Formatting Instructions For NeurIPS 2022

Tianyu Xie
School of Mathematical Sciences
Peking University
xietianyu@pku.edu.cn

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Simulation Studies

Although the authors of ARTICLE had done experiments to verify the validity of their propoesd methods, we will repeat their experiments in this review. In this section, we aim to evaluate performance of of various estimators discussed earlier in the article using artificial data, following the same settings in ARTICLE.

Data is generated from the following models:

$$\begin{cases} \delta^{\mathrm{L}}(X) = \tanh\left(\alpha^{\mathrm{T}} X\right) \\ \delta^{\mathrm{M}}(X) = \exp\left(\alpha^{\mathrm{T}} X\right) \\ \phi_i(X) = \mathrm{sigmoid}\left(\beta_i^{\mathrm{T}} X\right), i = 1, \ldots, 4 \\ {}_{\mathrm{OP}}{}^{\mathrm{CO}}(X) = \exp\left(\eta^{\mathrm{T}} X\right) \\ \Pr(Z = 1 \mid X) = \mathrm{sigmoid}\left(\gamma^{\mathrm{T}} X\right) \end{cases} \tag{1}$$

where we set $\alpha = (0, -1)^{\mathrm{T}}, \beta_i = (-0.4, 0.8)^{\mathrm{T}}, i = 1, \ldots, 4, \eta = (-0.4, 1)^{\mathrm{T}}$ and $\gamma = (0.1, -1)^{\mathrm{T}}$. The covariates X has two dimensions, including an intercept and another covariate generated from Uniform$(-1, 1)$. Our goal is to estimate α and then estimate the local average treatment effect, while other nuisance parameters are estimable but of no interest.

Similar to ARTICLE, we will fit models of the same functional form as in (1), although this is impossible in practical applications. We repeat the simulation studies of

- `mle`: the maximum likelihood estimator in ARTICLE;
- `drw`: the doubly robust estimator with the optimal weighting function in ARTICLE;
- `dru`: the doubly robust estimator with the identity weighting function in ARTICLE;

and cite the reported results of

- `reg.ogburn`: the outcome regression estimator in Ogburn et al. (2015, ğ 3.1);
- `drw.ogburn`: the doubly robust estimator in Ogburn et al. (2015, ğ 3.3) with the optimal weighting function;
- `dru.ogburn`: the doubly robust estimator in Ogburn et al. (2015, ğ 3.2) with the identity weighting function;
- `mle.wang`: the maximum likelihood estimator in Wang & Tchetgen Tchetgen (2018, ğ 4.1);

Preprint. Under review.

	LATE		MLATE	
Method	α_0	α_1	α_0	α_1
`mle.bth`	4.55 (0.44)	9.31 (0.93)	13.81 (0.81)	18.33 (1.32)
`mle.opc`	5.86 (0.48)	11.94 (0.95)	16.47 (0.85)	21.11 (1.35)
`mle.psc`	5.86 (0.48)	11.94 (0.95)	16.47 (0.85)	21.11 (1.35)
`mle.bad`	19.15 (0.69)	2.80 (0.62)	41.28 (1.46)	0.53 (1.05)
`dru.bth`	0.95 (0.45)	6.16 (1.04)	3.58 (1.00)	9.75 (1.79)
`dru.opc`	1.21 (0.43)	4.55 (0.97)	0.02 (1.05)	10.78 (1.99)
`dru.psc`	1.21 (0.43)	4.55 (0.97)	0.02 (1.05)	10.78 (1.99)
`dru.bad`	15.25 (0.70)	29.99 (1.80)	25.23 (1.43)	17.30 (2.56)
`drw.bth`	0.29 (1.04)	1.55 (1.20)	3.04 (1.24)	4.75 (1.77)
`drw.opc`	3.71 (1.27)	0.63 (1.45)	0.30 (1.10)	6.39 (1.78)
`drw.psc`	3.71 (1.27)	0.63 (1.45)	0.30 (1.10)	6.39 (1.78)
`drw.bad`	10.97 (0.56)	13.68 (1.19)	29.54 (1.65)	4.18 (2.59)

Table 1: bias $\times$ 100 (standard error $\times$ 100) of parameter estimation under different scenarios (repeated by us)

	LATE		MLATE	
Method	α_0	α_1	α_0	α_1
`mle.bth`	0.28 (0.35)	-3.5 (0.78)	-0.092 (0.71)	-3.0 (1.2)
`mle.bad`	-20 (0.42)	-15 (0.80)	-48 (1.2)	-18 (2.1)
`drw.bth`	0.55 (0.36)	-4.1 (0.82)	-0.54 (0.77)	-5.6 (1.5)
`drw.psc`	0.060 (0.38)	-5.9 (1.0)	-0.38 (1.2)	-12 (2.7)
`drw.opc`	0.55 (0.36)	-3.9 (0.79)	0.49 (0.75)	-5.3 (1.4)
`drw.bad`	-10 (0.40)	-9.6 (1.1)	-28 (1.4)	25 (3.3)
`dru.bth`	1.3 (0.44)	-5.8 (1.0)	1.8 (0.84)	-8.1 (1.7)
`reg.ogburn.bth`	-5.7 (1.6)	-2.9 (3.1)	1.8 (0.84)	-8.1 (1.7)
`reg.ogburn.bad`	-9.0 (0.25)	100 (0.23)	140 (5.6)	93 (3.6)
`drw.ogburn.bth`	0.10 (0.46)	-4.2 (0.99)	3.2 (1.4)	-13 (2.5)
`dru.ogburn.bth`	1.3 (0.45)	-5.8 (1.1)	1.9 (0.85)	-8.2 (1.7)
`dru.wang.bth`	1.3 (0.45)	-5.8 (1.1)	-	-
`ls.abadie.bth`	-0.19 (0.37)	-4.1 (0.93)	0.42 (0.79)	-11 (1.6)
`ls.abadie.bad`	-23 (0.88)	22 (1.2)	-32 (1.9)	7.7 (3.6)
`mle.crude`	-2.8 (0.10)	60 (0.19)	0.36 (0.25)	51 (0.42)

Table 2: bias $\times$ 100 (standard error $\times$ 100) of parameter estimation under different scenarios (reported by ARTICLE)

- `dru.wang`: the doubly robust estimator in Wang & Tchetgen Tchetgen (2018, ğ 4.4) with the identity weighting function;
- `ls.abadie`: the least squares estimator of Abadie (2003, ğ 4.2.1);
- `mle.crude`: the maximum likelihood estimator of the crude association on the additive or multiplicative scale (Richardson et al., 2017, ğ 2).

We also investigate the performance of these methods in misspecified scenarios. Consider two misspecified covariates: $X^\dagger$ includes an intercept and another covariate generated from Uniform$(-1, 1)$ independent of X; and X' includes $(\underbrace{1, \ldots, 1}_{0.5n}, \underbrace{0 \ldots, 0}_{0.5n})^{\mathrm{T}}$, $(\underbrace{0, \ldots, 0}_{0.1n}, \underbrace{1, \ldots, 1}_{0.9n})^{\mathrm{T}}$. We consider the following four misspecified scenarios, while the model of $\theta(X)$ is always correctly specified.

- `bth`: X is used in all nuisance models;
- `psc`: X is used in the instrumental density model, and X' is used in other nuisance models;
- `opc`: $X^\dagger$ is used in the instrumental density model, and X is used in other nuisance models;
- `bad`: $X^\dagger$ is used in the instrumental density model, and X' is used in other nuisance models.

Results repeated by us and reported by ARTICLE are presented in Table 1 and Table 2 respectively. Firstly, our results is different from ARTICLE, but they are roughly of the same magnitude. The

	LATE		MLATE	
Method	α_0	α_1	α_0	α_1
mle.bth	95.6	95.8	95.4	96.4
mle.bad	65.4	91.0	46.3	94.6
drw.bth	94.7	95.2	96.9	95.9
drw.psc	95.4	95.5	97.6	97.4
drw.opc	95.0	95.6	96.1	96.1
drw.bad	87.0	95.3	91.8	96.8
dru.bth	94.5	94.6	96.3	96.9
reg.ogburn.bth	98.0	99.9	99.6	100.0
reg.ogburn.bad	75.6	0.1	99.9	86.1
drw.ogburn.bth	97.0	98.1	98.5	98.4
dru.ogburn.bth	94.5	95.0	96.2	97.2
dru.wang.bth	94.4	94.7	-	-
dru.simple.bth	94.3	94.8	96.1	96.5
ls.abadie.bth	94.8	95.5	96.4	95.6
ls.abadie.bad	87.3	94.5	93.1	95.7
mle.crude	84.3	0.0	94.0	6.3

Table 3: Coverage probabilities $(\times 100)$ of confidence interval obtained from 500 bootstrap samples in selected scenarios (reported by ARTICLE)

reason for this difference is that we use different optimization methods and the covariates $X, X^\dagger$ are generated randomly. Secondly, in Table 2, the estimators `mle`, `drw`, `dru` generally have small bias and small standard error relative to other methods when all nuisance models are correctly specified, which vefiries the validity and efficiency of the method proposed by ARTICLE. Thirdly, in Table 2, we can also find that the performances of `dru.ogburn.bth`, `dru.wang.bth`, `dru.bth` are similar and less efficient than `drw.bth`, which implies that estimators using the optimal weighting function is more efficient than estimators using identity weighting function. At last, in Table 1, the two doubly robust estimators `drw` and `dru` performs worser than `mle` in the misspecification scenarios, which suggests the weighting function may yields less efficient estimations when model is misspecified. This phenomenon is not stated by ARTICLE, and they only partly reported their results.

Table 3 is the selected coverage probabilities of 95% confidence intervals obtained from the quantile bootstrap based on 500 bootstrap samples, reported by ARTICLE. We do not repeat the experiements of coverage probabilities since it cost too much time. Firstly, `mle` is less efficient in the misspecification scenarios when estimating the intercept α_0; in these scenarios, the doubly robust estimator with optimal weighting function `drw` seems better than `mle`. ARTICLE did not report the results of `dru` and did not explain why. Secondly, `mle`, `drw` and `dru` have higher converage probabilities than other methods, except for `reg.ogburn` and `drw.ogburn`. The authors of ARTICLE guessed it is due to misspecification of variation-dependent models.

2 Experiments on 401(k) Data

To explore the performances of estimators proposed by ARTICLE as well as other compared estimators, we use them to analyze the 401(k) data. According to Wikipedia, in the United States, a 401(k) plan is an employer-sponsored, defined-contribution, personal pension (savings) account, as defined in subsection 401(k) of the U.S. Internal Revenue Code. Collected by Employee Benefit Research Institute (EBRI) and the Investment Company Institute (ICI), 401(k) is the largest, most representative repository of information about individual 401(k) plan participant accounts. To be more specific, we aim to investigate whether the participation in 401(k) will reduce the paticipation in another retirement plan - Individual Retirement Accounts.

There may exists unobserved individual-level confounders such as the consuming willingness which influence the participation in 401(k) and Individual Retirement Accounts, thus directly estimating the treatment effect may yield invalid causal inferences. Therefore, we might wonder a instrumental variable irrelevant to the unobserved individual confounders. Following Ababie (2003), we choose eligibility for 401(k) as an instrument variable. Determined by companies, eligibility is weakly

relevant to the individual-level variables. The monotonicity and instrumental variable relevance assumptions holds since only those eligible for 401(k) may participate in it.

ARTICLE argued that eligibility for 401(k) has an impact on participation in Individual Retirement Accounts only through participation in 401(k) plans. However, it seems that this argument is problematic since the employees who is eligible for 401(k) is more likely to be eligible for Individual Retirement Accounts. In our review, we will set this problem aside, and repeat the real data studies in ARTICLE using the dataset provided by Abadie (2003). In the following, we will use '401(k) data' to refer to this dataset.

401(k) data consists of 9275 individuals from the Survey of Income and Program Participation of 1991. It has the following 11 variables:

- e401k; *=1 if eligble for 401(k).*
- inc; *annual income*.
- marr; *= 1 if married*.
- male; *=1 if male respondent* .
- age; *age of an individual, in years.*
- fsize; *family size.*
- nettfa; *net total fin. assets*.
- p401k; *=1 if participate in 401(k)* .
- pira; *=1 if have Individual Retirement Accounts* .
- incsq; inc^2.
- agesq; age^2.

Following the same setting as ARTICLE, we use the same instrumental variable model to estimate the multiplicative local average treatment effect of 401(k) participation on the participation in IRA. we use e401k as instrumental variable, p401k as treatment variable and pira as response variable. The covariate X include an intercept, inc, inc^2, age, marr and fsize. Since only those eligible for 401(k) will participate in it, i.e. $D(0) = 0$, there is no defiers and always takers. This implies the probability of an individual's being always taker conditional on his being always taker or never taker is zero,i.e. $\phi_2(X) \equiv 0$. In the parameterization of $\Pr(D, Y|Z)$, ϕ_4 is always multiplied by ϕ_2, thus the model of $\Pr(D, Y|Z)$ can be uniquely represented by $\delta^{\mathrm{M}}, \phi_1, \phi_3$ and ${}_{\mathrm{OP}}{}^{\mathrm{CO}}$.

The models for $\delta^{\mathrm{M}}, \phi_1, \phi_3$ and ${}_{\mathrm{OP}}{}^{\mathrm{CO}}$ are

$$\begin{cases} \delta^{\mathrm{M}}(X) = \exp(\alpha' X); \\ \phi_1(X) = \mathrm{sigmoid}(\beta_1' X); \\ \phi_3(X) = \mathrm{sigmoid}(\beta_2' X); \\ {}_{\mathrm{OP}}{}^{\mathrm{CO}}(X) = \exp(\eta^T X) \\ P(Z = 1|X) = \mathrm{sigmoid}(\gamma' X). \end{cases} \tag{2}$$

which are generally in log-linear family and logit-lieanr family. This two family is less flexible compared to machine learning models, such as energy based model. Future works may investigate the performances of these proposed methods with energy based model. We repeat the experiements of methods `mle` and `drw`, and cite the results of methods `drw.ogburn`, `dru.ogburn`, `ls.abadie` and `mle.crude` reported by ARTICLE. We construct 95% confidence intervals for coefficients in δ^{M} based on 500 bootstrap samples using different methods, in Figure 1 and Figure 2.

We find that the results repeat by us (Figure 1) are different from those reported by ARTICLE (Figure 2). Generally, C.I.s constructed by us is wider compared to those reported by ARTICLE, and the C.I.s based on `drw` is wider than `dru` in our results. We think this is due to differences in optimization methods, and in fact, we encountered some numerical problems in the experiments. To be specific, when ${}_{\mathrm{OP}}{}^{\mathrm{CO}}(X)$ is close to 1 or δ^{M} is close to 0, a value overflow will occur. We drop these cases in our experients, and this may introduce additional bias to the estimates.

In Figure 2, we find that `drw` and `mle` proposed by ARTICLE yield narrower C.I.s than other methods, when estimating all these six coefficients. Comparison between `drw` and `dru.ogburn` suggests the optimal weighting function is useful for reducing the variability of estimates. All the estimates

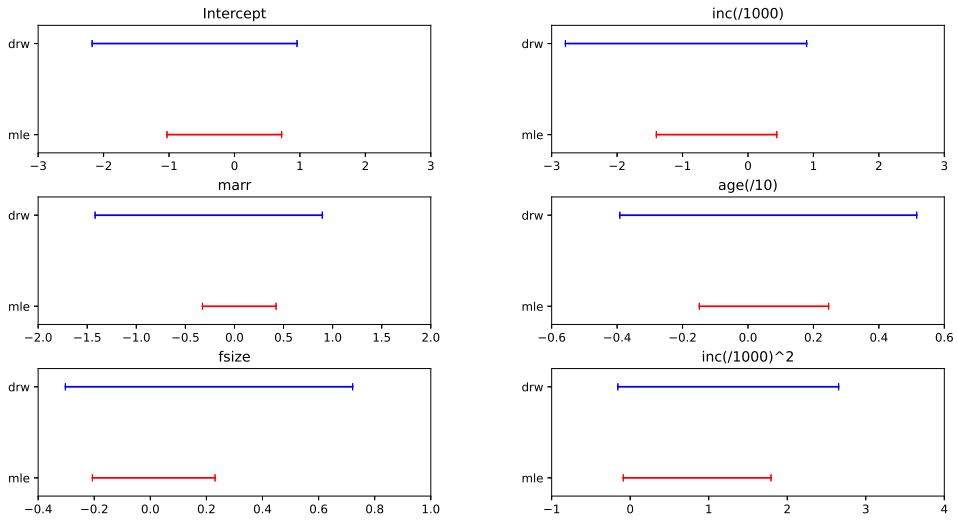

Figure 1: 95% Confidence intervals of estimates of coefficients in δ^{M} using different methods (repeated by us)

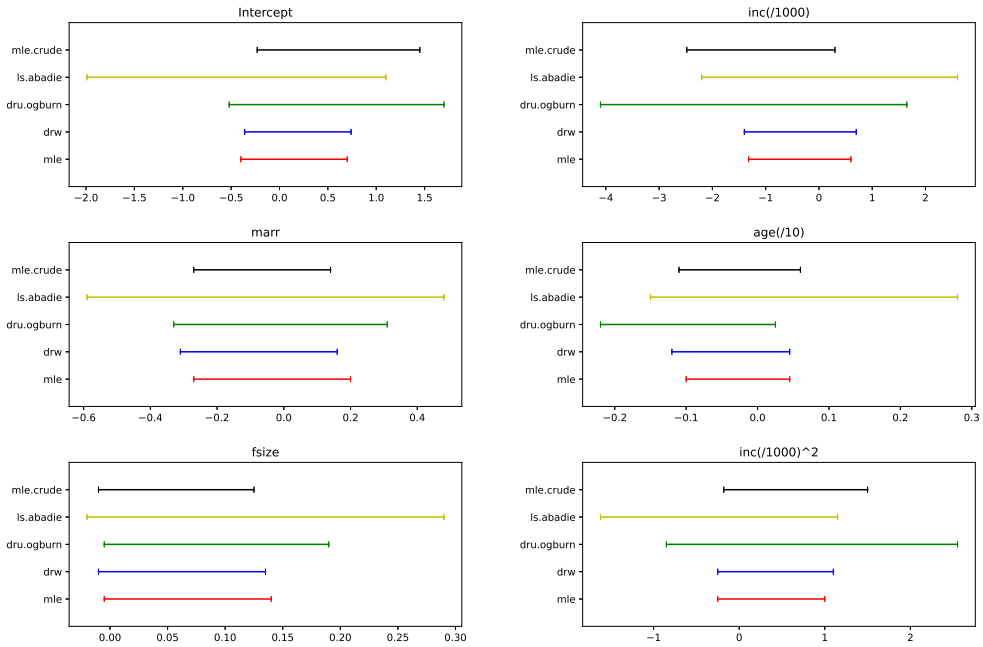

Figure 2: 95% Confidence intervals of estimates of coefficients in δ^{M} using different methods (reported by ARTICLE)

have direct interpretation, for example, the point estimation of the coefficient of marr (=-0.04, using `mle`) suggest for the married, the multiplicative effect of p401k on pira is -0.039 lesser than that for the unmarried.

At last, we cite the C.I.s for estimated MLATE for a typical married individual and a typical single individual using different (reported by ARTICLE) in Figure 3. The typical individuals are constructed by the median of individual covariates in each subgroup. For 401(k) data, (a) the typical married individual were 40 years old, had an annual income of $40530 and a family of size four; (b) the typical single individual were 39 years old, had an annual income of $23718 and had no other family members. The estimates of MLATE has a direct interpretation, for example, the point estimation of MLATE is 1.12 using `drw` suggests that for the typical married individual, participation in 401(k)

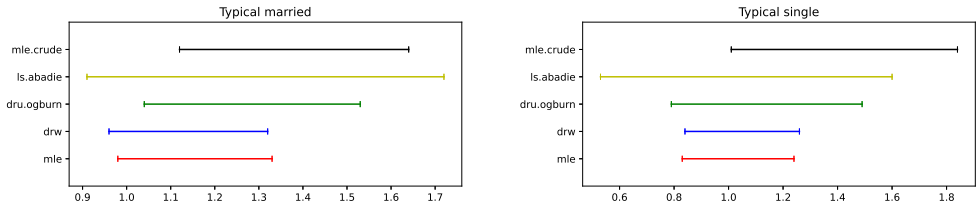

Figure 3: 95% Confidence intervals of estimated MLATE for a typical married individual and a typical single individual using different methods (reported by ARTICLE)

will increase the probability of participation in IRA by around 12%. Figure 3 also suggests that for the typical single individual, participation in 401(k) won't increase the probability of participation in IRA significantly. The last observation is, the two methods proposed by ARTICLE, `mle` and `drw`, yield narrower C.I.s than other methods thus have smaller variances.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Wang, L., Zhang, Y., Richardson, T. S., & Robins, J. M. (2021). Estimation of local treatment effects under the binary instrumental variable model. *Biometrika*, 108(4), 881-894.

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.