Learning Causal Models from Conditional Moment Restrictions by Importance Weighting

ICLR 2022

Masahiro Kato, CyberAgent, Inc. AILab / The University of Tokyo
Masaaki Imaizumi, The University of Tokyo
Kenichiro McAlinn, Temple University
Shota Yasui, Cyberagent, Inc.
Haruo Kakehi, Cyberagent, Inc.

Structural Equation Model

- What is structural equation model?
- Consider the following linear model between Y and X:

$$Y = X^\top \beta + \varepsilon, \qquad \mathbb{E}[X^\top \varepsilon] \neq 0.$$

- $\mathbb{E}[X^\top \varepsilon] \neq 0$ implies the correlation between ε and X.
- This situation is called endogeneity.
- In this case, an OLS estimator is not unbiased and consistent.
  - $X^\top \beta$ is not the conditional mean $\mathbb{E}[Y|X]$ ($\mathbb{E}[Y|X] \neq X^\top \beta$).
- This model is called structural equation model (Hansen (2022)).

Wage Equation

- The true wage equation:

$$\log(wage) = \beta_0 + years\ of\ education \times \beta_1 + ability \times \beta_2 + u,$$

$$\mathbb{E}[u | years\ of\ education, ability] = 0$$

- We cannot observe the "ability" and estimate the following model:

$$\log(wage) = \beta_0 + years\ of\ education \times \beta_1 + \varepsilon, \qquad \varepsilon = ability \times \beta_2 + u.$$

• If "$years\ of\ education$" is correlated with "$ability$,"

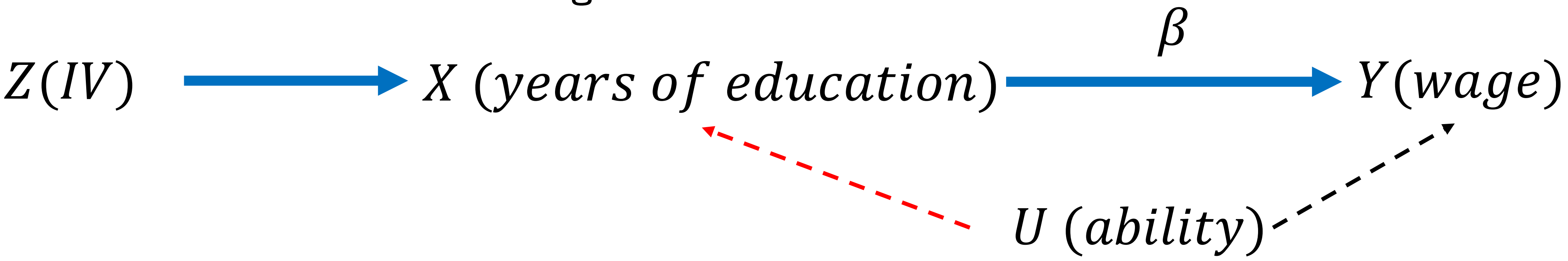
$$\mathbb{E}[\text{“}years\ of\ education\text{”} \times \varepsilon] \neq 0$$

→ We cannot consistently estimate β_1 with OLS.

Instrumental Variable (IV) Method

- By using IVs, we can estimate the parameter β.
- The IV is a random variable Z satisfying the following conditions:

1. Uncorrelated to the error term: $\mathbb{E}[Z^\top \varepsilon] = 0$.
2. Correlated with the endogeneous variable X.



Angrist and Krueger (1991)

- Estimation of the wage equation.
- IVs：correlated with the years of education and uncorrelated with the ability.
- **Angrist and Krueger (1991)**: use the education system in US.
- Enter school in the calendar year in which students turn 6.
- Require students to remain in school only until their 16th birthday,
- Attend school for different lengths of time depending on birthdays.
- Birthdays are irreverent to the ability of the students.

Nonparametric Instrumental Variable (NPIV) Regression

- A nonparametric version of IV problems (Newey and Powell (2003)):

$$Y = f^*(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] \neq 0.$$

  - Want to estimate the structural function f^*.
  - $\mathbb{E}[\varepsilon|X] \neq 0 \rightarrow$ least-squires does not yield consistent estimator.
- Instrumental variable Z: the condition for IVs: $\mathbb{E}[\varepsilon|Z] = 0$.
- Existing methods: two-stage least squares with series regression (Newey and Powell (2003)), minimax optimization (Dikkala et al. (2020)).

NPIV via Importance Weighting

- We solve the problem by using the conditional density ratio funtion.

Ex. Covariate shift adaptation by importance weighting (Shimodaira (2000)).

- From $\mathbb{E}_{Y,X}[\varepsilon|Z] = 0$, if we know $r^*(y,x|z) = \frac{p^*(y,x|z)}{p(y,x)}$, we can estimate f^* by minimizing an empirical approximation of $\mathbb{E}_Z\left[\left(\mathbb{E}_{Y,X}[\varepsilon|Z]\right)^2\right]$:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j=1}^{n} \big(Y_i - f(X_i)\big)\, r^*(y,x|z) \right)^2,$$

where $\mathcal{F}$ is the hypothesis class. We can use neural networks, RKHS, etc.

NPIV via Importance Weighting

- Estimate $r^*(y, x|z) = \frac{p^*(y,x|z)}{p(y,x)} = \frac{p^*(y,x,z)}{p(y,x)p(z)}$ as

$$r^* = \arg\min_r \mathbb{E}_Z\left[\mathbb{E}_{Y,X}\left[\left(r^*(Y,X|Z) - r(Y,X|Z)\right)^2\right]\right]$$
$$= \arg\min_r \mathbb{E}_Z\left[\mathbb{E}_{Y,X}\left[\left(r^*(Y,X|Z)\right)^2 - 2r^*(Y,X|Z)r(Y,X|Z) + r^2(Y,X|Z)\right]\right]$$
$$= \arg\min_r \mathbb{E}_Z\left[\mathbb{E}_{Y,X}[-2r^*(Y,X|Z)r(Y,X|Z) + r^2(Y,X|Z)]\right]$$
$$= \arg\min_r -2\mathbb{E}_Z\left[\mathbb{E}_{Y,X}[r(Y,X|Z)]\right] + \mathbb{E}_{Y,X,Z}[r^2(Y,X|Z)].$$

Ex. least-Squares Importance Fitting (LSIF, Kanamori et al. (2009))

Estimation Error Analysis

- Our goal is to show the upper bound of

$$\mathbb{E}\left[\left(f^*(X_i) - \hat{f}(X_i)\right)^2\right].$$

Lemma: estimation error of the conditional density ratio

Under appropriate conditions, we have

$$\sqrt{\mathbb{E}_Z\left[\mathbb{E}_{Y,X}\left[\left(r^*(Y_i, X_i|Z_i) - \hat{r}(Y_i, X_i|Z_i)\right)^2\right]\right]} = O_p\left(n^{-1/(2+\gamma)}\right),$$

where γ corresponds to the smoothness of r^*.

Estimation Error Analysis

- By using this lemma, we can obtain the following theorem.

Theorem: mean squared error of the conditional moment restrictions

Under appropriate conditions, we have

$$\mathbb{E}_Z\left[\left(\mathbb{E}_{Y,X}\left[Y_i - \hat{f}(X_i)\middle|Z_i\right]\right)^2\right] = O_p\left(n^{-1/(2+\gamma)}\right).$$

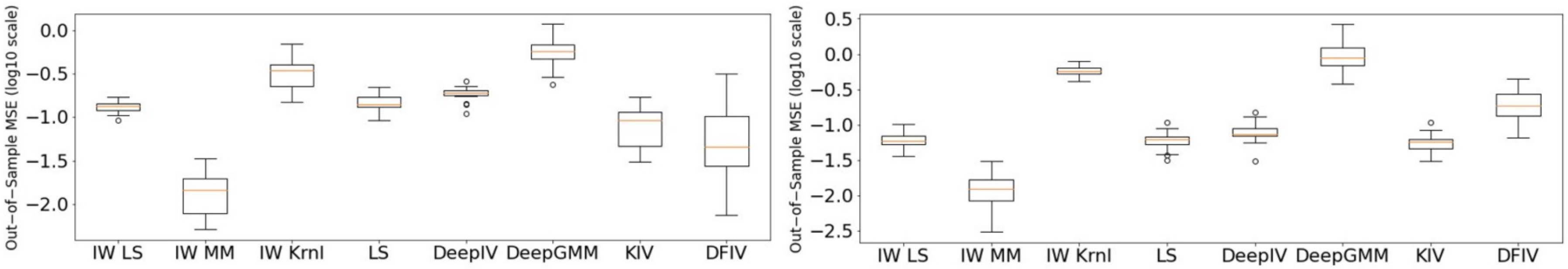
- From Dikkala et al. (2020), this theorem yields

$$\mathbb{E}\left[\left(f^*(X_i) - \hat{f}(X_i)\right)^2\right] = O_p\left(n^{-1/(2+\gamma)}\right)$$

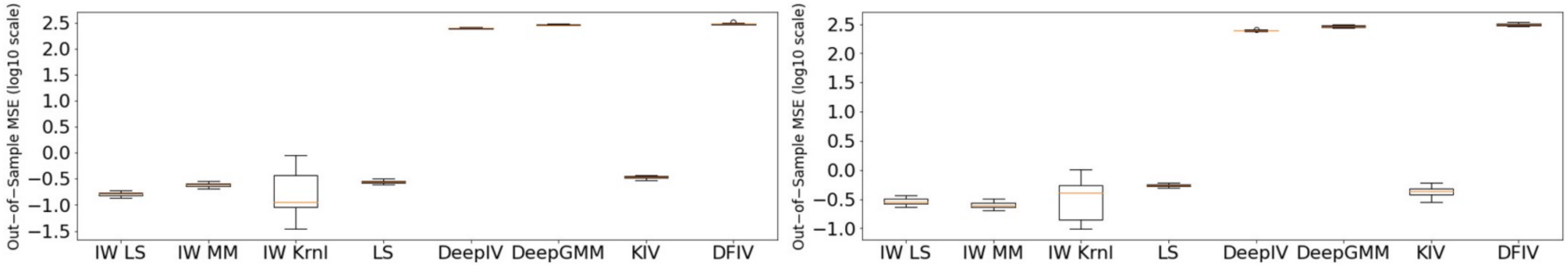
Experiments

- Investigate the empirical $\mathbb{E}\left[\left(f^*(X_i) - \hat{f}(X_i)\right)^2\right]$.

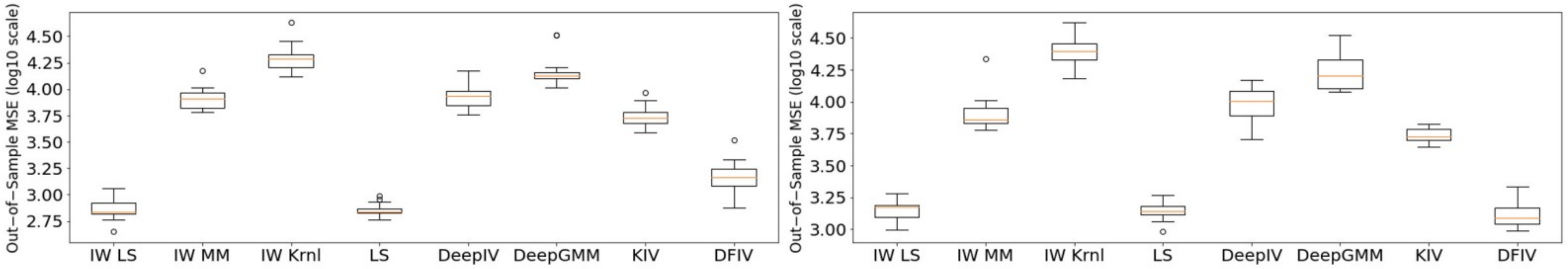
- Datasets: Newey and Powell (2003), Ai and Chen (2003), and Hartford et al, (2017).
- IW-LS: neural networks trained with the objective using a regularization to deal with a model that tends to overfit the approximated moment restrictions.
- IW-MM: neural networks trained with the objective in Section 2.
- IW-Krnl: RKHS trained with the objective in Section 2.
- We compare our proposed methods with LS (naïve least squares), DeepIV (Hartford et al, (2017)), DeepGMM (Benett et al. (2019)), KIV (Singh et al, (2019)), and DFIV (Xu et al. (2021)),



Dataset of Newey and Powell (2003)



Dataset of Ai and Chen (2003)



Dataset of Hartford et al. (2017)

Conclusion

- NPIV regression.
  - Structural (causal) model defined by the conditional moment restrictions.
  - Do not specify a specific model for the structural model.
- Our proposed method: importance weighting using the density ratio.
  - Estimate the conditional density ratio function using the least-squares.
  - Learn the nonparametric structural model by minimizing approximated conditional moment restrictions.

Reference

- Angrist, J. and Krueger, A. B. (1991), Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4):979–1014, 1991.
- Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. Econometrica, 71(5):1565–1578, 2003.
- Shimodaira, H. (2000), "Improving predictive inference under covariate shift by weighting the log-likelihood function," Journal of statistical planning and inference, 90, 227–244.
- Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. (2020), Minimax estimation of conditional moment models. In Advances in Neural Information Processing Systems, volume 33, pp. 12248–12262. Curran Associates, Inc.
- Kanamori, T., Hido, S., and Sugiyama, M. (2009), "A least-squares approach to direct importance estimation." Journal of Machine Learning Research, 10(Jul.):1391–1445.
- Hansen, B. E. (2022), Econometrics. 2022