Kernel Methods for Long Term Causal Inference

Treatment Effects, Dose Responses, Counterfactual Distributions

Rahul Singh

Simons Institute
MIT Economics

Algorithmic Aspects of Causal Inference 2022

Outline

1 Motivation

2 Model

3 Algorithm

4 Theory

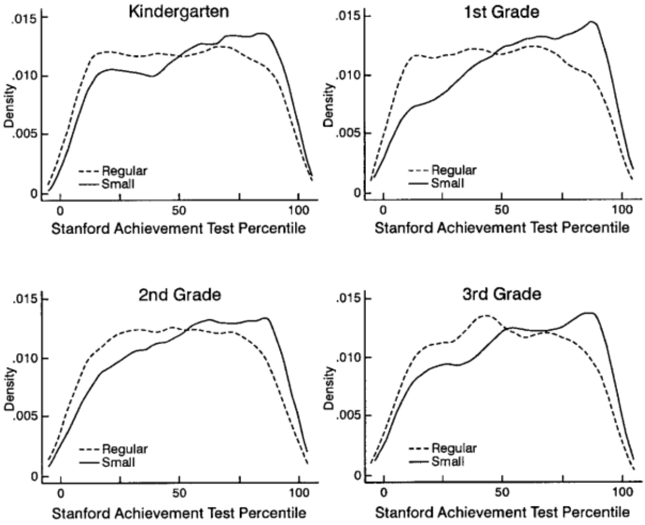
Motivation: Project STAR

© Courtesy of Tennessee State University/Project STAR

- Tennessee Student/Teacher Achievement Ratio experiment
- random assignment of class sizes for 1985 kindergarten cohort
- 11,600 elementary school students and teachers

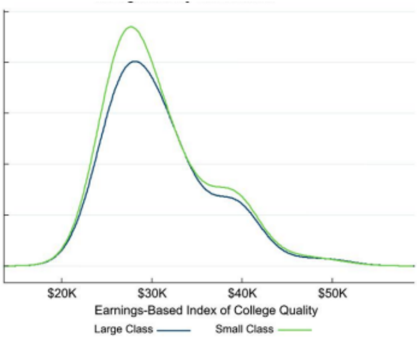
Motivation: Project STAR

- what are the effects?
- short term: test scores (Krueger 1999)

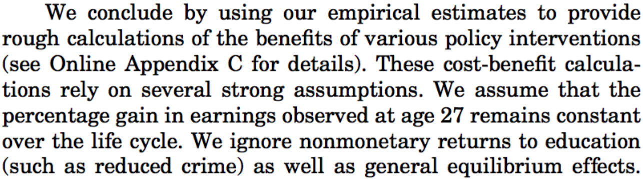


Motivation: Project STAR

- medium term: college quality and earnings at 27 (Chetty et al. 2011)



- long term: earnings at 65? (Chetty et al. 2011)



We conclude by using our empirical estimates to provide rough calculations of the benefits of various policy interventions (see Online Appendix C for details). These cost-benefit calculations rely on several strong assumptions. We assume that the percentage gain in earnings observed at age 27 remains constant over the life cycle. We ignore nonmonetary returns to education (such as reduced crime) as well as general equilibrium effects.

Motivation: Limitations of (quasi) experiments

The limited scope of (quasi) experiments is a more general issue

- experiments are expensive and therefore *short term*
  - Project STAR cost $12 million over 4 years (Word et al. 1990)
- even administrative data only takes us to the *medium term*
  - only up to age 27 (Chetty et al. 2011)
- but cost-benefit analysis depends on the *long term*
  - which policies pay for themselves (Hendren + Sprung-Keyser 2020)
- the "constant effect over time" assumption often fails
  - e.g. test score effects fade (Chetty et al. 2011)

So what can we say about long term effects in a principled way?

Main idea

- combine (quasi) experimental data with observational data
  1. experimental: treatment D and medium term M
  2. observational: medium term M and long term Y

- kernel estimators with closed forms and finite sample guarantees
  1. treatment effects: $\sqrt{n}$-consistency, Gaussian approx., efficiency
  2. dose responses: uniform consistency
  3. counterfactual distributions: convergence in distribution

- unifying framework for flexible estimation
  - key assumption: conditional distribution and regression are smooth
  - “multiple spectral robustness”

Related work

- statistical surrogates
  - long history (Prentice 1989; Begg + Leung 2000; Frangakis + Rubin 2002; Chen et al. 2009)
  - recent revival (Rosenman et al. 2018; Athey et al. 2020a, 2020b, Rosenman et al. 2020; Kallus + Mao 2020)

- semiparametric theory
  - multiply robust moment (Chen + Ritzwoller 2021)
  - imposing linearity and separability (Battocchi et al. 2021)
  - debiased machine learning (Chernozhukov et al. 2016; S. 2021)

- nonparametric theory
  - structural functions with sample selection (Das et al. 2003)
  - uniform analysis of causal kernel methods (S. et al 2020)
  - sequential mean embedding (S. et al. 2021)

The estimators and their guarantees are new

Outline

1 Motivation

2 Model

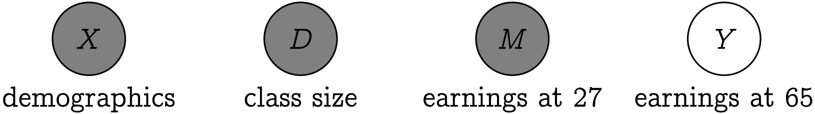
3 Algorithm

4 Theory

Model: Data setting

There are two data sets (Chetty et al. 2011)

1. Experimental ($G = 0$)



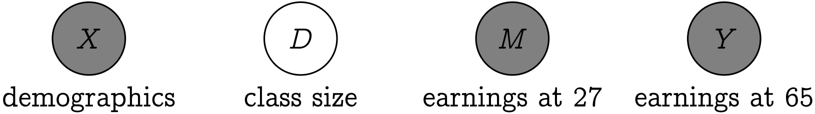
X demographics

D class size

M earnings at 27

Y earnings at 65

2. Observational ($G = 1$)



X demographics

D class size

M earnings at 27

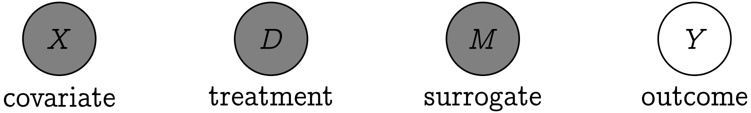
Y earnings at 65

What can we say about the effect of class size on earnings at 65?

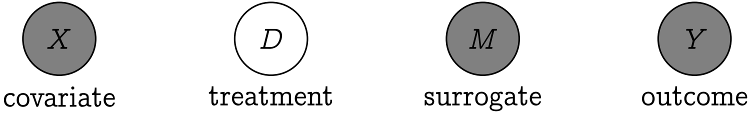
Model: Data setting

There are two data sets (Athey et al. 2020b)

1. Experimental ($G = 0$)



2. Observational ($G = 1$)

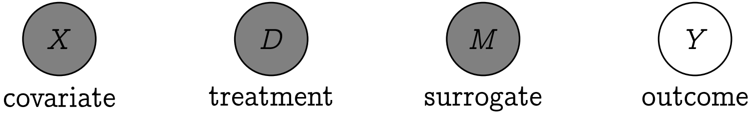


What can we say about the effect of treatment on outcome?

Model: Data setting

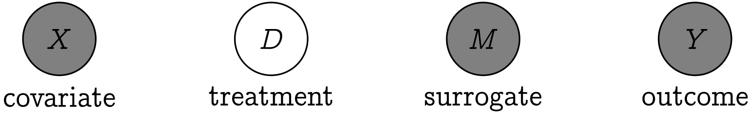
Define the fused data set

1. Experimental ($G = 0$)



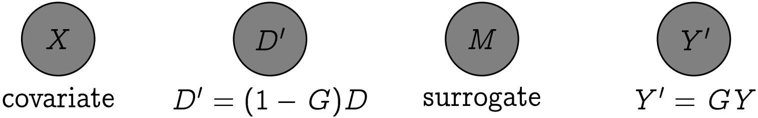
X covariate | D treatment | M surrogate | Y outcome

2. Observational ($G = 1$)



X covariate | D treatment | M surrogate | Y outcome

3. Fused



X covariate | D' $D' = (1 - G)D$ | M surrogate | Y' $Y' = GY$

We observe (G, X, D', M, Y'); c.f. (Rubin 1976; Heckman 1979)

Model: Identification

Assume the following (Athey et al. 2020b)

SUTVA

1. No interference: if $D = d$ then $M = M^{(d)}$ and $Y = Y^{(d)}$.

Treatment mechanism for experimental sample

1. Unconfounded treatment: $D \perp\!\!\!\perp \{M^{(d)}, Y^{(d)}\} | G = 0, X$.
2. Surrogacy: $D \perp\!\!\!\perp Y | G = 0, X, M$.
3. Treatment overlap: if $f(G = 0, x) > 0$ then $f(d | G = 0, x) > 0$.

Selection mechanism

1. Unconfounded selection: $G \perp\!\!\!\perp \{M^{(d)}, Y^{(d)}\} | X$.
2. Comparability: $G \perp\!\!\!\perp Y | X, M$.
3. Selection overlap: if $f(x) > 0$ then $\mathbb{P}(G = 0 | x) > 0$; if $f(x, m) > 0$ then $\mathbb{P}(G = 1 | x, m) > 0$.

Model: Identification

Consider the causal parameter $\theta_0(d) = \mathbb{E}[Y^{(d)}]$.

Theorem (Athey et al. 2020b): Under the stated assumptions

$$\theta_0(d) = \int \gamma_0(G = 1, x, m) \mathrm{d}\mathbb{P}(m | G = 0, x, d) \mathrm{d}\mathbb{P}(x).$$

Note that we use both data sets

1. Experimental ($G = 0$)

$$\mathbb{P}(m | G = 0, x, d) = \mathbb{P}(M = m | G = 0, X = x, D' = d)$$

2. Observational ($G = 1$)

$$\gamma_0(G = 1, x, m) = \mathbb{E}[Y' | G = 1, X = x, M = m]$$

Moreover, we only use (G, X, D', M, Y')

Model: Identification

Let's interpret

$$\theta_0(d) = \int \gamma_0(G=1,x,m)\mathrm{d}\mathbb{P}(m|G=0,x,d)\mathrm{d}\mathbb{P}(x).$$

1. treatment effect: nonlinear functional (Robins 1986; Chernozhukov et al. 2016; Chen + Ritzwoller 2021; S. 2021)
2. dose response: sequential partial mean (Newey 1994; S. et al. 2020, 2021)
3. counterfactual distribution: sequential "partial distribution" (Chernozhukov et al. 2013; S. et al. 2021)

- today I will focus on the dose response
- the paper covers all three
- extends to subpopulations and alternative populations (Stock 1989)

Outline

1 Motivation

2 Model

3 Algorithm

4 Theory

Algorithm: RKHS

Recall $\gamma_0(G = 1, x, m) = \mathbb{E}[Y' | G = 1, X = x, M = m]$ is a regression

I assume γ_0 is an element of a function space called a reproducing kernel Hilbert space (RKHS)

Define RKHSs for selection G, covariate X, and surrogate M then assume γ_0 is an element of the tensor product space

- define feature maps $\phi(g)$, $\phi(x)$, $\phi(m)$ for RKHSs $\mathcal{H}_{\mathcal{G}}$, $\mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_{\mathcal{M}}$
- define the tensor-product feature map $\phi(g) \otimes \phi(x) \otimes \phi(m)$ for RKHS $\mathcal{H} := \mathcal{H}_{\mathcal{G}} \otimes \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{M}}$
- I assume $\gamma_0 \in \mathcal{H}$, so $\gamma_0 : \mathcal{G} \times \mathcal{X} \times \mathcal{M} \to \mathbb{R}$ as desired
- formally, the product of three kernels gives a new kernel

Algorithm: RKHS

With this RKHS construction, I obtain a representation of the causal parameter as an inner product in $\mathcal{H}$. Recall

$$\theta_0(d) = \int \gamma_0(G=1,x,m)\mathrm{d}\mathbb{P}(m|G=0,x,d)\mathrm{d}\mathbb{P}(x).$$

<u>Theorem</u>: If $\gamma_0 \in \mathcal{H}$ then $\theta_0(d) = \langle \gamma_0, \mu_3(d) \rangle_{\mathcal{H}}$ where

$$\mu_1(0,x,d) = \int \phi(m)\mathrm{d}\mathbb{P}(m|G=0,x,d)$$

$$\mu_2(d) = \int [\phi(x) \otimes \mu_1(0,x,d)]\mathrm{d}\mathbb{P}(x)$$

$$\mu_3(d) = \phi(1) \otimes \mu_2(d)$$

Interpretation

- γ_0 is a regression from the observational sample
- μ_1 embeds $\mathbb{P}(m|G=0,x,d)$ from the experimental sample
- μ_2 reweights μ_1 using both samples
- μ_3 is a sequential embedding (S. et al. 2021)

Algorithm: Closed form

Theorem: The algorithm has a closed form solution. (Kimeldorf + Wahba 1974; Schölkopf et al. 2001)

Algorithm:

1. Calculate the empirical kernel matrices

$$K_{GG}, \quad K_{XX}, \quad K_{D'D'}, \quad K_{MM} \in \mathbb{R}^{n\times n}.$$

2. Calculate the vectors $\chi_i(d) \in \mathbb{R}^n$ equal to

$$K_{Xx_i} \odot [K_{MM}(K_{GG} \odot K_{XX} \odot K_{D'D'} + n\lambda_{\text{EXP}} I)^{-1}(K_{G0} \odot K_{Xx_i} \odot K_{D'd})].$$

3. Finally set

$$\hat{\theta}(d) = \frac{1}{n}\sum_{i=1}^{n}(Y')^{\top}(K_{GG} \odot K_{XX} \odot K_{MM} + n\lambda_{\text{OBS}} I)^{-1}\{K_{G1} \odot \chi_i(d)\}.$$

Algorithm: Tuning

There are two sets of hyperparameters

Ridge regression penalties $(\lambda_{\text{EXP}}, \lambda_{\text{OBS}})$

- theoretical values that balance bias and variance
- practical tuning procedure based on LOOCV
    - closed form solution
    - only compute estimator once

Kernel hyperparameters

- well known heuristics
- e.g. for Gaussian kernel, use median interpoint distance

Outline

1 Motivation

2 Model

3 Algorithm

4 Theory

Theory: Assumptions

Recall that

$$\theta_0(d) = \int \gamma_0(G = 1, x, m) \mathrm{d}\mathbb{P}(m | G = 0, x, d) \mathrm{d}\mathbb{P}(x).$$

I place assumptions on

1. regularity of original spaces
2. regularity of RKHSs
3. smoothness of $\mathbb{P}(m | G = 0, x, d)$ from the experimental sample
4. smoothness of $\gamma_0(G = 1, x, m)$ from the observational sample

Theory: Assumptions

Original spaces

- treatment, covariate, and surrogate spaces are Polish
  - separable and completelely metrizable topological spaces
  - may be reals, text, images, etc.
- outcome is bounded

RKHSs

- kernels are continuous and bounded
- feature maps are measurable
- covariate and surrogate kernels are characteristic
  - ensures injectivity of mean embeddings
  - implies uniqueness of RKHS representation

These weak conditions hold for Gaussian, spline, and Sobolev kernels (Sriperumbudur et al. 2010)

Theory: Assumptions

Define the conditional expectation operator for $\mathbb{P}(m|G=0,x,d)$

$$E_0 : f(\cdot) \mapsto \mathbb{E}[f(M)|G=\cdot, X=\cdot, D=\cdot]$$

I assume it is smooth: $c_{\text{EXP}} \in (1,2]$

Likewise I assume the regression $\gamma_0(G=1,x,m)$ is smooth: $c_{\text{OBS}} \in (1,2]$

Interpretation of $(c_{\text{EXP}}, c_{\text{OBS}})$

- approximation error assumption to analyze bias
- target is well approximated by the top eigenfunctions of its kernel
  - target is in the interior of its RKHS
- *source* condition (Smale + Zhou 2007; Caponnetto + de Vito 2007; Carrasco et al. 2007; Fischer + Steinwart 2020)

Theory: Assumptions

Example: Sobolev space

Let $\mathbb{H}^s$ be the Sobolev space over $\mathbb{R}^p$

- s is number of square integrable derivatives
- p is dimension of input space

$\mathbb{H}^s$ is an RKHS iff $s > \frac{p}{2}$ (Berlinet + Thomas-Agnan 2011)

Suppose

- $\mathcal{H} = \mathbb{H}^s$ is the RKHS of estimation
- the truth is $f_0 \in \mathbb{H}^{s_0}$

Then $c = \frac{s_0}{s}$ (Fischer + Steinwart 2020)

My analysis requires $s_0 > s > \frac{p}{2}$

Theory: Uniform consistency

Theorem: If

1. the stated conditions hold
2. $\lambda_{\text{EXP}} = n^{-\frac{1}{c_{\text{EXP}}+1}}$ and $\lambda_{\text{OBS}} = n^{-\frac{1}{c_{\text{OBS}}+1}}$

Then

$$\|\hat{\theta} - \theta_0\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c_{\text{EXP}}-1}{c_{\text{EXP}}+1}} + n^{-\frac{1}{2}\frac{c_{\text{OBS}}-1}{c_{\text{OBS}}+1}}\right)$$

Interpretation

- n is sample size
- $c_{\text{EXP}} \in (1,2]$ parametrizes smoothness of $\mathbb{P}(m|G=0,x,d)$
- $c_{\text{OBS}} \in (1,2]$ parametrizes smoothness of $\gamma_0(G=1,x,m)$
- at best $n^{-\frac{1}{6}}$ by setting $(c_{\text{EXP}}, c_{\text{OBS}}) = 2$

Theory: Uniform consistency

Theorem: If

1. the stated conditions hold
2. $\lambda_{\text{EXP}} = n^{-\frac{1}{c_{\text{EXP}}+1}}$ and $\lambda_{\text{OBS}} = n^{-\frac{1}{c_{\text{OBS}}+1}}$

Then

$$\|\hat{\theta} - \theta_0\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c_{\text{EXP}}-1}{c_{\text{EXP}}+1}} + n^{-\frac{1}{2}\frac{c_{\text{OBS}}-1}{c_{\text{OBS}}+1}}\right)$$

Interpretation

- sup norm guarantee
  - uniform across every treatment level
  - encodes caution when informing labor market policy
- slow rate of $n^{-\frac{1}{6}}$
  - minimal assumptions
  - does not directly depend on data dimension
  - can be compensated by a fast rate of $n^{-\frac{1}{3}}$...
- exact finite sample rates in the paper

Conclusion

I propose a new family of estimators for long term causal inference

- easily implemented due to closed form
- strong finite sample guarantees
    1. treatment effects: $\sqrt{n}$-consistency, Gaussian approx., efficiency
    2. dose responses: uniform consistency
    3. counterfactual distributions: convergence in distribution
- bridge between long term cost-benefit analysis and kernel methods
    - which labor market policies pay for themselves?
- part of a broader agenda
    - causal inference poses integral equations
    - kernel methods solve integral equations

I would love to talk more!

- email: rahul.singh@mit.edu