CSDS 452: Causality and Machine Learning

Lecture 10: Causal Effect Estimation with Machine Learning/Neural Network (2)

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Some useful resources for course projects

- EconML:
  - a Python package that applies ML for causal effect learning
  - https://econml.azurewebsites.net/
- Dowhy
  - a Python library for causal inference that supports explicit modeling and testing of causal assumptions.
  - based on a unified language for causal inference, combining causal graphical models and potential outcomes frameworks.
  - https://github.com/py-why/dowhy

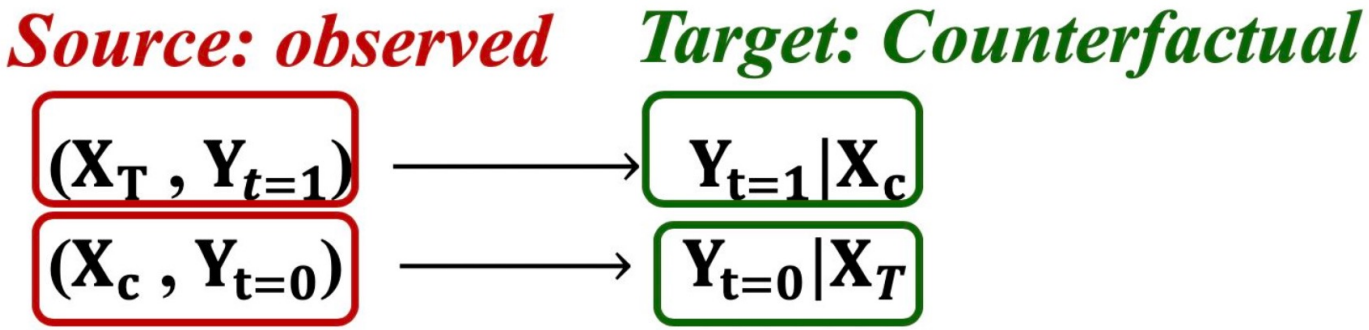
Outline

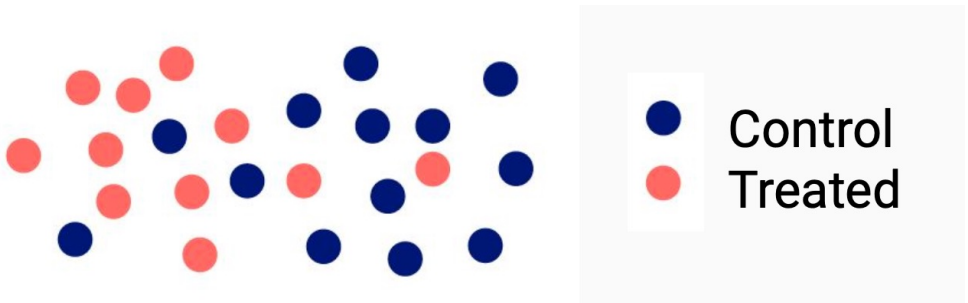
- Causal effect with NN:
  - Dragonnet
  - SITE
- Multi-treatment effect:
  - Deconfounding in Multiple Treatments

Why use NN for Causal Inference?

- In practice, neural networks are low-bias estimators
- Can approximate complex functions to estimate heterogeneous treatment effects
- Creative architectures might help with overlap issues
- Feature-extraction/de-confounding through representation learning
- Has potential to handle graphs, images, text, temporal data, …

Recap: Counterfactual inference & Domain adaptation





Recap: BNN, TARNet, and CFR

- **[Shalit et al., 2017] CFR** framework splits the net to outcomes.
- Balancing neural network (**BNN**) [Johansson et al., 2016] may underestimate the treatment effect (shrinkage estimation) in the outcome net due to regularization bias.

[Johansson et al., 2016]

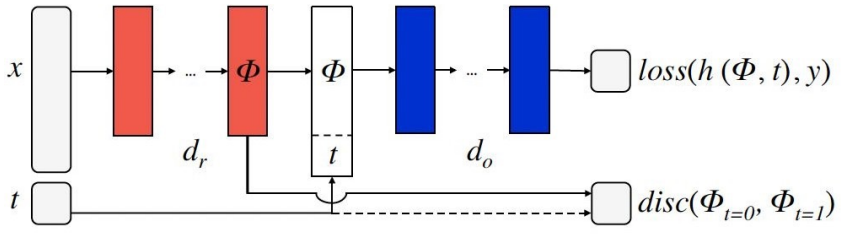


Figure 2. Neural network architecture.

[Shalit et al., 2017]

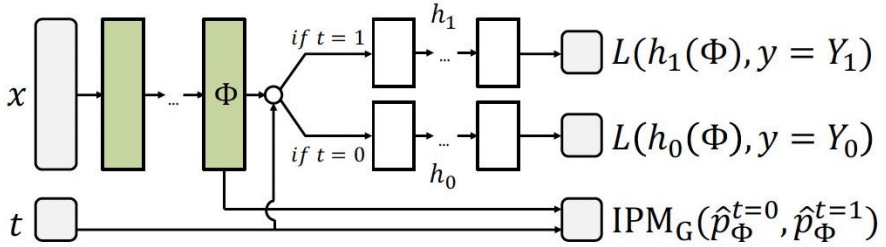


Figure 1. Neural network architecture for ITE estimation. L is a loss function, IPM_G is an integral probability metric. Note that only one of h_0 and h_1 is updated for each sample during training.

[Johansson et al., 2016] Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." International conference on machine learning. PMLR, 2016.

[Shalit et al., 2017] Shalit, Uri, Fredrik D. Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." International Conference on Machine Learning. PMLR, 2017.

CFR Conclusion

- Provide theoretical strong bounds on bias

$$\epsilon_{CF}(h,\Phi) \leq$$
$$(1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi)$$
$$+ B_\Phi \cdot IPM_G\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right),$$

- Strong empirical performance, slightly better than TARNet (No IPM)
- No consistency guarantees
  - This issue is addressed with a weighted version in follow-up work

Outline

- Causal effect with NN:
  - Dragonnet
  - SITE
- Multi-treatment effect:
  - Deconfounding in Multiple Treatments

Dragonnet

- Exploit the sufficiency of the propensity score for estimation adjustment

Theorem 2.1 (Sufficiency of Propensity Score). *If the average treatment effect ψ is identifiable from observational data by adjusting for X, i.e., $\psi = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1] - \mathbb{E}[Y \mid X, T = 0]]$, then adjusting for the propensity score also suffices:*

$$\psi = \mathbb{E}[\mathbb{E}[Y \mid g(X), T = 1] - \mathbb{E}[Y \mid g(X), T = 0]]$$

Shi C, Blei D, Veitch V. Adapting neural networks for the estimation of treatment effects[J]. NeurIPS, 2019, 32.

Dragonnet

- Add a head that predicts propensity score π
- Single neural forces representation to also tightly couple to propensity score

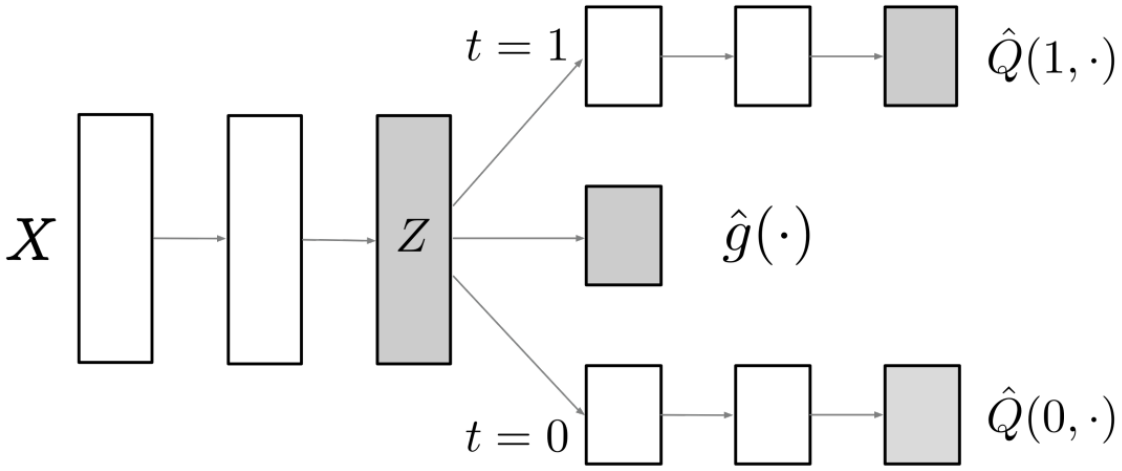


Figure 1: Dragonnet architecture.

Dragonnet

- Multiple objectives:
  - **Dragonnet**: Minimize outcome/propensity loss

$$\hat{R}(\theta; \boldsymbol{X}) = \frac{1}{n} \sum_i \underbrace{\left[(Q^{\mathrm{nn}}(t_i, x_i; \theta) - y_i)^2\right.}_{\text{Outcome prediction loss}} + \underbrace{\left.\alpha \mathrm{CrossEntropy}(g^{\mathrm{nn}}(x_i; \theta), t_i)\right]}_{\text{treatment prediction loss}}$$

[1] M. van der Laan and S. Rose. Targeted Learning: Causal Inference for Observational and Experimental Data. 2011.

Dragonnet

- Multiple objectives:
  - **Dragonnet**: Minimize outcome/propensity loss

$$\hat{R}(\theta;\boldsymbol{X}) = \frac{1}{n}\sum_i \underbrace{\left[(Q^{\mathrm{nn}}(t_i,x_i;\theta)-y_i)^2\right]}_{\text{Outcome prediction loss}} + \underbrace{\alpha\,\mathrm{CrossEntropy}(g^{\mathrm{nn}}(x_i;\theta),t_i)}_{\text{treatment prediction loss}}$$

  - Inspired by TMLE [1], which introduces an extra model parameter ε to perturb outcome prediction
  - **Targeted regularization**: Minimize (outcome + ε *propensity) against outcome

$$\tilde{Q}(t_i,x_i;\theta,\varepsilon) = Q^{\mathrm{nn}}(t_i,x_i;\theta) + \varepsilon\left[\frac{t_i}{g^{\mathrm{nn}}(x_i;\theta)} - \frac{1-t_i}{1-g^{\mathrm{nn}}(x_i;\theta)}\right]$$

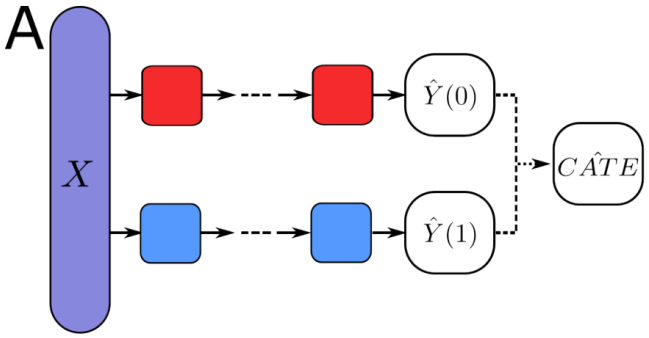
$$\gamma(y_i,t_i,x_i;\theta,\varepsilon) = (y_i-\tilde{Q}(t_i,x_i;\theta,\varepsilon))^2.$$

$$\hat{\theta},\hat{\varepsilon} = \underset{\theta,\varepsilon}{\mathrm{argmin}}\left[\hat{R}(\theta;\boldsymbol{X}) + \beta\frac{1}{n}\sum_i \gamma(y_i,t_i,x_i;\theta,\varepsilon)\right]$$

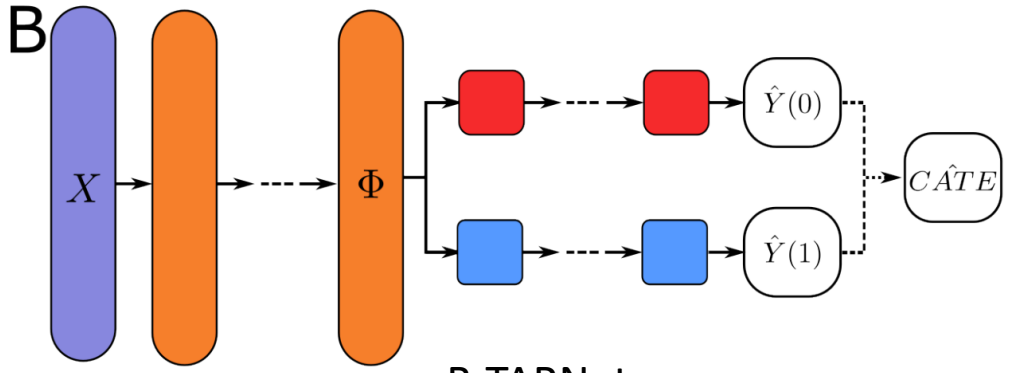
We can calculate a "doubly robust" estimate -- The effect estimate is consistent if either the modified outcome model or the propensity score is consistent

[1] M. van der Laan and S. Rose. Targeted Learning: Causal Inference for Observational and Experimental Data. 2011.

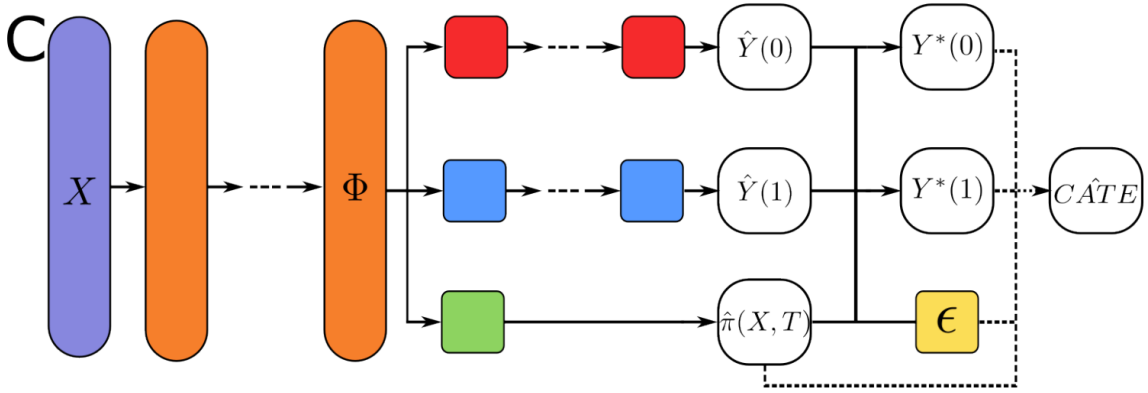
T-Learner, TARNet, and Draggonnet



A: T-Learner



B:TARNet



C: Dragonnet

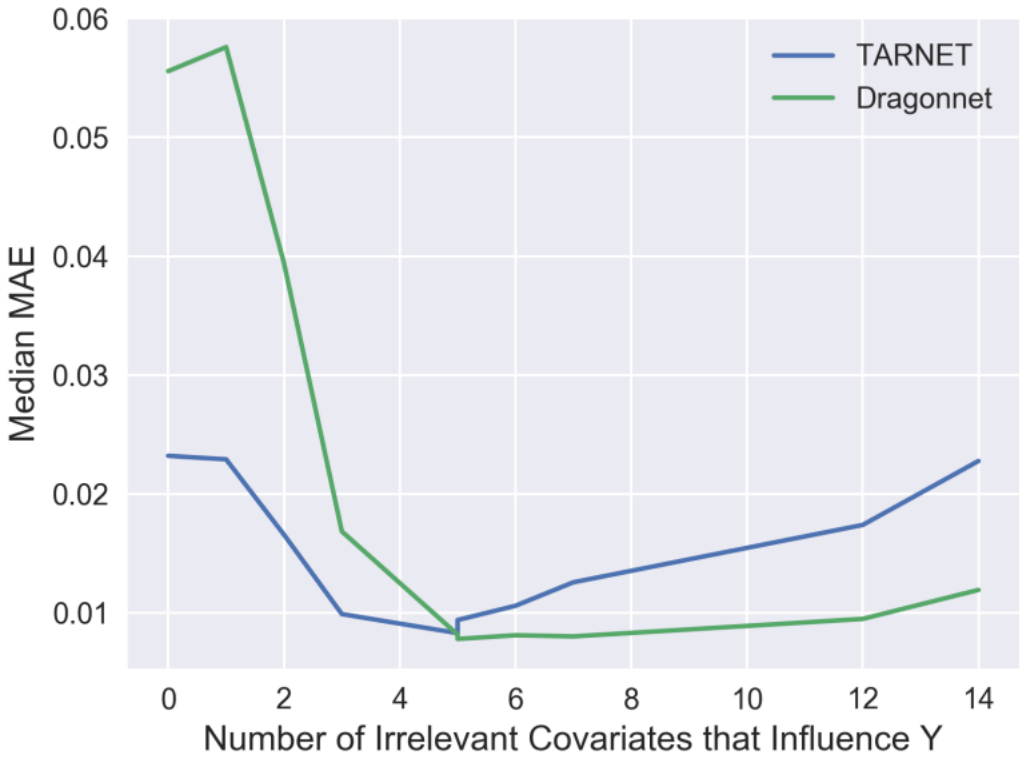
Ramachandra V. Deep learning for causal inference[J]. arXiv preprint arXiv:1803.00149, 2018.

Experiments of Dragonnet

Table 1: Dragonnet with targeted regularization is state-of-the-art among neural network methods on the IHDP benchmark data. Entries are mean absolute error (and standard error) across simulations. Estimators are computed with the training and validation data (Δ_{in}), heldout data (Δ_{out}), and all data (Δ_{all}). Note that using all the data for both training and estimation improves estimation relative to data splitting. Values from previous work are as reported in the cited papers.

Method	Δ_{in}	Δ_{out}	Δ_{all}
BNN [JSS16]	$0.37 \pm .03$	$0.42 \pm .03$	—
TARNET [SJS16]	$0.26 \pm .01$	$0.28 \pm .01$	—
CFR Wass[SJS16]	$0.25 \pm .01$	$0.27 \pm .01$	—
CEVAEs [Lou+17]	$0.34 \pm .01$	$0.46 \pm .02$	—
GANITE [YJS18]	$0.43 \pm .05$	$0.49 \pm .05$	—
baseline (TARNET)	$0.16 \pm .01$	$0.21 \pm .01$	$0.13 \pm .00$
baseline + t-reg	$0.15 \pm .01$	$0.20 \pm .01$	$0.12 \pm .00$
Dragonnet	$0.14 \pm .01$	$0.21 \pm .01$	$0.12 \pm .00$
Dragonnet + t-reg	$0.14 \pm .01$	$0.20 \pm .01$	$0.11 \pm .00$

Experiments of Dragonnet



Dragonnet improves over the baseline if many covariates are irrelevant for treatment.

Dragonnet Conclusions

- Excellent estimation of ATE in simulation
- Excellent asymptotic properties
- Asymptotic guarantees for ATE (not CATE)
- No simulation results reported for CATE

Outline

- Causal effect with NN:
  - Dragonnet
  - SITE
- Multi-treatment effect:
  - Deconfounding in Multiple Treatments

Integral probability metric (IPM)

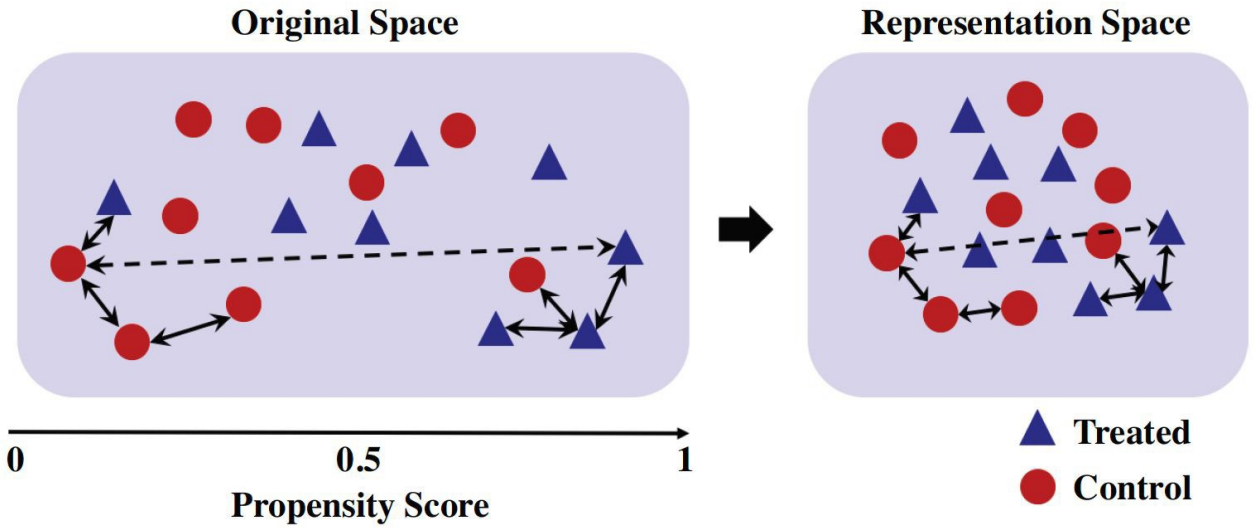
- Symmetric metric between probability distributions
- CFR uses two different metrics
  - Maximum Mean Discrepancy (MMD)
  - Wasserstein distance

Local similarity preserving based methods

- Limitation of most representation balancing methods: lost of local similarity
- Motivation: The latent space should encode:
  - The distribution in latent space is balanced
  - The similarity order information in X (because the performance of KNN is good)
  - For different data points, the strength of similarity should be different

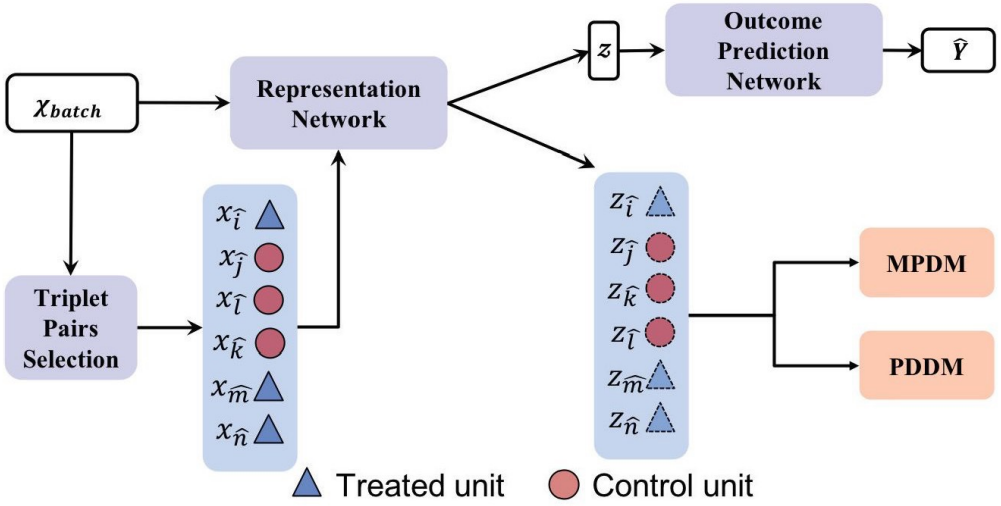
Local similarity preserving based methods

- A toy example



SITE

- Idea: Using triplet loss to preserve the local similarity



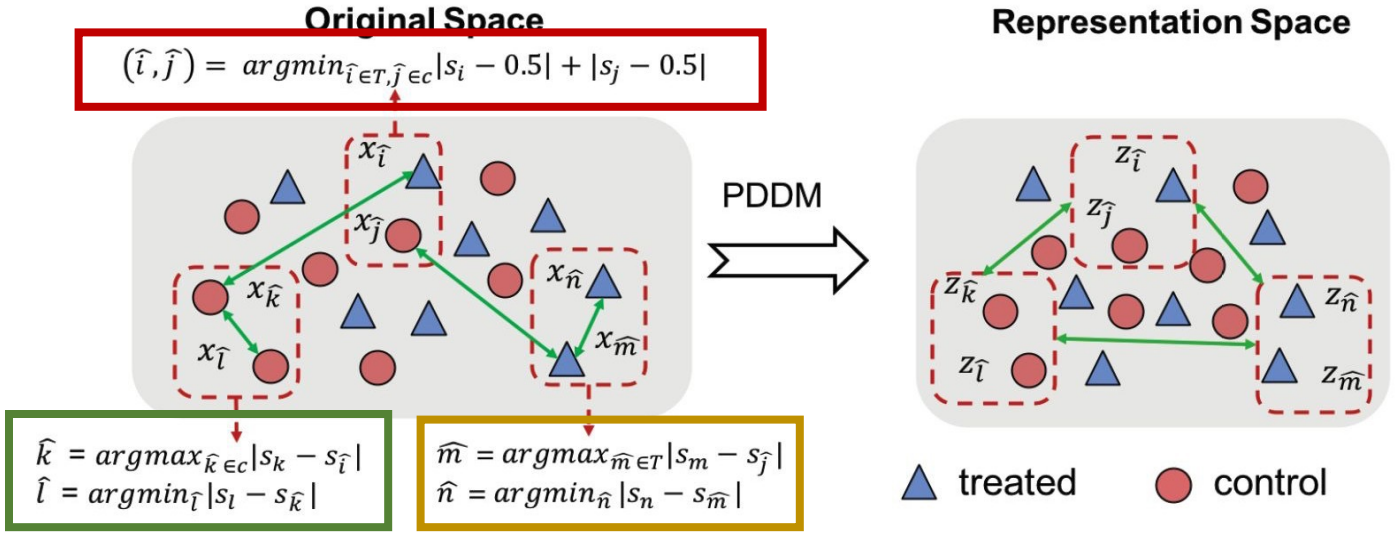
- Objective Function

$$\mathcal{L} = \mathcal{L}_{\text{FL}} + \beta\mathcal{L}_{\text{PDDM}} + \gamma\mathcal{L}_{\text{MPDM}} + \lambda||W||_2$$

L. Yao, et al. "Representation learning for treatment effect estimation from observational data." *NeurIPS* 2018.

SITE

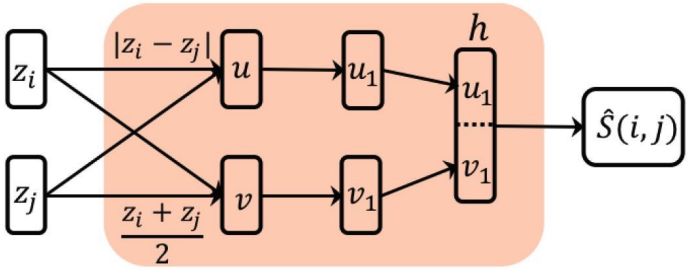
- Triplet pair selection: Closest to intermediate region; farthest control units; farthest treated units.



- s_i is the propensity score, which is the probability that a unit is in the treated group
- Propensity score can reflect the relative location of units in the original space

SITE

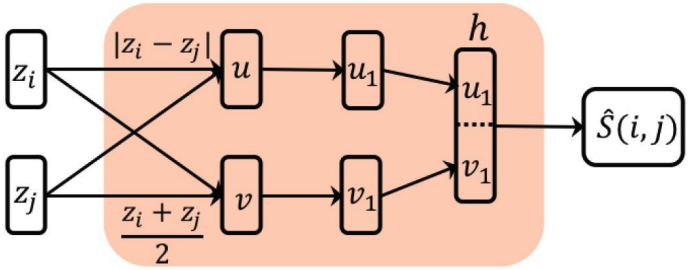
- Position-Dependent Deep Metric (PDDM):
  - The PDDM component measures the local similarity of two units based on their relative and absolute positions in the latent space



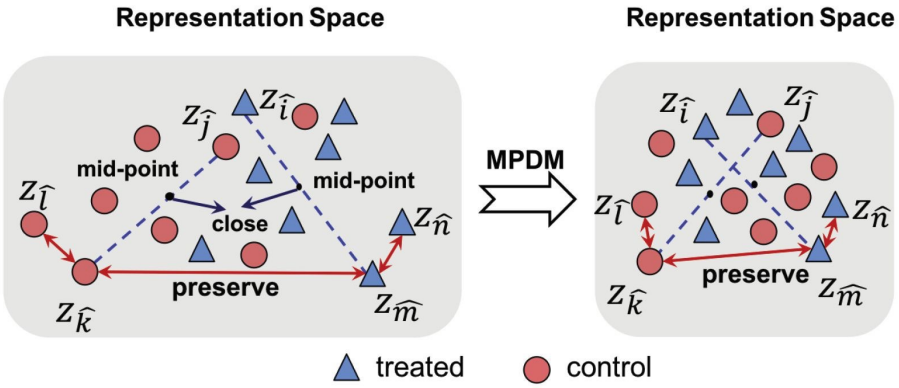
$$\mathcal{L}_{\text{PDDM}} = \frac{1}{5}\sum_{\hat{i},\hat{j},\hat{k},\hat{l},\hat{m},\hat{n}} \left[(\hat{S}(\hat{k},\hat{l}) - S(\hat{k},\hat{l}))^2 + (\hat{S}(\hat{m},\hat{n}) - S(\hat{m},\hat{n}))^2 + (\hat{S}(\hat{k},\hat{m}) - S(\hat{k},\hat{m}))^2 + (\hat{S}(\hat{i},\hat{m}) - S(\hat{i},\hat{m}))^2 + (\hat{S}(\hat{j},\hat{k}) - S(\hat{j},\hat{l}))^2\right],$$

SITE

- Position-Dependent Deep Metric (PDDM):
  - The PDDM component measures the local similarity of two units based on their relative and absolute positions in the latent space



- Middle Point Distance Minimization (MPDM):
  - Makes two mid-points close to each other
  - The MPDM balances the distribution in the latent space



$$\mathcal{L}_{\text{MPDM}} = \sum_{\hat{i},\hat{j},\hat{k},\hat{m}} \left(\frac{\mathbf{z}_{\hat{i}} + \mathbf{z}_{\hat{m}}}{2} - \frac{\mathbf{z}_{\hat{j}} + \mathbf{z}_{\hat{k}}}{2} \right)^2$$

SITE: Experiments

Table 1: Performance comparison on IHDP and Jobs Dataset.

	IHDP ($\sqrt{\epsilon_{\text{PEHE}}}$)		**Jobs** ($\mathcal{R}_{\text{pol}}$)	
Method	Within-sample	Out-of-sample	Within-sample	Out-of-sample
OLS/LR$_1$	10.761 ± 4.350	7.345 ± 2.914	0.310 ± 0.017	0.279 ± 0.067
OLS/LR$_2$	10.280 ± 3.794	5.245 ± 0.986	0.228 ± 0.012	0.733 ± 0.103
HSIC-NNM [5]	2.439 ± 0.445	2.401 ± 0.367	0.291 ± 0.019	0.311 ± 0.069
PSM [27]	7.188 ± 2.679	7.290 ± 3.389	0.292 ± 0.019	0.307 ± 0.053
k-NN [8]	4.432 ± 2.345	4.303 ± 2.077	0.230 ± 0.016	0.262 ± 0.038
Causal Forest [33]	4.732 ± 2.974	4.095 ± 2.528	0.232 ± 0.018	0.224 ± 0.034
BNN [18]	3.827 ± 2.044	4.874 ± 2.850	0.232 ± 0.008	0.240 ± 0.012
TARNet [30]	0.729 ± 0.088	1.342 ± 0.597	0.228 ± 0.004	0.234 ± 0.012
CFR-MMD [30]	0.663 ± 0.068	1.202 ± 0.550	$\mathbf{0.213 \pm 0.006}$	0.231 ± 0.009
CFR-WASS [30]	0.649 ± 0.089	1.152 ± 0.527	0.225 ± 0.004	0.225 ± 0.010
SITE (Ours)	$\mathbf{0.604 \pm 0.093}$	$\mathbf{0.656 \pm 0.108}$	0.224 ± 0.004	$\mathbf{0.219 \pm 0.009}$

Experiment on IHDP and Jobs dataset

Outline

- Causal effect with NN:
  - Dragonnet
  - SITE
- Multi-treatment effect:
  - Deconfounding in Multiple Treatments

A frivolous causal inference problem

- Data about movies: casts and revenue
- Goal: Understand the **causal effect** of putting an actor in a movie
- Causal: “What will the revenue be if we make a movie with a particular cast?”

The naive solution

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ... }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ... }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ... }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, ... }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ... }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ... }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ... }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ... }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ... }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, ... }	$1153M
⋮	⋮	⋮

- Naive solution: Fit a regression (or use deep learning)
- Actors are features; revenue is the response
- Estimates revenue as a function of which actors are cast

The naive solution

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ... }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ... }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ... }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, ... }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ... }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ... }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ... }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ... }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ... }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, ... }	$1153M
⋮	⋮	⋮

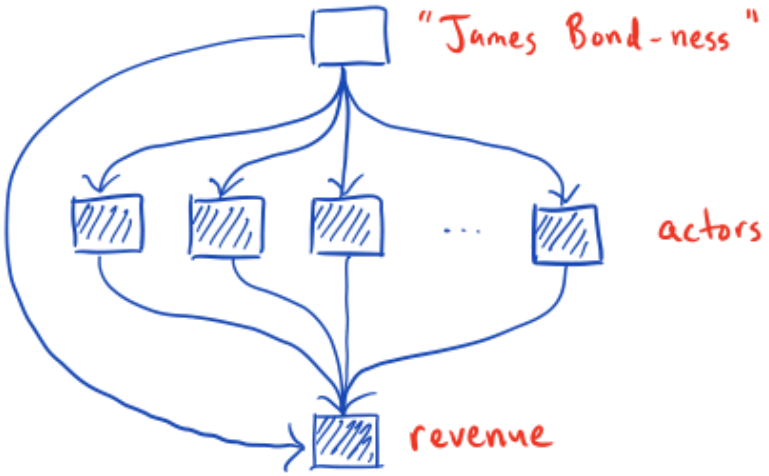
- But standard ML does not (necessarily) provide causal inferences
- Whether an *actor was cast* is different from *casting an actor*
- Causal inference is about **prediction under intervention**

[Hernan and Robins 2019; Imbens and Rubin 2015; Pearl 2009]


Metro Goldwyn Mayer
TRADE
MARK

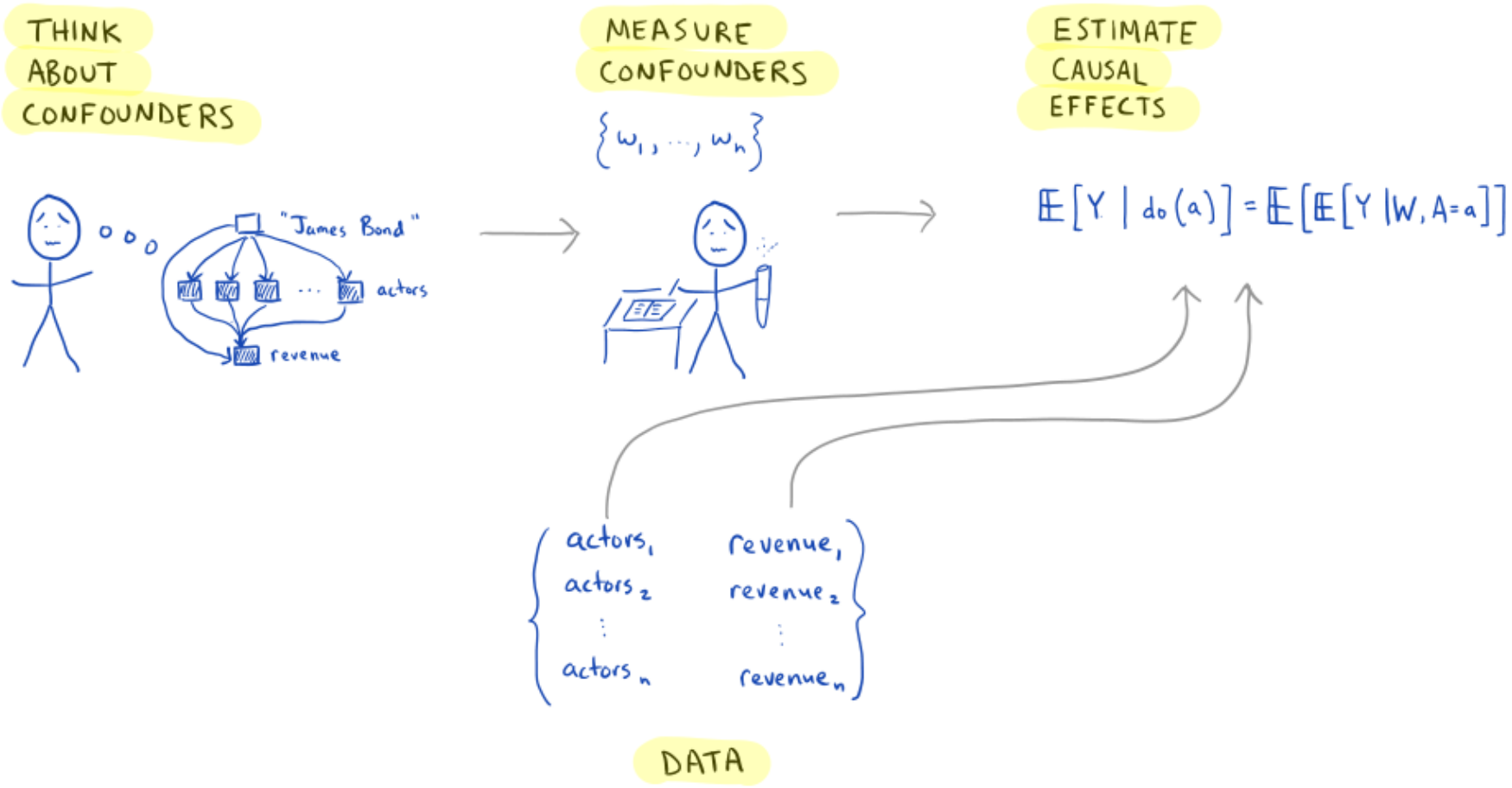

- James Bond movies are about James Bond, a British spy
- Cast James Bond, M, Q, Ms. Moneypenny
- M, Q, Ms Moneypenny only appear in Bond movies
- Bond movies always do well at the box office

The naive solution

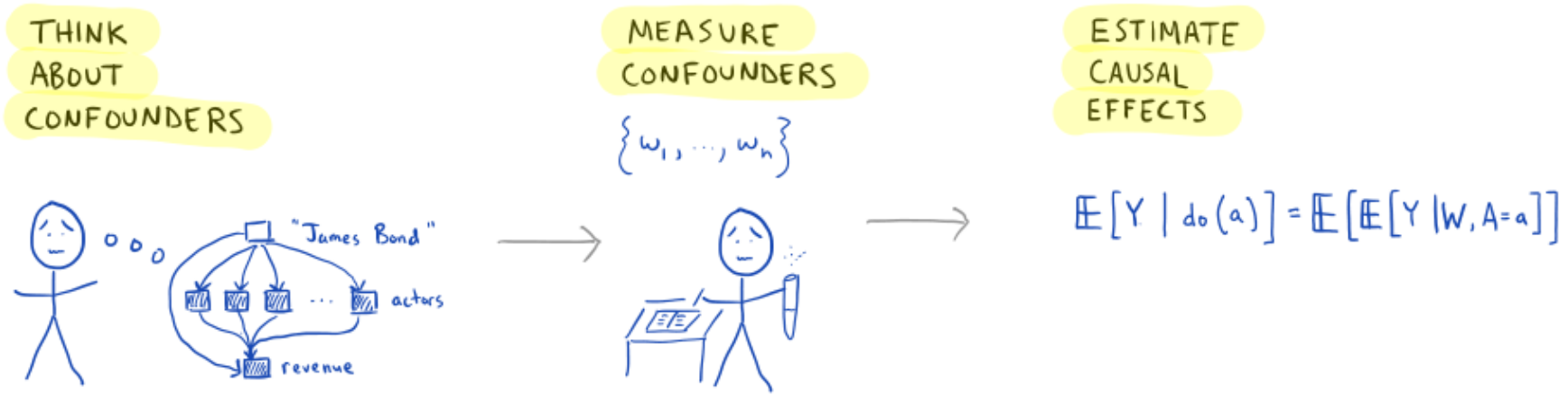


- James Bond-ness is an **unobserved confounder**.
- Confounders affect both the cast ("causes") and the revenue ("effect")
- Confounders bias "passive ML," when used to predict interventions.
  - Some actors overestimated; others are underestimated

The classical solution



The classical solution



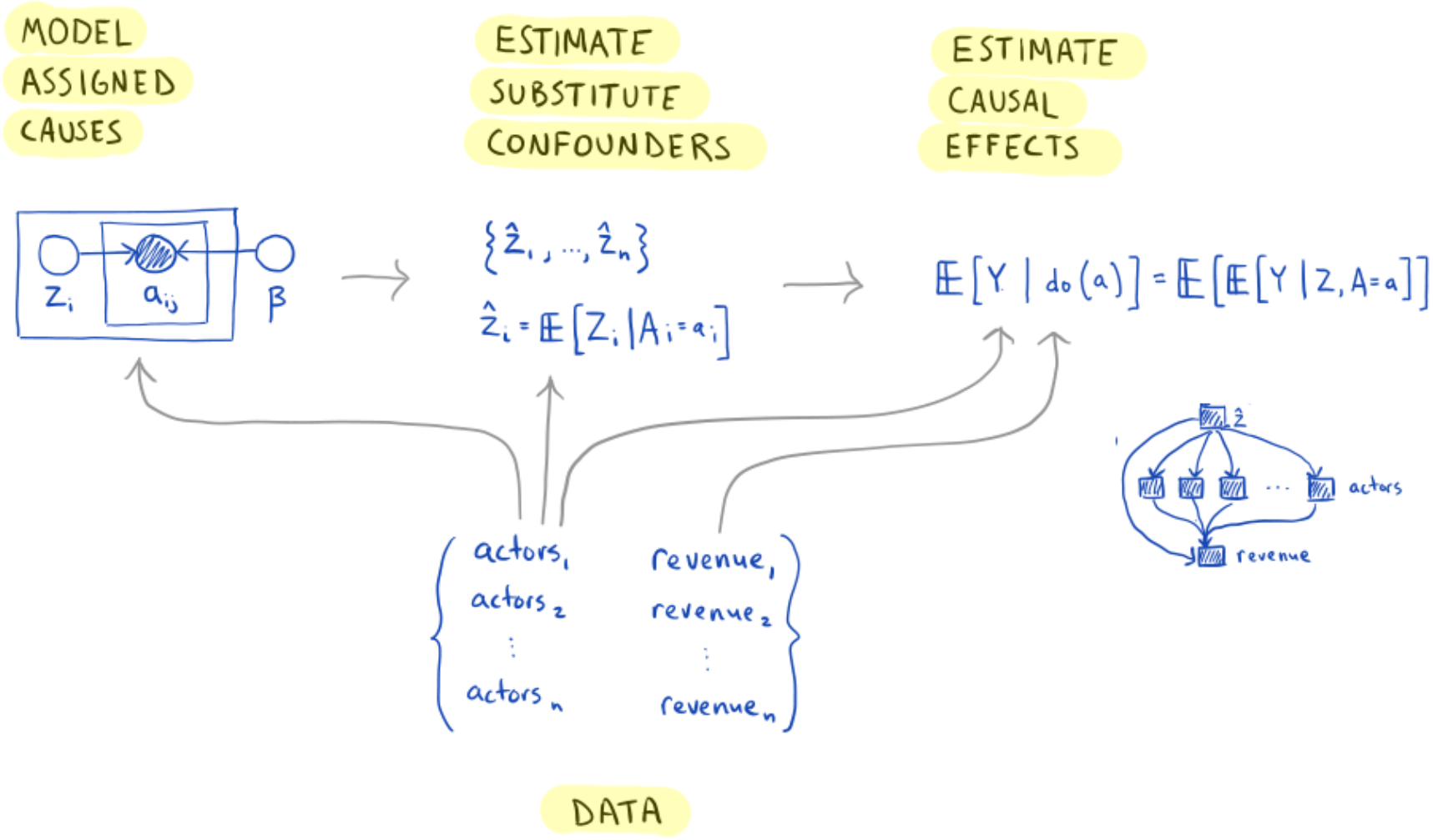
- This approach assumes that we measure **sufficient confounders**.
- But this assumption is **untestable**. [Imbens and Rubin 2015]

Multiple causal inference

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ... }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ... }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ... }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, ... }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ... }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ... }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ... }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ... }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ... }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, ... }	$1153M
⋮	⋮	⋮

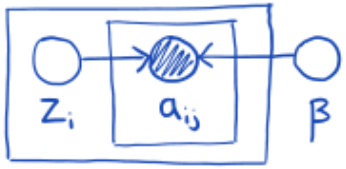
- But our problem is not classical.
- There are many causes (one per actor)—multiple causal inference
- **Multiple causes helps construct a variable that contains confounders.**

The deconfounder



The deconfounder

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \ldots, \hat{z}_n\}$$

$$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$$

ESTIMATE CAUSAL EFFECTS

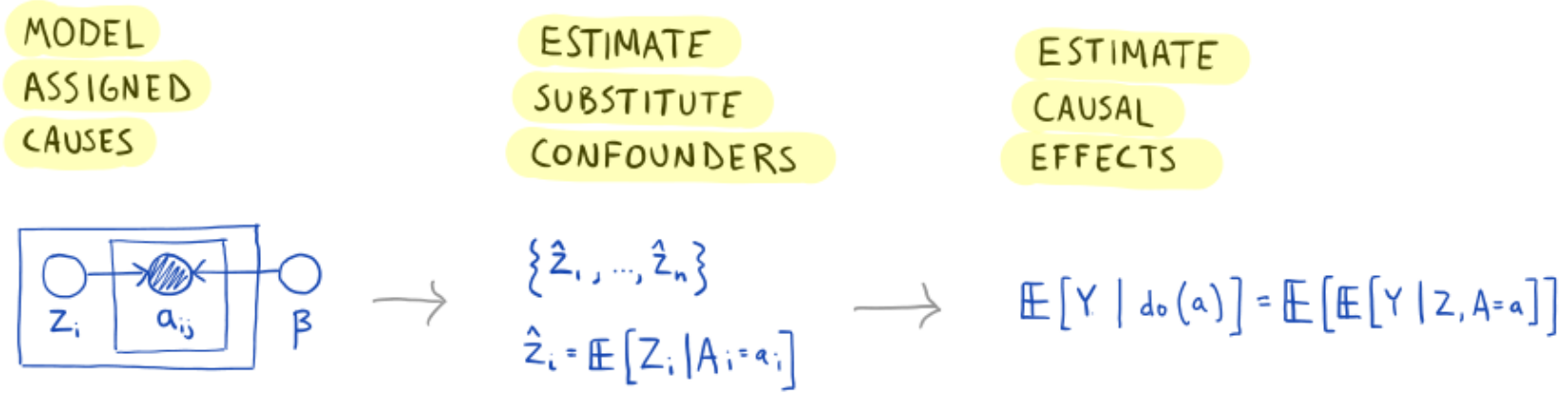
$$\mathbb{E}[Y \mid do(a)] = \mathbb{E}[\mathbb{E}[Y \mid Z, A = a]]$$

- Find, fit, and check a **factor model** of the assigned causes.
- Use the model to form **substitute confounders** for each individual.
- Use the substitute confounders in a **causal model** of the outcome.

(Note: There are still untestable assumptions!)


Metro Goldwyn Mayer
TRADE
MARK

Intuition and assumptions



- Intuition: “Multi-cause confounders” induce dependence among the causes.
- That dependence is encoded in the data; we can capture it with a factor model
- Assume: No unobserved single-cause confounders. (Other assumptions too)

Multiple causal inference (beyond James Bond)

- How do genes affect a trait?
- How do the players affect the game?
- How do prices of items affect how much money is spent?
- How do medicines affect lab measurements?
- How do neurons affect limb movement?

The deconfounder in more detail

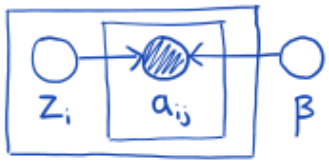
Multiple causal inference

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, … }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, … }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, … }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, … }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, … }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, … }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, … }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, … }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, … }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, … }	$1153M
⋮	⋮	⋮

- Observed dataset $\mathscr{D} = \{(\mathbf{a}_1, y_1), \ldots, (\mathbf{a}_n, y_n)\}$
  - assigned causes $\mathbf{a}_i = \{a_{i1}, \ldots, a_{im}\}$
  - outcome y_i
- Goal: Do causal inference, $\mathbb{E}[Y; \mathrm{do}(\mathbf{a})]$
  - "The expectation of Y in the model where we intervened on $\mathbf{a}$."

The deconfounder

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \ldots, \hat{z}_n\}$$

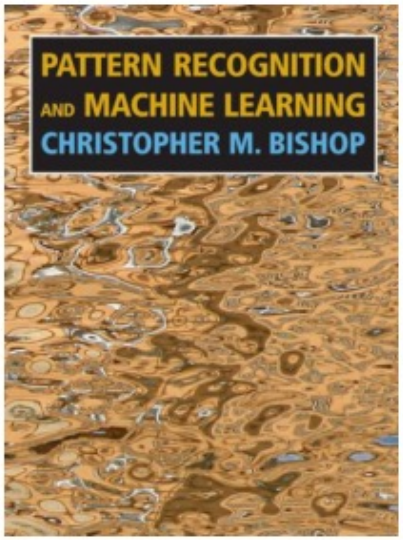
$$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$$

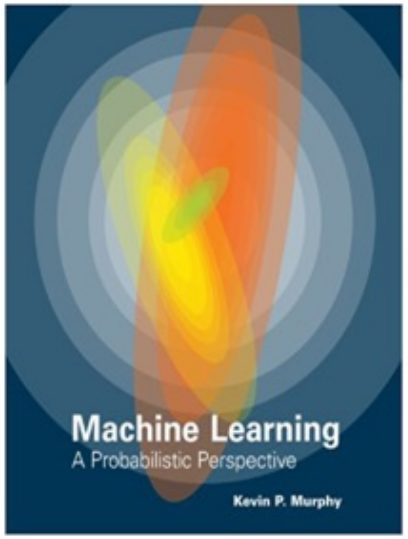
ESTIMATE CAUSAL EFFECTS

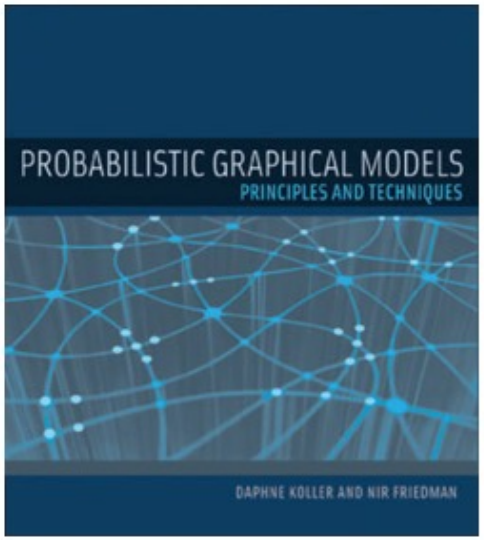
$$\mathbb{E}[Y \mid do(a)] = \mathbb{E}[\mathbb{E}[Y \mid Z, A = a]]$$

- Find, fit, and check a **factor model** of the movie casts.
- Use the factor model to form **substitute confounders** for each movie.
- Use the substitute confounders in a **causal model** of movie revenue.

Fit a probabilistic factor model





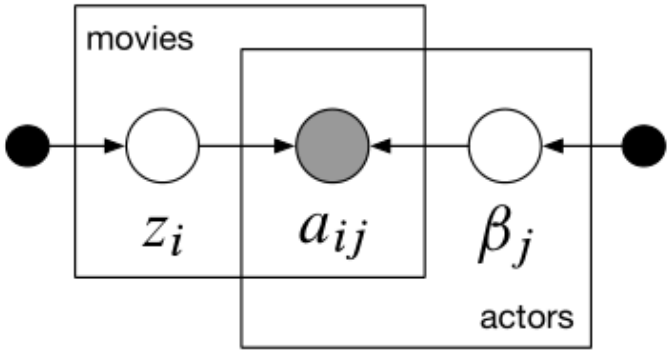


- A probabilistic factor model has the following form,

$$\begin{aligned} \beta_j &\sim \mathrm{p}(\beta_j) && j = 1, \ldots, m \\ z_i &\sim \mathrm{p}(z_i) && i = 1, \ldots, n \\ a_{ij} &\sim \mathrm{p}(a_{ij} \mid z_i, \beta_j). \end{aligned}$$

- E.g., mixtures, matrix factorization, deep generative models, topic models, ...

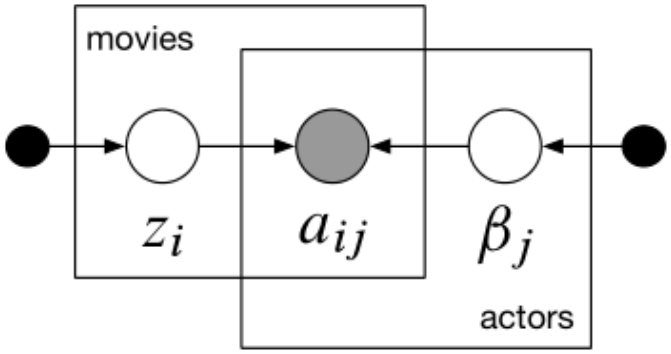
Poisson factorization [Gopalan+ 2015]



$$\beta_{jk} \sim \text{Gam}(a, b) \qquad i \in \{1, \ldots n\}$$
$$z_{ik} \sim \text{Gam}(a, b) \qquad j \in \{1, \ldots, m\}$$
$$a_{ij} \sim \text{Poi}(z_i^\top \beta_j) \qquad k \in \{1, \ldots, d\}$$

- Provides a generative model of the assigned causes a_{ij}.
- Can be approximated on large datasets with variational methods
- A Bayesian form of non-negative matrix factorization [Lee and Seung 1999]

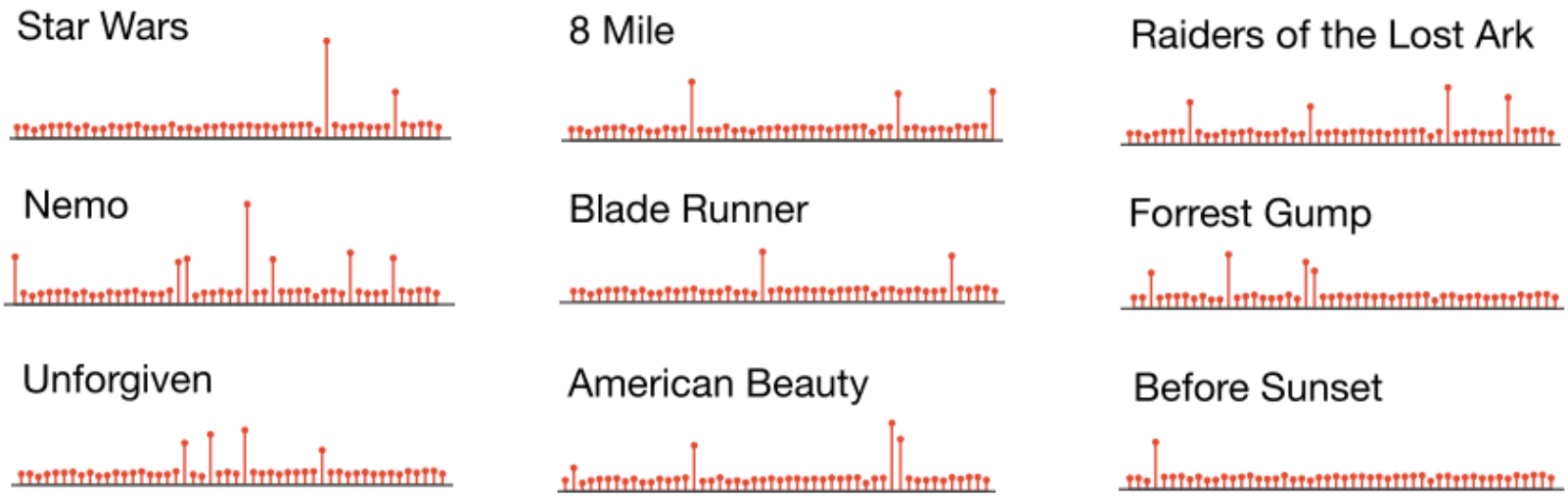
Poisson factorization [Gopalan+ 2015]



$$\beta_{jk} \sim \text{Gam}(a, b) \qquad i \in \{1, \ldots n\}$$
$$z_{ik} \sim \text{Gam}(a, b) \qquad j \in \{1, \ldots, m\}$$
$$a_{ij} \sim \text{Poi}(z_i^\top \beta_j) \qquad k \in \{1, \ldots, d\}$$

- Consider the dataset of casts $\mathbf{a}_{1:n}$.
- Approximate the posterior distribution $\text{p}(z_{1:n}, \beta_{1:m} \mid \mathbf{a}_{1:n})$.
- **We only model the actors $\mathbf{a}_i$; the outcome is not involved.**

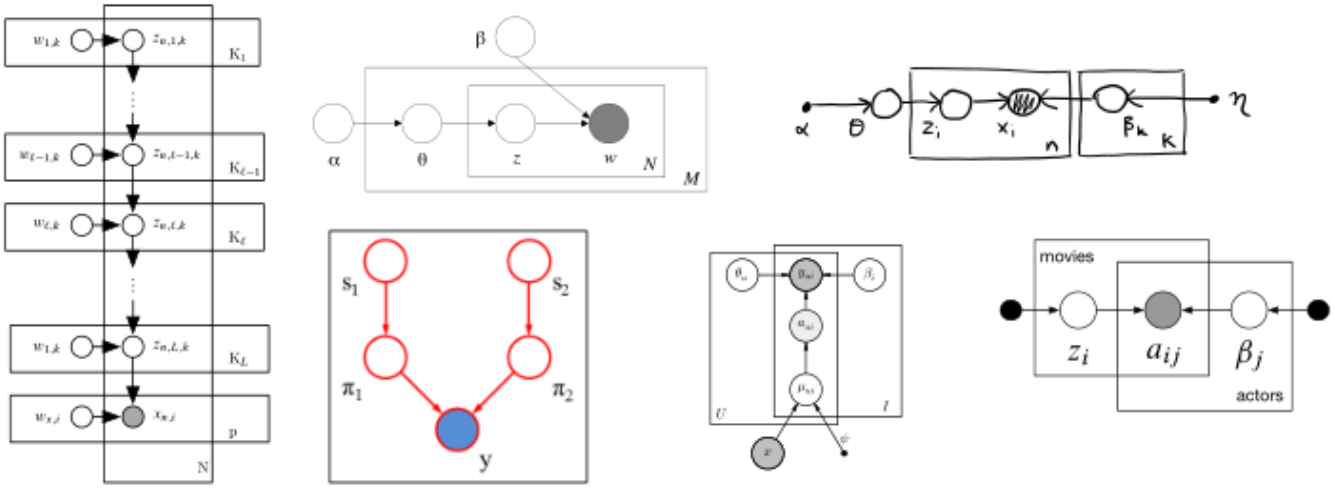
Check the factor model



- We want the **learned representation** to capture the distribution of actors.
- Estimate $\hat{z}_i = \mathbb{E}_{\text{model}}[Z \mid \mathbf{a}_i, \boldsymbol{\beta}]$. (Approximate inference is OK.)
- Check how well $\hat{z}_i$ captures the true distribution of the actors.

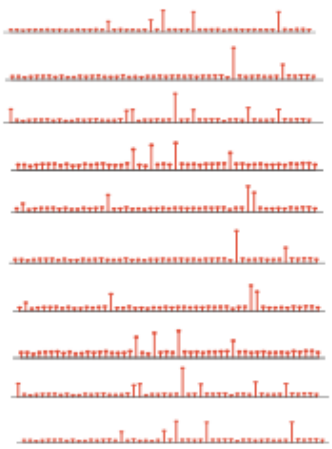
  [*Bayesian model criticism*: Rubin 1984; Gelfand+ 1992; Gelman+ 1996; ...]

Check the factor model



Model	Predictive score
Probabilistic PCA	0.14
Poisson factorization	0.16
Mixtures	0.01
Deep exponential families	0.19

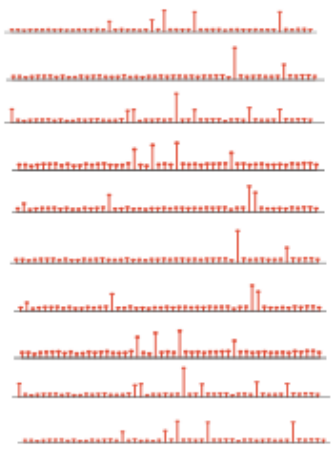
Do causal inference



{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, … }	$2788M
{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, … }	$1845M
{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, … }	$1520M
{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, … }	$1514M
{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, … }	$1506M
{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, … }	$1405M
{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, … }	$1274M
{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, … }	$1215M
{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, … }	$1157M
{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, … }	$1153M

- The representations $\hat{z}_i$ are **substitute confounders**.
- They are latent attributes of movie casts that the factorization has uncovered.
- Form an **augmented dataset** of triplets $(\mathbf{a}_i, y_i, \hat{z}_i)$.

Do causal inference



{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ... }	$2788M
{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ... }	$1845M
{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ... }	$1520M
{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, ... }	$1514M
{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ... }	$1506M
{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ... }	$1405M
{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ... }	$1274M
{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ... }	$1215M
{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ... }	$1157M
{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, ... }	$1153M

- Use the substitute confounders in a **causal inference**.
- E.g., fit regression from casts and confounders to revenue,

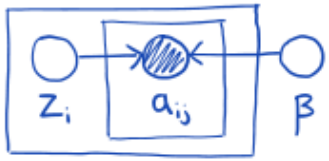
$$\mathbb{E}[Y \mid \mathbf{a}, \hat{z}] = \beta^\top \mathbf{a} + \eta^\top \hat{z}.$$

- Use adjustment to perform causal inference,

$$\mathbb{E}[Y \,;\, \mathrm{do}(\mathbf{a})] \approx \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}[Y \mid \mathbf{a}, \hat{z}_i].$$

What just happened?

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \ldots, \hat{z}_n\}$$

$$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$$

ESTIMATE CAUSAL EFFECTS

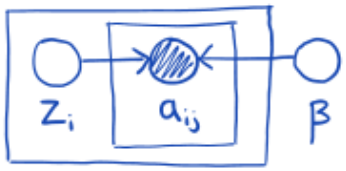
$$\mathbb{E}[Y \mid do(a)] = \mathbb{E}[\mathbb{E}[Y \mid Z, A = a]]$$

- We **modeled the causes** with a factor model.
- We used **learned representations** as substitutes for measured confounders.
- Idea: This exploratory method can correct for *some* unobserved confounding.

A little theory

The deconfounder

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$\{\hat{z}_1, \ldots, \hat{z}_n\}$

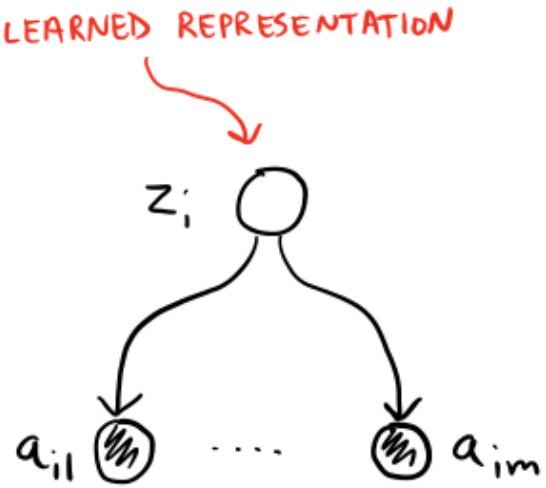
$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$

ESTIMATE CAUSAL EFFECTS

$\mathbb{E}[Y \mid do(a)] = \mathbb{E}[\mathbb{E}[Y \mid Z, A = a]]$

- Suppose we fit a **good factor model** of the assigned causes (the actors).
- Then its learned representation will contain **multi-cause confounders**.
- Main assumption: No unobserved single cause confounders.

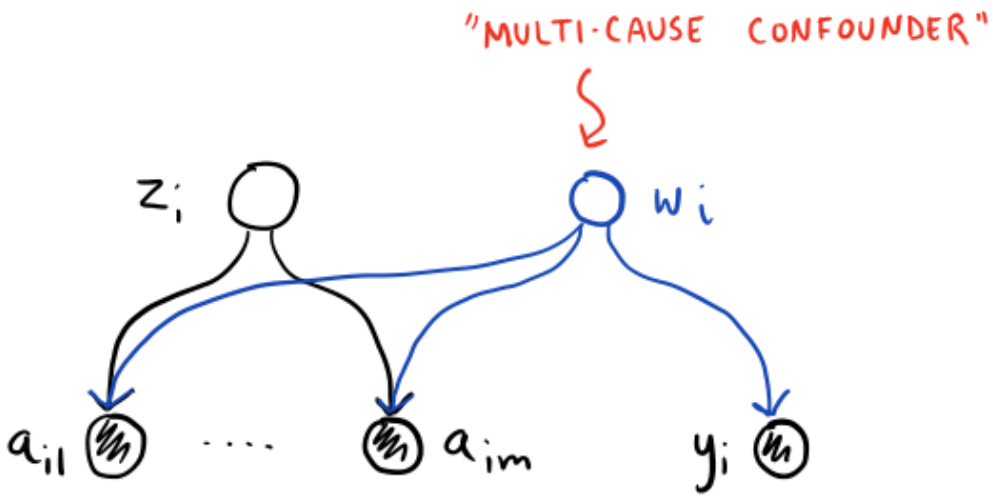
Intuition (through graphical models)



If we find a good factor model then

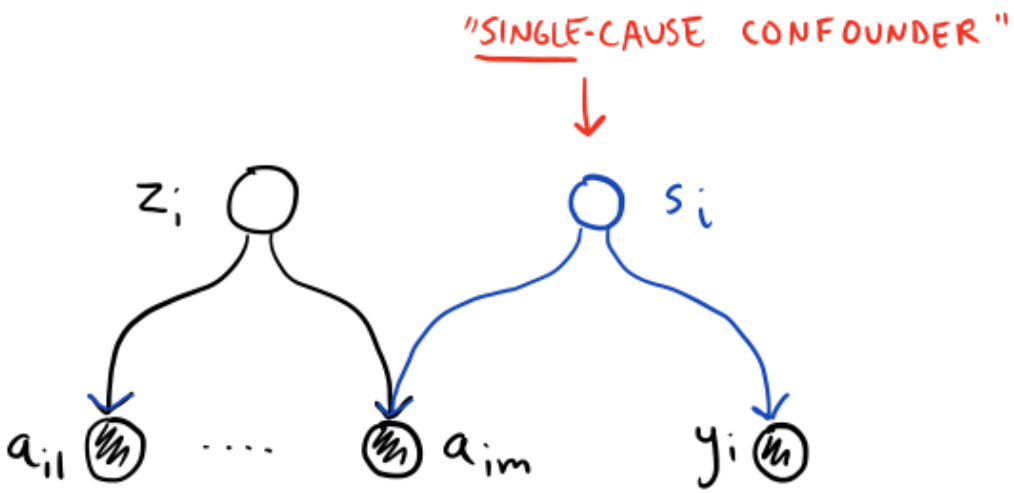
$$\mathrm{p}(a_{i1}, \ldots, a_{im} \,|\, z_i, \beta_{1:m}) = \prod_{j=1}^{m} \mathrm{p}(a_{ij} \,|\, z_i, \beta_j)$$

Intuition (through graphical models)



- There cannot be an additional unobserved **multi-cause confounder**.
- Contradiction: If one existed then the independence statement would not hold.

Intuition (through graphical models)



- There still might be a **single-cause confounder**.
- Reason: The conditional independence still holds.

THEOREM: THE DECONFOUNDER

Suppose $p_{\text{true}}(\mathbf{a})$ can be written $\int \mathrm{p}(z) \prod_j \mathrm{p}(a_j \mid z, \boldsymbol{\beta})\, \mathrm{d}z$.

Further assume

1. No unobserved single-cause confounders X
2. The causes "pinpoint" the substitute $Z = f(\mathbf{a})$
3. Some other assumptions (see the paper)

Then

$$\mathbb{E}[Y\,;\,\mathrm{do}(\mathbf{a})] - \mathbb{E}\left[Y\,;\,\mathrm{do}(\mathbf{a}')\right] = \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y \mid Z, X, \mathbf{a}\right] - \mathbb{E}_Y\left[Y \mid Z, X, \mathbf{a}'\right]\right].$$

(There has been further progress on identification; see references.)

The deconfounder

1. Find and fit a good probabilistic factor model of $\mathbf{A}_i$.
2. Use it to estimate Z_i, which renders the causes conditionally independent.
3. Use Z_i to help with causal inference.

- This theory motivates the algorithm.
- We assume that information about multi-cause confounders is embedded in the (observed) dependencies among the causes.
- The factor model extracts that information and uses it for causal inference.

The deconfounder

1. Find and fit a good probabilistic factor model of $\mathbf{A}_i$.
2. Use it to estimate Z_i, which renders the causes conditionally independent.
3. Use Z_i to help with causal inference.

- Genome-wide association studies: e.g., Pritchard+ 2000, Astle+ 2009, Yu+ 2006, Kang+ 2010, Song+ 2015, Haro+ 2015, Price+ 2006, Renaux+ 2020
- Econometrics in "factor-adjusted regression": e.g., Stock and Watson 2016, Gonclaves and Perron 2014, Cheng and Hansen 2015, Bai and Ng 2006
- Testing, covariance estimation, regression: e.g., Friguet+ 2009, Fan+ 2019, Shah and Meinshausen 2018, Cevid+ 2018, Guo+ 2020

The deconfounder

1. Find and fit a good probabilistic factor model of $\mathbf{A}_i$.
2. Use it to estimate Z_i, which renders the causes conditionally independent.
3. Use Z_i to help with causal inference.

How might the deconfounder go wrong?

- The factor model does not capture the distribution of causes. (It doesn't.)
- There is uncertainty about inference of z. (There is.)
- There is unmeasured single-cause confounding. (There probably is.)
- There is estimation variance. (Yes.)

The deconfounder

1. Find and fit a good probabilistic factor model of $\mathbf{A}_i$.
2. Use it to estimate Z_i, which renders the causes conditionally independent.
3. Use Z_i to help with causal inference.

My two cents

The deconfounder is an *exploratory method* that removes *some* sources of confounding bias. A better factor model captures more of the multi-cause confounding.

How to use a deconfounder: Condition on known confounders, both multi-cause and single-cause, and use domain knowledge to build a good factor model. Then use the deconfounder to explore hypothetical causal connections in your data.

Example: Genome-wide association studies (GWAS)

- GWAS is a problem of multiple causal inference
- How is genetic variation causally connected to a trait?
- For each individual: a trait and many measurements of the genome (SNPs).

Example: Genome-wide association studies (GWAS)

- Multiple-cause confounding is a problem.
- Non-causal SNPs may be highly correlated to causal SNPs
- Misestimates causal effects

Simulation study

ID (i)	SNP_1 ($a_{i,1}$)	SNP_2 ($a_{i,2}$)	SNP_3 ($a_{i,3}$)	SNP_4 ($a_{i,4}$)	SNP_5 ($a_{i,5}$)	SNP_6 ($a_{i,6}$)	SNP_7 ($a_{i,7}$)	SNP_8 ($a_{i,8}$)	SNP_9 ($a_{i,9}$)	$\cdots$	SNP_100K ($a_{i,100K}$)	Height (feet) (y_i)
1	1	0	0	1	0	0	1	2	0	$\cdots$	0	5.73
2	1	2	2	1	2	1	1	0	1	$\cdots$	2	5.26
3	2	0	1	1	0	1	0	1	1	$\cdots$	2	6.24
4	0	0	0	1	1	0	1	2	0	$\cdots$	0	5.78
5	1	2	1	1	1	0	1	0	0	$\cdots$	1	5.09
$\vdots$						$\vdots$						$\vdots$

- Generate SNPs a_{ij}, where each individual belongs to a latent group c_i.
- The true outcome is a trait y_i, drawn from

$$y_i = \sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_{c_i}),$$

  where many β_j are zero, i.e., non-causal SNPs.
- Confounded: the intercept λ_{c_i} and error ε_i are connected to the latent group.

Simulation study

	pred. score	**Real-valued outcome** RMSE$\times 10^2$	**Binary outcome** RMSE$\times 10^2$
No control	—		
Control for confounders*	—		
(G)LMM	—		
PPCA	0.14		
PF	0.15		
LFA	0.14		
Mixture	0.00		
DEF	0.20		

- We fit many factor models; none was the true model.
- Each provides different levels of predictive performance.
- All computation done in Edward [Tran+ 2018] .

Simulation study

	pred. score	**Real-valued outcome** RMSE$\times 10^2$	**Binary outcome** RMSE$\times 10^2$
No control	—	58.82	29.50
Control for confounders*	—	25.32	25.77
(G)LMM	—	35.18	28.87
PPCA	0.14	33.32	26.70
PF	0.15	33.38	26.84
LFA	0.14	33.93	26.83
Mixture	0.00	57.59	29.96
DEF	0.20	**26.47**	**25.91**

- Also fit outcome models with no control and with observed confounders
- The deconfounder provides good causal estimates.
- Predictive checks indicate downstream causal performance.

Is the theory correct?

- There has been some debate about the theory (Ogburn+ 2020, 2021).
- Key worry: By definition, we can *never* know anything about truly unobserved confounders from the observational data.
- The deconfounder does *not* challenge this indisputable fact.
- Rather, it finds a class of confounders that are *effectively observed*.
- Thus the deconfounder tries to *extract* effectively observed information.
- For a clarification of this theory, see Wang and Blei (2020).

Summary

- The deconfounder assumes there is information in the dependency among causes that is helpful for removing confounding bias.
- The algorithm tries to extract this information for causal inference. It uses unsupervised learning and Bayesian model criticism.

Caveat

- The deconfounder is not a turnkey solution to causal inference.
- It does not relieve the researcher from measuring confounders.
- It comes with uncheckable assumptions.

Neural network based multiple treatment effect estimators

- Deep embedding for multiple treatment effect [1]
  - Deep generative model
  - Leverage the dependency between multiple treatments
- Disentangled representation learning in multiple treatment environment [2]
  - Enhance human understanding for confounders
  - Treatments in different hierarchies

[1] Saini S K, et al. Multiple treatment effect estimation using deep generative model with task embedding[C]//WWW. 2019.
[2] Ma J, et al. Multi-cause effect estimation with disentangled confounder representation[C]//IJCAI. 2021.

References

- Some of the slide contents are from
  - Tutorial on deep learning for causal inference. Bernard Koch (SICSS-Los Angeles 19, 20, 21.
    https://www.youtube.com/watch?v=v9uf9rDYEMg
  - Ramachandra V. Deep learning for causal inference[J]. arXiv preprint arXiv:1803.00149, 2018.
  - AAAI 2020 Tutorial. Representation Learning for Causal Inference.
  - David M. Blei. The Blessing of Multiple Causes.

Reading Materials

- Shi C, et al. Adapting neural networks for the estimation of treatment effects[J]. NeurIPS, 2019.
  - https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf
- L. Yao, et al. "Representation learning for treatment effect estimation from observational data." *NeurIPS* 2018.
  - https://proceedings.neurips.cc/paper_files/paper/2018/file/a50abba8132a771917913390c3eb19fe7-Paper.pdf
- Wang Y, Blei D M. The blessings of multiple causes[J]. Journal of the American Statistical Association, 2019, 114(528): 1574-1596.
  - https://www.tandfonline.com/doi/pdf/10.1080/01621459.2019.1686987?casa_token=I3Kt9ikdVf4AAAAA:oceUkUyi16mj-3CZ3mE6qgjZ8n6l0Bqm4gW936LizPL21q_uR3FOljDuhfC9y_X3kOZWQHC9Tdxrxw

Thank you!
Questions?