UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Efficient representation adjustment

Alexander Mangulad Christgau

October 13, 2022

Ongoing work with Niels Richard Hansen
ETH-UCPH-TUM Workshop

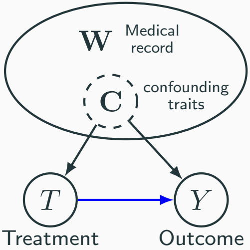
Motivation

Interested in a treatment effect $T \longrightarrow Y$.

Motivation

Interested in a treatment effect $T \longrightarrow Y$.

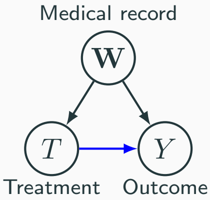
Confounders are indirectly measured via $\mathbf{W}$:



Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

[1] Chernozhukov et al. (2018)

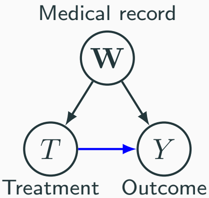
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

[1] Chernozhukov et al. (2018)

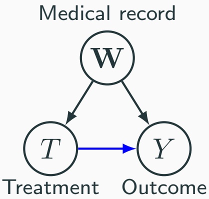
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

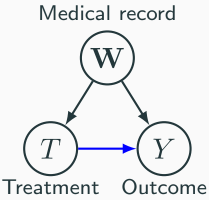
[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

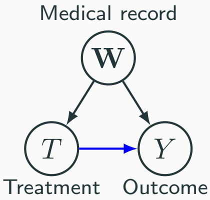
[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

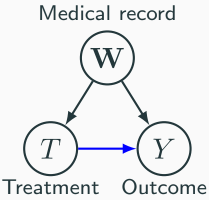
- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

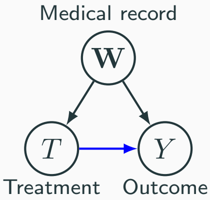
[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

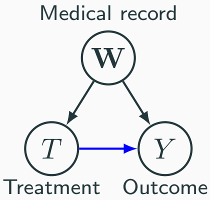
[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Motivation: challenge 1

Medical record is difficult to model.

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

- Use a pretrained *text embedding*.
- Do standard adjustment on embedding.
- "Double ML[1] with an extra step"



Embeddings need *finetuning*[2].

- Is it valid to finetune embedding once for all prediction tasks?
- Is there an "optimal" way to finetune the embedding?

[1] Chernozhukov et al. (2018)
[2] Veitch et al. (2020), Veitch et al. (2019)

Medical record is highly predictive of treatment assignment

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

- Can we formally distinguish information in $\mathbf{W}$?
- Can we leverage the existence of 'over-adjustments'?

Medical record is highly predictive of treatment assignment

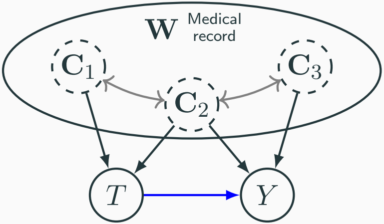
- Problematic for inverse propensity weighting.
- Suppose that confounding traits can categorized:

- Can we formally distinguish information in $\mathbf{W}$?
- Can we leverage the existence of 'over-adjustments'?

Motivation: challenge 2

Medical record is highly predictive of treatment assignment

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

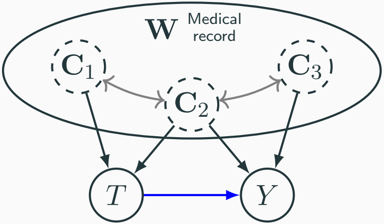


- Can we formally distinguish information in $\mathbf{W}$?
- Can we leverage the existence of 'over-adjustments'?

Motivation: challenge 2

Medical record is highly predictive of treatment assignment

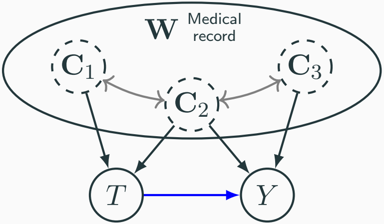
- Problematic for inverse propensity weighting.
- Suppose that confounding traits can categorized:



- Can we formally distinguish information in $\mathbf{W}$?
- Can we leverage the existence of 'over-adjustments'?

Medical record is highly predictive of treatment assignment

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



- Can we formally distinguish information in $\mathbf{W}$?
- Can we leverage the existence of 'over-adjustments'?

Motivation: synthesis

It can be natural to adjust for a transformation of $\mathbf{W}$ rather than $\mathbf{W}$ itself:

- For challenge 1: An embedding.
- For challenge 2: A projection onto a subset.

Objective: Formulate a general theory for adjustment that accomodates both settings.

Motivation: synthesis

It can be natural to adjust for a transformation of $\mathbf{W}$ rather than $\mathbf{W}$ itself:

- For challenge 1: An embedding.
- For challenge 2: A projection onto a subset.

Objective: Formulate a general theory for adjustment that accomodates both settings.

Adjusting for representations

Let $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- A representation of $\mathbf{W}$ is just a transformation $\mathbf{Z} = \varphi(\mathbf{W})$.
- Adjusting for $\mathbf{Z}$ means computing $\chi_t(\mathbf{Z}; P)$ where:

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

- Assume that $\mathbb{E}_P[Y|\mathrm{do}(T = t)] = \chi_t(\mathbf{W}; P)$.
- We want $\mathbf{Z}$ such that $\chi_t(\mathbf{W}; P) = \chi_t(\mathbf{Z}; P)$.

Adjusting for representations

Let $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- A representation of $\mathbf{W}$ is just a transformation $\mathbf{Z} = \varphi(\mathbf{W})$.
- Adjusting for $\mathbf{Z}$ means computing $\chi_t(\mathbf{Z}; P)$ where:

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

- Assume that $\mathbb{E}_P[Y | \mathrm{do}(T = t)] = \chi_t(\mathbf{W}; P)$.
- We want $\mathbf{Z}$ such that $\chi_t(\mathbf{W}; P) = \chi_t(\mathbf{Z}; P)$.

Adjusting for representations

Let $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- A representation of $\mathbf{W}$ is just a transformation $\mathbf{Z} = \varphi(\mathbf{W})$.
- Adjusting for $\mathbf{Z}$ means computing $\chi_t(\mathbf{Z}; P)$ where:

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

- Assume that $\mathbb{E}_P[Y|\mathrm{do}(T = t)] = \chi_t(\mathbf{W}; P)$.
- We want $\mathbf{Z}$ such that $\chi_t(\mathbf{W}; P) = \chi_t(\mathbf{Z}; P)$.

Adjusting for representations

Let $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- A representation of $\mathbf{W}$ is just a transformation $\mathbf{Z} = \varphi(\mathbf{W})$.
- Adjusting for $\mathbf{Z}$ means computing $\chi_t(\mathbf{Z}; P)$ where:

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

- Assume that $\mathbb{E}_P[Y | \mathrm{do}(T = t)] = \chi_t(\mathbf{W}; P)$.
- We want $\mathbf{Z}$ such that $\chi_t(\mathbf{W}; P) = \chi_t(\mathbf{Z}; P)$.

Adjusting for representations

Let $(T, \mathbf{W}, Y) \sim P$ for some $P \in \mathcal{P}$.

- A representation of $\mathbf{W}$ is just a transformation $\mathbf{Z} = \varphi(\mathbf{W})$.
- Adjusting for $\mathbf{Z}$ means computing $\chi_t(\mathbf{Z}; P)$ where:

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

- Assume that $\mathbb{E}_P[Y | \mathrm{do}(T = t)] = \chi_t(\mathbf{W}; P)$.
- We want $\mathbf{Z}$ such that $\chi_t(\mathbf{W}; P) = \chi_t(\mathbf{Z}; P)$.

Observation

- Adjusting for $\mathbf{Z}$ is theoretically equivalent to adjusting for any bimeasurable transformation of $\mathbf{Z}$.
- Adjustment depends only on information $\sigma(\mathbf{Z})$.

Observation

- Adjusting for $\mathbf{Z}$ is theoretically equivalent to adjusting for any bimeasurable transformation of $\mathbf{Z}$.
- Adjustment depends only on information $\sigma(\mathbf{Z})$.

Definition

Let $\mathcal{Z} \subseteq \sigma(\mathbf{W})$ be a σ-algebra.
We say $\mathcal{Z}$ is $\mathcal{P}$-*valid* if

$$\chi_t(\mathcal{Z}; P) = \chi_t(\mathbf{W}; P), \qquad \text{for all } t \text{ and } P.$$

We say $\mathcal{Z}$ is $\mathcal{P}$-*COS* if

$$b_t(\mathcal{Z}; P) = b_t(\mathbf{W}; P), \qquad P\text{-}a.s. \text{ for all } t \text{ and } P.$$

If there exists a representation $\mathbf{Z} = \varphi(\mathbf{W})$ such that $\mathcal{Z} = \sigma(\mathbf{Z})$, then $\mathcal{Z}$ is called a *description* of $\mathbf{W}$.

Relation to adjustment sets

Example

Suppose $\mathbf{W} \in \mathbb{R}^k$ and let $\mathcal{D}$ be a DAG on the nodes $(T, \mathbf{W}, Y)$. Assume $\mathcal{P} = \mathcal{M}(\mathcal{D})$ is the set of distributions that are Markovian with respect to $\mathcal{D}$.

Then:

- For any $\mathbf{Z} \subseteq \mathbf{W}$, the σ-algebra $\sigma(\mathbf{Z})$ is a description of $\mathbf{W}$.
- $\mathbf{Z}$ is a *valid adjustment* set if and only if $\sigma(\mathbf{Z})$ is $\mathcal{P}$-valid.

Relation to adjustment sets

Example

Suppose $\mathbf{W} \in \mathbb{R}^k$ and let $\mathcal{D}$ be a DAG on the nodes $(T, \mathbf{W}, Y)$. Assume $\mathcal{P} = \mathcal{M}(\mathcal{D})$ is the set of distributions that are Markovian with respect to $\mathcal{D}$.
Then:

- For any $\mathbf{Z} \subseteq \mathbf{W}$, the σ-algebra $\sigma(\mathbf{Z})$ is a description of $\mathbf{W}$.
- $\mathbf{Z}$ is a *valid adjustment* set if and only if $\sigma(\mathbf{Z})$ is $\mathcal{P}$-valid.

Example

Suppose $\mathbf{W} \in \mathbb{R}^k$ and let $\mathcal{D}$ be a DAG on the nodes $(T, \mathbf{W}, Y)$. Assume $\mathcal{P} = \mathcal{M}(\mathcal{D})$ is the set of distributions that are Markovian with respect to $\mathcal{D}$.
Then:

- For any $\mathbf{Z} \subseteq \mathbf{W}$, the σ-algebra $\sigma(\mathbf{Z})$ is a description of $\mathbf{W}$.
- $\mathbf{Z}$ is a *valid adjustment* set if and only if $\sigma(\mathbf{Z})$ is $\mathcal{P}$-valid.

Non-graphical example

Example

Assume $\mathbf{W} \in \mathbb{R}^k$ and

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

where $\alpha \in \mathbb{R}$ and $g \in C^1(\mathbb{R}_{\geq 0})$. Then

- $\mathbf{W}$ is the only valid adjustment set for (T, Y).
- $\sigma\big(\|\mathbf{W}\|\big)$ is a $\mathcal{P}$-COS description of $\mathbf{W}$.

Non-graphical example

Example

Assume $\mathbf{W} \in \mathbb{R}^k$ and

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

where $\alpha \in \mathbb{R}$ and $g \in C^1(\mathbb{R}_{\geq 0})$. Then

- $\mathbf{W}$ is the only valid adjustment set for (T, Y).
- $\sigma\big(\|\mathbf{W}\|\big)$ is a $\mathcal{P}$-COS description of $\mathbf{W}$.

Efficiency

Semiparametric efficiency bound: If $\mathcal{P}$ is sufficiently dense, all "reasonable" estimators of $\chi_t(\mathbf{W}; P)$ will have asymptotic variance of at least $\mathbb{V}_t(\mathbf{W}; P) :=$ *expression* (Hahn, 1998).

- We can improve the bound for $\mathcal{P} = \mathcal{M}(\mathcal{D})$!?[3]
- If $\mathcal{Z}$ is a $\mathcal{P}$-valid description of $\mathbf{W}$, then the efficiency bound is at most $\mathbb{V}_t(\mathcal{Z}; P)$.

[3] See e.g. Smucler et al. (2022).

Efficiency

Semiparametric efficiency bound: If $\mathcal{P}$ is sufficiently dense, all "reasonable" estimators of $\chi_t(\mathbf{W}; P)$ will have asymptotic variance of at least $\mathbb{V}_t(\mathbf{W}; P) :=$ *expression* (Hahn, 1998).

- We can improve the bound for $\mathcal{P} = \mathcal{M}(\mathcal{D})$!?[3]
- If $\mathcal{Z}$ is a $\mathcal{P}$-valid description of $\mathbf{W}$, then the efficiency bound is at most $\mathbb{V}_t(\mathcal{Z}; P)$.

[3] See e.g. Smucler et al. (2022).

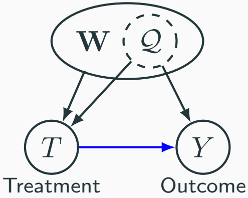
Efficiency

Semiparametric efficiency bound: If $\mathcal{P}$ is sufficiently dense, all "reasonable" estimators of $\chi_t(\mathbf{W}; P)$ will have asymptotic variance of at least $\mathbb{V}_t(\mathbf{W}; P) :=$ *expression* (Hahn, 1998).

- We can improve the bound for $\mathcal{P} = \mathcal{M}(\mathcal{D})$!?[3]
- If $\mathcal{Z}$ is a $\mathcal{P}$-valid description of $\mathbf{W}$, then the efficiency bound is at most $\mathbb{V}_t(\mathcal{Z}; P)$.

[3] See e.g. Smucler et al. (2022).

The conditional outcome algebra



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

The conditional outcome algebra

Theorem

For each $P \in \mathcal{P}$ define $\mathcal{Q}_P = \sigma\big(b_0(\mathbf{W}; P), b_1(\mathbf{W}; P)\big)$ and let

$$\mathcal{Q} \coloneqq \bigvee_{P \in \mathcal{P}} \mathcal{Q}_P.$$

A description $\mathcal{Z}$ is $\mathcal{P}$-COS if and only if $\mathcal{Z}$ contains $\mathcal{Q}$. Under additive noise on Y, it holds that

$$\mathbb{V}_t(\mathcal{Z}; P) - \mathbb{V}_t(\mathcal{Q}; P) = (\dots) \geq 0,$$

for all $\mathcal{P}$-COS descriptions $\mathcal{Z}$. In particular, the formula holds with $\mathcal{Z} = \sigma(\mathbf{W})$.

*Technical details about nullsets removed from theorem.

Summary

- There can be good reasons to transform a covariate $\mathbf{W}$ before adjustment:
  1. Embed $\mathbf{W}$ into euclidean space (practical)
  2. Remove overadjustment and redundant information (efficient)
- σ-algebras are an abstraction that account for equivalent representations.
- Many ideas for adjustment in DAGs generalize to similar non-graphical situations.

Some other topics (ongoing):

- General efficiency comparsion for descriptions.
- “Differentiable adjustment selection”.
- Estimation algorithms and asymptotic analysis.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 315–331.

Henckel, L., Perković, E., & Maathuis, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(2), 579–599. https://doi.org/https://doi.org/10.1111/rssb.12451

Rotnitzky, A., & Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models.. *J. Mach. Learn. Res.*, *21*(188), 1–86.

Smucler, E., Sapienza, F., & Rotnitzky, A. (2022). Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, *109*(1), 49–65.

Veitch, V., Sridhar, D., & Blei, D. (2020). Adapting text embeddings for causal inference. *Conference on Uncertainty in Artificial Intelligence*, 919–928.

Veitch, V., Wang, Y., & Blei, D. (2019). Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, *32*.

Comparison lemmas

Generalizations from Henckel et al. (2022) and Rotnitzky and Smucler (2020).

Lemma (Deletion of overadjustment)

Fix a $P \in \mathcal{P}$ *and let* $\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \subseteq \sigma(\mathbf{W})$ *be* σ*-algebras such that* $Y \perp\!\!\!\perp_P \mathcal{Z}_2 \mid T, \mathcal{Z}_1$. *Then* $\mathcal{Z}_1$ *is* P*-valid if and only if* $\mathcal{Z}_2$ *is* P*-valid. In any case,*

$$\mathbb{V}_t(\mathcal{Z}_2; P) - \mathbb{V}_t(\mathcal{Z}_1; P) = (\ldots) \geq 0.$$

Lemma (Supplementation with precision)

Fix $P \in \mathcal{P}$ *and let* $\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \subseteq \sigma(\mathbf{W})$ *be* σ*-algebras such that* $T \perp\!\!\!\perp_P \mathcal{Z}_2 \mid \mathcal{Z}_1$. *Then* $\mathcal{Z}_1$ *is* P*-valid if and only if* $\mathcal{Z}_2$ *is* P*-valid. In any case,*

$$\mathbb{V}_t(\mathcal{Z}_1; P) - \mathbb{V}_t(\mathcal{Z}_2; P) = (\ldots) \geq 0.$$