Causal Effect Estimation with Context and Confounders

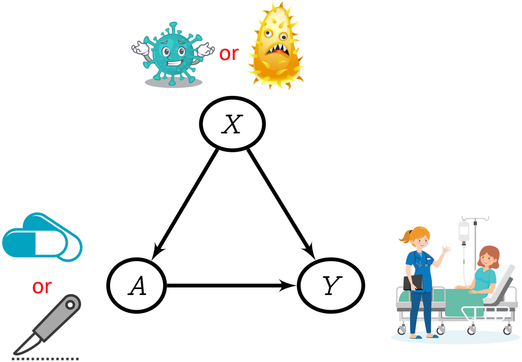
Arthur Gretton

Gatsby Computational Neuroscience Unit,
Deepmind

MLSS Okinawa, 2024

Observation vs intervention

Conditioning from observation: $\mathbb{E}[Y|A=a] = \sum_x \mathbb{E}[Y|a,x]p(x|a)$

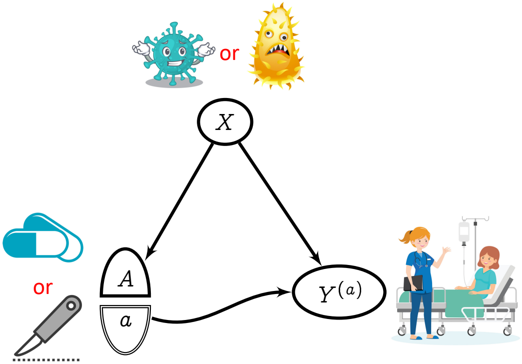


From our observations of historical hospital data:

- $P(Y = \text{cured}|A = \text{pills}) = 0.80$
- $P(Y = \text{cured}|A = \text{surgery}) = 0.72$

Observation vs intervention

Average causal effect (intervention): $\mathbb{E}[Y^{(a)}] = \sum_x \mathbb{E}[Y \mid a, x]\,p(x)$

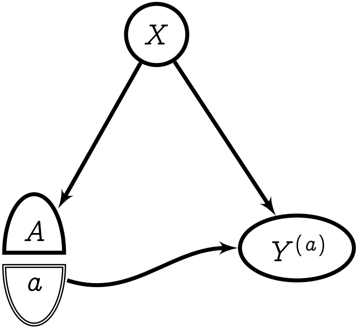


From our intervention (making all patients take a treatment):

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

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

Questions we will solve



Outline

First lecture: causal effect estimation, observed covariates:

- Average treatment effect (ATE), conditional average treatment effect (CATE), average treatment on treated (ATT), mediation effects.

Second lecture: causal effect estimation, hidden covariates:

- ... instrumental variables, proxy variables

What's new? What is it good for?

- Treatment A, covariates X, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations

Regression assumption: linear functions of features

All learned functions will take the form:

$$\begin{aligned} \gamma(x) &= \gamma^\top \varphi_\theta(x) \\ &\overset{\text{or}}{=} \langle \gamma, \varphi(x) \rangle_{\mathcal{H}} \end{aligned}$$

Regression assumption: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x) \stackrel{\text{or}}{=} \langle \gamma, \varphi(x) \rangle_{\mathcal{H}}$$

Option 1: Finite dictionaries of learned neural net features $\varphi_\theta(x)$ (linear final layer γ)

Xu, G., A Neural mean embedding approach for back-door and front-door adjustment. (ICLR 23)

Xu, Chen, Srinivasan, de Freitas, Doucet, G. Learning Deep Features in Instrumental Variable Regression. (ICLR 21)

Option 2: Infinite dictionaries of fixed kernel features:

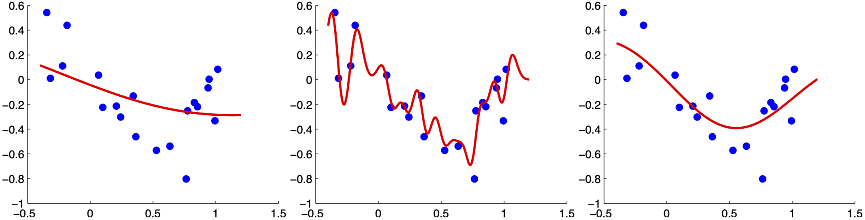
$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Kernel is feature dot product.

Singh, Xu, G. Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. (Biometrika 23)

Singh, Sahani, G. Kernel Instrumental Variable Regression. (NeurIPS 19)

Kernel ridge regression: reminder



Kernel ridge regression

Approximate $\gamma_0(x) := \mathbb{E}[Y | X = x]$ using ridge regression

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \sum_{i=1}^{n} (y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Kernel ridge regression

Approximate $\gamma_0(x) := \mathbb{E}[Y|X=x]$ using ridge regression

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}}\right)^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Representer theorem:

$$\gamma = \sum_{i=1}^{n} \alpha_i \varphi(x_i), \qquad \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j),$$

Kernel ridge regression

Approximate $\gamma_0(x) := \mathbb{E}[Y|X=x]$ using ridge regression

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Representer theorem:

$$\gamma = \sum_{i=1}^{n} \alpha_i \varphi(x_i), \qquad \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j),$$

Solution is

$$\begin{aligned} \hat{\alpha} &= \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \|y - K\alpha\|^2 + \lambda \alpha^\top K \alpha \\ &= (K + \lambda I_n)^{-1} y. \end{aligned}$$

Kernel ridge regression

Approximate $\gamma_0(x) := \mathbb{E}[Y|X=x]$ using ridge regression

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Representer theorem:

$$\gamma = \sum_{i=1}^{n} \alpha_i \varphi(x_i), \qquad \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j),$$

Solution is

$$\begin{aligned} \hat{\alpha} &= \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \|y - K\alpha\|^2 + \lambda \alpha^\top K \alpha \\ &= (K + \lambda I_n)^{-1} y. \end{aligned}$$

Prediction at new x: define $(k_{Xx})_i = k(x_i, x)$,

$$\hat{\gamma}(x) = \langle \hat{\gamma}, \varphi(x) \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \varphi(x) \right\rangle_{\mathcal{H}} = \hat{\alpha}^\top k_{Xx}$$

Model fitting: ridge regression

Approximate $\gamma_0(x) := \mathbb{E}[\,Y\,|\,X = x]$ from features $\varphi(x_i)$ and y_i:

$$\hat{\gamma} \;=\; \arg\min_{\gamma\in\mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|\gamma\|_{\mathcal{H}}^2 \right).$$

Kernel solution at x

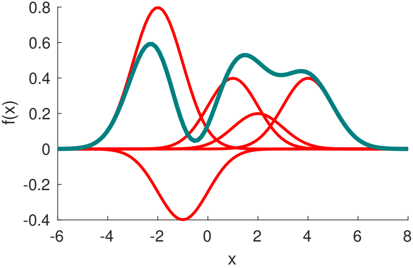
(as weighted sum of y)

$$\hat{\gamma}(x) = \sum_{i=1}^{n} y_i \beta_i(x)$$

$$\beta(x) = (K_{XX} + \lambda I)^{-1} k_{Xx}$$

$$(K_{XX})_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j)\rangle_{\mathcal{H}}$$

$$(k_{Xx})_i = k(x_i, x)$$



Observed covariates: (conditional) ATE

Kernel features (Biometrika 2023):

arXiv > econ > arXiv:2010.04855

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton

NN features (ICLR 2023):

arXiv > cs > arXiv:2210.06610

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

`https://github.com/liyuan9988/DeepFrontBackDoor/`

Observed covariates: (conditional) ATE

Kernel features (Biometrika 2023):


arXiv > econ > arXiv:2010.04855

Search...
Help | Adva

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton

NN features (ICLR 2023):


arXiv > cs > arXiv:2210.06610

Search...
Help | Advan

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/

Average treatment effect

Potential outcome (intervention):

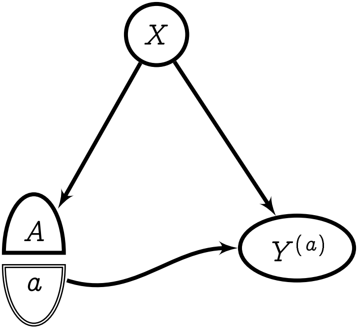
$$\mathbb{E}[Y^{(a)}] = \int \mathbb{E}[Y|a,x]\,dp(x)$$

(the average structural function; in epidemiology, for continuous a, the dose-response curve).

Assume: (1) Stable Unit Treatment Value Assumption (aka "no interference"), (2) Conditional exchangeability $Y^{(a)} \perp\!\!\!\perp A|X$. (3) Overlap.

Example: US job corps, training for disadvantaged youths:

- A: treatment (training hours)
- Y: outcome (percentage employment)
- X: covariates (age, education, marital status, ...)



Multiple inputs via products of kernels

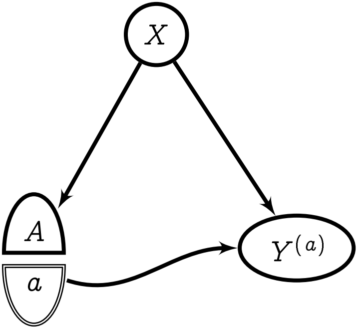
We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[Y | a, x]$$

Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

(argument of kernel/feature map indicates feature space)



Multiple inputs via products of kernels

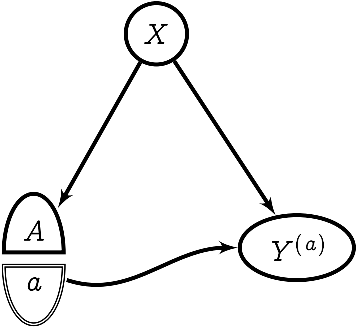
We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[Y | a, x]$$

Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

(argument of kernel/feature map indicates feature space)



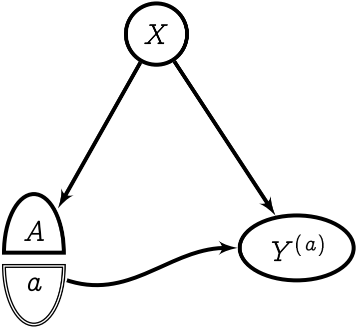
We use outer product of features ($\implies$ product of kernels):

$$\phi(x, a) = \varphi(a) \otimes \varphi(x) \qquad \mathfrak{K}([a, x], [a', x']) = k(a, a')k(x, x')$$

Multiple inputs via products of kernels

We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[\, Y \,|\, a, x \,]$$



Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

(argument of kernel/feature map indicates feature space)

We use outer product of features ($\implies$ product of kernels):

$$\phi(x, a) = \varphi(a) \otimes \varphi(x) \qquad \mathfrak{K}([a, x], [a', x']) = k(a, a')k(x, x')$$

Ridge regression solution:

$$\hat{\gamma}(x, a) = \sum_{i=1}^{n} y_i \beta_i(a, x), \quad \beta(a, x) = \left[K_{AA} \odot K_{XX} + \lambda I\right]^{-1} K_{Aa} \odot K_{Xx}$$

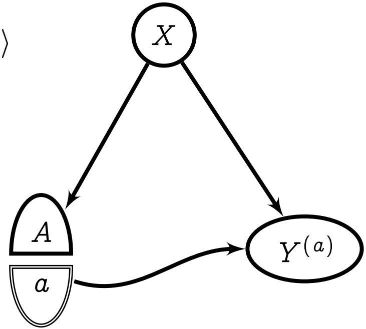
ATE (dose-response curve)

Well-specified setting:

$$\mathbb{E}[Y|a,x] =: \gamma_0(a,x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

ATE as feature space dot product:

$$\begin{aligned} \mathrm{ATE}(a) &= \mathbb{E}[\gamma_0(a,X)] \\ &= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle\right] \end{aligned}$$



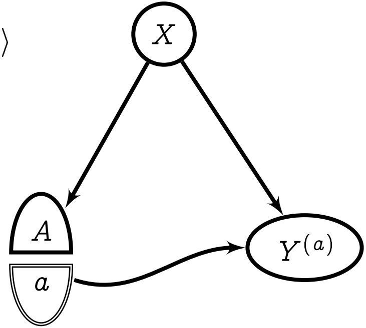
ATE (dose-response curve)

Well-specified setting:

$$\mathbb{E}[Y|a,x] =: \gamma_0(a,x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

ATE as feature space dot product:

$$\begin{aligned} \text{ATE}(a) &= \mathbb{E}[\gamma_0(a,X)] \\ &= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle\right] \\ &= \langle \gamma_0, \varphi(a) \otimes \underbrace{\mu_X}_{\mathbb{E}[\varphi(X)]} \rangle \end{aligned}$$



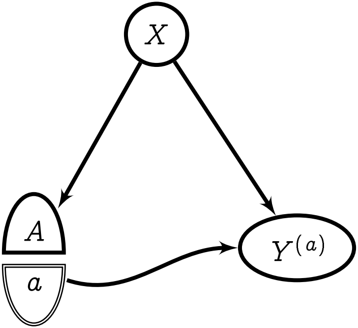
Feature map of probability $P(X)$,

$$\mu_X = [\ldots \mathbb{E}[\varphi_i(X)] \ldots]$$

ATE: example

US job corps: training for disadvantaged youths:

- X: covariate/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employment)



Empirical ATE:

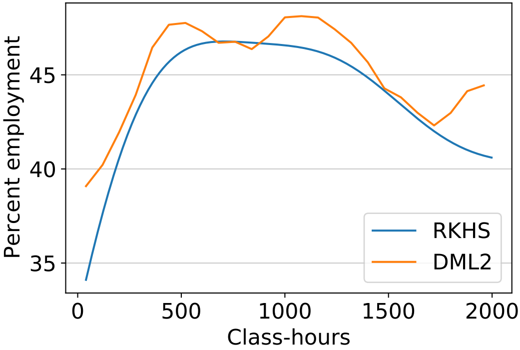
$$\widehat{\text{ATE}}(a) = \hat{\mathbb{E}}\left[\langle \hat{\gamma}_0, \varphi(X) \otimes \varphi(a)\rangle\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1}(K_{Aa} \odot K_{Xx_i})$$

Schochet, Burghardt, and McConnell (2008). Does Job Corps work? Impact findings from the national Job Corps study.

Singh, Xu, G (2022a).

ATE: results



- First 12.5 weeks of classes confer employment gain: from 35% to 47%.
- [RKHS] is our $\widehat{\mathrm{ATE}}(a)$.
- [DML2] Colangelo, Lee (2020), Double debiased machine learning nonparametric inference with continuous treatments.

Singh, Xu, G (2022a)

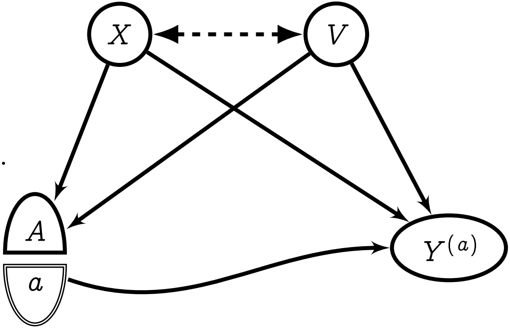
Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[\, Y \,|\, a, x, v] =: \gamma_0(a, x, v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle \,.$$

Conditional ATE

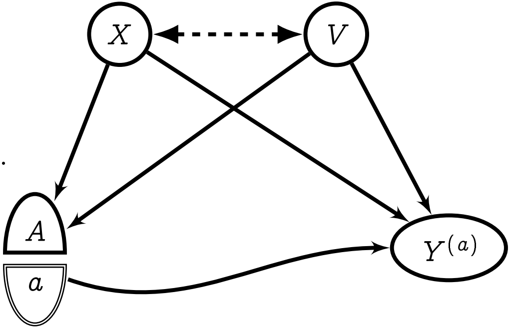
$$\mathrm{CATE}(a, v)$$
$$= \mathbb{E}\left[Y^{(a)} \middle| V = v \right]$$



Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[\,Y\,|\,a, x, v] =: \gamma_0(a, x, v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle \,.$$



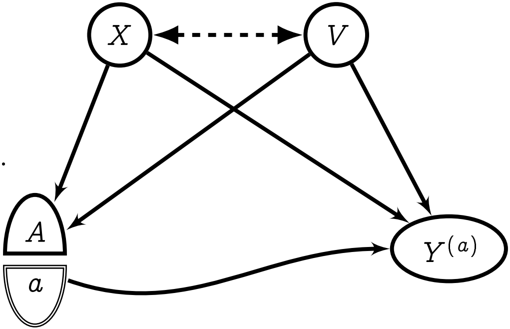
Conditional ATE

$$\text{CATE}(a, v)$$
$$= \mathbb{E}\left[Y^{(a)} \middle| V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$

Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[Y | a, x, v] =: \gamma_0(a, x, v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$



Conditional ATE

$$\mathrm{CATE}(a, v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$
$$= ...?$$

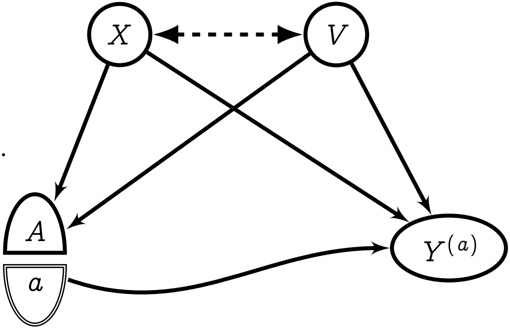
How to take conditional expectation?

Density estimation for $p(X | V = v)$? Sample from $p(X | V = v)$?

Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a)\otimes\varphi(x)\otimes\varphi(v)\rangle .$$



Conditional ATE

$$\mathrm{CATE}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)}\middle| V=v\right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a)\otimes\varphi(X)\otimes\varphi(V)\rangle \,|\, V=v\right]$$
$$= \langle \gamma_0, \varphi(a)\otimes \underbrace{\mathbb{E}[\varphi(X)|\,V=v]}_{\mu_{X|V=v}} \otimes\varphi(v)\rangle$$

Learn conditional mean embedding: $\mu_{X|V=v} := \mathbb{E}_X\left[\varphi(X)|\,V=v\right]$

Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_\mathcal{V} \rightarrow \mathcal{H}_\mathcal{X}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\operatorname{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \operatorname{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\operatorname{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \operatorname{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

A Smooth Operator

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\operatorname{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \mathrm{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)| V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Kernel ridge regression from $\varphi(v)$ to infinite features $\varphi(x)$:

$$\widehat{F} = \operatorname*{argmin}_{F \in HS} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \text{HS}(\mathcal{H}_{\mathcal{V}}, \mathcal{H}_{\mathcal{X}})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_{\mathcal{V}} \quad \forall h \in \mathcal{H}_{\mathcal{X}}$$

Kernel ridge regression from $\varphi(v)$ to infinite features $\varphi(x)$:

$$\widehat{F} = \underset{F \in HS}{\text{argmin}} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_{\mathcal{X}}} + \lambda_2 \|F\|^2_{HS}$$

Ridge regression solution:

$$\mu_{X|V=v} := \mathbb{E}[\varphi(X)|V = v] \approx \widehat{F}\varphi(v) = \sum_{\ell=1}^{n} \varphi(x_\ell)\beta_\ell(v)$$

$$\beta(v) = [K_{VV} + \lambda_2 I]^{-1} k_{Vv}$$

Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6]

$$E_0 = G_1 \circ T_1^{\frac{c_1-1}{2}}, \quad c_1 \in (1,2], \quad \|G_1\|_{HS}^2 \leq \zeta_1,$$

T_1 is covariance of features $\varphi(v)$:

- Eigenspectrum decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

Larger $c_1 \implies$ smoother $E_0 \implies$ easier problem.

[A] Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

[B] Singh, Xu, G (2022a)

Earlier consistency proofs for finite dimensional $\varphi(x)$:

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).

Caponnetto, De Vito (2007).

Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6]

$$E_0 = G_1 \circ T_1^{\frac{c_1 - 1}{2}}, \quad c_1 \in (1, 2], \quad \|G_1\|_{HS}^2 \leq \zeta_1,$$

T_1 is covariance of features $\varphi(v)$:

- Eigenspectrum decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

Larger $c_1 \implies$ smoother $E_0 \implies$ easier problem.

Consistency [A, Theorem 2, Theorem 3]

$$\left\| \widehat{E} - E_0 \right\|_{\mathrm{HS}} = O_P\left(n^{-\frac{1}{2} \frac{c_1 - 1}{c_1 + 1/b_1}} \right),$$

best rate is $O_P(n^{-1/4})$ (minimax)

[A] Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

[B] Singh, Xu, G (2022a)

Earlier consistency proofs for finite dimensional $\varphi(x)$:

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).

Caponnetto, De Vito (2007).

Consistency of CATE

Empirical CATE:

$$\hat{\theta}^{\text{CATE}}(a, v) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1}(K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X|V=v}} \odot K_{Vv})$$

Consistency of CATE

Empirical CATE:

$$\hat{\theta}^{\mathrm{CATE}}(a, v)$$
$$= Y^{\top}(K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1}(K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X|V=v}} \odot K_{Vv})$$

Consistency: [A, Theorem 2]

$$\|\hat{\theta}^{\mathrm{CATE}} - \theta_0^{\mathrm{CATE}}\|_{\infty} = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1//b}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1/b_1}}\right).$$

Follows from consistency of $\widehat{E}$ and $\hat{\gamma}$, under the assumptions:

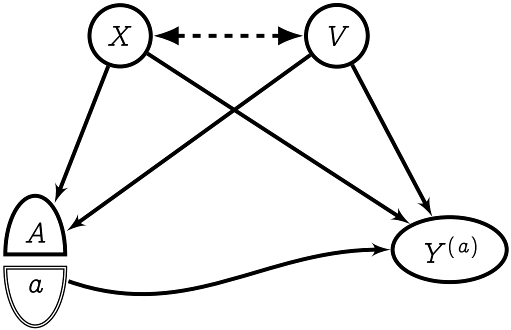
- $E_0 = G_1 \circ T_1^{\frac{c_1-1}{2}}$, $\|G_1\|_{HS}^2 \leq \zeta_1$,
- $\gamma_0 \in \mathcal{H}^c$.

[A] Singh, Xu, G (2022a)

Conditional ATE: example

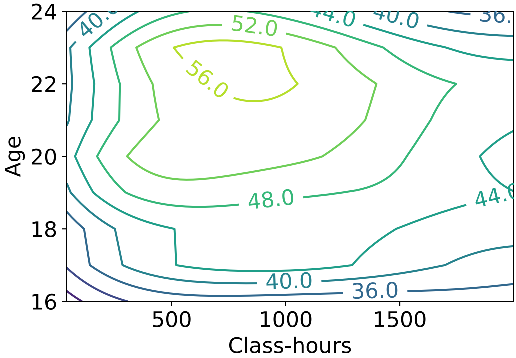
US job corps: training for disadvantaged youths:

- X: confounder/context (education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employed)
- V: age



Singh, Xu, G (2022a)

Conditional ATE: results



Average percentage employment $Y^{(a)}$ for class hours a, conditioned on age v. Given around 12-14 weeks of classes:

- 16 y/o: employment increases from 28% to at most 36%.
- 22 y/o: percent employment increases from 40% to 56%.

Singh, Xu, G (2022a)

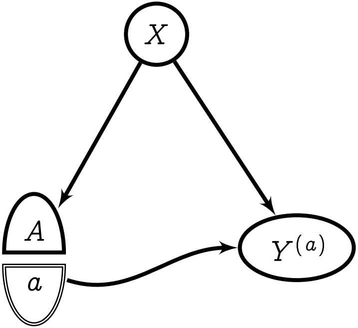
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[Y|a,x] = \gamma_0(a,x)$$

Average treatment on treated:

$$\theta^{ATT}(a, a')$$
$$= \mathbb{E}[y^{(a')}|A = a]$$



Empirical ATT:

$$\hat{\theta}^{\mathrm{ATT}}(a, a')$$

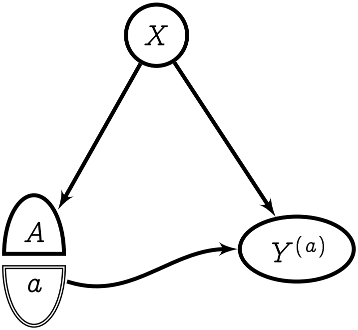
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[Y|a,x] = \gamma_0(a,x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

Average treatment on treated:

$$\theta^{ATT}(a, a')$$
$$= \mathbb{E}[y^{(a')}|A = a]$$



Empirical ATT:

$$\hat{\theta}^{\mathrm{ATT}}(a, a')$$

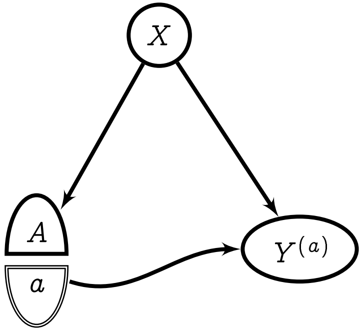
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[Y|a,x] = \gamma_0(a,x)$$

Average treatment on treated:

$$\begin{aligned}\theta^{ATT}(a, a') &= \mathbb{E}[y^{(a')}|A=a] \\ &= \mathbb{E}_P\left[\langle \gamma_0, \varphi(a') \otimes \varphi(X)\rangle \,|A=a\right] \\ &= \langle \gamma_0, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A=a]}_{\mu_{X|A=a}}\rangle\end{aligned}$$



Empirical ATT:

$$\hat{\theta}^{\text{ATT}}(a, a')$$

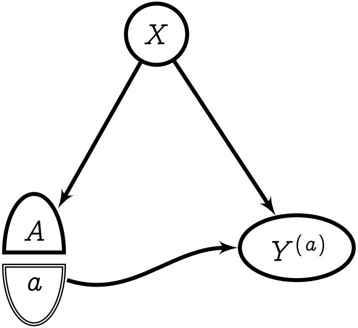
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[Y|a,x] = \gamma_0(a,x)$$

Average treatment on treated:

$$\begin{aligned}
&\theta^{ATT}(a, a') \\
&= \mathbb{E}[y^{(a')}|A=a] \\
&= \mathbb{E}_P\left[\langle \gamma_0, \varphi(a') \otimes \varphi(X)\rangle \,|A=a\right] \\
&= \langle \gamma_0, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A=a]}_{\mu_{X|A=a}}\rangle
\end{aligned}$$



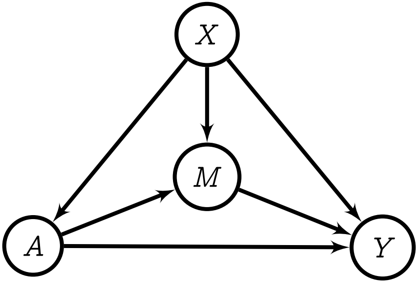
Empirical ATT:

$$\begin{aligned}
&\hat{\theta}^{\mathrm{ATT}}(a, a') \\
&= Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot \underbrace{K_{XX}(K_{AA} + n\lambda_1 I)^{-1} K_{Aa}}_{\text{from } \hat{\mu}_{X|A=a}})
\end{aligned}$$

Mediation analysis

- Direct path from treatment A to effect Y
- Indirect path $A \rightarrow M \rightarrow Y$
- X: context

Is the effect Y mainly due to A? To M?

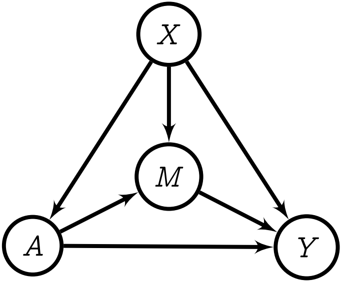


Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)

$\gamma_0(a, m, x) \approx \mathbb{E}[Y|A = a, M = m, X = x]$

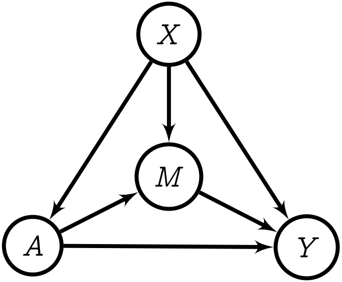


Singh, Xu, G (2022b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$\gamma_0(a, m, x) \approx \mathbb{E}[Y | A = a, M = m, X = x]$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_0(a', M, X) \mathrm{d}\mathbb{P}(M | A = a, X) d\mathbb{P}(X)$$

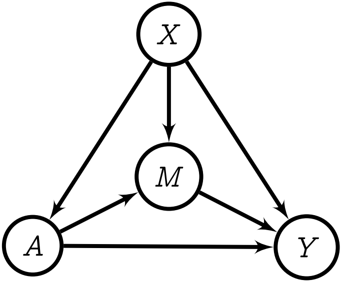
Effect of intervention a', with $M^{(a)}$ as if intervention were a

Singh, Xu, G (2022b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$\gamma_0(a, m, x) \approx \mathbb{E}[Y | A = a, M = m, X = x]$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_0(a', M, X) \mathrm{d}\mathbb{P}(M | A = a, X) d\mathbb{P}(X)$$
$$= \langle \gamma_0, \varphi(a') \otimes \mathbb{E}_P\{\mu_{M|A=a,X} \otimes \varphi(X)\} \rangle$$

Effect of intervention a', with $M^{(a)}$ as if intervention were a

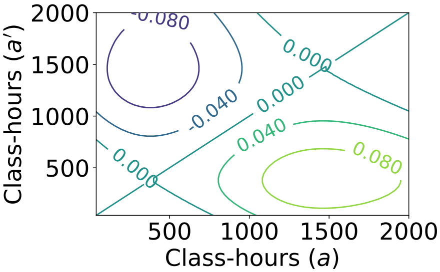
Singh, Xu, G (2022b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: results

Total effect:

$$\theta_0^{TE}(a, a')$$

$$:= \mathbb{E}[Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



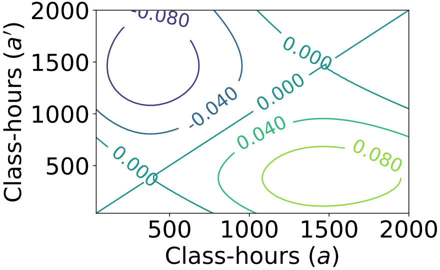
- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests

Singh, Xu, G (2022b)

Mediation analysis: results

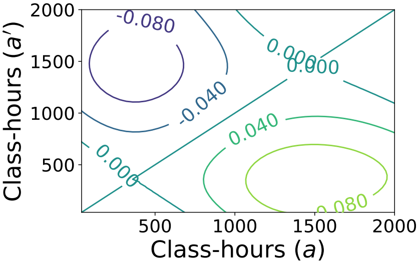
Total effect:

$$\theta_0^{TE}(a, a') := \mathbb{E}[Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



Direct effect:

$$\theta_0^{DE}(a, a') := \mathbb{E}[Y^{\{a', M^{(a)}\}} - Y^{\{a, M^{(a)}\}}]$$

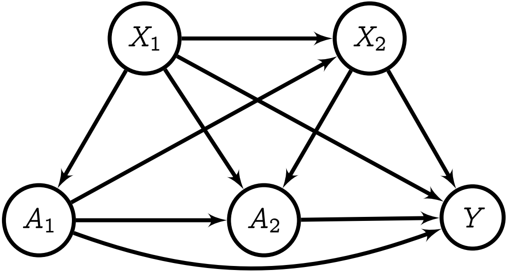


- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests
- Indirect effect mediated via employment effectively zero

Singh, Xu, G (2022b)

...dynamic treatment effect...

Dynamic treatment effect: sequence A_1, A_2 of treatments.



- potential outcomes $Y^{(a_1)}, Y^{(a_2)}, Y^{(a_1,a_2)}$,
- counterfactuals $\mathbb{E}\left[Y^{(a_1',a_2')} | A_1 = a_1, A_2 = a_2\right]$...

(c.f. the Robins G-formula)

Singh, Xu, G. (2022b) Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects

Conclusions

Kernel and neural net solutions:

- ...for ATE, CATE, dynamic treatment effects
- ...with treatment A, covariates X, V, proxies (W, Z) multivariate, "complicated"
- Convergence guarantees for kernels and NN

Next lecture:

- Unobserved covariates/confounders (IV and proxy methods)

Code available for all methods

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT