Learning Causal Models from Conditional Moment Restrictions by Importance Weighting

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

Motivation

■ **Learning causal models under conditional moment restrictions.**
e.g., Nonparametric instrumental variable (NPIV) regression.
- Approximate conditional moment restrictions.
- Learning models satisfying the approximated moment restrictions.

■ Classical approaches:
- Use sieve regression and kernel regression for the approximation.

→ It is difficult to use them when the dimension of the data is high..

➢ **Approximation of the moment restrictions by conditional density ratio.**
- We can use machine learning algorithms to estimate the density ratio.
- It is easy to apply to high-dimensional data with complicated models.

1. Causal Models under Moment Restrictions

■ For brevity, we only consider the NPIV problem.

■ We have observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, where

$$Y_i = f^*(X_i) + \varepsilon_i$$

➢ For $i \in \{1,2,..,n\}$, the following conditional moment restrictions hold:

$$\mathbb{E}[Y_i - f^*(X_i)|Z_i] = 0$$

- Z_i: **instrumental variable**.

■ Our goal is to learn $f^*(x)$ under the moment restrictions.
- f^* is called a structural function representing some causality.

2. Learning Models with Importance Weighting

■ **Learning causal models with importance weighting.**

1. Estimate the conditional density ratio.
2. Transform the conditional conditional restrictions into unconditional ones by using the conditional density ratio.
3. Learning causal models to satisfy the approximated restrictions

■ **Definition: Conditional density ratio function**

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

➢ By using the ratio, we obtain the conditional moment restrictions as

$$\mathbb{E}\left[(Y_i - f(X_i))r^*(Y_i, X_i|z)\right] = \int (y - f(x))\frac{p(y,x|z)}{p(y,\ x)}p(y,x)dydx = \mathbb{E}[Y_i - f(X_i)|z].$$

➢ **Because we cannot know the expectation and r^*, we use**

$$\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))\hat{r}(Y_i, X_i|z)$$

- $\hat{r}$ is an estimator of the conditional density ratio.
- We learn a causal model f^* as

$$\hat{f} = \underset{f\in\mathcal{F}}{\arg\min}\frac{1}{n}\sum_{j=1}^n\left(\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))\hat{r}(Y_i, X_i|Z_j)\right)^2$$

where $\mathcal{F}$is a hypothesis class.
- We can use various models, such as neural networks, for $\mathcal{F}$.

3. Conditional Density Ratio Estimation

■ **How to estimate the conditional density ratio r^*?**
→ We can directly estimate it without density estimation.

■ Consider the following mean squared error:

$$\frac{1}{2}\mathbb{E}_{Y,X}\left[\mathbb{E}_Z\left[\left(r^*(Y_i, X_i|Z_j) - r(Y_i, X_i|Z_i)\right)^2\right]\right]$$

■ Minimizing the above MSE is equal to minimizing

$$-\mathbb{E}_{Y,X,Z}\left[r(Y_i, X_i|Z_j)\right] + \frac{1}{2}\mathbb{E}_{Y,X}\left[\mathbb{E}_Z\left[r^2(Y_i, X_i|Z_j)\right]\right]$$

- This is because

$$\frac{1}{2}\mathbb{E}_{Y,X}\left[\mathbb{E}_Z\left[\left(r^*(Y_i, X_i|Z_j) - r(Y_i, X_i|Z_i)\right)^2\right]\right]$$

$$= \frac{1}{2}\mathbb{E}_{Y,X}\left[\mathbb{E}_Z\left[r^{*2}(Y_i, X_i|Z_j) - 2r^*(Y_i, X_i|Z_j)r(Y_i, X_i|Z_i) + r^2(Y_i, X_i|Z_i)\right]\right],$$

and

$$\mathbb{E}_{Y,X}\left[\mathbb{E}_Z\left[r^*(Y_i, X_i|Z_j)r(Y_i, X_i|Z_i)\right]\right] = -\mathbb{E}_{Y,X,Z}\left[r(Y_i, X_i|Z_j)\right].$$

➢ We can solve the problem without knowing the true r^*,
- By approximating the MSE with samples, we estimate r^* as

$$\hat{r} = \underset{r\in\mathcal{R}}{\arg\min}\left\{-\frac{1}{n}\sum_{i=1}^n r(Y_i, X_i|Z_i) + \frac{1}{2}\frac{1}{n}\sum_{j=1}^n\frac{1}{n}\sum_{i=1}^n r^2(Y_i, X_i|Z_j)\right\},$$

where $\mathcal{R}$ is a hypothesis class.
- We can use various models, such as neural networks, for $\mathcal{R}$.

4. Estimation Error Analysis

■ We show the upper bound of

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

■ To show the bound, we show the upper bound of

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

■ **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^*.

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

■ **Theorem**. Under appropriate conditions, we have

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

This yields

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

5. Learning with a Flexible Model

■ Consider using a very flexible model.
= Overfitting to the approximated moment restrictions.
→ There may be multiple solutions.

■ We want to choose the one with the highest prediction ability.

■ For this purpose, we also heuristically estimate f^* as

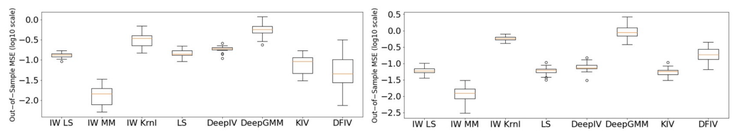
$$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 + \eta\frac{1}{n}\sum_{j=1}^n\left(\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))\hat{r}(Y_i, X_i|Z_j)\right)^2.$$

- $\eta \geq 0$: regularization coefficient.

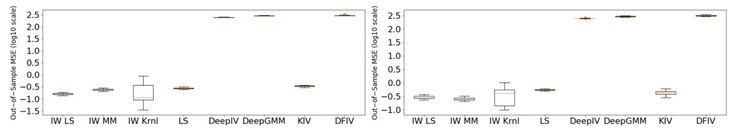
6. Experiments

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

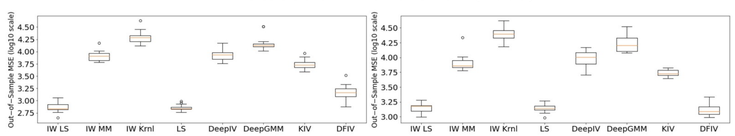
- We use three datasets from Newey and Powell (2003), Ai and Chen (2003), and Hartford et al, (2017).
- Show the box plot of the log 10 scaled squared error.
- IW-LS: neural networks trained with the objective in Section 5. Using a regularization term to deal with a flexibility of the model.
- IW-MM: 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)

References

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. Econometrica 2003.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 2003.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. ICML 2017.