MIT Economics

Kernel Instrumental Variable Regression

Rahul Singh, Maneesh Sahani, Arthur Gretton

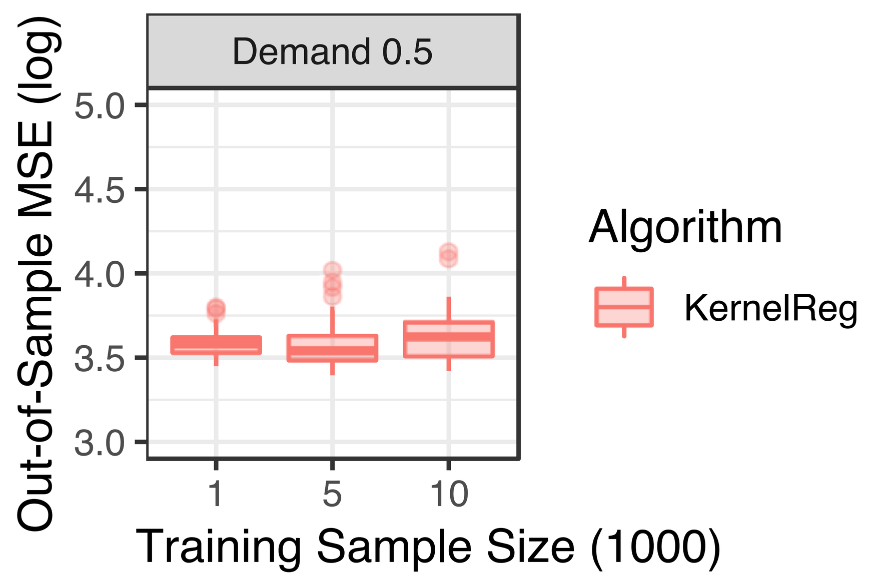
GATSBY

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 with smooth design or $< 10,000$ observations
- bridge between econometrics and machine learning

1. Motivation: Demand

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

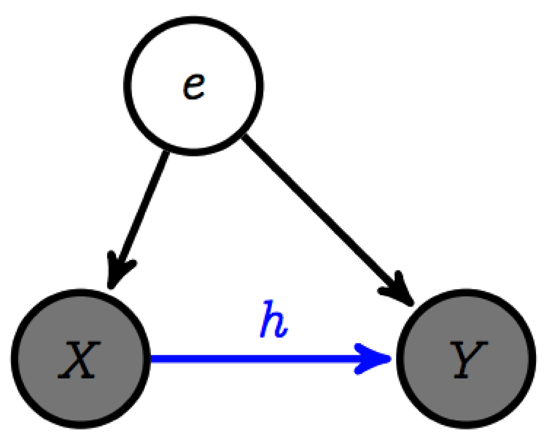


Kernel ridge regression on demand design

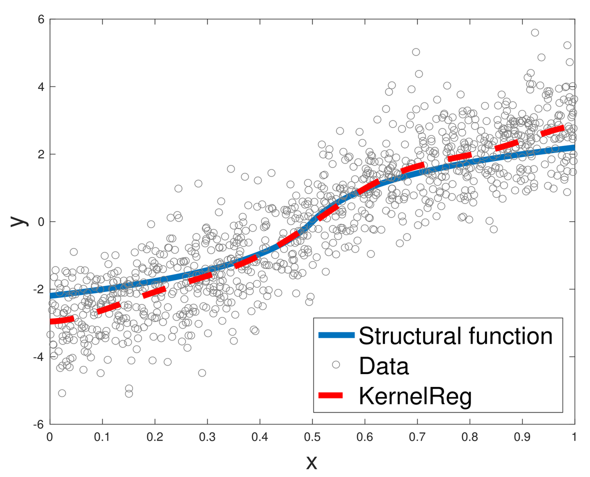
- learning gets worse as sample size increases
- what went wrong?

2. Confounding

- 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
- unobserved confounder $e \implies$ prediction $\neq$ counterfactual prediction
- regression is a badly biased estimator of h



Confounded DAG

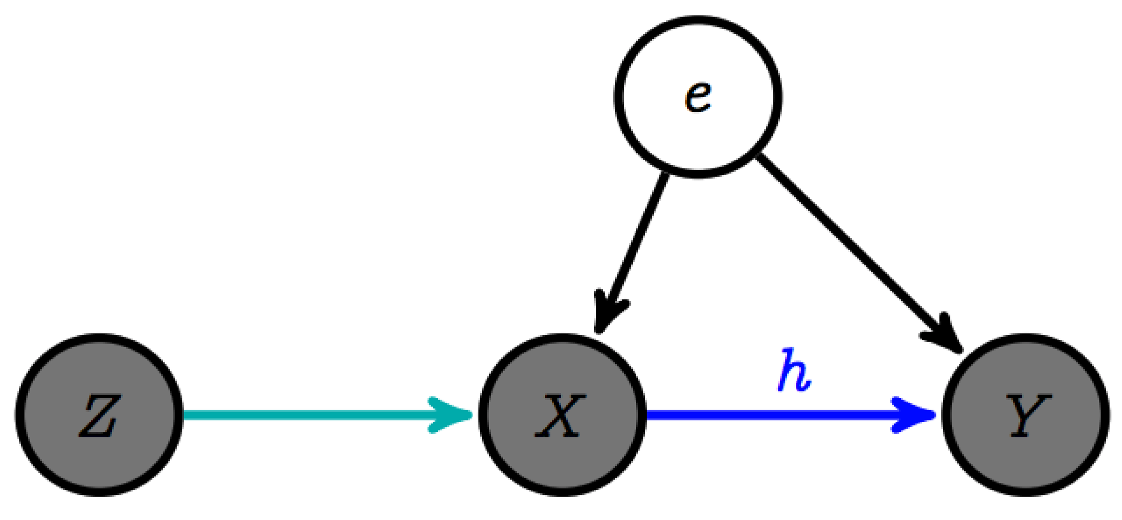


Sigmoid design

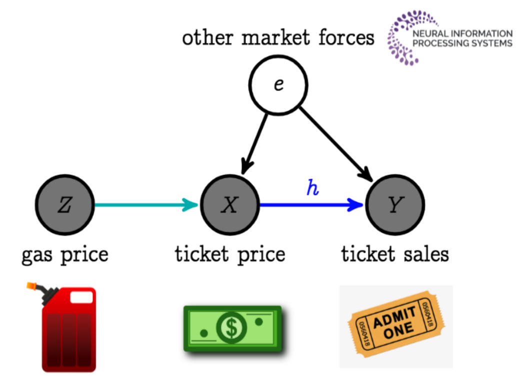
3. Instrumental variable

- instrument Z only influences Y via X

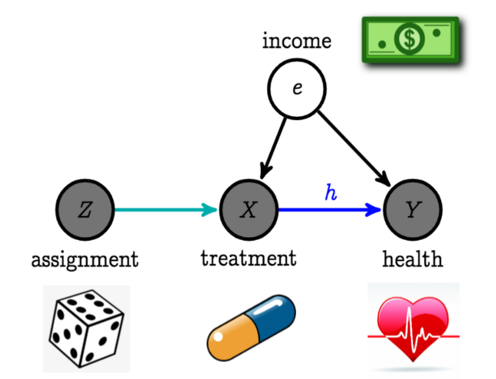
$$Y = h(X) + e, \quad \mathbb{E}[e|Z] = 0$$



IV DAG



e.g. Demand estimation



e.g. 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

closed form solution $\implies$ 3 lines of code

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

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 of $\psi(X)$
- $c \in (1, 2]$ smoothness of h
- $c_1 \in (1, 2]$ smoothness of μ
- asymmetric sample splitting is novel

Convergence rate

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

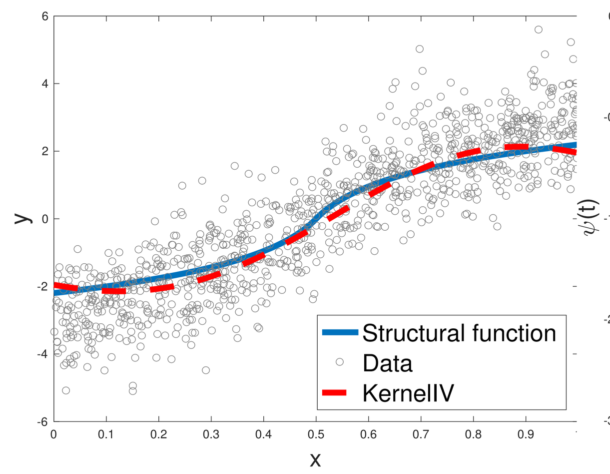
- $b \in (1, \infty]$ effective input dimension of $\psi(X)$
- $c \in (1, 2]$ smoothness of h
- learning with **confounded** data at the rate of learning with **unconfounded** data

6. Experiments

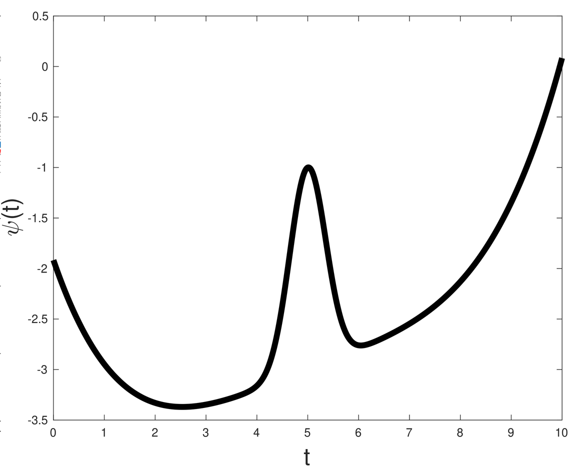
Sigmoid design

$$h(x) = \ln(|16x - 8| + 1) \cdot sgn(x - 0.5)$$

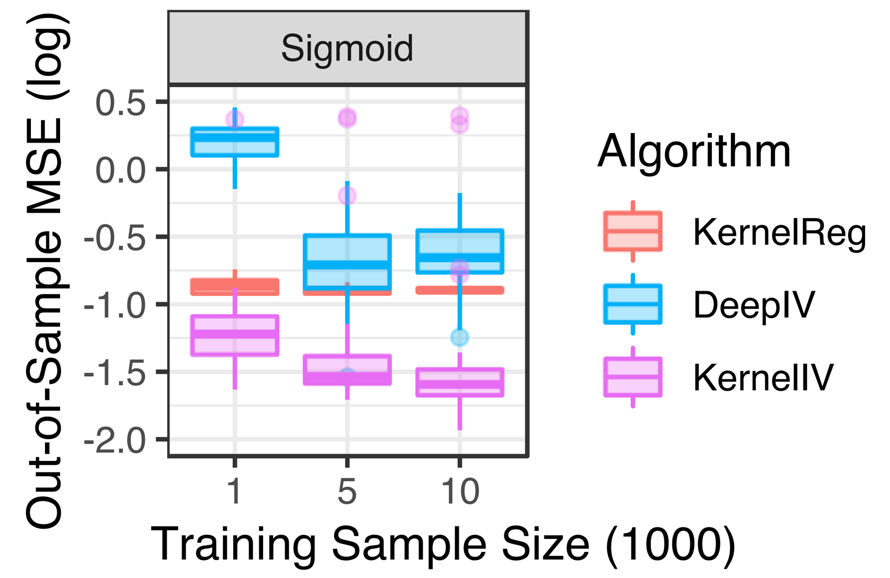
- KIV learns h despite unmeasured confounding
- KIV performs particularly well in smooth designs



KIV on sigmoid design



Demand nonlinearity $\psi(t)$

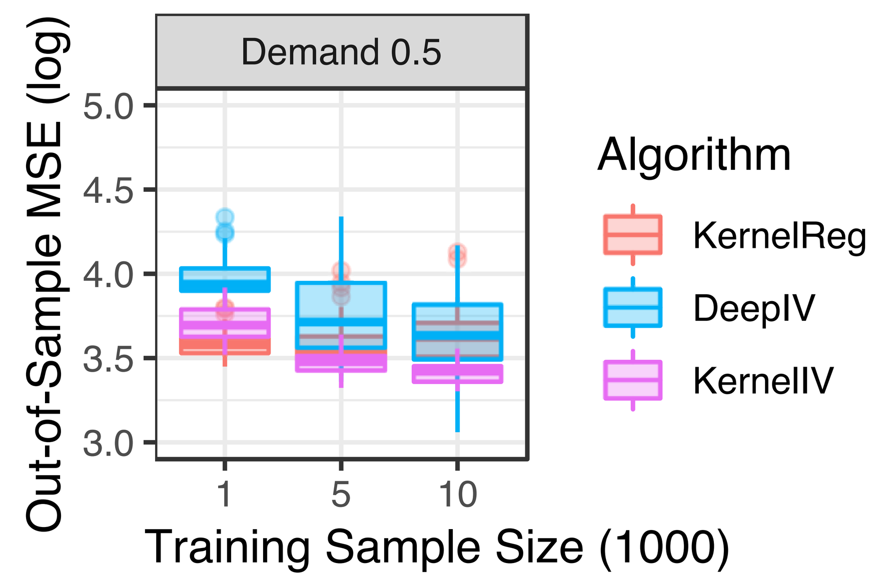


Comparison of methods on sigmoid design

Demand design

$$h(p, t, s) = 100 + (10 + p)s\psi(t) - 2p$$

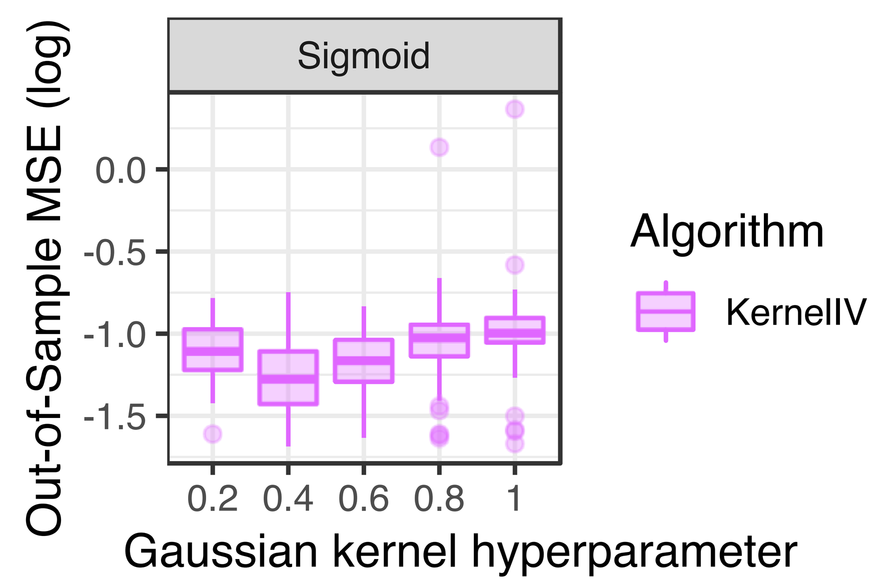
- ticket sales Y, ticket price P, time of year T, customer characteristics S, gas price C
- $X = (P, T, S)$ and $Z = (C, T, S)$
- KIV performs best



Comparison of methods on demand design

Tuning

- regularization (λ, ξ) by validation
- Gaussian kernel lengthscales by median inter-point distance



Robustness study

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. *ICML*, 1414–1423, 2017.
[3] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.
[4] Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *JMLR*, 17(152):1–40, 2016.