MIT Economics

Kernel Instrumental Variable Regression

Rahul Singh, Maneesh Sahani, Arthur Gretton

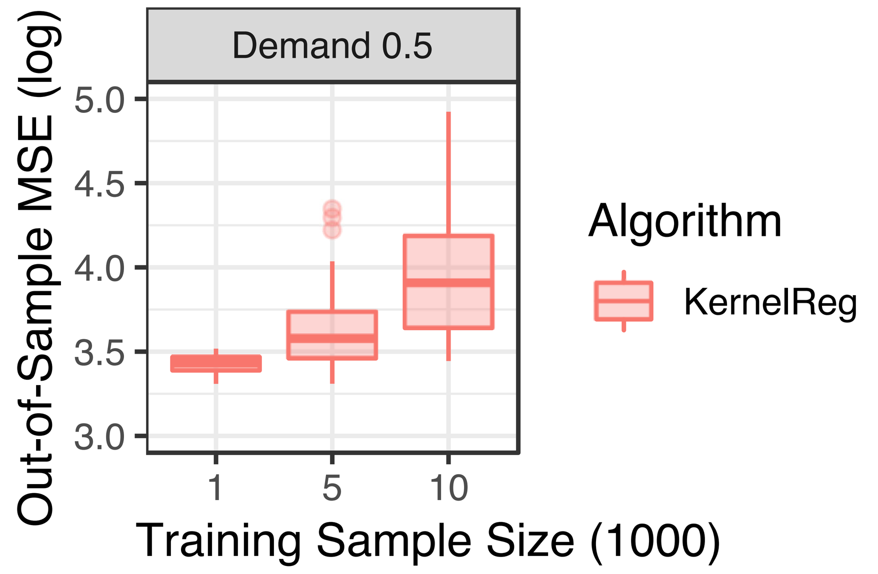
GATSBY

Abstract

- goal: learn **causal** relationship from **confounded** data
- we propose KIV
  1. computation: 3 lines of code
  2. statistical guarantee: minimax optimal
  3. performance: best when smooth design or $< 10,000$ observations
- bridge between econometrics and machine learning

1. Motivation: Demand

- predict airline ticket sales from airline ticket price, customer characteristics, time of year

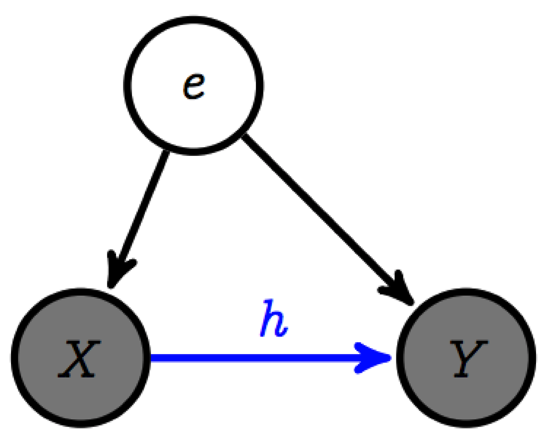


Kernel ridge regression on the demand design

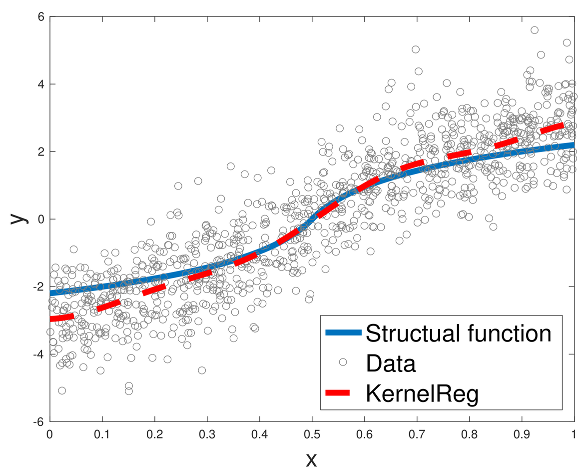
- learning gets worse as sample size increases
- what went wrong?

2. Confounding

- unobserved confounder e $\implies$ prediction $\neq$ counterfactual prediction
- goal: learn **causal** relationship h between input X and output Y
  - 'if we **intervened** on X, what would be the effect on Y?'
  - **counterfactual** prediction
- regression is a badly biased estimator of h



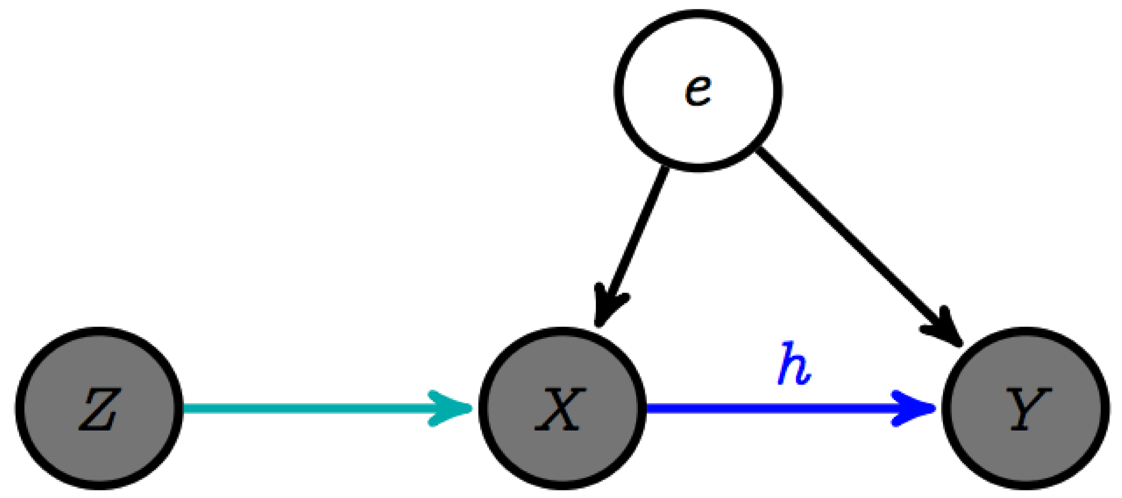
Confounded DAG



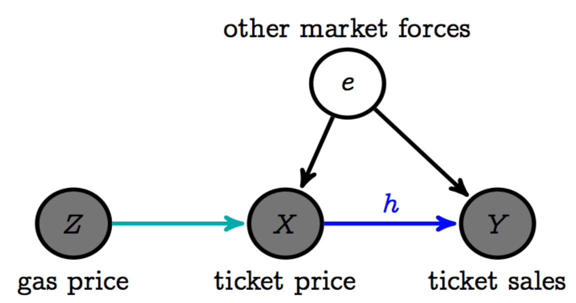
Sigmoid design

3. Instrumental variable

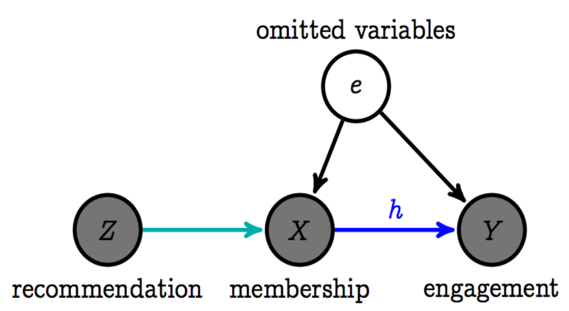
- instrument Z only influences Y via X, identifying h



IV DAG



Ex: Demand estimation



Ex: Imperfect compliance

4. Algorithm

KIV is a nonlinear generalization of 2SLS

1. kernel ridge regression of $\psi(X)$ on $\phi(Z)$
   - using n observations
   - construct $\mu(z) := \mathbb{E}[\psi(X)|Z=z]$
2. kernel ridge regression of Y on $\mu(Z)$
   - using remaining m observations
   - this is the estimator for h

note that

- allows nonlinearity among (X, Y, Z)
- closed form solution $\implies$ 3 lines of code

5. Theory

Sample splitting

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

- $b \in (1, \infty]$ effective input dimension
- $c \in (1, 2]$ smoothness of h
- $c_1 \in (1, 2]$ smoothness of μ
- asymmetric sample splitting is novel

Convergence rate

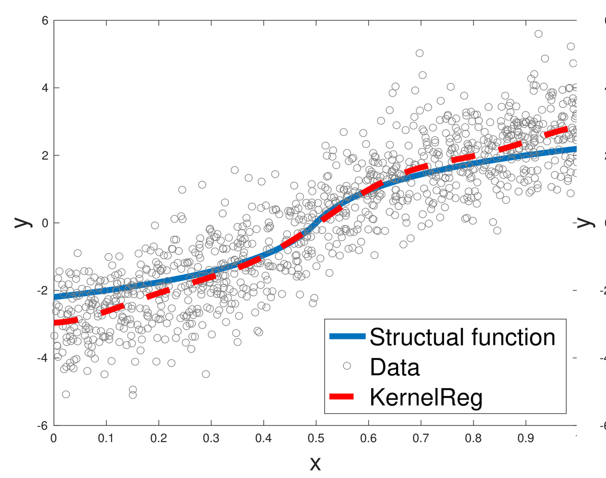
$$\mathcal{E}(\hat{h}) - \mathcal{E}(h) = O_p(m^{-\frac{bc}{bc+1}})$$

- $b \in (1, \infty]$ effective input dimension
- $c \in (1, 2]$ smoothness of h
- learning with **confounded** data at the rate of learning with **unconfounded** data

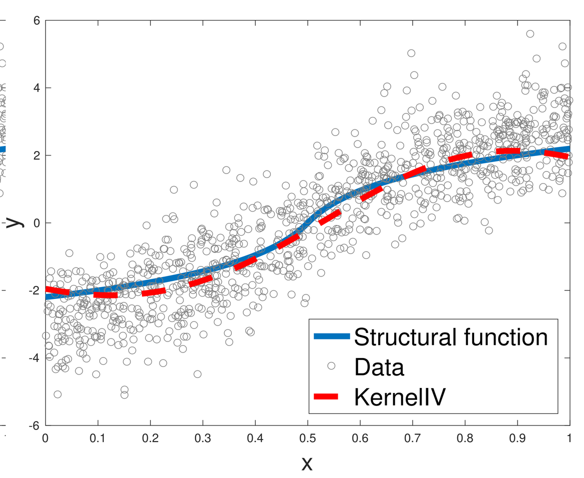
6. Experiments

One simulation

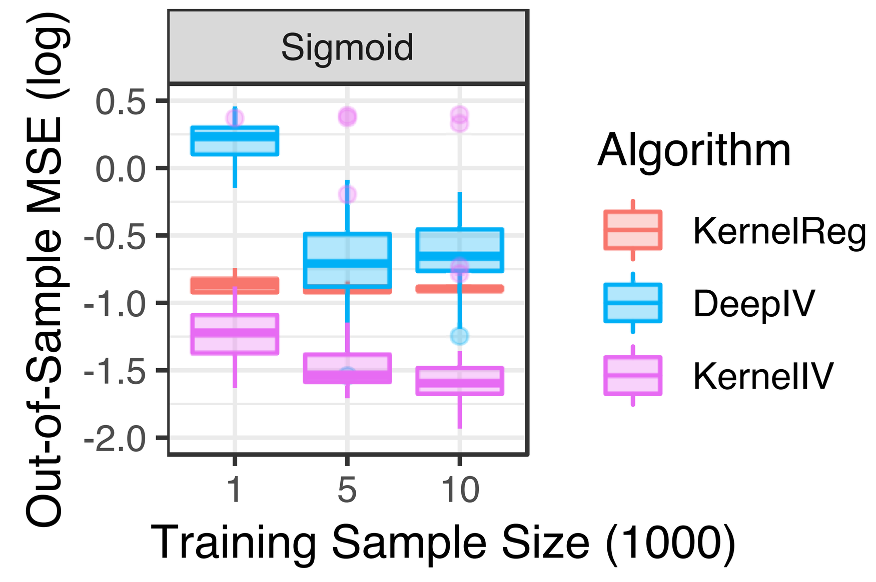
- KIV is learns h despite unmeasured confounding



Kernel ridge regression



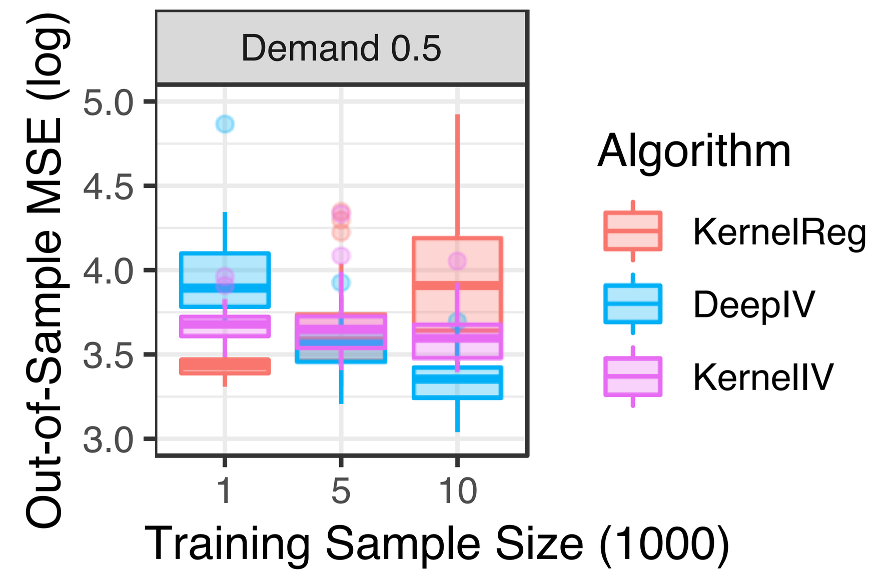
KIV



Comparison of methods on the sigmoid design

Many simulations

- in smooth designs, KIV performs best
- in highly nonlinear designs, KIV performs best when $< 10,000$ observations



Comparison of methods on the demand design

References

[1] W.K. Newey and J.L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[2] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. *International Conference on Machine Learning*, 1414–1423, 2017.

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

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

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