Kernel Instrumental Variable Regression

Kei Ishikawa

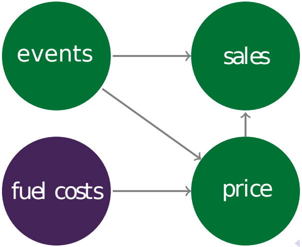
In Journal Club, May 15, 2020

Outline

Outline

Correlations and Causal Relations

- Suppose we want to estimate how much the sales of flight tickets will increase if we decrease their price.
- When we observe the price and sales (the number of tickets sold) of flight tickets, there should be a positive correlation between them.
  - The price and sales of the flight tickets are both affected by some event such as holidays, conference, outbreak of virus.
  - However, it is impossible to keep track of all kinds of events that can potentially impact sales.
  - If we use this positive correlation for predicting the sales of tickets, we can conclude that increase in the price would increase the sales (the number of tickets sold), which is very unrealistic.



Correlations and Causal Relations

- The variable "events" cannot be observed and is the cause positive of positive correlations. Such variables are called confounders.
- The variable "fuel costs" is independent of the confounder "events" (such as holidays). Such variables are called "Instrumental Variables" and can be used to estimate causal effects of price on sales.

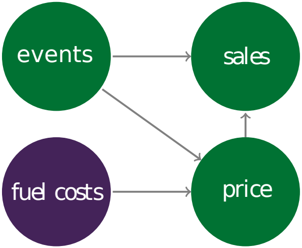


Figure: Demand estimation of flight tickets

Outline

Problem Setting (Flight Ticket Example)

- Model
  - $(sales) = \beta_0 + \beta_1 (price) + e$
  - $(price) = \delta_0 + \delta_1 (fuel\ costs) + \varepsilon$
  - $\mathbb{E}[e] = 0$ and $e \perp\!\!\!\perp (events)$, but $e \not\perp\!\!\!\perp (price)$ [1]
  - $\mathbb{E}[\varepsilon | (fuel\ price)] = 0$
- Problem
  - Estimate β_0 and β_1
- Difficulty
  - $e \not\perp\!\!\!\perp (price)$ makes OLS estimator biased

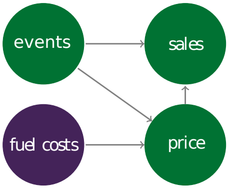
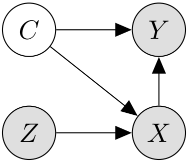


Figure: Demand estimation of flight tickets

[1] Weaker assumptions are possible.

Problem Setting

- Model
  - $Y = \beta' X + e$ and
    $X = \delta' Z + \varepsilon$,
    where $Y, e \in \mathbb{R}$, $X, \beta, \varepsilon \in \mathbb{R}^{d_x}$, $Z \in \mathbb{R}^{d_z}$, $\delta \in \mathbb{R}^{d_z \times d_x}$, and $d_z \geq d_x$.
  - $\mathbb{E}[e] = 0$ and $e \perp\!\!\!\perp Z$ but $\mathbb{E}[Xe] \neq 0$
  - $\mathbb{E}[Z\varepsilon] = 0$
- Problem
  - Estimate β.
- Difficulty
  - $\mathbb{E}[Xe] \neq 0$ makes OLS estimator of β biased



Ordinary Least Squares (OLS)

- OLS Estimator
  $\hat{\beta}^{\mathsf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \arg\min_{\beta \in \mathbb{R}^{d_x}} ||\mathbf{y} - \mathbf{X}\beta||_2^2$
  where $\mathbf{y} = (y_1, y_2, ..., y_n)'$ and $\begin{pmatrix} | & | & & | \\ x_1 & x_2 & ... & x_n \\ | & | & & | \end{pmatrix}'$
- Inconsistency of OLS Estimator

$$\begin{aligned} \hat{\beta}^{\mathsf{OLS}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e}) \\ &\xrightarrow{p.} \beta + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'e] \\ &\neq \beta \end{aligned}$$

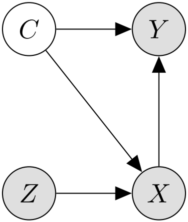
Two-Stage Least Squares (2SLS)

- Consider the model w.r.t. Z
  - $Y = \beta'(\delta' Z + \varepsilon) + e = \beta'(\delta' Z) + (\beta'\varepsilon + e)$
  - Since $\mathbb{E}[\beta'\varepsilon + e|Z] = 0$, OLS of $Y \sim \beta(\delta' Z)$ gives consistent estimate of β
  - $(\delta' Z)$ can be estimated consistently as $(\hat{\delta}^{\mathsf{OLS}\prime} Z)$ using OLS of $X \sim \delta' Z$
- 2SLS
  - Stage 1: $\hat{\delta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \arg\min_{\delta \in \mathbb{R}^{n \times d_z}} ||\mathbf{X} - \mathbf{Z}\delta||^2_{\mathsf{Fr}}$
  - Stage 2: $\hat{\beta}^{\mathsf{IV}} = (\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}'\mathbf{y} = \arg\min_{\beta \in \mathbb{R}^{d_x}} ||\mathbf{y} - \bar{\mathbf{X}}\beta||^2_2$ where $\bar{\mathbf{X}} = \hat{\delta}'\mathbf{Z}$

Outline

General Problem Setting

- Model
  - $Y = g(X) + e$
  - $\mathbb{E}[e|Z] = 0$, but $X \not\perp e$
- Problem
  - Estimate g.
- Solution
  - $Y = \mathbb{E}[g(X)|Z] + \underbrace{(\mathbb{E}[e|Z] + Y - \mathbb{E}[Y|Z])}_{\text{noise term having zero mean conditionally on } Z}$
  - Apply two-stage least squares
  - Stage 1: Estimate $\mathbb{P}[X|Z = \cdot\,]$ as $\hat{\mathbb{P}}[X|Z = \cdot\,]$
  - Stage 2: Estimate g by applying least squares to $Y \sim \hat{\mathbb{E}}[g(X)|Z]$ where $\hat{\mathbb{E}}[g(X)|Z]$ is calculated using $\hat{\mathbb{P}}[X|Z = \cdot\,]$



Kernel Methods and Notations

- Instead of using $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$,
  use feature $\psi(X) \in \mathcal{H}_{\mathcal{X}}$ and $\phi(Z) \in \mathcal{H}_{\mathcal{Z}}$ which satisfy
- Notations
  - $k_{\mathcal{X}}, k_{\mathcal{Z}}$: kernels over $\mathcal{X}$ and $\mathcal{Z}$
  - $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Z}}$: RKHS with kernel $k_{\mathcal{X}}, k_{\mathcal{Z}}$
  - $\psi(x) := k_{\mathcal{X}}(x, \cdot\,) \in \mathcal{H}_{\mathcal{X}}$: feature map
  - $\phi(z) := k_{\mathcal{Z}}(z, \cdot\,) \in \mathcal{H}_{\mathcal{Z}}$: feature map

Problem Setting

- Model
  - $Y = h(X) + e = H\psi(X) + e$
    where $Y, e \in \mathbb{R} = \mathcal{Y}$, $X \in \mathcal{X}$, $h \in \mathcal{H}_{\mathcal{X}}$ and $H : \mathcal{H}_{\mathcal{X}} \to \mathbb{R}$ is a linear operator
  - $\mathbb{E}[e|Z] = 0$ but $\mathbb{E}[\psi(X)e] \neq 0$
  - $\mathbb{E}[h(X)|Z = z]$ can be written as $\mathbb{E}[h(X)|Z = z] = [Eh](z)$ using a linear operator $E : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Z}}$.
- Problem
  - Estimate $h \in \mathcal{H}_{\mathcal{X}}$

Problem Setting

- Conditional Expectation Operator
  - Linear operator $E : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}$ satisfies $[Eh](z) = \mathbb{E}[h(X)|Z = z]$
  - Indeed, the adjoint of E satisfies
    $E^*\phi(z) = E^* k_\mathcal{Z}(z, \ \cdot\ ) = \mathbb{E}[\psi(X)|Z = z]$
    - ★ $\forall k_\mathcal{X}(x, \ \cdot\ ) \in \mathcal{H}_\mathcal{X}, k_\mathcal{Z}(z, \ \cdot\ ) \in \mathcal{H}_\mathcal{Z}$,
      $\langle E k_\mathcal{X}(x, \ \cdot\ ), k_\mathcal{Z}(z, \ \cdot\ )\rangle_{\mathcal{H}_\mathcal{Z}}$
      $= \langle \mathbb{E}[k_\mathcal{X}(x, X)|Z = \ \cdot\ ], k_\mathcal{Z}(z, \ \cdot\ )\rangle_{\mathcal{H}_\mathcal{Z}}$
      $= \mathbb{E}[k_\mathcal{X}(x, X)|Z = z]$
      $= \langle k_\mathcal{X}(x, \ \cdot\ ), \mathbb{E}[k_\mathcal{X}(\ \cdot\ , X)|Z = z]\rangle_{\mathcal{H}_\mathcal{X}}$
      $= \langle k_\mathcal{X}(x, \ \cdot\ ), E^* k(z, \ \cdot\ )\rangle_{\mathcal{H}_\mathcal{X}}$
- Model (re-formulated)
  - $Y = H\psi(X) + e$
    $\psi(X) = E^*\phi(Z) + \varepsilon$
    where $Y, e \in \mathbb{R} = \mathcal{Y}$, $X, \varepsilon \in \mathcal{X}$, $h \in \mathcal{H}_\mathcal{X}$,
    $H : \mathcal{H}_\mathcal{X} \to \mathbb{R}$ and $E^* : \mathcal{H}_\mathcal{Z} \to \mathcal{H}_\mathcal{X}$ are linear operators
  - $\mathbb{E}[e|Z] = 0$ but $\mathbb{E}[\psi(X)e] \neq 0$
  - $\mathbb{E}[\phi(Z) \otimes \varepsilon] = 0$.

Two-Stage Kernel Ridge Regression

- Consider conditional expectation w.r.t. Z
  - $\mathbb{E}[Y|Z]$
    $= \mathbb{E}[h(X)|Z] + \mathbb{E}[e|Z]$
    $= H\mathbb{E}[\psi(X)|Z] + \mathbb{E}[e|Z]$
    $= HE^*\phi(Z) + \mathbb{E}[e|Z]$
  - In other words,
    $Y = HE^*\phi(Z) + \underbrace{(\mathbb{E}[e|Z] + Y - \mathbb{E}[Y|Z])}_{\text{noise term having zero mean conditionally on } Z}$
  - Thus, if we have E^*, we can use kernel ridge regression of $Y \sim H(E^*\phi(Z))$ to estimate H
- Two-Stage Kernel Ridge Regression
  - Stage 1:
    $\hat{E}^n_\lambda = \arg \min_{E:\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}:\text{linear}} \mathcal{E}^n_\lambda(E)$,
    where $\mathcal{E}^n_\lambda(E) = \frac{1}{n}\sum_{i=1}^n ||\psi(x_i) - E^*\phi(z_i)||^2 + \lambda||E||^2$
  - Stage 2:
    $\hat{H}^n_\xi = \arg \min_{H:\mathcal{H}_\mathcal{X} \to \mathbb{R}:\text{linear}} \mathcal{E}^m_\xi(H)$,
    where $\mathcal{E}^m_\xi(H) = \frac{1}{m}\sum_{i=1}^m ||\tilde{y}_i - H\mu(\tilde{z}_i)||^2 + \xi||H||^2$

Two-Stage Kernel Ridge Regression

- Algorithm[1]
  Let X and Z be matrices of n observations. Let $\tilde{y}$ and $\tilde{Z}$ be a vector and matrix of m observations.

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

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

Two-Stage Kernel Ridge Regression

- Stage 1
  - Solution

$$\hat{E}^n_\lambda = \arg \min_{E:\mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}:\text{linear}} \frac{1}{n} \sum_{i=1}^{n} ||\psi(x_i) - E^* \phi(z_i)||^2 + \lambda ||E||^2$$
$$= \sum_{k,l} [(K_{ZZ} + n\lambda I)^{-1}]_{k,l} \langle \psi(x_k), \cdot \rangle_{\mathcal{H}_\mathcal{X}} \phi(z_l)$$

  - Derivation
    - ★ By the representer theorem, we can write $E : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}$ and its adjoint operator E^* as
      $E = \sum_{k,l} A_{k,l} \langle \psi(x_k), \cdot \rangle_{\mathcal{H}_\mathcal{X}} \phi(z_l)$
      $E^* = \sum_{k,l} \bar{A}_{k,l} \langle \phi(z_l), \cdot \rangle_{\mathcal{H}_\mathcal{Z}} \psi(x_k)$
    - ★ Thus,
      $\frac{1}{n} \sum_{i=1}^{n} ||\psi(x_i) - E^* \phi(z_i)||^2$
      $= \frac{1}{n} \sum_{i=1}^{n} ||\psi(x_i) - \sum_{k,l} A_{k,l} \langle \phi(z_l), \phi(z_i) \rangle_{\mathcal{H}_\mathcal{Z}} \psi(x_k)||^2$
      and
      $\lambda ||E||^2$
      $= \mathrm{tr}(EE^*)$
      $= \sum_{k,l} \sum_{\tilde{k},\tilde{l}} A_{k,l} \bar{A}_{\tilde{k},\tilde{l}} \langle \psi(x_k), \psi(x_{\tilde{k}}) \rangle_{\mathcal{H}_\mathcal{X}} \langle \phi(z_l), \phi(z_{\tilde{l}}) \rangle_{\mathcal{H}_\mathcal{Z}}$

Two-Stage Kernel Ridge Regression

- Stage 2:
  - Solution

$$
\begin{aligned}
\hat{H}^n_\xi &= \arg \min_{H:\mathcal{H}_\mathcal{X}\to\mathbb{R}:\text{linear}} \frac{1}{m}\sum_{i=1}^{m} ||\tilde{y}_i - H\hat{E}^{n*}_\lambda \phi(\tilde{z}_i)||^2 + \xi ||H||^2 \\
&= \sum_{k=1}^{n} \hat{\alpha}_k \langle \psi(x_k), \ \cdot \ \rangle_{\mathcal{H}_\mathcal{X}}
\end{aligned}
$$

where

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

Two-Stage Kernel Ridge Regression

- Stage 2:
  - Derivation
    - ★ Since the range of E^* is limited to the linear combination of $\psi(x_1), ..., \psi(x_n) \in \mathcal{H}_\mathcal{X}$, we can use the representer theorem and write $H : \mathcal{H}_\mathcal{X} \to \mathbb{R}$ as $H = \sum_{k=1}^n \alpha_k \langle \psi(x_k), \ \cdot \ \rangle_{\mathcal{H}_\mathcal{X}}$
    - ★ Letting $W = K_{XX}(K_{ZZ} + n\lambda I)^{-1} K_{Z\tilde{Z}}$, we can write the loss as

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m ||\tilde{y}_i - H\hat{E}_\lambda^{n*}\phi(\tilde{z}_i)||^2 + \xi||H||^2 \\
&= \frac{1}{m}\sum_{i=1}^m ||\tilde{y}_i - \sum_{k=1}^n \alpha_k \left\langle \psi(x_k), \ \hat{E}_\lambda^{n*}\phi(\tilde{z}_i) \right\rangle_{\mathcal{H}_\mathcal{X}} ||^2 + \xi|| \sum_{k=1}^n \alpha_k \langle \psi(x_k), \ \cdot \ \rangle_{\mathcal{H}_\mathcal{X}} ||^2 \\
&= \frac{1}{m}\sum_{i=1}^m ||\tilde{y}_i - \sum_{k=1}^n \alpha_k \left\langle \psi(x_k), \ \sum_{k,l} [(K_{ZZ} + n\lambda I)^{-1}]_{k,l} \langle \phi(z_{\tilde{l}}), \phi(\tilde{z}_i) \rangle_{\mathcal{H}_\mathcal{Z}} \psi(x_{\tilde{k}}) \right\rangle_{\mathcal{H}_\mathcal{X}} ||^2 \\
&\quad + \xi|| \sum_{k=1}^n \alpha_k \psi(x_k) ||^2_{\mathcal{H}_\mathcal{X}} \\
&= \frac{1}{m}||\mathbf{y} - W'\alpha||^2 + \xi\alpha' K_{XX}\alpha
\end{aligned}
$$

Two-Stage Kernel Ridge Regression

- Causal Effect Prediction:

$$\begin{aligned}\hat{H}_{\xi}^{n}\psi(x_{\mathsf{new}}) &= \sum_{k=1}^{n}\alpha_k\langle\psi(x_k),\psi(x_{\mathsf{new}})\rangle_{\mathcal{H}_{\mathcal{X}}} \\ &= \hat{\alpha}' K_{X x_{\mathsf{new}}}\end{aligned}$$

Outline

Toy Problems

- $Y = h(X) + e, \quad \mathbb{E}[e|Z] = 0$ but $e \not\perp\!\!\!\perp X$ (non-linear dependency).
- Linear design: $h(x) = 4x - 2$
- Sigmoid design: $h(x) = \ln(|16x - 8| + 1) \cdot sgn(x - 0.5)$
- Demand design: highly non-linear

Sigmoid Design

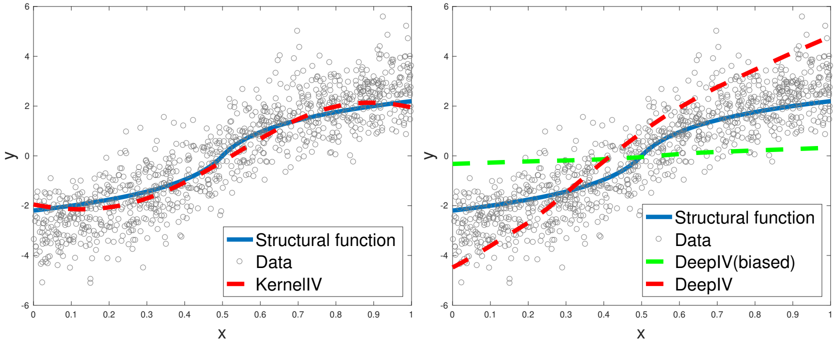


Figure: Kernel IV(left) and Deep IV(right) on the sigmoid design

Linear and Sigmoid Design

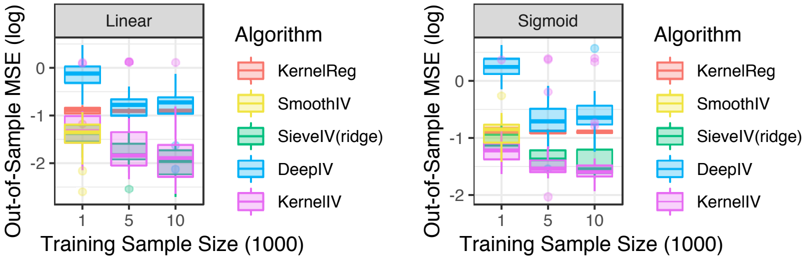


Figure: Linear(left) and sigmoid(right) designs

Demand Design

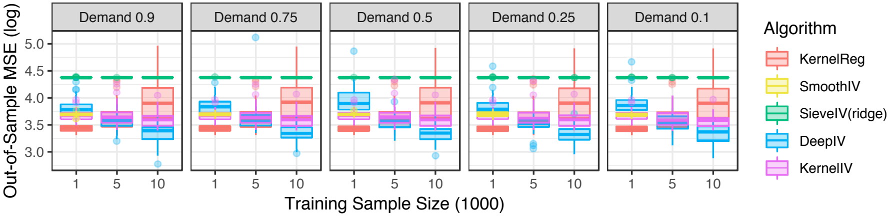


Figure: Demand design with different parameters (governing the strength of)

Fin

Outline

Remarks on Theoretical Results

- Theorem 2
  Under appropriate hypotheses, $\forall \delta \in (0,1)$, the following holds w.p. $1-\delta$:

$$\|E_\lambda^n - E_\rho\|_{\mathcal{H}_\Gamma} \leq r_E(\delta, n, c_1) := \frac{\sqrt{\zeta_1}(c_1+1)}{4^{\frac{1}{c_1+1}}} \left(\frac{4\kappa(Q + \kappa\|E_\rho\|_{\mathcal{H}_\Gamma}) \ln(2/\delta)}{\sqrt{n\zeta_1}(c_1-1)} \right)^{\frac{c_1-1}{c_1+1}}$$

$$\lambda = \left(\frac{8\kappa(Q + \kappa\|E_\rho\|_{\mathcal{H}_\Gamma}) \ln(2/\delta)}{\sqrt{n\zeta_1}(c_1-1)} \right)^{\frac{2}{c_1+1}}$$

- "Note that the convergence rate of E_λ^n is calibrated by c_1, which measures the smoothness of the conditional expectation operator $E : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}$."

Remarks on Theoretical Results

- Theorem 4
  Under appropriate hypotheses, by choosing $\lambda = n^{-\frac{1}{c_1+1}}$ and $n = m^{\frac{a(c_1+1)}{\iota(c_1-1)}}$ where $a > 0$, following convergence rate is achieved.
  1. If $a \leq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}(\hat{H}_\xi^m) - \mathcal{E}(H_\rho) = O_p(m^{-\frac{ac}{c+1}})$ with $\xi = m^{-\frac{a}{c+1}}$
  2. If $a \geq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}(\hat{H}_\xi^m) - \mathcal{E}(H_\rho) = O_p(m^{-\frac{bc}{bc+1}})$ with $\xi = m^{-\frac{b}{bc+1}}$
- "At $a = \frac{b(c+1)}{bc+1} < 2$, the convergence rate $m^{-\frac{bc}{bc+1}}$ is minimax optimal while requiring the fewest observations. This statistically efficient rate is calibrated by b, the effective input dimension, as well as c, the smoothness of structural operator H_ρ."

Reference

Rahul Singh, Maneesh Sahani, and Arthur Gretton.
Kernel instrumental variable regression.
In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.