Treatment effect estimation with neural network-based models

aidd school

Manuel Haußmann

Questions

- We have a potential new treatment D for disease X. Does it work?
- Alice has been diagnosed with disease X. Should she be treated with D?
- What if Bob had not been treated?
- ...

Overview

- RCT vs OS—*Don't we already have a perfect solution?*
- Potential Outcomes—*How to formally speak about the task?*
- Estimators—*What do we estimate and how?*
- Approaches—*An Overview on proposals in the literature*
- Outlook—*What remains to be done*

RCT vs OS: The "*gold standard*", Randomized controlled trials

- ✓ principled approach reducing potential bias
- ✓ well structured, specific data collection
- ϟ expensive, time consuming
- ϟ ethical constraints
- ϟ rarity of disease
- ϟ biased populations

OS: Observational Study

RCT vs OS: Effect estimation with electronic health records

- ✓ abundant data
- ✓ representative of the wider population
- ↯ confounding issues
- ↯ worse data quality

RCT vs OS: Effect estimation with electronic health records

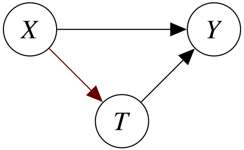
- ✓ abundant data
- ✓ representative of the wider population
- ↯ confounding issues
- ↯ worse data quality

$\Rightarrow$ Today: Focus on treatment effect estimation via observational data

Notation

For a patient i we observe...

- covariates $X_i \in \mathcal{X}$ *(e.g., age, gender, medical history, lab measurements,...)*
- a treatment assignment $T_i \in \mathcal{T}$ *(e.g., receive an operation, a specific drug dosage,...)*
    - Assume throughout that $\mathcal{T} = \{0, 1\}$
- an outcome $Y_i \in \mathcal{Y}$ *(e.g., time until death, recovery,...)*



Example

Patient	Age	Gender	Lab_1	...	Treated	Untreated
Alice	25	f	30 mg/l	...	?	?
Bob	32	m	13 mg/l	...	12 months	?
Charlie	21	m	58 mg/l	...	?	7 months
Denise	27	f	23 mg/l	...	?	14 months
Eve	40	f	17 mg/l	...	34 months	?

Potential Outcomes (i)

- Assume $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$
- $Y(0), Y(1) \in \mathcal{Y}$ are potential outcomes
  - *We observe only $Y_{Bob}(1)$, never the counterfactual $Y_{Bob}(0)$*
- Conditional average treatment effect (CATE)

$$\tau(x) \triangleq \mathbb{E}\left[Y_i(1) - Y_i(0) | X = x\right]$$

- Average treatment effect (ATE): $\mathbb{E}_{p(x)}\left[\tau(x)\right]$
- Average treatment effect on the treated (ATT): $\mathbb{E}_{p(x)}\left[\tau(x) | T = 1\right]$

See e.g., Peters et al. (2017) for a discussion on the relation to the do-calculus

Potential Outcomes (i)

- Assume $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$
- $Y(0), Y(1) \in \mathcal{Y}$ are potential outcomes
  - *We observe only $Y_{Bob}(1)$, never the counterfactual $Y_{Bob}(0)$*
- Conditional average treatment effect (CATE)

$$\tau(x) \triangleq \mathbb{E}\left[Y_i(1) - Y_i(0) | X = x\right]$$

- Average treatment effect (ATE): $\mathbb{E}_{p(x)}\left[\tau(x)\right]$
- Average treatment effect on the treated (ATT): $\mathbb{E}_{p(x)}\left[\tau(x) | T = 1\right]$

$\rightarrow$ *We are interested in the conditional average treatment effect*

See e.g., Peters et al. (2017) for a discussion on the relation to the do-calculus

ATE vs CATE

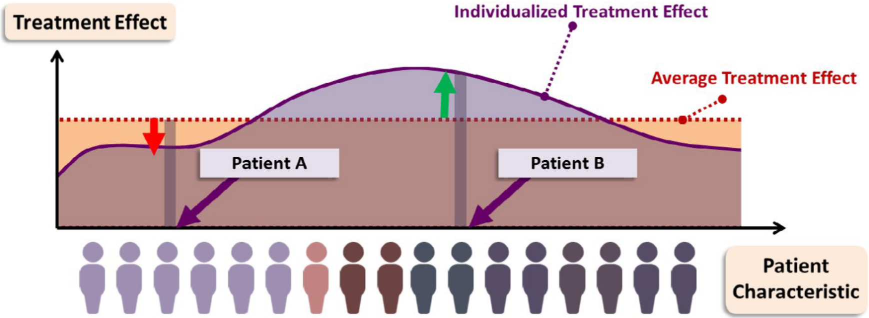


Figure via Bica et al. (2021)

Potential Outcomes (ii)

Assumptions for identifiability of causal effects

(i) Consistency $Y = TY(1) + (1 - T)Y(0)$

the potential outcome is the observed given a specific treatment

(ii) Unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T | X$ *(in an RCT:* $(Y(0), Y(1)) \perp\!\!\!\perp T$*)*

no hidden confounders $\rightarrow$ *can't be tested in practice*

(iii) Overlap $0 < \pi(x) < 1, \forall x \in \mathcal{X}$ where $\pi(x) \triangleq \mathbb{P}(T_i = 1 | X_i = x)$*(Propensity score)*

we need to observe treatment alternatives for an effect estimation

Potential Outcomes (ii)

Assumptions for identifiability of causal effects

(i) Consistency $Y = TY(1) + (1 - T)Y(0)$

the potential outcome is the observed given a specific treatment

(ii) Unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T|X$ *(in an RCT: $(Y(0), Y(1)) \perp\!\!\!\perp T$)*

no hidden confounders $\rightarrow$ can't be tested in practice

(iii) Overlap $0 < \pi(x) < 1, \forall x \in \mathcal{X}$ where $\pi(x) \triangleq \mathbb{P}(T_i = 1|X_i = x)$*(Propensity score)*

we need to observe treatment alternatives for an effect estimation

Unconfoundedness encourages a high dimensionality $\leftrightarrow$ Overlap encourages a low one

Potential Outcomes (iii) — Sidenote on Propensity Scores

- Propensity score: $\pi(x) \triangleq \mathbb{P}(T_i = 1 | X_i = x)$
- Balancing score: $b(X)$ such that $X \perp\!\!\!\perp Z | b(X)$
- Theorem: If $(Y(1), Y(0)) \perp\!\!\!\perp T | X$, then $(Y(1), Y(0)) \perp\!\!\!\perp T | b(X)$
- Theorem:[1] $\pi(x)$ is balancing and it is the "optimal" one.
- Use this to:
  1. Construct an estimator $\hat{\pi}(x)$
  2. Match two groups by the closeness of their estimated propensity scores
  3. Estimate the average treatment effect using the matched observations

[1] Rosenbaum and Rubin (1983)

Estimators – Two broad paths

Terminology following Curth et al., (2021)

The target: $\tau(x) = \mathbb{E}\left[Y(1) - Y(0)|X = x\right] = \mathbb{E}\left[Y(1)|X = x\right] - \mathbb{E}\left[Y(0)|X = x\right]$

1. *one-step plug-in learners*
   - Consider estimating $\mu_t(x) = \mathbb{E}\left[Y(t)|X = x\right]$
   - get $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$
2. *two-step learners*
   (i) Estimate $\eta = (\mu_0(x), \mu_1(x), \pi(x))$
   (ii) Construct pseudo-outcomes Y_η such that $\tau(x) = \mathbb{E}\left[Y_\eta|X = x\right]$

Estimators – One-step plugin learners

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Two broad approaches:

1. T-Learner: Learn separate models $\mu_0, \mu_1 : \mathcal{X} \to \mathcal{Y}$
2. S-Learner:
   (i) Augment the covariate space:
   Learn a joint model $\mu : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$, s.t., $\mu_t(x) \triangleq \mu(x, t)$
   (ii) Use a shared representation space:
   Learn $f_0(\cdot), f_1(\cdot), h(\cdot)$, s.t., $\mu_t(x) = f_t(h(x))$

Figure due to Curth et al. (2021)

Estimators – One-step plugin learners

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Two broad approaches:

1. T-Learner: Learn separate models $\mu_0, \mu_1 : \mathcal{X} \to \mathcal{Y}$
2. S-Learner:
   (i) Augment the covariate space:
   Learn a joint model $\mu : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$, s.t., $\mu_t(x) \triangleq \mu(x, t)$
   (ii) Use a shared representation space:
   Learn $f_0(\cdot), f_1(\cdot), h(\cdot)$, s.t., $\mu_t(x) = f_t(h(x))$

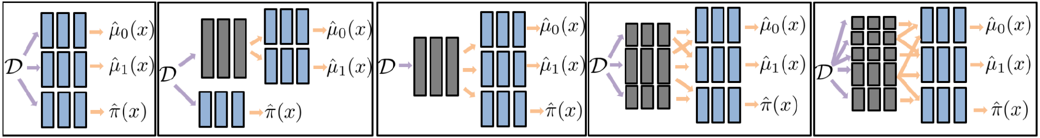
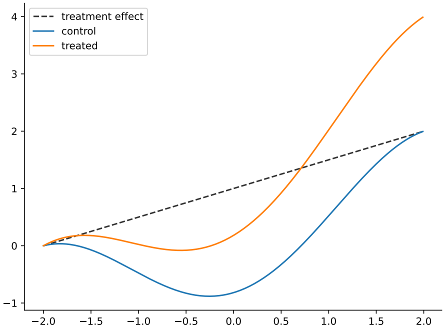


Figure due to Curth et al. (2021)

Estimators – Pseudo-outcomes

Why might we not be happy with them?

- T-Learners cannot take shared representations into account
- $\tau(x)$ might be simpler than $\mu_0(x), \mu_1(x)$



Estimators – Pseudo-outcomes: Regression Adjustment

Reminder, we consider two steps:

(i) Estimate $\mu_0(\cdot), \mu_1(\cdot), \pi(\cdot)$; *(ii)* Construct pseudo-observations Y_η to learn $\hat{\tau}$

Target: $\tau(x) = \mathbb{E}\left[Y_\eta | X = x\right]$

Three approaches for this task are...

- ...Regression adjustment $\rightarrow$ *unbiased if $\hat{\mu}$ is correct*
- ...Inverse Propensity weighting $\rightarrow$ *unbiased if $\hat{\pi}$ is correct*
- ...Doubly Robust Learner $\rightarrow$ *unbiased if either is unbiased*

1. Given $\hat{\mu}_0, \hat{\mu}_1$ impute treatment effects

$$D_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1) \qquad D_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0$$

2. Construct estimators $\hat{\tau}_1(x), \hat{\tau}_0(x)$
3. Estimate CATE as $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x) \quad (g(x) \in [0, 1])$

A simpler variant *(Curth et al., 2021)*

$$Y_{\hat{\eta}} = T(Y - \hat{\mu}_0(X)) + (1 - T)(\hat{\mu}_1(X) - Y)$$

Estimators – Pseudo-outcomes: Inverse propensity score weighting

Our pseudo-outcomes are given as

$$Y_{\hat{\eta}} = \left(\frac{T}{\hat{\pi}(X)} - \frac{1-T}{1-\hat{\pi}(X)} \right) Y$$

Estimators – Pseudo-outcomes: Inverse propensity score weighting

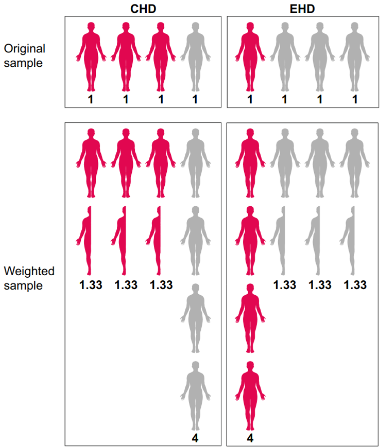


Figure due to Chesnaye et al. (2022)

Estimators – Pseudo-outcomes: Inverse propensity score weighting

Our pseudo-outcomes are given as

$$Y_{\hat{\eta}} = \left(\frac{T}{\hat{\pi}(X)} - \frac{1-T}{1-\hat{\pi}(X)} \right) Y$$

We get that

$$\mathbb{E}\left[Y_{\hat{\eta}} | X = x\right] = \frac{\pi(x)}{\hat{\pi}(x)} \mu_1(x) - \frac{1-\pi(x)}{1-\hat{\pi}(x)} \mu_0(x) = \tau(x),$$

if $\hat{\pi}(x) = \pi(x)$.

A downside: The variance explodes if $\pi(x)$ is close to zero/one

Combining the first two approaches we get

$$Y_{\hat{\eta}} = \left(\frac{T}{\hat{\pi}(X)} - \frac{1-T}{1-\hat{\pi}(X)}\right) Y + \left[\left(1 - \frac{T}{\hat{\pi}(X)}\right)\hat{\mu}_1(x) - \left(1 - \frac{1-T}{1-\hat{\pi}(X)}\right)\hat{\mu}_0(x)\right]$$

If $\hat{\pi} = \pi$ or $\hat{\mu}_t = \mu_t$ we get $\mathbb{E}\left[Y_{\hat{\eta}} | X = x\right] = \tau(x)$

A quick summary

- CATE: $\tau(x) = \mathbb{E}\left[Y(1) - Y(0) | X = x\right]$
- Step 1: Build estimators for μ_0, μ_1, π
- Step 2:
  - Estimate τ indirectly.
    Potential problems due to unnecessary complexity, but complete usage of $\mathcal{D}$
  - Estimate τ directly.
    Two step approach requires data split

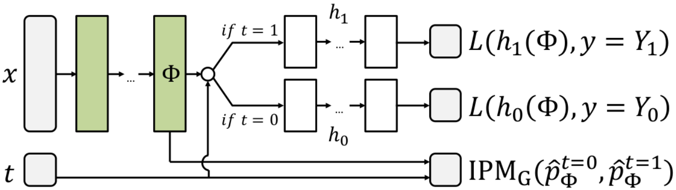
A quick summary

- CATE: $\tau(x) = \mathbb{E}\left[Y(1) - Y(0) | X = x\right]$
- Step 1: Build estimators for μ_0, μ_1, π
- Step 2:
  - Estimate τ indirectly.
    Potential problems due to unnecessary complexity, but complete usage of $\mathcal{D}$
  - Estimate τ directly.
    Two step approach requires data split

Note: So far we have not really cared about the estimation method

Common Approaches – Variations in the architectural structure

- Regularization within the representation space *(Shalit et al., 2017)*



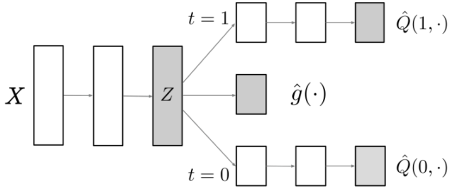
Increase the overlap by minimizing an Integral Probability Metric (IPM)

$$\min \mathsf{IPM}(p(\Phi|t = 1), p(\Phi|t = 0))$$

Known as *TARNet* and *CFRNet* *(with/without IPM)*

Common Approaches – Variations in the architectural structure

- Increasing the predictive constraints in the latent space *(Shi et al., 2019)*

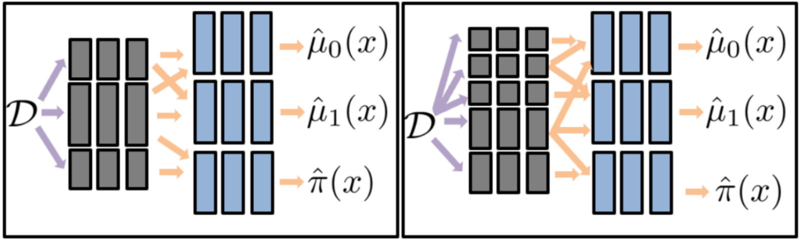


  - $Q \triangleq \mu$ and $g \triangleq \pi$
  - Predict the propensity score via the representation space
  - *(as well as an additional regularization on the loss)*

Known as *DragonNet*

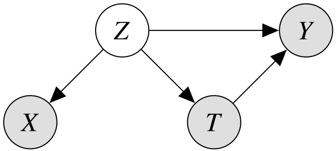
Common Approaches – Variations in the architectural structure

- Splitting the representation space *(Hassanpour and Greiner, 2020, Curth et al., 2021)*



Common Approaches – Generative models

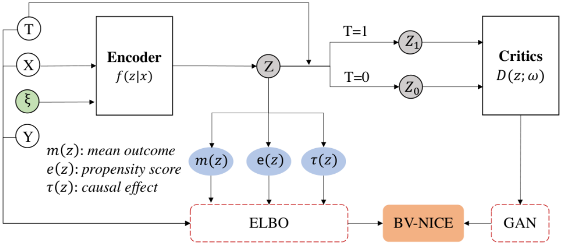
- Causal Effect Variational Autoencoder *(Louizos et al., 2017)*
  - Covariates X are a noisy view of latent covariates Z



  - Inference via amortized variational inference by optimizing the ELBO
  - But: See also Rissanen and Marttinen (2021) for a critique

Common Approaches – Generative models

- Balancing Variational Neural Inference for Causal Effects *(Lu et al., 2020)*



$$\mathbb{E}_{q(z)}\left[\log p(x|z) + \log p(y|z,t) + \log p(t|z)\right] - \mathrm{KL}\left(q(z|x,y,t) \parallel p(z)\right) - D(q_0, q_1)$$

(leave $\log p(x|z)$ optional; $m(\cdot), \tau(\cdot)$ part of R-learner for $\log p(y|z,t)$)

Outlook: Other questions to tackle

- Interpretability of the learned estimators *(E.g., Crabbé et al., 2022)*
  Doctors won't trust black-box predictors
- Uncertainty-aware models *(E.g., Jesson et al., 2020; 2021; 2022)*
  What about predictive uncertianties?
- Missing treatment information *(E.g., Kuzmanovic et al., 2023)*
  What about missing observations

Outlook: Other questions to tackle

- Further combinations of trial data with observational data
  - Combining RCT data with OS *(E.g., Hatt et al., 2022)*
    Can we use the complementary strengths?
  - External controls: Combination of single-arm trial data with hospital records
- Longitudinal structures *(E.g., Bica et al., 2020; Frauen et al., 2023)*
  What about time?
- Predictive guarantees (generalization bounds, etc.)
- ...

References

Bica et al., 2020 Time Series Deconfounder: Estimating Treatment Effects of Time in the Presence of Hidden Confounders

Bica et al., 2021 From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges

Chesnaye et al., 2022 An introduction to inverse probability of treatment weighting in observational research

Crabbé et al., 2022 Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability

Curth et al., 2021 Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms

Frauen et al., 2023 Estimating average causal effects from patient trajectories

References

Hassanpour et al., 2020 Learning Disentangled Representations for Counterfactual Regression

Hatt et al., 2022 Combining Observational and Randomized Data for Estimating Heterogeneous Treatment Effects

Jesson et al., 2020 Identifying causal-effect inference failure with uncertainty-aware models

Jesson et al., 2021 Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data

Jesson et al., 2022 Scalable Sensitivity and Uncertinty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions

Künzel et al., 2019 Metalearners for estimating heterogeneous treatment effects using machine learning

Kuzmanovic et al., 2022 Estimating Conditional Average Treatment Effects with Missing Treatment Information

References

Lu et al., 2020 Reconsidering Generative Objectives For Counterfactual Reasoning

Louizos et al., 2017 Causal Effect Inference with Deep Latent-Variable Models

Peters et al., 2017 Elements of causal inference

Rissanen and Marttinen, 2021 A Critical Loook at the Consistency of Causal Estimation with Deep Latent Variable Models

Rosebaum and Rubin, 1983 The central rol of the propensity score in observational studies for causal effects

Shalit et al., 2017 Estimating individual treatment effect: generalization bounds and algorithms

Shi et al., 2019 Adapting Neural Networks for the Estimation of Treatment Effects