Kernel Instrumental Variable Regression

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

[1]MIT Economics,
[2]Gatsby Unit, UCL

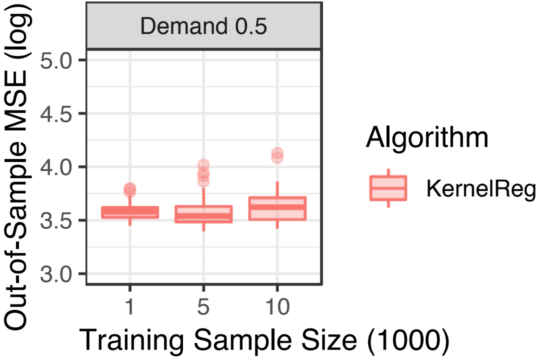
NeurIPS 2019

Motivation: demand estimation

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

Motivation: demand estimation

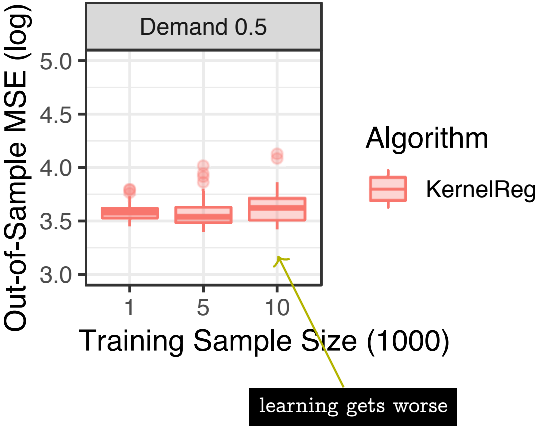
- predict ticket sales from price, customer characteristics, time of year



Kernel ridge regression on the demand design (Hartford et al. 2017)

Motivation: demand estimation

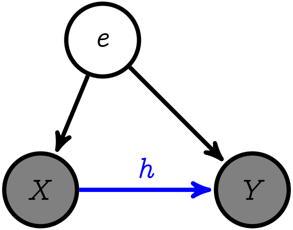
- predict ticket sales from price, customer characteristics, time of year



- what went wrong?

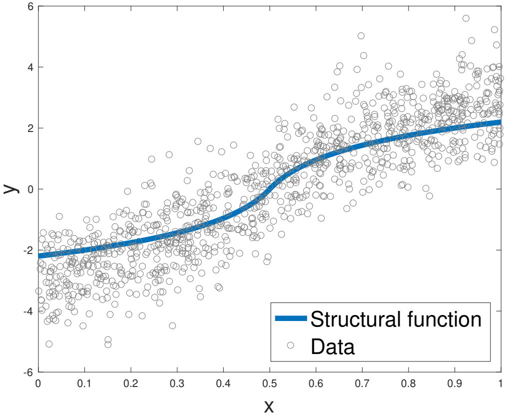
Confounding

- 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
  - $\mathbb{E}[Y|X] \neq h(X)$
  - regression is a badly biased estimator of h



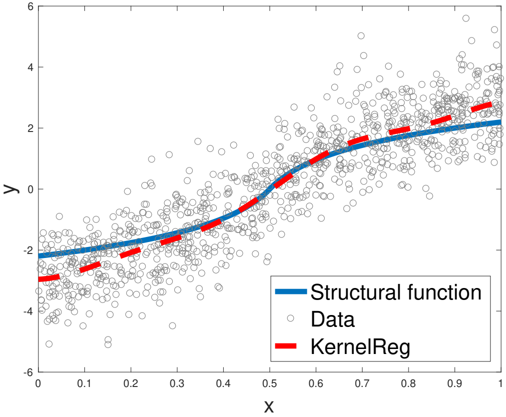
Confounded DAG

Confounding



Sigmoid design (Chen and Christensen 2018)

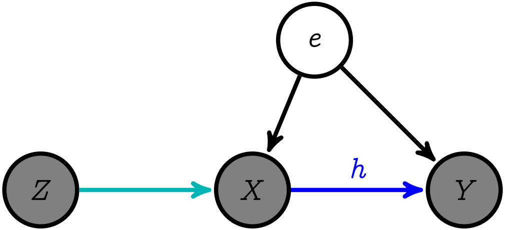
Confounding



Kernel ridge regression on the sigmoid design

Instrumental variable

- unobserved confounder $e \implies$ prediction $\neq$ counterfactual prediction
- goal: learn **causal** relationship h between input X and output Y
- instrument Z only influences Y via X, identifying h

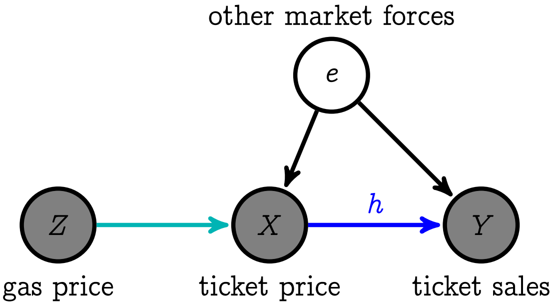


IV DAG

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

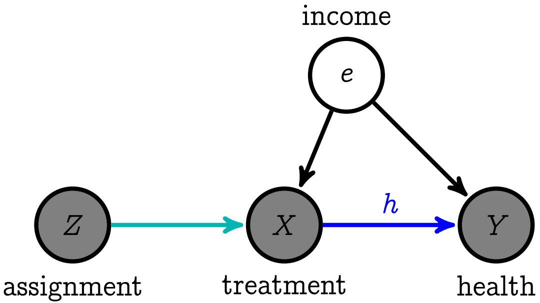
Example: Demand estimation

- goal: **causal** relationship between price and sales, e.g. airline tickets
- the original application (Wright 1928)



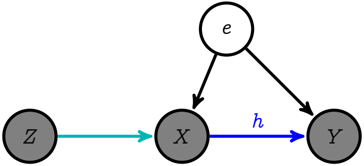
Example: Imperfect compliance

- goal: learn **causal** relationship between treatment and health
- relevant for digital platforms (Syrgkanis et al. 2019)



Algorithm: 2SLS

1. linear regression of X on Z
   - using n observations
   - construct $\bar{X}(z) := \mathbb{E}[X|Z = z]$, the conditional mean
2. linear regression of Y on $\bar{X}(Z)$
   - using remaining m observations
   - this is the estimator for h

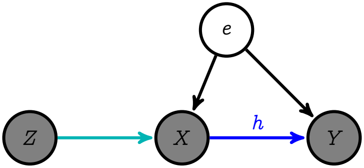


IV DAG

- imposes linearity among (X, Y, Z), assumes $\mathbb{E}[e \cdot Z] = 0$
- widely used in economics

Algorithm: KIV

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

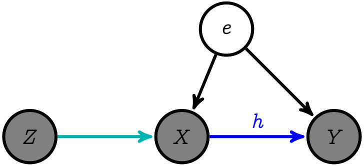


IV DAG

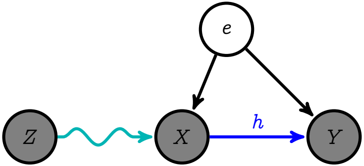
- allows nonlinearity among (X, Y, Z), assumes $\mathbb{E}[e|Z] = 0$
- closed form solution $\implies$ 3 lines of code

Theory: Sample splitting

- calibrate to smoothness of μ and h
- e.g. $n = m^{\alpha}$ where $\alpha > 1$ if



- e.g. $n = m^{\beta}$ where $\beta > \alpha > 1$ if



- exact formula in paper
- asymmetric sample splitting is novel

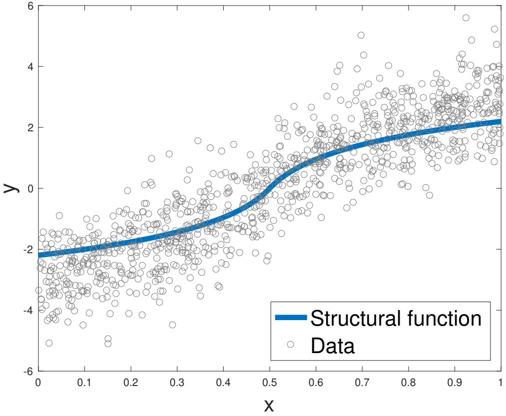
Theory: Convergence rate

using the sample splitting formula for (n, m),

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

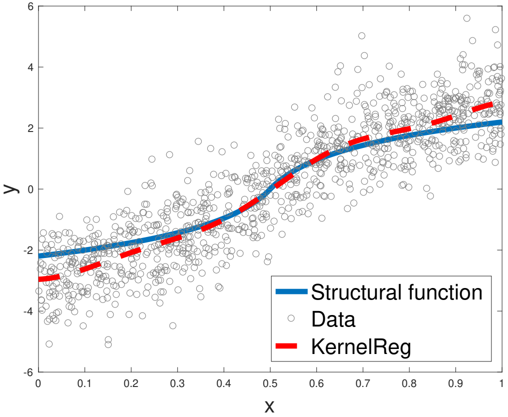
- $b \in (1, \infty]$ effective input dimension of $\psi(X)$
- $c \in (1, 2]$ smoothness of h
- learning with **confounded** data at the rate of learning with **unconfounded** data

Sigmoid design



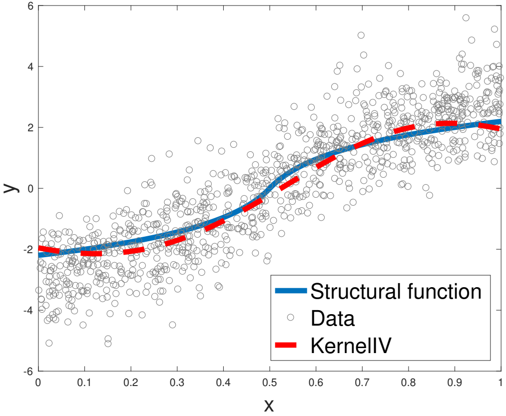
Sigmoid design (Chen and Christensen 2018)

Sigmoid design



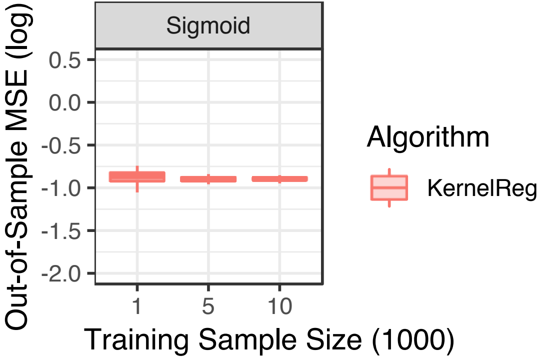
Kernel ridge regression on the sigmoid design

Sigmoid design



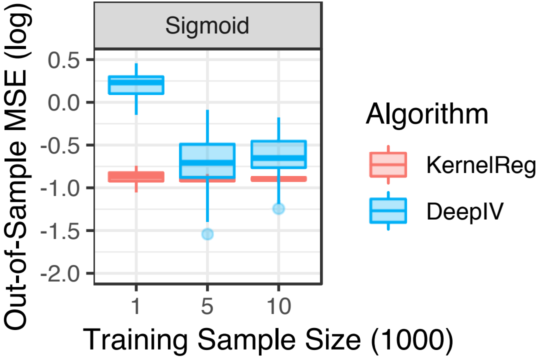
KIV on the sigmoid design

Sigmoid design



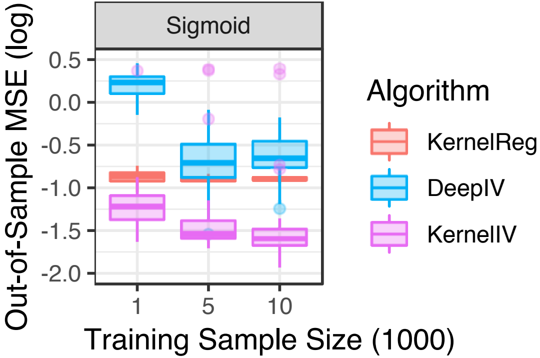
Comparison of methods varying training sample size

Sigmoid design



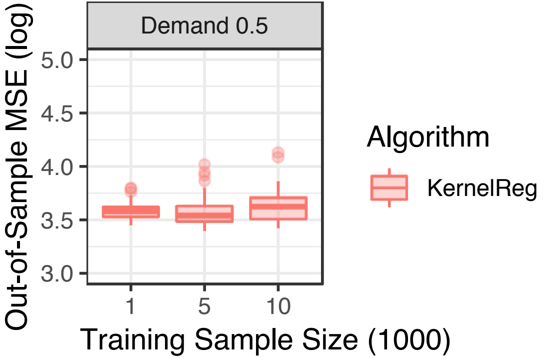
Comparison of methods varying training sample size

Sigmoid design



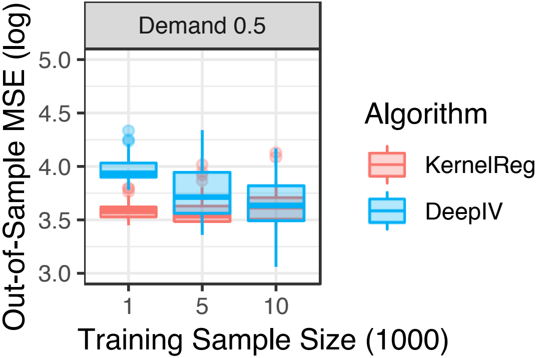
Comparison of methods varying training sample size

Demand design



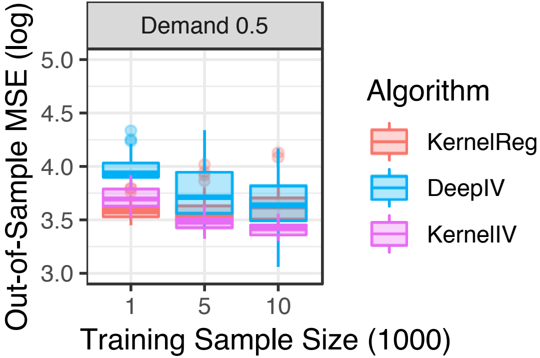
Comparison of methods varying training sample size

Demand design



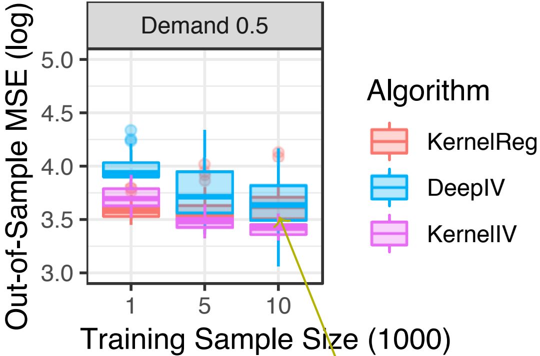
Comparison of methods varying training sample size

Demand design



Comparison of methods varying training sample size

Demand design



no statistical guarantees as yet

Conclusion

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

please visit us!

- poster #59, east exhibition hall B+C
- MATLAB code available for download