine the average treatment effect of maternal smoking during pregnancy on infants' birth weight using a dataset initially investigated by Almond et al. (2005) and subsequently analyzed by Cattaneo (2010), Liu et al. (2018), and Ghosh et al. (2021). The data comprises detailed observations on 4642 infants in Pennsylvania, USA, capturing various maternal and family background factors that could influence

TABLE 8
%RB and MSE(×100) for Model-Calibrated PEL Estimator.

t	n	Scenario	$\rho=0.3$ %RB	$\rho=0.3$ MSE	$\rho=0.5$ %RB	$\rho=0.5$ MSE	$\rho=0.7$ %RB	$\rho=0.7$ MSE
0.3	100	TTF	-1.8	694.6	-1.0	206.7	-0.6	72.9
		TFF	-3.1	670.2	-2.8	200.4	-2.5	70.6
		FTF	-1.7	728.4	-1.0	217.3	-0.7	76.9
		FFF	-6.0	682.8	-6.0	209.2	-5.6	76.0
	200	TTF	-2.4	340.6	-1.3	100.3	-1.0	36.3
		TFF	-3.4	334.5	-2.4	99.5	-2.0	36.4
		FTF	-2.4	354.7	-1.5	106.9	-1.0	38.3
		FFF	-7.0	342.1	-5.8	108.1	-5.3	40.9
	400	TTF	-2.3	156.3	-1.2	46.7	-0.7	16.6
		TFF	-2.9	156.3	-1.8	46.9	-1.3	16.8
		FTF	-2.2	164.5	-1.2	49.1	-0.7	17.4
		FFF	-6.5	164.6	-5.2	51.2	-4.5	19.2
0.5	100	TTF	2.1	455.5	1.2	138.2	0.7	50.4
		TFF	1.4	450.7	0.5	137.5	0.0	50.6
		FTF	2.1	458.2	1.2	139.0	0.7	50.7
		FFF	-2.9	447.4	-3.0	139.0	-3.1	52.4
	200	TTF	2.7	224.6	1.4	67.7	0.7	24.5
		TFF	2.2	224.6	0.8	67.8	0.2	24.5
		FTF	2.8	228.6	1.4	69.0	0.7	24.9
		FFF	-1.9	227.4	-2.8	70.1	-3.1	25.8
	400	TTF	-0.4	121.5	-0.2	36.8	-0.1	13.4
		TFF	-0.6	121.6	-0.5	36.9	-0.4	13.4
		FTF	-0.4	123.1	-0.2	37.3	-0.1	13.5
		FFF	-4.2	124.1	-3.6	38.3	-3.4	14.5
0.7	100	TTF	-0.8	549.2	-0.5	165.7	-0.3	59.7
		TFF	-1.3	545.8	-0.9	165.3	-0.8	59.7
		FTF	-0.8	547.7	-0.5	165.3	-0.3	59.6
		FFF	-5.5	551.2	-4.4	170.3	-3.7	62.3
	200	TTF	1.7	256.9	0.8	76.4	0.3	27.0
		TFF	1.2	256.8	0.4	76.5	0.0	27.2
		FTF	1.5	257.1	0.7	76.3	0.3	27.0
		FFF	-3.6	255.7	-3.3	78.3	-3.1	28.4
	400	TTF	2.6	125.8	1.4	37.9	0.8	13.6
		TFF	2.5	125.9	1.3	38.0	0.7	13.7
		FTF	2.6	126.2	1.4	38.0	0.8	13.7
		FFF	-1.4	127.1	-1.8	38.8	-2.1	14.3

birth weights. Key variables include the infant's birth weight (in grams), an indicator for low birth-weight babies (defined as birth weight $<$ 2500g), and the maternal smoking status (1=Smoking, 0=Non-smoking). Additionally, the data encompasses an array of covariates, such as mother's age, mother's marital status, indicators for alcohol consumption during pregnancy and previous infant mortality, as well as mother's education, father's education, number of prenatal care visits, mother's race, an indicator for the first-born child, and the months elapsed since the last birth. We designate the maternal smoking status as the treatment variable (T) and consider two types of outcomes: the infant's birth weight (continuous) and the indicator for low birth-weight babies (binary).

We consider the same propensity score and outcome regression models as in

Almond et al. (2005) where they revealed a strong negative impact of maternal smoking on infant birth weight, after controlling for other characteristics using propensity score subclassification and regression-adjusted methods. Table 9 reports the estimates of ATE when birth weight is the dependent variable, various approaches are considered including using difference in sample means (NAIVE), inverse probability weighting (IPW2 and AIPW2), and our proposed pseudo empirical likelihood methods (PEL and MCP). Wald-type and PEL ratio confidence intervals are reported for the IPW and PEL estimators respectively. Our findings align with the results in Almond et al. (2005). Ignoring the confounding factors inflates the impact of maternal smoking. The values of IPW2 and PEL estimators are identical as expected, indicating that maternal smoking reduces infant birth weight by approximately 235 grams. AIPW2 and the proposed MCP estimator are both doubly robust, their values are approximately equal indicating a slightly smaller reduction of 232 grams in birth weight due to maternal smoking, while MCP has a narrower 95% confidence interval. The bootstrap method yields wider confidence intervals in general.

TABLE 9
Estimated ATEs of maternal smoking on birth weight with confidence intervals.

Estimator	Estimate	95% CI	Bootstrap 95% CI
NAIVE	-275.25	-	-
IPW2	-235.42	(-287.25, -183.58)	-
PEL	-235.42	(-287.35, -183.48)	-
AIPW2	-231.96	(-285.89, -178.02)	(-288.57, -175.34)
MCP	-231.63	(-281.28, -181.97)	(-285.17, -178.07)

Table 10 is analogous to Table 9, but substitutes a binary indicator for birth weight below 2500 grams as the dependent variable. The ATE in this case, $E(Y_1) - E(Y_0) = P(Y_1 = 1) - P(Y_0 = 1)$, is the difference in risks of low birthweight. The results suggest that maternal smoking causes a significantly higher chance of having a low birth-weight baby. The increase in risk is estimated to be 5% based on IPW2 and PEL, and it is a bit higher as 5.7-5.8% according to the two doubly robust estimators while MCP again has a narrower confidence interval.

TABLE 10
Estimated ATEs of maternal smoking on low birth-weight baby with confidence intervals.

Estimator	Estimate	95% CI	Bootstrap 95% CI
NAIVE	0.061	-	-
IPW2	0.050	(0.024, 0.076)	-
PEL	0.050	(0.024, 0.077)	-
AIPW2	0.057	(0.029, 0.086)	(0.028, 0.087)
MCP	0.058	(0.032, 0.084)	(0.026, 0.090)

7. Additional remarks

Estimation methods discussed in this paper rely on the validity of the assumptions **A1** and **A2**. The most essential one, the SITA assumption **A1**, however, is not testable with the given sample data. A general guideline for practical use of the estimation methods is to include potential predictors for both the treatment assignment mechanism and the outcome variables during the data collection process. Moreover, Brookhart et al. (2006) demonstrated that incorporating outcome-related variables into the propensity score model can improve the efficiency of estimation while avoiding additional biases. The failure of the positivity assumption **A2** means certain subjects will have no chance to be included in the treatment group, a scenario similar to the under-coverage problem in survey sampling where some subjects have no chance to be included in the sample. Under-coverage problems are seemingly simple but notoriously difficult to address in survey sampling. They require additional information on the unknown population.

The simulation results seem to suggest that none of the methods uniformly outperforms the others for all scenarios. Our proposed PEL methods, however, have two major advantages. First, the maximum PEL estimators are obtained through a constrained maximization procedure, which allows the use of any suitable auxiliary information through the inclusion of additional constraints. When the target population is a well-defined finite population, auxiliary information may be available from census data in forms of known population counts or from reliable external sources such as existing large scale surveys. Calibration equations can readily be used to incorporate such information; see Chapters 6 and 8 of Wu and Thompson (2020) for further detail. Calibration with known population controls is also an effective strategy for dealing with undercoverage problems (Chen et al., 2023). Another direction of including additional constraints is to construct multiply robust estimators of the ATE by incorporating multiple working models (Han and Wang, 2013). This was briefly explored in simulation studies presented in Section 5.3, where multiple working regression models were used for model-calibration. The simulation results show that the model-calibrated PEL estimator is consistent if either the PS model or one of the working regression models for each of the potential outcomes is correctly specified. The simulated biases are acceptably small even if all models are misspecified. The concept of multiple robustness is very useful when one is uncertain about choices of regression models, for instance, whether certain interaction terms should be included in the model. All plausible models can then be considered.

Second, the PEL ratio confidence intervals are range-respecting and transformation invariant, which is an attractive feature for scenarios where, for instance, the true value of the ATE is confined within a particular range such as $[-1, 1]$. This is the case in the real data example presented in Section 6 where the outcome of interest is the binary indicator for low birth-weight, the ATE is the difference in risk and is confined within the range of $[-1, 1]$. The range-respecting property of our proposed PEL ratio confidence intervals becomes appealing. PEL ratio confidence intervals under the multiple robustness framework is a topic of interest for future investigations.

8. Appendix

8.1. Regularity conditions

Let $m_1(\boldsymbol{x}, \boldsymbol{\beta}_1)$ and $m_0(\boldsymbol{x}, \boldsymbol{\beta}_0)$ be respectively the mean functions of the outcome regression models for Y_1 and Y_0 given the covariate vector $\boldsymbol{x}$. The probability limit of the estimator $\hat{\boldsymbol{\beta}}_i$ for the vector of model parameters $\boldsymbol{\beta}_i$ under the assumed working model using the observed sample data is denoted as $\boldsymbol{\beta}_i^*$ for $i = 1, 0$.

R1 The treatment indicator T satisfies $E(T) = t \in (0, 1)$.
R2 The population satisfies $\mathrm{E}(Y_1^2) < \infty$, $\mathrm{E}(Y_0^2) < \infty$, and $\mathrm{E}(||\boldsymbol{x}||^2) < \infty$, where $|| \cdot ||$ denotes the l_2-norm.
R3 The population and the mean functions satisfy $\mathrm{E}\{m_1^2(\boldsymbol{x}, \boldsymbol{\beta}_1^*)\} < \infty$ and $\mathrm{E}\{m_0^2(\boldsymbol{x}, \boldsymbol{\beta}_0^*)\} < \infty$.
R4 For each $\boldsymbol{x}$, $\partial m_i(\boldsymbol{x}, \boldsymbol{\beta}_i)/\partial \boldsymbol{\beta}_i$ is continuous in $\boldsymbol{\beta}_i$ and $||\partial m_i(\boldsymbol{x}, \boldsymbol{\beta}_i)/\partial \boldsymbol{\beta}_i|| \leq h_i(\boldsymbol{x}, \boldsymbol{\beta}_i)$ for $\boldsymbol{\beta}_i$ in the neighborhood of $\boldsymbol{\beta}_i^*$, and $\mathrm{E}\{h_i(\boldsymbol{x}, \boldsymbol{\beta}_i^*)\} < \infty$, for $i = 0, 1$.
R5 For each $\boldsymbol{x}$, $\partial^2 m_i(\boldsymbol{x}, \boldsymbol{\beta}_i)/\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\top$ is continuous in $\boldsymbol{\beta}_i$ and $\max_{j,l} |\partial^2 m_i(\boldsymbol{x}, \boldsymbol{\beta}_i)/\partial \boldsymbol{\beta}_{ij} \partial \boldsymbol{\beta}_{il}| \leq k_i(\boldsymbol{x}, \boldsymbol{\beta}_i)$ for $\boldsymbol{\beta}_i$ in the neighborhood of $\boldsymbol{\beta}_i^*$, and $\mathrm{E}\{k_i(\boldsymbol{x}, \boldsymbol{\beta}_i^*)\} < \infty$, for $i = 0, 1$, where $| \cdot |$ denotes the absolute value.
R6 There exist c_1 and c_2 such that $0 < c_1 \leq \tau_j^0 \leq c_2 < 1$ for all units j, where $\tau_j^0 = \tau(\boldsymbol{x}_j; \boldsymbol{\alpha}^0)$ is the propensity score under the true propensity score model and $\boldsymbol{\alpha}^0$ is the true value for the vector of propensity score model parameters $\boldsymbol{\alpha}$.

Condition **R1** is commonly used in practice. Note that the sample size of the treatment group $n_1 = \sum_{j \in \mathcal{S}} T_j$ and then, $n_1/n = \sum_{j \in \mathcal{S}} T_j/n$, where n is the sample size. Under condition **R1**, we have n_1/n converges to $t \in (0, 1)$. Thus, there are no essential differences among $O_p(n_1^{-1/2})$, $O_p(n_0^{-1/2})$ and $O_p(n^{-1/2})$. Conditions **R2** and **R3** are the typical finite moment conditions. If the outcome regression models are linear, then $\mathrm{E}(||\boldsymbol{x}||^2) < \infty$ from **R2** implies **R3**. As stated in Lemma 11.2 of Owen (2001), if $Y_1, \ldots, Y_n$ are independent and identically distributed as the random variable Y, and $\mathrm{E}(Y^2) < \infty$, then $\max_{1 \leq i \leq n} |Y_i| = o_p(\sqrt{n})$. Conditions **R2** and **R3** play an important role in establishing the $O_p(n^{-1/2})$ order of the Lagrange multiplier for the constrained maximization problems. Conditions **R4** and **R5** are the usual smoothness and boundedness conditions (Wu and Sitter, 2001; Chen et al., 2020). They are automatically satisfied under linear outcome regression models. Condition **R6** states that each unit in the initial sample has a non-trivia probability of being assigned to either the treatment group or the control group.

8.2. Sandwich variance estimators for IPW estimators

Let $\boldsymbol{\Psi} = (\mu_1, \theta, \boldsymbol{\alpha}^\top)^\top$. The combination of estimators $\hat{\boldsymbol{\Psi}} = (\hat{\mu}_1, \hat{\theta}, \hat{\boldsymbol{\alpha}}^\top)^\top$, where $\hat{\boldsymbol{\alpha}}$ is the vector of estimators for the coefficients in the logistic regression model for propensity scores, is the solution to the estimating equations system

$$\mathbf{U}(\mathbf{\Psi}) = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} \frac{T_j(Y_{1j}-\mu_1)}{\tau_j} + \Delta_1\frac{T_j-\tau_j}{\tau_j} \\ \frac{(1-T_j)(Y_{0j}-(\mu_1-\theta))}{1-\tau_j} + \Delta_0\frac{\tau_j-T_j}{1-\tau_j} \\ \tilde{\boldsymbol{x}}_j(T_j-\tau_j) \end{pmatrix} = \frac{1}{n}\sum_{j\in\mathcal{S}}\mathbf{U}_j(\mathbf{\Psi}) = \mathbf{0}\,,$$

where $\tau_j = \tau(\boldsymbol{x}_j;\boldsymbol{\alpha}) = \{1+\exp(\tilde{\boldsymbol{x}}_j^\top\boldsymbol{\alpha})\}^{-1}\exp(\tilde{\boldsymbol{x}}_j^\top\boldsymbol{\alpha})$, as specified by logistic regression models. The Hájek-type IPW estimators $\hat{\mu}_1 = \hat{\mu}_{1\text{IPW2}}$ and $\hat{\theta} = \hat{\theta}_{\text{IPW2}}$ are obtained when setting $\mathbf{\Delta} = (\Delta_1,\Delta_0) = (0,0)$ in the estimating equations system. Similarly, the IPW estimators $\hat{\mu}_1 = \hat{\mu}_{1\text{IPW1}}$ and $\hat{\theta} = \hat{\theta}_{\text{IPW1}}$ can be derived by choosing $\mathbf{\Delta} = (\Delta_1,\Delta_0) = (\mu_1,\mu_1-\theta)$. This formulation is similar to the one used for proving Theorem 1 in Chen et al. (2020). Let $\mathbf{\Psi}^0 = (\mu_1^0,\theta^0,\{\boldsymbol{\alpha}^0\}^\top)^\top$ denote the vector of true values. Under the SITA assumption **A1** and the assumed propensity score model, we have $\mathrm{E}(\mathbf{U}(\mathbf{\Psi}^0)) = \mathbf{0}$ when $\mathbf{\Delta} = (0,0)$ or $\mathbf{\Delta} = (\mu_1,\mu_1-\theta)$. It follows that the estimator $\hat{\mathbf{\Psi}}$ is an m-estimator. Hence, $\hat{\mathbf{\Psi}} \overset{P}{\to} \mathbf{\Psi}^0$ (component-wise) under the regularity conditions **R2** and **R6**. Therefore, both types of IPW estimators are valid when the propensity score model is correctly specified. We have $\mathbf{U}(\hat{\mathbf{\Psi}}) = 0$ and $\hat{\mathbf{\Psi}} = \mathbf{\Psi}^0 + o_p(1)$ (component-wise).

Under the regularity conditions **R2** and **R6** and the fact that $\mathrm{E}(\mathbf{U}(\mathbf{\Psi}^0)) = \mathbf{0}$, we have $\mathbf{U}(\mathbf{\Psi}^0) = O_p(n^{-1/2})$. Taking the Taylor expansion to $\mathbf{U}(\hat{\mathbf{\Psi}})$ at $\mathbf{\Psi} = \mathbf{\Psi}^0$ yields

$$\hat{\mathbf{\Psi}} - \mathbf{\Psi}^0 = -\left[\mathbf{H}\left(\mathbf{\Psi}^0\right)\right]^{-1}\mathbf{U}\left(\mathbf{\Psi}^0\right) + o_p\left(n^{-1/2}\right) = -\left[\mathrm{E}\left\{\mathbf{H}\left(\mathbf{\Psi}^0\right)\right\}\right]^{-1}\mathbf{U}\left(\mathbf{\Psi}^0\right) + o_p\left(n^{-1/2}\right),$$

where $\mathbf{H}(\mathbf{\Psi}) = \partial\mathbf{U}(\mathbf{\Psi})/\partial\mathbf{\Psi}$. It follows that $\hat{\mathbf{\Psi}} - \mathbf{\Psi}^0 = O_p(n^{-1/2})$. Also, the theoretical asymptotic variance of $\hat{\mathbf{\Psi}}$ takes the sandwich form of

$$\mathrm{Var}\left(\hat{\mathbf{\Psi}}\right) = \left[\mathrm{E}\left\{\mathbf{H}\left(\mathbf{\Psi}^0\right)\right\}\right]^{-1}\mathrm{Var}\left\{\mathbf{U}\left(\mathbf{\Psi}^0\right)\right\}\left[\mathrm{E}\left\{\mathbf{H}\left(\mathbf{\Psi}^0\right)\right\}^\top\right]^{-1},$$

where $\mathrm{Var}\{\mathbf{U}(\mathbf{\Psi}^0)\} = n^{-1}\,\mathrm{E}\{\mathbf{U}_j(\mathbf{\Psi}^0)\mathbf{U}_j^\top(\mathbf{\Psi}^0)\}$ and $\mathbf{U}_j(\mathbf{\Psi})$ is defined as part of $\mathbf{U}(\mathbf{\Psi})$ at the beginning of this section. This implies that the variances of the IPW estimators $\hat{\theta}_{\text{IPW1}}$ and $\hat{\theta}_{\text{IPW2}}$ can be estimated by the second diagonal element of the corresponding sandwich variance estimator given by

$$\mathrm{var}\left(\hat{\mathbf{\Psi}}\right) = \left\{\mathbf{H}\left(\hat{\mathbf{\Psi}}\right)\right\}^{-1}\left\{\frac{1}{n^2}\sum_{j=1}^{n}\mathbf{U}_j\left(\hat{\mathbf{\Psi}}\right)\mathbf{U}_j^\top\left(\hat{\mathbf{\Psi}}\right)\right\}\left\{\mathbf{H}^\top\left(\hat{\mathbf{\Psi}}\right)\right\}^{-1},$$

where $\mathrm{var}(\cdot)$ denotes a variance estimator.

8.3. Proof of Proposition 1

Proof. Let $\mathbf{\Psi} = (\mu_1,\theta,\bar{m}_1,\bar{m}_0,\boldsymbol{\alpha}^\top,\boldsymbol{\beta}_1^\top,\boldsymbol{\beta}_0^\top)^\top$, where $\bar{m}_1$ and $\bar{m}_0$ are two additional nuisance parameters. The estimator $\hat{\mathbf{\Psi}} = (\hat{\mu}_1,\hat{\theta},\hat{\bar{m}}_1,\hat{\bar{m}}_0,\hat{\boldsymbol{\alpha}}^\top,\hat{\boldsymbol{\beta}}_1^\top,\hat{\boldsymbol{\beta}}_0^\top)^\top$ is the solution to the set estimating equations

$$\mathbf{U}(\boldsymbol{\Psi}) = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} \frac{T_j}{\tau_j}(Y_{1j} - m_{1j} + \bar{m}_1 - \mu_1) + \Delta_1 \frac{T_j - \tau_j}{\tau_j} \\ \frac{(1-T_j)}{1-\tau_j}(Y_{0j} - m_{0j} + \bar{m}_0 - (\mu_1 - \theta)) + \Delta_0 \frac{\tau_j - T_j}{1-\tau_j} \\ m_{1j} - \bar{m}_1 \\ m_{0j} - \bar{m}_0 \\ \tilde{\boldsymbol{x}}_j(T_j - \tau_j) \\ T_j(Y_{1j} - m_{1j})\tilde{\boldsymbol{x}}_j \\ (1-T_j)(Y_{0j} - m_{0j})\tilde{\boldsymbol{x}}_j \end{pmatrix} = \frac{1}{n}\sum_{j\in\mathcal{S}}\mathbf{U}_j(\boldsymbol{\Psi}) = \mathbf{0}\,. \tag{8.1}$$

Without loss of generality, we assume that the outcome regression models are linear. The arguments work for nonlinear outcome regression models.

The estimators $\hat{\mu}_1 = \hat{\mu}_{1\text{AIPW2}}$ and $\hat{\theta} = \hat{\theta}_{\text{AIPW2}}$ are obtained by setting $\boldsymbol{\Delta} = (\Delta_1, \Delta_0) = (0,0)$. Similarly, the estimators $\hat{\mu}_1 = \hat{\mu}_{1\text{AIPW1}}$ and $\hat{\theta} = \hat{\theta}_{\text{AIPW1}}$ can be derived by choosing $\boldsymbol{\Delta} = (\Delta_1, \Delta_0) = (\mu_1 - \bar{m}_1, \mu_1 - \theta - \bar{m}_0)$. Again, the formulation of (8.1) is similar to the one used for proving Theorem 1 in Chen et al. (2020).

Under the current framework, we have $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = O_p(n^{-1/2})$ for some value $\boldsymbol{\alpha}^*$ regardless of the correct specification of the propensity score model (White, 1982). Similarly, it is assumed that the estimator $\hat{\boldsymbol{\beta}}_i$ for $\boldsymbol{\beta}_i$ satisfies $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^* = O_p(n^{-1/2})$ for some value $\boldsymbol{\beta}_i^*$ regardless of the true outcome regression model, for $i = 0, 1$. Let μ_1^0 and θ^0 denote the true values of μ_1 and θ. Let $\boldsymbol{\Psi}^0 = (\mu_1^0, \theta^0, \mathrm{E}(m_{1j}^*), \mathrm{E}(m_{0j}^*),$
$\boldsymbol{\alpha}^{*\top}, \boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_0^{*\top})^\top$. We have $\mathrm{E}(\mathbf{U}(\boldsymbol{\Psi}^0)) = \mathbf{0}$ if either of the propensity score model and the set of outcome regression models is correctly specified. The estimator $\hat{\boldsymbol{\Psi}}$ is an m-estimator (Tsiatis, 2006) and $\hat{\theta}_{\text{AIPW1}}$ and $\hat{\theta}_{\text{AIPW2}}$ are doubly robust.

It follows that we have $\mathbf{U}(\hat{\boldsymbol{\Psi}}) = 0$ and $\hat{\boldsymbol{\Psi}} = \boldsymbol{\Psi}^0 + o_p(1)$ (component-wise) under the regularity conditions **R2-R3** and **R6** when one of the two sets of models is correctly specified.

By the central limit theorem and under regularity conditions **R2-R3** and **R6**, the result $\mathrm{E}(\mathbf{U}(\boldsymbol{\Psi}^0)) = \mathbf{0}$ leads to $\mathbf{U}(\boldsymbol{\Psi}^0) = O_p(n^{-1/2})$. Given that one of the two sets of models is correctly specified, taking the Taylor expansion of $\mathbf{U}(\hat{\boldsymbol{\Psi}})$ at $\boldsymbol{\Psi} = \boldsymbol{\Psi}^0$ yields

$$\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}^0 = -\left[\mathbf{H}\left(\boldsymbol{\Psi}^0\right)\right]^{-1}\mathbf{U}\left(\boldsymbol{\Psi}^0\right) + o_p\left(n^{-1/2}\right) = -\left[\mathrm{E}\left\{\mathbf{H}\left(\boldsymbol{\Psi}^0\right)\right\}\right]^{-1}\mathbf{U}\left(\boldsymbol{\Psi}^0\right) + o_p\left(n^{-1/2}\right),$$

where $\mathbf{H}(\boldsymbol{\Psi}) = \partial\mathbf{U}(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}$. The expansion leads to $\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}^0 = O_p(n^{-1/2})$. The theoretical asymptotic variance of $\hat{\boldsymbol{\Psi}}$ takes the sandwich form of

$$\mathrm{Var}\left(\hat{\boldsymbol{\Psi}}\right) = \left[\mathrm{E}\left\{\mathbf{H}\left(\boldsymbol{\Psi}^0\right)\right\}\right]^{-1}\mathrm{Var}\left\{\mathbf{U}\left(\boldsymbol{\Psi}^0\right)\right\}\left[\mathrm{E}\left\{\mathbf{H}\left(\boldsymbol{\Psi}^0\right)\right\}^\top\right]^{-1},$$

where $\mathrm{Var}\{\mathbf{U}(\boldsymbol{\Psi}^0)\} = n^{-1}\,\mathrm{E}\{\mathbf{U}_j(\boldsymbol{\Psi}^0)\mathbf{U}_j^\top(\boldsymbol{\Psi}^0)\}$. Consequently, the variances of the augmented IPW estimators $\hat{\theta}_{\text{AIPW1}}$ and $\hat{\theta}_{\text{AIPW2}}$ can be estimated by the second diagonal element of the corresponding sandwich variance estimator

$$\operatorname{var}\left(\hat{\boldsymbol{\Psi}}\right) = \left\{\mathbf{H}\left(\hat{\boldsymbol{\Psi}}\right)\right\}^{-1}\left\{\frac{1}{n^2}\sum_{j=1}^{n}\mathbf{U}_j\left(\hat{\boldsymbol{\Psi}}\right)\mathbf{U}_j^\top\left(\hat{\boldsymbol{\Psi}}\right)\right\}\left\{\mathbf{H}^\top\left(\hat{\boldsymbol{\Psi}}\right)\right\}^{-1},$$

where $\operatorname{var}(\cdot)$ denotes a variance estimator. Note that the aforementioned variance estimator for the augmented IPW $\hat{\theta}_{\text{AIPW1}}$ or $\hat{\theta}_{\text{AIPW2}}$ is doubly robust. As a matter of fact, the sandwich variance estimator is valid even if both sets of models are misspecified, since we can simply replace (μ_1^0, θ^0) in the above proof with the limit (μ_1^*, θ^*) under misspecified models. □

8.4. Proof of Theorem 1

Proof. Under the regularity condition **R1**, there is no need to distinguish among $O_p(n^{-1/2})$, $O_p(n_1^{-1/2})$, and $O_p(n_0^{-1/2})$. Following Wu and Yan (2012), the normalization constraint (C1) and the parameter constraint (C2) are equivalent to

$$\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} p_{ij} = 1\,,$$

$$\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} p_{ij}\boldsymbol{u}_{ij} = \mathbf{0}\,,$$

where $\boldsymbol{u}_{ij} = \boldsymbol{Z}_{ij} - \boldsymbol{\eta}$ with $\boldsymbol{Z}_{1j} = (1, Y_{1j}/w_1)^\top$, $\boldsymbol{Z}_{0j} = (0, -Y_{0j}/w_0)^\top$, and $\boldsymbol{\eta} = (w_1, \theta)^\top$.

Recall that $\hat{\boldsymbol{p}}_1(\theta) = (\hat{p}_{11}(\theta), \cdots, \hat{p}_{1n_1}(\theta))^\top$ and $\hat{\boldsymbol{p}}_0(\theta) = (\hat{p}_{01}(\theta), \cdots, \hat{p}_{0n_0}(\theta))^\top$ denote the maximizer of the joint pseudo-empirical likelihood function (3.1) subject to the normalization constraints (C1) and the parameter constraint (C2) for a fixed value of θ. Using the Lagrange multiplier method, we have $\hat{p}_{ij}(\theta) = \tilde{a}_{ij}/(1 + \hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij})$ for $i = 0, 1$ and $\hat{\boldsymbol{\lambda}}$ is the solution to

$$\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \frac{\tilde{a}_{ij}\boldsymbol{u}_{ij}}{(1 + \boldsymbol{\lambda}^\top \boldsymbol{u}_{ij})} = \mathbf{0}\,. \tag{8.2}$$

The above equations can be further rewritten as

$$\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij}\boldsymbol{u}_{ij} = \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \frac{\tilde{a}_{ij}\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^\top}{(1 + \boldsymbol{\lambda}^\top \boldsymbol{u}_{ij})}\boldsymbol{\lambda}\,.$$

Let $\boldsymbol{U} = \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij}\boldsymbol{u}_{ij} = (0, \hat{\theta}_{\text{IPW2}} - \theta)^\top$, which is of the order $O_p(n^{-1/2})$ under the assumed propensity score model if $\theta = \theta^0 + O_p(n^{-1/2})$. We have

$$\left\|\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \frac{\tilde{a}_{ij}\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^\top}{(1 + \boldsymbol{\lambda}^\top \boldsymbol{u}_{ij})}\right\| \geq \frac{1}{1 + \|\boldsymbol{\lambda}\| \max_{\{i,j\}} \|\boldsymbol{u}_{ij}\|}\left\|\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij}\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^\top\right\|,$$

where $||\cdot||$ denotes the ℓ_2-norm. The above results imply that

$$\|\boldsymbol{U}\| \geq \frac{1}{1+\|\boldsymbol{\lambda}\| \max_{\{i,j\}} \|\boldsymbol{u}_{ij}\|} \left\| \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^\top \right\| \|\boldsymbol{\lambda}\| \, .$$

It is known that $\|\boldsymbol{U}\| = O_p(n^{-1/2})$, and under the regularity condition **R2**, $\max_{\{i,j\}} \|\boldsymbol{u}_{ij}\| = o_p(n^{1/2})$ and $\| \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^\top \| = O_p(1)$. Therefore, we must have $\hat{\boldsymbol{\lambda}} = O_p(n^{-1/2})$. Since $\hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij} = o_p(1)$ uniformly for all i,j, we have an expansion to $\hat{\boldsymbol{\lambda}}$ as

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^\top \right)^{-1} \boldsymbol{U} + o_p(n^{-1/2}) = \boldsymbol{D}^{-1}\boldsymbol{U} + o_p(n^{-1/2}) \, ,$$

where $\boldsymbol{D} = \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^\top$ is a 2×2 matrix. The expansion to $\hat{\boldsymbol{\lambda}}$ is a crucial step to establish the asymptotic expansion to the PEL statistic $-2r_{\text{PEL}}(\theta)$, given below.

$$\begin{aligned}
-2r_{\text{PEL}}(\theta) &= 2n \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \log(1 + \hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij}) \\
&= 2n \sum_{i=0}^{1} w_i \sum_{j=1}^{n_i} \tilde{a}_{ij} \left(\hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij} - \frac{1}{2} \hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^\top \hat{\boldsymbol{\lambda}} \right) + o_p(1) \\
&= 2n \left(\hat{\boldsymbol{\lambda}}^\top \boldsymbol{U} - \frac{1}{2} \hat{\boldsymbol{\lambda}}^\top \boldsymbol{D} \hat{\boldsymbol{\lambda}} \right) + o_p(1) \\
&= n \boldsymbol{U}^\top \boldsymbol{D}^{-1} \boldsymbol{U} + o_p(1) \\
&= n \begin{pmatrix} 0 & \hat{\theta}_{\text{IPW2}} - \theta \end{pmatrix} \begin{pmatrix} d^{(11)} & d^{(12)} \\ d^{(21)} & d^{(22)} \end{pmatrix} \begin{pmatrix} 0 \\ \hat{\theta}_{\text{IPW2}} - \theta \end{pmatrix} + o_p(1) \\
&= n d^{(22)} (\hat{\theta}_{\text{IPW2}} - \theta)^2 + o_p(1) \, ,
\end{aligned} \tag{8.3}$$

where $d^{(22)}$ is the $(2,2)$-th element of $\boldsymbol{D}^{-1}$. The second equality in (8.3) holds since $\hat{\boldsymbol{\lambda}}^\top \boldsymbol{u}_{ij} = o_p(1)$ uniformly over i,j. A standard calculation yields that

$$d^{(22)} = \left\{ \theta^2 + 2 \Big[\sum_{j \in \mathcal{S}_1} \tilde{a}_{1j} Y_{1j}^2 + \sum_{j \in \mathcal{S}_0} \tilde{a}_{0j} Y_{0j}^2 - \hat{\theta}_{\text{IPW2}} \theta \Big] - (\hat{\mu}_{1\text{IPW2}} + \hat{\mu}_{0\text{IPW2}})^2 \right\}^{-1},$$

which converges in probability to

$$d_0^{(22)} = \left\{ 2[\mathrm{E}(Y_{1j}^2 + Y_{0j}^2) - (\mu_1^0)^2 - (\mu_0^0)^2] \right\}^{-1} ,$$

under the assumed propensity score model, with $\theta = \theta^0 = \mu_1^0 - \mu_0^0$.

By using the estimating equation techniques and the Taylor expansion, under the assumed propensity score model, we get

$$\hat{\theta}_{\text{IPW2}} - \theta^0 = \frac{1}{n} \sum_{j=1}^{n} \left[\frac{T_j}{\tau_j^0} (Y_{1j} - \mu_1^0) - \frac{1 - T_j}{1 - \tau_j^0} (Y_{0j} - \mu_0^0) - (\boldsymbol{J} - \boldsymbol{G}) \boldsymbol{C}^{-1} \tilde{\boldsymbol{x}}_j (T_j - \tau_j^0) \right] + o_p(n^{-1/2}) \, ,$$

where $\tau_j^0 = \tau(\tilde{\boldsymbol{x}}_j; \boldsymbol{\alpha}^0)$ is the propensity score under the true propensity score model with $\boldsymbol{\alpha}^0$ representing the true value of $\boldsymbol{\alpha}$, $\boldsymbol{C} = -\,\mathrm{E}[\tau_j^0(1-\tau_j^0)\tilde{\boldsymbol{x}}_j\tilde{\boldsymbol{x}}_j^\top]$, $\boldsymbol{J} = -\,\mathrm{E}[T_j(Y_{1j}-\mu_1^0)(1-\tau_j^0)\tilde{\boldsymbol{x}}_j^\top/\tau_j^0]$, and $\boldsymbol{G} = \mathrm{E}[(1-T_j)(Y_{0j}-\mu_0^0)(1-\tau_j^0)^{-1}\tau_j^0\tilde{\boldsymbol{x}}_j^\top]$.

It follows that, under the regularity conditions **R2** and **R6**, we have

$$\sqrt{n}\left(\hat{\theta}_{\text{IPW2}} - \theta^0\right) \xrightarrow{d} \mathrm{N}(0, V) \;\text{ as }\; n \to \infty\,,$$

where $V = n\,\mathrm{Var}(\hat{\theta}_{\text{IPW2}})$ and $\xrightarrow{d}$ denotes convergence in distribution. By Slutsky's theorem, we can conclude that under the assumed propensity score model,

$$-2r_{\text{PEL}}(\theta^0)/\hat{c} \xrightarrow{d} \chi_1^2\,,$$

where $\hat{c} = n\hat{d}_0^{(22)}\,\mathrm{var}(\hat{\theta}_{\text{IPW2}})$. This completes the proof. □

8.5. Proof of Theorem 2

Proof. We first derive an expression for the maximizer $\hat{\boldsymbol{p}}_1$. The original constrained optimization problem is equivalent to maximizing the pseudo-empirical likelihood function $nw_1\sum_{j\in\mathcal{S}_1}\tilde{a}_{1j}\log(p_{1j})$ under the normalization constraint $\sum_{j\in\mathcal{S}_1} p_{1j} = 1$ and the outcome regression model constraint $\sum_{j\in\mathcal{S}_1} p_{1j}\hat{m}_{1j} = \bar{\hat{m}}_1$. Under the normalization constraint, we can rewrite the model-calibration constraint as $\sum_{j\in\mathcal{S}_1} p_{1j}\hat{u}_{1j} = 0$. Let

$$\mathrm{L}(\boldsymbol{p}_1) = nw_1\sum_{j\in\mathcal{S}_1}\tilde{a}_{1j}\log(p_{1j}) - \lambda_{11}\left(\sum_{j\in\mathcal{S}_1} p_{1j} - 1\right) - \lambda_{12}\left(\sum_{j\in\mathcal{S}_1} p_{1j}\hat{u}_{1j}\right)\,,$$

where λ_{11} and λ_{12} are the Lagrange multipliers. Differentiating $\mathrm{L}(\boldsymbol{p}_1)$ with respect to p_{1j} and setting these equations to be zero leads to

$$\partial\mathrm{L}(\boldsymbol{p}_1)/\partial p_{1j} = nw_1\tilde{a}_{1j}/p_{1j} - \lambda_{11} - \lambda_{12}\hat{u}_{1j} = 0\,. \tag{8.4}$$

It follows that $nw_1\tilde{a}_{1j} - \hat{\lambda}_{11}\hat{p}_{1j} - \hat{\lambda}_{12}\hat{u}_{1j}\hat{p}_{1j} = 0$. By summing over all the subjects in $\mathcal{S}_1$, we get $\hat{\lambda}_{11} = nw_1$. Putting this back to (8.4) leads to

$$\hat{p}_{1j} = \frac{nw_1\tilde{a}_{1j}}{\hat{\lambda}_{11} + \hat{\lambda}_{12}\hat{u}_{1j}} = \frac{1}{\hat{\lambda}_{11}}\frac{nw_1\tilde{a}_{1j}}{1+\hat{\lambda}_1\hat{u}_{1j}} = \frac{\tilde{a}_{1j}}{1+\hat{\lambda}_1\hat{u}_{1j}}\,,$$

where $\hat{\lambda}_1 = \hat{\lambda}_{12}/\hat{\lambda}_{11}$ satisfies the model-calibration constraint $\sum_{j\in\mathcal{S}_1} p_{1j}\hat{u}_{1j} = 0$, i.e.,

$$\sum_{j\in\mathcal{S}_1}\frac{\tilde{a}_{1j}\hat{u}_{1j}}{1+\hat{\lambda}_1\hat{u}_{1j}} = 0\,. \tag{8.5}$$

(1) The propensity score model is correctly specified.

The equation (8.5) can be rewritten as

$$\sum_{j\in\mathcal{S}_1} \frac{\tilde{a}_{1j}\hat{u}_{1j}(1+\hat{\lambda}_1\hat{u}_{1j}-\hat{\lambda}_1\hat{u}_{1j})}{1+\hat{\lambda}_1\hat{u}_{1j}} = 0\,,$$

which implies

$$\hat{\lambda}_1 = \left(\sum_{j\in\mathcal{S}_1} \frac{\tilde{a}_{1j}\hat{u}_{1j}^2}{1+\hat{\lambda}_1\hat{u}_{1j}}\right)^{-1} \left(\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}\right). \tag{8.6}$$

On the other hand, it follows from (8.5) that

$$\left|\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}\right| = \left|\hat{\lambda}_1\right| \left|\sum_{j\in\mathcal{S}_1} \frac{\tilde{a}_{1j}\hat{u}_{1j}^2}{1+\hat{\lambda}_1\hat{u}_{1j}}\right| \geq \frac{\left|\hat{\lambda}_1\right|}{1+\left|\hat{\lambda}_1\right| u_1^*} \sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}^2\,,$$

where $u_1^* = \max_{j\in\mathcal{S}_1} |\hat{u}_{1j}|$. Now, we demonstrate that u_1^* is of order $o_p(n^{1/2})$. By the Taylor expansion of $\hat{u}_{1j}$ on $\hat{\boldsymbol{\beta}}_1$ around $\boldsymbol{\beta}_1^*$, under the regularity condition **R5**, we have

$$\hat{u}_{1j} = \hat{m}_{1j} - \bar{\hat{m}}_1 = (m_{1j}^* - \bar{m}_1^*) + \left(\frac{\partial(m_{1j}-\bar{m}_1)}{\partial\boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^*}\right)\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right) + o_p(1)\,,$$

where $\bar{m}_1 = n^{-1}\sum_{l=1}^n m_{1l}$ and $m_{1j}^* = m_1(\boldsymbol{x}_j, \boldsymbol{\beta}_1^*)$. Under the regularity conditions **R1** and **R3-R5**, it is true that $u_1^* = o_p(n^{1/2})$. Similarly, by Taylor expansions, we get that $|\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}|$ is of the order $O_p(n^{-1/2})$, and $\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}^2$ is of order $O_p(1)$ under the regularity conditions **R1** and **R3-R4**. Thus, we derive the order of $\hat{\lambda}_1$ to be $O_p(n^{-1/2})$, which further implies $\hat{\lambda}_1\hat{u}_{1j} = o_p(1)$ uniformly over all $j \in \mathcal{S}_1$. The above asymptotic analysis leads to

$$\hat{\lambda}_1 = \left(\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}^2\right)^{-1} \left(\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}\right) + o_p\left(n^{-1/2}\right).$$

Under regularity conditions **R1-R3**, the maximum PEL estimator for μ_1 has an asymptotic expansion given by

$$\begin{aligned}
\hat{\mu}_{1\text{MCP}} &= \sum_{j\in\mathcal{S}_1} \hat{p}_{1j}Y_{1j} = \sum_{j\in\mathcal{S}_1} \frac{\tilde{a}_{1j}}{1+\hat{\lambda}_1\hat{u}_{1j}} Y_{1j} \\
&= \sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\left(1-\hat{\lambda}_1\hat{u}_{1j}\right) Y_{1j} + o_p(n^{-1/2}) \\
&= \sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}Y_{1j} + \hat{B}\left\{\frac{1}{n}\sum_{j\in\mathcal{S}} \hat{m}_{1j} - \sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{m}_{1j}\right\} + o_p\left(n^{-1/2}\right),
\end{aligned}$$

where $\hat{B} = \{\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}^2\}^{-1}\{\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}Y_{1j}\}$. The expression for $\hat{\mu}_{1\text{MCP}}$ shows that

$$\hat{\mu}_{1\text{MCP}} = \hat{\mu}_{1\text{IPW2}} - \hat{B}\left(\sum_{j\in\mathcal{S}_1} \tilde{a}_{1j}\hat{u}_{1j}\right) + o_p(n^{-1/2}) = \mu_1^0 + o_p(1) ,$$

where $\hat{B} = \left\{\text{Var}\left(m_{1j}^*\right)\right\}^{-1} \text{Cov}\left(m_{1j}^*, Y_{1j}\right) + o_p(1) = O_p(1)$.

(2) The outcome regression model is correctly specified.

The maximum PEL estimator can be expressed as

$$\hat{\mu}_{1\text{MCP}} = \sum_{j\in\mathcal{S}_1} \hat{p}_{1j}Y_{1j} = \sum_{j\in\mathcal{S}_1} \hat{p}_{1j}\left(Y_{1j} - \hat{m}_{1j}\right) + \sum_{j\in\mathcal{S}_1} \hat{p}_{1j}\hat{m}_{1j} .$$

Applying the Taylor expansion to the first term of the right hand side of the above equation gives

$$\begin{aligned}\sum_{j\in\mathcal{S}_1} \hat{p}_{1j}\left(Y_{1j} - \hat{m}_{1j}\right) &= \sum_{j=1}^{n} \frac{T_j\tilde{a}_{1j}}{1+\hat{\lambda}_1\hat{u}_{1j}}\left(Y_{1j} - \hat{m}_{1j}\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} \frac{T_j\left(Y_{1j} - m_{1j}^*\right)}{\tau_j^*\left(1+\lambda_1^* u_{1j}^*\right)} \Big/ \frac{1}{n}\sum_{j=1}^{n}\frac{T_j}{\tau_j^*} + o_p(1) ,\end{aligned}$$

which converges in probability to

$$\text{E}\left(\frac{T_j\left(Y_{1j} - m_{1j}^*\right)}{\tau_j^*\left(1+\lambda_1^* u_{1j}^*\right)}\right) \Big/ \text{E}\left(\frac{T_j}{\tau_j^*}\right) ,$$

under regularity conditions **R1-R4**, where $\tau_j^* = \tau(\tilde{\boldsymbol{x}}_j, \boldsymbol{\alpha}^*)$, $\lambda_1^* = \lambda(\boldsymbol{\alpha}^*, \boldsymbol{\beta}_1^*)$, and $u_{1j}^* = m_{1j}^* - \bar{m}_1^*$. When the outcome regression model is correctly specified, by the law of total expectation conditional on $\boldsymbol{x}_j$, we get $\text{E}[\{\tau_j^*(1+\lambda_1^* u_{1j}^*)\}^{-1}\{T_j(Y_{1j} - m_{1j}^*)\}] = 0$. Therefore, $\hat{\mu}_{1\text{MCP}} = \sum_{j\in\mathcal{S}_1}\hat{p}_{1j}\hat{m}_{1j} + o_p(1)$.

On the other hand, from the model-calibration constraint, we obtain that

$$\sum_{j\in\mathcal{S}_1} \hat{p}_{1j}\hat{m}_{1j} = \frac{1}{n}\sum_{j\in\mathcal{S}} \hat{m}_{1j} .$$

Under the regularity condition **R4**, applying the Taylor expansion immediately leads to

$$\sum_{j\in\mathcal{S}_1} \hat{p}_{1j}\hat{m}_{1j} = \frac{1}{n}\sum_{j\in\mathcal{S}} m_{1j}^* + o_p(1) \xrightarrow{P} \text{E}\left(m_{1j}^*\right) = \mu_1^0 ,$$

where $\xrightarrow{P}$ denotes convergence in probability. Thus, when the outcome regression model is correct, $\hat{\mu}_{1\text{MCP}}$ is a consistent estimator of μ_1. □

8.6. Proof of Theorem 3

Proof. We first note that $\hat{p}_{ij}(\theta) = \tilde{a}_{ij}/\{1 + \hat{\boldsymbol{\lambda}}^\top \boldsymbol{g}_{ij}(\theta)\}$, where $\hat{\boldsymbol{\lambda}}$ is the solution to

$$\sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \frac{\tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)} = \mathbf{0}\,.$$

We also note that $\hat{\boldsymbol{\lambda}}$ is a function of θ, so it can be denoted as $\hat{\boldsymbol{\lambda}}(\theta)$. If the propensity score model is correctly specified, for $\theta = \theta^0 + O_p(n^{-1/2})$ and under regularity conditions **R1-R5**, we have

(i) $\sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta) = O_p(n^{-1/2})$;

(ii) $\max_{i,j} \left\|\boldsymbol{g}_{ij}(\theta)\right\| = o_p(n^{1/2})$;

(iii) $\sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)\boldsymbol{g}_{ij}(\theta)^\top = O_p(1)$.

The results (i)-(iii) altogether imply that $\hat{\boldsymbol{\lambda}}(\theta) = O_p(n^{-1/2})$. Let

$$Q_{n1}(\theta, \boldsymbol{\lambda}) = \sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \frac{\tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)}$$

and

$$Q_{n2}(\theta, \boldsymbol{\lambda}) = \sum_{i=0}^{1} w_i \sum_{j \in \mathcal{S}_i} \frac{\tilde{a}_{ij}}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)} \left\{\frac{\partial \boldsymbol{g}_{ij}(\theta)}{\partial \theta^\top}\right\}^\top \boldsymbol{\lambda}\,.$$

The estimators $\hat{\theta}_{\text{MCP}}$ and $\hat{\boldsymbol{\lambda}}_{\text{MCP}} = \hat{\boldsymbol{\lambda}}(\hat{\theta}_{\text{MCP}})$ satisfy that $Q_{n1}(\hat{\theta}_{\text{MCP}}, \hat{\boldsymbol{\lambda}}_{\text{MCP}}) = 0$ and $Q_{n2}(\hat{\theta}_{\text{MCP}}, \hat{\boldsymbol{\lambda}}_{\text{MCP}}) = 0$. Applying Taylor expansions to $Q_{n1}(\hat{\theta}_{\text{MCP}}, \hat{\boldsymbol{\lambda}}_{\text{MCP}})$ and $Q_{n2}(\hat{\theta}_{\text{MCP}}, \hat{\boldsymbol{\lambda}}_{\text{MCP}})$ at $(\theta^0, \mathbf{0})$ yields

$$Q_{n1}\left(\theta^0, \mathbf{0}\right) + \frac{\partial Q_{n1}\left(\theta^0, \mathbf{0}\right)}{\partial \theta}\left(\hat{\theta}_{\text{MCP}} - \theta^0\right) + \frac{\partial Q_{n1}\left(\theta^0, \mathbf{0}\right)}{\partial \boldsymbol{\lambda}}\left(\hat{\boldsymbol{\lambda}}_{\text{MCP}} - \mathbf{0}\right) + o_p\left(\sigma_n\right) = \mathbf{0}\,,$$
$$Q_{n2}\left(\theta^0, \mathbf{0}\right) + \frac{\partial Q_{n2}\left(\theta^0, \mathbf{0}\right)}{\partial \theta}\left(\hat{\theta}_{\text{MCP}} - \theta^0\right) + \frac{\partial Q_{n2}\left(\theta^0, \mathbf{0}\right)}{\partial \boldsymbol{\lambda}}\left(\hat{\boldsymbol{\lambda}}_{\text{MCP}} - \mathbf{0}\right) + o_p\left(\sigma_n\right) = \mathbf{0}\,,$$

where $\sigma_n = \|\hat{\theta}_{\text{MCP}} - \theta^0\| + \|\hat{\boldsymbol{\lambda}}_{\text{MCP}}\| = O_p(n^{-1/2})$. Thus, a standard calculation leads to

$$\begin{pmatrix} -Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p\left(n^{-1/2}\right) \\ o_p\left(n^{-1/2}\right) \end{pmatrix} = \boldsymbol{S}_{n1} \begin{pmatrix} \hat{\boldsymbol{\lambda}}_{\text{MCP}} \\ \hat{\theta}_{\text{MCP}} - \theta^0 \end{pmatrix},$$

where

$$\boldsymbol{S}_{n1} \xrightarrow{P} \begin{pmatrix} -\boldsymbol{W} & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^\top & 0 \end{pmatrix}.$$

By noting that $Q_{n1}\left(\theta^0, \mathbf{0}\right) = O_p(n^{-1/2})$, the above equation implies that

$$\begin{pmatrix} \hat{\boldsymbol{\lambda}}_{\text{MCP}} \\ \hat{\theta}_{\text{MCP}} - \theta^0 \end{pmatrix} = \begin{pmatrix} -\boldsymbol{P}_1 & \sigma \boldsymbol{W}^{-1}\boldsymbol{\Gamma} \\ \sigma \boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1} & \sigma \end{pmatrix} \begin{pmatrix} -Q_{n1}\left(\theta^0, \mathbf{0}\right) \\ 0 \end{pmatrix} + o_p(n^{-1/2})\,,$$

where $\sigma = (\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}\boldsymbol{\Gamma})^{-1}$ and $\boldsymbol{P}_1 = \boldsymbol{W}^{-1} - \sigma \boldsymbol{W}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}$.

On the other hand, we have the expression for $Q_{n1}\left(\theta^0, \mathbf{0}\right)$ that

$$Q_{n1}\left(\theta^0, \mathbf{0}\right) = \frac{1}{n}\sum_{j\in\mathcal{S}} \boldsymbol{h}_j + o_p(n^{-1/2}) \,,$$

where $\boldsymbol{h}_j$ is defined in Section 4.2. Note that $\mathrm{E}(\boldsymbol{h}_j) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{h}_j) < \infty$ under regularity conditions **R2-R3** and **R6**. It follows that $\sqrt{n}Q_{n1}\left(\theta^0, \mathbf{0}\right) \overset{d}{\to} \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Omega}\right)$. Therefore,

$$\sqrt{n}\left(\hat{\theta}_{\text{MCP}} - \theta^0\right) = -\sigma \boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}\sqrt{n}Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(1) \overset{d}{\to} \mathrm{N}\left(0, V_1\right) \,,$$

where $V_1 = \sigma^2 \boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1}\boldsymbol{\Omega}\boldsymbol{W}^{-1}\boldsymbol{\Gamma}$. □

8.7. Proof of Theorem 4

Proof. The Lagrange multiplier $\boldsymbol{\lambda}$ is a function of θ, i.e., $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\theta)$, which is the solution to

$$\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \frac{\tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)} = \mathbf{0} \,. \tag{8.7}$$

For $\theta = \theta^0 + O_p(n^{-1/2})$, we have $\boldsymbol{\lambda} = O_p(n^{-1/2})$ as argued in Section 8.6. Taking the Taylor expansion of (8.7) around $\boldsymbol{\lambda} = \mathbf{0}$ gives

$$\mathbf{0} = \sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta) - \sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)\boldsymbol{g}_{ij}(\theta)^\top \boldsymbol{\lambda} + o_p(n^{-1/2}) \,,$$

which implies that, at $\theta = \theta^0$,

$$\begin{aligned}\boldsymbol{\lambda} &= \left[\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)\boldsymbol{g}_{ij}(\theta)^\top\right]^{-1} \left(\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\boldsymbol{g}_{ij}(\theta)\right) + o_p(n^{-1/2}) \\ &= \boldsymbol{W}^{-1}Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(n^{-1/2}) \,.\end{aligned}$$

which further leads to

$$\begin{aligned}-2\ell_{\text{PEL}}(\hat{\boldsymbol{p}}_1(\theta), \hat{\boldsymbol{p}}_0(\theta)) &= nA + 2n\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\log\left\{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)\right\} \\ &= nA + 2n\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\left\{\boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta) - \frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}(\theta)\boldsymbol{g}_{ij}(\theta)^\top \boldsymbol{\lambda}\right\} + o_p(1) \\ &= nA + nQ_{n1}\left(\theta^0, \mathbf{0}\right)^\top \boldsymbol{W}^{-1}Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(1) \,,\end{aligned}$$

where $A = -2\left\{\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij}\log\tilde{a}_{ij}\right\}$.

It is shown in Theorem 3 that $\hat{\boldsymbol{\lambda}}_{\text{MCP}} = \boldsymbol{P}_1 Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(n^{-1/2})$, therefore,

$$\begin{aligned} -2\ell_{\text{PEL}}\left(\hat{\boldsymbol{p}}_1\left(\hat{\theta}_{\text{MCP}}\right), \hat{\boldsymbol{p}}_0\left(\hat{\theta}_{\text{MCP}}\right)\right) =& nA + 2n\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i} \tilde{a}_{ij} \log\left\{1 + \hat{\boldsymbol{\lambda}}_{\text{MCP}}^\top \boldsymbol{g}_{ij}\left(\hat{\theta}_{\text{MCP}}\right)\right\} \\ =& nA + nQ_{n1}\left(\theta^0, \mathbf{0}\right)^\top \boldsymbol{P}_1 \boldsymbol{W} \boldsymbol{P}_1 Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(1) \\ =& nA + nQ_{n1}\left(\theta^0, \mathbf{0}\right)^\top \boldsymbol{P}_1 Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(1) \,. \end{aligned}$$

We can conclude that, when $\theta = \theta^0$,

$$-2r_{\text{PEL}}(\theta) = n\sigma Q_{n1}\left(\theta^0, \mathbf{0}\right)^\top \boldsymbol{W}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1} Q_{n1}\left(\theta^0, \mathbf{0}\right) + o_p(1) \xrightarrow{d} \boldsymbol{Q}^\top \boldsymbol{M} \boldsymbol{Q} \,,$$

where $\boldsymbol{M} = \sigma \boldsymbol{\Omega}^{1/2} \boldsymbol{W}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \boldsymbol{W}^{-1} \boldsymbol{\Omega}^{1/2}$, $\boldsymbol{Q} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}_4)$, and $\boldsymbol{I}_4$ denotes the 4×4 identity matrix. This implies that $-2r_{\text{PEL}}(\theta^0) \xrightarrow{d} \delta\chi_1^2$, where δ is the non-zero eigenvalue of $\boldsymbol{M}$. □

8.8. Justification of the bootstrap procedure

We provide a theoretical justification for the proposed bootstrap procedure under a correctly specified propensity score model. We show that the bootstrap version statistic $-2r_{\text{PEL}}^{[b]}(\hat{\theta}_{\text{MCP}})$, conditional on the given sample $\mathcal{S}$, follows approximately a scaled χ_1^2 distribution when the initial sample size n is large, and the ratio of the scaling constant $\delta^{[b]}$ over δ (given in Theorem 4) converges in probability to 1 as $n \to \infty$. The theoretical arguments used here are similar to the justification of the bootstrap procedure for the generalized pseudo empirical likelihood inferences described in Tan and Wu (2015).

First, we can argue that each bootstrap sample leads to a similar expansion in the form of

$$\begin{pmatrix} \hat{\boldsymbol{\lambda}}_{\text{MCP}}^{[b]} \\ \hat{\theta}_{\text{MCP}}^{[b]} - \hat{\theta}_{\text{MCP}} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{P}_1^{[b]} & \left(\boldsymbol{W}^{[b]}\right)^{-1}\boldsymbol{\Gamma}\sigma^{[b]} \\ \sigma^{[b]}\boldsymbol{\Gamma}^\top\left(\boldsymbol{W}^{[b]}\right)^{-1} & \sigma^{[b]} \end{pmatrix} \begin{pmatrix} -Q_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right) \\ 0 \end{pmatrix} + o_p(n^{-1/2})$$

when the initial sample size n is large, where $\hat{\theta}_{\text{MCP}}^{[b]}$ is the bootstrap version of the point estimator obtained via maximizing the joint pseudo-empirical likelihood function $\ell_{\text{PEL}}^{[b]}(\boldsymbol{p}_1, \boldsymbol{p}_0)$ subject to the normalization and model-calibration constraints, and $\hat{\boldsymbol{\lambda}}_{\text{MCP}}^{[b]}$ is the corresponding value for the parameter $\boldsymbol{\lambda}$. Let $\boldsymbol{W}^{[b]}$ denote the limit of $\sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i^{[b]}} \tilde{a}_{ij}^{[b]} \boldsymbol{g}_{ij}^{[b]}$
$(\hat{\theta}_{\text{MCP}})\boldsymbol{g}_{ij}^{[b]}(\hat{\theta}_{\text{MCP}})^\top$. We have $\sigma^{[b]} = (\boldsymbol{\Gamma}^\top(\boldsymbol{W}^{[b]})^{-1}\boldsymbol{\Gamma})^{-1}$ and $\boldsymbol{P}_1^{[b]} = (\boldsymbol{W}^{[b]})^{-1} - \sigma^{[b]}(\boldsymbol{W}^{[b]})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top(\boldsymbol{W}^{[b]})^{-1}$. These quantities are just the bootstrap versions of the σ, $\boldsymbol{P}_1$ and $\boldsymbol{W}$. The bootstrap version of $Q_{n1}(\theta, \boldsymbol{\lambda})$ is

$$Q_{n1}^{[b]}(\theta, \boldsymbol{\lambda}) = \sum_{i=0}^{1} w_i \sum_{j\in\mathcal{S}_i^{[b]}} \frac{\tilde{a}_{ij}^{[b]} \boldsymbol{g}_{ij}^{[b]}(\theta)}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}_{ij}^{[b]}(\theta)} \,,$$

and $Q_{n1}^{[b]}(\hat{\theta}_{\text{MCP}}, \mathbf{0})$ is also asymptotically normally distributed.

Second, we can argue that $(\hat{\theta}_{\text{MCP}}, \boldsymbol{\lambda}_B^{[b]})$ satisfies that $Q_{n1}^{[b]}(\hat{\theta}_{\text{MCP}}, \boldsymbol{\lambda}_B^{[b]}) = 0$, where $\boldsymbol{\lambda}_B^{[b]} = \boldsymbol{\lambda}(\hat{\theta}_{\text{MCP}})$ based on $Q_{n1}^{[b]}(\theta, \boldsymbol{\lambda})$. Applying the Taylor expansion to $Q_{n1}^{[b]}(\hat{\theta}_{\text{MCP}}, \boldsymbol{\lambda}_B^{[b]}) = \mathbf{0}$ at $\boldsymbol{\lambda}_B^{[b]} = \mathbf{0}$ gives that $\boldsymbol{\lambda}_B^{[b]} = (\boldsymbol{W}^{[b]})^{-1} Q_{n1}^{[b]}(\hat{\theta}_{\text{MCP}}, \mathbf{0}) + o_p(n^{-1/2})$. This yields the expression

$$2\ell_{\text{PEL}}^{[b]}\left(\hat{\boldsymbol{p}}_1\left(\hat{\theta}_{\text{MCP}}\right), \hat{\boldsymbol{p}}_0\left(\hat{\theta}_{\text{MCP}}\right)\right) = nA^{[b]} - nQ_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right)^\top (\boldsymbol{W}^{[b]})^{-1} Q_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right) + o_p(1),$$

where $A^{[b]} = \sum_{i=0}^{1} \sum_{j \in \mathcal{S}_i^{[b]}} \tilde{a}_{ij}^{[b]} \log \tilde{a}_{ij}^{[b]}$. On the other hand, we compute $\ell_{\text{PEL}}^{[b]}(\boldsymbol{p}_1(\theta), \boldsymbol{p}_0(\theta))$ at $\theta = \hat{\theta}_{\text{MCP}}^{[b]}$ to obtain that

$$2\ell_{\text{PEL}}^{[b]}\left(\hat{\boldsymbol{p}}_1\left(\hat{\theta}_{\text{MCP}}^{[b]}\right), \hat{\boldsymbol{p}}_0\left(\hat{\theta}_{\text{MCP}}^{[b]}\right)\right) = nA^{[b]} - nQ_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right)^\top \boldsymbol{P}_1^{[b]} Q_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right) + o_p(1).$$

Finally, the above results lead to

$$\begin{aligned} -2r_{\text{PEL}}^{[b]}(\hat{\theta}_{\text{MCP}}) &= n\sigma^{[b]} Q_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right)^\top (\boldsymbol{W}^{[b]})^{-1} \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top (\boldsymbol{W}^{[b]})^{-1} Q_{n1}^{[b]}\left(\hat{\theta}_{\text{MCP}}, \mathbf{0}\right) + o_p(1) \\ &\xrightarrow{d} \boldsymbol{Q}^\top \boldsymbol{M}^{[b]} \boldsymbol{Q}, \end{aligned}$$

where $\boldsymbol{M}^{[b]} = \sigma^{[b]} (\boldsymbol{\Omega}^{[b]})^{1/2} (\boldsymbol{W}^{[b]})^{-1} \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top (\boldsymbol{W}^{[b]})^{-1} (\boldsymbol{\Omega}^{[b]})^{1/2}$ and $\boldsymbol{Q} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}_4)$. Moreover, $\boldsymbol{\Omega}^{[b]}$ is the asymptotic variance matrix of $Q_{n1}^{[b]}(\hat{\theta}_{\text{MCP}}, \mathbf{0})$. This implies that $-2r_{\text{PEL}}^{[b]}(\theta) \xrightarrow{d} \delta^{[b]} \chi_1^2$ if $\theta = \hat{\theta}_{\text{MCP}}$, where $\delta^{[b]}$ is the non-zero eigenvalue of $\boldsymbol{M}^{[b]}$. The matrix $\boldsymbol{M}^{[b]}$ is the bootstrap sample version of the matrix $\boldsymbol{M}$ from which we obtain δ. Therefore, $\delta^{[b]}/\delta$ converges in probability to 1 as $n \to \infty$.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor and two reviewers for constructive comments and suggestions which led to an improved final version of the paper.

Funding

C.W. was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute. L.Z. was supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267. MR2194325

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120:1031–1083.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Rmer, T. S. (2006). Variable selection for propensity score models. *Am J Epidemiol*, 163:1149–1156.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734. MR2538768

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155:138–154. MR2607191

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science*, 29:380–396. MR3264551

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–123. MR1225218

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406. MR1707846

Chen, M., Thompson, M. E., and Wu, C. (2018). Empirical likelihood methods for complex surveys with data missing-by-design. *Statistica Sinica*, 28:2027–2048. MR3839002

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104:439–453. MR3698264

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115:2011–2021. MR4189773

Chen, Y., Li, P., and Wu, C. (2023). Dealing with undercoverage for nonprobability survey samples. *Survey Methodology*, 49:497–515.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382. MR1173804

Duan, X. and Yin, G. (2017). Ensemble approaches to estimating the population mean with missing response. *Scandinavian Journal of Statistics*, 44:899–917. MR3730020

Ghosh, T., Ma, Y., and Luna, X. D. (2021). Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Statistica Sinica*, 31:821–842. MR4286196

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

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statisti -cal inference*, V. Godambe & D. Sprott, eds. Toronto: Holt, Rinehart and Winston. page 236.

Hall, P. and Scala, B. L. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58:109.

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, 148:101–110. MR3174151

Han, P. (2014b). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109:1159–1173. MR3265688

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103:683–700. MR3551792

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100:417–430. MR3068443

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189. MR1995826

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685. MR0053460

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc. A*, 171:481–502. MR2427345

Liu, J., Ma, Y., and Wang, L. (2018). An alternative robust estimator of average treatment effect in causal inference. *Biometrics*, 74:910–923. MR3860712

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249. MR0946049

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18:90–120. MR1041387

Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19:1725–1747. MR1135146

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

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22:300–325. MR1272085

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560. MR1766776

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866. MR1294730

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55. MR0742974

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524. MR0763575

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1120. MR1731478

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *American Statistician*, 56(1):29–38. MR1939394

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag New York. MR1140409

Tan, Z. and Wu, C. (2015). Generalized pseudo empirical likelihood inferences for complex surveys. *The Canadian Journal of Statistics*, 43:1–17. MR3324425

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York: Springer. MR2233926

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25. MR0640163

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90:935–951. MR2024768

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14:1057–1067. MR2126339

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, 31:239–243.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34:359–375. MR2328549

Wu, C. and Rao, J. N. K. (2010). Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. *Statistics and Probability Letters*, 80:1472–1478. MR2669748

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193. MR1952731

Wu, C. and Thompson, M. E. (2020). *Sampling Theory and Practice*. Springer, Cham. MR4180686

Wu, C. and Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and Its Interface*, 5:345–354. MR2979214

Zhang, S., Han, P., and Wu, C. (2023). Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *International Statistical Review*, 91:165–192. MR4629023

Zhao, Q. and Percival, D. (2017). Entropy Balancing is Doubly Robust. *Journal of Causal Inference*, 5(1):1–19. MR4323812