Nonparametric Instrumental Variable Regression through Stochastic Gradients and Kernel Methods

Student: Caio Lins
Advisor: Yuri Saporito

EMAp – FGV

December 15, 2023

FGV EMAp

Summary

Instrumental Variable Regression

Nonparametric Instrumental Variable Regression

Stochastic Approximate Gradient Descent IV

Instrumental Variables

- Regression: $Y = h^\star(X) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$. Find $h^\star$.

Instrumental Variables

- Regression: $Y = h^\star(X) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$. Find $h^\star$.
- Minimizing $\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2]$ gives

$$\widehat{h}(X) = \mathbb{E}[Y \mid X] = h^\star(X) + \mathbb{E}[\varepsilon \mid X].$$

Instrumental Variables

- Regression: $Y = h^\star(X) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$. Find $h^\star$.
- Minimizing $\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2]$ gives

$$\widehat{h}(X) = \mathbb{E}[Y \mid X] = h^\star(X) + \mathbb{E}[\varepsilon \mid X].$$

- What if $\mathbb{E}[\varepsilon \mid X] \neq 0$? That is, if X is *endogenous*?

Instrumental Variables

- Suppose we have access to a variable Z such that

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. Z influences X, that is, $Z \not\perp\!\!\!\perp X$;

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. Z influences X, that is, $Z \not\perp\!\!\!\perp X$;
  2. Z is *exogenous*, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. Z influences X, that is, $Z \not\perp X$;
  2. Z is *exogenous*, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

  Z is called an *instrumental variable*.

Instrumental Variables

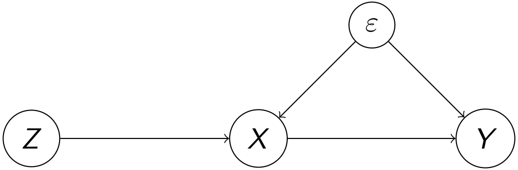
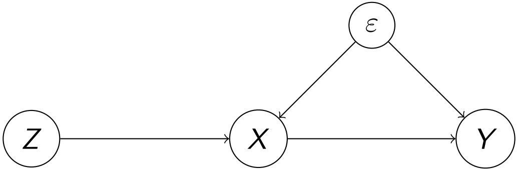


Figure: Causal diagram representing an instrumental variable for an endogenous covariate.

Instrumental Variables

Example:

- X = is smoker?
- Y = general health.
- Z = tax rate on tobacco products.
- ε = depression, self care.



Linear Model

$$\mathbf{Y} = \mathbf{X}\beta^\star + \varepsilon.$$

Linear Model

$$\mathbf{Y} = \mathbf{X}\beta^\star + \varepsilon.$$

- $\widehat{\beta}_{\mathrm{OLS}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$ is biased and inconsistent;

Linear Model

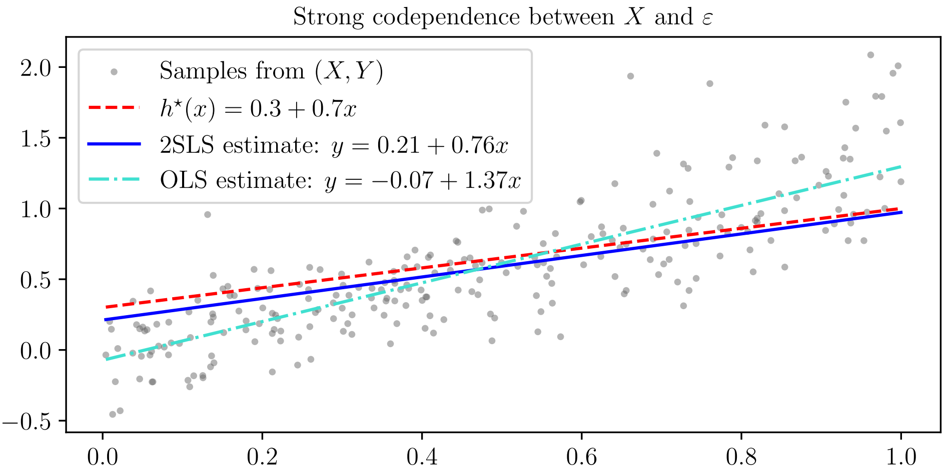
$$\mathbf{Y} = \mathbf{X}\beta^\star + \varepsilon.$$

- $\widehat{\beta}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is biased and inconsistent;
- $\widehat{\beta}_{\mathrm{IV}} = (\widehat{\mathbf{X}}^\top\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^\top\mathbf{Y}$, where

$$\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X},$$

  is unbiased and consistent.
- Two Stages Least Squares (2SLS).

Linear Model



Summary

Instrumental Variable Regression

Nonparametric Instrumental Variable Regression

Stochastic Approximate Gradient Descent IV

Problem formulation

$$Y = h^\star(X) + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon \mid X] \neq 0,\ Z \not\perp X \text{ and } \mathbb{E}[\varepsilon \mid Z] = 0.$$

- We have access to $\{(X_i, Y_i, Z_i)\}$;

Problem formulation

$$Y = h^\star(X) + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon \mid X] \neq 0,\ Z \not\perp X \text{ and } \mathbb{E}[\varepsilon \mid Z] = 0.$$

- We have access to $\{(X_i, Y_i, Z_i)\}$;
- We assume $h^\star \in L^2(X) = \left\{h : \mathbb{E}[h(X)^2] < \infty\right\}$.

Problem formulation

$$Y = h^\star(X) + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon \mid X] \neq 0,\ Z \not\perp\!\!\!\perp X \text{ and } \mathbb{E}[\varepsilon \mid Z] = 0.$$

- We have access to $\{(X_i, Y_i, Z_i)\}$;
- We assume $h^\star \in L^2(X) = \{h : \mathbb{E}[h(X)^2] < \infty\}$.
- Conditioning in Z:

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) \mid Z] \iff r = \mathcal{P}[h^\star],$$

where $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P} : L^2(X) \to L^2(Z)$ is the *conditional expectation operator*:

$$\mathcal{P}[h](z) = \mathbb{E}[h(X) \mid Z = z].$$

Problem formulation

$$Y = h^\star(X) + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon \mid X] \neq 0,\ Z \not\perp X \text{ and } \mathbb{E}[\varepsilon \mid Z] = 0.$$

- We have access to $\{(X_i, Y_i, Z_i)\}$;
- We assume $h^\star \in L^2(X) = \left\{h : \mathbb{E}[h(X)^2] < \infty\right\}$.
- Conditioning in Z:

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) \mid Z] \iff r = \mathcal{P}[h^\star],$$

where $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P} : L^2(X) \to L^2(Z)$ is the *conditional expectation operator*:

$$\mathcal{P}[h](z) = \mathbb{E}[h(X) \mid Z = z\,].$$

- Ill posed problem.

Classical Approaches

Problem: $r = \mathcal{P}[h^\star]$, where $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z]$.

Classical Approaches

Problem: $r = \mathcal{P}[h^\star]$, where $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z]$.

- Nonlinear Two Stages Least Squares [3]:
  - $h^\star(x) \approx \sum_{j=1}^{J} \gamma_j p_j(x)$;
  - $\mathbb{E}[p_j(X) \mid Z = z\,] \approx \sum_{i=1}^{n} a_{ji} q_i(z)$.

Classical Approaches

Problem: $r = \mathcal{P}[h^\star]$, where $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z]$.

- Nonlinear Two Stages Least Squares [3]:
  - $h^\star(x) \approx \sum_{j=1}^{J} \gamma_j p_j(x)$;
  - $\mathbb{E}[p_j(X) \mid Z = z\,] \approx \sum_{i=1}^{n} a_{ji} q_i(z)$.
- Iterated Tikhonov regularization [1]:
  - $\arg\min_h \|\mathcal{P}[h] - r\|_{L^2(Z)}^2 + \alpha\|h\|_{L^2(X)}^2 = (\mathcal{P}^*\mathcal{P} + \alpha I)^{-1}\mathcal{P}^*[r]$;
  - $h_{k+1} = (\mathcal{P}^*\mathcal{P} + \alpha I)^{-1}[\mathcal{P}^* r + h_k]$

Summary

Instrumental Variable Regression

Nonparametric Instrumental Variable Regression

Stochastic Approximate Gradient Descent IV

Risk measure

We know that $h^\star$ satisfies

$$r = \mathcal{P}[h^\star].$$

Risk measure

We know that $h^\star$ satisfies

$$r = \mathcal{P}[h^\star].$$

- Define the risk

  $$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

  E.g., if $\ell(y, y') = (y - y')^2$:

  $$\begin{aligned} \mathcal{R}(h) &= \mathbb{E}\left[(r(Z) - \mathcal{P}[h](Z))^2\right] \\ &= \mathbb{E}\left[(\mathcal{P}[h - h^\star](Z))^2\right]. \end{aligned}$$

Gradient Descent

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

Gradient Descent

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

- We showed that

$$\begin{aligned}\nabla\mathcal{R}(h)(x) &= \mathcal{P}^*[\partial_2(r(Z), \mathcal{P}[h](Z))] \\ &= \mathbb{E}[\Phi(x, Z)\partial_2(r(Z), \mathcal{P}[h](Z))],\end{aligned}$$

where $\Phi(x, z) = \frac{p_{XZ}(x,z)}{p_X(x)p_Z(z)}$.

Gradient Descent

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

- We showed that

$$\begin{aligned} \nabla\mathcal{R}(h)(x) &= \mathcal{P}^*[\partial_2(r(Z), \mathcal{P}[h](Z))] \\ &= \mathbb{E}[\Phi(x, Z)\partial_2(r(Z), \mathcal{P}[h](Z))], \end{aligned}$$

  where $\Phi(x, z) = \frac{p_{XZ}(x,z)}{p_X(x)p_Z(z)}$.
- The term in blue is a stochastic gradient.

Gradient Descent

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

- We showed that

$$\begin{aligned} \nabla\mathcal{R}(h)(x) &= \mathcal{P}^*[\partial_2(r(Z), \mathcal{P}[h](Z))] \\ &= \mathbb{E}[\Phi(x, Z)\partial_2(r(Z), \mathcal{P}[h](Z))], \end{aligned}$$

  where $\Phi(x, z) = \frac{p_{XZ}(x,z)}{p_X(x)p_Z(z)}$.
- The term in blue is a stochastic gradient.
- Problem: do not know $\Phi, \mathcal{P}$ nor r. Only have access to $\{(X_i, Y_i, Z_i)\}$.

Projected Gradient Descent

- Our regularization: look for solutions in $\mathcal{H} \subseteq L^2(X)$;

Projected Gradient Descent

- Our regularization: look for solutions in $\mathcal{H} \subseteq L^2(X)$;
- $\mathcal{H}$ is convex, closed and bounded.

Projected Gradient Descent

- Our regularization: look for solutions in $\mathcal{H} \subseteq L^2(X)$;
- $\mathcal{H}$ is convex, closed and bounded.
- E.g. for $A > 0$:

$$\mathcal{H} = \left\{ h \in L^2(X) : |h(x)| \leq A \ \forall x \right\}.$$

Stochastic Approximate Gradient Descent

- $\nabla\mathcal{R}(h)(x) = \mathbb{E}[\Phi(x, Z)\partial_2(r(X), \mathcal{P}[h](Z))]$, but we do not know $\Phi, r, \mathcal{P}$;

Stochastic Approximate Gradient Descent

- $\nabla \mathcal{R}(h)(x) = \mathbb{E}[\Phi(x, Z)\partial_2(r(X), \mathcal{P}[h](Z))]$, but we do not know $\Phi, r, \mathcal{P}$;
- Assume we have estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$, so that

$$u_h(x) = \widehat{\Phi}(x, Z)\partial_2(\widehat{r}(Z), \widehat{\mathcal{P}}[h](Z))$$

is an *approximate stochastic gradient* for $\mathcal{R}$ at h.

Kernel Methods

- Kernel Ridge Regression to compute $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$;

Kernel Methods

- Kernel Ridge Regression to compute $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$;
- Reproducing Kernel Hilbert Space (RKHS) as a class of approximating functions $\Longrightarrow$ closed form solutions;

Kernel Methods

- Kernel Ridge Regression to compute $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$;
- Reproducing Kernel Hilbert Space (RKHS) as a class of approximating functions $\implies$ closed form solutions;
- $\widehat{\mathcal{P}}$ is tricky. We used KIV's first stage [4].

SAGD–IV

Algorithm 1: SAGD–IV

input : Samples $\{(\boldsymbol{z}_m)_{m=1}^M\}$. Estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$. Sequence of learning rates $(\alpha_m)_{m=1}^M$.

output: $\widehat{h}$

for $1 \leq m \leq M$ **do**

  Set $u_m = \widehat{\Phi}(\cdot, \boldsymbol{z}_m)\partial_2 \ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)$;

  Set $\widehat{h}_m = \mathrm{proj}_{\mathcal{H}}\left[\widehat{h}_{m-1} - \alpha_m u_m\right]$;

end

Set $\widehat{h} = \frac{1}{M}\sum_{m=1}^M \widehat{h}_m$;

Theory

Theorem (SAGD–IV convergence rate)

Under suitable assumptions on $\ell, \mathcal{H}, \mathcal{P}$ *and the estimators* $\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}$, *we have*

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^{M}\alpha_m + \tau\sqrt{\zeta},$$

Theory

Theorem (SAGD–IV convergence rate)

Under suitable assumptions on $\ell, \mathcal{H}, \mathcal{P}$ and the estimators $\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}$, we have

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^{M}\alpha_m + \tau\sqrt{\zeta},$$

Where

$$\zeta = \left\|\Phi - \widehat{\Phi}\right\|^2_{L^2(\mathbb{P}_X \otimes \mathbb{P}_Z)} + \|r - \widehat{r}\|^2_{L^2(Z)} + \left\|\mathcal{P} - \widehat{\mathcal{P}}\right\|^2_{\mathrm{op}},$$

$$\xi = \frac{3}{2}\left\|\widehat{\Phi}\right\|^2_\infty \left(C_0^2 + L^2\|\widehat{r}\|^2_{L^2(Z)} + L^2 D^2\left\|\widehat{\mathcal{P}}\right\|^2_{\mathrm{op}}\right),$$

$$\tau = 2D\max\left\{3(C_0^2 + L^2\mathbb{E}[Y^2] + L^2D^2), 2L^2\left\|\widehat{\Phi}\right\|^2_\infty, 2L^2D^2\left\|\widehat{\Phi}\right\|^2_\infty\right\}.$$

Theory

- The bound

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^{M}\alpha_m + \tau\sqrt{\zeta},$$

suggests (α_m) should satisfy

$$M\alpha_M \to \infty \quad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M}\alpha_m \to 0.$$

Theory

- The bound

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^{M}\alpha_m + \tau\sqrt{\zeta},$$

suggests (α_m) should satisfy

$$M\alpha_M \to \infty \quad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M}\alpha_m \to 0.$$

- Additional samples from just Z can already increase estimator's quality.

Practice

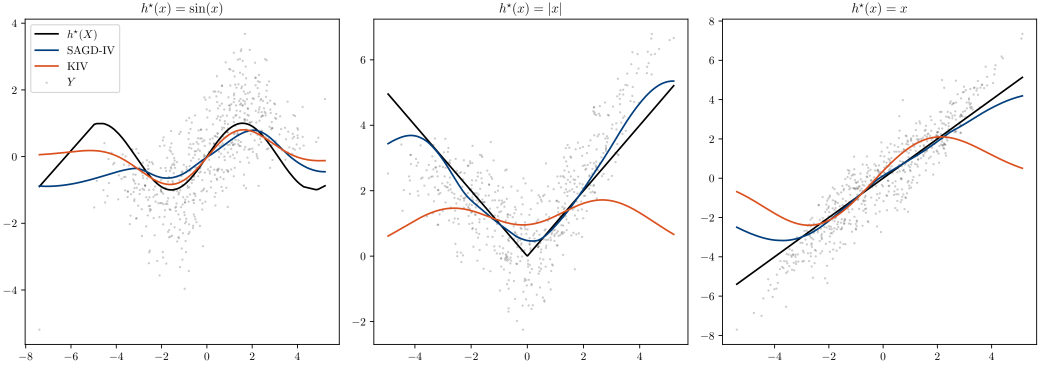


Figure: Benchmark against Kernel Instrumental Variable (KIV) [4]

Future Work

- Robust benchmarks against other recent methods;
- Application to discrete outcome models: $Y = \mathbf{1}\left\{h^{\star}(X) + \varepsilon > 0\right\}$.

References

[1] S. Darolles et al. "Nonparametric Instrumental Regression". In: *Econometrica* 79.5 (2011), pp. 1541–5165.

[2] Yuri R. Fonseca and Yuri F. Saporito. *Statistical Learning and Inverse Problems: A Stochastic Gradient Approach*. 2022. arXiv: 2209.14967 [stat.ML].

[3] Whitney K. Newey and James L. Powell. "Instrumental Variable Estimation of Nonparametric Models". In: *Econometrica* 71.5 (2003), pp. 1565–1578. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1555512 (visited on 07/03/2023).

[4] Rahul Singh, Maneesh Sahani, and Arthur Gretton. "Kernel Instrumental Variable Regression". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

Thank You!