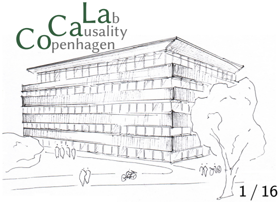
Efficient adjustment for complex covariates:
Gaining efficiency with DOPE

Alexander Mangulad Christgau
Joint work with Niels Richard Hansen

European Causal Inference Meeting

April 2024

UNIVERSITY OF COPENHAGEN



Outline

1. Covariate adjustment and motivation
2. Efficiency theory
3. Debiased Outcome-adapted Propensity Estimation

Adjusted treatment-specific mean of outcome

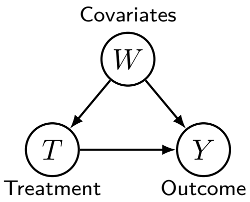
Template observation $T \in \{0,1\}$, $W \in \mathbb{R}^p$, $Y \in \mathbb{R}$.

- Adjusted mean for treatment $t \in \{0,1\}$ is

$$\chi_t \coloneqq \mathbb{E}\big[\mathbb{E}[\,Y \mid T = t, W]\big].$$

- Under the assumption of *no unmeasured confounding*:

$$\chi_t = \mathbb{E}\big[Y \mid \mathrm{do}(T = t)\big] \quad \text{and} \quad \chi_1 - \chi_0 = \mathrm{ATE}.$$



Estimation of adjusted mean

- Given i.i.d. observations $(T_i, W_i, Y_i)_{i\in[n]}$, we could produce an estimate $\widehat{g}(t, \cdot\,)$ of

  $$g(t, \cdot\,) := \mathbb{E}[\,Y \,|\, T = t, W = \cdot\,]$$

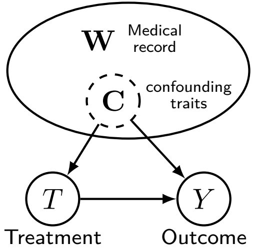
  and then use the estimator $\widehat{\chi}_t^{\mathrm{reg}} := \frac{1}{n}\sum_{i=1}^n \widehat{g}(t, W_i)$.
- $\widehat{\chi}_t^{\mathrm{reg}}$ is not $\sqrt{n}$-consistent – unless restrictive assumptions on g.
- The *AIPW estimator* is given by

  $$\widehat{\chi}_t^{\mathrm{aipw}} := \frac{1}{n}\sum_{i=1}^n \left(\widehat{g}(t, W_i) + \frac{\mathbb{1}(T_i = t)(Y_i - \widehat{g}(t, W_i))}{\widehat{m}_t(W_i)} \right),$$

  where $\widehat{m}_t(\cdot) = \widehat{\mathbb{P}}(T = t \,|\, W = \cdot)$ is an estimate of $m_t(\cdot) = \mathbb{P}(T = t \,|\, W = \cdot)$.

Challenges with complex and unstructured covariates

To meet the assumption of no unmeasured confounding, suppose we record a large/complex covariate $\mathbf{W}$.

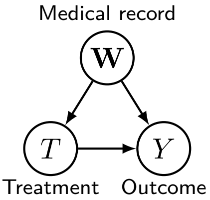


Problem 1

Medical record is difficult to model.

If $\mathbf{W}$ is a text variable we could:

- Apply a pretrained text embedding $\rightarrow \mathbf{Z} = \varphi(\mathbf{W}) \in \mathbb{R}^d$.
- Do standard adjustment based on $\mathbf{Z}$.



Embeddings need fine-tuning[1].

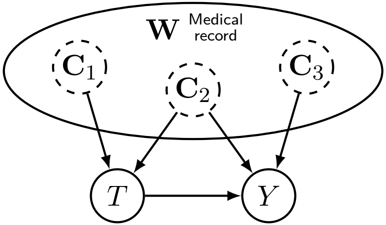
Is there an optimal way to fine-tune the embedding?

[1] Veitch, Wang, and Blei (2019) and Veitch, Sridhar, and Blei (2020)

Problem 2

Medical record might be highly predictive of treatment assignment

- Problematic for inverse propensity weights.
- Suppose that confounding traits can categorized:



- Can we formally and practically distinguish the information in $\mathbf{W}$?

Adjustment for covariate transformations

It can be sensible to adjust for $\mathbf{Z} = \varphi(\mathbf{W})$ rather than $\mathbf{W}$ itself.
Suppose $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- The adjusted mean for $\mathbf{Z}$ is denoted by:

$$\chi_t(\mathbf{Z}; P) := \mathbb{E}_P[\mathbb{E}_P[Y \mid T = t, \mathbf{Z}]]$$

- We want $\mathbf{Z}$ such that

$$\mathbf{Z} \textit{ valid transformation} \quad \overset{\text{def}}{\iff} \quad \chi_t(\mathbf{Z}; P) = \chi_t(\mathbf{W}; P) \text{ for all } P \in \mathcal{P}.$$

Connections to adjustment sets

Example 1

Given a DAG $\mathcal{D}$ on the nodes $(T, \mathbf{W}, Y) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$, let $\mathcal{P}$ contain all distributions consistent w.r.t. $\mathcal{D}$. Assume $\mathbf{W}$ is a *valid adjustment set*, equivalent with

$$\forall P \in \mathcal{P}: \qquad \chi_t(\mathbf{W}, P) = \mathbb{E}_P[\, Y \,|\, \mathrm{do}(T = t)\,].$$

Then

- A subset $\mathbf{Z} \subseteq \mathbf{W}$ is a valid adjustment set if and only if it is valid as a transformation.
- The *optimal adjustment set* is obtained by a pruning procedure[2]:
  iteratively remove covariates that are not predictive of outcome conditionally on treatment and remaining covariates.

[2] Henckel, Perković, and Maathuis (2022) and Rotnitzky and Smucler (2020)

Non-graphical example

Example 2

Suppose again that $\mathbf{W} \in \mathbb{R}^p$ and

$$Y = g\big(T, \|\mathbf{W}\|\big) + \varepsilon_Y, \qquad \mathbb{E}[\varepsilon_Y | T, \mathbf{W}] = 0,$$

for an unknown function g. Then

- $\mathbf{W}$ is *a priori* the only valid adjustment set.
- $\mathbf{Z} = \|\mathbf{W}\|$ is a valid transformation of $\mathbf{W}$. Follows from:

$$\mathbb{E}[Y \mid T, \mathbf{W}] = g(T, \mathbf{Z}) = \mathbb{E}[Y \mid T, \mathbf{Z}].$$

- For AIPW, should we estimate the propensity score based on $\mathbf{W}$ or $\mathbf{Z}$?

Efficiency bound

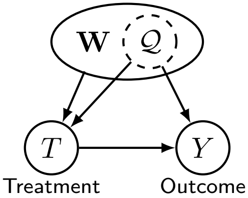
Hahn (1998) semiparametric efficiency bound:
If $\mathcal{P}$ is sufficiently dense, all RAL estimators of $\chi_t(\mathbf{W}; P)$ will have asymptotic variance of at least

$$\mathbb{V}_t(\mathbf{W}; P) := \mathrm{Var}_P\Big(\text{influence function of } \chi_t(\mathbf{W}; P)\Big).$$

- The bound can usually be improved if $\mathcal{P}$ is consistent w.r.t. a DAG (Example 1).
- Under rate conditions, AIPW/TMLE achieve this efficiency bound.
- If $\mathbf{Z}$ is a valid transformation of $\mathbf{W}$, then the efficiency bound is at most $\mathbb{V}_t(\mathbf{Z}; P)$.
- Note that $\chi_t(\mathbf{Z}; P)$ and $\mathbb{V}_t(\mathbf{Z}; P)$ depend only on the information in $\mathbf{Z} = \varphi(\mathbf{W})$.

Conditional outcome information

1. The information in $\sigma(\mathbf{W})$ that is "minimally sufficient" for prediction of $Y \mid T = t$ should be more efficient than $\mathbf{W}$ for adjustment:



2. We construct this information $\mathcal{Q}$ and prove that it is optimal to adjust for among all transformations that preserve the conditional outcome.
3. In particular, for all $P \in \mathcal{P}$ it holds that $\mathbb{V}_t(\mathcal{Q}; P) \leq \mathbb{V}_t(\mathbf{W}; P)$.

Debiased Outcome-adapted Propensity Estimation

General estimation framework for any outcome regression method with predictions of the form

$$\widehat{\mathbb{E}}[Y \mid T = t, \mathbf{W} = \mathbf{w}] = \widehat{h}(t, \widehat{\varphi}(\mathbf{w})).$$

Abridged version:

1. regress $(Y_i)_{i\in[n]}$ onto $(T_i, \mathbf{W}_i)_{i\in[n]}$ to obtain estimates $\widehat{\varphi}(\cdot)$ and $\widehat{h}(\cdot,\cdot)$.
2. regress $(T_i)_{i\in[n]}$ onto $(\widehat{\varphi}(\mathbf{W}_i))_{i\in[n]}$ to obtain an estimate of the form $\widehat{m}_t(\cdot) = \widetilde{m}_t(\widehat{\varphi}(\cdot))$.
3. Compute AIPW estimate based on nuisance estimates $\widehat{h}(t, \widehat{\varphi}(\cdot))$ and $\widehat{m}_t(\cdot)$.

Related to existing works and estimation methods[3].

[3] Shortreed and Ertefaie (2017), van de Geer (2019), Ju, Benkeser, and van der Laan (2020), and Benkeser, Cai, and van der Laan (2020).

Example in single-index model

For $\mathbf{W} \in \mathbb{R}^p$, the single-index model assumes that

$$Y = h(T, \mathbf{W}^\top \theta) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid T, \mathbf{W}] = 0,$$

where $h(\cdot, \cdot)$ and $\theta \in \mathbb{R}^p$ are unknown.

- Can apply DOPE with $\widehat{\varphi}(\mathbf{w}) = \mathbf{w}^\top \widehat{\theta}$, and with h and θ estimated using any single-index regression method.
- Paper contains a simulation study of this procedure as well as application to real data.

Conclusion

- Some ideas for efficient adjustment in graphical models generalize to a non-graphical setting.
- It can pay off to adjust for an outcome-adapted transformation in certain situations.
- See arXiv preprint[4] for further details, in particular: general information-based adjustment framework, efficiency results, asymptotic analysis of DOPE and an application to real data.

Thank you for listening 😎

[4] Christgau and Hansen (2024)

References

- Benkeser, David, Weixin Cai, and Mark J van der Laan (2020). "A Nonparametric Super-Efficient Estimator of the Average Treatment Effect". In: Statistical Science 35.3, pp. 484–495.
- Christgau, Alexander Mangulad and Niels Richard Hansen (2024). "Efficient adjustment for complex covariates: Gaining efficiency with DOPE". In: arXiv preprint arXiv:2402.12980.
- Hahn, Jinyong (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: Econometrica, pp. 315–331.
- Henckel, Leonard, Emilija Perković, and Marloes H. Maathuis (2022). "Graphical criteria for efficient total effect estimation via adjustment in causal linear models". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84.2, pp. 579–599.
- Ju, Cheng, David Benkeser, and Mark J van der Laan (2020). "Robust inference on the average treatment effect using the outcome highly adaptive lasso". In: Biometrics 76.1, pp. 109–118.
- Rotnitzky, Andrea and Ezequiel Smucler (2020). "Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models.". In: J. Mach. Learn. Res. 21.188, pp. 1–86.
- Shortreed, Susan M and Ashkan Ertefaie (2017). "Outcome-adaptive lasso: variable selection for causal inference". In: Biometrics 73.4, pp. 1111–1122.
- van de Geer, Sara (2019). "On the asymptotic variance of the debiased Lasso". In: Electronic Journal of Statistics 13, pp. 2970–3008.
- Veitch, Victor, Dhanya Sridhar, and David Blei (2020). "Adapting text embeddings for causal inference". In: Conference on Uncertainty in Artificial Intelligence. PMLR, pp. 919–928.
- Veitch, Victor, Yixin Wang, and David Blei (2019). "Using embeddings to correct for unobserved confounding in networks". In: Advances in Neural Information Processing Systems 32.