NPIV Estimation through Stochastic Gradients and Kernel Methods

Student: Caio Lins
Advisor: Yuri Saporito

EMAp – FGV

October 18, 2023

Summary

NPIV estimation

Our approach

Where we are at

Next steps

Instrumental Variables

- Consider a generic regression problem:

$$Y = h^{\star}(X) + \varepsilon,$$

where $\mathbb{E}[\varepsilon] = 0$ and we wish to estimate $h^{\star}$.

Instrumental Variables

- Consider a generic regression problem:

$$Y = h^\star(X) + \varepsilon,$$

  where $\mathbb{E}[\varepsilon] = 0$ and we wish to estimate $h^\star$.
- What happens if $\varepsilon \not\perp\!\!\!\perp X$? That is, $\mathbb{E}[\varepsilon \mid X] \neq 0$?

Instrumental Variables

- Consider a generic regression problem:

$$Y = h^\star(X) + \varepsilon,$$

  where $\mathbb{E}[\varepsilon] = 0$ and we wish to estimate $h^\star$.
- What happens if $\varepsilon \not\perp\!\!\!\perp X$? That is, $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathbb{E}[(Y - h(X))^2]$ over h gives biased results.

Instrumental Variables

- Consider a generic regression problem:

$$Y = h^\star(X) + \varepsilon,$$

  where $\mathbb{E}[\varepsilon] = 0$ and we wish to estimate $h^\star$.
- What happens if $\varepsilon \not\perp\!\!\!\perp X$? That is, $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathbb{E}[(Y - h(X))^2]$ over h gives biased results.

$$Y = \underbrace{h^\star(X) + \mathbb{E}[\varepsilon \mid X]}_{= f(X) \text{ for some } f} + (\varepsilon - \mathbb{E}[\varepsilon \mid X]).$$

Instrumental Variables

- Consider a generic regression problem:

$$Y = h^\star(X) + \varepsilon,$$

  where $\mathbb{E}[\varepsilon] = 0$ and we wish to estimate $h^\star$.
- What happens if $\varepsilon \not\perp\!\!\!\perp X$? That is, $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathbb{E}[(Y - h(X))^2]$ over h gives biased results.

$$Y = \underbrace{h^\star(X) + \mathbb{E}[\varepsilon \mid X]}_{= f(X) \text{ for some } f} + (\varepsilon - \mathbb{E}[\varepsilon \mid X]).$$

- We end up estimating f instead of $h^\star$!

Instrumental Variable

- Suppose we have access to a variable Z such that

Instrumental Variable

- Suppose we have access to a variable Z such that
  1. $Z \not\perp\!\!\!\perp X$, i.e., $\mathbb{E}[X \mid Z]$ is not constant,

Instrumental Variable

- Suppose we have access to a variable Z such that
  1. $Z \not\perp\!\!\!\perp X$, i.e., $\mathbb{E}[X \mid Z]$ is not constant,
  2. Z affects Y only through X,

Instrumental Variable

- Suppose we have access to a variable Z such that
  1. $Z \not\perp\!\!\!\perp X$, i.e., $\mathbb{E}[X \mid Z]$ is not constant,
  2. Z affects Y only through X,
  3. $\varepsilon \perp\!\!\!\perp Z$, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

Instrumental Variable

- Suppose we have access to a variable Z such that
  1. $Z \not\perp\!\!\!\perp X$, i.e., $\mathbb{E}[X \mid Z]$ is not constant,
  2. Z affects Y only through X,
  3. $\varepsilon \perp\!\!\!\perp Z$, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

  Z is called an *instrumental variable*.

Instrumental Variable

- Suppose we have access to a variable Z such that
  1. $Z \not\perp\!\!\!\perp X$, i.e., $\mathbb{E}[X \mid Z]$ is not constant,
  2. Z affects Y only through X,
  3. $\varepsilon \perp\!\!\!\perp Z$, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

  Z is called an *instrumental variable*.
- How does it help us?

Instrumental Variable

- Structural equation:
$$Y = h^{\star}(X) + \varepsilon,$$
  where $\mathbb{E}[\varepsilon \mid Z] = 0$.

Instrumental Variable

- Structural equation:
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$.
- Consider minimizing $\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right]$ over h.

Instrumental Variable

- Structural equation:
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$.
- Consider minimizing $\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right]$ over h.
- Since
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid Z],$$

Instrumental Variable

- Structural equation:
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$.
- Consider minimizing $\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right]$ over h.
- Since
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid Z],$$
We have
$$\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}\left[(h^\star - h)(X) \mid Z\right])^2\right].$$

Instrumental Variable

- Structural equation:
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$.
- Consider minimizing $\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right]$ over h.
- Since
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid Z],$$
We have
$$\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}\left[(h^\star - h)(X) \mid Z\right])^2\right].$$
- $\mathcal{R}(h) = 0 \iff \mathbb{E}\left[(h^\star - h)\,(X) \mid Z\right] = 0 \iff \mathbb{E}[h^\star(X) \mid Z] = \mathbb{E}[h(X) \mid Z]$.

Instrumental Variable

- Structural equation:
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$.
- Consider minimizing $\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right]$ over h.
- Since
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid Z],$$
We have
$$\mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}\left[(h^\star - h)(X) \mid Z\right])^2\right].$$
- $\mathcal{R}(h) = 0 \iff \mathbb{E}\left[(h^\star - h)(X) \mid Z\right] = 0 \iff \mathbb{E}[h^\star(X) \mid Z] = \mathbb{E}[h(X) \mid Z]$.
- Still does *not* imply $h = h^\star$, but reduces bias *if Z is a good instrument*.

Example

$$\underbrace{\text{Grades}}_{Y} = h^{\star}\big(\underbrace{\text{Attends tutoring sessions?}}_{X}\big) + \varepsilon.$$

Example

$$\underbrace{\text{Grades}}_{Y} = h^{\star}\big(\underbrace{\text{Attends tutoring sessions?}}_{X}\big) + \varepsilon.$$

- Natural ability is a confounding variable: maybe only people who struggle a lot go to tutoring sessions.

Example

$$\underbrace{\text{Grades}}_{Y} = h^\star\big(\underbrace{\text{Attends tutoring sessions?}}_{X}\big) + \varepsilon.$$

- Natural ability is a confounding variable: maybe only people who struggle a lot go to tutoring sessions.
- Z = Lives close to school?

Example

$$\underbrace{\text{Grades}}_{Y} = h^{\star}\big(\underbrace{\text{Attends tutoring sessions?}}_{X}\big) + \varepsilon.$$

- Natural ability is a confounding variable: maybe only people who struggle a lot go to tutoring sessions.
- $Z =$ Lives close to school?
  1. $Z \not\perp\!\!\!\perp X$,
  2. Z affects Y only through X,
  3. $\varepsilon \perp\!\!\!\perp Z$.

Example

$$\underbrace{\text{Grades}}_{Y} = h^\star\big(\underbrace{\text{Attends tutoring sessions?}}_{X}\big) + \varepsilon.$$

- Natural ability is a confounding variable: maybe only people who struggle a lot go to tutoring sessions.
- $Z =$ Lives close to school?
  1. $Z \not\perp\!\!\!\perp X$,
  2. Z affects Y only through X, (Kind of)
  3. $\varepsilon \perp\!\!\!\perp Z$.

NPIV estimation

- Stands for “Nonparametric Instrumental Variable estimation”.

NPIV estimation

- Stands for "Nonparametric Instrumental Variable estimation".
- No assumptions on some parametric form for $h^\star$.

Summary

NPIV estimation

Our approach

Where we are at

Next steps

Problem formulation

- We have

$$Y = h^\star(X) + \varepsilon,$$

where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.

Problem formulation

- We have
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Equivalently,
$$r_0(Z) = \mathcal{T}[h^\star](Z),$$
with

Problem formulation

- We have
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Equivalently,
$$r_0(Z) = \mathcal{T}[h^\star](Z),$$
with
    - $r_0(Z) = \mathbb{E}[Y \mid Z]$,
    - $\mathcal{T}[h](Z) = \mathbb{E}[h(X) \mid Z]$.

Problem formulation

- We have
$$Y = h^\star(X) + \varepsilon,$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Equivalently,
$$r_0(Z) = \mathcal{T}[h^\star](Z),$$
with
  - $r_0(Z) = \mathbb{E}[Y \mid Z]$,
  - $\mathcal{T}[h](Z) = \mathbb{E}[h(X) \mid Z]$.
- Risk measure:
$$\mathcal{R}(h) = \mathbb{E}\left[\frac{1}{2}\left(\mathbb{E}\left[Y - h(X) \mid Z\right]\right)^2\right] = \mathbb{E}\left[\frac{1}{2}\left(r_0(Z) - \mathcal{T}[h](Z)\right)^2\right].$$

Stochastic Gradients

- It turns out that $\nabla \mathcal{R}(h)(X) = \mathcal{T}^*[\mathcal{T}[h] - r_0](X)$.

Stochastic Gradients

- It turns out that $\nabla\mathcal{R}(h)(X) = \mathcal{T}^*[\mathcal{T}[h] - r_0](X)$.
- Immediate idea:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow h_{t-1} - \alpha_t \nabla\mathcal{R}(h_{t-1}) \quad \text{for t } \geq 1 \ . \end{cases}$$

Stochastic Gradients

- Problem: *We don't observe* r_0 *neither know how to compute* $\mathcal{T}^*$ *nor* $\mathcal{T}$.

Stochastic Gradients

- Problem: *We don't observe* r_0 *neither know how to compute* $\mathcal{T}^*$ *nor* $\mathcal{T}$. We only have access to joint independent samples from X, Y and Z.

Stochastic Gradients

- Problem: *We don't observe r_0 neither know how to compute $\mathcal{T}^*$ nor $\mathcal{T}$.* We only have access to joint independent samples from X, Y and Z.
- Solution 1: "No problem, we estimate everything!" ...doable, but horrible, since $\mathcal{T}^*[\mathcal{T}[h] - r_0]$ involves plugin estimates into other estimates. Goodbye theoretical guarantees.

Stochastic Gradients

- Solution 2: Notice that

$$\nabla \mathcal{R}(h)(X) = \mathbb{E}_Z \left[\Phi(X, Z)(\mathcal{T}[h](Z) - r_0(Z)) \right],$$

where $\Phi(x, z) = \frac{p(x,z)}{p(x)p(z)}$.

Stochastic Gradients

- Solution 2: Notice that

$$\nabla \mathcal{R}(h)(X) = \mathbb{E}_Z \left[\Phi(X, Z)(\mathcal{T}[h](Z) - r_0(Z)) \right],$$

  where $\Phi(x, z) = \frac{p(x,z)}{p(x)p(z)}$.
- Second idea: *Now* we estimate everything:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow \widehat{\Phi}(\cdot, Z_i) \left(\widehat{\mathcal{T}[h_{t-1}]}(Z_i) - \widehat{r}_0(Z_i) \right). \end{cases}$$

Stochastic Gradients

- Solution 2: Notice that

$$\nabla\mathcal{R}(h)(X) = \mathbb{E}_Z\left[\Phi(X,Z)(\mathcal{T}[h](Z) - r_0(Z))\right],$$

  where $\Phi(x,z) = \frac{p(x,z)}{p(x)p(z)}$.
- Second idea: *Now* we estimate everything:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow \widehat{\Phi}(\cdot, Z_i)\left(\widehat{\mathcal{T}[h_{t-1}]}(Z_i) - \widehat{r}_0(Z_i)\right). \end{cases}$$

- ...Not pretty, but manageable, since we no longer have iterated conditional expectations

Stochastic Gradients

- Solution 2: Notice that

$$\nabla\mathcal{R}(h)(X) = \mathbb{E}_Z\left[\Phi(X,Z)(\mathcal{T}[h](Z) - r_0(Z))\right],$$

where $\Phi(x,z) = \frac{p(x,z)}{p(x)p(z)}$.

- Second idea: *Now* we estimate everything:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow \widehat{\Phi}(\cdot, Z_i)\left(\widehat{\mathcal{T}[h_{t-1}]}(Z_i) - \widehat{r}_0(Z_i)\right). \end{cases}$$

- ...Not pretty, but manageable, since we no longer have iterated conditional expectations (but must estimate ratio of densities).

Summary

NPIV estimation

Our approach

Where we are at

Next steps

Prototype

Prototype gave reasonable results

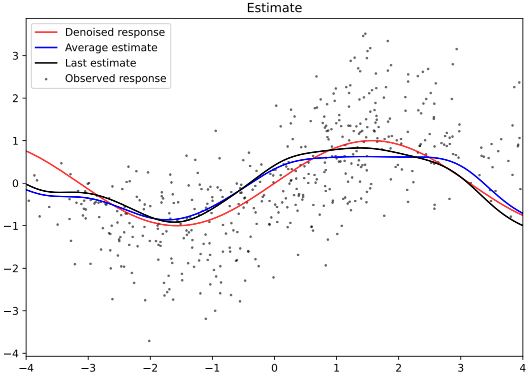

Figure: In red we have $h^\star = \sin$, in black we have h_N and in blue, $\frac{1}{N}\sum_{t=1}^{N} h_N$.

Theoretical properties

- Still working on convergence guarantees.

Theoretical properties

- Still working on convergence guarantees.
- This is helping us find better ways to estimate Φ and $\mathcal{T}$ (mainly RKHS methods).

Summary

NPIV estimation

Our approach

Where we are at

Next steps

Next steps

- Finalize convergence guarantees.

Next steps

- Finalize convergence guarantees.
- Implement modifications which the theory points to.

Next steps

- Finalize convergence guarantees.
- Implement modifications which the theory points to.
- Benchmark against current methods.

References

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

[2] 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).

Thank You!