NPIV Estimation through Stochastic Gradients and Kernel Methods

Student: Caio Lins
Advisor: Yuri Saporito

EMAp – FGV

October 19, 2023

Summary

NPIV estimation

Our approach

Results

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 $\varepsilon \not\perp\!\!\!\perp X$? More specifically, if $\mathbb{E}[\varepsilon \mid X] \neq 0$?

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 $\varepsilon \not\perp\!\!\!\perp X$? More specifically, if $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2]$ 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 $\varepsilon \not\perp\!\!\!\perp X$? More specifically, if $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2]$ over h gives biased results:

$$\underset{W = h(X) \text{ for some } h}{\arg\min} \mathbb{E}[(Y - W)^2] = \mathbb{E}[Y \mid X] = h^\star(X) + \underbrace{\mathbb{E}[\varepsilon \mid X]}_{\neq 0}.$$

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 $\varepsilon \not\perp\!\!\!\perp X$? More specifically, if $\mathbb{E}[\varepsilon \mid X] \neq 0$?
- Minimizing $\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2]$ over h gives biased results:

$$\operatorname*{arg\,min}_{W = h(X) \text{ for some } h} \mathbb{E}[(Y - W)^2] = \mathbb{E}[Y \mid X] = h^\star(X) + \underbrace{\mathbb{E}[\varepsilon \mid X]}_{\neq 0}.$$

- We end up estimating $h^\star(X) + \mathbb{E}[\varepsilon \mid X]$. Other problems may occur.

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. $p_{X|Z}(x \mid z)$ is not constant in z, that is, Z influences X;

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. $p_{X|Z}(x \mid z)$ is not constant in z, that is, Z influences X;
  2. ε is uncorrelated with Z, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. $p_{X|Z}(x \mid z)$ is not constant in z, that is, Z influences X;
  2. ε is uncorrelated with Z, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. $p_{X|Z}(x \mid z)$ is not constant in z, that is, Z influences X;
  2. ε is uncorrelated with Z, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

  Z is called an *instrumental variable*.

Instrumental Variables

- Suppose we have access to a variable Z such that
  1. $p_{X|Z}(x \mid z)$ is not constant in z, that is, Z influences X;
  2. ε is uncorrelated with Z, that is, $\mathbb{E}[\varepsilon \mid Z] = 0$.

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

Instrumental Variables

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

Instrumental Variables

- 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 Variables

- 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.
- Compare:

$$\mathrm{MSE}(h) = \mathbb{E}[(Y - h(X))^2] \quad \text{v.s.} \quad \mathcal{R}(h) = \mathbb{E}\left[(\mathbb{E}[Y - h(X) \mid Z])^2\right].$$

Instrumental Variables

- Since

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid X] + \cancelto{0}{\mathbb{E}[\varepsilon \mid Z]} = \mathbb{E}[h^\star(X) \mid Z],$$

Instrumental Variables

- Since

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid X] + \cancelto{0}{\mathbb{E}[\varepsilon \mid Z]} = \mathbb{E}[h^\star(X) \mid Z],$$

We have

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

Instrumental Variables

- Since

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) + \varepsilon \mid Z] = \mathbb{E}[h^\star(X) \mid X] + \cancelto{0}{\mathbb{E}[\varepsilon \mid Z]} = \mathbb{E}[h^\star(X) \mid Z],$$

  We have

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

- $\mathcal{R}(h) = 0 \iff \mathbb{E}\left[(h^\star - h)(X) \mid Z\right] = 0 \iff \mathbb{E}[h^\star(X) \mid Z] = \mathbb{E}[h(X) \mid Z].$

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. $\varepsilon \perp\!\!\!\perp Z$.

NPIV estimation

- Stands for “Nonparametric Instrumental Variable estimation”.

NPIV estimation

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

Summary

Problem Formulation

- We have
$$Y = h^\star(X) + \varepsilon, \tag{1}$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.

Problem Formulation

- We have
$$Y = h^\star(X) + \varepsilon, \tag{1}$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Let $\mathcal{P}[h] = \mathbb{E}[h(X) \mid Z = \cdot]$, that is
$$\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z].$$

Problem Formulation

- We have
$$Y = h^\star(X) + \varepsilon, \tag{1}$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Let $\mathcal{P}[h] = \mathbb{E}[h(X) \mid Z = \cdot]$, that is
$$\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z].$$
- Let $r_0(Z) = \mathbb{E}[Y \mid Z]$. Then (1) is equivalent to:
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) \mid Z] \iff r_0 = \mathcal{P}[h^\star].$$

Problem Formulation

- We have
$$Y = h^\star(X) + \varepsilon, \tag{1}$$
where $\mathbb{E}[\varepsilon \mid Z] = 0$, and want to estimate $h^\star$.
- Let $\mathcal{P}[h] = \mathbb{E}[h(X) \mid Z = \cdot]$, that is
$$\mathcal{P}[h](Z) = \mathbb{E}[h(X) \mid Z].$$
- Let $r_0(Z) = \mathbb{E}[Y \mid Z]$. Then (1) is equivalent to:
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) \mid Z] \iff r_0 = \mathcal{P}[h^\star].$$
- We wish to "invert" $\mathcal{P}$.
- Risk measure:
$$\mathcal{R}(h) = \mathbb{E}\left[\frac{1}{2}\left(\mathbb{E}\left[Y - h(X) \mid Z\right]\right)^2\right] = \mathbb{E}\left[\frac{1}{2}\left(r_0(Z) - \mathcal{T}[h](Z)\right)^2\right].$$

Problem Formulation

- Our risk measure:

$$\begin{aligned}\mathcal{R}(h) &= \frac{1}{2}\mathbb{E}\left[(\mathbb{E}\left[Y - h(X) \mid Z\right])^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[(r_0(Z) - \mathcal{P}[h](Z))^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[(r_0 - \mathcal{P}[h])(Z)^2\right].\end{aligned}$$

Stochastic Gradients

- We showed that $\nabla \mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$, where

$$\mathcal{P}^*[f](X) = \mathbb{E}[f(Z) \mid X].$$

Stochastic Gradients

- We showed that $\nabla\mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$, where

$$\mathcal{P}^*[f](X) = \mathbb{E}[f(Z) \mid X].$$

- Immediate idea:

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

Stochastic Gradients

- We showed that $\nabla \mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$, where

$$\mathcal{P}^*[f](X) = \mathbb{E}[f(Z) \mid X].$$

- Immediate idea:

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

- There are problems...

Stochastic Gradients

- Another idea: notice that

$$
\begin{aligned}
\mathcal{P}^*[f](x) &= \mathbb{E}\left[f(Z) \mid X = x\right] \\
&= \int_{\mathcal{Z}} f(z) p(z \mid x) \, \mathrm{d}z \\
&= \int_{\mathcal{Z}} f(z) \frac{p(x, z)}{p(x) p(z)} p(z) \, \mathrm{d}z \\
&= \mathbb{E}_Z\left[f(Z) \Phi(x, Z)\right],
\end{aligned}
$$

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

Stochastic Gradients

- Another idea: notice that

$$
\begin{aligned}
\mathcal{P}^*[f](x) &= \mathbb{E}\left[f(Z) \mid X = x\right] \\
&= \int_{\mathcal{Z}} f(z) p(z \mid x) \, \mathrm{d}z \\
&= \int_{\mathcal{Z}} f(z) \frac{p(x, z)}{p(x)p(z)} p(z) \, \mathrm{d}z \\
&= \mathbb{E}_Z \left[f(Z)\Phi(x, Z)\right],
\end{aligned}
$$

  where $\Phi(x, z) = \frac{p(x,z)}{p(x)p(z)}$.
- In the spirit of SGD, $f(Z)\Phi(x, Z)$ is a stochastic estimate for $\mathcal{P}^*[f](x)$.

Stochastic Gradients

- Substitute $\nabla \mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$ for

$$\widehat{\Phi}(\cdot, Z)\left(\widehat{\mathcal{P}}[h](Z) - \widehat{r}_0(Z)\right).$$

Stochastic Gradients

- Substitute $\nabla \mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$ for

$$\widehat{\Phi}(\cdot, Z)\left(\widehat{\mathcal{P}}[h](Z) - \widehat{r}_0(Z)\right).$$

- New algorithm:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow h_{t-1} - \alpha_t \widehat{\Phi}(\cdot, z_t)\left(\widehat{\mathcal{P}}[h](z_t) - \widehat{r}_0(z_t)\right) \quad \text{for t } \geq 1 \ . \end{cases}$$

Stochastic Gradients

- Substitute $\nabla\mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$ for

$$\widehat{\Phi}(\cdot, Z)\left(\widehat{\mathcal{P}}[h](Z) - \widehat{r}_0(Z)\right).$$

- New algorithm:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow h_{t-1} - \alpha_t \widehat{\Phi}(\cdot, z_t)\left(\widehat{\mathcal{P}}[h](z_t) - \widehat{r}_0(z_t)\right) & \text{for t } \geq 1 \ . \end{cases}$$

- We chose to use RKHS (kernel) methods for computing $\widehat{\Phi}, \widehat{\mathcal{P}}$ and $\widehat{r}_0$.

Stochastic Gradients

- Substitute $\nabla\mathcal{R}(h) = \mathcal{P}^*[\mathcal{P}[h] - r_0]$ for

$$\widehat{\Phi}(\cdot, Z)\left(\widehat{\mathcal{P}}[h](Z) - \widehat{r}_0(Z)\right).$$

- New algorithm:

$$\begin{cases} h_0 \equiv 0, \\ h_t \leftarrow h_{t-1} - \alpha_t \widehat{\Phi}(\cdot, z_t)\left(\widehat{\mathcal{P}}[h](z_t) - \widehat{r}_0(z_t)\right) \quad \text{for t } \geq 1 . \end{cases}$$

- We chose to use RKHS (kernel) methods for computing $\widehat{\Phi}, \widehat{\mathcal{P}}$ and $\widehat{r}_0$.
- One dataset with samples from (X, Z, Y) to compute $\widehat{\Phi}, \widehat{\mathcal{P}}, \widehat{r}_0$, and one dataset with samples from Z to conduct SGD loop.

Summary

Practical Results

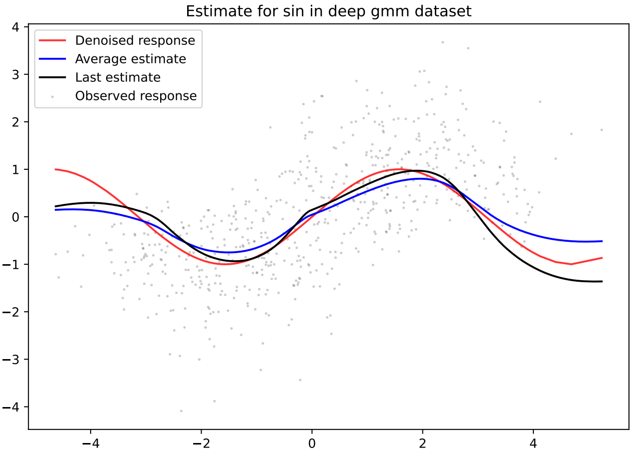

Figure: In red we have $h^\star = \sin$, in black we have $\widehat{h}_N$ and in blue, $h = \frac{1}{N}\sum_{t=1}^{N} h_N$. Results produced with 600 joint samples of (X, Z, Y) and 2000 more samples of Z only.

Practical Results

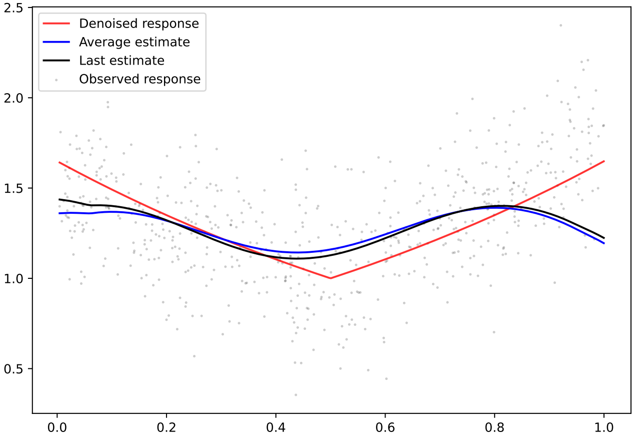

Figure: In red we have $h^{\star}(x) = \exp(|x|)$, in black we have $\widehat{h}_N$ and in blue, $h = \frac{1}{N}\sum_{t=1}^{N} h_N$. Results produced with 600 joint samples of (X, Z, Y) and 2000 more samples of Z only.

Practical Results

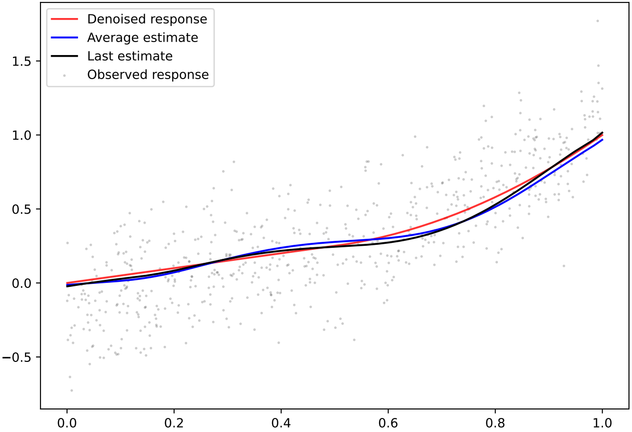

Figure: In red we have $h^\star(x) = 2((x - 1/2)^+)^2 + x/2$, in black we have $\widehat{h}_N$ and in blue, $h = \frac{1}{N}\sum_{t=1}^{N} h_N$. Results produced with 600 joint samples of (X, Z, Y) and 2000 more samples of Z only.

Theoretical properties

Letting $\bar{h} = \frac{1}{N}\sum_{t=1}^{N} h_t$, we have

$$\mathbb{E}_{z_{1:N}}\left[\mathcal{R}\left(\bar{h}\right)\right] \leq \frac{D^2}{2N\alpha_N} + \mathcal{O}_p(1)\frac{1}{N}\sum_{t=1}^{N}\alpha_t + \mathcal{O}_p(1)\left(\left\|\widehat{\Phi} - \Phi\right\|^2 + \left\|\widehat{r_0} - r_0\right\|^2 + \left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|^2\right)^{1/2}.$$

Theoretical properties

Letting $\bar{h} = \frac{1}{N}\sum_{t=1}^{N} h_t$, we have

$$\mathbb{E}_{z_{1:N}}\left[\mathcal{R}\left(\bar{h}\right)\right] \leq \frac{D^2}{2N\alpha_N} + \mathcal{O}_p(1)\frac{1}{N}\sum_{t=1}^{N}\alpha_t$$
$$+ \mathcal{O}_p(1)\left(\left\|\widehat{\Phi} - \Phi\right\|^2 + \|\widehat{r}_0 - r_0\|^2 + \left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|^2\right)^{1/2}.$$

Choose $(\alpha_t)_{t=1}^{\infty}$ so that $N\alpha_N \to \infty$ but $\frac{1}{N}\sum_{t=1}^{N}\alpha_t \to 0$ as $N \to \infty$.

Summary

NPIV estimation

Our approach

Results

Next steps

Next steps

- Benchmark against current methods.
- Discrete outcome models.

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!