Kernel Instrumental Variable Regression

Rahul Singh[1] Maneesh Sahani[2] Arthur Gretton[2]

[1]Department of Economics
Massachusetts Institute of Technology

[2]Gatsby Computational Neuroscience Unit
University College London

NeurIPS, 2019

The main idea

- IV regression: strategy to learn causal relationship from **confounded** observational data
- RKHS methods: ML with statistical guarantees
- we propose RKHS approach to IV regression
  1. easily implemented (3 lines of code)
  2. consistent and minimax optimal
  3. outperforms state-of-the-art alternatives
- bridge between econometrics and machine learning

Outline

Outline

1. Framework
   - IV
   - RKHS
2. Algorithm
   - Estimation
   - Convergence rates
3. Simulations
   - Sigmoid design
   - Demand design

Problem setting
Confounding

- we wish to learn h, the nonlinear structural relationship between input X and output Y
  - 'if we **intervened** on X, what would be the effect on Y?'
  - **counterfactual** prediction
- observations of (X, Y) confounded by unobserved e
- (single-stage) regression of Y on X is a badly biased estimator of h

Confounding

Sigmoid design

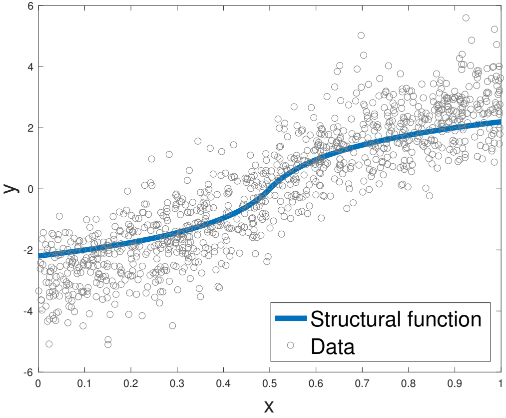

Figure: Sigmoid design from Chen and Christensen (2018). X is confounded. Note the additional correlation in observed (X, Y)

Confounding

Naive, single-stage approach

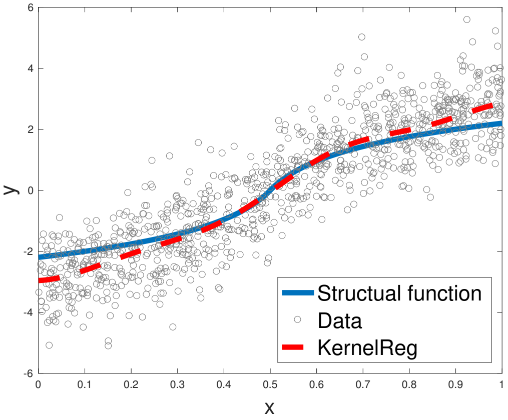

Figure: Kernel ridge regression on the sigmoid design

Problem setting
Instrumental variable

- we wish to learn h, the nonlinear structural relationship between input X and output Y
- observations of (X, Y) confounded by unobserved e
- instrument Z is independent of Y conditional on (X, e)
  - intuitively: Z only influences Y via X, identifying h

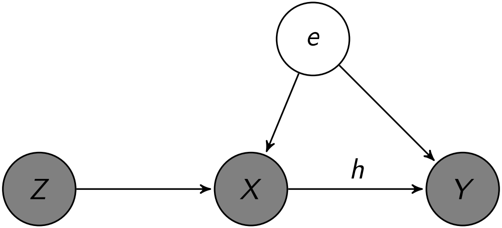

Figure: IV DAG

Examples

Economists and epidemiologists use instrumental variables to overcome issues of

- imperfect compliance
- selection bias
- strategic interaction

Example
Imperfect compliance

In RCTs, patients do not always do what they are assigned

- Y is patient health
- X is consumption of the drug
- e is (unobserved) patient income
- Z is assignment of the drug

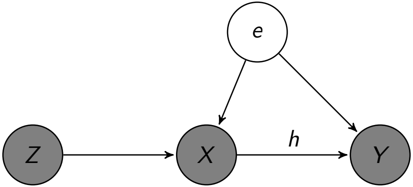

Figure: IV DAG

Example
Selection bias

Sometimes there is a natural experiment (Angrist 1990)

- Y is lifetime earnings
- X is military service
- e is (unobserved) 'ability'
- Z is draft lottery number

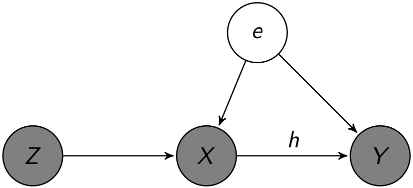

Figure: IV DAG

Example
Strategic interaction

The original application is demand estimation (Wright 1928)

- Y is sales
- X is price
- e is market forces
- Z is supply cost shifter

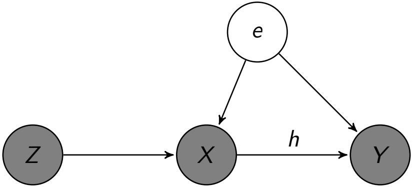

Figure: IV DAG

Outline

1. Framework
   - IV
   - RKHS
2. Algorithm
   - Estimation
   - Convergence rates
3. Simulations
   - Sigmoid design
   - Demand design

Formalizing IV

Newey and Powell (2003)

Assume

$$Y = h(X) + e, \quad \mathbb{E}[e|Z] = 0$$

Observations:

1. if $X = Z$, then reduces to unconfounded input case
2. helps to identify h when X is confounded
3. includes scenario of both confounded and unconfounded inputs
4. includes scenario of multiplicative error

Operator equation

Taking conditional expectations of both sides yields

$$\mathbb{E}[Y|Z] = \mathbb{E}_{X|Z} h(X)$$

Note that

- LHS is a function
- RHS is a composition
  1. *stage 1* conditional expectation operator $\mathbb{E}_{X|Z}$
  2. *stage 2* structural function h

Solving this operator equation is an ill-posed problem

To recover a well-posed problem, we impose smoothness and Tikhonov regularization

RKHS approach

Operator equation is

$$\mathbb{E}[Y|Z] = \mathbb{E}_{X|Z} h(X)$$

In our model

- LHS is a function. $\mathbb{E}[Y|Z] : \mathcal{Z} \to \mathbb{R}$
  - we assume it lives in RKHS $\mathcal{H}_{\mathcal{Z}}$
  - corresponding feature map $\phi : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}}$
- structural function $h : \mathcal{X} \to \mathbb{R}$
  - we assume it lives in RKHS $\mathcal{H}_{\mathcal{X}}$
  - corresponding feature map $\psi : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}$
- conditional expectation operator $E : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Z}}$
  - we assume it lives in tensor-product RKHS $\mathcal{H}_{\Gamma} \subset \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Z}}$

RKHS recap

- an RKHS is a smooth subset of L_2
  - it has a penalized inner product
  - for appropriate choice of feature map, it is dense in L_2 (universal)
- the prior $\mathcal{P}(b, c)$ is an assumption with two parameters
  1. $b \in (1, \infty]$. Bigger b means smaller effective input dimension
  2. $c \in (1, 2]$. Bigger c means smoother CEF
- typically used as hypothesis space for single-stage regression
  - pairs well with Tikhonov regularization
  - closed form solution due to representer theorem

Outline

1. Framework
   - IV
   - RKHS
2. Algorithm
   - Estimation
   - Convergence rates
3. Simulations
   - Sigmoid design
   - Demand design

The algorithm

Theory

$$\boxed{\mathbb{E}[Y|Z] = \mathbb{E}_{X|Z}h(X)}$$

Partition observations into 2 distinct subsets of sizes n and m

1. estimate conditional expectation operator E

$$E_\lambda^n = \operatorname*{argmin}_{E \in \mathcal{H}_\Gamma} \mathcal{E}_\lambda^n(E)$$

$$\mathcal{E}_\lambda^n(E) = \frac{1}{n}\sum_{i=1}^{n} \|\psi(x_i) - E^*\phi(z_i)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda\|E\|^2_{\mathcal{H}_\Gamma}$$

2. estimate structural function h using estimate of E

$$h_\xi^m = \operatorname*{argmin}_{h \in \mathcal{H}_\mathcal{X}} \hat{\mathcal{E}}_\xi^m(h)$$

$$\hat{\mathcal{E}}_\xi^m(h) = \frac{1}{m}\sum_{i=1}^{m} [y_i - [E_\lambda^n h](z_i)]^2 + \xi\|h\|^2_{\mathcal{H}_\mathcal{X}}$$

The algorithm
Practice

Repeated applications of the representer theorem yield a closed form solution.
Let X and Z be matrices of n observations. Let $\tilde{y}$ and $\tilde{Z}$ be a vector and matrix of m observations.

$$\begin{aligned} W &= K_{XX}(K_{ZZ} + n\lambda I)^{-1} K_{Z\tilde{Z}} \\ \hat{\alpha} &= (WW' + m\xi K_{XX})^{-1} W\tilde{y} \\ h_{\xi}^{m}(x) &= (\hat{\alpha})' K_{Xx} \end{aligned}$$

where K_{XX} and K_{ZZ} are the empirical kernel matrices i.e.

$$\begin{aligned} [K_{XX}]_{ij} &= \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{H}_{\mathcal{X}}} \\ [K_{ZZ}]_{ij} &= \langle \phi(z_i), \phi(z_j) \rangle_{\mathcal{H}_{\mathcal{Z}}} \end{aligned}$$

Outline

Stage 1 convergence rate

$$\boxed{\mathbb{E}[Y|Z] = \mathbb{E}_{X|Z} h(X)}$$

Under the regularity conditions

1. $\mathcal{X}$ and $\mathcal{Z}$ are Polish spaces
2. feature maps ϕ and ψ are continuous, bounded, and measurable
3. the problem is correctly-specified; $E_\rho \in \mathcal{H}_\Gamma$
4. a stage 1 prior $\mathcal{P}(c_1)$ is satisfied

Choose $\lambda = n^{-\frac{1}{c_1+1}}$. Then

$$\|E_\lambda^n - E_\rho\|_{\mathcal{H}_\Gamma} = O_p\left(n^{-\frac{1}{2}\cdot\frac{c_1-1}{c_1+1}}\right)$$

Stage 1 convergence rate

Discussion

The stage 1 convergence rate is

$$\|E_\lambda^n - E_\rho\|_{\mathcal{H}_\Gamma} = O_p\left(n^{-\frac{1}{2}\cdot\frac{c_1-1}{c_1+1}}\right)$$

- $c_1 \in (1, 2]$ measures the smoothness of E_ρ
- at best, $\|E_\lambda^n - E_\rho\|_{\mathcal{H}_\Gamma} = O_p\left(n^{-\frac{1}{6}}\right)$
- we actually have an exact, non-asymptotic bound
- it may not be the tightest possible bound
- the argument extends classic proofs by Smale and Zhou (2005, 2007)

Stage 2 convergence rate

$$\boxed{\mathbb{E}[Y|Z] = \mathbb{E}_{X|Z}h(X)}$$

Under the regularity conditions above as well as

1. $\mathcal{Y}$ is a Polish space and Y is bounded a.s.
2. $z \mapsto \mathbb{E}_{X|Z=z}\psi(X)$ is injective (characteristic property)
3. the stage 2 operator family is bounded in trace norm and Hölder continuous in operator norm
4. the problem is correctly-specified; $h_\rho \in \mathcal{H}_X$
5. a stage 2 prior $\mathcal{P}(b, c)$ is satisfied

Choose $\lambda = n^{-\frac{1}{c_1+1}}$, $\xi = m^{-\frac{b}{bc+1}}$, and $n = m^{\frac{b(c+1)}{bc+1}\cdot\frac{(c_1+1)}{\iota(c_1-1)}}$. Then

$$\mathcal{E}(h_\xi^m) - \mathcal{E}(h_\rho) = O_p(m^{-\frac{bc}{bc+1}})$$

Stage 2 convergence rate

Discussion

The stage 2 convergence rate is

$$\mathcal{E}(h_\xi^m) - \mathcal{E}(h_\rho) = O_p(m^{-\frac{bc}{bc+1}})$$

- $b \in (1, \infty]$ measures the effective input dimension
- $c \in (1, 2]$ measures the smoothness of h_ρ
- we actually have an exact, non-asymptotic bound
- the argument extends proofs by Szabó et al. (2016)

Stage 2 convergence rate

Discussion

The stage 2 convergence rate is

$$\mathcal{E}(h_\xi^m) - \mathcal{E}(h_\rho) = O_p(m^{-\frac{bc}{bc+1}})$$

- KIV achieves the minimax optimal rate of single-stage regression (Caponnetto and De Vito 2007)
- KIV learns the causal relationship with **confounded** data equally well as single-stage regression learns the causal relationship with **unconfounded** data

Stage 2 convergence rate

Discussion

The optimal ratio between stage 1 and stage 2 samples sizes is

$$n = m^{\frac{b(c+1)}{bc+1} \cdot \frac{(c_1+1)}{\iota(c_1-1)}}$$

- $\iota \in (0, 1]$ is the Hölder continuity parameter
- use more samples in stage 1 than stage 2
- as far as we know, this is a novel prescription for IV

Outline

Sigmoid design

Chen and Christensen (2018)

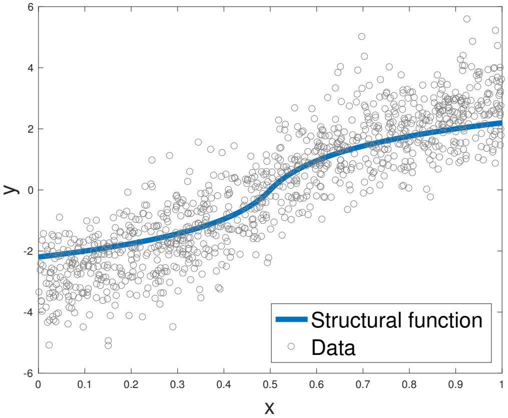

Figure: X is confounded. Additional correlation in observed (X, Y)

RKHS regression

Black box machine learning

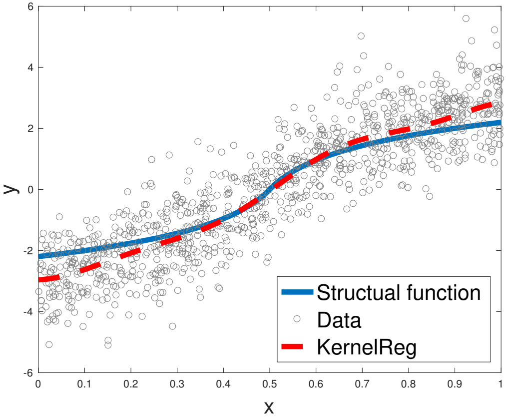

Figure: Single-stage regression with an infinite dictionary of basis functions, Tikhonov regularization

Sieve IV

Newey and Powell (2003)

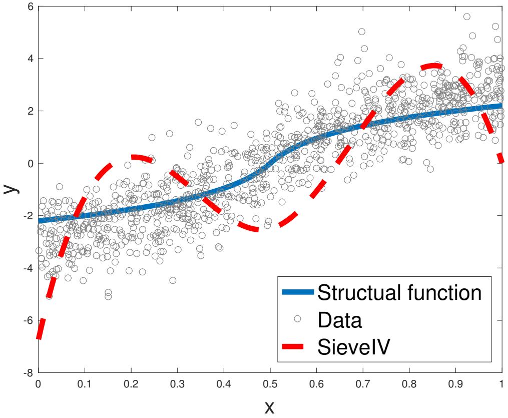

Figure: Two-stage regression with a finite dictionary of basis functions, spectral cut-off regularization

Smooth IV

Carrasco, Florens, and Renault (2007)

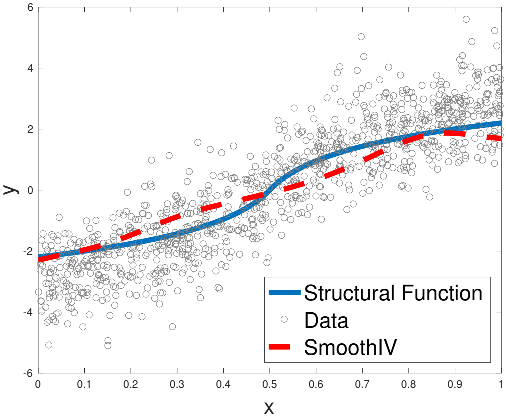

Figure: Nadaraya-Watson style kernel smoothing, Tikhonov regularization

Deep IV
Hartford et al. (2017)

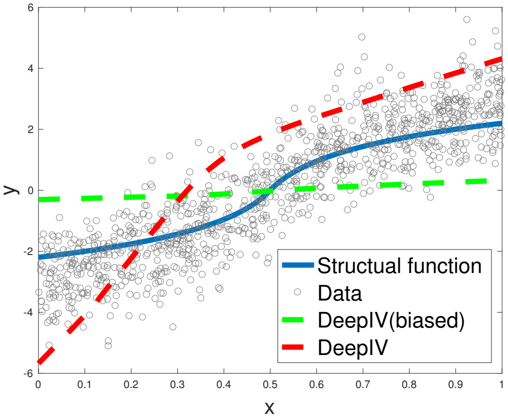

Figure: Neural nets

KIV

Singh, Sahani, and Gretton (2019)

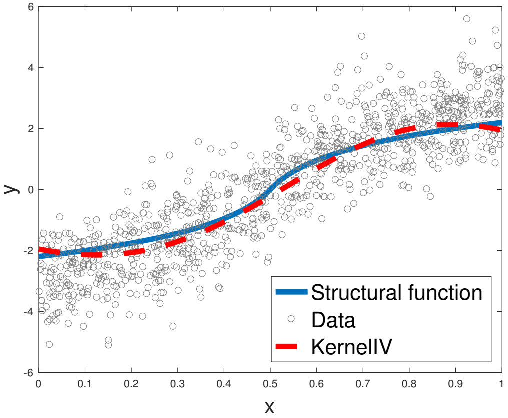

Figure: Two-stage regression with an infinite dictionary of basis functions, Tikhonov regularization

Sieve IV
Updated by Singh, Sahani, and Gretton (2019)

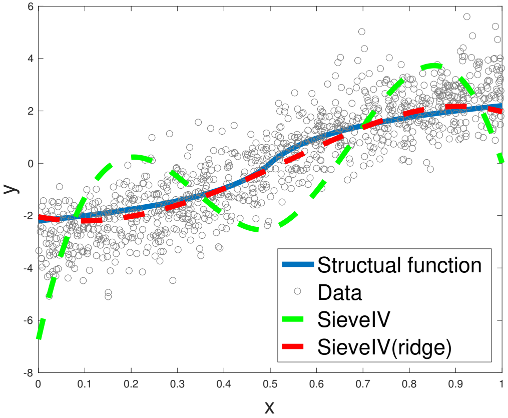

Figure: iseTwo-stage regression with an finite dictionary of basis functions, Tikhonov regularization

Comparison of methods

Sigmoid design

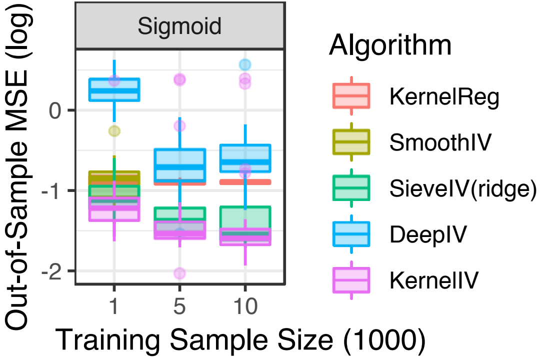

Figure: Comparison of methods varying training sample size

Outline

Comparison of methods
Demand design

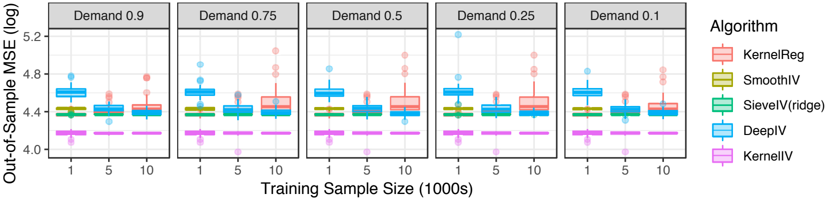

Figure: Comparison of methods varying training sample size

Summary

- IV regression: strategy to learn causal relationship from **confounded** observational data
- RKHS methods: ML with statistical guarantees
- we propose RKHS approach to IV regression
  1. easily implemented (3 lines of code)
  2. consistent and minimax optimal
  3. outperforms state-of-the-art alternatives
- bridge between econometrics and machine learning

Going ahead

- consider mis-specified case
- prove stage 2 convergence in RKHS norm (implies sup norm)
- inference: derive uniform confidence bands
- open to ideas!

For Further Reading I

A. Caponnetto and E. De Vito
Optimal rates for the regularized least-squares algorithm.
Foundations of Computational Mathematics, 7(3):331–368, 2007.

W. Newey and J. Powell
Instrumental variable estimation of nonparametric models.
Econometrica, 71(5):1565–1578, 2003.

S. Smale and D. Zhou
Shannon sampling II: Connections to learning theory.
Applied and Computational Harmonic Analysis, 19(3):285–302, 2005.

For Further Reading II

S. Smale and D. Zhou
Learning theory estimates via integral operators and their approximations.
Constructive Approximation, 26(2):153–172, 2007.

Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton
Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40, 2016.