CMPT 983 Fall 2022
Causal Inference

Martin Ester

Simon Fraser University

SFU

Motivation

- Want to know the effect of a treatment.
- Examples: new drug, mask policy, . . .
- Standard approach: Randomized Controlled Trial (RCT). Assign participants randomly to treatment / no treatment (control) and evaluate the difference of the average effects in the treatment and the control group.
- RCTs are expensive and often unethical.
- But observational data may be available.
- How to estimate the treatment effect from observational data?

SFU

Motivation

- Observational data records only the factual outcome.
- If we can predict the counterfactual outcome, then we can compute the treatment effect.
- Counterfactual outcome cannot be predicted by simply training a model from the observed data, due to selection bias.
- Selection bias is due to confounders that affect the treatment assignment and the outcome. The observed outcomes are caused by the treatment and by the confounder.

Motivation

Simpson's Paradox

Evaluate the success of two treatments A and B for kidney stones

Treatment Stone size	A	B
Small	**81/87 = 93%**	234/270 = 87%
Large	**192/263 = 73%**	55/80 = 69%
Overall	273/350 = 78%	**289/350 = 83%**

Treatment A is more successful for both small and large stones but is overall less effective.
Treatment B is preferred for small stones (easy) while treatment A is preferred for large stones (hard).

Definitions

- Dataset of observations
  $D = \{(x_i, t_i, y_i) | 1 \leq i \leq n\}$
- Unit: entity to which treatment is applied
- X: covariates (multi-dimensional)
- T: treatment (treatment assignment)
  in the binary case: 1=treatment/0=no treatment
- Y: outcome,
- Confounder
  Variable C that causally affects the treatment assignment and the outcome.

Definitions

- Potential outcome
  In the binary case: every unit has two potential outcomes: $Y(T = 1) = Y(1)$ and $Y(T = 0) = Y(0)$
  outcomes on unit *i*: $Y_i(1)$ and $Y_i(0)$
- Observed outcome
  outcome of the treatment that was actually applied
  $T = t$: observed outcome is Y_t
- Counterfactual outcome
  The hypothetical, missing potential outcome.
  $T = t$: counterfactual outcome is Y_{1-t}

SFU

Definitions

- Individual Treatment Effect (ITE) on unit *i* is defined as the difference of the potential outcomes:

$$ITE_i = Y_i(1) - Y_i(0)$$

- Average treatment effect (ATE)

$$ATE = E[Y(1) - Y(0)]$$

- Conditional average treatment effect (CATE) common measurement when the treatment effect varies across different subgroups

$CATE = E[Y(1) - Y(0)|X]$, e.g.

$$CATE_{male} = E[Y(1) - Y(0)|X = male]$$
$$CATE_{female} = E[Y(1) - Y(0)|X = female]$$

Identifiability

- Estimating treatment effects from observational data is hard.
- The treatment effect is identifiable only under certain assumptions. Otherwise, the estimate may be biased.
- Stable Unit Treatment Value Assumption
  The outcome of unit *i* depends only on the treatment(s) of unit *i*, not the treatments of the other units.
- Positivity and Overlap
  The treatment assignment is not deterministic.

$$0 < P(T = t|X = x) < 1$$

SFU

Identifiability

- Ignorability
  The potential outcome $Y(t)$ is independent of the treatment assignment T given the covariates X.
  $$\{Y(1), Y(0)\} \perp T|X$$
- Confounder
  A variable that affects both the treatment assignment and the outcome.
  Ignorability means that there is no unobserved (not part of X) confounder.
  → Ignorability also called Unconfoundedness.

1

Identifiability

- Selection bias
  The existence of confounders leads to selection bias:
  $$P(T|X) \neq P(T)$$

→ In an RCT, $P(T|X) = P(T)$.

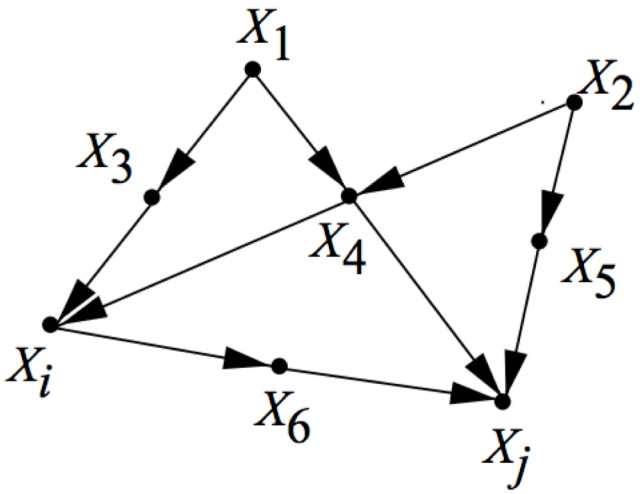
- Selection bias implies that a model trained on factual outcomes will be inaccurate when predicting counterfactual outcomes.

$$1$$

Identifiability

- Causal graph
  A directed acyclic graph where nodes represent variables and edges represent probabilistic dependencies.
- Backdoor criterion
  A set of variables (nodes) *X* satisfies the backdoor criterion for *T* and *Y* if no node in *X* is a descendent of *T* and *X* blocks every path between *T* and *Y* that contains a directed edge into *T*.

Identifiability



Both $\{X_3, X_4\}$ and $\{X_4, X_5\}$ satisfy the backdoor criterion for X_i and X_j

But $\{X_4\}$ does not

Identifiability

- If *X* satisfies the back-door criterion relative to *T* and *Y,* then the causal effect of *T* on *Y* is identifiable.
- To estimate the treatment effect, condition on (adjust for) *X.*

$$ATE = E_X[E_Y[Y(1)|X] - E_Y[Y(0)|X]]$$

- If the backdoor criterion is satisfied, the potential outcome of two units with the same covariates is the same.
- The treatment assignment mechanism is also identical.
- If all confounders are observed, i.e. are part of *X*, *X* satisfies the backdoor criterion.

Identifiability

- Propensity score

$$g(X) = P(T = 1|X)$$

- Sufficiency of the propensity score
  If the ATE is identifiable from observational data by adjusting for *X*, then adjusting for the propensity score g(*X*) also suffices to identify the ATE.

$$ATE = E_X\left[E_Y[Y(1)|g(X)] - E_Y[Y(0)|g(X)]\right]$$

→ Sufficiency holds only for ATE, not for ITE.

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

SFU

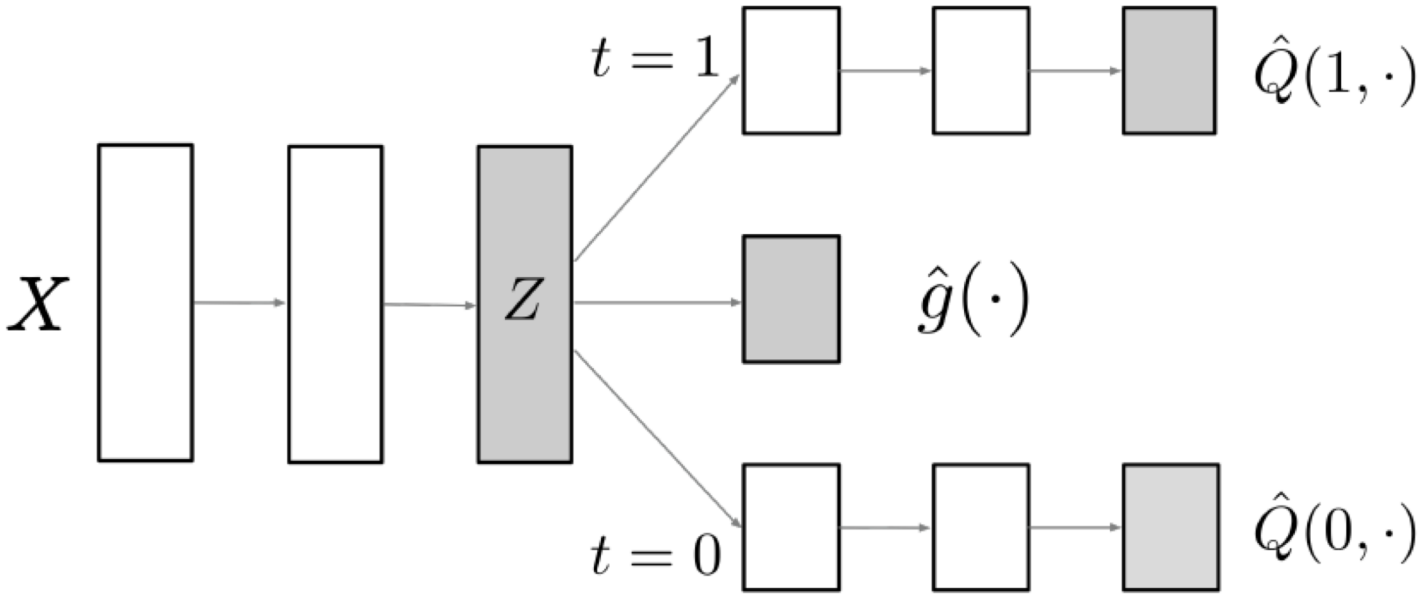
- How to exploit the sufficiency of the propensity score for a neural network?
- DragonNet
- ATE of a binary treatment
- It suffices to adjust for only the information in *X* that is relevant for predicting the treatment.
- The parts of *X* that are not relevant for predicting the treatment are irrelevant for the estimation of the causal effect, and are effectively noise for the adjustment. As such, conditioning on these parts may hurt the performance.

Adapting Neural Networks for the Estimation of Treatment Effects

SFU

- A multi-task neural network predicting the propensity score and the conditional outcomes $T(1)$ and $T(0)$.
- Learn a latent representation *Z* of *X* which works for all three prediction tasks.
- Trade off accuracy of outcome prediction for accuracy of prediction of the propensity score.
- 2-hidden layer neural networks for each of the outcome models.
- A simple linear map followed by a sigmoid for the propensity score model to tightly couple the representation to the estimated propensity score.

Adapting Neural Networks for the Estimation of Treatment Effects



$$argmin_\theta \ \frac{1}{n}\sum_i \left[(Q^{\text{nn}}(t_i, x_i; \theta) - y_i)^2 + \alpha \text{CrossEntropy}(g^{\text{nn}}(x_i; \theta), t_i)\right]$$

Adapting Neural Networks for the Estimation of Treatment Effects

SFU

- How to evaluate the performance of causal inference methods?
- There are no datasets with ground truth treatment effects.
- Use semi-synthetic datasets:
  real-life covariates
  treatment assignment and outcome are synthetic
    sampled from distributions $f(X)$ and $g(X)$
    that depend on X
- Measure the MAE or RMSE
  comparing the predicted and the ground truth outcome.

Adapting Neural Networks for the Estimation of Treatment Effects

SFU

- Infant Health and Development Program (IHDP) dataset
- Single binary treatment, continuous outcome.
- Mimics a study on infant development.
- Treatment = child had home visit from a trained provider.
- 24 covariates
- Outcome is cognitive test scores, and the goal is to measure the causal effect of the home visits.
- Benchmark contains ten replications of a study.

Disentangled representations for counterfactual regression [Hassanpour et al. 2019]

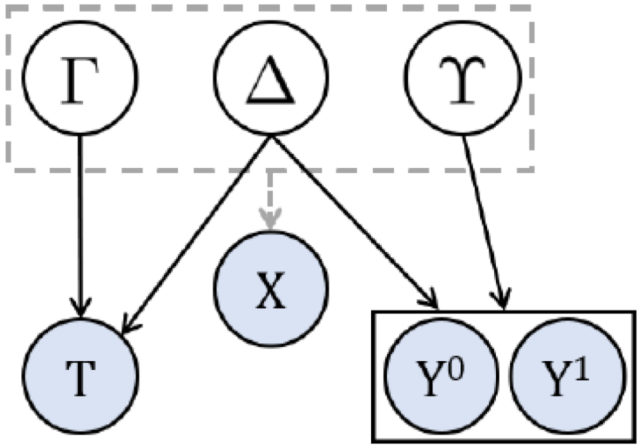
SFU

- Estimation of ITE of a binary treatment
- Outcome binary or continuous
- Propensity score is not sufficient
- Learn a disentangled representation Z consisting of three independent components:
  1) Δ influences T and Y
  2) Γ influences only T
  3) Υ influences only Y

Disentangled representations for counterfactual regression

SFU

- Graphical model



Disentangled representations for counterfactual regression

SFU

- Neural network consists of
  - three representation learning networks

$$\Gamma(x), \Delta(x), \text{and } \Upsilon(x)$$

→ $Z =< \Gamma, \Delta, \Upsilon >$

  - two outcome prediction networks

$$h^0(\Delta(x), \Upsilon(x)) \text{ and } h^1(\Delta(x), \Upsilon(x))$$

  - one treatment prediction network

$$\pi_0\big(t \mid \Gamma(x), \Delta(x)\big)$$

Disentangled representations for counterfactual regression

SFU

- Representation learning loss

  According to the graphical model, $\Upsilon \perp\!\!\!\perp T$

$$\Longrightarrow \quad \Pr(\Upsilon \mid T) = \Pr(\Upsilon) \quad \Longrightarrow \quad \Pr(\Upsilon \mid T=0) = \Pr(\Upsilon \mid T=1)$$

  use Maximum Mean Discrepancy (MMD)
  to calculate dissimilarity between the two
  conditional distributions of Υ given t=0 versus t=1

- Outcome prediction loss
  can only be calculated on the observed outcomes
  L2-loss or log-loss

- Treatment prediction loss
  Cross entropy loss

Disentangled representations for counterfactual regression

SFU

- Overall loss

$$
\begin{aligned}
J(\Gamma, \Delta, \Upsilon, h) \; = \; & \frac{1}{N} \sum_{i=1}^{N} \omega\big(t_i, \Delta(x_i)\big) \cdot \mathcal{L}\big[\, y_i,\, h^{t_i}\big(\Delta(x_i), \Upsilon(x_i)\big)\,\big] \\
& + \alpha \cdot \mathtt{disc}\Big(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\Big) \\
& + \beta \cdot \frac{1}{N} \sum_{i=1}^{N} -\log\big[\, \pi_0\big(t_i \,|\, \Gamma(x_i), \Delta(x_i)\big)\,\big] \\
& + \lambda \cdot \mathfrak{Reg}(h^0, h^1, \pi_0)
\end{aligned}
$$

Causal Effect VAE [Louizos et al., 2017]

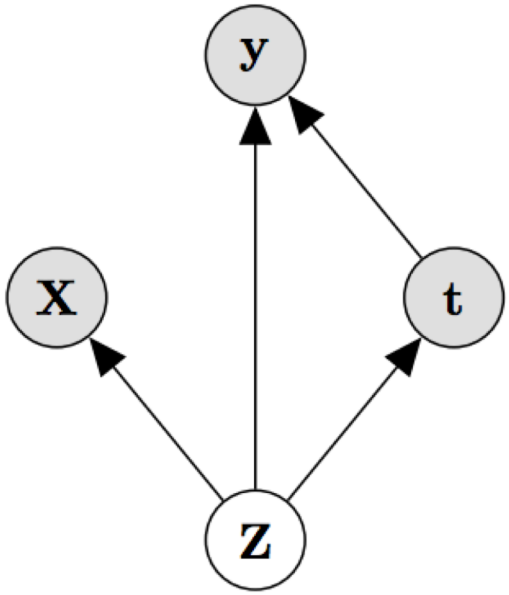
- Estimation of ITE of a binary treatment
- Does not assume unconfoundedness.
- Uses a latent representation *Z* as a proxy for unobserved confounders.
- Uses VAE to learn the latent representation.

--> CEVAE

- VAE make weaker assumptions about the data generating process and the structure of the hidden confounders.

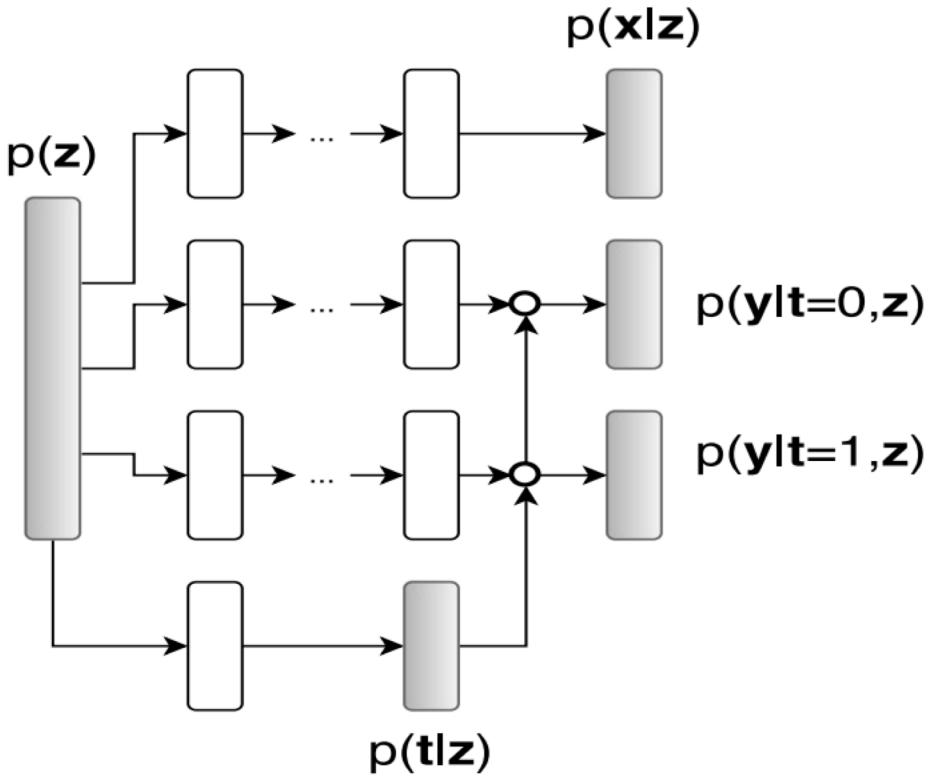
Causal Effect VAE

- Graphical model



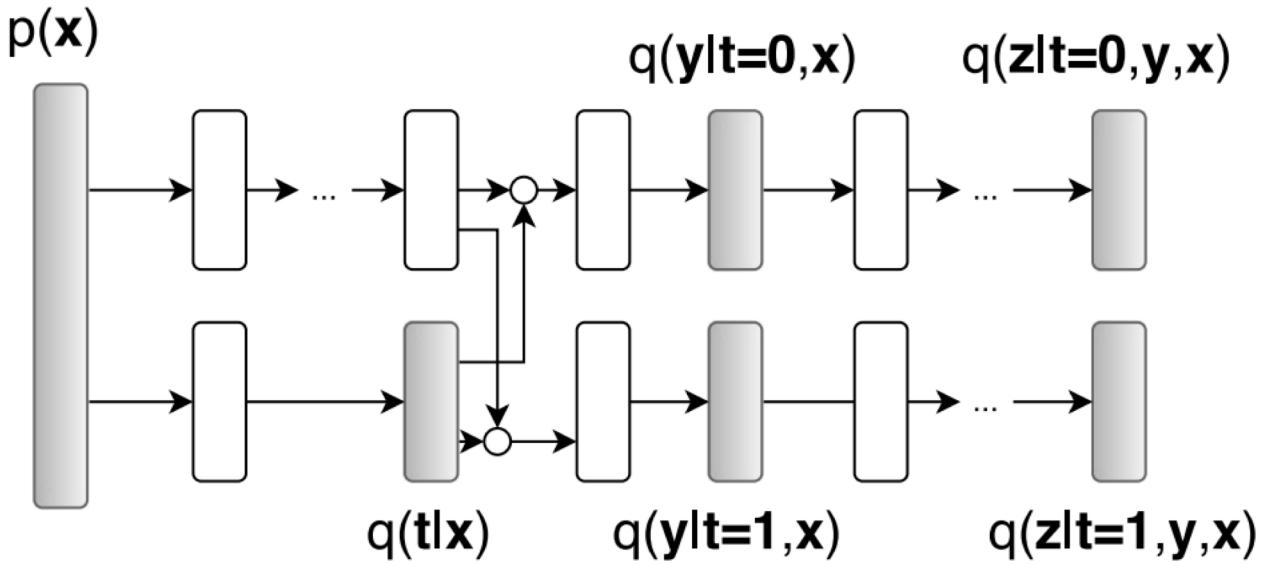
→*Z* is a proxy for unobserved confounders

Causal Effect VAE



Model network
recover $p(\mathbf{x}, \mathbf{z}, t, y)$

Causal Effect VAE



Inference network
infer $p(y, \mathbf{z}|\mathbf{x}, t)$

Directions for Future Research

SFU

- Critical analysis of latent variable models in the presence of unobserved confounders
  [Rissanen et al., 2021]
  Presents counter-examples where CEVAE fails to identify the treatment effect.
- Continuous treatments
  Often treatments can be applied at different doses.
  Not just one treatment (and one control) group but infinitely many treatment groups
  [Schwab et al., 2020]
  [Nie et al., 2021]

Directions for Future Research

SFU

- Multiple treatments
  In many applications, multiple treatments are applied simultaneously.
  Treatments and their effects interact with each other.
  [Egami et al, 2018] [Wang et al., 2019]

References

- Egami, N., & Imai, K. (2018). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*.
- Hassanpour, N., & Greiner, R. (2019, September). Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.
- Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, *30*.
- Nie, L., Ye, M., Liu, Q., & Nicolae, D. (2021). Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*.
- Rissanen, S., & Marttinen, P. (2021). A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, *34*, 4207-4217.

References

- Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., & Karlen, W. (2020, April). Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 5612-5619).
- Shi, C., Blei, D., & Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, *32*.
- Wang, Y., & Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, *114*(528), 1574-1596.