A Brief Introduction to Doubly Robust Methods

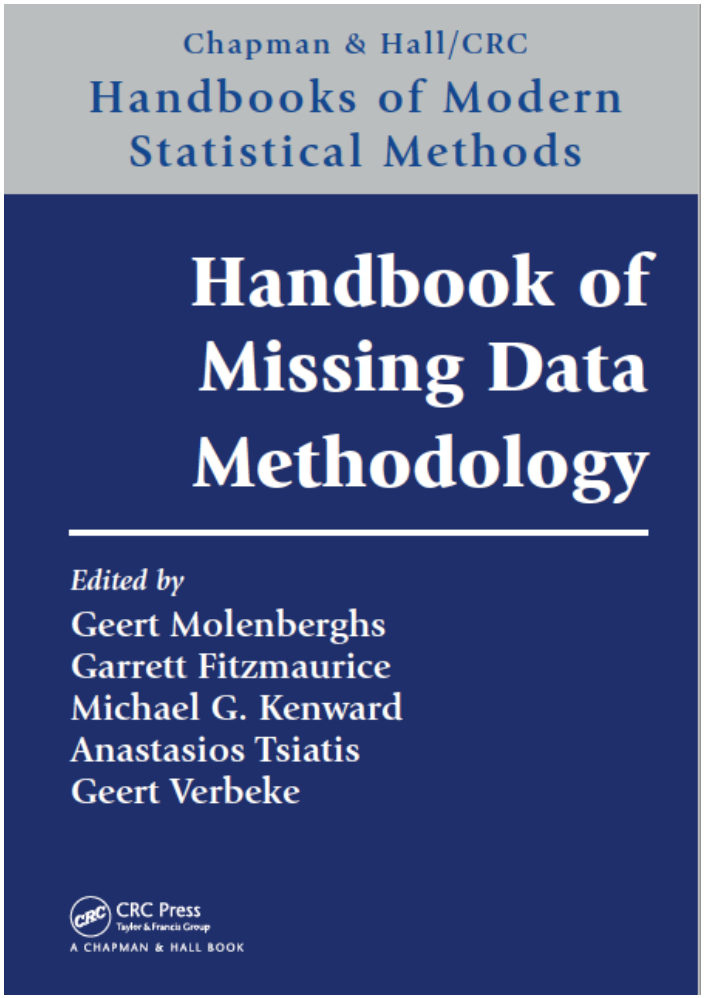
Peng Wu

Beijing International Center for Mathematical Research,
Peking University

March 20, 2022

Main references

Seaman, S. R. & Vansteelandt, S. (2018), 'Introduction to double robust methods for incomplete data', Statistical Science 33, 184-197.



Our two works:

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, Xiao-Hua Zhou (2021), "Causal Analysis Framework for Recommendation", arXiv:2201.06716.

Peng Wu, Haoxuan Li, Yan Lyu & Xiao-Hua Zhou (2022), 'Doubly Robust Collaborative Targeted Learning for Recommendation on Data Missing Not at Random', arXiv.

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

Setup

Data: $\{(Y_i, T_i, X_i) : i = 1, ..., n\}$ is an i.i.d. sample,

- Y: outcome variable;
- $T \in \{0, 1\}$: binary treatment;
- X: covariates.
- (Y^0, Y^1): potential outcomes.

Causal parameter: For brevity, we focus on

$$\mu = \mathbb{E}(Y^1).$$

This is a basic scenario of missing outcome data.

T_i	X_i	Y_i	Y_i^1	Y_i^0
1	✓	✓	✓	
1	✓	✓	✓	
1	✓	✓	✓	
0	✓	✓		✓
0	✓	✓		✓
0	✓	✓		✓

Identifiability

Unconoundedness assumption: $T \perp Y^1|X$. Define

$$\pi(X) = \mathbb{P}(T = 1|X), \quad m(X) = \mathbb{E}(Y|X, T = 1).$$

Both of them can be estimated based on observed data.

- **IPW**: inverse probability weighting

$$\mathbb{E}[Y^1] = \mathbb{E}[\frac{TY}{\pi(X)}] \tag{1}$$

- **OR**: outcome regression

$$\mathbb{E}[Y^1] = \mathbb{E}_X\left[\mathbb{E}(Y|X, T = 1)\right] = \mathbb{E}_X[m(X)]. \tag{2}$$

- **AIPW**: augmented inverse probability weighting

$$\mathbb{E}[Y^1] = \mathbb{E}\Big[\frac{T}{\pi(X)}Y + (1 - \frac{T}{\pi(X)})m(X).\Big] \tag{3}$$

Proof of (1), (2) and (3):

Naive estimator

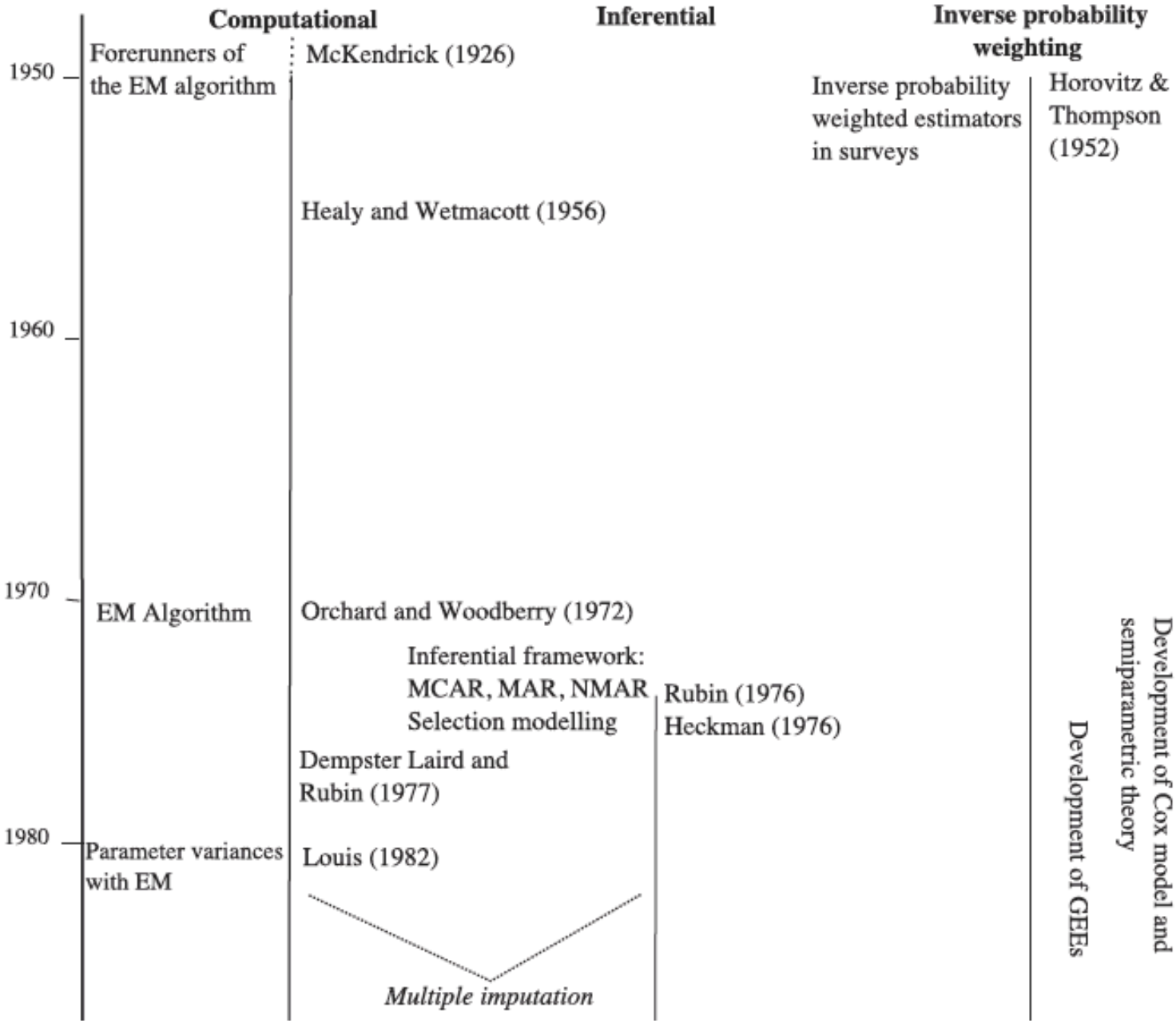
A natural estimator for μ, using the observed data, is the complete-case sample ($T = 1$) average, namely

$$\hat{\mu}_{complete} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} \xrightarrow{P} \mathbb{E}[Y|T=1] = \mathbb{E}[Y^1|T=1] \tag{4}$$

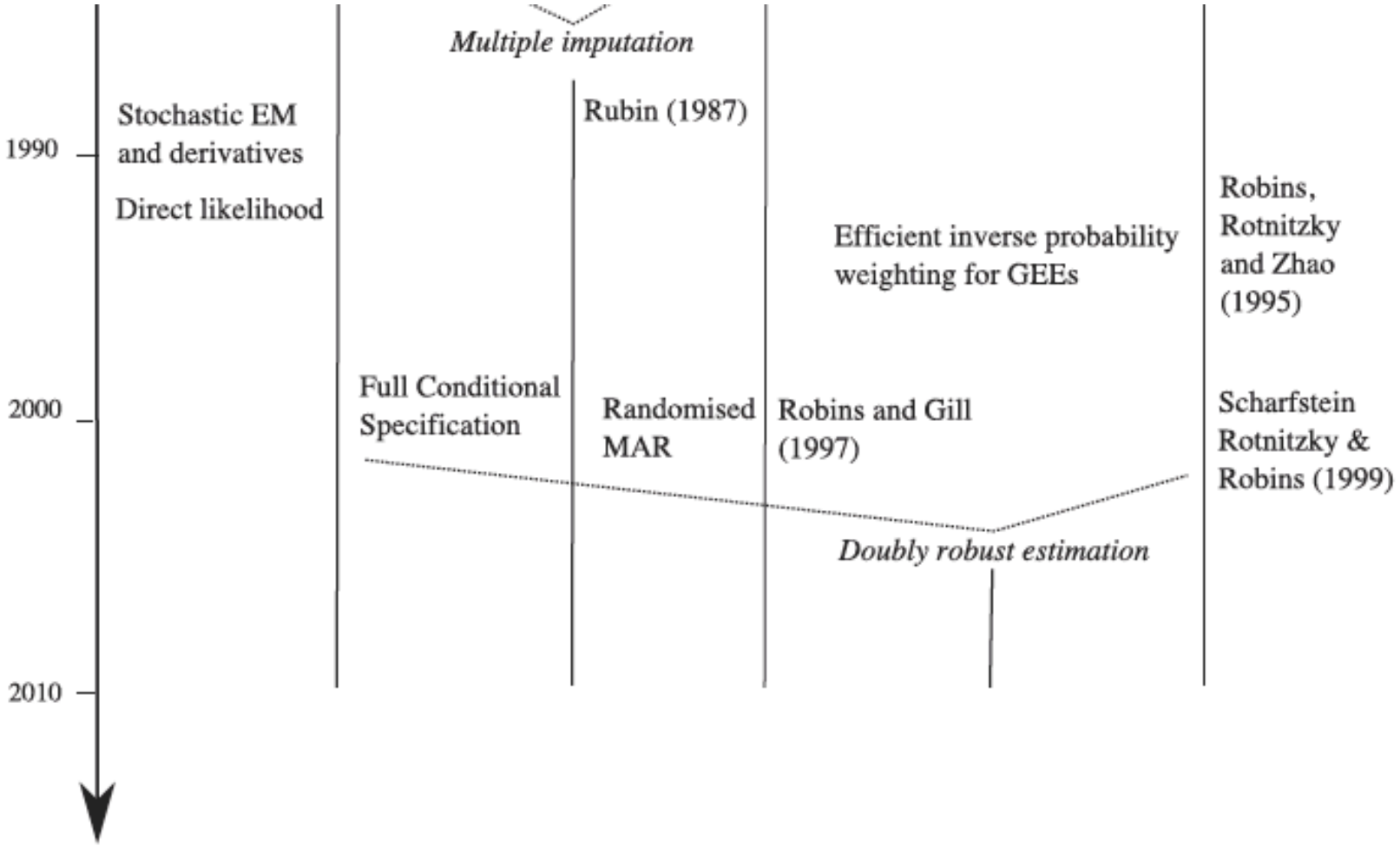
- For RCT, $\mu = \mathbb{E}[Y^1|T=1]$;
- For OBS, $\mu \neq \mathbb{E}[Y^1|T=1]$.

Methods for handling missing data

- Likelihood and Bayesian methods.
- Multiple imputation methods.
- Weighting methods.


Computational
Inferential
Inverse probability weighting
1950
1960
1970
1980
Forerunners of the EM algorithm
McKendrick (1926)
Inverse probability weighted estimators in surveys
Horovitz & Thompson (1952)
Healy and Wetmacott (1956)
EM Algorithm
Orchard and Woodberry (1972)
Inferential framework: MCAR, MAR, NMAR Selection modelling
Rubin (1976)
Heckman (1976)
Dempster Laird and Rubin (1977)
Parameter variances with EM
Louis (1982)
Multiple imputation
Development of GEEs
Development of Cox model and semiparametric theory

Brief history



Weighting methods will be the primary focus in this slides.

IPW

Let $\hat{\pi}(X_i)$ be the estimate of $\pi(X_i)$.

- **IPW:**

$$\hat{\mu}_{ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)},$$

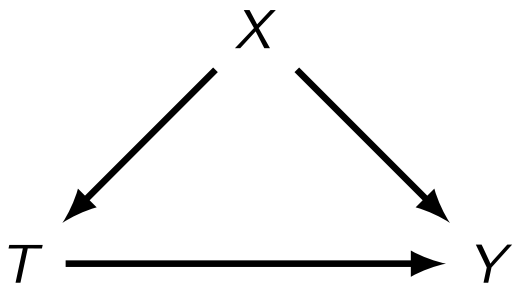
- **Normalized IPW:**

$$\hat{\mu}_{nipw} = \sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} \Big/ \sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}$$

Remarks:

- The IPW estimators is unstable in the presence of small propensities.

OR



IPW pay no heed to relationships between X and Y. Outcome regression (OR) model Y from X directly to predict the missing values.

$$\hat{\mu}_{or} = n^{-1} \sum_{i=1}^{n} \hat{m}(X_i). \tag{5}$$

or

$$\hat{\mu}_{or} = n^{-1} \sum_{i=1}^{n} \{ T_i Y_i + (1 - T_i) \hat{m}(X_i) \}. \tag{6}$$

OR

Remarks.

- OR suffers the problem of implicitly making extrapolation, in which OR builds model within complete cases, while using the predicted values to construct the estimator of μ. Extrapolation is dangerous, and its good performance is partly a matter of luck.

	stability	without extrapolation	efficiency
IPW	$\times$	✓	$\times$
OR	✓	$\times$	✓

- 倾向性得分模型和回归模型, 哪一个更容易指定正确?
- 当两个模型都指定错误时, 哪一个会产生更糟糕的估计?

Tan, Z. (2007), 'Comment: Understanding OR, PS and DR', Statistical Science, 22 (4), 560–568.

AIPW

IPW estimator formally introduced in 1952. Nevertheless, for many years, IPW gained little acceptance in the missing data literature because of its imprecision and instability in the presence of extreme weights ($\pi(X)$ close to 0) relative to OR.

This has changed drastically over the past two decade, since the seminal work of Robins et al. (1994), who demonstrated how the precision of IPW estimator could be greatly improved.

$$\hat{\mu}_{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(X_i)} + \frac{1}{n}\sum_{i=1}^{n}\{1 - \frac{T_i}{\hat{\pi}(X_i)}\}\hat{m}(X_i) \tag{7}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\{Y_i - \hat{m}(X_i)\}, \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i^1 + \frac{1}{n}\sum_{i=1}^{n}\frac{\{T_i - \hat{\pi}(X_i)\}}{\hat{\pi}(X_i)}\{Y_i - \hat{m}(X_i)\} \tag{9}$$

where

- $n^{-1}\sum_{i=1}^{n}\{1 - \frac{T_i}{\hat{\pi}(X_i)}\}\hat{m}(X_i)$ is called the augmentation term. It improves the efficiency of IPW estimator.
- $n^{-1}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\{Y_i - \hat{m}(X_i)\}$ is called the correction term. It use IPW to estimate how much $\hat{\mu}_{or}$ overestimates (or underestimates) μ and then subtracts this.

Proposition 1. The following statements hold:
(a). $\hat{\mu}_{aipw}$ is doubly robust.
(b). $\hat{\mu}_{aipw}$ achieves the semiparametric variance bound.
(c). If the OR model $m(X;\beta)$ of $m(X)$ is correctly specified and β is efficiently estimated, then

$$asy.var(\hat{\mu}_{or}) \leq asy.var(\hat{\mu}_{aipw}) \leq asy.var(\hat{\mu}_{ipw})$$

where $asy.var$ denotes asymptotic variance.

□

It can be seen from proposition 1 that

- Compared with $\hat{\mu}_{ipw}$,

  if PS model is misspecified but OR model is correctly specified, $\hat{\mu}_{aipw}$ tends to have a smaller bias.

  if both the models are correctly specified, $\hat{\mu}_{aipw}$ is more efficient.

- Compared with $\hat{\mu}_{or}$,

  if OR model is misspecified but PS model is correctly specified, $\hat{\mu}_{aipw}$ tends to have a smaller bias.

  if both modes are correctly specified, $\hat{\mu}_{or}$ is more efficient.

  Thus, it involves the bias-variance trade-off.

AIPW

There are some issues with AIPW estimator:

- $\hat{\mu}_{aipw}$ may lie outside its parameter space (e.g. outside $[0, 1]$ when Y is binary). Even when guaranteed to lie within its parameter space, it may not be within the range of the observed Y values.
- when OR model is misspecified, there is no guarantee that $\hat{\mu}_{aipw}$ will be at least as efficient as the $\hat{\mu}_{ipw}$.
- Poor finite-sample behavior when PS is close to zero for some units;

In practical applications, both models PS and OR are likely to be at least mildly misspecified.

Comparisons of IPW, OR and AIPW

	stability	without extrapolation	doubly robust	efficiency	boundedness
IPW	×	✓	×	×	×
NIPW	*o*	✓	×	×	✓
OR	✓	×	×	✓	✓
AIPW	*o*	*o*	✓	*o*	×

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

Basic idea

The construction of a DR outcome regression estimator is based on regression imputations $\hat{m}(X)$ computed in a manner that ensures that

$$n^{-1}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\{Y_i-\hat{m}(X_i)\}=0. \tag{10}$$

and then use

$$\hat{\mu}_{dr.or}=n^{-1}\sum_{i=1}^{n}\hat{m}(X_i)$$

as a resulting estimator of μ.

As the left-hand side of (10) is the "correction term" in (8), this ensures that $\hat{\beta}_{dr.or}$ reduces to a OR estimator and consequently doubly robust.

Example

Scharfstein et al. (1999) considered a generalized model with

$$m(X;\beta) = \varphi\{\beta^T v(X)\},$$

where $v(X)$ is some known vector vector functions of X, φ is a known link function. The authors fit β with a extended model,

$$m(X;\beta) = \varphi\{\beta^T v(X) + \eta \frac{1}{\hat{\pi}(X)}\} \tag{11}$$

The corresponding estimating equation is

$$(\hat{\beta},\hat{\eta}) : \sum_{i=1}^{n} T_i \cdot \left\{ Y_i - \varphi\{\beta^T v(X) + \eta/\hat{\pi}(X)\} \right\} \cdot \begin{pmatrix} v(X) \\ 1/\hat{\pi}(X) \end{pmatrix} = 0,$$

which implies restriction (10). The resulting estimator is given as

$$\hat{\mu}_{dr.or} = n^{-1} \sum_{i=1}^{n} \varphi\{\hat{\beta}^T v(X_i) + \hat{\eta} \frac{1}{\hat{\pi}(X_i)}\}$$

Advantages

	stability	without extrapolation	doubly robust	efficiency	boundedness
IPW	×	✓	×	×	×
NIPW	*o*	✓	×	×	✓
OR	✓	×	×	✓	✓
AIPW	*o*	*o*	✓	*o*	×
DR.OR	✓	*o*	✓	✓	✓

References

More examples can be found in the following references:

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

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

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, 523–539.

Mark J. van der Laan & Sherri Rose (2011). "Targeted Learning: Causal Inference for Observational and Experimental Data". Springer.

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

Basic idea

Let $\hat{\mu}_{or} = n^{-1}\sum_{i=1}^{n}\hat{m}(X_i)$ is the OR estimator of μ. To achieve DR property, the estimator $\hat{\pi}(X)$ is computed in a manner that ensures that

$$n^{-1}\sum_{i=1}^{n}\{1-\frac{T_i}{\hat{\pi}(X_i)}\}\{\hat{m}(X_i)-\hat{\mu}_{or}\}=0 \tag{12}$$

and then use

$$\hat{\mu}_{dr.ipw} = \sum_{i=1}^{n}\frac{T_iY_i}{\hat{\pi}(X_i)}\Big/\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}$$

as an estimator of μ.

Advantages

	stability	without extrapolation	doubly robust	efficiency	boundedness
IPW	×	✓	×	×	×
NIPW	*o*	✓	×	×	✓
OR	✓	×	×	✓	✓
AIPW	*o*	*o*	✓	*o*	×
DR.OR	✓	*o*	✓	✓	✓
DR.IPW	*o*	✓	✓	*o*	✓

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

More references

For more extended doubly robust methods, interested readers can refer to the following literature:

Tan, Z. (2006a), 'A distributional approach for causal inference using propensity scores', Journal of the American Statistical Association 101(476), 1619–1637.

Tan, Z. (2006b), "Regresion and weighting methods for causal inference using instrumental variables", Journal of the American Statistical Association 101, 1607–1618.

Tan, Z. (2010), "Bounded, efficient, and doubly robust estimation with inverse weighting", Biometric 92(2), 1–22.

Tan, Z. (2011), "Efficient restricted estimators for conditional mean models with missing data", Biometrika 98, 663–684.

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

In this section, we introduce some latest doubly robust methods.

Vermeulen, K. & Vansteelandt, S. (2015), 'Bias-reduced doubly robust estimation', Journal of the American Statistical Association 110, 1024–1036.

Vermeulen, K. & Vansteelandt, S. (2016), 'Data-adaptive bias-reduced doubly robust estimation', Int. J. Biostat. 12, 253–282.

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

Peng Wu, Zhiqiang Tan, Wenjie Hu & Xiao-Hua Zhou (2021), 'Model-assisted inference for covariate-Specific treatment effects with high-dimensional data', arXiv:2105.11362.

Basic idea

Rather than seeking directly to minimize the asymptotic variance, "Bias-Reduced DR estimator" uses the estimators $\hat{\alpha}$ and $\hat{\beta}$ obtained by locally minimizing the squared asymptotic bias of $\hat{\mu}_{aipw}$ when both models $\pi(X;\alpha)$ and $m(X;\beta)$ are misspecified. This makes the bias-reduced DR estimator less sensitive than the standard DR estimator to mild model misspecification.

Example

Consider a linear OR model and and a logistic PS mode,

$$m^*(X) := \mathbb{E}[Y|T=1,X] = m(X;\beta) = \beta^T X, \tag{13}$$

$$\pi^*(X) = \mathbb{P}(T=1 \mid X) = \pi(X;\alpha) = [1+\exp\{-\alpha^T X\}]^{-1}, \tag{14}$$

Assume α and β are solved by minimizing

$$L_{or}(\alpha,\beta) = \tilde{E}_n[T\frac{1-\pi(X;\alpha)}{\pi(X;\alpha)}\{Y-\beta^T X\}^2] \tag{15}$$

$$L_{ps}(\alpha) = \tilde{E}_n[T\exp\{-\alpha^T X\} + (1-T)\alpha^T X], \tag{16}$$

where $\tilde{E}_n(X) = n^{-1}\sum_{i=1}^n X_i$.

Example

Key point 1: Define

$$\bar{\alpha} = \arg\min_{\alpha} \mathbb{E}[L_{ps}(\alpha)], \quad \bar{\beta} = \arg\min_{(\beta)} \mathbb{E}[L_{or}(\bar{\alpha}, \beta)].$$

- If model (15) is correctly specified, then $\pi(X;\bar{\alpha}) = \pi^*(X)$; $\pi(X;\bar{\alpha}) \neq \pi^*(X)$ otherwise.
- If model (14) is correctly specified, then $m(X;\bar{\beta}) = m^*(X)$; $m(X;\bar{\beta}) \neq m^*(X)$ otherwise.

However, whether PS/OR model is correctly specified or not,

$$\hat{\alpha} \to \bar{\alpha}, \quad \hat{\beta} \to \bar{\beta}.$$

Example

Let

$$\varphi(\alpha,\beta)=\frac{TY}{\pi(X;\alpha)}-\frac{T-\pi(X;\alpha)}{\pi(X,\alpha)}m(X,\beta).$$

$$\hat{\mu}_{aipw}=\tilde{E}_n[\varphi(\hat{\alpha},\hat{\beta})]$$

Applying Taylor expansion leads to that

$$\hat{\mu}_{aipw}=\tilde{E}_n[\varphi(\bar{\alpha},\bar{\beta})]+(\hat{\alpha}-\bar{\alpha})^T\Delta_1+(\hat{\beta}-\bar{\beta})^T\Delta_2+o_p(n^{-1/2}),$$

$$n^{1/2}(\hat{\mu}_{aipw}-\mu)=n^{1/2}\tilde{E}_n[\varphi(\bar{\alpha},\bar{\beta})-\mu]+n^{1/2}(\hat{\alpha}-\bar{\alpha})^T\Delta_1+n^{1/2}(\hat{\beta}-\bar{\beta})^T\Delta_2+o_p(1),$$

where $\Delta_1=\frac{\partial}{\partial\alpha}\tilde{E}_n[\varphi(\bar{\alpha},\bar{\beta})]$, $\Delta_2=\frac{\partial}{\partial\beta}\tilde{E}_n[\varphi(\bar{\alpha},\bar{\beta})]$. Under models (14) and (15),

$$\Delta_1=\tilde{E}_n\{(1-\frac{T}{\pi(X;\bar{\alpha})})X\},$$

$$\Delta_2=\tilde{E}_n\{T\frac{1-\pi(X;\bar{\alpha})}{\pi(X;\bar{\alpha})}(Y-\bar{\beta}^TX)X\},$$

Example

Key point 2: Under loss functions (16) and (17), setting $\partial L_{ps}(\alpha)/\partial\alpha = 0$ and $\partial L_{or}(\hat{\alpha}, \beta)/\partial\beta = 0$ yield that

$$\tilde{E}_n\{(1 - \frac{T}{\pi(X;\hat{\alpha})})X\} = 0,$$

$$\tilde{E}_n\{T\frac{1-\pi(X;\hat{\alpha})}{\pi(X;\hat{\alpha})}(Y - \hat{\beta}^T X)X\} = 0.$$

General framework

Without specifications of (14), (15), (16), (17).

Let $Z = (T, X, Y)$, $U_\alpha(Z;\alpha)$ and $U_\beta(Z;\beta)$ are the score functions for α and β, namely, $\hat{\alpha}$ and $\hat{\beta}$ are solutions to estimating equations

$$\tilde{E}_n\{U_\alpha(Z;\hat{\alpha})\} = 0, \qquad \tilde{E}_n\{U_\beta(Z;\hat{\beta})\} = 0$$

General framework

Recall that the AIPW estimator is given as

$$\hat{\mu}_{aipw}(\hat{\alpha}, \hat{\beta}) = \tilde{E}_n\{\frac{TY}{\pi(X;\hat{\alpha})} - \frac{T - \pi(X;\hat{\alpha})}{\pi(X,\hat{\alpha})} m(X,\hat{\beta})\},$$

The score function of $\hat{\mu}_{aipw}(\hat{\alpha}, \hat{\beta})$ is

$$\phi(Z;\mu,\alpha,\beta) = \frac{TY}{\pi(X;\alpha)} - \frac{T - \pi(X;\alpha)}{\pi(X,\alpha)} m(X,\beta) - \mu.$$

Denote $\bar{\alpha} = \text{plim}(\hat{\alpha})$, $\bar{\beta} = \text{plim}(\hat{\beta})$, the probability limits of $\hat{\alpha}$ and $\hat{\beta}$.

- if $\pi(X;\alpha)$ is correctly specified, i.e., $\pi(X) = \pi(X;\alpha_0)$, then $\bar{\alpha} = \alpha_0$; $\bar{\alpha} \neq \alpha_0$ otherwise.
- Similarly, $\bar{\beta} = \beta_0$ if $m(X;\beta)$ is correctly specified; $\bar{\beta} \neq \beta_0$ otherwise.

General framework

Proposition 2. Under suitable regularity conditions, a first-order asymptotic expansion of $\hat{\mu}_{aipw}(\hat{\alpha}, \hat{\beta})$ is given by

$$n^{1/2}\{\hat{\mu}_{aipw}(\hat{\alpha}, \hat{\beta}) - \mu\} = n^{1/2}\tilde{E}_n\{\tilde{\phi}(Z; \mu, \bar{\alpha}, \bar{\beta})\} + o_p(1), \tag{17}$$

where

$$\begin{aligned}
\tilde{\phi}(Z; \mu, \bar{\alpha}, \bar{\beta}) &= \phi(Z; \mu, \bar{\alpha}, \bar{\beta}) \\
&- \mathbb{E}\left(\frac{\partial \phi}{\partial \alpha}\right)\left\{ I_p - \mathbb{E}^{-1}\left(\frac{\partial U_\alpha}{\partial \alpha}\right)\mathbb{E}\left(\frac{\partial U_\alpha}{\partial \beta}\right)\mathbb{E}^{-1}\left(\frac{\partial U_\beta}{\partial \beta}\right)\frac{\partial U_\beta}{\partial \alpha}\right\}^{-1} \\
&\times\left\{\mathbb{E}^{-1}\left(\frac{\partial U_\alpha}{\partial \alpha}\right)U_\alpha(Z; \bar{\alpha}, \bar{\beta}) - \mathbb{E}^{-1}\left(\frac{\partial U_\alpha}{\partial \alpha}\right)\mathbb{E}\left(\frac{\partial U_\alpha}{\partial \beta}\right)\mathbb{E}^{-1}\left(\frac{\partial U_\beta}{\partial \beta}\right)U_\beta(Z; \bar{\alpha}, \bar{\beta})\right\} \\
&- \mathbb{E}\left(\frac{\partial \phi}{\partial \beta}\right)\left\{ I_q - \mathbb{E}^{-1}\left(\frac{\partial U_\beta}{\partial \beta}\right)\mathbb{E}\left(\frac{\partial U_\beta}{\partial \alpha}\right)\mathbb{E}^{-1}\left(\frac{\partial U_\alpha}{\partial \alpha}\right)\frac{\partial U_\alpha}{\partial \beta}\right\}^{-1} \\
&\times\left\{\mathbb{E}^{-1}\left(\frac{\partial U_\beta}{\partial \beta}\right)U_\beta(Z; \bar{\alpha}, \bar{\beta}) - \mathbb{E}^{-1}\left(\frac{\partial U_\beta}{\partial \beta}\right)\mathbb{E}\left(\frac{\partial U_\beta}{\partial \alpha}\right)\mathbb{E}^{-1}\left(\frac{\partial U_\alpha}{\partial \alpha}\right)U_\alpha(Z; \bar{\alpha}, \bar{\beta})\right\}
\end{aligned}$$

where all gradients are evaluated at $(Z; \mu, \bar{\alpha}, \bar{\beta})$.

General framework

Specifically, if

$$\partial U_\alpha(Z;\alpha)/\partial\beta \equiv 0, \quad \partial U_\beta(Z;\beta)/\partial\alpha \equiv 0$$

$$\tilde{\phi}(Z;\mu,\bar{\alpha},\bar{\beta}) = \phi(Z;\mu,\bar{\alpha},\bar{\beta}) - \mathbb{E}(\frac{\partial\phi}{\partial\alpha})\mathbb{E}^{-1}(\frac{\partial U_\alpha}{\partial\alpha})U_\alpha(Z;\bar{\alpha}) - \mathbb{E}(\frac{\partial\phi}{\partial\beta})\mathbb{E}^{-1}(\frac{\partial U_\beta}{\partial\beta})U_\beta(Z;\bar{\beta})$$

□

General framework

With $\pi_\alpha(X;\alpha) = \partial\pi(X;\alpha)/\partial\alpha$ and $m_\beta(X;\beta) = \partial\pi(X;\beta)/\partial\beta$, we obtain that

$$E\{\frac{\phi(Z;\mu,\bar{\alpha},\bar{\beta})}{\partial\alpha}\} = E\Big[\frac{\pi^*(X)}{\pi^2(X;\bar{\alpha})}\{m(X;\bar{\beta}) - m^*(X)\}\pi_\alpha(X;\bar{\alpha})\Big]$$

$$E\{\frac{\phi(Z;\mu,\bar{\alpha},\bar{\beta})}{\partial\beta}\} = E\Big[\Big\{1 - \frac{\pi^*(X)}{\pi(X;\bar{\alpha})}\Big\}m_\beta(X;\bar{\beta})\Big],$$

regardless of whether the models are specified correctly. Both of them are zero if $\pi^*(X) = \pi(X;\bar{\alpha})$ and $m^*(X) = m(X;\bar{\beta})$.

When both working model are misspecified, the first-order asymptotic bias is

$$\text{bias}(\alpha,\beta;\mu) = \mathbb{E}[\phi(Z;\mu,\alpha,\beta)].$$

Suppose there exists a vector $(\bar{\alpha}_{BR},\bar{\beta}_{BR})$ such that

$$\mathbb{E}\{\partial\phi(Z;\mu,\bar{\alpha}_{BR},\bar{\beta}_{BR})/\partial\alpha\} = 0$$

$$\mathbb{E}\{\partial\phi(Z;\mu,\bar{\alpha}_{BR},\bar{\beta}_{BR})/\partial\beta\} = 0$$

General framework

The following theorem then shows that $(\bar{\alpha}_{BR}, \bar{\beta}_{BR})$ locally minimizes the squared first-order asymptotic bias.

Theorem 1. Under suitable regularity conditions, $(\bar{\alpha}_{BR}, \bar{\beta}_{BR})$ locally minimizes

$$\text{bias}^2(\alpha, \beta; \mu).$$

Proof.

$$\begin{aligned}\frac{\partial \text{bias}^2(\alpha, \beta; \mu)}{\partial \alpha} &= 2\text{bias}(\alpha, \beta; \mu)\frac{\partial \text{bias}(\alpha, \beta; \mu)}{\partial \alpha} \\ &= 2\text{bias}(\alpha, \beta; \mu)\mathbb{E}[\frac{\partial \phi(Z; \mu, \alpha, \beta)}{\partial \alpha}]\end{aligned}$$

and likewise for β.

□

General framework

$(\bar{\alpha}_{BR}, \bar{\beta}_{BR})$ need to be estimated. Define the estimators $\hat{\alpha}_{BR}$ and $\hat{\beta}_{BR}$ as the solutions to the estimating equations

$$\tilde{E}_n\{\partial\phi(Z;\mu,\hat{\alpha}_{BR},\hat{\beta}_{BR})/\partial\alpha\} = 0 \tag{18}$$

$$\tilde{E}_n\{\partial\phi(Z;\mu,\hat{\alpha}_{BR},\hat{\beta}_{BR})/\partial\beta\} = 0. \tag{19}$$

When the gradient of $\phi(Z;\mu,\hat{\alpha}_{BR},\hat{\beta}_{BR})$ with respect to α and β depends on the unknown population value μ, a preliminary consistent DR estimator $\hat{\mu}_{dr}^{prel}$ is substituted for μ. If $\text{plim}\hat{\mu}_{dr}^{prel} = \mu^* \neq \mu$, the values $(\hat{\alpha}_{BR}, \hat{\beta}_{BR})$ no longer minimize $\text{bias}^2(\alpha,\beta;\mu)$ but instead minimize $\text{bias}^2(\alpha,\beta;\mu^*)$.

Adaptive Bias-Reduced DR estimator

A limitation of the proposed approach is that it demands working models of the same dimension because the gradient of $\phi(\alpha, \beta)$ with respect to β is used as an estimating function for α and vice versa. This can be overcome by targeting asymptotic bias reduction in the direction of a single nuisance parameter, for instance in the direction of α. This is the method of "Data-adaptive Bias-Reduced DR estimator".

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

Setup

AIPW estimator can be incorporated in a more general frameworks. Assume that random variables $Z_1, ..., Z_n$ are i.i.d. random variables. Suppose θ is the solution to estimating equations of the form

$$\sum_{i=1}^{n} U(Z_i; \theta) = 0,$$

where $E[U(Z; \theta_0)] = 0$, θ_0 is the true value of θ. the observed data are:

$$\{Z_i^{(1)}, T_i, Z_i^{(2)} T_i\}, i = 1, ..., n.$$

Target: estimate θ_0.

Naive estimator of θ

$$\hat{\theta}_{naive}: \quad \sum_{i=1}^{n} T_i U(Z_i;\theta) = 0. \tag{20}$$

Is $\hat{\theta}_{naive}$ unbiased?

- MCAR, it is unbiased.
- MAR, it depends on $U(Z,\theta)$. Assume $Z = (X, Y^1)$, Y^1 is missing when $T = 0$, then the joint distribution of (T, Y^1) conditional on X is

$$\begin{aligned}\mathbb{P}(Y^1, R \mid X) &= \mathbb{P}(T \mid Y^1, X) \cdot \mathbb{P}(Y^1 \mid X, \theta) \\ &= \mathbb{P}(T \mid Y^1, X) \cdot \mathbb{P}(Y \mid X, T = 1, \theta)\end{aligned}$$

IPW estimator of θ

$\mathbb{P}(T=1|Z) = \mathbb{P}(T=1|Z^{(1)}) \equiv \pi(Z^{(1)})$. $\pi(Z^{(1)};\alpha)$ is specified model for $\pi(Z^{(1)})$. Then IPW estimating equation is given as

$$\hat{\theta}_{ipw}: \quad \sum_{i=1}^{n} \frac{T_i}{\pi(Z_i^{(1)};\hat{\alpha})} U(Z_i;\theta) = 0. \tag{21}$$

Why is it unbiased?

Construction of DR regression coefficients

Then AIPW estimating equation is given as

$$\hat{\theta}_{aipw}: \quad \sum_{i=1}^{n} \frac{T_i}{\pi(Z_i^{(1)};\hat{\alpha})} U(Z_i;\theta) + \{1 - \frac{T_i}{\pi(Z_i^{(1)};\hat{\alpha})}\}\phi(Z_i^{(1)};\theta) = 0. \qquad (22)$$

Semiparametric theory shows the optimally efficient choice of $\phi(Z_i^{(1)};\theta)$ is

$$\phi_{opt}(Z_i^{(1)};\theta) = E\{U(Z;\theta) \mid Z^{(1)}, T = 1\}$$

Example

Example 1. For estimating $\mu = \mathbb{E}[Y]$, $Z = (X, Y)$. $U(Z_i; \theta) = Y_i - \theta$; $\phi(Z_i^{(1)}; \theta) = \mathbb{E}[Y|X, T = 1] - \theta$.

Example 2. $Z = (X, Y)$, $U(Z_i; \theta) = X(Y - X^T\theta)$;

$$\begin{aligned}\phi(Z_i^{(1)}; \theta) &= \mathbb{E}[X(Y - X^T\theta) \mid X, T = 1] \\ &= X[\mathbb{E}(Y \mid X, T = 1) - X^T\theta].\end{aligned}$$

Estimation

In practice, $E\{U(Z;\theta) \mid Z^{(1)}, R=1\}$ is unknown. So, a imputation model $\phi(Z^{(1)};\theta,\beta)$ for $E\{U(Z;\theta) \mid Z^{(1)}, T=1\}$ is specified. Let $\hat{\beta}$ denote an estimator of β based the complete cases. Now, θ can be estimated as the solution to

$$\sum_{i=1}^{n} \frac{T_i}{\pi(Z_i^{(1)};\hat{\alpha})} U(Z_i;\theta) + \{1 - \frac{T_i}{\pi(Z_i^{(1)};\hat{\alpha})}\}\phi(Z_i^{(1)};\theta,\hat{\beta}) = 0, \tag{23}$$

denoted as $\hat{\theta}_{aipw}$. And it can be shown that

- $\hat{\theta}_{aipw}$ is doubly robust.
- when $\pi(Z^{(1)};\alpha)$ and $\phi(Z^{(1)};\theta,\beta)$ are correctly specified, $\hat{\theta}_{aipw}$ is locally efficient over the class of estimators that solve equations (25) for the given $U(Z_i;\theta)$ and arbitrary $\phi(Z_i^{(1)};\theta)$.

However, the efficiency of $\hat{\theta}_{aipw}$ also depends on the choice of function $u(Z;\theta)$.

Efficiency

Assumes

$$E(Y|X) = \mu(X;\theta).$$

where $\mu(X;\theta)$ is a known vector function of X. This is known as a restricted moment model and is usually fitted using generalized estimating equations. The locally efficient estimating equation is

$$D^T(X)V^{-1}(X)\{Y - \mu(X,\theta)\} = 0, \tag{24}$$

where

$$D(X) = \frac{\partial \mu(X,\theta)}{\partial \theta}, \qquad V(X) = \text{var}(Y|X)$$

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

Recommender Systems:

- E-commerce: Taobao, Amazon.
- Movie/Video: Douban, Netflix, YouTube.
- Music: Pandora, Last.fm.
- Social Network: Wechat, QQ, Facebook, Twitter.
- Reader: Google Reader, arXiv.
- Advertisement: Tiktok, Facebook
- .....

	ML 100K	Coat Shopping	Yahoo! R3
#users	943	290	15400
#items	1682	300	1000
#MNAR ratings	100000	6960	311704
#MAR ratings	0	4640	54000

Figure 1: Common datasets in RS

Missing rate is very high:

- ML 100K: $100000/(943 * 1682) = 0.063$;
- Coat Shopping: $6960/(290 * 300) = 0.080$;
- Yahoo! R3: $311704/(15400 * 1000) = 0.020$.

	Item 1	Item 2	Item 3	...	Item M
User 1	0.5	2.2	1.0	...	2.7
User 2	2.2	0.6	0.9	...	0.7
User 3	2.2	0.2	3.4	...	0.2
...	...	...	...	...	...
User N	1.9	1.0	0.2	...	0.3

Figure 2: Data structure (ratings).

Black denotes observed data; Red represents unobserved data.

- **Target**: predict the ratings for all the user-item pair.
- **Challenges**: selection bias, data sparsity.

Setup

Denote with $u \in \{1, ..., U\}$ the users and with $i \in \{1, ..., I\}$ the items.

Framework 1. missing data (without defining potential outcomes)

Example: video websites.

- r_{ui}: the true rating of user u for video i.
- o_{ui}: observing indictor. $o_{ui} = 1 \iff r_{ui}$ is observed
- main challenge: selection bias.

o_{ui}	x_{ui}	r_{ui}
1	✓	✓
1	✓	✓
1	✓	✓
0	✓	
0	✓	
0	✓	

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, Xiao-Hua Zhou (2021), "Causal Analysis Framework for Recommendation", arXiv:2201.06716.

Framework 2. Potential outcomes (binary treatment)

Example: advertising.

- r_{ui}: $r_{ui} = 1$ if u click on item i; $r_{ui} = 0$ otherwise.
- o_{ui}: $o_{ui} = 1$ if item i is exposed to u; $o_{ui} = 0$ otherwise.
- It usually assumes that $r_{ui} \equiv 0$ if $o_{ui} = 0$.
- main challenge: selection bias or confounding bias.
- of interest: $r_{ui}(1)$.

o_{ui}	x_{ui}	r_{ui}	$r_{ui}(1)$	$r_{ui}(0)$
1	✓	✓	✓	0
1	✓	✓	✓	0
1	✓	✓	✓	0
0	✓	✓		0
0	✓	✓		0
0	✓	✓		0

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, Xiao-Hua Zhou (2021), "Causal Analysis Framework for Recommendation", arXiv:2201.06716.

Problem Formulation. $\mathcal{U} = \{u\}$ and $\mathcal{I} = \{i\}$ denote the set of users and items.

- *Unit*: a user-item pair (u, i).
- *Target population*: the set of all user-item pairs $\mathcal{D} = \mathcal{U} \times \mathcal{I}$.
- *Feature*: the feature $x_{u,i}$ describes user-item pair (u, i).
- *Treatment*: $o_{u,i} \in \{1, 0\}$. $o_{u,i} = 1$ or 0 denotes item i is exposed to user u or not.
- *Outcome*: the feedback $r_{u,i}$ of user-item pair (u, i).
- *Potential outcome*: $r_{u,i}(o)$ for $o \in \{0, 1\}$. It is the outcome that would be observed if $o_{u,i}$ had been set to o.

Estimand: Let $\mathbb{P}$ ($\mathbb{E}$) be the distribution (expectation) on the target population.

$$\mathbb{E}(r_{u,i}(1) \mid x_{u,i})$$

Model:

- $\hat{r}_{u,i}(1) = f_\phi(x_{u,i})$ be a model with parameters ϕ, predicting $\mathbb{E}(r_{u,i}(1) \mid x_{u,i})$.

Ideal Loss: If all potential outcomes $\{r_{u,i}(1) : (u,i) \in \mathcal{D}\}$ were observed, the ideal loss function for training ϕ is

$$\mathcal{L}_{ideal}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} e_{u,i}, \tag{25}$$

where $e_{u,i}$ is the prediction error, such as the least square loss:

$$e_{u,i} = (\hat{r}_{u,i}(1) - r_{u,i}(1))^2. \tag{26}$$

Naive Estimator: Biased.

$$\mathcal{L}_{naive}(\phi) = |\mathcal{O}|^{-1} \sum_{(u,i)\in\mathcal{O}} e_{u,i},$$

where $\mathcal{O} = \{(u,i) \mid (u,i) \in \mathcal{D}, o_{u,i} = 1\}$ be the set of exposed units.

Inverse propensity score (IPS) estimator: Large variance.

$$\mathcal{L}_{IPS}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e_{u,i}}{\hat{p}_{u,i}}, \tag{27}$$

where $p_{u,i} := \mathbb{P}(o_{u,i} = 1 \mid x_{u,i})$ is the propensity score.

	Item 1	Item 2	Item 3	...	Item M
User 1	0.5*1.3			...	
User 2			0.9*2.7	...	
User 3		0.2*3.4		...	0.2*1.4
...	...	...	...	...	...
User N			0.2*3.9	...	0.3*1.2

Figure 3: IPS estimaor

Doubly Robust Joint Learning (DR-JL):

$$\mathcal{L}_{DR}(\phi, \theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left[\hat{e}_{u,i} + \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} \right], \tag{28}$$

where $\hat{e}_{u,i} = g_\theta(x_{u,i})$ fits the prediction error $e_{u,i}$ using $x_{u,i}$, i.e., it estimates $\mathbb{E}[e_{u,i}|x_{u,i}]$.

Joint-Learning:

- given $\hat{\theta}$, ϕ is updated by minimizing $\mathcal{L}_{DR}(\phi, \hat{\theta})$;
- given $\hat{\phi}$, θ is updated by minimizing

$$\mathcal{L}_e^{DR-JL}(\phi, \theta) = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(\hat{e}_{u,i} - e_{u,i})^2}{\hat{p}_{u,i}}. \tag{29}$$

DR-TMLE

- We propose **DR-TMLE** that effectively captures the merits of both EIB and DR, by leveraging the targeted maximum likelihood estimation (TMLE) technique.
- Furthermore, we propose a novel RCT-free collaborative targeted learning algorithm for DR-TMLE, called **DR-TMLE-TL**, which updates the propensity model adaptively to reduce the bias of imputed errors.
- Both theoretical analysis and experiments demonstrate the advantages of the proposed methods compared with existing debiasing methods.

Peng Wu, Haoxuan Li, Yan Lyu & Xiao-Hua Zhou (2022), 'Doubly Robust Collaborative Targeted Learning for Recommendation on Data Missing Not at Random', arXiv.

DR-TMLE

Table 1: Comparison of various debiasing estimators

	Doubly Robust	Robust to Small Propensities	Boundedness	Without Extrapolation	Low Variance
IPS	×	×	×	✓	×
SNIPS	×	*o*	✓	✓	×
EIB	×	✓	✓	×	✓
DR	✓	×	×	*o*	*o*
DR-TMLE (ours)	✓	✓	✓	*o*	✓

Note: symbols ✓, *o* and × denotes good, medium and bad, respectively.

Semi-synthetic Experiments

Table 2: Mean and standard deviation of the relative error on various methods

	Naive	EIB	IPS	DR	DR-TMLE
ONE	0.0688 ± 0.0025	0.5442 ± 0.0016	0.0338 ± 0.0033	0.0140 ± 0.0034	**0.0053 ± 0.0026**
THREE	0.0790 ± 0.0028	0.5878 ± 0.0017	0.0390 ± 0.0037	0.0180 ± 0.0037	**0.0035 ± 0.0025**
FIVE	0.1027 ± 0.0028	0.6167 ± 0.0018	0.0511 ± 0.0033	0.0150 ± 0.0034	**0.0066 ± 0.0032**
ROTATE	0.1378 ± 0.0011	0.2533 ± 0.0004	0.0696 ± 0.0026	0.0401 ± 0.0016	**0.0325 ± 0.0015**
SKEW	0.0265 ± 0.0021	0.3584 ± 0.0007	0.0129 ± 0.0027	0.0101 ± 0.0027	**0.0029 ± 0.0020**
CRS	0.1062 ± 0.0022	0.1443 ± 0.0007	0.0526 ± 0.0026	0.0237 ± 0.0025	**0.0193 ± 0.0025**

Real-world Experiments.

Table 3: MSE, AUC, NDCG@5, and NDCG@10 on the MAR test set of COAT and YAHOO.

	COAT				YAHOO			
	MSE	AUC	NDCG@5	NDCG@10	MSE	AUC	NDCG@5	NDCG@10
Base Model	0.2448	0.7047	0.5912	0.6667	0.2496	0.6699	0.6347	0.7636
+ IPS	0.2304	0.6985	0.5980	0.6749	0.2501	0.6845	0.6449	0.7697
+ SNIPS	0.2410	0.7066	0.5978	0.6761	0.2502	0.6867	0.6509	0.7724
+ DR	0.2359	0.7031	0.6213	0.6967	**0.2420**	0.6867	0.6613	0.7791
+ DR-JL	0.2365	0.7039	0.6063	0.6857	0.2500	0.6850	0.6414	0.7673
+ DR-TL	0.2349	0.7102	0.6253	0.6933	0.2494	0.6808	0.6334	0.7622
+ DR-TMLE	**0.2161**	**0.7170**	**0.6348**	**0.6999**	**0.2115**	**0.7044**	**0.7008**	**0.8016**
+ DR-TMLE-JL	**0.2151**	**0.7236**	**0.6388**	**0.7047**	0.2577	**0.7036**	**0.6786**	**0.7884**
+ DR-TMLE-TL	**0.2119**	**0.7339**	**0.6526**	**0.7112**	**0.2472**	**0.7057**	**0.6758**	**0.7871**
+ MRDR-JL	0.2160	0.7203	0.6406	0.7035	**0.2496**	0.6842	0.6487	0.7717
+ MRDR-TL	0.2155	0.7200	0.6427	0.7047	**0.2494**	0.6805	0.6345	0.7623
+ MRDR-TMLE-JL	**0.2114**	**0.7278**	**0.6498**	**0.7101**	0.2568	**0.7027**	**0.6826**	**0.7899**
+ MRDR-TMLE-TL	**0.2114**	**0.7316**	**0.6428**	**0.7088**	0.2517	**0.7002**	**0.6759**	**0.7873**

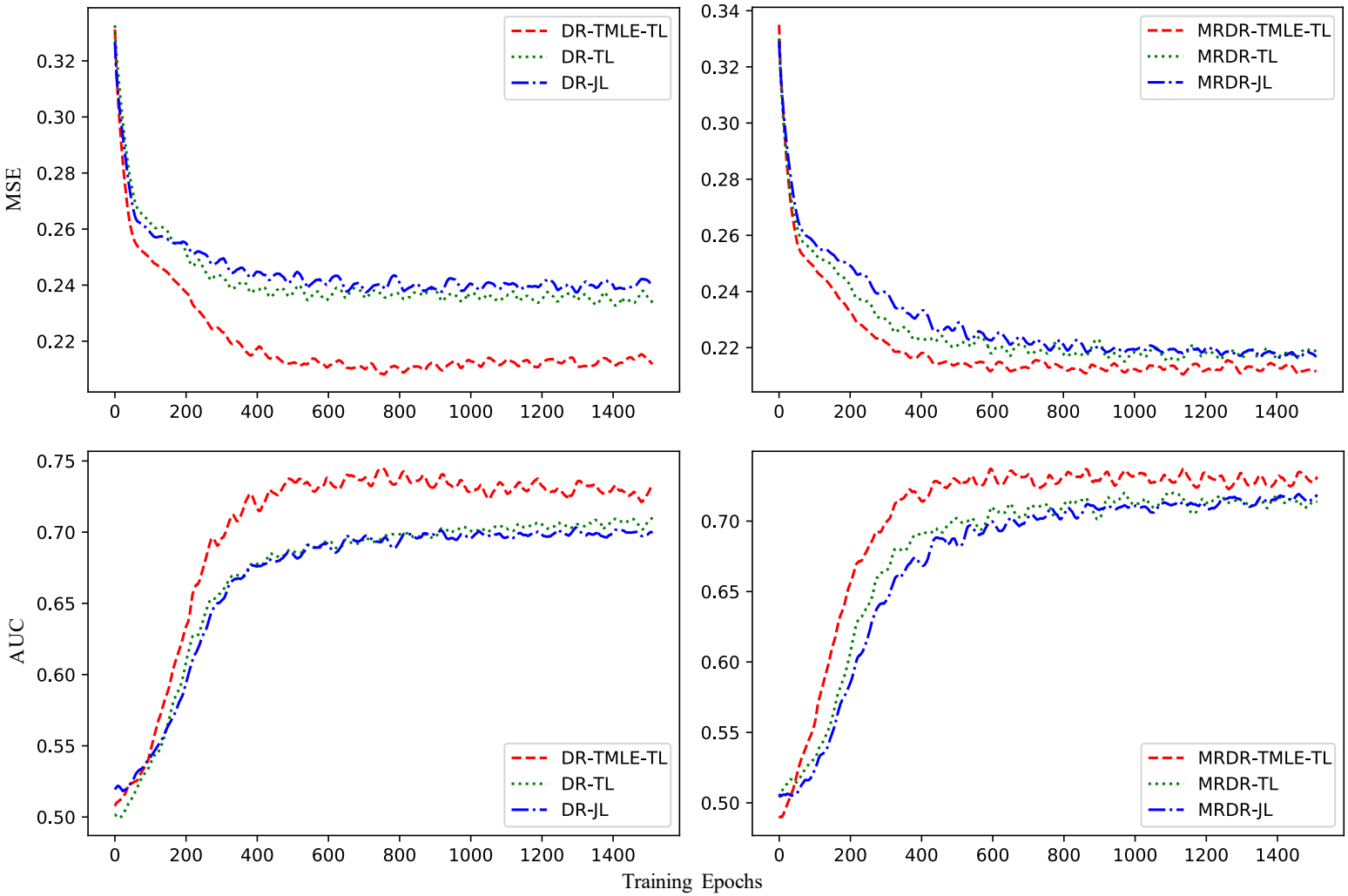


Figure 4: Advantage of TMLE collaborative targeted learning with DR (left) and MRDR (right) as the initialization, comparing with targeted learning without TMLE steps (TL) and JL.

1 Problem Statement and Preliminaries

2 Doubly robust OR

3 Doubly robust bounded IPW

4 Bias-Reduced DR estimator

5 DR methods for estimating regression coefficient

6 DR methods in recommender system

7 Discussion

1. **Estimation v.s. Prediction**
   - Doubly robust estimating equation (mean and regression coefficient)
   - Doubly robust loss function (regression coefficient).

2. **Other DR methods.**
   - Quantile Regression.
   - Instrumental variable.

3. **Future Directions**
   - Consider the problem of unobserved confounder.
   - Bias-Reduced DR Loss function, Multiple Robust.