Proxy Methods for Causal Effect Estimation with Hidden Confounders

Arthur Gretton

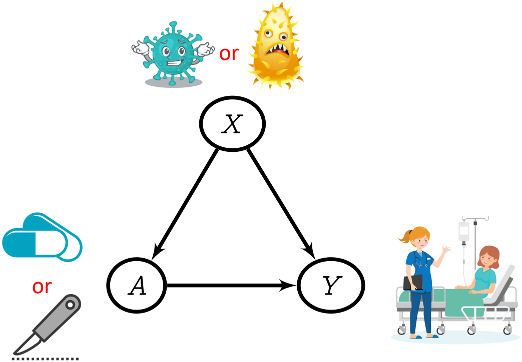
Gatsby Computational Neuroscience Unit
Google Deepmind

UCL Centre for Data Science Symposium, 2023

Introduction: observation vs intervention

Conditioning from observation:

$\mathbb{E}[Y|A=a] = \sum_{x\in\{0,1\}} \mathbb{E}[Y|a,x]p(x|a)$

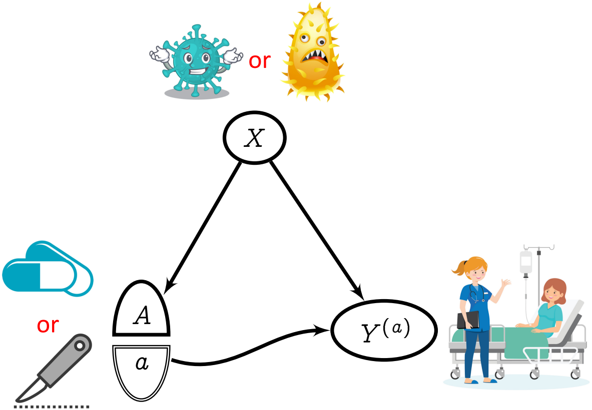


From our *observations* of historical hospital data:

- $P(Y = \text{cured}|A = \text{pills}) = 0.85$
- $P(Y = \text{cured}|A = \text{surgery}) = 0.72$

Introduction: observation vs intervention

Average causal effect (intervention): $\mathbb{E}[Y^{(a)}] = \sum_{x \in \{0,1\}} \mathbb{E}[Y | a, x] p(x)$

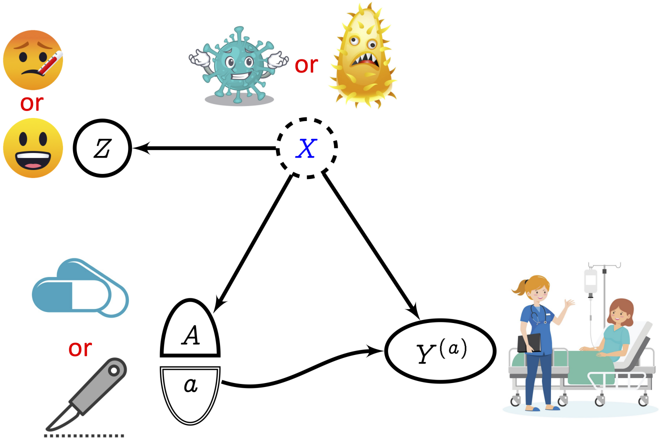


From our *intervention* (making all patients take a treatment):

- $P(Y^{(\text{pills})} = \text{cured}) = 0.64$
- $P(Y^{(\text{surgery})} = \text{cured}) = 0.75$

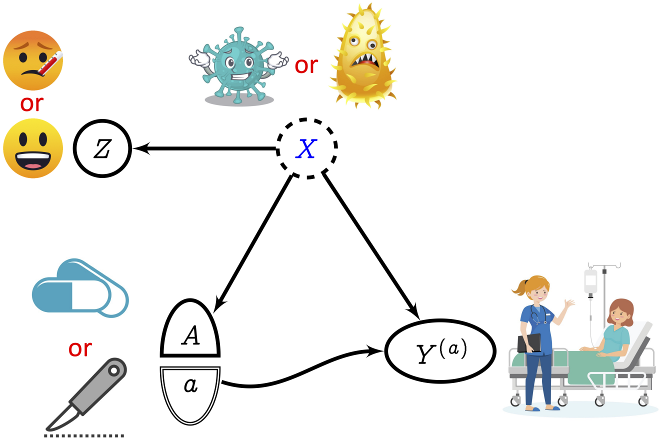
Richardson, Robins (2013), Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

We observe symptom Z, not disease X



- $P(Z = \text{fever} | X = \text{mild}) = 0.2$
- $P(Z = \text{fever} | X = \text{severe}) = 0.8$

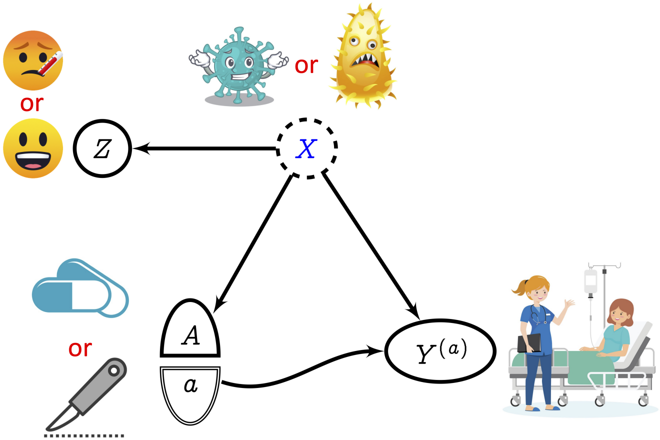
We observe symptom Z, not disease X



- $P(Z = \text{fever} | X = \text{mild}) = 0.2$
- $P(Z = \text{fever} | X = \text{severe}) = 0.8$

Could we just write: $P(Y^{(a)}) \stackrel{?}{=} \sum_{z \in \{0,1\}} \mathbb{E}[Y | a, z] p(z)$

We observe symptom Z, not disease X



Results are very bad:

- $\sum_{z \in \{0,1\}} \mathbb{E}[\text{cured} | \text{pills}, z] p(z) = 0.8 \quad (\neq 0.64)$
- $\sum_{z \in \{0,1\}} \mathbb{E}[\text{cured} | \text{surgery}, z] p(z) = 0.73 \quad (\neq 0.75)$

Correct answer impossible without observing X

Pearl (2010), On Measurement Bias in Causal Inference

Outline

Causal effect estimation, with hidden covariates X:

- Use proxy variables (negative controls)

What's new? What is it good for?

- Treatment A, proxy variables, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations
- Don't ~~meet your heroes~~ model your hidden variables!

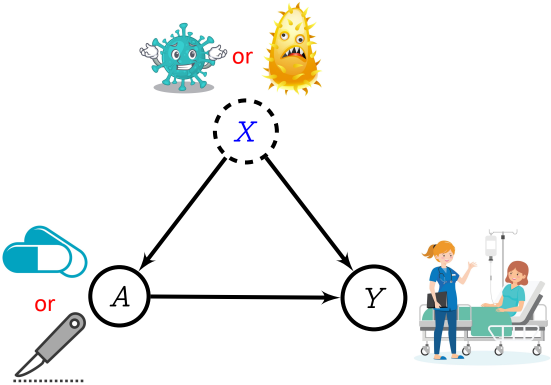
Proxy/Negative Control Methods

Proxy variables: health example

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome



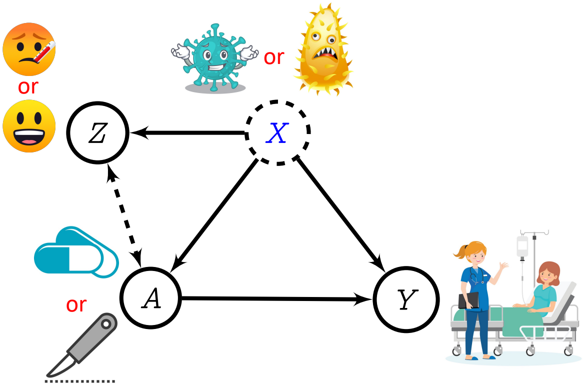
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms



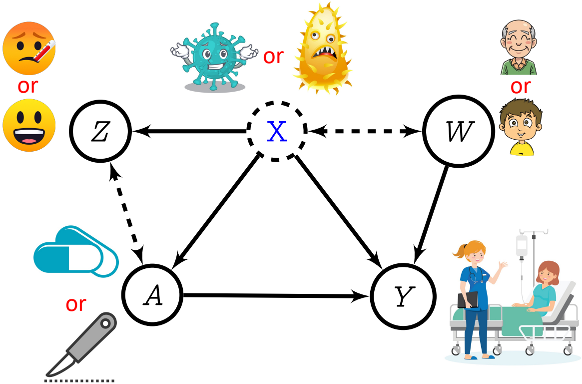
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms
- W: age



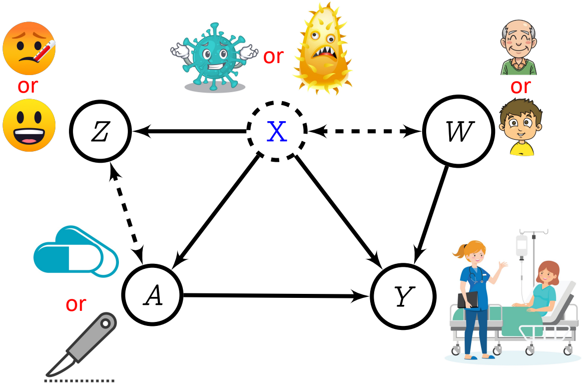
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms
- $:W$ age



$\Longrightarrow$ Can recover $\mathbb{E}(Y^{(a)})$ from observational data!

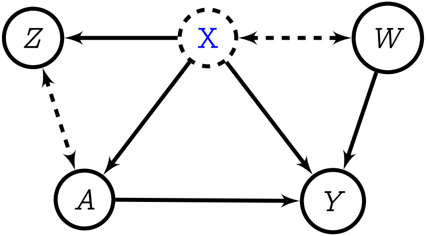
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: general setting

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: treatment proxy
- W outcome proxy

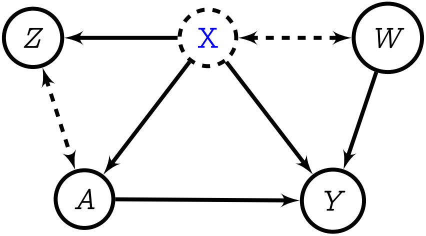


Proxy variables: general setting

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: treatment proxy
- W outcome proxy



Structural assumptions:

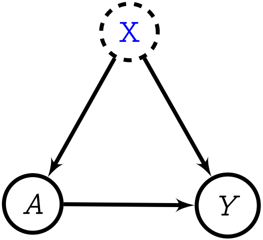
$$W \perp\!\!\!\perp (Z, A) | X$$

$$Y \perp\!\!\!\perp Z | (A, X)$$

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



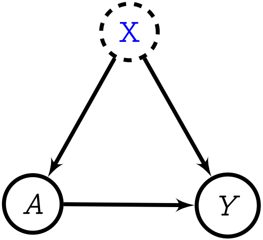
If X were observed,

$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i)$$

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



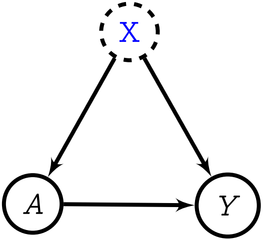
If X were observed,

$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



If X were observed,

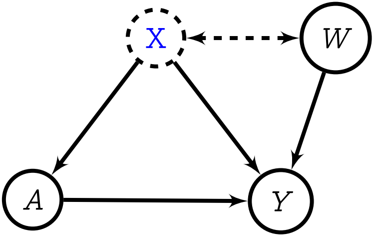
$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

Goal: "get rid of the blue" X

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



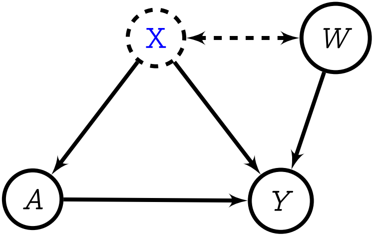
For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

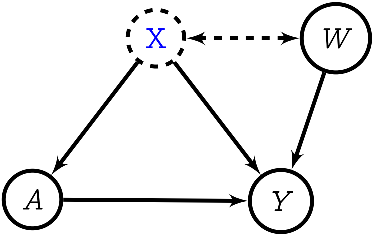
.....then

$$P(Y^{(a)}) = P(Y|X,a)P(X)$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w}\underbrace{P(W|X)}_{d_w \times d_x}$$

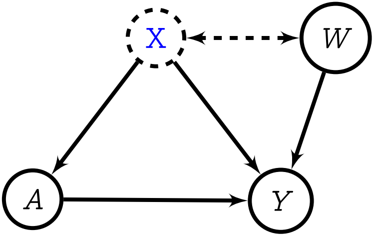
.....then

$$\begin{aligned} P(Y^{(a)}) &= P(Y|X,a)P(X) \\ &= H_{w,a}P(W|X)P(X) \end{aligned}$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

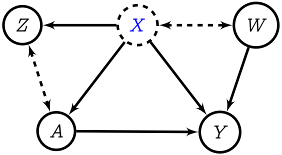
.....then

$$\begin{aligned} P(Y^{(a)}) &= P(Y|X,a)P(X) \\ &= H_{w,a} P(W|X)P(X) \\ &= H_{w,a} P(W) \end{aligned}$$

...now project onto $p(X|Z,a)$

From last slide,

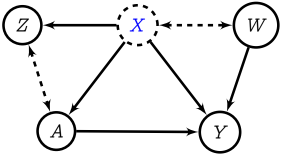
$$P(Y|X,a) \qquad = H_{w,a} P(W|X)$$



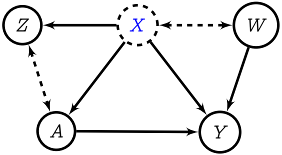
...now project onto $p(X|Z,a)$

From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a}P(W|X)\underbrace{p(X|Z,a)}_{d_x \times d_z}$$



...now project onto $p(X|Z,a)$



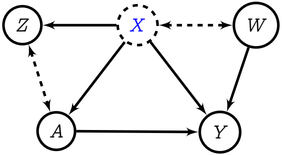
From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X)\underbrace{p(X|Z,a)}_{d_x \times d_z}$$

Because $W \perp\!\!\!\perp (Z,A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

...now project onto $p(X|Z,a)$



From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X)\underbrace{p(X|Z,a)}_{d_x \times d_z}$$

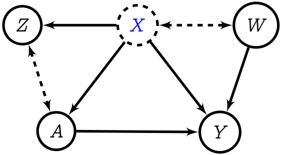
Because $W \perp\!\!\!\perp (Z,A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

Because $Y \perp\!\!\!\perp Z|(A,X)$,

$$P(Y|X,a)p(X|Z,a) = P(Y|Z,a)$$

...now project onto $p(X|Z,a)$



From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X) \underbrace{p(X|Z,a)}_{d_x \times d_z}$$

Because $W \perp\!\!\!\perp (Z,A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

Because $Y \perp\!\!\!\perp Z|(A,X)$,

$$P(Y|X,a)p(X|Z,a) = P(Y|Z,a)$$

Solve for $H_{w,a}$:

$$P(Y|Z,a) = H_{w,a} P(W|Z,a)$$

Everything observed!

Proxy/Negative Control Methods in the Real World

Unobserved confounders: proxy methods

Kernel features (ICML 2021):

arXiv.org > cs > arXiv:2105.04544

Computer Science > Machine Learning

[Submitted on 10 May 2021 (v1), last revised 9 Oct 2021 (this version, v4)]

Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, Krikamol Muandet







NN features (NeurIPS 2021):

arXiv.org > cs > arXiv:2106.03907

Computer Science > Machine Learning

[Submitted on 7 Jun 2021 (v1), last revised 7 Dec 2021 (this version, v2)]

Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Liyuan Xu, Heishiro Kanagawa, Arthur Gretton


Code for NN and kernel proxy methods:

https://github.com/liyuan9988/DeepFeatureProxyVariable/

Main theorem

If X were observed, we would write (average treatment effect)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a, x)p(x)dx.$$

....but we do not observe X.

Main theorem

If X were observed, we would write (average treatment effect)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a, x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(y|a, z) = \int_w h_y(w, a)p(w|a, z)dw$$

- "Primary task" $\mathbb{E}(y|a, z)$, "auxiliary task" $p(w|a, z)$, linked by h_y
- All variables observed, X not seen *or modeled.*

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Main theorem

If X were observed, we would write (average treatment effect)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a, x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(y|a, z) = \int_w h_y(w, a)p(w|a, z)dw$$

- "Primary task" $\mathbb{E}(y|a, z)$, "auxiliary task" $p(w|a, z)$, linked by h_y
- All variables observed, X not seen *or modeled.*

Average treatment effect via $p(w)$:

$$\mathbb{E}(Y^{(a)}) = \int_w h_y(a, w)p(w)dw$$

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Main theorem

If X were observed, we would write (average treatment effect)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a, x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(y|a, z) = \int_w h_y(w, a)p(w|a, z)dw$$

- "Primary task" $\mathbb{E}(y|a, z)$, "auxiliary task" $p(w|a, z)$, linked by h_y
- All variables observed, X not seen *or modeled.*

Average treatment effect via $p(w)$:

$$\mathbb{E}(Y^{(a)}) = \int_w h_y(a, w)p(w)dw$$

Challenge: need to parametrize and solve for h_y

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Link function NN parametrization

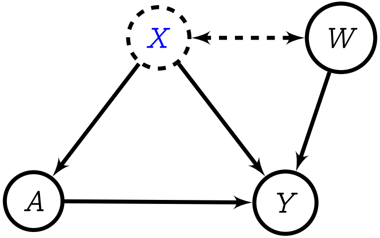
The link function takes the form

$$h_y(a,w) = \gamma^\top [\varphi_\theta(w) \otimes \varphi_\xi(a)]$$

Assume we have:

- output proxy NN features $\varphi_\theta(w)$
- treatment NN features $\varphi_\xi(a)$
- linear final layer γ

(argument of feature map indicates feature space)



Link function NN parametrization

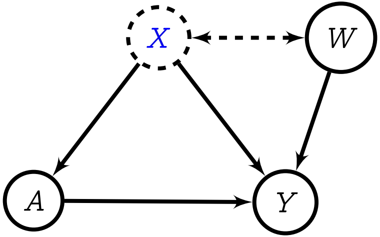
The link function takes the form

$$h_y(a, w) = \gamma^\top [\varphi_\theta(w) \otimes \varphi_\xi(a)]$$

Assume we have:

- output proxy NN features $\varphi_\theta(w)$
- treatment NN features $\varphi_\xi(a)$
- linear final layer γ

(argument of feature map indicates feature space)



Questions:

- Why feature map $\varphi_\theta(w) \otimes \varphi_\xi(a)$?
- Why final linear layer γ?

Both are necessary (next slides)!

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: minimize

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: minimize

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

....where

$$\mathbb{E}_{W|A,Z} h_y(W, A) = \mathbb{E}_{W|A,Z} \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(A)) \right]$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: minimize

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

....where

$$\begin{aligned} \mathbb{E}_{W|A,Z} h_y(W, A) &= \mathbb{E}_{W|A,Z} \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(A)) \right] \\ &= \gamma^\top \Big(\underbrace{\mathbb{E}_{W|A,Z} [\varphi_\theta(W)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(A) \Big) \end{aligned}$$

(this is why linear γ and feature map $\varphi_\theta(w) \otimes \varphi_\xi(a)$)

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: minimize

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

....where

$$\begin{aligned} \mathbb{E}_{W|A,Z} h_y(W, A) &= \mathbb{E}_{W|A,Z} \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(A)) \right] \\ &= \gamma^\top \Big(\underbrace{\mathbb{E}_{W|A,Z} [\varphi_\theta(W)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(A) \Big) \end{aligned}$$

How to get conditional feature mean $\mathbb{E}_{W|A,Z} [\varphi_\theta(W)]$?

Density estimation for $p(W|a, z)$? Sample from $p(W|a, z)$?

Deaner (2021).

Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).

Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: minimize

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

....where

$$\begin{aligned} \mathbb{E}_{W|A,Z} h_y(W, A) &= \mathbb{E}_{W|A,Z} \left[\gamma^\top \left(\varphi_\theta(W) \otimes \varphi_\xi(A) \right) \right] \\ &= \gamma^\top \Big(\underbrace{\mathbb{E}_{W|A,Z} [\varphi_\theta(W)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(A) \Big) \end{aligned}$$

How to get conditional feature mean $\mathbb{E}_{W|A,Z} [\varphi_\theta(W)]$?

Density estimation for $p(W|a, z)$? Sample from $p(W|a, z)$?

Answer: ridge regression (again!)

Deaner (2021).

Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).

Xu, Kanagawa, G. (2021).

Learning the auxiliary task

We have

$$\mathbb{E}_{W|a,z}\left[\varphi_\theta(W)\right] = \hat{F}_{\theta,\zeta}\varphi_\zeta(a,z)$$

where $\hat{F}_{\theta,\zeta} \in \mathbb{R}^{d_\theta \times d_\zeta}$ minimizes Stage 1 RR loss:

$$\mathbb{E}_{W,A,Z}\left\|\varphi_\theta(W) - F\varphi_\zeta(A,Z)\right\|^2 + \lambda_1\|F\|^2$$

Xu, Kanagawa, G. (2021).

Learning the auxiliary task

We have

$$\mathbb{E}_{W|a,z}[\varphi_\theta(W)] = \hat{F}_{\theta,\zeta}\varphi_\zeta(a,z)$$

where $\hat{F}_{\theta,\zeta} \in \mathbb{R}^{d_\theta \times d_\zeta}$ minimizes Stage 1 RR loss:

$$\mathbb{E}_{W,A,Z}\|\varphi_\theta(W) - F\varphi_\zeta(A,Z)\|^2 + \lambda_1\|F\|^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt ϕ_θ, ϕ_ζ:

$$\hat{F}_{\theta,\zeta} = C_{W,AZ}(C_{AZ} + \lambda_1 I)^{-1} \qquad C_{W,AZ} = \mathbb{E}[\varphi_\theta(W)\phi_\zeta^\top(A,Z)]$$

$$C_{AZ} = \mathbb{E}[\phi_\zeta(A,Z)\phi_\zeta^\top(A,Z)]$$

Plug $\hat{F}_{\theta,\zeta}$ into S1 loss, take gradient steps for ζ (...but not θ...)

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y-\gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right]\otimes\varphi_\xi(A)\right)\right)^2+\lambda_2\|\gamma\|^2$$

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y-\gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right]\otimes\varphi_\xi(A)\right)\right)^2+\lambda_2\|\gamma\|^2$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)]\approx\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$
- Substitute $\hat{\gamma}$ into Stage 2, minimize wrt θ, ξ
    - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y-\gamma^{\top}\left(\mathbb{E}_{W|A,Z}\left[\varphi_{\theta}(W)\right]\otimes\varphi_{\xi}(A)\right)\right)^{2}+\lambda_{2}\|\gamma\|^{2}$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_{\theta}(W)]\approx\hat{F}_{\theta,\zeta}\phi_{\zeta}(A,Z)$$

Solution procedure: for γ,θ,ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_{\zeta}(A,Z)$ and $\varphi_{\xi}(A)$
- Substitute $\hat{\gamma}$ into Stage 2, minimize wrt θ,ξ
    - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_{θ}.
    - Iterate between θ,ξ and ζ

Xu, Kanagawa, G. (2021).

Final algorithm

Stage 2 RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y-\gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right]\otimes\varphi_\xi(A)\right)\right)^2+\lambda_2\|\gamma\|^2$$

Stage 1 regression (auxiliary): NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)]\approx\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ,θ,ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$
- Substitute $\hat{\gamma}$ into Stage 2, minimize wrt θ,ξ
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.
  - Iterate between θ,ξ and ζ

Key point: features $\varphi_\theta(W)$ learned specially for primary task:

$$\mathbb{E}(Y|a,Z)=\int_w h_y(W,a)p(W|a,Z)dw$$

Contrast with autoencoders/sampling: must reconstruct/sample all of W.

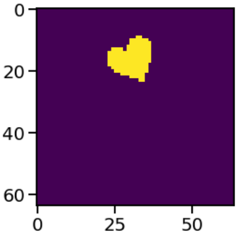
Xu, Kanagawa, G. (2021).

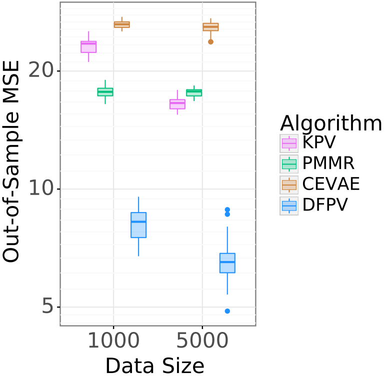
Experiments

Synthetic experiment, adaptive neural net features

dSprite example:

- $X = \{\mathtt{scale}, \mathtt{rotation}, \mathtt{posX}, \mathtt{posY}\}$
- Treatment A is the image generated (with Gaussian noise)
- Outcome Y is quadratic function of A with multiplicative confounding by $\mathtt{posY}$.
- $Z = \{\mathtt{scale}, \mathtt{rotation}, \mathtt{posX}\}$, $W =$ noisy image sharing $\mathtt{posY}$
- Comparison with CEVAE (Louzios et al. 2017)



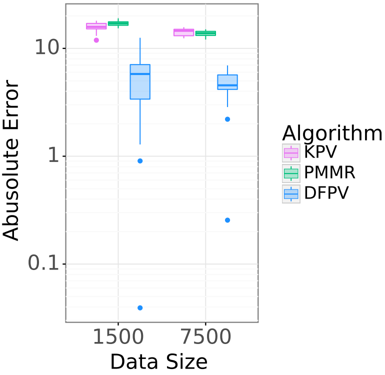


Louizos, Shalit, Mooij, Sontag, Zemel, Welling, Causal Effect Inference with Deep Latent-Variable Models (2017)

Confounded offline policy evaluation

Synthetic dataset, demand prediction for flight purchase.

- Treatment A is ticket price.
- Policy $A \sim \pi(Z)$ depends on fuel price.

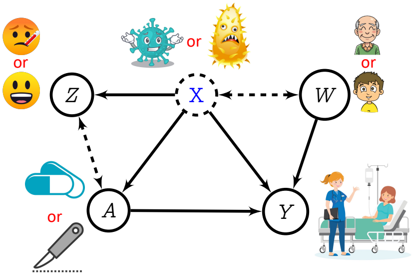


Conclusion

Causal effect estimation with unobserved X, (possibly) complex nonlinear effects on A, Y

We need to observe:

- Treatment proxy Z (interacts with A, but not directly with Y)
- Outcome proxy W (no direct interaction with A, can affect Y)

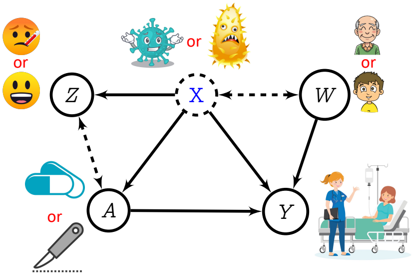


Conclusion

Causal effect estimation with unobserved X, (possibly) complex nonlinear effects on A, Y

We need to observe:

- Treatment proxy Z (interacts with A, but not directly with Y)
- Outcome proxy W (no direct interaction with A, can affect Y)



Key messages:

- Don't model or sample from latents X
- Don't model all of W, only relevant features
- "Ridge regression is all you need"

Code available:

https://github.com/liyuan9988/DeepFeatureProxyVariable/

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Questions?

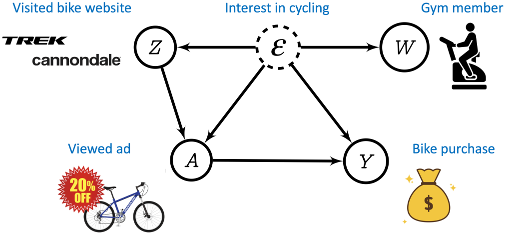

THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Web ads example

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- ε: "interest in cycling"
- A: bike ad on browser
- Y: purchase
- Z: visit to bike website
  $\implies$ cookies
- W membership of gym



Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Tennenholtz, Mannor, Shalit (2020), OPE in Partially Observed Environments.

Uehara, Sekhari, Lee, Kallus, Sun (2022) Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a, \varepsilon)p(\varepsilon)d\varepsilon.$$

....but we do not observe ε.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)d\varepsilon.$$

....but we do not observe ε.

Main theorem: Assume we solved:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

Both $p(y|a,z)$ and $p(w|a,z)$ are in terms of observed quantities.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)d\varepsilon.$$

....but we do not observe ε.

Main theorem: Assume we solved:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

Both $p(y|a,z)$ and $p(w|a,z)$ are in terms of observed quantities.

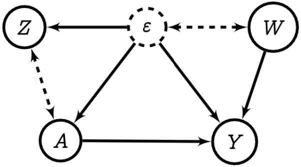
Average treatment effect via $p(w)$:

$$p(y^{(a)}) = \int h_y(a,w)p(w)dw$$

Proof (1)

Because $W \perp\!\!\!\perp (Z, A) | \varepsilon$, we have

$$p(w|a, z) = \int p(w|\varepsilon) p(\varepsilon|a, z) d\varepsilon$$



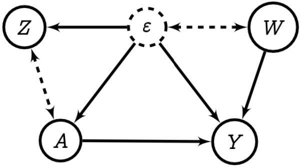
Proof (1)

Because $W \perp\!\!\!\perp (Z, A) | \varepsilon$, we have

$$p(w|a, z) = \int p(w|\varepsilon) p(\varepsilon|a, z) d\varepsilon$$

Because $Y \perp\!\!\!\perp Z | (A, \varepsilon)$ we have

$$p(y|a, z) = \int p(y|a, \varepsilon) p(\varepsilon|a, z) d\varepsilon$$



Proof (3)

Given the solution h_y to:

$$p(y|a, z) = \int h_y(w, a) p(w|a, z) dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

Proof (3)

Given the solution h_y to:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

$$\int p(y|a,\varepsilon)p(\varepsilon|a,z)d\varepsilon = \int h_y(w,a)\int p(w|\varepsilon)p(\varepsilon|a,z)d\varepsilon\, dw$$

Proof (3)

Given the solution h_y to:

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

$$\int p(y|a,\varepsilon) p(\varepsilon|a,z) d\varepsilon = \int h_y(w,a) \int p(w|\varepsilon) p(\varepsilon|a,z) d\varepsilon\, dw$$

This implies:

$$p(y|a,\varepsilon) = \int h_y(w,a) p(w|\varepsilon) dw$$

under identifiability condition

$$\mathbb{E}[f(\varepsilon)|A=a, Z=z] = 0,\ \forall (z,a) \iff f(\varepsilon) = 0,\ \mathbb{P}_{\varepsilon|A=a} \text{ a.s.} \quad (\triangle)$$

Proof (4)

From last slide,

$$p(y|a,\varepsilon) = \int h_y(w,a)p(w|\varepsilon)dw$$

Thus

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)du$$

Proof (4)

From last slide,

$$p(y|a,\varepsilon) = \int h_y(w,a)p(w|\varepsilon)dw$$

Thus

$$\begin{aligned} p(y|do(a)) &= \int_u p(y|a,\varepsilon)p(\varepsilon)du \\ &= \int_u \left[\int h_y(w,a)p(w|\varepsilon)dw\right] p(\varepsilon)d\varepsilon \end{aligned}$$

Proof (4)

From last slide,

$$p(y|a,\varepsilon) = \int h_y(w,a)p(w|\varepsilon)dw$$

Thus

$$\begin{aligned} p(y|do(a)) &= \int_u p(y|a,\varepsilon)p(\varepsilon)du \\ &= \int_u \left[\int h_y(w,a)p(w|\varepsilon)dw \right] p(\varepsilon)d\varepsilon \\ &= \int h_y(w,a)p(w)dw \end{aligned}$$

How not to do 2SLS for proxy methods

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in HS} \mathbb{E}_{w,a,z} \|\phi(w) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

Deaner (2021).

Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).

Xu, Kanagawa, G. (2021).

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in HS} \mathbb{E}_{w,a,z} \|\phi(w) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

Average treatment effect estimate:

$$\mathbb{E}_y(y|do(a)) = \langle h_{\lambda_2}, \phi(a) \otimes \mu_W \rangle ,$$

where $\mu_W = \mathbb{E}_W \phi(W)$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

How *not* to do it

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W,A|a,z} \right\rangle \right)^2 + \lambda_2 \|h\|_{\mathcal{H}}^2$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in \mathcal{G}} \mathbb{E}_{w,a,z} \|\phi(w) \otimes \phi(a) - F[\phi(a) \otimes \phi(z)]\|_{\mathcal{H}_{\mathcal{W}}}^2 + \lambda_1 \|F\|_{HS}^2$$

which gives us

$$\mu_{W,A|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

How *not* to do it

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W,A|a,z} \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in \mathcal{G}} \mathbb{E}_{w,a,z} \|\phi(w) \otimes \phi(a) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_\mathcal{W}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

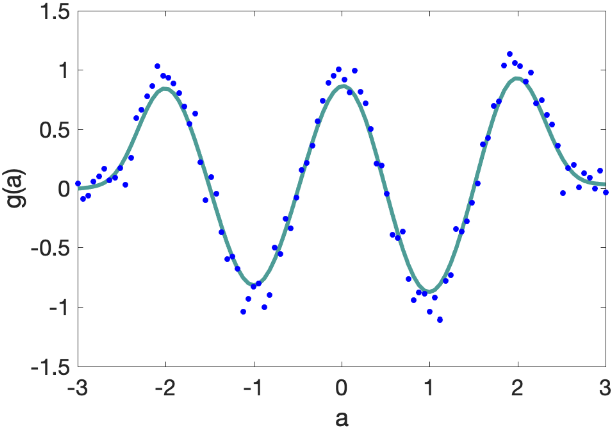
$$\mu_{W,A|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

Problem: ridge regressing from $\phi(a)$ to $\phi(a)$.

Theoretical issue: $\mathcal{I}_{\mathcal{H}_\mathcal{A}}$ is not Hilbert-Schmidt so consistency of F not established.

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

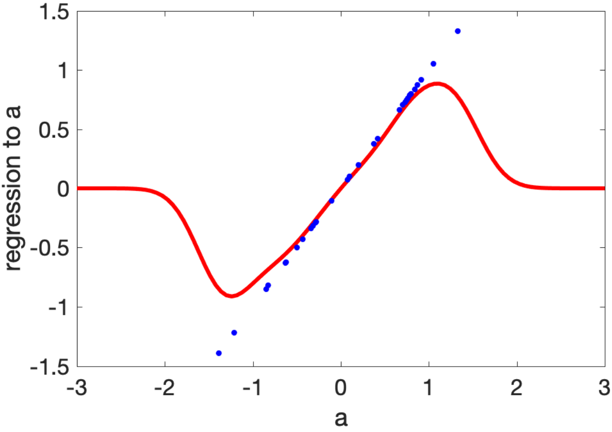
$a \sim \mathcal{N}(0, \sigma^2).$

Stage 2:

$a \sim \mathcal{U}[-3, 3].$

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

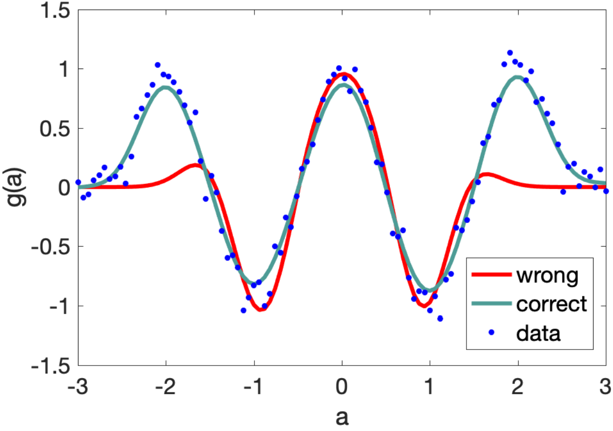
$$a \sim \mathcal{N}(0, \sigma^2).$$

Stage 2:

$$a \sim \mathcal{U}[-3, 3].$$

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

$$a \sim \mathcal{N}(0, \sigma^2).$$

Stage 2:

$$a \sim \mathcal{U}[-3, 3].$$

Failures of identifiability assumptions (1)

Recall (one of the) identifiability assumptions:

$$\mathbb{E}[f(\varepsilon)|A = a, Z = z] = 0,\ \mathbb{P}_{Z|A=a}\ \text{a.s.} \iff f(\varepsilon) = 0,\ \mathbb{P}_{\varepsilon|A=a}\ \text{a.s.} \quad (\triangle)$$

For conciseness, assume conditioning on some a.

Failure 1: $Z \perp\!\!\!\perp \varepsilon$ (no information about ε in proxy)

$$g(\varepsilon) = \tilde{g}(\varepsilon) - \mathbb{E}_{\varepsilon}\tilde{g}(\varepsilon)$$
$$\mathbb{E}(g(\varepsilon)|Z) = \mathbb{E}g(\varepsilon) = 0.$$

Failures of identifiability assumptions (2)

Failure 2: "exploitable invariance" of $p(\varepsilon|z)$

$$\varepsilon \sim \mathcal{N}(0,1),$$
$$Z = |\varepsilon| + \mathcal{N}(0,1),$$

where $p(\varepsilon|z) \propto p(z|\varepsilon)p(\varepsilon)$ symmetric in ε. Consider square integrable *antisymmetric* function $g(\varepsilon) = -g(-\varepsilon)$. Then

$$\begin{aligned}
&\int_{-\infty}^{\infty} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon \\
&= \int_{-\infty}^{0} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon + \int_{0}^{\infty} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon \\
&= 0.
\end{aligned}$$

If distribution of $\varepsilon|Z$ retains the same "symmetry class" over a set of Z with nonzero measure, then the assumption is violated by $g(\varepsilon)$ with zero mean on this class.