Part 3: Opportunities and Challenges

Hang Wu

Dr. May D. Wang

Dept. Biomedical Engineering, Georgia Tech

Link: https://github.com/hang-wu/CI

Georgia Tech

CREATING THE NEXT

Outline

- Recent Advances
- Challenges and Opportunities

Recent Advances

Georgia Tech
CREATING THE NEXT

Regression

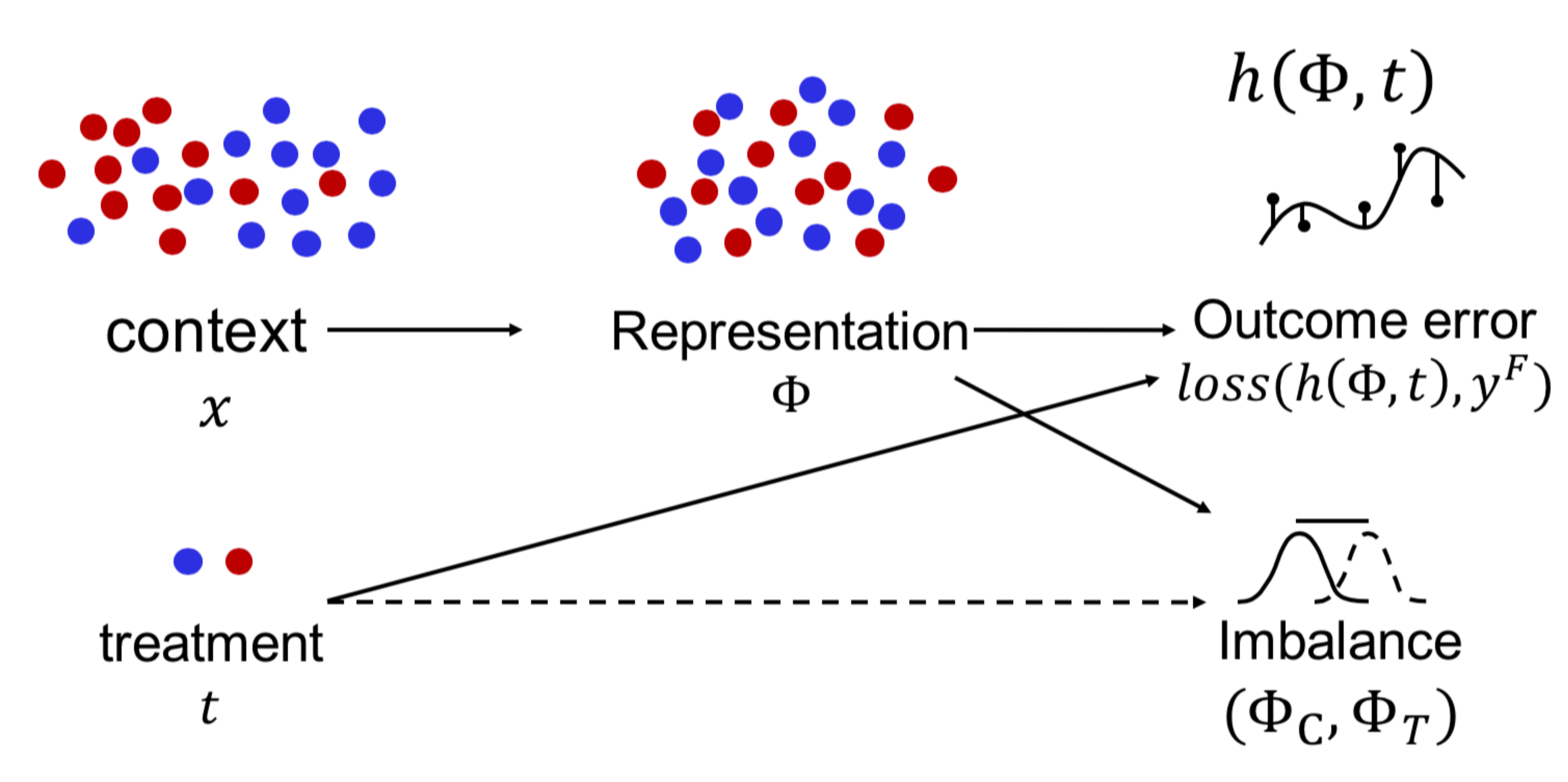
- We build a regression model as $Y \sim X, T$

$$E(Y|X,T) = \alpha_1 X_1 + \alpha_2 X_2 + \cdots \alpha_n X_n + \alpha_T T$$

- Then the causal effect is interpreted as α_T
- Assumptions:
  - No hidden confounding
  - Covariates are not correlated
  - Model correctness: e.g. what if the true model is non-linear?

Georgia Tech
CREATING THE NEXT

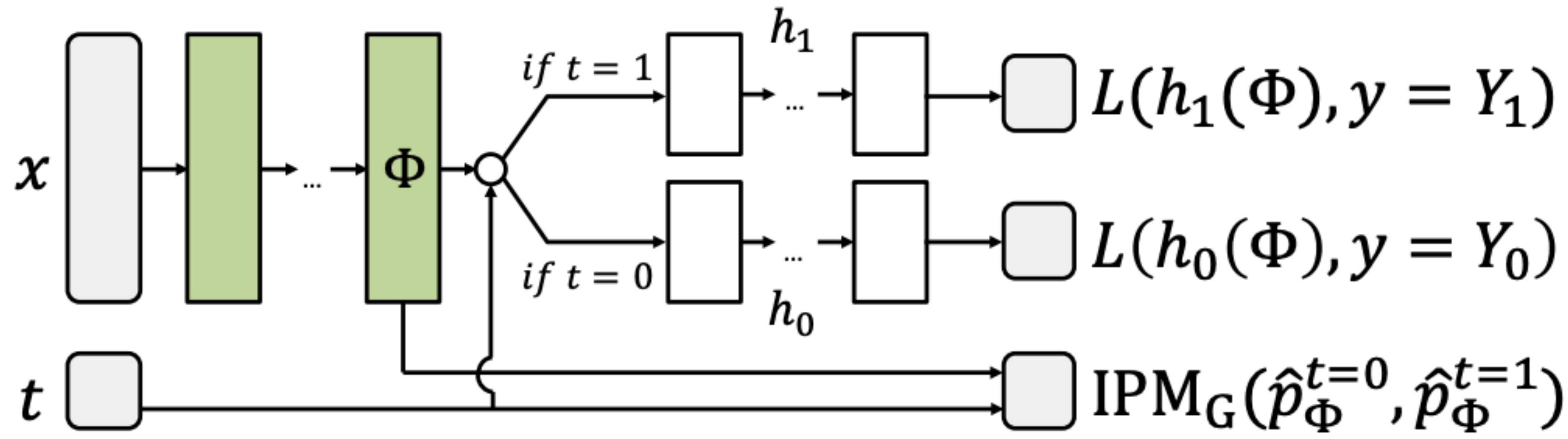
Representation Learning



Johansson, F., Shalit, U., & Sontag, D. (2016, June). Learning representations for counterfactual inference. In *International conference on machine learning* (pp. 3020-3029).

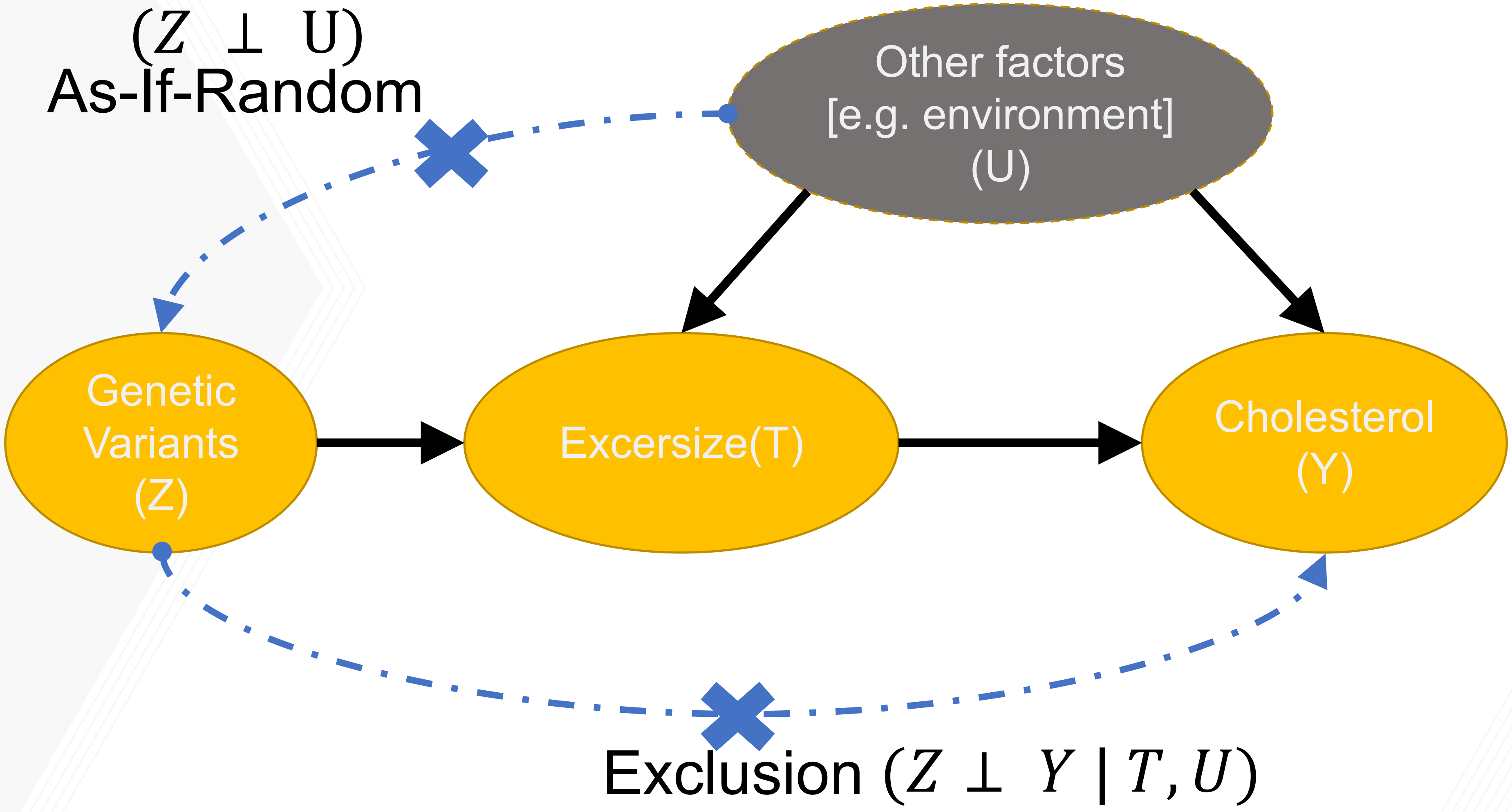
Representation Learning

- We can follow the regression approach
  - Instead of using linear models, fit regression models $y \sim x, t$ using deep neural networks
  - To balance the confounding, add additional



Johansson, F., Shalit, U., & Sontag, D. (2016, June). Learning representations for counterfactual inference. In *International conference on machine learning* (pp. 3020-3029).

Recap: Instrument Variables



Deep Instrument Variables

- Exclusion of instrument variables
  - $E[y|x,z] = \int g(t,x)dF(t|x,z)$
- $g(t,x)$: the causal effect function we want to estimate
  - $g(t=1,x) - g(t=0,x)$
- $E[y|x,z]$ - conditional expectation
  - Can be estimated from the data
- $F(t|x,z)$ – conditional density function
  - Can be estimated from the data

Deep Instrument Variables

$$E[y|x,z] = \int g(t,x)dF(t|x,z)$$

- We can solve an inverse problem

$$\min \sum \left(y_i - \int g(t,x_i)dF(t|x_i,z_i)\right)^2$$

  - Where the $\min$ operator is w.r.t $g(t,x_i)$

Recap: two stage least square model

- We assume linear models
  - $t = \beta z + \epsilon$
  - $g(t, x) = \tau t$
- Thus, $\int g(t, x) dF(t|x, z) = \int \tau t \, dF(t|\cancel{x,} z) = \tau E[t|z]$
- To estimate $E[t|z]$
  - We can solve a linear regression problem
  - $E[t|z] = \hat{\beta} z$
  - Stage 1

Recap: two stage least square model

- Stage 1: regress $t \sim z$ => $\hat{\beta}$
- Stage 2: regress $y \sim \hat{\beta} z$=> τ
- $\min \sum \left(y_i - \int g(t, x_i) dF(t|x_i, z_i) \right)^2 = \min_{\tau} \sum (y_i - \tau \hat{\beta} z)^2$

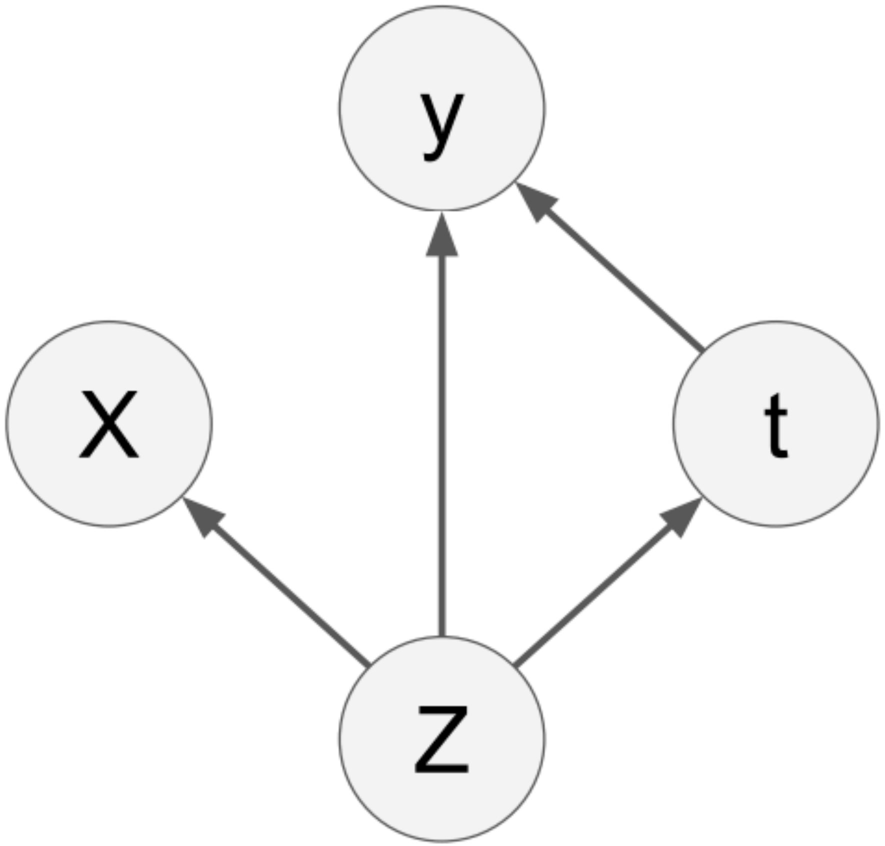
Georgia Tech
CREATING THE NEXT

Deep IV

- $\min \sum \left(y_i - \int g(t, x_i) dF(t|x_i, z_i)\right)^2$
- We can use two neural networks to parametrize $g(t, x)$ and $dF(t|x, z)$
- Now we have two generic supervised machine learning tasks
- And an inverse problem with L2 loss

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. arXiv preprint arXiv:1612.09596.

Latent Variable



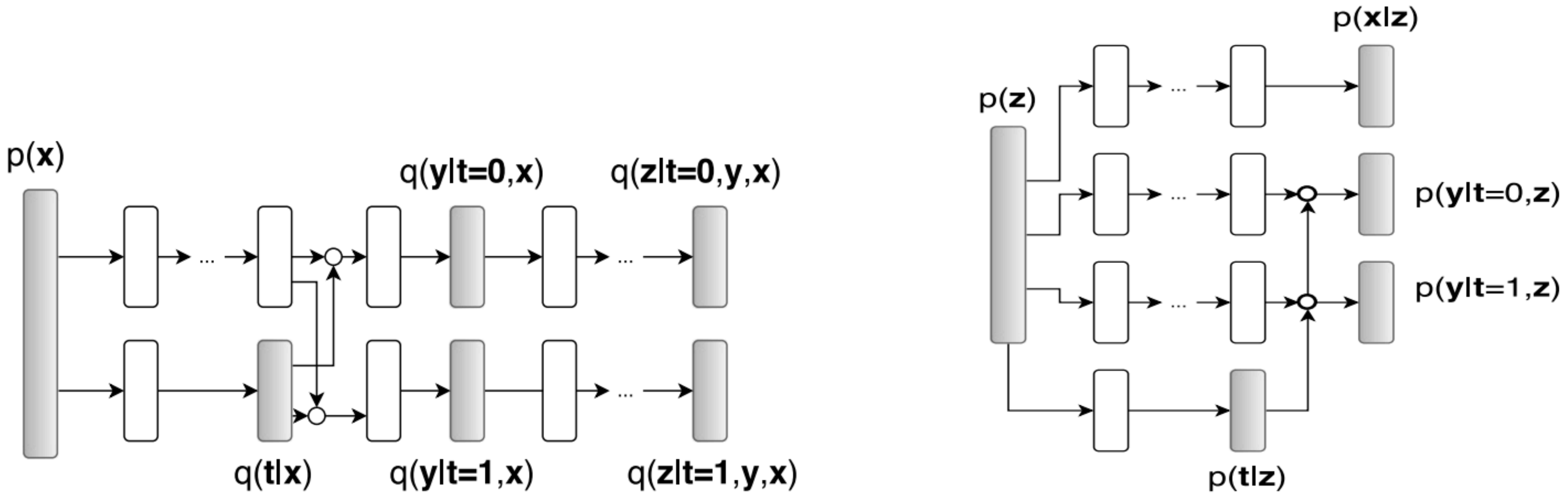
- Notations
  - X: features
  - Y: outcomes
  - T: Treatment
  - Z: hidden variables

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (pp. 6446-6456).

Causal Effect with Latent Variable

- $P(y|X, do(t=1))$
- $= \int_Z P(y, Z|X, do(t=1))dZ$
- $= \int_Z P(y|X, t=1, Z)P(Z|X)dZ$ [using do-calculus]
- $= E_{p(z|x)}[P(y|X, t=1, Z)]$
- Converts to estimation of two conditional distribution estimation
  - $p(z|x)$
  - $P(y|X, t=1, Z)$

Variational Inference for Latent Variable Models



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$.

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

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (pp. 6446-6456).

Key takeaways

- Make appropriate assumptions:
  - Instrumental variables
  - Latent variables
- Under the assumptions, we can turn
  - Causal inference => statistical estimation
- With statistical estimation problem,
  - Deep learning can be applied to improve over linear models

Challenges and Opportunities

Challenges: Counterfactual Nature

- As the we can never observe all the potential outcomes, but one out of the potential outcomes
- How do we evaluate the effectiveness of the proposed algorithm?
- How do we convince people our conclusion is correct?

Opportunities

- Data-side
  - Build benchmark (semi-synthetic) datasets in biomedical data to evaluate different algorithms
- Algorithm-side
  - Develop measures (e.g. statistical tests) or other validations methods to evaluate different algorithms
  - Design expert-in-the-loop algorithms

Challenges: Domain Causal Knowledge Discovery and Integration

- In bio/biomedical research, we have cumulated considerable domain knowledge
  - Protein-Protein Interactions
  - Cell signaling pathway
- How can we integrate knowledge into our algorithm for
  - Causal effect estimation
  - Causal effect estimation algorithm validation

Challenges and Opportunities: Heterogeneity

- Recent research shows that learning subject-specific causal models can improve the biological relevance and the interpretability of models
- Common assumptions to start with:
  - People with similar clinical observations should have similar underlying causal models
- Can we design algorithms that can address the heterogeneity of causal models among individuals/ populations?

Challenges: Dynamics

- Causal models can also change over time.
- So far, we have only discussed the case for static models
- What if patients' characteristics change over time and the treatment effect model is also changing?

Opportunities: Dynamics of Causal Inference

- Design dynamic causal modeling
  - E.g. dynamic Bayesian networks
  - E.g. dynamic causal models
  - E.g. Markov models

- Study causality for time series data
  - Granger causality

Challenges: Heterogenous Treatment Effect

- We have focused most on average treatment effect
- $E_i[Y_i(1) - Y_i(0)]$
- Sometimes, it might be more helpful to identify for unit i
    - $Y_i(1) - Y_i(0)$
    - So that we can make more personalized treatment recommendations

Opportunities: Personalized Medicine

- Understanding subject-specific causal models can help design tailored medical treatment to individuals
  - For example, if we know what genes trigger the symptoms
- Understanding causal models can help guide drug discovery

Opportunities: Causal Inference and Data Integration

- Different datasets contain dataset specific biases, such as confounding, selection bias, and cross-population bias
- Causal inference provides a general framework for study how we can extract invariant information from different datasets

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345-7352.

Opportunities: Causal Inference and Machine Learning

- Conventional supervised learning focus on predictive models
  - The learned models can exploit spurious correlations/ nuisance factors
- How to extract causal relationships for predictions
  - The benefit is the model will be more robust to changing environments

Resources - Softwares

- DoWhy (Structural causal models)
  - https://github.com/Microsoft/dowhy
- EconML (Potential Outcome)
  - https://github.com/microsoft/EconML
- CausalML (Potential Outcome)
  - https://github.com/uber/causalml

Resources - Dataset

- Common benchmark dataset
  - Infant Health and Development Program (IHDP):https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP
  - Job training: https://rugg2.github.io/Lalonde%20dataset%20-%20Causal%20Inference.html
  - Causality repo for model discovery: http://www.causality.inf.ethz.ch/repository.php
  - ACIC Causal Challenge: https://sites.google.com/view/ACIC2019DataChallenge/data-challenge

Discussions

https://github.com/hang-wu/CI