Biased-sample empirical likelihood weighting for missing data problems: an alternative to inverse probability weighting

Yukun Liu

KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China

Yan Fan†

School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China

Summary. Inverse probability weighting (IPW) is widely used in many areas when data are subject to unrepresentativeness, missingness, or selection bias. An inevitable challenge with the use of IPW is that the IPW estimator can be remarkably unstable if some probabilities are very close to zero. To overcome this problem, at least three remedies have been developed in the literature: stabilizing, thresholding, and trimming. However the final estimators are still IPW type estimators, and inevitably inherit certain weaknesses of the naive IPW estimator: they may still be unstable or biased. We propose a biased-sample empirical likelihood weighting (ELW) method to serve the same general purpose as IPW, while completely overcoming the instability of IPW-type estimators by circumventing the use of inverse probabilities. The ELW weights are always well defined and easy to implement. We show theoretically that the ELW estimator is asymptotically normal and more efficient than the IPW estimator and its stabilized version for missing data problems. Our simulation results and a real data analysis indicate that the ELW estimator is shift-equivariant, nearly unbiased, and usually outperforms the IPW-type estimators in terms of mean square error.

Keywords: causal inference, empirical likelihood, inverse probability weighting, missing data

1. Introduction

Inverse probability weighting (IPW) has long been accepted as the standard estimation procedure under unequal probability samplings with and without replacement ever since the work of Hansen and Hurwitz (1943) and Horvitz and Thompson (1952). IPW always produces an unbiased or asymptotically unbiased estimator with an elegant expression, regardless of the complexity of the underlying sampling plan, and this method therefore enjoys great popularity. As well as survey sampling, it has been widely used in many other areas, including missing data problems (Robins et al., 1994; Wooldridge, 2007; Tan, 2010; Kim and Shao, 2021), treatment effect estimation or program evaluation (Rosenbaum and Rubin, 1983; Rosenbaum, 2002; Imbens and Wooldridge, 2009; Hirano et al., 2003; Cattaneo, 2010; Young et al., 2019; Zhao, 2019; Tan, 2020), personalized medicine (Zhang et al., 2012; Jiang et al., 2017), and survival data analysis (Robins and Rotnitzky, 1992; Robins, 1993; Bang and Tsiatis, 2000; Ma and Yin, 2011; Dong et al., 2020), where IPW is renamed inverse probability of censoring weighting. In recent years, accompanied by

†*Address for correspondence:* Yan Fan, School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China. Email: fanyan212@126.com

optimal subsampling, the IPW method has also proved to be an effective approach to validate statistical inferences for big data (Wang et al., 2018, 2019; Yu et al., 2022).

Through weighting the observations by the reciprocal of a certain probability of inclusion in the sample, the IPW estimator is able to account for unrepresentativeness, missingness or selection bias caused by non-random lack of information or non-random selection of observations. However, the IPW estimator can be highly unstable if there are extremely small probabilities, which can result in biased estimation or poor finite-sample performance of the accompanying asymptotic-normality-based inference (Busso et al., 2014; Kang and Schafer, 2007; Robins et al., 2007; Imbens and Wooldridge, 2009; Cao et al., 2009; Han et al., 2019). As pointed out by Robins et al. (2007) with regard to double-robust estimators (which are IPW-type estimators) in missing data problems, 'Whenever the "inverse probability" weights are highly variable, ..., a small subset of the sample will have extremely large weights relative to the remainder of the sample. In this setting, no estimator of the marginal mean $\mu = \mathbb{E}(Y)$ can be guaranteed to perform well.' In casual inference with observational studies, this is the well-known limited- or non-overlap problem in covariate distributions in different treatment groups (Crump et al., 2009; Khan and Tamer, 2010; Yang and Ding, 2018). The IPW estimator becomes inflated disproportionately or even breaks down in survival analysis when the number of patients at risk in the tails of the survival curves of censoring times is too small (Robins and Finkelstein, 2000; Dong et al., 2020). To guarantee that the IPW estimator possesses consistency, asymptotic normality, and satisfactory finite-sample performance, it is usual to impose an unnatural lower boundedness assumption on the probabilities (Rosenbaum and Rubin, 1983; Mccaffrey et al., 2013; Sun and Tchetgen Tchetgen, 2018), although tiny probabilities are frequently encountered in practice, especially when the propensity scores are estimated from data (Yang and Ding, 2018; Ma and Wang, 2020).

To overcome this notorious problem, at least three remedies have been proposed in the literature: stabilizing, thresholding, and trimming. The stabilizing method (Hájek, 1971) rescales the IPW estimator so that the weights sum to 1 (Kang and Schafer, 2007). Although straightforward, it can often sharply reduce the instability of the IPW estimator. The thresholding method, proposed by Zong et al. (2019) in the context of survey sampling, replaces those probabilities that are less than a given threshold by that threshold while keeping others unchanged. The parameter of interest is then estimated by IPW with the modified probabilities. Zong et al. (2019) proposed an easy-to-use threshold determining procedure and showed that, in general, the resulting IPW estimator works better than the naive IPW estimator. This method can reduce the negative effect of highly heterogeneous inclusion probabilities, and hence leads to improved estimation efficiency, although at the cost of an estimation bias. The trimming method excludes those observations with probabilities less than a given threshold or, equivalently, sets their weights to zero (Crump et al., 2009). Ma and Wang (2020) systematically investigated the large-sample behaviour of the IPW estimator after trimming and found it to be sensitive to the choice of trimming threshold and subject to a non-negligible bias. They proposed a bias-corrected and trimmed IPW estimator, which depends on an adaptively trimming threshold and a bandwidth. Inappropriate choices of the trimmed threshold and the bandwidth may affect the performance of their estimator. More importantly, the bias correction technique depends on the target quantity to be weighted, which makes their method inapplicable to weighted optimization problems, such as optimal treatment regime estimation (Zhang et al., 2012).

The final point estimators of the stabilizing, trimming, and thresholding methods are all based on IPW, although they adopt different strategies to reduce the detrimental effect of extremely small probabilities. These IPW-type estimators inevitably inherit certain weaknesses of the naive IPW estimator: they are either still unstable or biased. Also, the accompanying intervals, regardless of whether they are asymptotic-normality-based or resampling-based, often exhibit much undercoverage. See our simulation results in Section 3.

In this paper, we propose a biased-sample empirical likelihood weighting (ELW) estimation method to serve the same general purpose as IPW in handling incomplete or biased data while overcoming its instability. We systematically investigate its finite- and large-sample properties in the context of missing data problems, although it is generally applicable. The proposed ELW estimation method has several advantages over the IPW-type methods and the usual empirical likelihood (EL) (Owen, 1988, 1990, 2001).

(a) The ELW method circumvents the use of inverse probabilities and therefore never suffers from extremely small or even zero selection probabilities. It takes the maximum EL estimates of the probability masses of a multinomial distribution as weights, which always range from 0 to 1. This is the most significant advantage of the ELW method over IPW and its variants.
(b) The ELW weights are always well defined. By contrast, the usual EL weights suffer from the well-known convex hull constraint or the empty-set problem: they are undefined if the origin lies outside the convex hull of certain transformed data points (Tsao, 2004; Chen et al., 2008; Liu and Chen, 2010).
(c) Like the stabilized IPW estimator, the ELW weights always sum to 1, which gives the ELW estimator the nice property of shift-equivariance. Unfortunately, the naive IPW estimator, the trimmed IPW estimator of Zong et al. (2019), and the IPW estimator of Ma and Wang (2020) are all sensitive to a location shift in the response or the parameter of interest.
(d) The ELW weights are very convenient to calculate. Their calculation involves only solving a univariate rational equation, which can be done efficiently by the commonly used bisection algorithm. In contrast to the IPW estimator of Ma and Wang (2020), the ELW estimator is free of any tuning parameter and is hence more computationally efficient. The ELW weights depend only on the propensity scores and the full data size, and therefore the ELW method is directly applicable to survey sampling and weighted optimization problems.
(e) As we shall show, the ELW estimator is theoretically more efficient than the IPW estimator for missing data problems. This is a bonus of ELW, since the construction of the ELW weights makes use of side information. Our simulation results indicate that the ELW estimator often has smaller mean square errors and the accompanying interval has better coverage accuracy in most cases.

A crucial requirement of ELW is knowledge of the size of the finite population of interest or a larger independent and identically distributed sample that includes the observed data as a subsample. This is also required by the original IPW method and some of its variants, and is available in most situations. For example, in missing data problems, the size of the overall dataset is clearly known, and in survey sampling, the size of the finite population from which the sample was drawn is usually known a priori, since we need to construct a sampling frame before sampling. This mild requirement implies that the ELW method has many potential applications beyond missing data problems, sample surveys and casual inference.

The remainder of this article is organized as follows. In Section 2, we introduce the ELW method by estimating the parameter defined through just-identified estimating equations when data are subject to missingness. A simulation study and a real-life data analysis are conducted in Sections 3 and 4 to demonstrate the usefulness and advantage of the ELW method. Section 5 concludes with some discussion. All technical proofs, an extension of the ELW method to unequal probability samplings, large-sample properties of the ELW method under over-identified estimating equations, and additional simulation results can be found in the Supplementary Material. The R codes for reproducing all the computational results are available in online supplementary material.

2. Empirical likelihood weighting

We introduce the ELW method by solving missing data problems. Special cases of missing data problems include treatment effect estimation in observation studies under the potential outcome framework of Rubin (1974), as well as program evaluation in economics and other social sciences. Let $Z = (Y, X)$, with Y being a response variable that is subject to missingness and X an always-observed covariate. Denote by D a non-missingness indicator, with $D = 1$ if Y is observed and 0 otherwise. For ease of exposition, for the time being, we assume that the conditional non-missingness probability or the propensity score $\pi(Z) = P(D = 1|Z)$ is completely known and always positive, although our method allows $\pi(Z)$ to take zero values. The case with unknown propensity score is considered in Section 2.4. Suppose that the parameter of interest θ is an r-dimensional vector defined as the solution to $\mathbb{E}\{g(Z,\theta)\} = 0$, where $g(Z,\theta)$ is an s-dimensional compatible estimating function. We consider only the just-identified case (i.e. $s = r$); the over-identified case (i.e. $s > r$) is discussed in the supplementary material.

Denote the data by $\{(D_i, D_iZ_i), i = 1, 2, \ldots, N\}$, with $Z_i = (Y_i, X_i)$ or simply $\{z_i, i = 1, 2, \ldots, n\}$, where $z_i = (y_i, x_i)$ and $n = \sum_{j=1}^N D_j$; the covariates X_i with $D_i = 0$ do not come into play in most of this paper. The data $\{z_i, i = 1, 2, \ldots, n\}$ is in fact a biased sample of the underlying population if all $\pi(z_i)$ are not equal. The IPW estimator of θ is the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\pi(Z_i)}g(Z_i,\theta) = \frac{1}{N}\sum_{i=1}^{n}\frac{g(z_i,\theta)}{\pi(z_i)} = 0. \tag{1}$$

If $g(Z,\theta)$ can be expressed as $f(Z) - \theta$, then the IPW estimator is actually the Hájek estimator, or stabilized IPW (SIPW) estimator

$$\hat{\theta}_{\text{SIPW}} = \frac{\sum_{i=1}^{N} D_i f(Z_i)/\pi(Z_i)}{\sum_{j=1}^{N} D_j/\pi(Z_j)} = \frac{\sum_{i=1}^{n} f(z_i)/\pi(z_i)}{\sum_{j=1}^{n} 1/\pi(z_j)}. \tag{2}$$

Hereafter we also denote the solution to (1) by $\hat{\theta}_{\text{SIPW}}$ in general. The original version of IPW estimator is

$$\hat{\theta}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N} D_i\frac{f(Z_i)}{\pi(Z_i)} = \frac{1}{N}\sum_{i=1}^{n}\frac{f(z_i)}{\pi(z_i)}. \tag{3}$$

The expression (3) for the IPW estimator indicates that it becomes extremely unstable when some of the $\pi(Z_i)$ with $D_i = 1$ are close to zero, and that the terms with $D_i = 0$ actually contribute nothing to it. Since the size N is known, the zero-value D_i together with the other single-value D_i contain information about $\mathbb{E}(D) = \mathbb{E}\{\mathbb{E}(D|Z)\} = \mathbb{E}\{\pi(Z)\}$. The IPW estimator and its variants ignore such side information, and are not able to utilize it as well, and they consequently have potential losses of efficiency. As a popular and flexible non-parametric technique, EL (Owen, 1988, 1990, 2001) can conveniently and efficiently make use of side information to achieve improvements in efficiency. This motivates us to develop the ELW estimation method to serve the same purpose as the IPW estimator, while overcoming its instability and improving its estimation efficiency.

2.1. ELW estimator

Let the distribution function of Z be $F(z) = \text{pr}(Z \le z)$, where the inequality holds element-wise for vector-valued Z. To estimate θ, the solution to $\int g(z,\theta)dF(z) = 0$, it suffices to estimate $F(z)$. We consider the problem of estimating F by discarding those Z_i with $D_i = 0$, although these quantities may be partially accessible. The likelihood based on the remaining data is

$$\tilde{L} = (1-\alpha)^{N-n} \cdot \prod_{i=1}^{N} \{\pi(Z_i) dF(Z_i)\}^{D_i}, \tag{4}$$

where $\alpha = \mathrm{pr}(D=1) = \mathbb{E}\{\pi(Z)\}$ is the marginal non-missingness probability.

We use EL to handle the distribution $F(z)$. The basic idea of EL is to model $F(z)$ by a discrete distribution or a multinomial distribution assigning probability mass p_i to a datum Z_i, i.e., $F(z) = \sum_{i=1}^{N} p_i I(Z_i \le z)$, where the inequalities hold element-wise. Replacing $dF(Z_i)$ with p_i and taking logarithms of (4), we have the biased-sample empirical log-likelihood

$$\tilde{\ell} = \sum_{i=1}^{N} [D_i \log(p_i) + D_i \log\{\pi(Z_i)\} + (1-D_i)\log(1-\alpha)] \tag{5}$$

as the observed data $\{Z_i : D_i = 1\}$ is a biased sample of $\{Z_i : 1 \le i \le N\}$. Those p_i that are feasible satisfy

$$p_i \ge 0, \quad \sum_{i=1}^{N} p_i = 1, \quad \sum_{i=1}^{N} p_i\{\pi(Z_i) - \alpha\} = 0. \tag{6}$$

We emphasize that although those Z_i with $D_i = 0$ appear in $\tilde{\ell}$ and $\sum_{i=1}^{N} p_i\{\pi(Z_i)-\alpha\} = 0$, they have no likelihood contribution or any influence on the resulting EL method.

The proposed EL estimator of $F(z)$, or equivalently of the p_i, is obtained by maximizing the empirical log-likelihood (5) subject to (6). For fixed α, the maximum of the log-EL in (5) subject to (6) is attained at

$$p_i = \frac{1}{n} \frac{D_i}{1 + \lambda(\alpha)\{\pi(Z_i) - \alpha\}}, \tag{7}$$

where $\lambda(\alpha)$ satisfies

$$\frac{1}{n} \sum_{i=1}^{N} \frac{D_i}{1 + \lambda(\alpha)\{\pi(Z_i) - \alpha\}} \{\pi(Z_i) - \alpha\} = 0. \tag{8}$$

Putting (7) into (5) gives the profile log-EL of α (up to a constant that is independent of α)

$$\ell(\alpha) = \sum_{i=1}^{N} \{-D_i \log[1 + \lambda(\alpha)\{\pi(Z_i) - \alpha\}] + (1-D_i)\log(1-\alpha)\}.$$

This immediately gives $\hat{\alpha} = \arg\max \ell(\alpha)$, the EL estimator of α. Accordingly, the EL estimators of p_i and $F(z)$ are

$$\hat{p}_i = \frac{1}{n} \frac{D_i}{1 + \lambda(\hat{\alpha})\{\pi(Z_i) - \hat{\alpha}\}} \tag{9}$$

and $\hat{F}(z) = \sum_{i=1}^{N} \hat{p}_i I(Z_i \le z)$. Finally, the EL estimator or the ELW estimator $\hat{\theta}_{\mathrm{ELW}}$ of θ is the solution to

$$\int g(z,\theta) d\hat{F}(z) = \sum_{i=1}^{N} \hat{p}_i g(Z_i, \theta) = 0. \tag{10}$$

Obviously, both $\hat{F}(z)$ and $\hat{\theta}$ are well-defined statistics because $\hat{p}_i = D_i\hat{p}_i$.

When calculating the proposed EL estimator of $F(z)$, or equivalently of the p_i, we may maximize the empirical log-likelihood (5) with respect to p_i's, α and θ subject to both (6) and $\sum_{i=1}^N p_i g(Z_i,\theta) = 0$. Because the dimension of g is equal to that of θ, the resulting $\hat{p}_i$ and $\hat{\theta}_{\mathrm{ELW}}$ are exactly the same.

Compared with the usual EL, a remarkable feature of the likelihood in (4) is to include α, which has many advantages. First, including α, the likelihood in (4) can automatically incorporate the auxiliary information carried by N. Otherwise we have to construct new estimating equations to make use of this information. Second, after including α in the full likelihood, it is reasonable to construct the constraint $\sum_{i=1}^N p_i\{\pi(Z_i) - \alpha\} = 0$. The existence of this equation guarantees that the resulting estimator is consistent or can correct selection bias; otherwise the resulting estimator is inconsistent. In the next subsection, we show that including α, the resulting EL weights $\hat{p}_i$'s are always well defined, and that the ELW method can be quickly calculated by commonly-used softwares.

2.2. Practical implementation

The key to calculating the proposed EL estimators, including the EL estimator $\hat{F}$ of F and the ELW estimator $\hat{\theta}_{\mathrm{ELW}}$, is to calculate $\hat{\alpha}$ by maximizing $\ell(\alpha)$. This necessitates a double iterative algorithm because $\ell(\alpha)$ involves an implicit function $\lambda(\alpha)$, and thus it seems to be rather a difficult task. We find a more convenient solution, in which we need only solve a univariate equation.

Mathematically, $\hat{\alpha} = \arg\max \ell(\alpha)$ is a solution to

$$0 \;=\; \sum_{i=1}^N \left[\frac{D_i\lambda}{1+\lambda\{\pi(Z_i)-\alpha\}} - \frac{1-D_i}{1-\alpha}\right]. \tag{11}$$

Combining (8) and (11) gives

$$\lambda = \frac{N-n}{n(1-\alpha)}. \tag{12}$$

Putting this expression into (8) leads to an equivalent equation for α:

$$\sum_{i=1}^N \frac{D_i(\pi(Z_i)-\alpha)}{n/N + (1-n/N)\pi(Z_i) - \alpha} \;=\; 0. \tag{13}$$

As (13) has multiple roots, it is necessary to identify the interval containing the desired root. Denote the observed Z_i by $z_1, \ldots, z_n$ and define $\xi_i = n/N + (1-n/N)\pi(z_i)$ for $i = 1, 2, \ldots, n$. Equation (13) is further equivalent to $K(\alpha) = 0$, where $K(\alpha) = \sum_{i=1}^n \{\pi(z_i) - \alpha\}/(\xi_i - \alpha)$. Because $\alpha \in (0,1)$, $\xi_i \geq n/N$, and n/N is a consistent estimator of α, the desired root of $K(\alpha) = 0$ should lie between 0 and $\min \xi_i$. Actually, there must exist one and only one solution to $K(\alpha) = 0$ between 0 and $\min \xi_i$. Because $\xi_i \geq \pi(z_i)$, it follows that $K\{\min \pi(z_i)\} \geq 0$, $\lim_{\alpha \uparrow \min \xi_i} K(\alpha) = -\infty$, and that $K(\alpha)$ is strictly decreasing between 0 and $\min \xi_i$. By the intermediate value theorem, there must exist one and only one solution, denoted by $\hat{\alpha}$, in $[\min \pi(z_i), \min \xi_i)$ such that $K(\hat{\alpha}) = 0$. It is worth noting that if all the $\pi(z_i)$ are equal and equal to α_0, then $\hat{\alpha} = \alpha_0$ and the resulting $\hat{p}_i$ are all equal to $1/n$, and the ELW estimator reduces to the solution to $(1/n)\sum_{i=1}^n g(z_i, \theta) = 0$. Otherwise, all $\pi(z_i)$ $(i = 1, 2, \ldots, n)$ are not equal to each other, and $\hat{\alpha}$, $\hat{p}_i$, and $\hat{\theta}_{\mathrm{ELW}}$ are all non-degenerate.

The proposed ELW estimation procedure can be implemented by Algorithm 1. The first and third steps involve only closed-form calculations, the second step can be efficiently achieved by a bi-section search algorithm, and the last step has the same calculation burden as those of the IPW-type methods. These imply that the ELW procedure is easy to implement.

```
Algorithm 1: ELW estimation procedure
  Input: The missing dataset $\{(D_i, D_iZ_i, \pi(Z_i)) : i = 1, 2, \ldots, N\}$.
  Output: The ELW estimate, $\hat{\theta}_{\mathrm{ELW}}$, of $\theta$.
    Step 1. Calculate $n = \sum_{i=1}^{N} D_i$, $\zeta_l = \min\{\pi(Z_i) : D_i = 1, i = 1, 2, \ldots, N\}$ and
    $\zeta_u = n/N + (1 - n/N)\zeta_l$.
    Step 2. Calculate $\hat{\alpha}$ by solving (13) in the interval $[\zeta_l, \zeta_u)$, and calculate
    $\lambda(\hat{\alpha}) = (N - n)/\{n(1 - \hat{\alpha})\}$.
    Step 3. Calculate $\hat{p}_i = D_i n^{-1}[1 + \lambda(\hat{\alpha})\{\pi(Z_i) - \hat{\alpha}\}]^{-1}$ for $i = 1, 2, \ldots, N$.
    Step 4. Obtain $\hat{\theta}_{\mathrm{ELW}}$ by solving the equation $\sum_{i=1}^{N} \hat{p}_i D_i g(Z_i, \theta) = 0$ or minimizing
    $\| \sum_{i=1}^{N} \hat{p}_i D_i g(Z_i, \theta)\|^2$ with respect to $\theta$.
```

2.3. Finite- and large-sample properties

The non-zero EL weights are $(1 - \hat{\alpha})/\{N(\xi_i - \hat{\alpha})\}$ for $1 \le i \le n$. We use the maximum weight ratio $\kappa = (\max_{1\le i\le n} \xi_i - \hat{\alpha})/(\min_{1\le i\le n} \xi_i - \hat{\alpha})$ among the non-zero EL weights to quantify the dispersion between the EL weights. The following lemma establishes an upper bound on κ.

LEMMA 2.1. *Suppose $\pi(z_i)$ $(1 \le i \le n)$ take $m \ge 2$ distinct values $\pi_{(1)} < \ldots < \pi_{(m)}$ $(m \ge 2)$. If there exists $\varepsilon \in (0, 1)$ such that $\pi_{(m)} - \pi_{(1)} > \varepsilon$ and $n/N < 1 - \varepsilon$, then $\kappa \le N/\varepsilon^3$.*

Lemma 2.1 indicates that the ELW method works even if the smallest $\pi(z_i)$ is as small as zero. However, the maximum weight ratio of the IPW estimator has no such a guarantee, and the IPW estimator becomes extremely unstable when some of the $\pi(z_i)$ are close to zero. In particular, it fails to work when $\min_{1\le i\le n} \pi(z_i)$ is exactly zero. Our ELW estimator successfully and completely overcomes this issue, which is its most significant advantage over the traditional IPW estimator in finite-sample performance.

Next, we show that asymptotically our ELW estimator is unbiased and more efficient than the IPW estimator. This is a bonus of using ELW, and also a significant advantage that it has over the conventional IPW in large-sample performance. We make the following assumptions on the function $g(Z, \theta)$.

CONDITION 1. *(i) θ_0 is the unique solution to $\mathbb{E}\{g(Z, \theta)\} = 0$. (ii) The parameter space is a compact set $\Theta \subset \mathbb{R}^r$, $g(Z, \theta)$ is a continuous function of θ for every Z, and there exists a function $\bar{g}(Z)$ such that $\mathbb{E}\{\bar{g}(Z)\} < \infty$ and $\sup_{\theta\in\Theta} \|g(Z, \theta)\| \le \bar{g}(Z)$. (iii) $g(Z, \theta)$ has a continuous partial derivative $g_1(Z, \theta) = \partial g(Z, \theta)/\partial\theta^\top$ in a neighborhood of θ_0 for each Z. There exists a positive function $\bar{g}_1(Z)$ such that $\mathbb{E}\{\bar{g}_1(Z)\} < \infty$ and $\|g_1(Z, \theta)\|_F \le \bar{g}_1(Z)$ for all Z and for β in the neighborhood, where $\| \cdot \|_F$ is the Frobenius norm. (iv)The $r \times r$ matrix $K = \mathbb{E}\{g_1(Z, \theta_0)\}$ is nonsingular.*

We denote $A^{\otimes 2} = AA^\top$ for a vector or matrix A, and define $B_{gg} = \mathbb{E}\{g^{\otimes 2}(Z, \theta_0)/\pi(Z)\}$, $B_{11} = \mathbb{E}\{1/\pi(Z)\}$, and $B_{g1} = \mathbb{E}\{g(Z, \theta_0)/\pi(Z)\}$. When $g(Z, \theta) = f(Z) - \theta$, we define $B_{ff} = \mathbb{E}\{f^{\otimes 2}(Z)/\pi(Z)\}$ and $B_{f1} = \mathbb{E}\{f(Z)/\pi(Z)\}$.

THEOREM 2.1. *Let $\alpha_0 \in (0, 1)$ be the truth of α. Suppose that Condition 1 is satisfied, $\mathbb{V}\mathrm{ar}\{\pi(Z)|D = 1\} > 0$ and that B_{11} and B_{gg} are both finite. Also suppose that the conditional inclusion probabilities $\pi(Z_i)$ are known. As N goes to infinity,* *(a)* $\sqrt{N}(\hat{\theta}_{\rm ELW}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\rm ELW})$*, where* $\Sigma_{\rm ELW}=K^{-1}\{B_{gg}-B_{g1}^{\otimes 2}/(B_{11}-1)\}(K^{-1})^{\top}$;

(b) $\sqrt{N}(\hat{\theta}_{\rm SIPW}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\rm SIPW})$*, where* $\Sigma_{\rm SIPW}=K^{-1}B_{gg}(K^{-1})^{\top}$*;*

(c) The ELW estimator $\hat{\theta}_{\rm ELW}$ *is more efficient than the SIPW estimator* $\hat{\theta}_{\rm SIPW}$*, i.e.* $\Sigma_{\rm ELW}\le\Sigma_{\rm SIPW}$*, where the equality holds only if* $\pi(Z)$ *is degenerate.*

(d) If $g(Z,\theta)=f(Z)-\theta$*, then* $\sqrt{N}(\hat{\theta}_{\rm IPW}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\rm IPW})$ *with* $\Sigma_{\rm IPW}=B_{ff}-\theta_0^{\otimes 2}$ *and* $\Sigma_{\rm ELW}=(B_{ff}-\theta_0^{\otimes 2})-(B_{f1}-\theta_0)^{\otimes 2}/(B_{11}-1)$*; the ELW estimator* $\hat{\theta}_{\rm ELW}$ *is also more efficient than the IPW estimator* $\hat{\theta}_{\rm IPW}$.

In Theorem 2.1, we treat the marginal non-missingness probability α as a fixed and unknown parameter, which needs to be estimated. The assumption $\mathbb{V}{\rm ar}\{\pi(Z)|D=1\}>0$ guarantees that with probability tending to 1, the observed propensity scores are not all equal to each other, and so the ELW estimator is non-degenerate. Theorem 2.1 indicates that the ELW estimator is more efficient than the SIPW estimator. A likelihood explanation for this result is as follows. Let $z_1,\ldots,z_n$ be the Z_i's with $D_i=1$. As the foundation of our ELW method, the full likelihood $\tilde{L}$ in equation (4) is proportional to $L_m\times L_c$, where $L_m=\binom{N}{n}\alpha^n(1-\alpha)^{N-n}$ is a marginal likelihood, and $L_c=\prod_{i=1}^n\{\pi(z_i)dF(z_i)/\alpha\}$ is a conditional likelihood. When $\pi(z_i)$'s are known, the nonparametric maximum conditional likelihood estimator of F is $\tilde{F}(z)=\{\sum_{i=1}^n I(z_i\le z)/\pi(z_i)\}/\{\sum_{j=1}^n 1/\pi(z_j)\}$, therefore $\hat{\theta}_{\rm SIPW}$ is the maximum conditional likelihood estimator of θ. The ELW estimator $\hat{\theta}_{\rm ELW}$ is the maximum full likelihood estimator of θ. With the additional L_m, our ELW method automatically makes use of the auxiliary information carried by N, and hence is more efficient than the SIPW estimator.

A reasonable estimator of $\Sigma_{\rm ELW}$ is required in the construction of Wald-type confidence intervals for θ. Inspired by the fact that $\hat{p}_i\approx D_i/\{N\pi(Z_i)\}$, we propose to estimate $\Sigma_{\rm ELW}$ with the ELW method by

$$\widehat{\Sigma}_{\rm ELW}=\hat{K}^{-1}\{\hat{B}_{gg}-\hat{B}_{g1}^{\otimes 2}/(\hat{B}_{11}-1)\}(\hat{K}^{-1})^{\top}, \tag{14}$$

where $\hat{K}=\sum_{i=1}^N\hat{p}_i g_1(Z_i,\hat{\theta}_{\rm ELW})$, $\hat{B}_{11}=N\sum_{i=1}^N(\hat{p}_i)^2$, $\hat{B}_{g1}=N\sum_{i=1}^N g(Z_i,\hat{\theta}_{\rm ELW})(\hat{p}_i)^2$, and $\hat{B}_{gg}=N\sum_{i=1}^N\{g(Z_i,\hat{\theta}_{\rm ELW})\}^{\otimes 2}(\hat{p}_i)^2$. It is worth stressing that the ELW-based variance estimator is again insensitive to small probabilities, since it circumvents the use of inverse probabilities.

2.4. Estimated propensity score

In many situations, such as missing data problems and causal inference, the propensity score is unknown and needs be estimated from the observed data. The ELW and IPW estimators have different large-sample behaviours if we take the variability of the estimated propensity score into account. Suppose that $\pi(\cdot)$ is parametrically modelled by $\pi(Z,\beta)$.

CONDITION 2. *(i) There exists* β_0 *such that* $\pi(Z,\beta_0)=\pi(Z)$ *for all* Z*, and the function* $\pi(Z,\beta)$ *is continuously differentiable in* β *in a neighborhood of* β_0*. Let* $\pi_1(Z,\beta)=\partial\pi(Z,\beta)/\partial\beta^{\top}$*. (ii) There exist a positive constant* ε *and positive functions* $\bar{\pi}(Z)$ *and* $\bar{\pi}_1(Z)$ *such that* $\bar{\pi}(Z)\le\inf_{\beta:\|\beta-\beta_0\|\le\varepsilon}\pi(Z,\beta)$*,* $\sup_{\beta:\|\beta-\beta_0\|\le\varepsilon}\|\pi_1(Z,\beta)\|\le\bar{\pi}_1(Z)$*,* $\mathbb{E}\{\pi(Z)/(\bar{\pi}(Z))^2\}<\infty$*,* $\mathbb{E}\left\{\pi(Z)\bar{g}(Z)/\bar{\pi}(Z)\right\}<\infty$*, and* $\mathbb{E}\left\{\pi(Z)\bar{g}(Z)\bar{\pi}_1(Z)/\{\bar{\pi}(Z)\}^2\right\}<\infty$*, where* $\bar{g}$ *is given in Condition 1.*

Condition 2(i) holds when the non-missingness indicator D follows Logistic and Probit models, which are commonly used in the literature. Condition 2(ii) together with the other conditions guarantees the consistency of $\hat{\alpha}$ and hence the consistency of the ELW estimator. Under Condition 2(i), $\pi(Z,\beta_0)=\pi(Z)$ and therefore $B_{11}=\mathbb{E}\{1/\pi(Z,\beta_0)\}$, $B_{g1}=\mathbb{E}\{g(Z)/\pi(Z,\beta_0)\}$, and $B_{gg}=\mathbb{E}[\{g(Z)\}^{\otimes 2}/\pi(Z,\beta_0)]$. We define $B_{1\bar{\pi}}=\mathbb{E}\{\pi_1(Z,\beta_0)/\pi(Z)\}$

and $B_{g\dot{\pi}} = \mathbb{E}\{g(Z,\theta_0)\pi_1(Z,\beta_0)/\pi(Z)\}$. In the case of $g(Z,\theta) = f(Z) - \theta$, define $B_{f\dot{\pi}} = \mathbb{E}\{f(Z)\pi_1(Z,\beta_0)/\pi(Z)\}$.

THEOREM 2.2. *Assume Conditions 1 and 2 and that $\hat{\beta}$ satisfies $\hat{\beta}-\beta_0 = N^{-1}\sum_{i=1}^N h(D_i,Z_i)+o_p(N^{-1/2})$, where the influence function $h(D,Z)$ has zero mean. Suppose that the truth α_0 of α satisfies $0<\alpha_0<1$ and $\mathbb{V}\text{ar}\{\pi(Z,\beta_0)|D=1\}>0$. As N goes to infinity,*

(a) $\sqrt{N}(\hat{\theta}_{\text{ELW}}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\text{ELW},e})$, *where* $\Sigma_{\text{ELW},e} = K^{-1}\Omega(K^{-1})^\top$ *with*

$$\Omega = \mathbb{V}\text{ar}\left\{\frac{Dg(Z,\theta_0)}{\pi(Z,\beta_0)} + \frac{B_{g1}}{B_{11}-1}\left(1-\frac{D}{\pi(Z,\beta_0)}\right) + \left(\frac{B_{g1}B_{1\dot{\pi}}}{B_{11}-1} - B_{g\dot{\pi}}\right)h(D,Z)\right\};$$

(b) $\sqrt{N}(\hat{\theta}_{\text{SIPW}}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\text{SIPW},e})$, *where*

$$\Sigma_{\text{SIPW},e} = K^{-1}\mathbb{V}\text{ar}\left\{\frac{Dg(Z,\theta_0)}{\pi(Z,\beta_0)} - B_{g\dot{\pi}}h(D,Z)\right\}(K^{-1})^\top;$$

(c) In the case of $g(Z,\theta) = f(Z)-\theta$, $\sqrt{N}(\hat{\theta}_{\text{IPW}}-\theta_0)\xrightarrow{d} N(0,\Sigma_{\text{IPW},e})$, *where*

$$\Sigma_{\text{IPW},e} = \mathbb{V}\text{ar}\left\{\frac{Df(Z)}{\pi(Z,\beta_0)} - B_{f\dot{\pi}}h(D,Z)\right\}.$$

When the propensity score is known, Theorem 2.1 establishes the asymptotic normalities of the ELW, IPW and SIPW estimators. We find that $\Sigma_{\text{ELW}} \le \Sigma_{\text{SIPW}}$ and $\Sigma_{\text{ELW}} \le \Sigma_{\text{IPW}}$, indicating that the ELW estimator is asymptotically more efficient than the IPW and SIPW estimators. According to Theorem 2.2, the ELW, IPW and SIPW estimators still follow asymptotic normal distributions when the propensity score involves a finite-dimensional unknown parameter. However, in general, the inequality $\Sigma_{\text{ELW},e} \le \Sigma_{\text{SIPW},e}$ or $\Sigma_{\text{ELW},e} \le \Sigma_{\text{IPW},e}$ does not hold any longer. This implies that the efficiency gain of the ELW estimator over the IPW and SIPW estimators is no longer guaranteed.

If the response Y is missing at random (Rubin, 1976) and the covariate X is observed, then $\pi(Z,\beta)$ depends on $Z=(Y,X)$ through only X. We may estimate β by its maximum likelihood estimator $\hat{\beta}$, i.e. the maximizer of $\sum_{i=1}^N[D_i\log\pi(Z_i,\beta)+(1-D_i)\pi\{1-\pi(Z_i,\beta)\}]$. In this case,

$$h(D,Z) \quad = \quad \frac{D-\pi(Z,\beta_0)}{\pi(Z,\beta_0)\{1-\pi(Z,\beta_0)\}}(\tilde{B}_{\dot{\pi}\dot{\pi}})^{-1}\pi_1(Z,\beta_0),$$

where $\tilde{B}_{\dot{\pi}\dot{\pi}} = \mathbb{E}[\{\pi_1(Z,\beta_0)\}^{\otimes 2}/\{\pi(Z,\beta_0)(1-\pi(Z,\beta_0))\}]$. The asymptotic variance of the ELW estimator is

$$\Sigma_{\text{ELW},e} = K^{-1}\left\{B_{gg} - \frac{B_{g1}^{\otimes 2}}{B_{11}-1} - \left(\frac{B_{g1}B_{1\dot{\pi}}}{B_{11}-1} - B_{g\dot{\pi}}\right)(\tilde{B}_{\dot{\pi}\dot{\pi}})^{-1}\left(\frac{B_{g1}B_{1\dot{\pi}}}{B_{11}-1} - B_{g\dot{\pi}}\right)^\top\right\}(K^{-1})^\top.$$

Again, an ELW estimator can be constructed for $\Sigma_{\text{ELW},e}$. If $\pi(Z,\beta)$ is mis-specified, the desirable properties of the ELW, IPW and SIPW estimators in Theorem 2.2 disappear. A nonparametric or semiparametric model may be used for $\pi(X)$ to alleviate the risk of model mis-specification. If Y is missing not at random, namely $\pi(Z,\beta)$ depends on Y, the estimation of β becomes much more challenging as β may not be identifiable. Under additional assumptions such as the existence of an instrument (Wang et al., 2014), various estimation methods for β and θ have been developed based on data that are missing not at random. For a more comprehensive discussion on this issue, see Kim and Shao (2021).

2.5. Resampling-based interval estimation

Based on Theorems 2.1 and 2.2, we can construct Wald-type confidence intervals for θ once a consistent estimator for the asymptotic variance is available. The asymptotic normality of the ELW, IPW, and SIPW estimators requires that both $B_{11} = \mathbb{E}[\{\pi(Z)\}^{-1}]$ and B_{gg} are finite and well defined. If this is violated, the Wald-type confidence intervals may not have the promised coverage probability. This dilemma can be overcome by resampling. We propose to construct confidence intervals for θ by the resampling method in Algorithm 2.

```
Algorithm 2: Wald confidence region based on resampling and ELW
  Input: The missing dataset $\{(D_i, D_i Z_i, \pi(Z_i)) : i = 1, 2, \ldots, N\}$. Calculate the ELW
      estimator $\hat{\theta}_{\mathrm{ELW}}$ and the proposed variance estimator $\widehat{\Sigma}_{\mathrm{ELW}}$, and define
      $T_N = \sqrt{N}(\widehat{\Sigma}_{\mathrm{ELW}})^{-1/2}(\hat{\theta}_{\mathrm{ELW}} - \theta_0)$.
  Output: Wald confidence region for $\theta$ based on resampling and ELW
      Step 1. Draw $M \ll N$ (e.g. $M = \sqrt{N}$) observations, say $(D_i^*, D_i^* Z_i^*, \pi(Z_i^*))$
      $(1 \le i \le M)$, from the original sample by simple random sampling without
      replacement.
      Step 2. Calculate the counterparts of $\hat{\theta}_{\mathrm{ELW}}$ and $\widehat{\Sigma}_{\mathrm{ELW}}$ based on the subsample,
      denoted by $\hat{\theta}^*_{\mathrm{ELW}}$ and $\widehat{\Sigma}^*_{\mathrm{ELW}}$ . Construct $T_M^* = \sqrt{M}(\widehat{\Sigma}^*_{\mathrm{ELW}})^{-1/2}(\hat{\theta}^*_{\mathrm{ELW}} - \hat{\theta}_{\mathrm{ELW}})$.
      Step 3. Repeat Steps 1 and 2 $B = 1000$ times and denote the resulting test statistics
      by $\{T^*_{M,i} : i = 1, 2, \ldots, B\}$. Let $t_i^* = \|T^*_{M,i} - \bar{T}^*\|$, where $\bar{T}^* = (1/B)\sum_{i=1}^{B} T^*_{M,i}$. Denote
      the $(1-a)$ empirical quantile of the $t_i^*$ by $q^*_{1-a}$. Then a $(1-a)$-level confidence region
      for $\theta$ can be constructed as $\{\theta : \|\sqrt{N}(\widehat{\Sigma}_{\mathrm{ELW}})^{-1/2}(\hat{\theta}_{\mathrm{ELW}} - \theta) - \bar{T}^*\| \le q^*_{1-a}\}$.
```

In the case of the estimated propensity score $\widehat{\pi}(Z_i)$, we replace $\pi(Z_i)$ and $\widehat{\Sigma}_{\mathrm{ELW}}$ by $\widehat{\pi}(Z_i)$ and $\widehat{\Sigma}_{\mathrm{ELW},e}$, respectively. The ELW variance estimator $\widehat{\Sigma}_{\mathrm{ELW}}$ converges in probability to Σ_{ELW}, which is assumed to be positive definite. This, together with Theorems 2.1 and 2.2, implies that T_N converges in distribution to the standard normal, an obviously continuous distribution. By Corollary 2.1 of Politis and Romano (1994), the empirical distribution of T_M^* is a uniformly consistent estimator of the distribution of T_N, which is formally summarized in Theorem 2.3. This validates the interval estimator produced by Algorithm 2.

THEOREM 2.3. *Assume the conditions in Theorem 2.1 (for a known propensity score) or those in Theorem 2.2 (for an estimated propensity score) are satisfied. As* $N \to \infty$, *if* $M \to \infty$ *and* $M/N \to 0$, *then* $\sup_{t \ge 0} |P(T_N \le t) - P^*(T_M^* \le t)| = o_p(1)$, *where* P^* *is the conditional probability given the original sample.*

3. Simulation study

We consider the parameter θ corresponding to $g(Z, \theta) = Y - \theta$ with $Z = (Y, X)$, and conduct simulations to investigate the finite-sample performance of the proposed ELW estimator and the accompanying asymptotic-normality-based interval estimator. For comparison, we also take into account the IPW estimator, the SIPW estimator, and some popular variants of the IPW estimator:

(a) The modified IPW estimator of Zong et al. (2019) (ZZZ for short): $\hat{\theta}_{\mathrm{ZZZ}} = N^{-1}\sum_{i=1}^{N} D_i Y_i/\tilde{\pi}_i$, where $\tilde{\pi}_i = \max\{\pi_{(K)}, \pi(X_i)\}$, K is the maximum i such that $\pi_{(i)} \le 1/(i+1)$, and $\{\pi_{(1)}, \ldots, \pi_{(N)}\}$ are the propensity scores in increasing order.

(b) The trimmed IPW estimator of Crump et al. (2009) (CHIM for short):

$$\hat{\theta}_{\mathrm{CHIM}} = \sum_{i=1}^{N} \frac{D_i Y_i}{\pi(X_i)} \cdot I\{\alpha \leq \pi(X_i) \leq 1-\alpha\} \Bigg/ \sum_{i=1}^{N} I\{\alpha \leq \pi(X_i) \leq 1-\alpha\},$$

where α is obtained by minimizing a variance term and $I(\cdot)$ is the indicator function.

(c) The IPW estimator of Ma and Wang (2020) with $s = 1$ and $s = 2$, denoted by MW1 and MW2, respectively. Following Ma and Wang (2020), we set the tuning parameters b_N and h_N in MW1 and MW2 to the respective solutions of $b_N^s N^{-1} \sum_{i=1}^N I\{\pi(X_i) \leq b_N\} = 1/(2N)$ and $h_N^5 \sum_{i=1}^N I\{\pi(X_i) \leq h_N\} = 1$. For details, see the discussion below Theorem 3 of Ma and Wang (2020) and Section III of their supplementary material.

We simulate data from Example 1. All numbers reported in this simulation study are calculated based on $M = 5000$ simulated random samples.

EXAMPLE 1. *Instead of generating X, we generate the propensity score $\pi(X)$ from $P(\pi(X) \leq u) = u^{\gamma-1}$ $(0 \leq u \leq 1)$ with $\gamma = 1.5$ or 2.5. Given $\pi(X)$, we generate Y from $Y = \mu\{\pi(X)\} + c \cdot (\eta - 4)/\sqrt{8}$, where $c = 1$ or 0.1, and $\eta \sim \chi_4^2$, and the missingness status D of Y follows the Bernoulli distribution with success probability $\pi(X)$. Four choices of $\mu(t)$ are considered: $\mu(t) = \cos(2\pi t)$ (Model 1), $\mu(t) = 1 - t$ (Model 2), $\mu(t) = \cos(2\pi t) + 5$ (Model 3), and $\mu(t) = 6 - t$ (Model 4). The full data size is $N = 2000$ and the parameter of interest is $\theta = \mathbb{E}(Y)$.*

This example is a modified version of Example 1 in Section III of the supplementary material of Ma and Wang (2020), who considered the cases with $\gamma = 1.5$, $c = 1$, and $N = 2000$ for Models 1 and 2. The parameter γ $(\gamma > 1)$ controls the tail behaviour of $1/\pi(X)$. When $\gamma > 2$, the tail is light and $\mathbb{E}\{1/\pi(X)\} = (\gamma - 1)/(\gamma - 2)$ is finite. In this case, if g is bounded, then the conditions in Theorem 2.1 are generally fulfilled, and the asymptotic normalities of the ELW, IPW, and SIPW estimators are guaranteed. However, in the case of $1 < \gamma \leq 2$, the tail is heavy and $\mathbb{E}\{1/\pi(X)\} = \infty$, which violates the conditions of Theorem 2.1: the ELW, IPW, and SIPW estimators no longer follow asymptotically normal distributions. The constant c controls the influence of the random error on the response variable; a smaller c leads to a smaller noise. Models 3 and 4 are simply Models 1 and 2 with a mean shift.

Point estimation As a measure of the finite-sample performance of a generic estimator $\tilde{\theta}$, we define its scaled root mean square error (RMSE) as $\mathrm{RMSE}(\tilde{\theta}) = \sqrt{N} \times \{(1/M) \sum_{j=1}^M (\tilde{\theta}_j - \theta)^2\}^{1/2}$, where $\tilde{\theta}_j$ is the estimate $\tilde{\theta}$ based on the jth simulated random sample. Table 1 presents a comparison of the RMSEs of the seven estimators. Figure 1 displays the boxplots of the estimators under comparison (minus the true parameter value) when data were generated from Example 1 with $\gamma = 1.5$ and $c = 1$ and 0.1. For clearer presentation, we ignore the boxplots of the IPW estimator, because it fluctuates too dramatically.

In terms of RMSE, ELW outperforms IPW, SIPW, ZZZ, and CHIM in almost all scenarios. The only exception is the scenario with $\gamma = 2.5, c = 1$ for Model 1, where the RMSE (1.17) of ELW is slightly greater than the minimum RMSE (1.14) of IPW, SIPW, ZZZ, and CHIM. The boxplots also indicate that ELW is always nearly unbiased in all scenarios. ELW also outperforms MW1 and MW2 in most cases. The only exceptions are the scenarios with $\gamma = 2.5$ for Model 1 and those with $\gamma = 1.5, c = 1$ for Models 1 and 2. In the least favourable scenario ($\gamma = 1.5, c = 1$, Model 2), the RMSE of ELW is greater than those of MW1 and MW2 by at most $(5.13 - 3.78)/3.78 \approx 35.7\%$. By contrast, the RMSEs of MW1 and MW2 can be more than 12 times that of ELW; see the scenario with $\gamma = 1.5$ and $c = 0.1$ for Model 4. Although MW1 and MW2 often have smaller RMSEs than ELW for Model 1, the boxplots in Figure 1 indicates that they tend to have either non-ignorable biases or larger variances.

Models 3 and 4 are simply Models 1 and 2 with a mean shift. When we change Models 1 and 2 to Models 3 and 4, respectively, and keep the remaining settings unchanged, the boxplots demonstrate that ELW clearly performs the best: it not only is nearly unbiased, but also has the smallest variance. Meanwhile, ELW, CHIM, and SIPW have nearly unchanged RMSEs. This makes sense, because their weights all sum to 1. Unfortunately, IPW, ZZZ, MW1, and MW2 are all very sensitive to a mean shift in the data generating process, since their weights do not sum to 1.

When c decreases from 1 to 0.1, the influence of random error become negligible and we expect all methods to exhibit better performance. Indeed, all methods have decreasing RMSEs, except for IPW. ELW has the largest rates of decline in RMSE: these rates are at least 69% and 42% when $\gamma = 1.5$ and 2.5, respectively. However, the RMSEs of ZZZ, MW1, and MW2 have nearly no reduction for Models 3 and 4. ELW performs in the most stable manner, whereas the other methods have either extremely large fluctuations or remarkable biases.

When γ increases from 1.5 to 2.5, ELW clearly outperforms the competitors in all scenarios except those for Model 1. All methods exhibit similar performance for Models 1 and 2. However for Models 3 and 4, IPW, ZZZ, MW1, and MW2 have much larger fluctuations, compared with their performance for both Models 1 and 2. This indicates that they are sensitive to a mean shift, which is undesirable.

Roughly speaking, among the seven estimators under comparison, the ELW estimator is the most reliable in almost all scenarios. Both the RMSE results and the boxplots indicate that MW1 and MW2 can exhibit very different performances. In other words, the performance of the method of Ma and Wang (2020) can be affected by the choice of underlying tuning parameters. We have also conducted simulations for $N = 50$ and 500, $\gamma = 1.3$ and 1.9, and we even considered the case with estimated propensity scores. See Section 8 of the supplementary material for the RMSE results, the corresponding boxplots and additional results. The general findings are similar.

Interval estimation Two confidence intervals for θ can be constructed based on the ELW estimator $\hat{\theta}_{\mathrm{ELW}}$. One is the Wald confidence interval (ELW-an for short) based on the asymptotic normality of $\hat{\theta}_{\mathrm{ELW}}$, where the asymptotic variance is estimated using the ELW method. The other is the resampling-based interval estimator (ELW-re) given in Section 2.5. Similar intervals (SIPW-an and SIPW-re) can be constructed when the SIPW estimator takes the place of the ELW estimator in the estimations of both θ and the asymptotic variances. We compare these four confidence intervals with those of Ma and Wang (2020) based on their resampling method and the MW1 and MW2 point estimators, which are denoted by MW1-re and MW2-re, respectively. We exclude the IPW-based confidence intervals because the IPW point estimator is dramatically unstable.

We generate random data of size $N = 2000$ from Example 1, and calculate the coverage probabilities and average lengths of the eight confidence intervals at the 95% confidence level. The results are displayed in Figure 2. MW2-re has the most accurate coverage accuracy, followed by ELW-re when $\gamma = 1.5$ and $c = 1.0$ for Models 1 and 2. When Models 1 and 2 are replaced by Models 3 and 4, the coverage probabilities of ELW-re remain nearly unchanged; however, those for MW1-re and MW2-re decrease sharply by more than 5% and 10%, respectively. When c decreases from 1.0 to 0.1, the coverage accuracy of ELW-re becomes better or is still acceptable, although both MW1-re and MW2-re perform much more poorly. With different tuning parameters, MW1-re and MW2-re often have quite different coverage probabilities and average lengths, which again shows that the performance of the method of Ma and Wang (2020) can be greatly affected by different choices of tuning parameters. SIPW-re has very close coverage probabilities to

ELW-re in most cases, whereas its average lengths are generally much greater than those of the latter.

As expected, all asymptotic-normality-based Wald intervals exhibit remarkable undercoverage when $\gamma = 1.5$, because the asymptotic normalities are generally violated. In the meantime, all resampling-based intervals have improved performance. When γ increases to 2.5, all intervals except MW1-re and MW2-re have very desirable coverage accuracy, and the asymptotic-normality-based intervals have close or even better coverage probabilities compared with the resampling-based intervals.

In summary, the ELW point estimator has the most reliable overall performance, is shift-equivariant, and is nearly unbiased in all cases. The proposed resampling-based ELW interval estimator often has desirable coverage accuracy and short lengths in missing data problems, whether the proportion of extremely small propensity scores is small or large.

4. Real data analysis

LaLonde (1986) estimated the impact of the National Supported Work Demonstration, a labour training programme, on post-intervention income levels, using data from a randomized evaluation of the programme. To further demonstrate the superiority of the proposed ELW method, we analyse the `LLvsPSID` data from the `R` package `cem`, which is the Lalonde set of treated units versus PSID (Panel Study of Income Dynamics) control individuals. The data consist of 2787 observations (297 from treated units and 2490 from control units) on 12 variables: *treated* (treatment indicator), *age* (age), *education* (years of education), *black* (race, indicator variable), *married* (marital status, indicator variable), *nodegree* (indicator variable of not possessing a degree), *re74* (real earnings in 1974), *re75* (real earnings in 1975), *re78* (real earnings in 1978), *hispanic* (ethnic, indicator variable), *u74* (unemployment in 1974, indicator variable), and *u75* (unemployment in 1975, indicator variable). The variable *re78* is the post-treatment outcome.

Let $Y = re78/10\,000$ be the response, let $D =$ *treated*, and let $Y(d)$ denote the response of an individual whose treatment status is $D = d$. We shall not address the original treatment effect estimation problem. Instead, we take the data as missing data and wish to estimate the average earnings of the treated in 1978. In other words, the parameter of interest is $\theta = \mathbb{E}\{Y(1)\}$. We first estimate the propensity scores by fitting a linear logistic regression model of the treatment indicator D on the remaining eight variables (excluding D, Y, and *re78*). With the fitted propensity scores, the IPW, SIPW, MW1, MW2, and ELW point estimates are 0.65, 0.92, 0.72, 0.70, and 1.11, respectively, and the corresponding resampling-based interval estimates at the 95% level are $[-12.68, 8.00]$, $[-4.16, 3.15]$, $[-3.04, 1.26]$, $[-0.27, 1.04]$, and $[-8.86, 6.27]$, respectively. If we replace all Y by $Y+5$, the point estimates become 4.16, 5.92, 5.81, 5.56, and 6.11 with interval estimates being $[-27.31, 22.76]$, $[0.90, 8.16]$, $[0.92, 7.16]$, $[5.05, 6.22]$, and $[-3.87, 11.44]$, respectively. As expected, the SIPW and ELW point estimates are shift-equivariant, but the IPW estimator and the MW estimators are not.

Figure 3 displays the fitted propensity scores of both the treated and control groups. A clump of near-zero propensity scores in the treated group implies that the standard IPW estimator is dramatically unstable. The excessive number of near-zero propensity scores in both groups indicates that the distribution of the inverse propensity score has a very heavy right tail similar to that in the simulation scenario with $\gamma = 1.5$ in Example 1. According to our simulation experience in the case of $\gamma = 1.5$, the ELW point estimator is always unbiased or nearly unbiased, and its performance is the most stable in most cases. By contrast, the other estimators SIPW, MW1, and MW2 may have either much larger RMSEs or large biases. The ELW-re interval has the most desirable and much better coverage accuracy than the other intervals. These observations makes it reasonable to believe that the ELW point and interval estimates, 1.11 and $[-8.86, 6.27]$, are the most preferable for the estimation of $\theta = \mathbb{E}\{Y(1)\}$.

We have extended the ELW method to unequal probability samplings with and without replacements in Section 5 of the supplementary material. Poisson, pivotal, and PPS samplings are three popular unequal probability samplings. Here we regard the observations in the `LLvsPSID` data with non-zero *re75* as a finite population, and conduct Poisson, pivotal, and PPS samplings with inclusion probabilities proportional to *re75*. We take the parameter of interest to be the mean of $Y = re78/10\,000 + a$, with $a = 0$ or 2. Table 2 presents the simulated RMSE results based on 5000 simulation repetitions with a sample size (in pivotal and PPS samplings) or ideal sample size (in Poisson sampling) of 200. The ELW estimator has the smallest RMSEs under Poisson sampling, regardless of whether $a = 0$ or 2, and under pivotal and PPS samplings when $a = 2$. It also uniformly outperforms SIPW under all three samplings. When $a = 0$, its performance can be inferior to those of IPW and ZZZ, which, however, are highly sensitive to a location shift in Y. The ELW estimator again has the best overall performance under unequal probability sampling.

5. Discussion

The focus of this paper is the development of a better weighting method than IPW. We have developed the ELW method for parameter estimation under just-identified estimating equations, and shown that the ELW method is always well defined, easy to calculate and more stable than IPW. The foundation of the ELW method is the biased-sample EL. If calculation convenience or burden is not an issue, we can use the biased-sample likelihood ratio function to conduct interval estimation and hypothesis testing. When calculating the biased-sample likelihood ratio function, we need to fix the parameter value, which makes the corresponding maximum likelihood has a high probability of having no definition. If the likelihood ratio function has no definition, the ELW approach fails to work. This numerical issue is inherited from the standard EL (Chen et al., 2008). In the case of over-identified estimating equations (Qin and Lawless, 1994), the non-definition problem becomes even more serious and the calculation burden becomes even heavier because the objective function involves a vector-valued implicit function and its maximization generally requires double optimizations.

We extend the ELW method to unequal probability samplings with and without replacement in Section 5 of the supplementary material. The ELW estimator is still asymptotically normal in unequal probability samplings, and is more efficient than both the IPW and SIPW estimators when the sampling is without replacement. When the sampling is with replacement, the ELW estimator is still more efficient than the SIPW estimator. Although we cannot tell which of the ELW and IPW estimators wins in this situation, our simulation results indicate that the ELW estimator usually has smaller mean square errors than the IPW and SIPW estimators.

We systematically investigate the large-sample properties of the ELW estimator and the ELW likelihood ratio statistic under over-identified estimating equations for missing data problems and unequal probability samplings. See Section 9 of the supplementary material. The ELW likelihood ratio statistic has a limiting central chisquare distribution for missing data problems with known propensity score and unequal probability sampling with replacement. Otherwise, its limiting distribution is usually a weighted chisquare distribution.

Data availability and funding statement

The `LLvsPSID` data underlying this article are available in the `R` package `cem`. The research was supported by the National Natural Science Foundation of China (12171157, 11971300, 71931004, 32030063, 11771144), the Natural Science Foundation of Shanghai (19ZR1420900), and the 111 Project (B14019).

Acknowledgements

The authors thank the Editor, the Associate Editor, and two referees for valuable comments and suggestions that have led to significant improvements in the paper.

References

Bang, H. and Tsiatis, A. A. (2000) Estimating medical costs with censored data. *Biometrika*, **87**, 329–343.

Busso, M., DiNardo, J. E. and McCrary, J. (2014) New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, **96**, 885–897.

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

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

Chen, J., Variyath, A. M. and Abraham, B. (2008) Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, **17**, 426–443.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**, 187–199.

Dong, G., Mao, L., Huang, B., Gamalo-Siebers, M., Wang, J., Yu, G. and Hoaglin, D. C. (2020) The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of Biopharmaceutical Statistics*, **30**(5), 882–899.

Hájek, J. (1971) Discussion of 'An essay on the logical foundations of survey sampling, Part One' by D. Basu. In *Foundations of Statistical Inference* (eds. V. P. Godambe and D. A. Sprott), vol. 236. Toronto: Holt, Rinehart and Winston.

Han, P., Kong, L., Zhao, J. and Zhou, X. (2019) A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society, Series B*, **82**(2), 305–333.

Hansen, M. and Hurwitz, W. (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14**, 333–362.

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

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

Imbens, G. W. and Wooldridge, J. (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, **47**(1), 5–86.

Jiang, R., Lu, W., Song, R. and Davidian, M. (2017) On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society, Series B*, **79**, 1165–1185.

Kang, J. D. and Schafer, J. L. (2007) A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, 523–539.

Khan, S. and Tamer, E. (2010) Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, **78**(6), 2021–2042.

Kim, J. K., and Shao, J. (2021) *Statistical methods for handling incomplete data.* Chapman and Hall/CRC.

LaLonde, R. J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, **76**, 604–620.

Liu, Y. and Chen, J. (2010) Adjusted empirical likelihood with high-order precision. *Annals of Statistics*, **38** (3), 1341–1362.

Ma, Y. and Yin, G. (2011) Censored quantile regression with covariate measurement errors. *Statistica Sinica*, **21**, 949–971.

Mccaffrey, D. F., Lockwood, J. R. and Setodji, C. M. (2013) Inverse probability weighting with error-prone covariates. *Biometrika*, **100**(3), 671–680.

Ma, X. and Wang, J. (2020) Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, **115**(532), 1851–1860.

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

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

Owen, A. B. (2001) *Empirical Likelihood.* New York: Chapman and Hall.

Politis, D. N. and Romano, J. P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, **22** (4), 2031–2050.

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

Robins, J. M. (1993) Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pp. 24–33. Alexandria, VA: American Statistical Association.

Robins, M. and Finkelstein, D. M. (2000) Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, **56**, 779–788.

Robins, J. M. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology – Methodological Issues* (eds. N. Jewell, K. Dietz and V. Farewell), pp. 297–331. Boston: Birkhäuser.

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.

Robins, J., Sued, M., Lei-Gomez, Q, and Rotnitzky, A. (2007) Comment: Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statistical Science*, **22**(4), 544–559.

Rosenbaum, P. R. (2002) *Observational Studies.* New York: Springer.

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

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

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

Sun, B., and Tchetgen Tchetgen, E. J. (2018) On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association*, **113**, 369–379.

Tan, Z. (2010) Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, **97**(3), 661–682.

Tan, Z. (2020) Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, **48**(2), 811–837.

Tsao, M. (2004) Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Annals of Statistics* **32** 1215–1221.

Wang, H., Yang, M. and Stufken, J. (2019) Information-based optimal subdata selection for big data linear regression, *Journal of the American Statistical Association*, **114**, 393–405.

Wang, H., Zhu, R. and Ma, P. (2018) Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, **113**, 829–844.

Wang, S., Shao, J., and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, **24**, 1097–1116.

Wooldridge, J. M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, **141**, 1281–1301.

Yang, S. and Ding, P. (2018) Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, **105**(2), 487–493.

Young, J. G., Logan, R. W., Robins, J. M. and Hernán, M. A. (2019) Inverse probability weighted estimation of risk under representative interventions in observational studies. *Journal of the American Statistical Association*, **114**, 938–947.

Yu, J., Wang, H., Ai, M. and Zhang, H. (2022) Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, **117**, 265–276.

Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2012) A robust method for estimating optimal treatment regimes. *Biometrics*, **68**, 1010–1018.

Table 1. Simulated RMSEs of the estimators under comparison when data are generated from Example 1 and $N = 2000$. Smallest RMSEs are highlighted in bold.

γ	c	Model	IPW	SIPW	ZZZ	CHIM	MW1	MW2	ELW
1.5	1.0	1	24.72	8.05	6.05	8.00	4.96	**4.35**	5.51
1.5	1.0	2	17.89	6.17	5.95	6.17	4.84	**3.78**	5.13
1.5	1.0	3	69.08	7.49	27.27	7.49	18.29	9.59	**5.21**
1.5	1.0	4	110.80	6.49	27.11	6.49	18.31	9.83	**5.21**
1.5	0.1	1	14.76	4.89	4.48	4.87	3.00	2.52	**1.60**
1.5	0.1	2	26.23	2.16	4.44	2.15	2.88	1.36	**0.71**
1.5	0.1	3	68.12	4.74	27.04	4.73	17.81	8.94	**1.61**
1.5	0.1	4	140.05	2.21	26.86	2.19	18.02	9.03	**0.74**
2.5	1.0	1	2.11	2.11	1.97	2.11	1.93	**1.87**	2.02
2.5	1.0	2	2.06	1.81	1.90	1.81	1.89	1.82	**1.72**
2.5	1.0	3	7.64	2.15	6.77	2.15	6.33	5.32	**2.05**
2.5	1.0	4	8.14	1.85	7.31	1.85	6.87	6.01	**1.70**
2.5	0.1	1	1.49	1.33	1.14	1.33	1.07	**0.98**	1.17
2.5	0.1	2	1.22	0.69	1.01	0.69	0.93	0.77	**0.42**
2.5	0.1	3	7.63	1.31	6.60	1.31	6.21	5.14	**1.18**
2.5	0.1	4	8.26	0.68	7.13	0.68	6.80	5.85	**0.42**

Table 2. Simulated RMSEs of the IPW, SIPW, ZZZ and ELW estimators when data were generated from the `LLvsPSID` dataset with $n = 200$ with Y replaced by $Y + a$.

	IPW	ZZZ	IPW	ZZZ	SIPW	ELW
	$a = 0$		$a = 2$			
Poisson sampling	9.35	8.44	19.27	16.33	8.41	6.14
Pivotal sampling	5.07	3.91	12.17	7.63	7.15	4.66
PPS sampling	5.46	4.13	13.86	8.19	8.70	5.51

Zhao, Q. (2019) Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, **47**(2), 965–993.

Zong, X., Zhu, R. and Zou, G. (2019) Improved Horvitz–Thompson estimator in survey sampling. *Survey Methodology*, **45**(1), 165–184.

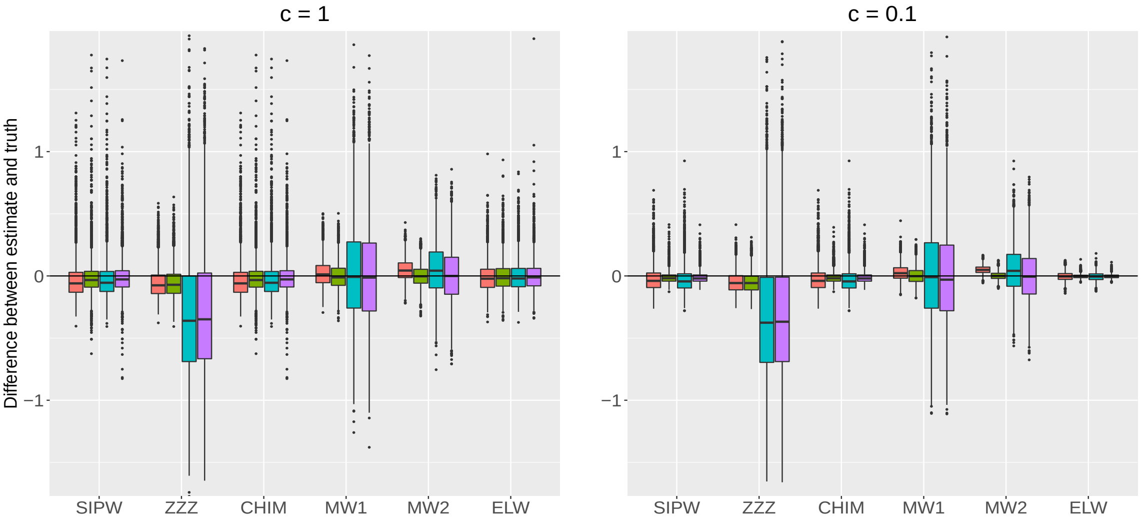

Fig. 1. Boxplots of the SIPW, ZZZ, CHIM, MW1, MW2, and ELW estimators (minus the true parameter values) when data were generated from Example 1 with $N = 2000$, $\gamma = 1.5$, and $c = 1$, 0.1. For each case of c and each method, the four boxplots from left to right and in red, green, blue and purple correspond to models 1-4, respectively.

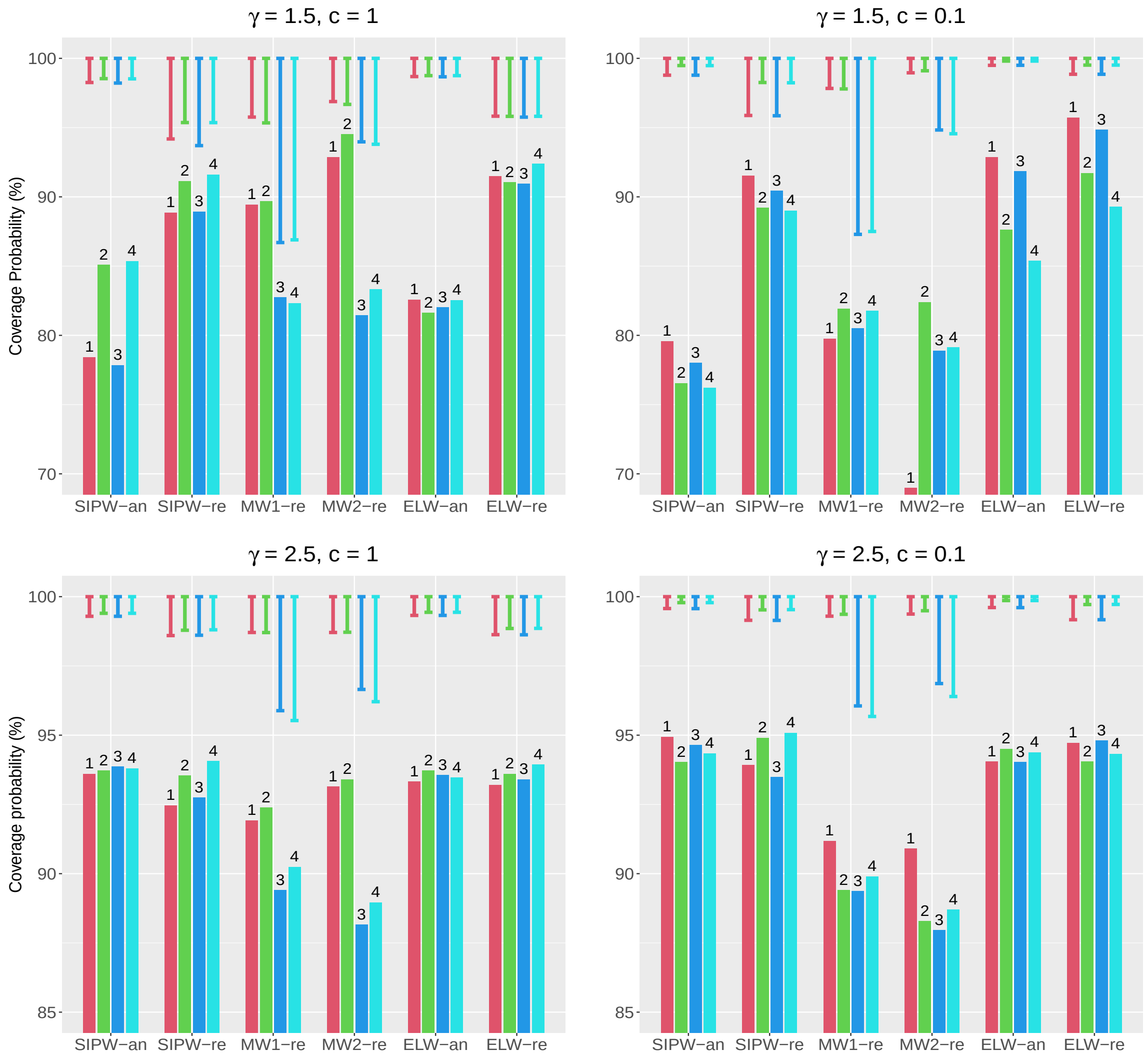

Fig. 2. Simulated coverage probabilities (%) of the interval estimators under comparison (SIPW-an, SIPW-re, MW1-re, MW2-re, ELW-an and ELW-re) when data were generated from Example 1 with full data size $N = 2000$ and different choices of γ and c. The number above each bar is model number. The length of the segment above each bar equals five times the average length of the corresponding interval estimator.

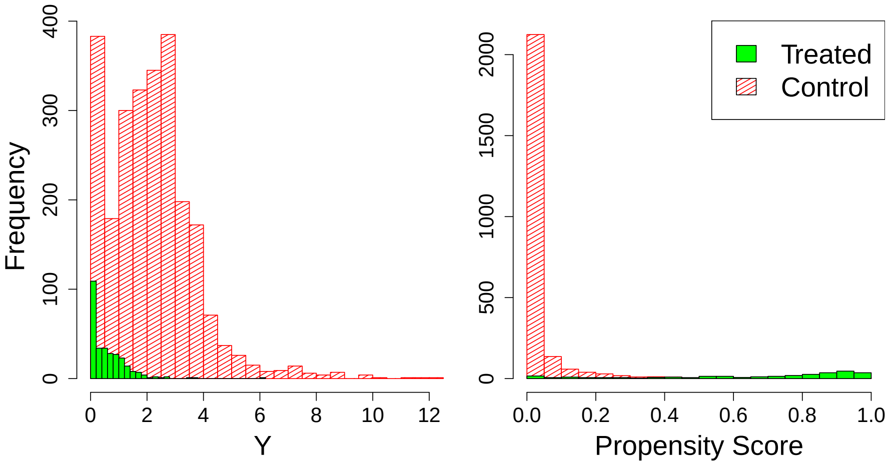

Fig. 3. Histograms of the variable Y and the fitted propensity score in the treated and control groups, based on the `LLvsPSID` data.