Journal of Modern Applied Statistical Methods

Volume 19 | Issue 2 Article 4

A Simulation study of some Modern Weighting Methods for Estimating Treatment Effects in Observational Studies

Lateef Babatunde Amusa
Department of Statistics, University of Ilorin, Ilorin, Nigeria, amusa.ib@unilorin.edu.ng

Temesgen Zewotir
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Delia North
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Recommended Citation

Lateef Babatunde Amusa, Temesgen Zewotir, Delia North (2020). A Simulation study of some Modern Weighting Methods for Estimating Treatment Effects in Observational Studies. Journal of Modern Applied Statistical Methods, 19(2), https://doi.org/10.56801/Jmasm.V19.i2.4

Journal of Modern Applied Statistical Methods
2020, Vol. 19, No. 2,
Doi: 10.56801/Jmasm.V19.i2.4

ISSN 1538 – 9472

A Simulation study of some Modern Weighting Methods for Estimating Treatment Effects in Observational Studies

Lateef Babatunde Amusa
Department of Statistics, University of Ilorin, Ilorin, Nigeria

Temesgen Zewotir
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Delia North
Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Using extensive Monte Carlo simulations, we compared the performance of entropy balancing, empirical balancing calibration weighting (EBCW), covariate balancing propensity score, and the inverse probability treatment weighting methods, via the performance of treatment effects estimation. The consistently superior performance of EBCW and entropy balancing leads us to recommend these two modern weighting techniques.

Keywords: Monte Carlo, Treatment Effects, Weighting, Propensity score, Entropy balancing.

Introduction

Estimating average treatment effects is essential in the evaluation of a treatment or intervention. It is particularly straightforward in experiments, but very complicated in observational studies, where the treatment assignment is not random. The complication comes from the fact that treatment exposure may be associatedwith background covariates that are also associated with the potential response. This may substantially introduce covariates imbalance in these treatment groups.There is thus the need for methods that adjust for such confounding of the background covariates, for a reliable causal inference to be made from observational data.

Considering the concerns above, the use of propensity score methods as an adjustment method has become prevalent in applied studies(Austin, 2014; Dehejia & Wahba, 2002; Guo, Barth, & Gibbons, 2006; Guo & Fraser, 2010). Specifically, propensity score matching methods have been widely used, until more recently, propensity score weighting methods have taken centre stage in estimating causal effects(Guo et al., 2006; Hirshberg & Zubizarreta, 2017). Weighting methods, unlike matching methods, have the apparent advantage of not discarding units, which in turn reduces the standard error of estimates. Weighting approachadjusts for

confounding by constructing weights for individual units to match the target population. When the weightsare constructed from the inverse of propensity scores, the methodis referred to as the inverse probability weighting (IPW). IPW is the most common weighting method to applied researchers and practitioners, especially in the medical and health sciences(Austin & Stuart, 2015).

Despite the popularity and simplicity of the IPW, it heavily relies on the correct specification of the propensity score model(Kang & Schafer, 2007). Though Lee and Stuart (2010) suggested machine learning methodsas a promising alternative to logistic regression for propensity score estimation, they are data-hungry andmore computationally intensive in most cases. More recently, the construction of weights took a different dimension:the weights are chosen using some optimization algorithm to perfectly balance the covariates, subject to some specified constraints (Chan, Yam, & Zhang, 2016; Hainmueller, 2012; Imai & Ratkovic, 2014).These optimization-based techniques have an inbuilt facility of directly incorporating a balance condition for the moments the covariates in the estimation procedure, thereby ensuring perfect covariate balance. Accordingly, the conventional balance checking is not necessary for these methods. It is important to study these methods and demonstrate their effectiveness, as they have been given relatively little attention in applied studies. Although a few prior studies (Setodji et al., 2017; Harvey et al., 2017; Wyss et al., 2014) compared these optimization-based techniques with the IPW method, there is a paucity of research in that regard.

This paper, using the IPW method as a benchmark, aims toprovide a comparative study of main weighting methods that guarantee a near-perfect covariates balance, in terms of performance of their treatment effects estimation. This was achieved through Monte Carlo simulations using the average treatment effect among the treated (ATT) as the estimand of interest.Without loss of generality, the focus is on entropy balancing (Hainmueller, 2012), covariate balancing propensity score (Imai &Ratkovic, 2014), and empirical balancing calibration weighting (Chan, Yam, & Zhang, 2016) methods.

Methodology

The performance of the widely used Inverse probability weighting method was comparedwith some modern weighting methods, namely; entropy balancing,covariate balancing propensity scores, and empirical balancing calibration weighting. These comparisons were made by assessing the performance of treatment effects estimation,via the absolute biases and root mean squared errors (RMSE) of estimated treatment effects.

Consider the unit i (i =1, ..., n); we assumed that there is a binary treatment variableZ_i, coded 1 and 0 for treated and control groups, respectively. Let Y_i(z): z $\epsilon\{0,1\}$ be the potential outcome variable value, that is, the value of the outcome variable if Z_i = z, also known as a counterfactual outcome (Rubin, 1974). This implies that $Y_i = Z_iY_i(1) + (1 - Z_i)Y_i(0)$ is observed. Let X_i be a vector of pre-treatment covariates.

SATT = $\frac{1}{n_z}\sum_{i\epsilon Z} TE_i$, where $TE_i = Y_i(1) - Y_i(0)$.

Simulation scheme

- Two different phases of simulations were conducted overall. For each Phase and model, 1000 datasets, each of varying sample sizeswere simulated. The performance of the estimated treatment effects was assessed by its absolute bias, calculated asE ($|\hat{\gamma} - \gamma|$), andthe root mean square error (RMSE), calculatedas E ($(\hat{\gamma} - \gamma)^2$).
  All simulations were done using the R statistical package version 3.5.1.

Phase 1:In this Phase, the previous simulation design was replicated with slight modifications(Abdia, Kulasekera, Datta, Boakye, & Kong, 2017; Leacy & Stuart, 2014).Under this Phase, two scenarios were considered. In estimating the ATT. The two Scenarios were:

Scenario I: Tenstandard normal distributed variables, X_1, X_2, …, X_{10}were generated and includedas confounders. This scenario is aimed at capturing what is obtainable as a first resort in practice, where all covariates are included in the treatment assignment model and the outcome regression model in a linear fashion. Treatment assignment was modelled as

$$\text{Logit}(\pi_i) = \alpha_0 + \sum_k^{10} \alpha_k X_k. \quad (1)$$

Outcome variable was generated, with true ATT ($\gamma = 2.568$), as

$$Y_i = \beta_0 + \gamma Z_i + \sum_k^{10} \beta_k X_k + \varepsilon_i, \varepsilon_i \sim N\left(0, \frac{\text{Var } E(Y/X)}{50}\right) \quad (2)$$

Scenario II: Not all the covariatesX_1, X_2, …, X_{10}were includedin both the treatment assignment and outcome models. Further, interactions of some covariates were included in the models. Treatment assignment was modelled as

$$\text{Logit}(\pi_i) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_1^2 + \alpha_5 X_2^2 + \alpha_6 X_2 X_3 + \alpha_7 X_1 X_2 X_3 \quad (3)$$

Outcome variable was generated with the true ATT ($\gamma = 2.568$) as

$$Y_i = \beta_0 + \beta_1 X_1^2 Z_i + \beta_2 X_4 Z_i + \beta_3 X_1 X_4 (1 - Z_i) + \beta_4 X_5 (1 - Z_i) + \varepsilon_i, \varepsilon_i \sim N\left(0, \frac{\text{Var } E(Y/X)}{50}\right) \quad (4)$$

For each subject *i*, in both scenarios, treatment status was generated as $Z_i \sim$ Bernoulli (π_i). The non-zero coefficients are set to reflect low, medium, high and very high effect sizes, respectively.

Phase 2: In thisPhase, the simulation was made to be as realistic as possible by simulatingfrom real-life data. Treatment and outcome variables were generated from the covariates of the Lalonde-CPS data. The dataset is a merger of program

participants (treated units) from Lalonde's experimental data(LaLonde, 1986)and the control group drawn from the Current Population Survey (CPS) participants. The dataset comprises ten covariates including age (*age*), number of years of education (*edu*), real earnings in 1974 (*re74*) and 1975 (*re75*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), marital status (*married*), no high school diploma/degree (*nodeg*), hispanic race (*hisp*), and black race (*black*). Treatment variable was generated similarly to Phase I. Outcome variable was generated with the true ATT ($\gamma = 1000$), and coefficients being the ones from fitting such model to the real data as

$$Y = \gamma T + \beta_1 \text{ age} + \beta_2 \text{edu} + \beta_3 \text{re74} + \beta_4 \text{re75} + \beta_5 \text{ married} + \beta_6 \text{ black} + \beta_7 \text{hisp} +$$

$$\beta_8 \text{nodeg} + \beta_9 \text{ u74} + \beta_{10} \text{u75} + \varepsilon, \varepsilon_i \sim N(0, 10). \quad (5)$$

Weighting methods

In this section, the weighting methods that wereincluded in the simulation study are briefly described.

Inverse Probability Treatment Weighting (IPW)

The propensity score, defined by e (x) = P(Z = 1 | X),0 < e < 1, is the probability of a subject or unit receiving the treatment of interest given the observed baseline covariates(Rosenbaum, 1983).In estimating the ATT,the IPW is definedas fixing treated units' weight at unity, and the control units as $\frac{e\,(x)}{1-e\,(x)}$(Imbens, 2004).In IPW, each unit's weight equals the reciprocal of the probability of receiving the treatment thatthe unit received. It is unlikely for the propensity score to be known in practice, so it is routinely estimated using a parametric model, like the logistic model. Therefore, the success of the IPW largely rests on the correct specification of the propensity score model. The IPW estimator is defined as:

$$\hat{\gamma}_{ATT} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} Y_i (1-Z_i) \frac{e(X_i)}{1-e(X_i)}}{\sum_{i=1}^{n} (1-Z_i) \frac{e(X_i)}{1-e(X_i)}} \quad (6)$$

Entropy balancing (ebal)

Entropy balancing is a preprocessing method that can guarantee covariates balance via a reweightingscheme thatassigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraintsthat are imposed on the sample moments of the covariate distributions (Hainmueller, 2012). The ATT is estimated using the difference in mean outcomes between the treated and reweighted control group. After assuming uniform weights w_i, the following reweighting scheme reweights the control group:

$$\min_{w_i} H(w) = \min_{w_i} \sum_{i|T=0} w_i \log (w_i / q_i) \quad (7)$$

Subject to the following balance and normalization constraints:

$\sum_{i|T=0} w_i c_{ri}(X_i) = m_r$, for r = 1, ..., R (7.1)

$\sum_{i|T=0} w_i = 1$ (7.2)

$w_i \geq 0$ for all i, (7.3)

Where $q_i = \frac{1}{n_0}$ is a vector of the base weights, and m_r represents a set of R balance constraints imposed on the moments of the reweighted control group, and $c_{ri}(X_i)$.

Covariate Balancing Propensity Scores (CBPS)

The CBPS (Imai & Ratkovic, 2014) is a robust version of the propensity score weighting methods. It was developed to be robust to mild misspecification of the propensity score model by exploiting the dual properties of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. The CBPS is operationalized by using IPW, with the moment condition, with $\tilde{X}_i$ = f (X_i) being a vector-valued measurable function of X_i:

$$E\left\{\frac{Z_i \tilde{X}_i}{e_\beta X_i} - \frac{(1-Z_i)\tilde{X}_i}{1-e_\beta(X_i)}\right\} = 0 \quad (8)$$

It further ensures that each covariate's first moment is balanced even when the model is misspecified, by setting $\tilde{X}_i = X_i$. Also, the first and second moments will be balanced if $\tilde{X}_i = (X_i^T X_i^{2T})^T$. In estimating the ATT, the control group units are weighted such that it matches that of the treated group by utilizing the moment condition, which becomes

$$E\left\{Z_i \tilde{X}_i - \frac{e_\beta(X_i)(1-Z_i)\tilde{X}_i}{1-e_\beta(X_i)}\right\} = 0 \quad (9)$$

Empirical Balancing Calibration Weighting (EBCW)

The EBCW (Chan et al., 2016) is a globally efficient nonparametric method of estimating treatment effects. It was built on the notion of achieving efficiency by solely balancing the covariate distributions without a direct estimation of the propensity score or outcome regression function. The EBCW estimator belongs to the general class of calibration estimators which minimize the overall distance between the final weights to a given vector of design weights, subject to moment constraints. It considers a vector of misspecified uniform design weights d*= (1,1, …, 1) and constructs weights w by solving:

$$\text{Min} \sum_{i=1}^{n} D\ (w_i, 1) \text{ subject to}$$

$$\frac{1}{n}\sum_{i=1}^{n} Z_i\, w_i \text{ u } (X_i) \qquad \text{and}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)w_i \text{ u } (X_i), \text{ where D(.) is a distance metric.}$$

Results

In this section, the results obtained from analyzing the simulated datasetsare presented and explained. For all considered scenarios, as expected and in alignment with statistical theory, the RMSEs of the weighting methods decreased with increasing sample sizes. In addition to the tabular representation of the simulation results, the performances of the considered weighting methods, are depicted in the graphical plots provide in Figures 1,2, 3, 4, and 5, for better understanding.

For Phase 1 (Scenario I), in Table 1 and Figure 1, there is the case of the propensity score model being assumed false, and where the IPW method has been established to perform worse(Kang & Schafer, 2007). As expected, the IPW method yielded the highest RMSEs across all sample sizes. In terms of bias, the weighting methods, except for CBPS, had minimal values and areclose to zero. As CBPS is a direct improvement over IPW, it increases the efficiency of estimates at the expense of an additional increase in bias. The EBCW outperformed the others in terms of bias;however, EBCW and entropy balancing methods performed similarly, with both methodsdominating the other two methods in terms of RMSE.

Table 1: Simulation results for estimating treatment effects under Phase 1, Scenario I.

		Sample size			
	Method	500	1000	2000	5000
Absolute bias	IPW	0.0448	0.0284	0.0079	0.0106
	EBAL	0.0096	0.0072	0.0029	0.0008
	CBPS	0.4148	0.2730	0.1628	0.0908
	EBCW	0.0023	0.0039	0.0017	0.0008
RMSE	IPW	0.7024	0.5787	0.3574	0.2593
	EBAL	0.1435	0.0992	0.0692	0.0435
	CBPS	0.5186	0.3509	0.2225	0.1244
	EBCW	0.1431	0.0992	0.0692	0.0435

Table 2: Simulation results for estimating treatment effects under Phase 1, Scenario II.

		Sample size			
	Method	500	1000	2000	5000
Absolute bias	IPW	0.0169	0.0137	0.0284	0.0444
	EBAL	0.0322	0.0027	0.0198	0.0405
	CBPS	0.0138	0.0051	0.0179	0.0369
	EBCW	0.0323	0.0028	0.0198	0.0405
RMSE	IPW	0.7692	0.6370	0.4624	0.3227
	EBAL	0.6617	0.5003	0.3608	0.2471
	CBPS	0.5170	0.4148	0.3200	0.2358
	EBCW	0.6624	0.5005	0.3608	0.2471

For Phase 1 (Scenario II), in Table 2 and Figure 2, the propensity score model is assumed known. In terms of bias, all the weighting methods had minimal values and are close to zero. Though there is no clear pattern of superiority or otherwise among the weighting methods, the CBPS method produced the smallest absolute biases at samples 1000, 2000, and 5000. There is evidence of a substantial gain in bias reduction for the CBPS method as compared to the situation where a false propensity score model was assumed. In terms of RMSE, CBPSoutperformed the other techniques across the considered sample sizes. Further, when thecorrect propensity score model was specified, the IPW method did not do better than the entropy balancing and EBCW.

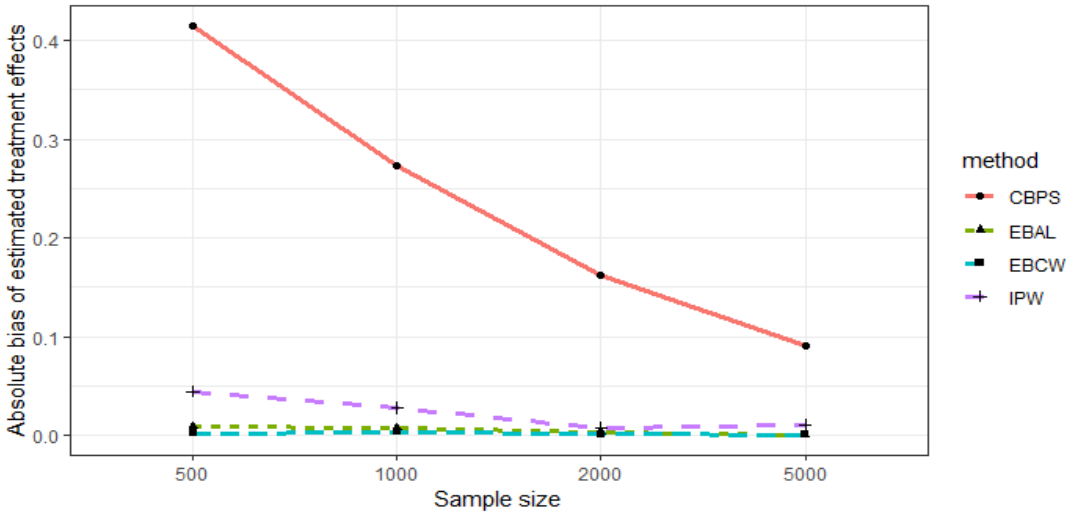

Figure 1: Absolute bias of treatment effects treatment effects of the weighting methods and varying sample sizes under Phase 1, Scenario I.

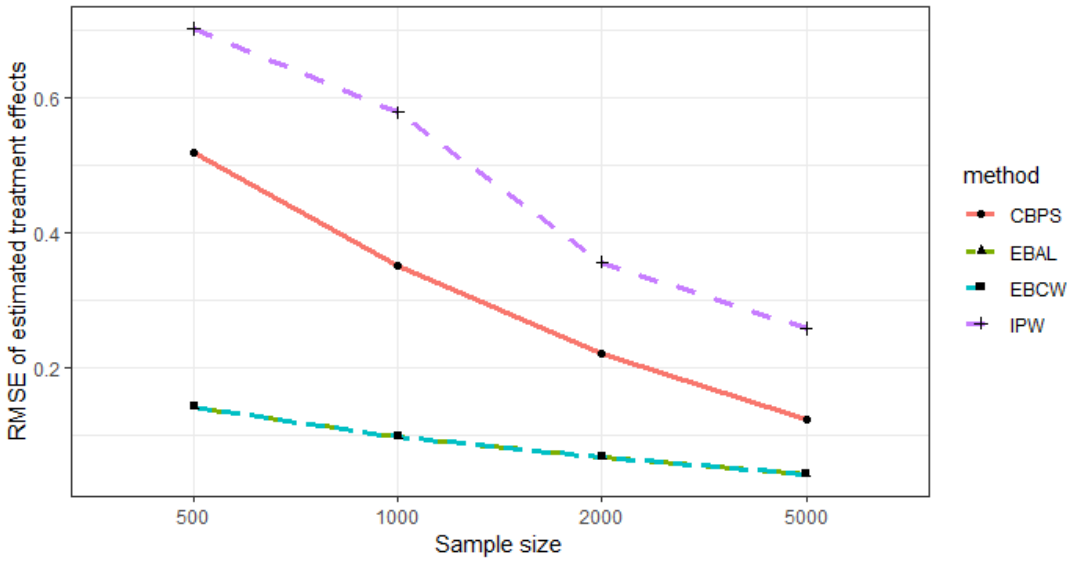

Figure 2: RMSE of treatment effects of the weighting methods and different sample sizes under Phase 1, Scenario I.

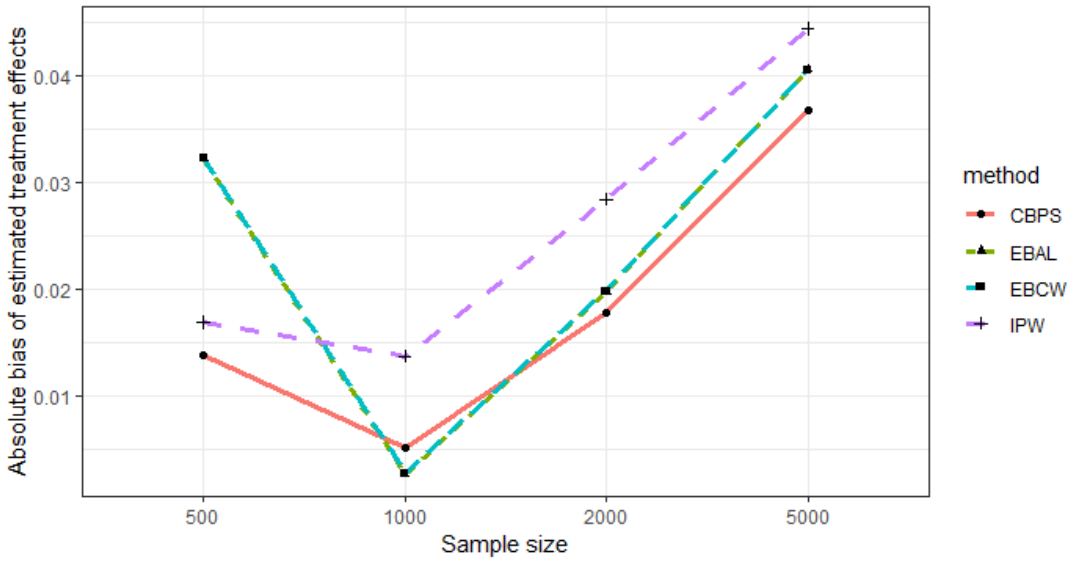

Figure 3: Absolute bias of treatment effects of the weighting methods and different sample sizes under Phase 1, Scenario II.

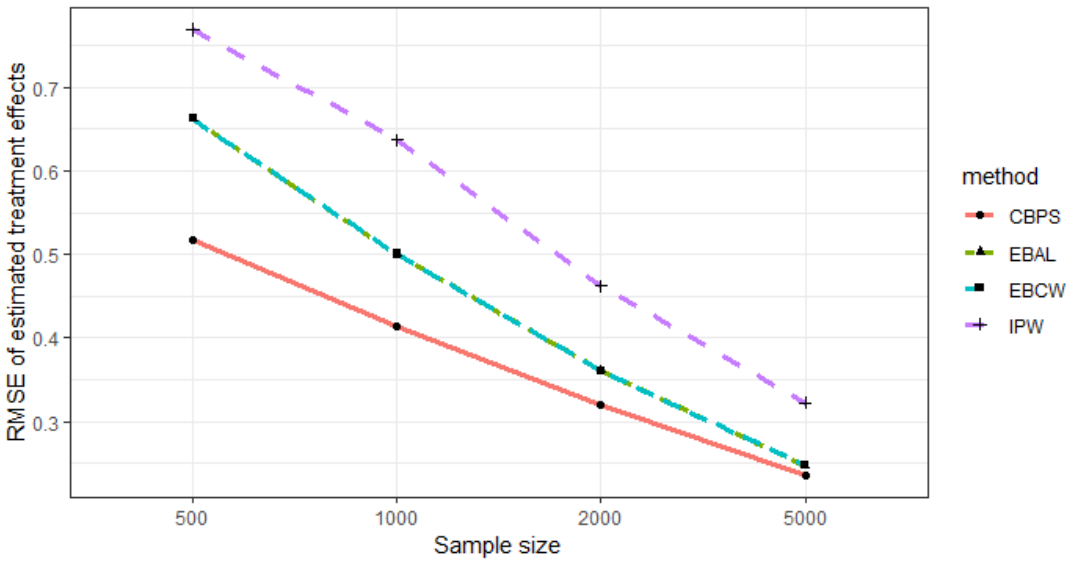

Figure 4: RMSE of treatment effects of the weighting methods and differentsample sizes under Phase 1, Scenario II.

Finally, we report the results from Phase 2 (Table 3 and Figure 5). The relatively higher values of the bias and RMSE, for this scenario, is due to the high variance of the Lalonde-CPS's outcome variable. The EBCW method, followed by entropy balancing, performed best in terms of bias and RMSE. IPW method had the worst performance. Relative to IPW, theEBCW, entropy balancing and CBPS methodshad 99.87%,99.39%, and 55.52% lower absolute biases, respectively; while they respectively had99.42%, 99.4% and 71.28% lower RMSEvalues.

Table 3: Simulation results for estimating treatment effects under Phase 2.

	Method			
	IPW	EBAL	CBPS	EBCW
Absolute bias	25.8828	0.1576	11.5517	0.0315
RMSE	224.6080	1.3439	64.5137	1.3109

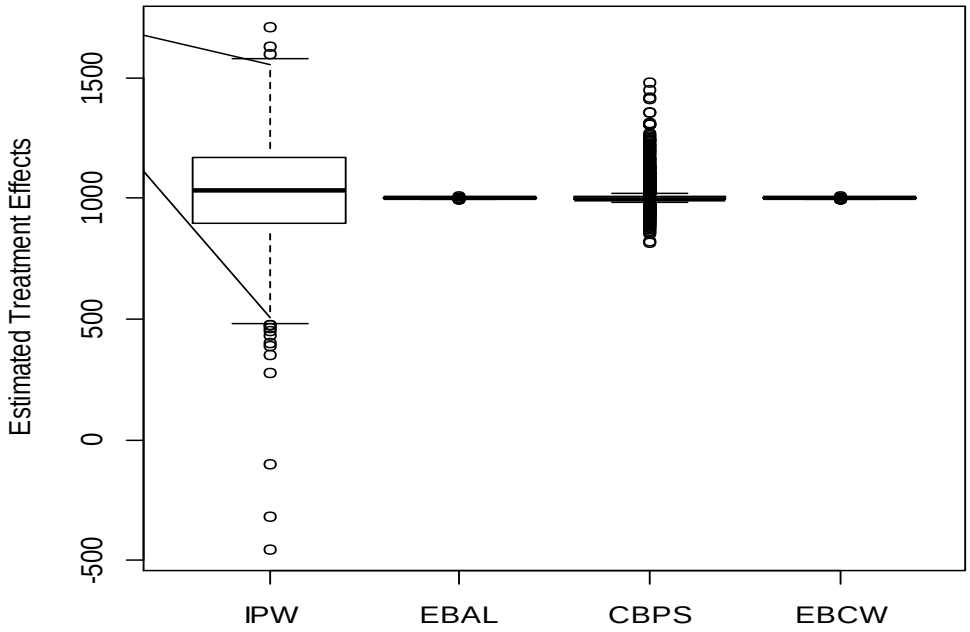

Figure 5: Estimated treatment effects for the weighting methods, under Phase 2.

Conclusion

The inverse probability of weighting (IPW) method has been the standardweighting technique for estimating treatment effects in observational studies. More recently, modern weighting techniques (unlike the IPW) that do not require correct specification of the treatment assignment model and can guarantee perfect covariates balance were introduced. There is a paucity of research studying the comparative effectiveness of these modern weighting techniques.This study addresses the gap in the existing literature and provides important information on each of them, as well as demonstrating their comparative superiority in the estimationof treatment effects. With specific reference to the average treatment effects among the treated (ATT) as the estimand of interest, we included three of these methods in our simulation study.

In this comparative study, using the IPW as a benchmark, we investigated the performance of entropy balancing, EBCW, and CBPS methods in bias reduction and efficiency gains in the estimation of treatment effects. Whether the excellent performances of these modern weighting methods translate to a better estimation of treatment effects, has not been extensively studied.Though the EBCW estimates produced quantitatively better results (slightly better than entropy balancing in most cases), the consistently superior performance of EBCW and entropy balancing leads us to recommend these two modern weighting techniques for future consideration in the estimation of causal treatment effects.

The excellent performance of the CBPS method in the case where the correct propensity score is assumed true indicates that CBPS, relative to IPW (being methods that depend on the propensity score), takes better advantage of correctly specifying the propensity score model to perform optimally.These results support the findings of Imai and Ratkovic (2014). However, knowing the correct propensity score model in practice is a tall order. It takes a highly skilled user to specify what is close to the correct propensity score model.

In this study, only the basic, off-the-shelf versions of each of the weighting methods were utilized, since that is what most applied practitioners would likely do. For instance, while techniques like entropy balancing, CBPS, and EBCW, produced large weights, the estimation algorithms could further trim extreme weights.

To our knowledge, this study is the first to jointly study the performance of these modern weighting methods, relative to the conventional IPW method. Previous studies like (Setodji et al., 2017; Harvey et al., 2017; Wyss et al., 2014) either compared two modern weighting methods or compared only one of the methods with the IPW method. Findings from our simulations are reliable and generalizable because they were based on established and accessible designs, as well as being consolidated with simulations from real-life data. It would be interesting to expand the simulation scenarios and to accommodate other estimands in future studies.We also recommend to extensively study the effect of weight trimming on the performance of these modern weighting methods, as it was done for PS weighting in a previous study (Lee et al. 2011).

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. Biometrical Journal, 59(5), 967-985.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine, 33(6), 1057-1069.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine, 34(28), 3661-3679.

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(3), 673-700.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics, 84(1), 151-161.

Guo, S., Barth, R., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. Children and Youth Services Review, 28, 357–383.

Guo, S., & Fraser, M. W. (2010). Propensity score analysis; Statistical methods and applications: SAGE Publications.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25-46.

Harvey, R. A., Hayden, J. D., Kamble, P. S., Bouchard, J. R., & Huang, J. C. (2017). A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. Pharmacoepidemiology and drug safety, 26(4), 368-377.

Hirshberg, D. A., & Zubizarreta, J. R. (2017). On Two Approaches to Weighting in Causal Inference. Epidemiology, 28(6), 812-816.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 243-263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86(1), 4-29.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science, 22(4), 523-539.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American economic review, 604-620.

Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. Statistics in Medicine, 33(20), 3488-3508.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. PloS one,6(3), e18174.

Rosenbaum, P. R. (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(5), 688.

Setodji, C. M., McCaffrey, D. F., Burgette, L. F., Almirall, D., & Griffin, B. A. (2017). The right tool for

the job: Choosing between covariate balancing and generalized boosted model propensity scores. Epidemiology (Cambridge, Mass.), 28(6), 802.