Kernel Methods for Causal Effect Estimation

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Google DeepMind

Modern Kernel Methods, EMS 2026

Talk on behalf of Mattes Mollenhauer

Outline

Recap, kernel ridge regression

Causal effect estimation, observed covariates:

- Heterogeneous response curve (HR)

Causal effect estimation, hidden covariates:

- ... instrumental variables (and 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

Model fitting: kernel ridge regression

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

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

Model fitting: kernel ridge regression

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

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

Infinite dimensional solution at x:

$$\hat{\gamma}(x) = \widehat{C}_{YX}(\widehat{C}_X + \lambda)^{-1}\varphi(x)$$

$$\widehat{C}_{YX} = \frac{1}{n}\sum_{i=1}^{n}[y_i\, \varphi(x_i)]$$

$$\widehat{C}_X = \frac{1}{n}\sum_{i=1}^{n}[\varphi(x_i) \otimes \varphi(x_i)]$$

Notation : $(f \otimes g)h = f\langle g, h\rangle_{\mathcal{H}_\mathcal{X}}$

Model fitting: kernel ridge regression

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

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

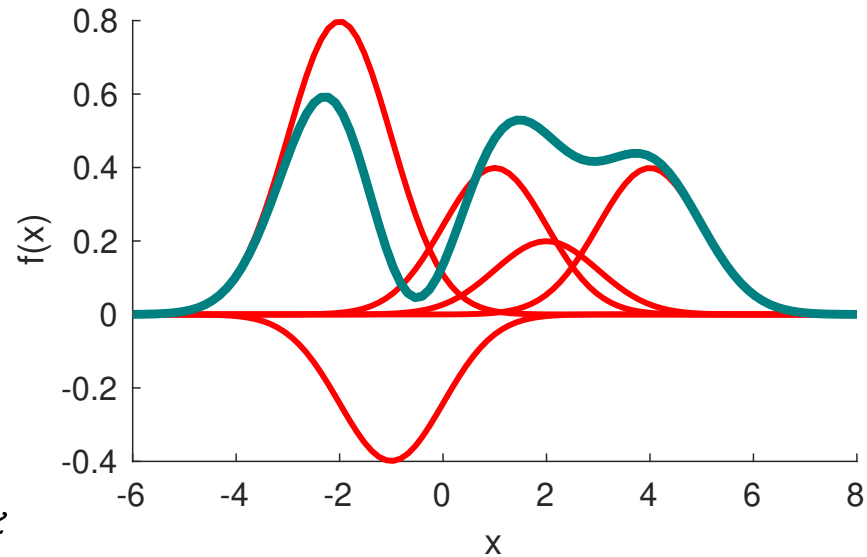
Kernel solution at x
(as weighted sum of y)

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

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

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

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



KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X)g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

$\eta_i\, e_i(x) = \int k(x, x')e_i(x')dP_X(x')$, assume for simplicity: $\mathrm{supp}(P_X) = \mathcal{X}$

Assume problem well specified:

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X)g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

$\eta_i\, e_i(x) = \int k(x, x')e_i(x')dP_X(x')$, assume for simplicity: $\mathrm{supp}(P_X) = \mathcal{X}$

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}$, $\quad c \in (1, 2]$

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X)g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

$\eta_i e_i(x) = \int k(x, x') e_i(x') dP_X(x')$, assume for simplicity: $\mathrm{supp}(P_X) = \mathcal{X}$

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}$, $\quad c \in (1, 2]$

$$\gamma_* = \sum_i \gamma_{*,i}\, e_i \quad \text{where} \quad \|\gamma_*\|^2_{\mathcal{H}_\mathcal{X}^c} = \sum_{i=1}^{\infty} \frac{\gamma_{*,i}^2}{\eta_i^c} < \infty$$

  - Larger $c \implies$ smoother $\gamma_* \implies$ easier problem.

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X)g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

$\eta_i e_i(x) = \int k(x, x') e_i(x') dP_X(x')$, assume for simplicity: $\mathrm{supp}(P_X) = \mathcal{X}$

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}$, $\quad c \in (1, 2]$

$$\gamma_* = \sum_i \gamma_{*,i}\, e_i \quad \text{where} \quad \|\gamma_*\|^2_{\mathcal{H}_\mathcal{X}^c} = \sum_{i=1}^{\infty} \frac{\gamma_{*,i}^2}{\eta_i^c} < \infty$$

  - Larger $c \implies$ smoother $\gamma_* \implies$ easier problem.
- Eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$
  - Larger $b \implies$ easier problem

KRR: consistency in RKHS norm

Consistency for models in class $\mathcal{M}(c, b)$ [A, Theorem 1.ii]

$$\|\hat{\gamma} - \gamma_*\|^2_{L_2(\mathcal{X})} = O_P\left(n^{-\frac{c}{c+1/b}}\right) \qquad \|\hat{\gamma} - \gamma_*\|^2_{\mathcal{H}_\mathcal{X}} = O_P\left(n^{-\frac{c-1}{c+1/b}}\right)$$

Example:

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{H}_\mathcal{X} := W^{\nu,2}(\mathcal{X})$ Sobolev space with $\nu > \frac{d}{2}$ (Matérn kernel) gives $b = \frac{2\nu}{d}$ [B, Corollary 2].

Then $\gamma_* \in \mathcal{H}^c_\mathcal{X}$ with $c \in (1, 2]$ gives consistency in RKHS and L_2.

[A] Fischer, Steinwart (2020). Sobolev norm learning rates for regularized least-squares algorithms.
[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm

Causal effects, observed covariates

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

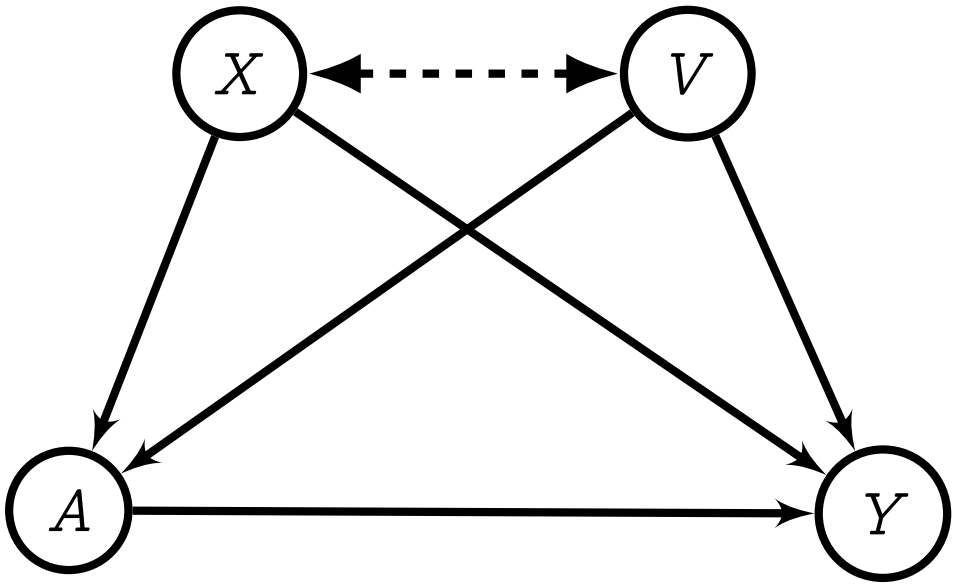
Code for NN and kernel causal estimation with observed covariates:

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

Heterogeneous response curve

US job corps:

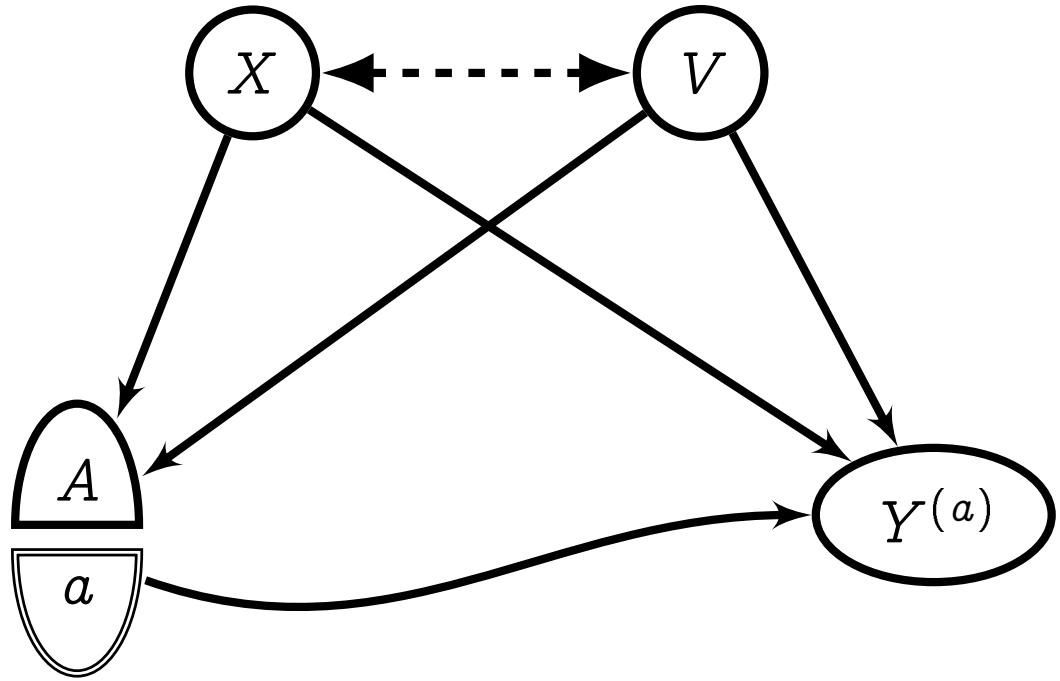
- X: confounder/context (education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employed)
- V: age



Heterogeneous response curve

US job corps:

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



Intervention: $\mathbb{E}\left[Y^{(a)} | V = v\right]$

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

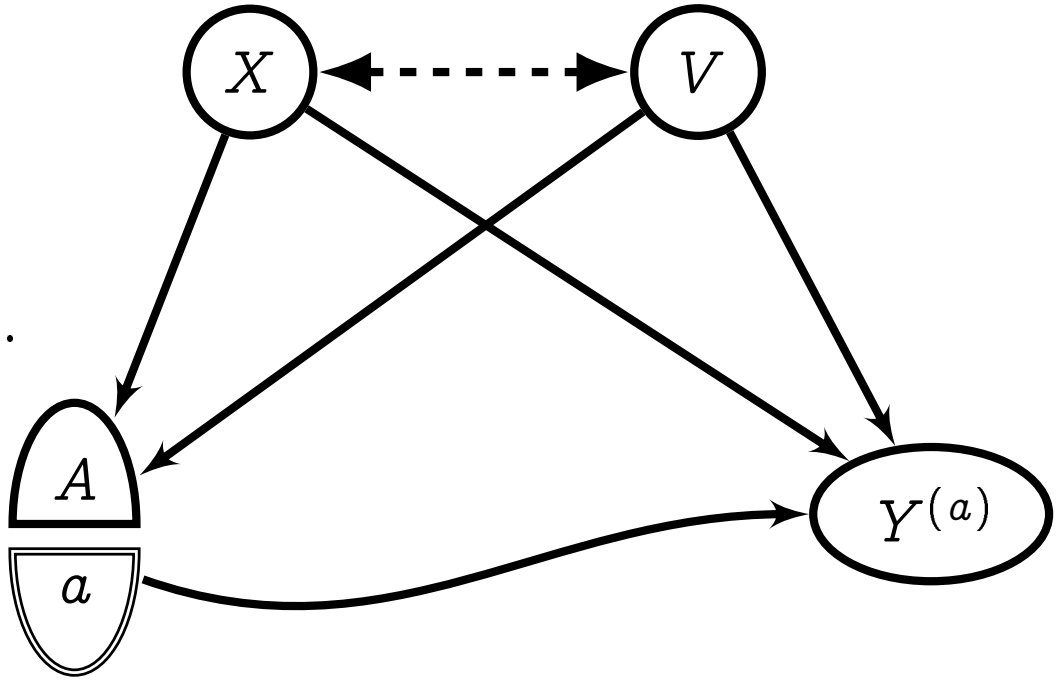
Heterogeneous response curve

Well-specified setting:

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

Heterogeneous response:

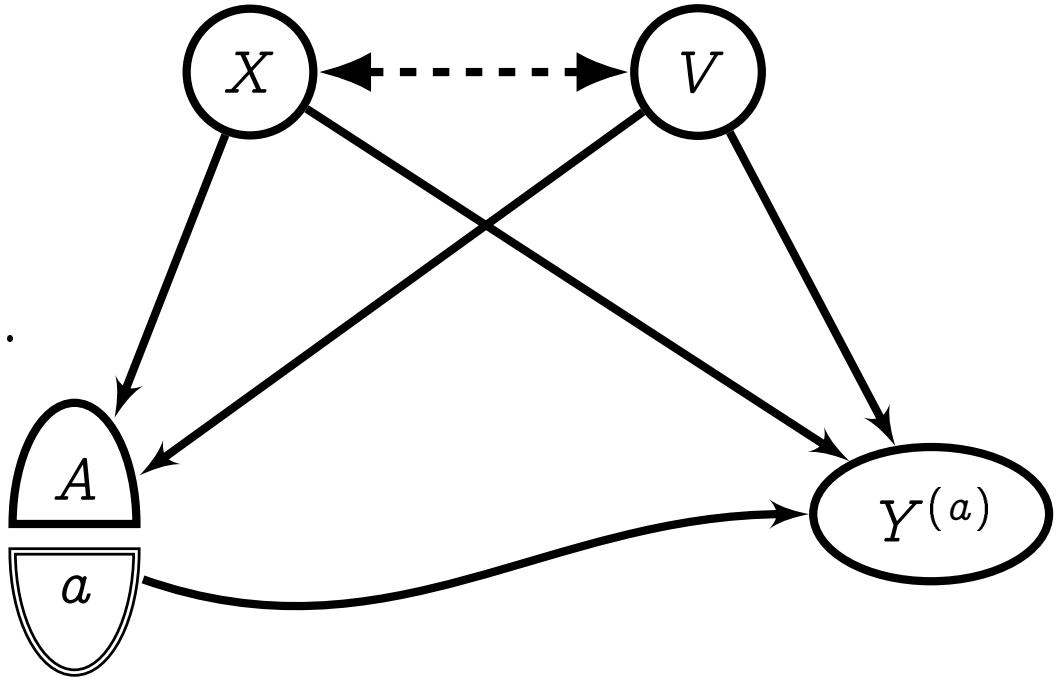
$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$



Heterogeneous response curve

Well-specified setting:

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



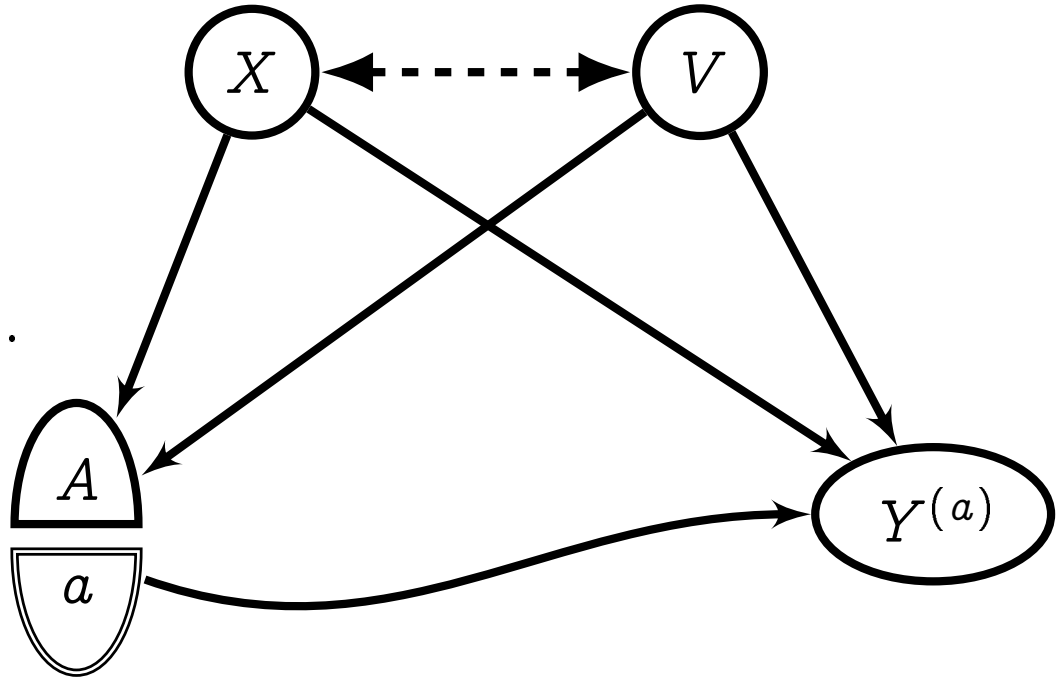
Heterogeneous response:

$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$

Heterogeneous response curve

Well-specified setting:

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



Heterogeneous response:

$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$
$$= \ldots ?$$

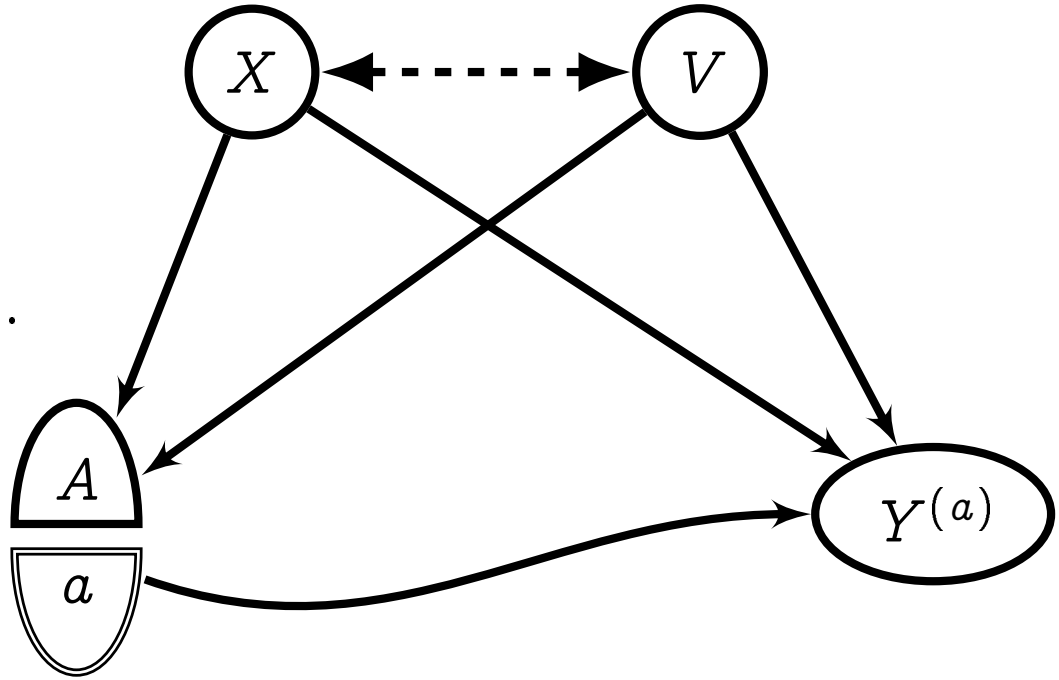
How to take conditional expectation?

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

Heterogeneous response curve

Well-specified setting:

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



Heterogeneous response:

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

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

Regressing from feature space to feature space

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

$$F_* \varphi(v) = \mathbb{E}_X \left[\varphi(X) \middle| V = v \right] =: \mu_{X|V=v}$$

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 $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

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 $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

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 $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

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 $\varphi(x)$:

$$\widehat{F} = \operatorname*{argmin}_{F \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

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 $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

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 $\varphi(x)$:

$$\widehat{F} = \operatorname*{argmin}_{F \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

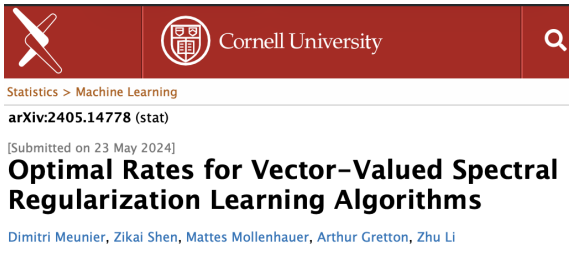
Ridge regression solution:

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

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

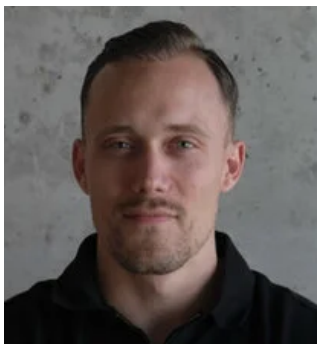
Consistency of Vector Valued Least Squares

NeurIPS (2022)



NeurIPS (2024)

JMLR (2024)



Consistency of Vector Valued Least Squares

Define C_V as covariance of features $\varphi(v)$,

$$C_V = \mathbb{E}(\varphi(V) \otimes \varphi(V)), \qquad \mathbb{E}_V(f(V)g(V)) = \langle f, C_V g \rangle_{\mathcal{H}_\mathcal{V}}$$

- Assume problem well specified [C, Assumption 4]

$$F_* \in S_2(\mathcal{H}_\mathcal{V}^{c_1}, \mathcal{H}_\mathcal{X}), \quad c_1 \in (1, 2],$$

  larger $c_1 \implies$ smoother $F_* \implies$ easier problem.
- Eigenspectrum of C_V decays as $\eta_j(C_V) \sim j^{-b_1}$, $b_1 \geq 1$.

[A] Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized CME Learning
[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm
[C] Singh, Xu, G (2023)
Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).
Caponnetto, De Vito (2007).

Consistency of Vector Valued Least Squares

Define C_V as covariance of features $\varphi(v)$,

$$C_V = \mathbb{E}(\varphi(V) \otimes \varphi(V)), \qquad \mathbb{E}_V(f(V)g(V)) = \langle f, C_V g \rangle_{\mathcal{H}_\mathcal{V}}$$

- Assume problem well specified [C, Assumption 4]

$$F_* \in S_2(\mathcal{H}_\mathcal{V}^{c_1}, \mathcal{H}_\mathcal{X}), \quad c_1 \in (1, 2],$$

larger $c_1 \implies$ smoother $F_* \implies$ easier problem.

- Eigenspectrum of C_V decays as $\eta_j(C_V) \sim j^{-b_1}$, $b_1 \geq 1$.

Consistency [A, Theorem 2, Theorem 3] (minimax)

$$\left\| \widehat{F} - F_* \right\|_{\text{HS}}^2 = O_P\left(n^{-\frac{c_1 - 1}{c_1 + 1/b_1}} \right),$$

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

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

[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm

[C] Singh, Xu, G (2023)

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

Caponnetto, De Vito (2007).

Heterogeneous Response from Samples

Empirical HR:

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

[A] Singh, Xu, G. (2023)

Heterogeneous Response from Samples

Empirical HR:

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

Consistency: [A, Theorem 2]

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

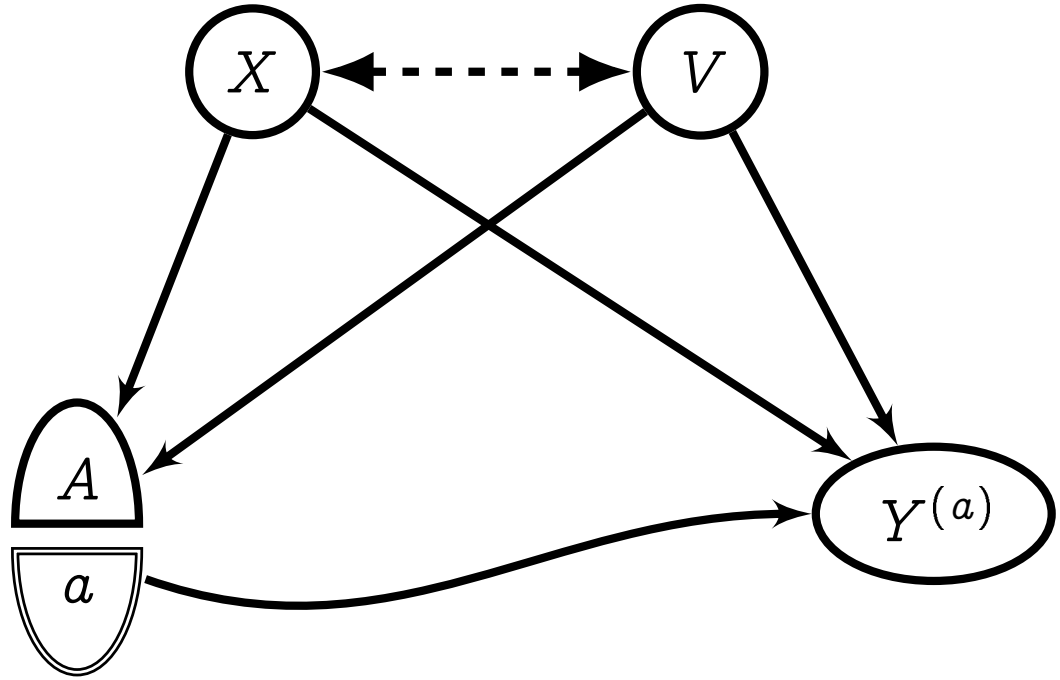
Follows from consistency of $\widehat{F}$ and $\hat{\gamma}$

[A] Singh, Xu, G. (2023)

Heterogeneous response: example

US job corps:

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

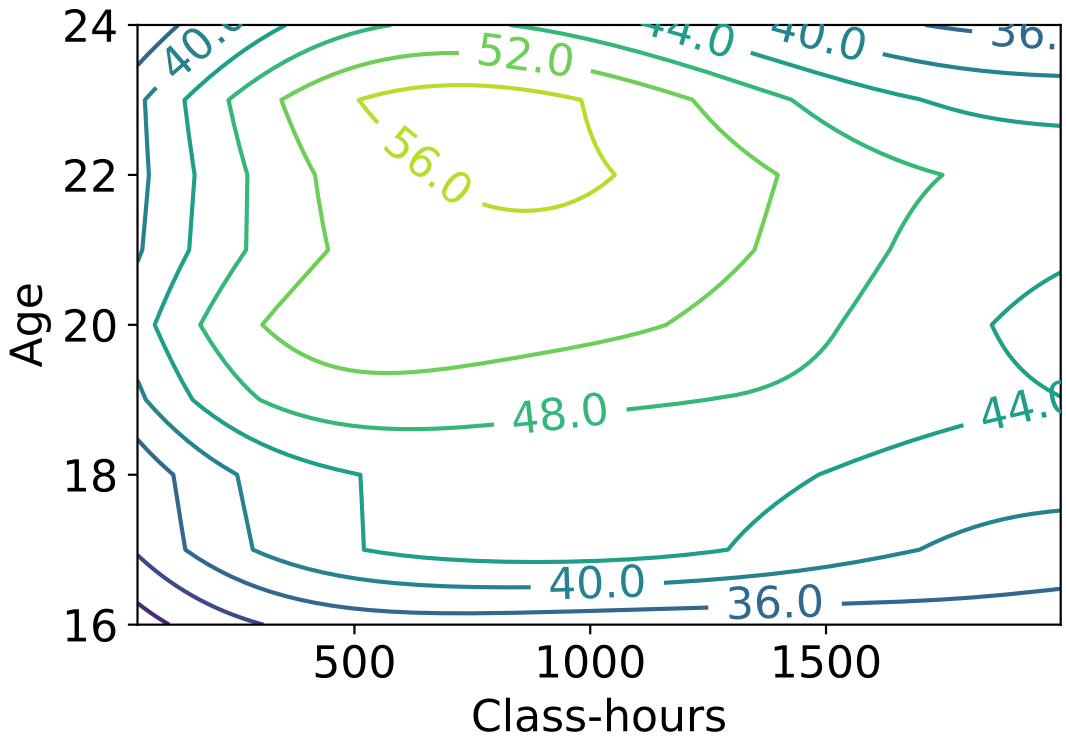


Empirical HR:

$$\widehat{\mathrm{HR}}(a, v) = \langle \hat{\gamma}_0, \varphi(a) \otimes \underbrace{\hat{F}\varphi(v)}_{\widehat{\mathbb{E}}[\varphi(X)\,|\,V=v]} \otimes \varphi(v) \rangle$$

[A] Singh, Xu, G. (2023)

Heterogeneous response: 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%.

[A] Singh, Xu, G. (2023)

Instrumental variable regression

Instrumental variable regression with kernels

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

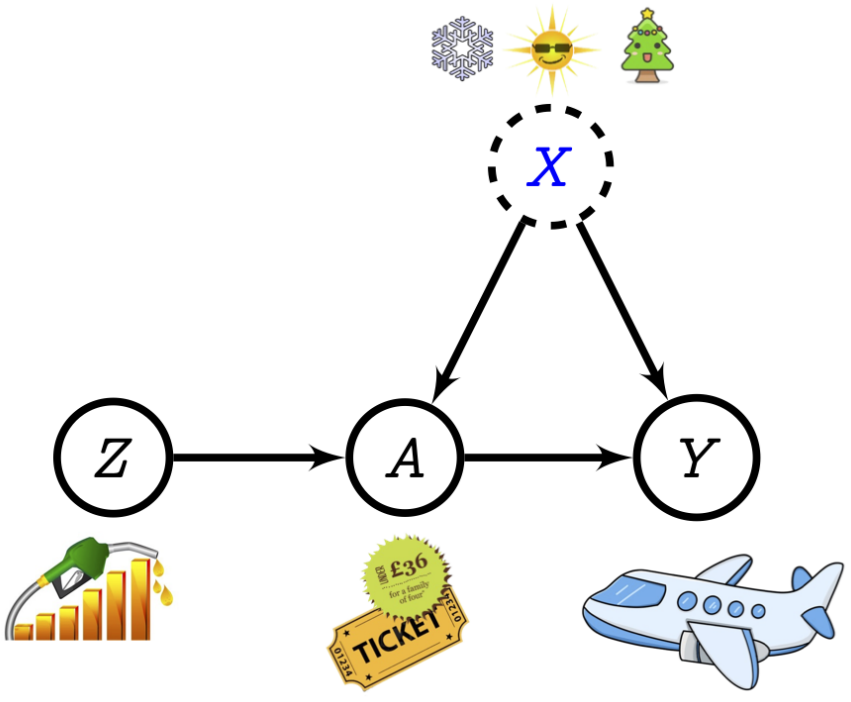
Assumptions

$$\mathbb{E}[X|Z] = 0$$

$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

$$Y = \langle \gamma, \phi(A) \rangle + X$$

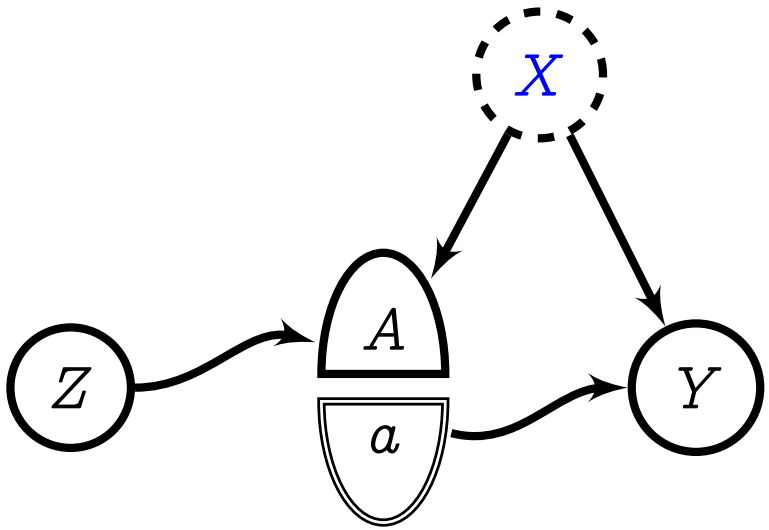


Instrumental variable regression with kernels

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

Assumptions

$$\mathbb{E}[X|Z] = 0$$
$$Z \not\perp\!\!\!\perp A$$
$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$
$$Y = \langle \gamma, \phi(A) \rangle + X$$

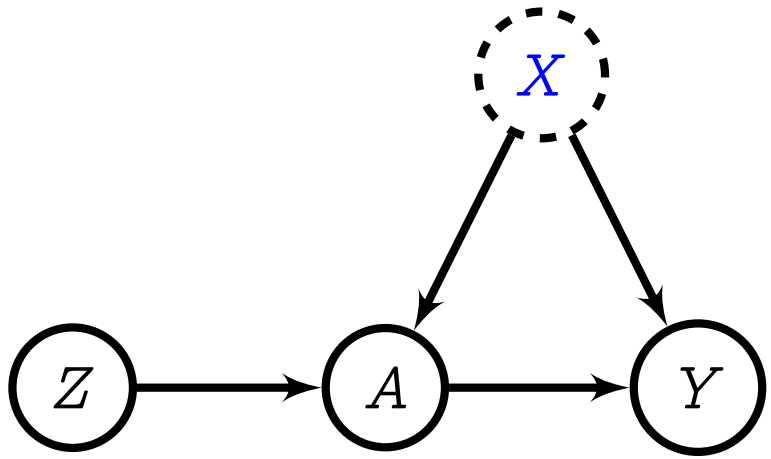


Dose-response curve:

$$\mathrm{DR}(a) = \int \mathbb{E}(Y|X,a)\,dp(X) = \langle \gamma, \phi(a) \rangle$$

Instrumental variable regression with kernels

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



Assumptions

$$\mathbb{E}[X|Z] = 0$$

$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

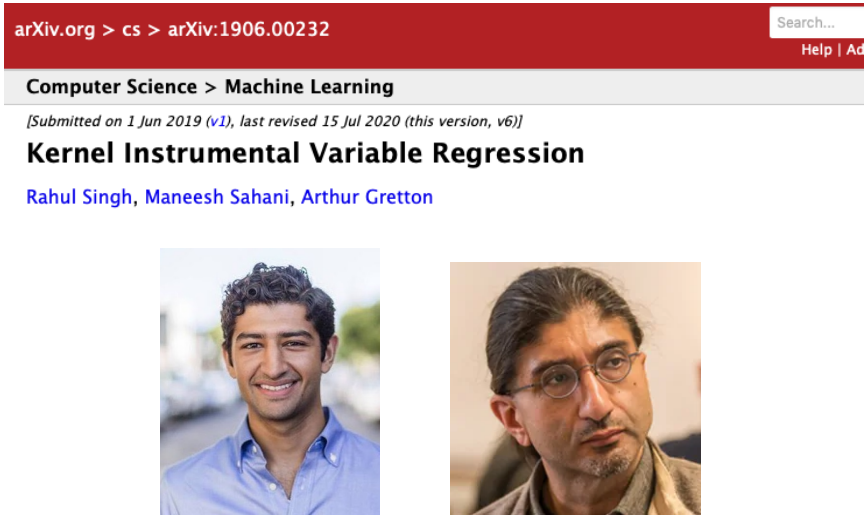
$$Y = \langle \gamma, \phi(A) \rangle + X$$

IV regression: Condition both sides on Z,

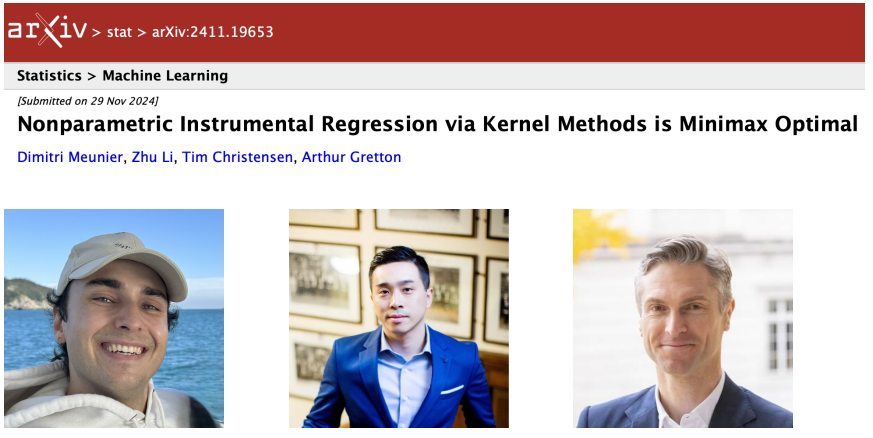
$$\mathbb{E}[Y|Z] = \mathbb{E}\left[\langle \gamma, \varphi(A)]\rangle \,|Z\right] + \underbrace{\mathbb{E}[X|Z]}_{=0}$$

Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):



Consistency (submitted):



Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/

IV using kernel features

Stage 2 regression (IV): Solve for γ with RR loss:

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \left\langle \gamma, \widehat{\mathbb{E}}\left[\varphi(A)|z_\ell\right] \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

Stage 1 regression: define $\widehat{F} \in S_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{A}})$:

$$\widehat{\mathbb{E}}[\varphi(A)|z] = \widehat{F}\varphi(z)$$

Two stage least squares

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regression

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Stage 2 loss naturally expressed in weak metric:

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)}$$

where $\mathcal{T}\gamma = \mathbb{E}(\gamma(A)|Z)$.

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Stage 2 loss naturally expressed in weak metric:

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)}$$

where $\mathcal{T}\gamma = \mathbb{E}(\gamma(A)|Z)$.

By Jensen,

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)} \leq \|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Weak control does not imply strong control: need additional link assumption.

Strong/weak geometry, link condition

Strong geometry:

$$\begin{aligned}\|\gamma\|^2_{L_2(A)} &= \mathbb{E}\left[\langle\gamma, \varphi(A)\rangle^2_{\mathcal{H}_\mathcal{A}}\right] \\ &= \langle\gamma, C_A\gamma\rangle_{\mathcal{H}_\mathcal{A}}\end{aligned}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Strong/weak geometry, link condition

Strong geometry:

$$\|\gamma\|^2_{L_2(A)} = \mathbb{E}\left[\langle \gamma, \varphi(A)\rangle^2_{\mathcal{H}_A}\right] = \langle \gamma, C_A \gamma\rangle_{\mathcal{H}_A}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Weak geometry:

$$\|\mathcal{T}\gamma\|^2_{L_2(Z)} = \mathbb{E}\left[\mathbb{E}\left[\langle \gamma, \varphi(A)\rangle_{\mathcal{H}_A} | Z\right]^2\right] = \langle \gamma, C_M \gamma\rangle_{\mathcal{H}_A}$$

where

$$C_M = \mathbb{E}\left[\mu_{A|Z} \otimes \mu_{A|Z}\right],$$
$$\mu_{A|Z} := \mathbb{E}\left[\varphi(A)|Z\right].$$

Strong/weak geometry, link condition

Strong geometry:

$$\|\gamma\|^2_{L_2(A)} = \mathbb{E}\left[\langle \gamma, \varphi(A)\rangle^2_{\mathcal{H}_\mathcal{A}}\right] = \langle \gamma, C_A \gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Weak geometry:

$$\|\mathcal{T}\gamma\|^2_{L_2(Z)} = \mathbb{E}\left[\mathbb{E}\left[\langle \gamma, \varphi(A)\rangle_{\mathcal{H}_\mathcal{A}} \,|\, Z\right]^2\right] = \langle \gamma, C_M \gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_M = \mathbb{E}\left[\mu_{A|Z} \otimes \mu_{A|Z}\right], \qquad \mu_{A|Z} := \mathbb{E}\left[\varphi(A)|Z\right].$$

Link condition: $\exists \zeta > 0$, $\theta \geq 1$ such that

$$C_A \preceq \zeta C_M^{1/\theta}$$

Earlier link conditions two-sided $C_A^\gamma \asymp C_M$: Nair, Pereverzev, Tautenhahn (2005); Chen, Reiss (2011).

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of $\gamma_* \in \mathcal{H}^c_{\mathcal{A}}$, b effective dimension, θ ill-posedness.

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of $\gamma_* \in \mathcal{H}^c_{\mathcal{A}}$, b effective dimension, θ ill-posedness.

Assume: Stage 1 samples $m = n^\alpha$ sufficiently large relative to Stage 2 samples n. Then

$$\|\hat{\gamma} - \gamma_*\|^2_{L_2(A)} = O_P(n^{-\frac{c}{c+1/b+\theta-1}}).$$

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of $\gamma_* \in \mathcal{H}_{\mathcal{A}}^c$, b effective dimension, θ ill-posedness.

Assume: Stage 1 samples $m = n^\alpha$ sufficiently large relative to Stage 2 samples n. Then

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}^2 = O_P(n^{-\frac{c}{c+1/b+\theta-1}}).$$

Lower bound (minimax):

$$\inf_{\hat{\gamma}} \sup_{\text{NPIV models in } \mathcal{M}(c,b,\theta)} \|\hat{\gamma} - \gamma_*\|_{L_2(A)}^2 \gtrsim n^{-\frac{c}{c+1/b+\theta-1}}$$

For $\theta = 1$ recover nonparametric rate for regression.

Conclusions

Kernel (and neural net) solutions:

- ...for dose-response, heterogeneous response, dynamic treatment effects
- ...with treatment A, covariates X, V, multivariate, "complicated"
- Convergence guarantees

Unobserved confounding:

- IV (and proxy) methods

Code available for all methods

Research support

Work supported by:

The Gatsby Charitable Foundation

Google DeepMind

Google DeepMind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT