Electronic Journal of Statistics
Vol. 18 (2024) 2687–2720
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2263

Robust propensity score weighting estimation under missing at random

Hengfang Wang[1], **Jae Kwang Kim**[*2], **Jeongseop Han**[3] and **Youngjo Lee**[3]

[1]*School of Mathematics and Statistics & Fujian Provincial Key Laboratory of Statistics and Artificial Intelligence, Fujian Normal University, Fuzhou, Fujian, 350117, China,* *e-mail:* hengfang@fjnu.edu.cn

[2]*Department of Statistics, Iowa State University, Ames, IA, 50011, U.S.A.,* *e-mail:* jkim@iastate.edu

[3]*Department of Statistics, Seoul National University, Seoul, 08826, Korea (South),* *e-mail:* jshan411@gmail.com; youngjomail@gmail.com

Abstract: Missing data is frequently encountered in many areas of statistics. One popular approach to address this issue is through the use of propensity score weighting. However, correctly specifying the statistical model can be a daunting task. Doubly robust estimation is attractive, as the consistency of the estimator is guaranteed when either the outcome regression model or the propensity score model is correctly specified. In this paper, we first employ information projection to develop an efficient and doubly robust estimator via indirect model calibration. The resulting propensity score estimator can be equivalently expressed as a doubly robust regression imputation estimator by imposing the internal bias calibration condition in estimating the regression parameters. In addition, using the γ-divergence measure, we generalize the information projection to allow for outlier-robust propensity score estimation. The study includes the presentation of certain asymptotic properties and findings from a simulation study, which demonstrate that the proposed method enables robust inference, not only in cases of various model assumptions being violated but also in the presence of outliers. A real-life application is also presented using data from the Conservation Effects Assessment Project.

MSC2020 subject classifications: Primary 62D10, 62F35.
Keywords and phrases: Covariate balancing, information projection, γ-power divergence, missing data.

Received September 2023.

1. Introduction

Missing data represents a fundamental challenge in statistics. Ignoring missing data can lead to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weak generalizability of findings, as highlighted by Dong and Peng (2013). In addition, the missing data framework is very useful in defining problems in other research areas, such as

arXiv: 2306.15173
*Corresponding author.

causal inference or offline policy evaluation. Propensity score (PS) weighting is a popular method to handle missing data. It often employs a response propensity model, but correct specification of the statistical model can be challenging in the presence of missing data. How to make the propensity score weighting method less dependent on the response propensity model is an important practical problem.

In the quest for robust PS estimation, we find two distinct avenues. One approach embraces flexibility, employing nonparametric or semiparametric models to construct robust propensity scores. The nonparametric kernel method (Hahn, 1998), the sieve logistic regression method (Hirano, Imbens and Ridder, 2003), the general calibration method of Chan, Yam and Zhang (2016) using increasing dimension of the basis functions, the generalized covariate balancing estimator using tailored loss functions (Zhao, 2019) and the random forest approach of Dagdoug, Goga and Haziza (2023) are examples of the robust propensity score estimation method using flexible models. The other path, which we explore in this paper, leans on the outcome regression model explicitly to achieve doubly robust estimation. The literature is replete with investigations in this arena, with notable contributors like Bang and Robins (2005), Cao, Tsiatis and Davidian (2009), Kim and Haziza (2014), Han and Wang (2013), Chen and Haziza (2017), and Yang, Kim and Song (2020).

Our journey takes the latter route as we construct a unified framework for doubly robust PS estimation under the setup of missing at random (Rubin, 1976). To achieve the goal, we apply the information projection (Csiszár and Shield, 2004) to the PS weighting problem, while adhering to covariate-balancing constraints. Specifically, the initial PS weights are constructed from the working PS model, but the balancing constraints, crucial for achieving double robustness in the PS weighting estimator, are derived from the working outcome regression model. This maneuver can be perceived as indirect model calibration. Information projection is used to obtain the augmented PS model.

Remarkably, the PS weighting estimator that is obtained from information projection can also be cast as a special type of regression imputation estimator, incorporating balancing functions as covariates in the outcome regression model and the regression coefficients are computed using the weights computed from the final PS weights. This algebraic equivalence, known as self-efficiency, forms the cornerstone of our journey toward outlier-robust PS estimation. In practice, outliers do exist and the presence of outliers can significantly undermine the efficiency of estimators. While outlier-robust procedures are well studied in the literature on regression or classification (Stefanski, Carroll and Ruppert, 1986; Wu and Liu, 2007; Zhang, Jiang and Chai, 2010), their application in the context of missing data remains underexplored. To our best knowledge, the construction of an outlier-robust PS weighting estimator has not been investigated in the literature.

To fill in this important research gap, we embark on the creation of an outlier-robust regression imputation technique and subsequently align it with a PS weighting estimator by imposing the self-efficiency condition. In doing so, we set forth on a path previously untraveled in the literature – a journey toward

constructing an outlier-robust PS estimator. Specifically, the information projection approach to developing the augmented PS model and obtaining the doubly robust PS estimation is based on the Kullback-Leibler divergence. The γ-power divergence, originally proposed by Basu et al. (1998) and further developed by Eguchi (2021), is a generalization of the Kullback-Leibler divergence to expand the class of statistical models by introducing an additional scale parameter γ. Furthermore, it is well known that the statistical model derived from the γ-power divergence produces robust inferences against outliers. We adopt the γ-power divergence to develop an outlier-robust regression imputation estimator. By self-efficiency, the regression estimator can be expressed as a PS weighting estimator. Therefore, we obtain an outlier-robust PS weighting estimator. The resulting estimator is also robust against misspecification of the response probability model.

The structure of the paper is as follows. In Section 2, the basic setup and the research problem are introduced. In Section 3, we develop a balancing constraint to adjust propensity weights and construct an augmented propensity model using information projection. In Section 5, we present the use of the γ-power divergence to enlarge the class of propensity score models and develop outlier-robust doubly robust estimators. Results from two limited simulation studies are presented in Section 6. A real-life application using data from the Conservation Effects Assessment Project is presented in Section 7. Some concluding remarks are made in Section 8. All required proofs are presented in the Appendix.

2. Basic setup

Suppose that there are n independently and identically distributed realizations of $(\boldsymbol{X}, Y, \delta)$, denoted by $\{(\boldsymbol{x}_i, y_i, \delta_i) : i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is a vector of observed covariates and δ_i is the missingness indicator associated with unit i. In particular, $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Thus, instead of observing $(\boldsymbol{x}_i, y_i, \delta_i)$, we only observe $(\boldsymbol{x}_i, \delta_i, \delta_i y_i)$ for $i = 1, \ldots, n$. We are interested in estimating $\theta = E(Y)$ from the observed sample.

Suppose that we have a working outcome regression (OR) model given by

$$m_0(\boldsymbol{x}; \boldsymbol{\beta}) \in \text{span}\{b_0(\boldsymbol{x}), b_1(\boldsymbol{x}), \ldots, b_L(\boldsymbol{x})\} \tag{2.1}$$

for some given basis functions $b_1(\boldsymbol{x}), \ldots, b_L(\boldsymbol{x})$ and $b_0(\boldsymbol{x}) = 1$. The model (2.1) can be expressed equivalently as

$$y_i = m(\boldsymbol{x}_i; \boldsymbol{\beta}) + \varepsilon_i, \;\; m(\boldsymbol{x}_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 b_1(\boldsymbol{x}_i) + \ldots + \beta_L b_L(\boldsymbol{x}_i)$$

for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_L)^{\mathrm{T}}$, where ε_i is an error term that is independent of $\boldsymbol{x}_i$ and satisfies $E(\varepsilon_i) = 0$. Furthermore, the assumption of missing at random (Rubin, 1976) implies that ε_i and δ_i are conditionally independent given $\boldsymbol{x}_i$.

We are interested in using the following propensity score weighting (PSW) estimator

$$\hat{\theta}_{\mathrm{PSW}} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \omega_i y_i. \tag{2.2}$$

The following lemma provides a sufficient condition for the unbiasedness of the propensity score weighting estimators under the model (2.1).

Lemma 2.1. *Assume that the response mechanism is missing at random. If the weights ω_i's satisfy*

$$\sum_{i=1}^{n} \delta_i \omega_i \left[b_0(\boldsymbol{x}_i), b_1(\boldsymbol{x}_i), \ldots, b_L(\boldsymbol{x}_i)\right] = \sum_{i=1}^{n} \left[b_0(\boldsymbol{x}_i), b_1(\boldsymbol{x}_i), \ldots, b_L(\boldsymbol{x}_i)\right], \tag{2.3}$$

then $\hat{\theta}_{\mathrm{PSW}}$ in (2.2) is unbiased for θ under the OR model in (2.1).

By Lemma 2.1, condition (2.3) is the key condition that incorporates the OR model into the PSW estimator. Condition (2.3) is often called the covariate balancing condition (Imai and Ratkovic, 2014) in the missing data literature. It is closely related to the calibration estimation in survey sampling (Deville and Särndal, 1992). Under the OR model in (2.1), the covariate-balancing condition in (2.3) implies that

$$\sum_{i=1}^{n} \delta_i \omega_i E(Y_i \mid \boldsymbol{x}_i) = \sum_{i=1}^{n} E(Y_i \mid \boldsymbol{x}_i), \tag{2.4}$$

which is often referred to as the model calibration (Wu and Sitter, 2001). To distinguish from the direct model calibration of Wu and Sitter (2001), we may call (2.3) the *indirect model calibration*.

The indirect model calibration condition also implies the following algebraic equivalence.

Lemma 2.2. *If the weights ω_i satisfy (2.3), then we have*

$$\sum_{i=1}^{n} \delta_i \omega_i y_i = \sum_{i=1}^{n} \{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}\}, \tag{2.5}$$

where $\hat{\boldsymbol{\beta}}$ satisfies

$$\sum_{i=1}^{n} \delta_i \left(\omega_i - 1\right)\left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}\right) = 0, \tag{2.6}$$

and $\boldsymbol{b}_i = (1, b_1(\boldsymbol{x}_i), \ldots, b_L(\boldsymbol{x}_i))^{\mathrm{T}}$.

Lemma 2.2 means that, under (2.6), the PSW estimator that satisfies the covariate-balancing condition is algebraically equivalent to a regression imputation estimator using the balancing functions as covariates in the outcome regression model. In other words, regression imputation under the outcome re-

gression model in (2.1) can be viewed as a PSW estimator where the propensity score weights ω_i and the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ are related by equation (2.6). Equation (2.5) means that the final propensity score weights ω_i do not directly use the outcome regression model (2.1) for imputation, but implement regression imputation indirectly. Condition (2.5) is referred to as the self-efficiency condition.

We now wish to achieve double robustness by including the PS model. Suppose that we have the following working PS model

$$P(\delta = 1 \mid \boldsymbol{x}) = \pi_1(\boldsymbol{x}; \boldsymbol{\phi}_0)$$

for some $\boldsymbol{\phi}_0$ with a known function $\pi_1(\cdot) \in (0, 1)$. An example is the logistic regression model such as $\text{logit}\{\pi_1(\boldsymbol{x}; \boldsymbol{\phi}_0)\} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\phi}_0$. Under the working propensity score model, the PS weights are computed as $\omega_i = \{\pi_1(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})\}^{-1} := \hat{d}_i$, where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\phi}$. However, the PS weight $\hat{d}_i$ does not necessarily satisfy the balancing condition in (2.3). Therefore, to reduce the bias due to model misspecification in the propensity score model, it makes sense to impose the balancing condition in the final weighting. To achieve this goal, Hainmueller (2012) proposed the so-called entropy balancing method that minimizes

$$Q(\boldsymbol{\omega}) = \sum_{i=1}^{n} \delta_i \omega_i \log(\omega_i / \hat{d}_i), \tag{2.7}$$

subject to the balancing constraint in (2.3). Chan, Yam and Zhang (2016) generalized this idea further to develop a general calibration weighting method that satisfies the covariance balancing property with increasing dimensions of basis functions $\boldsymbol{b}(\boldsymbol{x})$. However, the choice of distance measure is somewhat unclear. In the next section, we adopt the information projection approach to develop a unified approach to doubly robust propensity score weighting.

3. Proposed method

We now develop an alternative approach to modifying the initial propensity score weights to satisfy the covariate balancing condition in (2.3). The proposed approach consists of two parts. One is the modeling part and the other is the parameter estimation part. In the modeling part, we use information projection innovatively to obtain the propensity score model incorporating the moment constraints from the outcome regression model. Information projection is a powerful tool for incorporating the moment constraints into the final model.

3.1. Information projection

Instead of using a distance measure for propensity weights, as in (2.7), we use the Kullback-Leibler divergence measure properly to develop the information

projection under some moment constraint obtained from the outcome regression model. Because the Kullback-Leibler divergence is defined in terms of density functions, we need to formulate the weighting problem as an optimization problem for density functions.

To apply the information projection, using the Bayes theorem, we obtain

$$\frac{P(\delta = 0 \mid \boldsymbol{x})}{P(\delta = 1 \mid \boldsymbol{x})} = \frac{1-p}{p} \times \frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})},$$

where $p = P(\delta = 1)$, and $f_k(\boldsymbol{x}) = f(\boldsymbol{x} \mid \delta = k)$ is the density function for $\boldsymbol{x}$ given $\delta = k$ for $k = 0, 1$. Thus, the propensity score (PS) weight can be written as

$$\omega(\boldsymbol{x}) \equiv \frac{1}{P(\delta = 1 \mid \boldsymbol{x})} = 1 + c \times \frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})}. \tag{3.1}$$

where $c = 1/p - 1$. For a fixed f_1, the propensity weight function $\omega(\boldsymbol{x})$ is completely determined by f_0. Furthermore, assume that the basis function $\boldsymbol{b}(\boldsymbol{x})$ in the outcome regression model in (2.1) is integrable in the sense that

$$pE_1\{\boldsymbol{b}(\boldsymbol{X})\} + (1-p)E_0\{\boldsymbol{b}(\boldsymbol{X})\} = E\{\boldsymbol{b}(\boldsymbol{X})\}, \tag{3.2}$$

where $E_k\{\boldsymbol{b}(\boldsymbol{X})\} = \int \boldsymbol{b}(\boldsymbol{x}) f_k(\boldsymbol{x}) d\mu(\boldsymbol{x})$ for $k = 0, 1$ and μ is the Lebesgue measure. Thus, for a given f_1, equation (3.2) can serve as a constraint on f_0 when $E\{\boldsymbol{b}(\boldsymbol{X})\}$ is known.

Let $\pi_1^{(0)}(\boldsymbol{x})$ be the PS function corresponding to a "working" model for $\pi_1(\boldsymbol{x}) = P(\delta = 1 \mid \boldsymbol{x})$. For a fixed f_1, by (3.1), we can define $f_0^{(0)}$ to satisfy

$$\frac{1}{\pi_1^{(0)}(\boldsymbol{x})} = 1 + c \times \frac{f_0^{(0)}(\boldsymbol{x})}{f_1(\boldsymbol{x})}. \tag{3.3}$$

We can treat $f_0^{(0)}$ as the baseline density for f_0 derived from the working PS function $\pi_1^{(0)}(\boldsymbol{x})$. Our objective is to modify $f_0^{(0)}$ to satisfy (3.2). Finding the density function f_0 satisfying the balancing condition in (3.2) can be formulated as an optimization problem using the Kullback-Leibler divergence. Given the baseline density $f_0^{(0)}$, the Kullback-Leibler divergence of f_0 at $f_0^{(0)}$ is defined by

$$D_{\mathrm{KL}}\left(f_0 \| f_0^{(0)}\right) = \int \log\left(\frac{f_0}{f_0^{(0)}}\right) f_0 d\mu. \tag{3.4}$$

Given $f_0^{(0)}$, we wish to find the minimizer of $D_{\mathrm{KL}}(f_0 \| f_0^{(0)})$ subject to (3.2). The Lagrangian function can be formulated as

$$\begin{aligned}\mathcal{L}(f_0, \boldsymbol{\lambda}) = &\int \log\left(f_0 / f_0^{(0)}\right) f_0 d\mu \\ &+ \boldsymbol{\lambda}^{\mathrm{T}} \left[p \int \boldsymbol{b}(\boldsymbol{x}) f_1(\boldsymbol{x}) d\mu + (1-p) \int \boldsymbol{b}(\boldsymbol{x}) f_0(\boldsymbol{x}) d\mu - E\{\boldsymbol{b}(\boldsymbol{X})\}\right],\end{aligned} \tag{3.5}$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier. As f_1 is fixed and $E\{\boldsymbol{b}(\boldsymbol{X})\}$ is known, by taking the derivative with respect to f_0 in (3.5), one may arrive at $f_0^*(\boldsymbol{x};\boldsymbol{\lambda}) \propto f_0^{(0)}(\boldsymbol{x}) \exp\{\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\lambda}\}$. Furthermore, the density constraint $\int f_0^*(\boldsymbol{x};\boldsymbol{\lambda})d\boldsymbol{x} = 1$ leads to

$$f_0^*(\boldsymbol{x};\boldsymbol{\lambda}) = f_0^{(0)}(\boldsymbol{x}) \frac{\exp\{\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\lambda}\}}{\int \exp\{\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\lambda}\} f_0^{(0)}(\boldsymbol{x}) d\mu(\boldsymbol{x})}, \tag{3.6}$$

with abuse of notation for $\boldsymbol{\lambda}$. By (3.6), we obtain

$$\frac{f_0^*(\boldsymbol{x};\boldsymbol{\lambda})}{f_1(\boldsymbol{x})} = \frac{f_0^{(0)}(\boldsymbol{x})}{f_1(\boldsymbol{x})} \times \frac{\exp\{\sum_{j=1}^{L} \lambda_j b_j(\boldsymbol{x})\}}{\int \exp\{\sum_{j=1}^{L} \lambda_j b_j(\boldsymbol{x})\} f_0^{(0)}(\boldsymbol{x}) d\mu(\boldsymbol{x})}. \tag{3.7}$$

Combining the above results, we can establish the following lemma.

Lemma 3.1. *Given the PS function $\pi_1^{(0)}(\boldsymbol{x})$ from the working PS model, the final PS function minimizing the Kullback-Leibler divergence in (3.4) subject to the balancing condition (3.2) is given by*

$$\pi_1^*(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{\pi_1^{(0)}(\boldsymbol{x})}{\pi_1^{(0)}(\boldsymbol{x}) + \{1 - \pi_1^{(0)}(\boldsymbol{x})\} \exp\left\{\lambda_0 + \sum_{j=1}^{L} \lambda_j b_j(\boldsymbol{x})\right\}}, \tag{3.8}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_L$ are the Lagrange multipliers satisfying (3.2).

In (3.8), the Lagrange multiplier $\boldsymbol{\lambda}$ is determined to satisfy the balancing condition (3.2) with

$$E_0\{\boldsymbol{b}(\boldsymbol{X})\} = \frac{\int \boldsymbol{b}(\boldsymbol{x}) O^{(0)}(\boldsymbol{x}) \exp\{\sum_{j=1}^{L} \lambda_j b_j(\boldsymbol{x})\} f_1(\boldsymbol{x}) d\mu(\boldsymbol{x})}{\int O^{(0)}(\boldsymbol{x}) \exp\{\sum_{j=1}^{L} \lambda_j b_j(\boldsymbol{x})\} f_1(\boldsymbol{x}) d\mu(\boldsymbol{x})}.$$

where $O^{(0)}(\boldsymbol{x}) = \{\pi_1^{(0)}(\boldsymbol{x})\}^{-1} - 1$. If the working propensity score model is correct, then the balancing constraint (3.2) is already satisfied. In this case, we have $\lambda_j \equiv 0$ for all $j = 1, \ldots, L$. Thus, the information projection is used to obtain the augmented propensity score model (Kim and Riddles, 2012).

Figure 1 presents a graphical summary of obtaining the augmented PS model in (3.8). Using the relationship in (3.1), we can transform $\pi_1(x)$ to $f_0(x)$ and then apply the information projection to obtain $f_0^*(x)$ in (3.6), which in turn finds $\pi_1^*(x)$ in (3.8) by applying (3.1) again.

In practice, the base functions for the "working" outcome model are chosen by standard model selection techniques. For example, Shortreed and Ertefaie (2017) used adaptive Lasso to select the covariates in the outcome regression model.

3.2. Model parameter estimation

We now discuss parameter estimation in the augmented propensity score model in (3.8). Suppose that the working propensity score model is given by $\pi_1^{(0)}(\boldsymbol{x};\boldsymbol{\phi})$

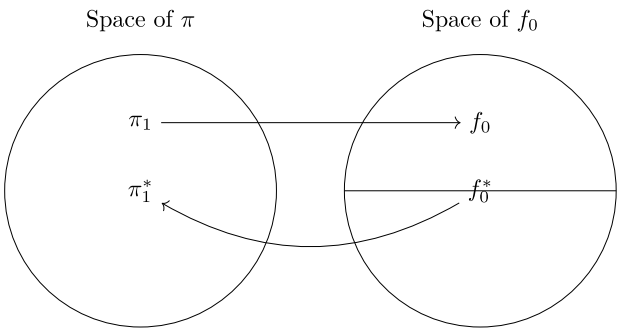

FIG 1. *A graphical illustration of obtaining $\pi_1^*(x)$ in (3.8) using the information projection for f_0, where the line running halfway across the right circle represents the space of f_1 satisfying the covariate-balancing constraints.*

with the unknown parameter $\boldsymbol{\phi}$. We can estimate $\boldsymbol{\phi}$ by maximizing

$$\ell(\boldsymbol{\phi}) = \sum_{i=1}^{n} \left[\delta_i \log \left\{ \pi_1^{(0)}(\boldsymbol{x}_i; \boldsymbol{\phi}) \right\} + (1 - \delta_i) \log\{1 - \pi_1^{(0)}(\boldsymbol{x}_i; \boldsymbol{\phi})\} \right]$$

with respect to $\boldsymbol{\phi}$. Note that the propensity score model has nothing to do with the OR model in (2.1).

Now, to incorporate the OR model in (2.1), we use the augmented propensity score model in (3.8) to obtain the final propensity score weights

$$\hat{\omega}_i = 1 + (\hat{d}_i - 1) \exp \left\{ \sum_{j=0}^{L} \hat{\lambda}_j b_j(\boldsymbol{x}_i) \right\}, \tag{3.9}$$

where $\hat{d}_i = \{\pi_1^{(0)}(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})\}^{-1}$ and $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_L$ are computed from the calibration equation:

$$\sum_{i=1}^{n} \delta_i \hat{\omega}_i b_j(\boldsymbol{x}_i) = \sum_{i=1}^{n} b_j(\boldsymbol{x}_i), \quad \forall j = 0, 1, \ldots, L. \tag{3.10}$$

By construction, the PS weights in (3.9) satisfies the covariate-balancing property on the basis functions in the outcome regression model in (2.1). By Lemma 2.1, the covariate balancing property implies unbiasedness of the PSW estimator under the OR model. On the other hand, if the working PS model is correct, then $\hat{\lambda}_j \to 0$ in probability as $n \to \infty$ for $j = 0, 1, \ldots, L$. In this case, $\hat{\omega}_i$ will converge to $\hat{d}_i$ and the PSW estimator is consistent under the PS model.

As discussed in Section 2, the propensity score estimator satisfying the indirect model calibration condition can be expressed as a regression imputation estimator. Since $\hat{\omega}_i$ satisfy (3.10), by Lemma 2.2, we can obtain

$$\frac{1}{n} \sum_{i=1}^{n} \delta_i \hat{\omega}_i y_i = \frac{1}{n} \sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i) \hat{y}_i\}, \tag{3.11}$$

where $\hat{y}_i = \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and

$$\hat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{n} \delta_i \left(\hat{\omega}_i - 1\right) \boldsymbol{b}_i \boldsymbol{b}_i^{\mathrm{T}}\right\}^{-1} \left\{\sum_{i=1}^{n} \delta_i \left(\hat{\omega}_i - 1\right) \boldsymbol{b}_i y_i\right\}. \tag{3.12}$$

Note that, by (3.9), $\hat{\boldsymbol{\beta}}$ in (3.12) can be expressed as the minimizer of

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 (\hat{d}_i - 1)\hat{g}_i, \tag{3.13}$$

where $\hat{g}_i = \exp(\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}})$. The first term $\hat{d}_i - 1$ is the adjustment term to incorporate the working PS model into the regression imputation. The second term $\hat{g}_i = \exp(\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}})$ is to achieve covariate balance in the PS weights, which is a sufficient condition for self-efficiency.

Now, if $g_i = 1$ is used in (3.13), then the self-efficiency in (3.11) will not be satisfied. Instead, we only obtain

$$\frac{1}{n}\sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i)\hat{y}_i\} = \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{y}_i + \delta_i \hat{d}_i \left(y_i - \hat{y}_i\right)\right\}. \tag{3.14}$$

Condition (3.14) is called the *internal bias calibration* (IBC), which was originally termed by Firth and Bennett (1998) in the context of design-consistent estimation of model parameters under complex sampling. The imputation estimator satisfying the IBC condition (3.14) achieves consistency even when the outcome regression model is incorrect, as long as the working PS model is correct. Thus, the IBC condition is a sufficient condition for the double robustness of the regression imputation estimator.

For the choice of $\hat{g}_i = \exp(\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}})$, under the PS model, we obtain $\hat{g}_i \to 1$ in probability as $n \to \infty$. Thus, the IBC condition in (3.14), or double robustness of the regression imputation estimator, holds approximately.

Equation (3.11) deserves further discussion. In the PSW estimation, for a given $\hat{\boldsymbol{\phi}}$, the final estimator $\hat{\theta}_{\mathrm{PSW}}$ is a function of estimated nuisance parameter $\hat{\boldsymbol{\lambda}}$ while the regression imputation estimator is a function of other estimated nuisance parameter $\hat{\boldsymbol{\beta}}$. Thus, $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}$ are in one-to-one correspondence with each other through the following estimating equation:

$$\sum_{i=1}^{n} \delta_i (y_i - \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})\boldsymbol{b}_i (\hat{d}_i - 1)\exp(\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}}) = \boldsymbol{0}. \tag{3.15}$$

In summary, the proposed method can be implemented in the following steps.

1. Specify a working PS model $\pi_1^{(0)}(\boldsymbol{x}; \boldsymbol{\phi})$ and estimate $\boldsymbol{\phi}$ using the ML estimation method to get $\hat{d}_i = \{\pi_1^{(0)}(\boldsymbol{x}; \hat{\boldsymbol{\phi}})\}^{-1}$.
2. Specify a working OR model in (2.1) to construct the augmented PS model in (3.8).
3. Compute $\hat{\boldsymbol{\lambda}}$ from the calibration equation in (3.10) to obtain $\hat{\omega}_i$ in (3.9).
4. Also, compute $\hat{\boldsymbol{\beta}}$ from (3.12) to obtain the regression imputation estimator in (3.11).

4. Statistical properties

Now we formally describe the asymptotic properties of the augmented PSW estimator using the final propensity score weight (3.9) with the calibration constraint in (3.10). By the algebraic equivalence established in Lemma 2.2, the result is directly applicable to the regression imputation estimator using the regression coefficient in (3.12).

Let $\pi_1^{(0)}(\boldsymbol{x};\boldsymbol{\phi})$ be the working propensity score model, where $\boldsymbol{\phi} \in \mathbb{R}^p$. We can estimate $\boldsymbol{\phi}$ by solving

$$\boldsymbol{U}_{1,n}(\boldsymbol{\phi}) \equiv \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})} - 1\right\}\boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{\phi}) = \boldsymbol{0} \tag{4.1}$$

for some $\boldsymbol{h}(\boldsymbol{x};\boldsymbol{\phi})$ such that the solution to (4.1) exists uniquely almost everywhere. The estimating equation (4.1) for $\boldsymbol{\phi}$ includes the score equation for $\boldsymbol{\phi}$ as a special case. For a given $\hat{\boldsymbol{\phi}}$, let $\hat{\boldsymbol{U}}_2(\boldsymbol{\lambda} \mid \hat{\boldsymbol{\phi}})$ be the estimating equation for $\boldsymbol{\lambda}$. By (2.3), we can estimate $\boldsymbol{\lambda}$ by solving

$$\boldsymbol{U}_{2,n}(\boldsymbol{\lambda} \mid \hat{\boldsymbol{\phi}}) \equiv \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}},\boldsymbol{\lambda})\boldsymbol{b}_i - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}_i = \boldsymbol{0}, \tag{4.2}$$

where

$$\omega\left(\boldsymbol{x}_i;\boldsymbol{\phi},\boldsymbol{\lambda}\right) = 1 + \left\{\frac{1}{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})} - 1\right\}\exp\left(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\lambda}\right). \tag{4.3}$$

Further, define

$$\mathbf{G}_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left\{1 - \frac{\delta_i}{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})}\right\}\left\{\frac{\partial \boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}^{\mathrm{T}} + \frac{\delta_i}{\{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})\}^2}\frac{\partial \pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\boldsymbol{h}^{\mathrm{T}}(\boldsymbol{x}_i;\boldsymbol{\phi})\right].$$

We have the following assumptions before the statement of our main theorem.

Assumption 4.1. *The estimating equation $\boldsymbol{U}_{1,n}(\boldsymbol{\phi}) = \boldsymbol{0}$ has a unique solution $\hat{\boldsymbol{\phi}}$ and there exists a function $\boldsymbol{U}_1(\boldsymbol{\phi})$ such that $\boldsymbol{U}_{1,n}(\boldsymbol{\phi}) \to \boldsymbol{U}_1(\boldsymbol{\phi})$ uniformly as $n \to \infty$ and $\boldsymbol{U}_1(\boldsymbol{\phi}) = \boldsymbol{0}$ has a unique solution $\boldsymbol{\phi}^*$. Further, The estimating equation $\boldsymbol{U}_{2,n}(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*) = \boldsymbol{0}$ has a unique solution $\hat{\boldsymbol{\lambda}}$ and there exists a function $\boldsymbol{U}_2(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*)$ such that $\boldsymbol{U}_{2,n}(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*) \to \boldsymbol{U}_2(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*)$ uniformly as $n \to \infty$ and $\boldsymbol{U}_2(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*) = \boldsymbol{0}$ has a unique solution $\boldsymbol{\lambda}^*$.*

Assumption 4.2. *The limiting function $\boldsymbol{U}_1(\boldsymbol{\phi})$ is differentiable and $\partial_{\boldsymbol{\phi}}\boldsymbol{U}_1(\boldsymbol{\phi})$ is continuous on a compact set $\mathcal{G}_1$ containing $\boldsymbol{\phi}^*$. In addition, the limiting function $\boldsymbol{U}_2(\boldsymbol{\phi} \mid \boldsymbol{\phi}^*)$ is differentiable and $\partial_{\boldsymbol{\lambda}}\boldsymbol{U}_2(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*)$ is continuous on a compact set $\mathcal{G}_2$ containing $\boldsymbol{\lambda}^*$.*

Assumption 4.3. *The matrices* $\partial_{\boldsymbol{\phi}} \boldsymbol{U}_1(\boldsymbol{\phi}) \mid_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}$ *and* $\partial_{\boldsymbol{\lambda}} \boldsymbol{U}_2(\boldsymbol{\lambda} \mid \boldsymbol{\phi}^*) \mid_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*}$ *are both non-singular.*

Assumption 4.4. *There exists a function* $\mathbf{G}(\boldsymbol{\phi})$ *such that* $\mathbf{G}_n(\boldsymbol{\phi}) \to \mathbf{G}(\boldsymbol{\phi})$ *uniformly as* $n \to \infty$ *and* $\mathbf{G}(\boldsymbol{\phi}^*)$ *is non-singular.*

Assumptions 4.1 – 4.3 are commonly used in estimating equation theory and also ensure the existence of $\boldsymbol{\lambda}^*$ and $\boldsymbol{\phi}^*$; see Newey and McFadden (1994) and Kim and Rao (2009) for more details. Note that $\boldsymbol{\lambda}^*$ and $\boldsymbol{\phi}^*$ are the probability limits of $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\phi}}$, respectively. Assumption 4.3 also guarantees the existence of $\boldsymbol{\beta}^*$, which is the probability limit of $\hat{\boldsymbol{\beta}}$ in (3.12).Assumption 4.4 ensures the existence of the nuisance parameter $\boldsymbol{\kappa}^*$ defined in the following Theorem 4.1, which helps us represent the influence function for the proposed estimator. If the maximum likelihood estimation is used to estimate $\boldsymbol{\phi}$, then $\boldsymbol{\phi}^*$ can be understood as the minimizer of the cross-entropy

$$H(\pi_1, \pi_1^{(0)}(\boldsymbol{\phi}))=-E\left[\pi_1(\boldsymbol{x}) \log\left\{\pi_1^{(0)}(\boldsymbol{x};\boldsymbol{\phi})\right\} + \{1-\pi_1(\boldsymbol{x})\} \log\left\{1-\pi_1^{(0)}(\boldsymbol{x};\boldsymbol{\phi})\right\}\right]$$

with respect to $\boldsymbol{\phi}$, where $\pi_1(\boldsymbol{x}) = P(\delta = 1 \mid \boldsymbol{x})$ is the true response probability. Now, using (4.3), we can treat

$$\hat{\theta}_{\mathrm{APSW}} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \omega(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) y_i := \hat{\theta}_{\mathrm{APSW}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}}) \tag{4.4}$$

as a function of $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})$ and apply the standard Taylor linearization to obtain the following theorem, whose proof is presented in the Appendix.

Theorem 4.1. *Let* $\hat{\theta}_{\mathrm{APSW}}$ *in (4.4) be the PSW estimator under the augmented PS model in (3.8) with* $\hat{\boldsymbol{\phi}}$ *and* $\hat{\boldsymbol{\lambda}}$ *satisfying (4.1) and (4.2), respectively. Under the regularity conditions described in the Appendix, we have*

$$\hat{\theta}_{\mathrm{APSW}} - \theta = \frac{1}{n}\sum_{i=1}^{n} \eta(\boldsymbol{x}_i, y_i, \delta_i) + o_p(n^{-1/2}), \tag{4.5}$$

where

$$\begin{aligned}\eta(\boldsymbol{x}_i, y_i, \delta_i) = \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^* - \theta + \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*\right) \\ + \left\{1 - \frac{\delta_i}{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi}^*)}\right\} \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\kappa}^*,\end{aligned} \tag{4.6}$$

$\boldsymbol{h}_i = \boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{\phi})$, $\omega(\boldsymbol{x};\boldsymbol{\phi},\boldsymbol{\lambda})$ *is defined in (4.3), and* $\boldsymbol{\kappa}^*$ *is the probability limit of* $\hat{\boldsymbol{\kappa}}$ *that satisfies*

$$\sum_{i=1}^{n} \delta_i (\hat{\pi}_{1,i}^{-1} - 1)\left\{\exp(\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}})(y_i - \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \boldsymbol{h}_i^{\mathrm{T}}\hat{\boldsymbol{\kappa}}\right\}\left[\frac{\partial}{\partial\boldsymbol{\phi}} \mathrm{logit}\left\{\pi_1^{(0)}(\boldsymbol{x}_i;\boldsymbol{\phi})\right\}\Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}\right] = \mathbf{0}, \tag{4.7}$$

where $\hat{\pi}_{1,i} = \pi_1^{(0)}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})$.

In (4.6), $\eta(\boldsymbol{x}, y, \delta)$ is called the influence function of $\hat{\theta}_{\mathrm{APSW}}$ (Tsiatis, 2006). Note that we can obtain

$$\begin{aligned} E\{\eta(\boldsymbol{X}, Y, \delta)\} = & E\{\delta\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)Y\} - \theta \\ & + E\left[\{1 - \delta\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)\}\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\beta}^*\right] \\ & + E\left[\left\{1 - \frac{\delta}{\pi_1^{(0)}(\boldsymbol{X}; \boldsymbol{\phi}^*)}\right\}\boldsymbol{h}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\kappa}^*\right], \end{aligned}$$

where the second term is equal to zero by the definition of $\boldsymbol{\lambda}^*$ and the third term is equal to zero by the definition of $\boldsymbol{\phi}^*$. If the outcome regression model is correctly specified, we have

$$\begin{aligned} E\{\delta\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)Y\} &= E\left\{\delta\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\beta}^*\right\} \\ &= E\left\{\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\beta}^*\right\} = E(Y), \end{aligned}$$

where the second equality follows from the definition of $\boldsymbol{\lambda}^*$. Furthermore, the correct specification of the outcome regression model gives $\boldsymbol{\kappa}^* = \mathbf{0}$. Thus, we can summarize the asymptotic result in the outcome regression model as follows.

Corollary 4.1. *Suppose that the regularity conditions in Theorem 4.1 hold and the outcome regression model is correctly specified. Then,*

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{APSW}} - \theta\right) \xrightarrow{\mathcal{L}} N(0, V_{\mathrm{OR}}), \tag{4.8}$$

where

$$V_{\mathrm{OR}} = \mathrm{var}\{E(Y \mid \boldsymbol{X})\} + E\left[\delta\{\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)\}^2 \,\mathrm{var}(Y \mid \boldsymbol{X})\right]. \tag{4.9}$$

Now, consider the case where the propensity score model is correctly specified. In this case, we obtain $\boldsymbol{\lambda}^* = \mathbf{0}$ and $\pi_1^{(0)}(\boldsymbol{x}; \boldsymbol{\phi}^*) = P(\delta = 1 \mid \boldsymbol{x})$. Thus,

$$E\{\delta\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)Y\} = E\left[P(\delta = 1 \mid \boldsymbol{X})\{\pi_1^{(0)}(\boldsymbol{X}; \boldsymbol{\phi}^*)\}^{-1}Y\right] = E(Y).$$

Therefore, we can obtain the following result.

Corollary 4.2. *Suppose that the regularity conditions in Theorem 4.1 hold and the propensity score model is correctly specified. Then,*

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{APSW}} - \theta\right) \xrightarrow{\mathcal{L}} N(0, V_{\mathrm{PS}}), \tag{4.10}$$

where

$$V_{\mathrm{PS}} = \mathrm{var}(Y) + E\left[\left\{\frac{1}{\pi_1(\boldsymbol{X})} - 1\right\}\{Y - \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\beta}^* - \boldsymbol{h}^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{\kappa}^*\}^2\right]$$

and $\pi_1(\boldsymbol{X}) = P(\delta = 1 \mid \boldsymbol{X})$.

Thus, the efficiency gain using the estimated propensity score function over the true one is visible only when the PS model is true but the OR model is incorrect.

If the two models, the OR model and the PS model, are correctly specified, then the two variance forms are equal to

$$\operatorname{var}\{\eta(\boldsymbol{X}, Y, \delta)\} = \operatorname{var}\{E(Y \mid \boldsymbol{X})\} + E\left[\{P(\delta = 1 \mid \boldsymbol{X})\}^{-1} \operatorname{var}(Y \mid \boldsymbol{X})\right],$$

which is equal to the lower bound of the semiparametric efficient estimator considered in Robins, Rotnitzky and Zhao (1994). The influence function in (4.6) can also be used to develop the linearized variance estimation formula for $\hat{\theta}_{\mathrm{APSW}}$ regardless of whether the outcome regression model or the propensity score model holds.

In summary, the influence function in (4.6) can be reduced to the following limits under each model as follows:

$$\begin{array}{ll} \text{Correct OR:} & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \boldsymbol{\beta} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{0} \\ & \eta(\boldsymbol{x}_i, y_i, \delta_i) = \delta_i \underbrace{\omega_i^*}_{\pi_{1,i}^{-1}} y_i - \theta + (1 - \delta_i \underbrace{\omega_i^*}_{\pi_{1,i}^{-1}}) \boldsymbol{b}_i^{\mathrm{T}} \overbrace{\boldsymbol{\beta}^*} + (1 - \delta_i / \underbrace{\pi_{1,i}^{(0)}(\phi^*)}_{\pi_{1,i}}) \boldsymbol{h}_i^{\mathrm{T}} \overbrace{\boldsymbol{\kappa}^*}, \\ \text{Correct RP:} & \end{array} \tag{4.11}$$

where $\pi_{1,i} = P(\delta = 1 \mid \boldsymbol{x}_i)$, the upward arrows point to the limits under the correct OR model and the downward arrows points to the limits under the correct RP model.

In the next section, we address how to obtain the robust propensity score weighting estimator against model misspecification of the propensity score model and outliers in the outcome regression model.

5. Adding outlier-robustness

In many real cases, outliers exist in addition to missingness. In this case, imposing robustness against outliers is an important practical problem. This section discusses how to allow robust inference against outliers in the outcome variable in the context of doubly robust estimator. In the presence of outliers, one may use the heavy-tailed distribution (e.g. t-distribution) (Lange, Little and Taylor, 1989) to allow robust inference against outliers. However, it is not straightforward to extend the indirect model calibration condition to the t-distribution.

Basu et al. (1998) introduced the density power divergence as a generalization of Kullback-Leibler divergence to expand the class of statistical models by introducing an additional scale parameter γ. Density power divergence is also called γ-power divergence (Eguchi, 2021). We can use the γ-power divergence to develop an outlier robust imputation method. Specifically, we employ the

following M-estimator to handle the misspecification of the propensity score model. Define

$$Q_\gamma(\boldsymbol{\beta} \mid \boldsymbol{\lambda}) = \sum_{i=1}^{n} \delta_i(\hat{d}_i - 1)\Psi_\gamma\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right) g_i(\boldsymbol{\lambda}), \tag{5.1}$$

where Ψ_γ is an objective function that reduces the effect of outliers, and $g_i = \exp(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\lambda})$ is the adjustment term to achieve the self-efficiency. Thus, under some suitable choice of Ψ_γ, the resulting estimator is not only doubly robust, but also outlier-robust. Note that the Huber-type loss function could be used in (5.1), but it is computationally less attractive as the Huber loss function is non-smooth.

To incorporate the γ-divergence, we now consider the following minimization problem,

$$\begin{aligned}&\arg\min_{(\boldsymbol{\beta},\sigma^2)} Q_\gamma(\boldsymbol{\beta};\sigma^2 \mid \boldsymbol{\lambda}) \\ &\quad = -(2\pi\sigma^2)^{-\frac{\gamma}{1+\gamma}} \sum_{i=1}^{n} \delta_i(\hat{d}_i - 1) \exp\left\{-\frac{\gamma}{2\sigma^2}(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\right\} g_i(\boldsymbol{\lambda}).\end{aligned} \tag{5.2}$$

That is, in (5.1), we use

$$\Psi_\gamma(x) = -(2\pi\sigma^2)^{-\frac{\gamma}{1+\gamma}} \exp\left\{-\frac{\gamma}{2\sigma^2}x^2\right\}. \tag{5.3}$$

As a result, the optimization problem (5.2) can be solved by the iterated reweighted least squares method. That is, we use

$$\sum_{i=1}^{n} \delta_i\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)(\hat{d}_i - 1) g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\boldsymbol{b}_i = \mathbf{0} \tag{5.4}$$

$$\sum_{i=1}^{n} \delta_i\left\{\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 - \frac{\sigma^2}{1+\gamma}\right\}(\hat{d}_i - 1) g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta},\sigma^2) = 0 \tag{5.5}$$

to estimate $\boldsymbol{\beta}$ and σ^2 for given $\boldsymbol{\lambda}$, where

$$q_{\gamma,i}(\boldsymbol{\beta},\sigma^2) = \exp\{-0.5\gamma\sigma^{-2}(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\}. \tag{5.6}$$

Note that during the iterative reweighted least squares estimation procedure, if $|y_i - \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}| \gg 0$, then the effect of the i-th subject for the estimation of $\boldsymbol{\beta}$ will be greatly mitigated as $\hat{w}_{\gamma,i}$ will be smaller, indicating the robustness of our proposed method. The parameter $\gamma > 0$ plays the role of the tuning parameter for robust estimation. As the value of γ increases, the estimator becomes more robust but less efficient.

Let $\hat{\boldsymbol{\beta}}_q$ and $\hat{\sigma}_q^2$ be the solution to the above estimating equations, (5.4) and (5.5). Further, let $\hat{q}_{\gamma,i} = q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)$. We can construct robust imputation with $\hat{y}_i = \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_q$ easily. Furthermore, we can use the robust imputation to construct

robust propensity score weights with calibration constraints. Specifically, by (5.4), we can obtain

$$\sum_{i=1}^{n} \delta_i \left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q \right) \boldsymbol{b}_i (\hat{d}_i - 1) g_i(\hat{\boldsymbol{\lambda}}) \hat{q}_{\gamma,i} = \mathbf{0}. \tag{5.7}$$

In a similar fashion of derivation for Lemma 2.2, it yields that

$$\sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i)\hat{y}_i\} = \sum_{i=1}^{n} \delta_i \hat{\omega}_{\gamma,i} y_i, \tag{5.8}$$

where $\hat{y}_i = \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q$ and

$$\omega_{\gamma,i}(\boldsymbol{x}_i; \boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2, \hat{\boldsymbol{\phi}}) = 1 + (d_i(\hat{\boldsymbol{\phi}}) - 1) g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta}, \sigma^2). \tag{5.9}$$

Note that (5.8) is the self-efficiency of the propensity score weights in (5.9). The final PS weights satisfy

$$\sum_{i=1}^{n} \delta_i \hat{\omega}_{\gamma,i} \boldsymbol{b}_i = \sum_{i=1}^{n} \boldsymbol{b}_i. \tag{5.10}$$

Thus, for given $\hat{q}_{\gamma,i}$'s, we can compute $\hat{\boldsymbol{\lambda}}$ from calibration constraints in (5.10). That is, we can treat (5.4), (5.5), and (5.10) as the estimating equations for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\lambda}$. Note that (5.9) is equivalent to (3.9) for $q_{\gamma,i} = 1$. The additional factor $q_{\gamma,i}(\boldsymbol{\beta}, \sigma^2)$ controls the effect of outliers in the final weighting.

Let $\boldsymbol{\beta}^*$ and σ^{*2} be the probability limits of $\hat{\boldsymbol{\beta}}_q$ and $\hat{\sigma}_q^2$, respectively. The following theorem presents the Taylor linearization for our proposed estimator in (5.8).

Theorem 5.1. *Let* $\hat{\theta}_{\mathrm{APSW},\gamma} = n^{-1} \sum_{i=1}^{n} \delta_i \hat{\omega}_{\gamma,i} y_i$ *be the propensity score weighting estimator under the augmented propensity score model in (3.8) with* $\hat{\boldsymbol{\phi}}$ *and* $\hat{\boldsymbol{\lambda}}$ *satisfying (4.1) and (5.7),* $\hat{\boldsymbol{\beta}}_q$ *and* $\hat{\sigma}_q^2$ *minimizing* (5.2) *respectively. Under the regularity conditions described in the Appendix, we have*

$$\hat{\theta}_{\mathrm{APSW},\gamma} - \theta = \frac{1}{n} \sum_{i=1}^{n} \eta_\gamma(\boldsymbol{x}_i, y_i, \delta_i) + o_p(n^{-1/2}), \tag{5.11}$$

where

$$\begin{aligned} \eta_\gamma(\boldsymbol{x}_i, y_i, \delta_i) = {} & \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu} - \theta + \delta_i \omega_{\gamma,i}^* (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}^*) + \{1 - \delta_i d_i(\boldsymbol{\phi}^*)\} \boldsymbol{\kappa}^{*\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\phi}^*) \\ & + \delta_i \{\omega_{\gamma,i}^* - 1\} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\zeta}^* \\ & + \delta_i \{\omega_{\gamma,i}^* - 1\} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 \nu^* \left\{ (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 - \sigma^{*2}/(1 + \gamma) \right\}, \end{aligned} \tag{5.12}$$

where $\omega_{\gamma,i}^* = \omega_{\gamma,i}(\boldsymbol{x}_i; \boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2}, \boldsymbol{\phi}^*)$ *and* $\boldsymbol{\mu}^*$, $\boldsymbol{\kappa}^*$, ν^*, $\boldsymbol{\zeta}^\star$ *are the probability limits of the solutions for the equations presented in the Appendix.*

Note that the first line and the last line has the same structure as in (4.6) of Theorem 4.1. The other two lines are the additional part of the influence functions that reflect the effect of the gamma-divergence model. Specifically, the second line reflects the estimation effect of $\hat{\boldsymbol{\beta}}_q$, and the third line reflects the estimation effect of $\hat{\sigma}_q^2$.

Regarding the choice of the tuning parameter γ, we can use cross-validation to choose the optimal γ. Specifically, we can randomly partition the sample into K-groups $(S_1, \cdots, S_K)$ and then compute

$$\text{MSPE}(\gamma) = \sum_{k=1}^{K} \sum_{i \in S_k} \delta_i \left\{ y_i - \hat{y}_i^{(-k)}(\gamma) \right\}^2$$

where $\hat{y}_i^{(-k)}(\gamma) = \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q^{(-k)}$ and $\hat{\boldsymbol{\beta}}_q^{(-k)}$ is obtained by solving the same estimation formula using the sample in S_k^c. The minimizer of $\text{MSPE}(\gamma)$ can be used as the optimal value of the tuning parameter. Furthermore, instead of cross-validation, we can use other approximation-based methods (Sugasawa and Yonekura, 2021; Basak, Basu and Jones, 2020) that are computationally less expensive. As we can find in our simulation study in Section 6, the optimal choice of γ is not critical. The simulation result shows similar performances for different values of γ.

For variance estimation, we can use the influence function in (5.12) to construct a linearization variance estimator. Specifically, the linearization variance estimator for $\hat{\theta}_{\text{APSW},\gamma}$ for a given γ is

$$\hat{\text{V}}(\hat{\theta}_{\text{APSW},\gamma}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\eta}_\gamma(\boldsymbol{x}_i, y_i, \delta_i) - \bar{\eta}_\gamma)^2, \tag{5.13}$$

where

$$\begin{aligned} &\hat{\eta}_\gamma(\boldsymbol{x}_i, y_i, \delta_i) \\ &= \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\mu}} + \delta_i \hat{\omega}_{\gamma,i} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\mu}}) + \left\{ 1 - \delta_i d_i(\hat{\boldsymbol{\phi}}) \right\} \hat{\boldsymbol{\kappa}}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}}) \\ &\quad + \delta_i \{ \hat{\omega}_{\gamma,i} - 1 \} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q) \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\zeta}} \\ &\quad + \delta_i \{ \hat{\omega}_{\gamma,i} - 1 \} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 \hat{\nu} \left\{ (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 - \hat{\sigma}_q^2 / (1+\gamma) \right\}, \end{aligned} \tag{5.14}$$

$\hat{\omega}_{\gamma,i} = \omega_{\gamma,i}(\boldsymbol{x}_i; \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2, \hat{\boldsymbol{\phi}})$ and

$$\bar{\eta}_\gamma = \frac{1}{n} \sum_{i=1}^{n} \hat{\eta}_\gamma(\boldsymbol{x}_i, y_i, \delta_i).$$

Note that $\hat{\boldsymbol{\beta}}_q$, $\hat{\sigma}_q^2$ and $\hat{\boldsymbol{\lambda}}$ are estimated from (5.4), (5.5) and (5.10), $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\kappa}}$, $\hat{\nu}$ and $\hat{\boldsymbol{\zeta}}$ can be estimated from the procedure listed in the Appendix. A flowchart of the inference procedure for the robust augmented PSW estimator is presented in Fig 2, where the equations (E.2) and (E.6) are estimating equations presented in Appendix E.

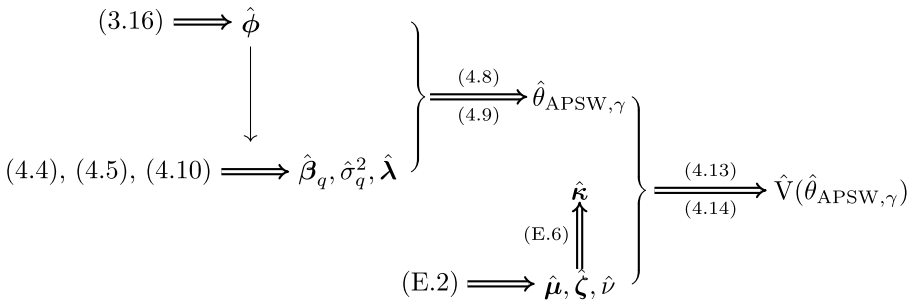

FIG 2. *Illustration of the inference procedure for the robust augmented PSW estimator.*

6. Empirical studies

6.1. Simulation study

We examine the performance of proposed methods under various propensity score models. Let $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^{\mathrm{T}}$, where $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_{5\times 5})$ follow the standard multivariate normal distribution. The simulation experiment can be described as a 2×2 factorial design with two factors. The sample size is 100 and the Monte Carlo sample size is $1,000$. The first factor is the outcome regression model (OM1, OM2) and the second factor is the propensity score model (PM1, PM2).

The models for generating Y are described as follows: OM1, where $Y \mid \boldsymbol{X}$ follows a normal distribution with mean $E(Y \mid \boldsymbol{X}) = 1 + X_1 + X_2 + X_3 + X_4 + X_5$ and variance 1; OM2, where $Y \mid \boldsymbol{X}$ follows a normal distribution with $E(Y \mid \boldsymbol{X}) = 1 + 0.25\{\cos(3\pi X_1) + 1\}X_2^3 + 0.25\{\sin(3\pi X_4) - 1\}X_3^3 + \cos(3\pi X_5)$ and variance 1. For calibration, we use $\boldsymbol{b}(\boldsymbol{X}) = (1, X_1, X_2, X_3, X_4, X_5)^{\mathrm{T}}$ as the calibration variable. Thus, the implicit model calibration using $\boldsymbol{b}(\boldsymbol{X}) = (1, X_1, X_2, X_3, X_4, X_5)^{\mathrm{T}}$ is justified under the OM1 outcome model, but it is incorrectly specified under the OM2 outcome model. In addition, the models for generating δ can be described as follows: PM1, where $\delta \mid \boldsymbol{X}$ follows Bernoulli distribution with $\mathrm{logit}\{P(\delta = 1 \mid \boldsymbol{X})\} = \phi_0 + 0.25X_1 + 0.25X_2 + 0.25X_3 + 0.25X_4 + 0.25X_5$, where ϕ_0 is chosen to achieve 60% response rates; PM2, where $\delta \mid \boldsymbol{X}$ follows the Bernoulli distribution with $P(\delta = 1 \mid \boldsymbol{X}) = 0.8$ if $a + X_1 + X_2 > 0$, and $P(\delta = 1 \mid \boldsymbol{X}) = 0.4$ otherwise, where a is chosen to achieve 60% response rates. Further, we use

$$\mathrm{logit}\left\{\pi_1^{(0)}(\boldsymbol{X}; \boldsymbol{\phi})\right\} = \phi_0 + \phi_1 X_1 + \phi_2 X_2 + \phi_3 X_3 + \phi_4 X_4 + \phi_5 X_5. \tag{6.1}$$

as the working model for the propensity score function, where $\mathrm{logit}(x) = \log\{x/(1-x)\}$. Thus, the working propensity score model is correctly specified under PM1 only.

For each sample, we compute the following propensity score estimators.

(i) *Mean of the complete cases* (CC): $\hat{\theta}_{\mathrm{CC}} = n_1^{-1} \sum_{i=1}^{n} \delta_i Y_i$, that is, mean of observed Y_i's, where $n_1 = \sum_{i=1}^{n} \delta_i$.

(ii) *Maximum likelihood estimation* (MLE): the classical propensity score estimator $\hat{\theta}_{\mathrm{PSW}} = n^{-1} \sum_{i=1}^{n} \delta_i \hat{d}_i Y_i$ using $\hat{d}_i = 1/\pi_1^{(0)}(\boldsymbol{X}_i; \hat{\boldsymbol{\phi}})$ as the estimated propensity score weight, where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$ from the working propensity score model in (6.1).

(iii) *Entropy balancing method in Hainmueller (2012)*: $\hat{\theta}_{\mathrm{HM}} = n^{-1} \sum_{i=1}^{n} \delta_i \hat{\omega}_i Y_i$, where $\hat{\omega}_i$ is obtained from (2.7).

(iv) *Regularized calibration method in Tan (2019)*: $\hat{\theta}_{\mathrm{Tan}} = n^{-1} \sum_{i=1}^{n} \delta_i \{\pi_1^{(0)}(\boldsymbol{X}_i; \hat{\boldsymbol{\phi}})\}^{-1} Y_i$, where the logistic regression coefficients $\hat{\boldsymbol{\phi}}$ are obtained by minimizing $\ell_{\mathrm{CAL}}(\boldsymbol{\phi}) = \sum_{i=1}^{n} [\delta_i \exp\{-\boldsymbol{\phi}^{\mathrm{T}} f(\boldsymbol{X}_i)\} - (1 - \delta_i) \boldsymbol{\phi}^{\mathrm{T}} f(\boldsymbol{X}_i)]$, where we take the identity mapping for $f(\cdot)$ in this simulation.

(v) *Augmented propensity score weighting estimator* (APS): $\hat{\theta}_{\mathrm{APSW}} = n^{-1} \sum_{i=1}^{n} \delta_i \hat{\omega}_i^* Y_i$ using $\hat{\omega}_i^*$ in (3.9).

(vi) *Augmented propensity score weighting estimator with γ-divergence* (APS_γ): the augmented propensity score weighting estimator in (5.8) using $\gamma = 0.3, 0.5, 0.7, 1$.

TABLE 1

Bias, standard error (S.E.), root mean square error (RMSE) for estimators comparison with sample size 100 *and Monte Carlo sample* 1,000. *All criteria are multiplied by 10. The errors from outcome models are generated from i.i.d. standard Gaussian distribution. CC: mean of the complete cases; MLE: maximum likelihood estimation; APS: augmented propensity score weighting estimator;* APS_γ*: augmented propensity score weighting with* γ*-divergence estimator, for* $\gamma = 0.3, 0.5, 0.7, 1$*; HM: entropy balancing method in Hainmueller (2012); Tan: regularized calibration method in Tan (2019)*

Model	Criteria	Method								
		CC	MLE	APS	$\mathrm{APS}_{0.3}$	$\mathrm{APS}_{0.5}$	$\mathrm{APS}_{0.7}$	APS_1	HM	Tan
OM1PM1	Bias	4.62	0.04	−0.01	−0.00	−0.02	−0.02	−0.00	−0.01	0.03
	S.E.	3.26	3.02	2.77	2.72	2.72	2.75	2.72	2.76	2.77
	RMSE	5.65	3.02	2.77	2.72	2.71	2.74	2.72	2.76	2.77
OM2PM1	Bias	0.03	−0.02	−0.03	−0.02	−0.01	0.01	−0.03	−0.03	−0.02
	S.E.	2.75	2.95	2.78	2.48	2.48	2.47	2.45	2.70	2.69
	RMSE	2.75	2.95	2.78	2.48	2.47	2.47	2.45	2.70	2.69
OM1PM2	Bias	3.72	−0.54	−0.02	−0.01	−0.05	−0.05	−0.02	−0.03	0.01
	S.E.	3.28	3.51	2.76	2.76	2.76	2.71	2.73	2.75	2.75
	RMSE	4.96	3.55	2.76	2.76	2.76	2.72	2.73	2.75	2.75
OM2PM2	Bias	1.24	−0.61	−0.22	−0.08	−0.03	−0.04	−0.04	−0.26	0.01
	S.E.	2.76	4.19	2.93	2.60	2.58	2.59	2.55	2.81	2.83
	RMSE	3.03	4.23	2.94	2.60	2.58	2.59	2.54	2.82	2.83

Figure 3 shows the results of the simulation study above, where the dashed line represents the true parameter $\theta = E(Y)$. Table 1 presents the corresponding bias, standard error, and root mean square error for each method. Since the consistency of all estimators is guaranteed under OM1PM1, we can see that all estimators give similar performances except for the mean of the complete cases method. However, under OM1PM2, the consistency of the generalized linear model method is no longer guaranteed, while that of other estimators is still

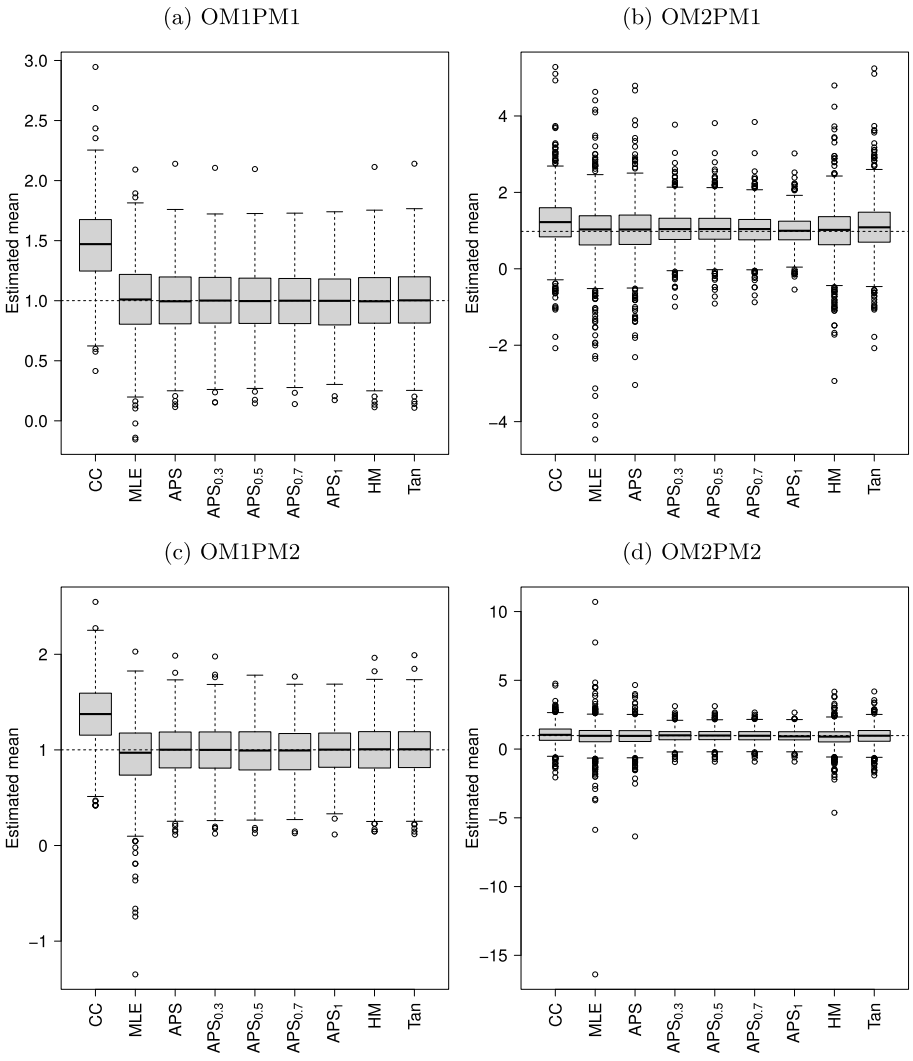

FIG 3. *Boxplots for estimators comparison with sample size* 100 *and Monte Carlo sample* 1,000. *The errors from outcome models are generated from i.i.d. standard Gaussian distribution. CC: mean of the complete cases; MLE: maximum likelihood estimation; APS: augmented propensity score weighting estimator;* APS_γ*: augmented propensity score weighting with* γ*-divergence estimator, for* $\gamma = 0.3, 0.5, 0.7, 1$*; HM: entropy balancing method in Hainmueller (2012); Tan: regularized calibration method in Tan (2019).*

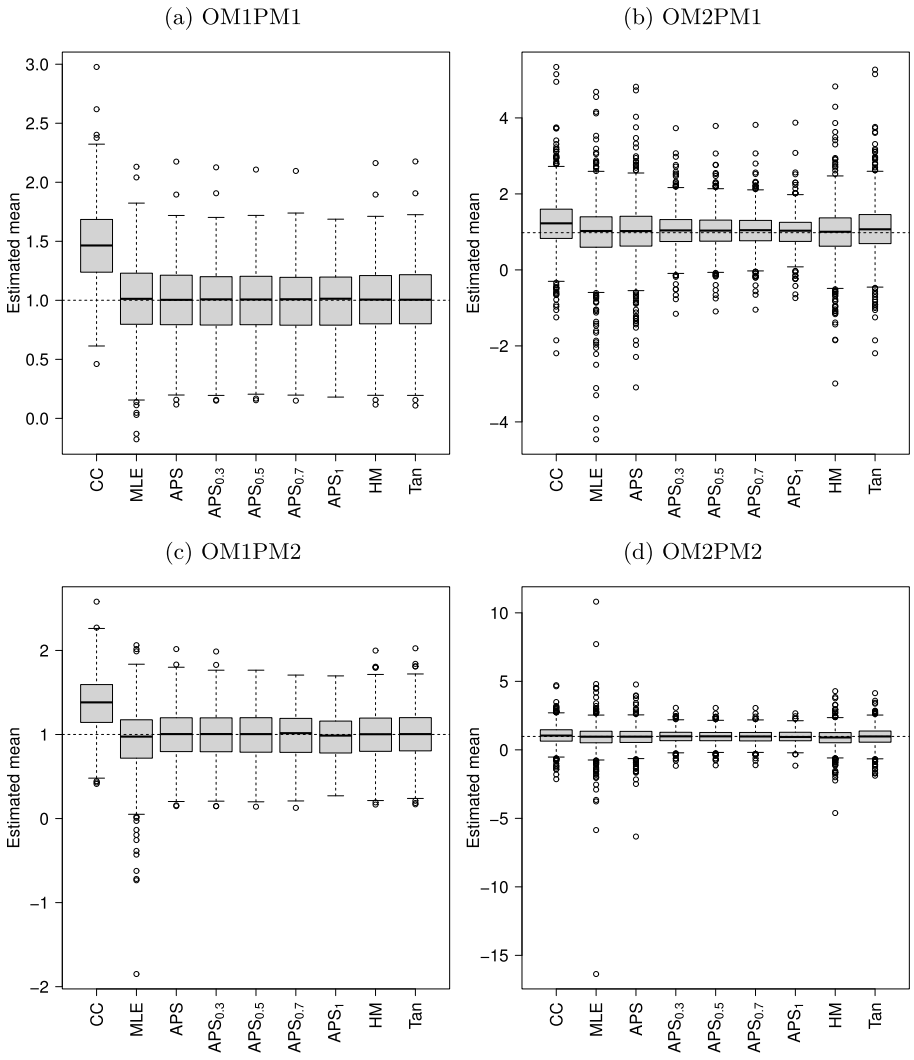

FIG 4. *Boxplots for estimators comparison with sample size* 100 *and Monte Carlo sample* $1,000$*; 20% of the samples are contaminated with the additive noise that eight times a random variable generated from Student's t distribution with degree freedom of 8. The errors from outcome models are generated from i.i.d. standard Gaussian distribution. CC: mean of the complete cases; MLE: maximum likelihood estimation; APS: augmented propensity score weighting estimator;* APS_γ*: augmented propensity score weighting with* γ*-divergence estimator, for* $\gamma = 0.3, 0.5, 0.7, 1$*; HM: entropy balancing method in Hainmueller (2012); Tan: regularized calibration method in Tan (2019).*

TABLE 2

Bias, standard error (S.E.), root mean square error (RMSE) for estimators comparison with sample size 100 *and Monte Carlo sample* 1,000*; 20% of the samples are contaminated with the additive noise that eight times a random variable generated from Student's t distribution with degree freedom of 8. The errors from outcome models are generated from i.i.d. standard Gaussian distribution. All criteria are multiplied by 10. CC: mean of the complete cases; MLE: maximum likelihood estimation; APS: augmented propensity score weighting estimator;* APS_γ*: augmented propensity score weighting with* γ*-divergence estimator, for* $\gamma = 0.3, 0.5, 0.7, 1$*; HM: entropy balancing method in Hainmueller (2012); Tan: regularized calibration method in Tan (2019)*

Model	Criteria	Method								
		CC	MLE	APS	$\text{APS}_{0.3}$	$\text{APS}_{0.5}$	$\text{APS}_{0.7}$	APS_1	HM	Tan
OM1PM1	Bias	4.62	0.00	−0.05	−0.10	−0.09	−0.14	−0.13	−0.04	0.69
	S.E.	6.48	6.62	6.59	4.52	4.46	4.44	4.36	6.51	6.50
	RMSE	7.95	6.62	6.59	4.52	4.46	4.44	4.36	6.51	6.53
OM2PM1	Bias	0.03	−0.06	−0.06	−0.01	−0.07	−0.01	−0.18	−0.06	−0.05
	S.E.	6.11	6.51	6.54	4.45	4.38	4.37	4.21	6.43	6.24
	RMSE	6.11	6.51	6.53	4.45	4.37	4.37	4.21	6.42	6.23
OM1PM2	Bias	3.66	−0.64	−0.14	−0.07	−0.02	−0.04	0.05	−0.14	0.38
	S.E.	6.39	7.43	6.54	4.56	4.62	4.55	4.55	6.45	6.43
	RMSE	7.36	7.45	6.54	4.56	4.61	4.55	4.54	6.45	6.44
OM2PM2	Bias	1.18	−0.71	−0.34	−0.18	−0.16	−0.13	−0.05	−0.37	0.17
	S.E.	6.04	7.67	6.57	4.37	4.39	4.30	4.37	6.44	6.15
	RMSE	6.15	7.69	6.58	4.37	4.39	4.30	4.36	6.44	6.15

valid. In OM1PM2, the covariate balancing condition becomes critical, and so all methods satisfying the covariate balancing will be unbiased. In OM2PM1, the proposed augmented PSW estimator is still consistent, as the propensity score model is correctly specified. Our proposed augmented PSW estimator with γ-divergence has relatively lower root mean square error compared to other estimators. Our proposed augmented PSW estimator with γ-divergence is biased in this scenario due to the misspecification of the outcome model, where other estimators are built under the correct specified response model. In OM2PM2, our proposed augmented PSW estimator with γ-divergence is robust against other estimators, with fewer outliers and the smallest root mean square error.

We also investigated the performance of the proposed linearization variance estimator in (5.13). We computed confidence intervals based on asymptotic normality. The results are presented in Table 3. The performances are satisfactory when the ourcome regression model is specified correctly.

In addition, we check the robustness of the augmented propensity score weighting estimator with γ-divergence against outliers. After the data are generated under the same 2×2 factorial design as in the previous simulation setting, additional noise eight times a random variable generated from Student's t distribution with degree freedom of 8 is added to 20% of the observed outcomes. Again, the sample size is 100 and the Monte Carlo sample size is 1,000.

Figure 4 shows the performance of various estimators, where the dashed line represents the true mean of $y_i's$. Table 2 presents the corresponding bias, standard error, and root mean square error for each method. Compared to other estimators, in all four scenarios, our proposed augmented propensity score weight-

TABLE 3
Linearized variance estimation for gamma divergence method where the errors are i.i.d. from standard Gaussian distribution

Model	Nominal coverage rate	Method			
		$\text{APS}_{0.3}$	$\text{APS}_{0.5}$	$\text{APS}_{0.7}$	$\text{APS}_{1.0}$
OM1RM1	90%	88.5 %	89.3 %	89.0 %	90.4 %
	95%	94.2 %	94.5 %	94.4 %	96.0 %
OM1RM2	90%	89.2 %	89.8 %	90.6 %	90.3 %
	95%	94.5 %	94.8 %	95.2 %	95.1 %

ing estimator with γ-divergence apparently gives the smallest root mean square error, which validates our robustness claim in Section 5.

6.2. Real data application

We further check our proposed estimators for artificial missingness with real data from the California API program (`http://api.cde.ca.gov/` or survey package (Lumley, 2010) in R (R Core Team, 2022)). Within the data set, standardized student tests are performed to calculate the API for California schools. Specifically, we select the API for year 2000 (api00) as the response variable. For covariates, we set the API for year 1999 (api99) as X_1, the percentage of students eligible for subsidized meals (meals) as X_2, the percentage of English language learners (ell) as X_3, the average level of parental education (avg) as X_4, the percentage of fully qualified teachers (full) as X_5, and the number of students enrolled (enroll) as X_6, where the words in parentheses are the abbreviations for variable names in the dataset. We artificially created the missingness with the following two response mechanisms: PM3, where $\delta \mid \boldsymbol{X}$ follows the Bernoulli distribution with $\text{logit}\{P(\delta = 1 \mid \boldsymbol{X})\} = \phi_0 + 2X_1 + X_2 + 0.5X_3$, and ϕ_0 is chosen to achieve the 60% response rates; PM4, where $\delta \mid \boldsymbol{X}$ follows the Bernoulli distribution with $P(\delta = 1 \mid \boldsymbol{X}) = 0.8$, if $a + 2X_1 + X_2 + X_3 + X_4 + X_5 + X_6 > 0$, $P(\delta = 1 \mid \boldsymbol{X}) = 0.4$ otherwise, and a is chosen to achieve the response rates 60%.

We compare the same estimators as in previous subsections with 1000 Monte Carlo samples. In particular, the mean of the complete cases method does not behave well in both cases and will affect the scale of the resultant box plots, so we do not show its performance. Furthermore, we adopt the MSPE as the 5-fold cross-validation criterion for the selection strategy of γ, as described in Section 5. For each Monte Carlo sample, we select the best γ from $\{0.1, 0.2, \ldots, 1\}$ and denote the corresponding estimator as APS_γ. The results are summarized with box plots in Figure 5, where the dashed line denotes the true population mean. As we can see, the estimator APS_γ always gives the unbiased estimate with a relatively low root mean square error, outperforming other estimators.

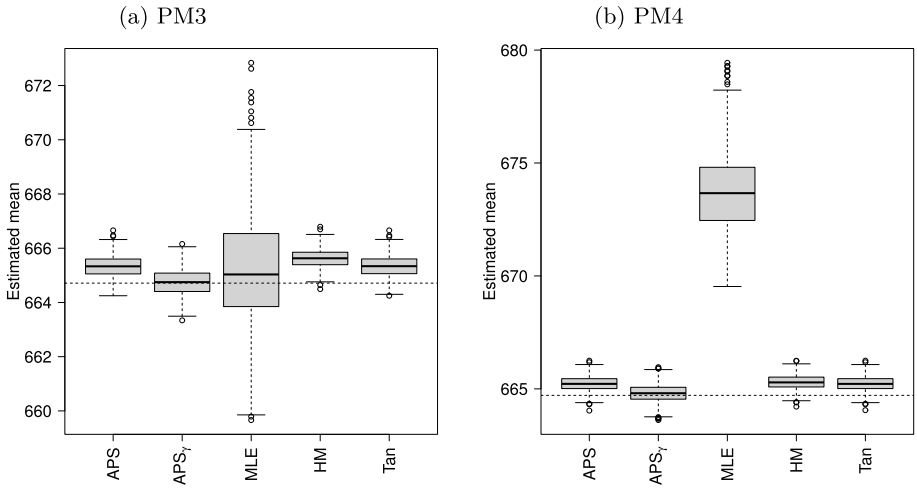

FIG 5. *Estimators comparison in real data with* 1000 *Monte Carlo sample, with artificially missingness mechanism PM3 and PM4. APS: augmented propensity score weighting estimator;* APS_γ*: augmented propensity score weighting with* γ*-divergence estimator, where* γ *is chosen by the cross-validation method based on MSPE criterion; MLE: maximum likelihood estimation; HM: entropy balancing method in Hainmueller (2012); Tan: regularized calibration method in Tan (2019).*

7. Application to CEAP data

The Conservation Effects Assessment Project (CEAP) is a program initiated by the United States Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS). CEAP collects and analyzes data from various sources, including field studies, monitoring sites, and modeling, to evaluate the impact of water and wind erosion. Further details of the CEAP data can be found in Berg and Yu (2019). The farmer's interview data together with the NRI data serve as input to the revised Universal Soil Loss Equation (RUSLE2) to generate a measure of sheet and rill erosion. In addition, RUSLE2 is also an advancement of a traditional approximation called the Universal Soil Loss Equation (USLE). For the CEAP sample, the RUSLE2 suffers from missingness due to the refusal of the farmer interviewed, while the corresponding USLE is available.

We are interested in estimating the population mean of RUSLE2 using USLE as an auxiliary variable for calibration weighting in Arkansas. Specifically, the total sampled points are 1509 and the fully observed are 406, with observed rate 26.9%. The normal quantile-quantile (Q-Q) plot generated from linear models based on the complete cases is presented in Sub-figure (a) of Figure 6. By the normal Q-Q plot, numerous outlier data points are evident, and the residuals deviate from satisfying the assumption of a normal distribution. Therefore, given this scenario, we have strong reasons to place greater trust in our suggested robust approach. The associated mean estimation result is presented in Sub-

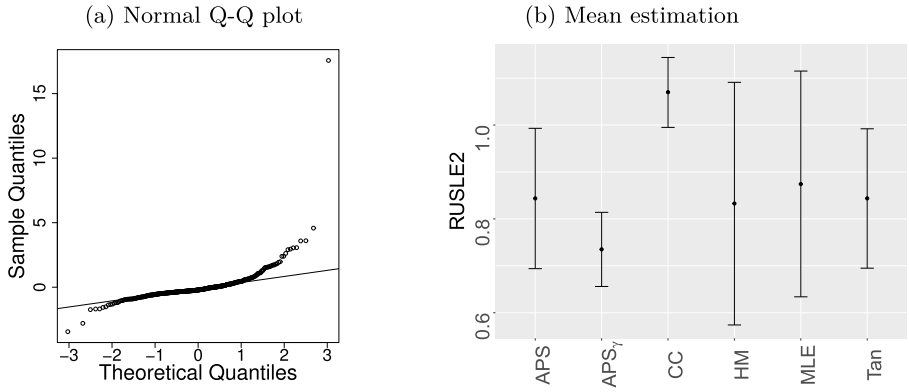

FIG 6. *Quantile-quantile plot and mean estimation with 95% confidence band for CEAP data in Arkansas.*

figure (b) of Figure 6, where the confidence 95% confidence intervals for the APS method is constructed by (4.5) and (4.6) and the APS$_\gamma$ is constructed by (5.13) and (5.14). In this presentation, our proposed method using gamma divergence exhibits a narrow 95% confidence interval and yields a low RUSLE2 value.

8. Concluding remarks

We have applied the information projection to obtain an augmented PS model that allows for doubly robust estimation. We have introduced self-efficiency to obtain the algebraic equivalence between the PSW estimator and the regression imputation estimator. In addition, γ-power divergence can be used to obtain an outlier-robust regression imputation estimator, which can be expressed as a PSW estimator under self-efficiency. In practice, an efficient and outlier-robust PSW estimator is very attractive as the existence of outliers can often damage the efficiency of the result estimator.

The tuning parameter is determined to balance the trade-off between statistical efficiency and robustness against outliers. In the future, further theoretical investigation on the effect of the choice of parameter γ on the final estimation will be considered.

There are several directions for further extensions of the proposed method. The proposed method is directly applicable to the calibration weighting in survey sampling (Fuller, 2009; Tillé, 2020). Other divergence measures such as Hellinger divergence (Antoine and Dovonon, 2021) can be also considered in the information projection.

The proposed method is based on the assumption of missing at random. Extension to nonignorable nonresponse can also be an interesting research direction. Furthermore, the proposed method can be used for causal inference, including the estimation of the average treatment effect from observational stud-

ies (Yang and Ding, 2020; Chen et al., 2021, 2023). Developing tools for causal inference using the proposed method will be an important extension of this research.

Appendix A: Proof of Lemma 2.1

If the balancing condition (2.3) holds, then the propensity score weighting estimator can be expressed as

$$\begin{aligned}\hat{\theta}_{\mathrm{PSW}} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega_i y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega_i\left(m_i+\varepsilon_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega_i m_i + \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega_i\varepsilon_i \\ &= \frac{1}{n}\sum_{i=1}^{n} m_i + \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega_i\varepsilon_i,\end{aligned}$$

where $m_i = m(\boldsymbol{x}_i;\boldsymbol{\beta})$. Then

$$\hat{\theta}_{\mathrm{PSW}} - \hat{\theta}_{\mathrm{com}} = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i\omega_i - 1\right)\varepsilon_i.$$

Under the MAR assumption, we can obtain

$$E\left(\hat{\theta}_{\mathrm{PSW}} - \hat{\theta}_{\mathrm{com}} \mid \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{\delta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i\omega_i - 1\right)E(\varepsilon_i \mid \boldsymbol{x}_i) = 0.$$

Thus, we can conclude that $\hat{\theta}_{\mathrm{PSW}}$ is unbiased for θ.

Appendix B: Proof of Lemma 2.2

Note that we can express the covariate-balancing condition in (2.3) as

$$\sum_{i=1}^{n}\left(1-\delta_i\omega_i\right)\boldsymbol{b}_i = \mathbf{0}. \tag{B.1}$$

Thus, as long as (B.1) is satisfied, we can write

$$\begin{aligned}\hat{\theta}_{\mathrm{PSW}} &= n^{-1}\sum_{i=1}^{n}\delta_i\omega_i y_i + n^{-1}\sum_{i=1}^{n}\left(1-\delta_i\omega_i\right)\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta} \\ &= n^{-1}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right\} + n^{-1}\sum_{i=1}^{n}\delta_i(\omega_i-1)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\end{aligned}$$

for any $\boldsymbol{\beta}$. Thus, as long as (2.6) hold, we can obtain (2.5).

Appendix C: Proof of Theorem 4.1

Since $\hat{\boldsymbol{\lambda}}$ satisfies (4.2), we can obtain the following equivalence

$$\begin{aligned}\hat{\theta}_{\mathrm{APSW}} &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\omega^*(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\lambda}})y_i - \boldsymbol{U}_{2,n}^{\mathrm{T}}(\hat{\boldsymbol{\lambda}}\mid\hat{\boldsymbol{\phi}})\boldsymbol{\beta}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta_i\omega^*(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\lambda}})(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\right\}\\ &:= \hat{\theta}_{\mathrm{APSW}}(\hat{\boldsymbol{\lambda}};\boldsymbol{\beta})\end{aligned}$$

for any $\boldsymbol{\beta}$. Now,

$$\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\theta}_{\mathrm{APSW}}(\boldsymbol{\lambda};\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\hat{\pi}_{1,i}^{-1}-1\right)\exp\left(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\lambda}\right)\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{b}_i.$$

Thus, for the choice of $\hat{\boldsymbol{\beta}}$ in (18), we can obtain

$$E\left\{\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\theta}_{\mathrm{APSW}}(\boldsymbol{\lambda};\boldsymbol{\beta}^*)\right\} = \mathbf{0}.$$

Thus, the uncertainty associated with $\hat{\boldsymbol{\lambda}}$ is asymptotically negligible at $\boldsymbol{\beta}=\boldsymbol{\beta}^*$. Thus, we obtain

$$\hat{\theta}_{\mathrm{APSW}}(\boldsymbol{\lambda}^*;\boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \delta_i\hat{\omega}^*(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}},\boldsymbol{\lambda}^*)(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\right\} + o_p(n^{-1/2}). \tag{C.1}$$

Now, to apply Taylor linearization with respect to $\boldsymbol{\phi}$, we define

$$\begin{aligned}\hat{\theta}_\ell(\hat{\boldsymbol{\phi}};\boldsymbol{\kappa}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \delta_i\hat{\omega}^*(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}},\boldsymbol{\lambda}^*)(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\right\} - \boldsymbol{U}_{1,n}^{\mathrm{T}}(\hat{\boldsymbol{\phi}})\boldsymbol{\kappa}\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i y_i + (1-\delta_i)\left(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\kappa}\right)\\ &\qquad +\delta_i\left(\hat{\pi}_{1,i}^{-1}-1\right)\left\{\exp(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\lambda}^*)(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*) - \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\kappa}\right\}\Big].\end{aligned}$$

Thus, we can choose $\boldsymbol{\kappa}^*$ to satisfy

$$E\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\theta}_\ell(\boldsymbol{\phi};\boldsymbol{\kappa}^*)\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right\} = \mathbf{0},$$

which gives the final linearization form as in (4.5).

Appendix D: Regularity conditions for Theorem 5.1

Before the statement of regularity conditions for Theorem 5.1, we first define a few quantities. Let

$$\boldsymbol{U}_{3,n}(\boldsymbol{\lambda},\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{\phi}^*) = \begin{pmatrix} \frac{1}{n}\left[\sum_{i=1}^n \delta_i \left\{1+\{d_i(\boldsymbol{\phi}^*)-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\right\}\boldsymbol{b}_i - \sum_{i=1}^n \boldsymbol{b}_i\right] \\ \frac{1}{n}\sum_{i=1}^n \delta_i\{d_i(\boldsymbol{\phi}^*)-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{b}_i \\ \frac{1}{n}\sum_{i=1}^n \delta_i\{d_i(\boldsymbol{\phi}^*)-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\left\{(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2-\frac{\sigma^2}{1+\gamma}\right\} \end{pmatrix}. \tag{D.1}$$

Further, define

$$\boldsymbol{s}_{11} = -\sum_{i=1}^n \delta_i \left\{\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\right\}\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{s}_{12} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{s}_{13} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\left\{(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2-\frac{\sigma^2}{1+\gamma}\right\}\boldsymbol{b}_i,$$

$$\boldsymbol{s}_{21} = -\sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\left\{\frac{\gamma}{\sigma^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\right\}\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{s}_{22} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\left\{\frac{\gamma}{\sigma^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2-1\right\}\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{s}_{23} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})\left\{\frac{\gamma}{\sigma^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2-2\right\}\boldsymbol{b}_i,$$

$$\boldsymbol{s}_{31} = -\sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\frac{\gamma}{2(\sigma^2)^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{s}_{32} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2)\frac{\gamma}{2(\sigma^2)^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^3\boldsymbol{b}_i^{\mathrm{T}},$$

$$s_{33} = \sum_{i=1}^n \delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta},\sigma^2) \times\left[\frac{\gamma}{2(\sigma^2)^2}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2\left\{(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2-\frac{\sigma^2}{1+\gamma}\right\}-\frac{1}{1+\gamma}\right].$$

Let

$$\mathbf{S}_n(\boldsymbol{\lambda},\boldsymbol{\beta},\sigma^2 \mid \hat{\boldsymbol{\phi}}) = \begin{pmatrix} \boldsymbol{s}_{11} & \boldsymbol{s}_{12} & \boldsymbol{s}_{13} \\ \boldsymbol{s}_{21} & \boldsymbol{s}_{22} & \boldsymbol{s}_{23} \\ \boldsymbol{s}_{31} & \boldsymbol{s}_{32} & \boldsymbol{s}_{33} \end{pmatrix}.$$

In addition with the regularity conditions in Theorem 1, we need the following additional assumptions.

Assumption D.1. *The estimating equation* $\boldsymbol{U}_{3,n}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*) = \mathbf{0}$ *has a unique solution* $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)$ *and there exists a function* $\boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*)$ *such that* $\boldsymbol{U}_{3,n}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*) \to \boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*)$ *uniformly as* $n \to \infty$ *and* $\boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*) = \mathbf{0}$ *has a unique solution* $(\boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2})$.

Assumption D.2. *The limiting function* $\boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*)$ *is differentiable and* $\partial_{(\boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}} \boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*)$ *is continuous on a compact set* $\mathcal{G}_3$ *containing* $(\boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2})$.

Assumption D.3. *The matrix* $\partial_{(\boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}} \boldsymbol{U}_3(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*) \mid_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*, \boldsymbol{\beta}=\boldsymbol{\beta}^*, \sigma^2=\sigma^{*2}}$ *is non-singular.*

Assumption D.4. *There exists a function* $\mathbf{S}(\boldsymbol{\phi})$ *such that* $\mathbf{S}_n(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*) \to \mathbf{S}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\phi}^*)$ *uniformly as* $n \to \infty$ *and* $\mathbf{S}(\boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2} \mid \boldsymbol{\phi}^*)$ *is non-singular.*

Assumptions D.1 – D.3 are commonly used in estimating equation theory and also ensure the existence of $\boldsymbol{\lambda}^*$, $\boldsymbol{\beta}^*$ and σ^{*2}. Assumption D.4 ensures the existence of the nuisance parameter $\mu^*, \boldsymbol{\zeta}^*$ and ν^* defined in Theorem 5.1, which helps us represent the influence function for the proposed robust estimator.

Appendix E: Proof of Theorem 5.1

By the normal equations, we have

$$
\begin{aligned}
&\theta_{\mathrm{APSW},\gamma}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2 \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}, \boldsymbol{\zeta}, \nu) \\
&= \frac{1}{n} \sum_{i=1}^{n} \delta_i \omega_{\gamma,i}(\boldsymbol{x}_i; \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2, \hat{\boldsymbol{\phi}}) y_i \\
&\quad - \frac{\boldsymbol{\mu}^{\mathrm{T}}}{n} \left[\sum_{i=1}^{n} \delta_i \left\{ 1 + \{ d_i(\hat{\boldsymbol{\phi}}) - 1 \} g_i(\hat{\boldsymbol{\lambda}}) q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2) \right\} \boldsymbol{b}_i - \sum_{i=1}^{n} \boldsymbol{b}_i \right] \\
&\quad + \frac{\boldsymbol{\zeta}^{\mathrm{T}}}{n} \sum_{i=1}^{n} \delta_i \{ d_i(\hat{\boldsymbol{\phi}}) - 1 \} g_i(\hat{\boldsymbol{\lambda}}) q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q) \boldsymbol{b}_i \\
&\quad + \frac{\nu}{n} \sum_{i=1}^{n} \delta_i \{ d_i(\hat{\boldsymbol{\phi}}) - 1 \} g_i(\hat{\boldsymbol{\lambda}}) q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 \left\{ (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 - \frac{\hat{\sigma}_q^2}{1+\gamma} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu} + \frac{1}{n} \sum_{i=1}^{n} \delta_i \omega_{\gamma,i}(\boldsymbol{x}_i; \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2, \hat{\boldsymbol{\phi}})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}) \\
&\quad + \frac{\boldsymbol{\zeta}^{\mathrm{T}}}{n} \sum_{i=1}^{n} \delta_i \{ d_i(\hat{\boldsymbol{\phi}}) - 1 \} g_i(\hat{\boldsymbol{\lambda}}) q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q) \boldsymbol{b}_i \\
&\quad + \frac{\nu}{n} \sum_{i=1}^{n} \delta_i \{ d_i(\hat{\boldsymbol{\phi}}) - 1 \} g_i(\hat{\boldsymbol{\lambda}}) q_{\gamma,i}(\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 \left\{ (y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q)^2 - \frac{\hat{\sigma}_q^2}{1+\gamma} \right\}
\end{aligned}
\tag{E.1}
$$

for any $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\zeta} \in \mathbb{R}^p$ and $\nu \in \mathbb{R}$. In addition, tedious derivation will lead

$$\begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\lambda}} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda} \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}, \boldsymbol{\zeta}, \nu) \\ \frac{\partial}{\partial \boldsymbol{\beta}} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda} \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}, \boldsymbol{\zeta}, \nu) \\ \frac{\partial}{\partial \sigma^2} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda} \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}, \boldsymbol{\zeta}, \nu) \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \boldsymbol{s}_{10} + \boldsymbol{s}_{11}\boldsymbol{\mu} + \boldsymbol{s}_{12}\boldsymbol{\zeta} + \boldsymbol{s}_{13}\nu \\ s_{20} + \boldsymbol{s}_{21}\boldsymbol{\mu} + \boldsymbol{s}_{22}\boldsymbol{\zeta} + s_{23}\nu \\ \boldsymbol{s}_{30} + \boldsymbol{s}_{31}\boldsymbol{\mu} + \boldsymbol{s}_{32}\boldsymbol{\zeta} + \boldsymbol{s}_{33}\nu \end{pmatrix} \quad \text{(E.2)}$$

where

$$\begin{aligned} \boldsymbol{s}_{10} &= \sum_{i=1}^{n} \delta_i \left\{ \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta}, \sigma^2) y_i \right\} \boldsymbol{b}_i, \\ \boldsymbol{s}_{20} &= \sum_{i=1}^{n} \delta_i \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta}, \sigma^2) \left\{ \frac{\gamma}{\sigma^2} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}) \right\} y_i \boldsymbol{b}_i, \\ \boldsymbol{s}_{30} &= \sum_{i=1}^{n} \delta_i \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}) q_{\gamma,i}(\boldsymbol{\beta}, \sigma^2) \frac{\gamma}{2(\sigma^2)^2} (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta})^2 y_i, \end{aligned}$$

and other quantities are defined in Appendix D.

Therefore, we can choose $\boldsymbol{\mu}^*$, $\boldsymbol{\zeta}^*$ and ν^* such that

$$E \left\{ \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\lambda}} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}^*, \boldsymbol{\zeta}^*, \nu^*) \\ \frac{\partial}{\partial \boldsymbol{\beta}} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}^*, \boldsymbol{\zeta}^*, \nu^*) \\ \frac{\partial}{\partial \sigma^2} \theta_{\mathrm{APSW},\gamma}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma^2 \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}^*, \boldsymbol{\zeta}^*, \nu^*) \end{pmatrix} \Bigg|_{\substack{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*, \\ \boldsymbol{\beta}=\boldsymbol{\beta}^*, \\ \sigma^2=\sigma^{*2}}} \right\} = \mathbf{0}, \quad \text{(E.3)}$$

where ensures that

$$\begin{aligned} &\theta_{\mathrm{APSW},\gamma}(\boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2} \mid \hat{\boldsymbol{\phi}}; \boldsymbol{\mu}^*, \boldsymbol{\zeta}^*, \nu^*) \\ &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}^* + \frac{1}{n} \sum_{i=1}^{n} \delta_i \omega_{\gamma,i}(\boldsymbol{x}_i; \boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2}, \hat{\boldsymbol{\phi}})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}^*) \\ &\quad + \frac{\boldsymbol{\zeta}^{*\mathrm{T}}}{n} \sum_{i=1}^{n} \delta_i \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}^*) q_{\gamma,i}(\boldsymbol{\beta}^*, \sigma^{*2})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{b}_i \\ &\quad + \frac{\nu^*}{n} \sum_{i=1}^{n} \delta_i \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}^*) q_{\gamma,i}(\boldsymbol{\beta}^*, \sigma^{*2})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 \left\{ (y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 - \frac{\sigma^{*2}}{1+\gamma} \right\} \\ &= \theta_{\mathrm{APSW},\gamma}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2, \hat{\boldsymbol{\phi}}) + o_p(n^{-1/2}). \end{aligned} \quad \text{(E.4)}$$

Finally, we have

$$\begin{aligned} &\theta_{\mathrm{APSW},\gamma}(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2}; \boldsymbol{\mu}^*, \boldsymbol{\zeta}^*, \nu^*, \boldsymbol{\kappa}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}^* + \frac{1}{n} \sum_{i=1}^{n} \delta_i \omega_{\gamma,i}(\boldsymbol{x}_i; \boldsymbol{\lambda}^*, \boldsymbol{\beta}^*, \sigma^{*2}, \hat{\boldsymbol{\phi}})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\mu}^*) \\ &\quad + \frac{\boldsymbol{\zeta}^{*\mathrm{T}}}{n} \sum_{i=1}^{n} \delta_i \{d_i(\hat{\boldsymbol{\phi}}) - 1\} g_i(\boldsymbol{\lambda}^*) q_{\gamma,i}(\boldsymbol{\beta}^*, \sigma^{*2})(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{b}_i \end{aligned}$$

$$+\frac{\nu^*}{n}\sum_{i=1}^{n}\delta_i\{d_i(\hat{\boldsymbol{\phi}})-1\}g_i(\boldsymbol{\lambda}^*)q_{\gamma,i}(\boldsymbol{\beta}^*,\sigma^{*2})(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2\left\{(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2-\frac{\sigma^{*2}}{1+\gamma}\right\}$$
$$-\boldsymbol{U}_{1,n}^{\mathrm{T}}(\hat{\boldsymbol{\phi}})\boldsymbol{\kappa} \tag{E.5}$$

holds for any $\boldsymbol{\zeta}\in\mathbb{R}^p$. Apparently,

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\phi}}\theta_{\mathrm{APSW},\gamma}(\boldsymbol{\phi}\mid\boldsymbol{\lambda}^*,\boldsymbol{\beta}^*,\sigma^{*2};\boldsymbol{\mu}^*,\boldsymbol{\zeta}^*,\nu^*,\boldsymbol{\kappa})\\
&=-\frac{1}{n}\sum_{i=1}^{n}\delta_i d_i^2(\boldsymbol{\phi})g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta}^*,\sigma^{*2})(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\mu}^*)\frac{\partial}{\partial\boldsymbol{\phi}}\pi_1^{(0)}(\boldsymbol{x}_i,\boldsymbol{\phi})\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\delta_i d_i^2(\boldsymbol{\phi})g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta}^*,\sigma^{*2})(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)(\boldsymbol{\zeta}^{*\mathrm{T}}\boldsymbol{b}_i)\frac{\partial}{\partial\boldsymbol{\phi}}\pi_1^{(0)}(\boldsymbol{x}_i,\boldsymbol{\phi})\\
&\quad-\frac{\nu^*}{n}\sum_{i=1}^{n}\delta_i d_i^2(\boldsymbol{\phi})g_i(\boldsymbol{\lambda})q_{\gamma,i}(\boldsymbol{\beta}^*,\sigma^{*2})(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2\\
&\qquad\qquad\times\left\{(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2-\frac{\sigma^{*2}}{1+\gamma}\right\}\frac{\partial}{\partial\boldsymbol{\phi}}\pi_1^{(0)}(\boldsymbol{x}_i,\boldsymbol{\phi})\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\left[\{\delta_i d_i(\boldsymbol{\phi})-1\}\left\{\frac{\partial\boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right\}^{\mathrm{T}}-\delta_i d_i^2(\boldsymbol{\phi})\frac{\partial\pi_1^{(0)}(\boldsymbol{x}_i,\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\boldsymbol{h}^{\mathrm{T}}(\boldsymbol{x}_i;\boldsymbol{\phi})\right]\boldsymbol{\kappa}.
\end{aligned}\tag{E.6}$$

As long as we choose $\boldsymbol{\kappa}^*$ such that

$$E\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\theta_{\mathrm{APSW},\gamma}(\boldsymbol{\phi}\mid\boldsymbol{\lambda}^*,\boldsymbol{\beta}^*,\sigma^{*2};\boldsymbol{\mu}^*,\boldsymbol{\zeta}^*,\nu^*,\boldsymbol{\kappa}^*)\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right\}=\mathbf{0}, \tag{E.7}$$

the linearization form is attained.

In a nutshell, we can derive estimates for $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\zeta}}$, $\hat{\nu}$, and $\hat{\boldsymbol{\kappa}}$ by solving the estimation equations, which result from equating the formula in (E.2) to $\mathbf{0}$ and that in (E.6) to $\mathbf{0}$.

Appendix F: Computational complexity

The main computation involved in our proposed method is solving the estimating equation. We apply the Newton-Raphson algorithm to solve the estimating equation. For each step, the computational complexity is $O(N^3)$, where N is the length of root. Recall that $\boldsymbol{\lambda}\in\mathbb{R}^{L+1}$ and $\boldsymbol{\phi}\in\mathbb{R}^p$. By the flowchart 1 in Fig 2 in our revised manuscript, in each iteration, the computational complexity for the augmented PSW estimator is $O(n\max\{p+L\}+p^3+L^3)$, where the first term is the complexity to compute the estimating equation and the last two terms are the complexity to solve the estimating equation. Therefore, suppose we set K as the maximum iterations for the Newton-Raphson method, the final computational complexity is $O(K(n\max\{p+L\}+p^3+L^3))$. In a similar fashion,

the computational complexity for the robust augmented PSW estimator is also $O(K(n\max\{p+L\}+p^3+L^3))$. As a result, the computational complexity is linear in n.

Acknowledgments

The authors would like to thank the Editor, Associate Editor, and two anonymous referees for their constructive comments. We also thank Dr. Emily Berg for sharing the CEAP data for analysis.

Funding

Hengfang Wang was partially supported by the Open Research Fund of Key Laboratory of Analytical Mathematics and Applications (Fujian Normal University), Ministry of Education, P. R. China (JAM2403) and Middle-aged and Young Teachers' Training Program (SDPY2023024).

Jae Kwang Kim's research was partially supported by a grant from US National Science Foundation (2242820), a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, and from the U.S. Department of Agriculture's National Resources Inventory, Cooperative Agreement NR203A750023C006, Great Rivers CESU 68-3A75-18-504.

References

Antoine, B. and Dovonon, P. (2021). Robust estimation with exponentially tilted Hellinger distance. *Journal of Econometrics* **224** 330–344. MR4303784

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973. MR2216189

Basak, S., Basu, A. and Jones, M. C. (2020). On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics* **48** 536-556. DOI: 10.1080/02664763.2020.1736524. MR4205987

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **85** 549–559. MR1665873

Berg, E. and Yu, C. (2019). Semiparametric quantile regression imputation for a complex survey with application to the Conservation Effects Assessment Project. *Survey methodology* **45** 249–270.

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** 723–734. MR2538768

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B* **78** 673-700. MR3506798

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

Chen, C., Shen, B., Liu, A., Wu, R. and Wang, M. (2021). A multiple robust propensity score method for longitudinal analysis with intermittent missing data. *Biometrics* **77** 519–532. MR4307652

Chen, C., Chen, S., Ye, Z., Shi, X. and Ma, T. (2023). The Effect of Substance Use on Brain Ageing: A New Causal Inference Framework for Incomplete and Massive Phenomic Data. *arXiv preprint arXiv:2303.03520*.

Csiszár, I. and Shield, P. C. (2004). Information Theory and Statistics: A Tutorial. *Foundations and Trends in Communications and Information Theory* **1** 417–528.

Dagdoug, M., Goga, C. and Haziza, D. (2023). Model-Assisted Estimation Through Random Forests in Finite Population Sampling. *Journal of the American Statistical Association* **118** 1234-1251. MR4595490

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

Dong, Y. and Peng, C. J. Y. (2013). Principled missing data methods for researchers. *SpringerPlus* **2** 222.

Eguchi, S. (2021). Pythagoras theorem in information geometry and applications to generalized linear models. *Handbook of Statistics* **45** 15–42. MR4382424

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B* **60** 3–21. MR1625672

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, NJ.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica* **66** 315–331. MR1612242

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* **20** 25–46.

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

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

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B* **76** 243–263. MR3153941

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24** 375–394. MR3183689

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96** 917–932. MR2767279

Kim, J. K. and Riddles, M. K. (2012). Some theory for propensity-score-

adjustment estimators in survey sampling. *Survey Methodology* **38** 157–165.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84** 881–896. MR1134486

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley and Sons.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4** 2111–2245. MR1315971

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525

Stefanski, L. A., Carroll, R. J. and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **73** 413–424. MR0855901

Sugasawa, S. and Yonekura, S. (2021). On Selection Criteria for the Tuning Parameter in Robust Divergence. *Entropy* **23** 1147. https://doi.org/10.3390/e23091147. MR4319509

Tan, Z. (2019). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika* **107** 137-158. MR4064145

R Core Team (2022). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations*. John Wiley & Sons.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer-Verlag, New York. MR2233926

Wu, Y. C. and Liu, Y. F. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association* **102** 974–983. MR2411659

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

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association* **115** 1540–1554. MR4143484

Yang, S., Kim, J. K. and Song, R. (2020). Doubly Robust Inference when Combining Probability and Non-probability Samples with High-dimensional Data. *Journal of the Royal Statistical Society: Series B* **82** 445–465. MR4084171

Zhang, C. M., Jiang, Y. and Chai, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika* **97** 551–566. MR2672483

Zhao, Q. (2019). Covariate balancing propensity score by Tailored loss functions. *The Annals of Statistics* **47** 965–993. MR3909957