Learning Causal Models from Conditional Moment Restrictions by Importance Weighting

ICLR 2022

Masahiro Kato, CyberAgent, Inc. AI Lab / 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 $X$:
  \[ Y = X^T \beta + \varepsilon, \quad \mathbb{E}[X^T \varepsilon] \neq 0. \]
- $\mathbb{E}[X^T \varepsilon] \neq 0$ implies the correlation between $\varepsilon$ and $X$.
- This situation is called endogeneity.
- In this case, an OLS estimator is not unbiased and consistent.
  - $X^T \beta$ is not the conditional mean $\mathbb{E}[Y|X]$ ($\mathbb{E}[Y|X] \neq X^T \beta$).
- This model is called structural equation model (Hansen (2022)).
Wage Equation

- The true wage equation:
  \[ \log(wage) = \beta_0 + \text{years of education} \times \beta_1 + \text{ability} \times \beta_2 + u, \]
  \[ \mathbb{E}[u|\text{years of education, ability}] = 0 \]

- We cannot observe the “ability” and estimate the following model:
  \[ \log(wage) = \beta_0 + \text{years of education} \times \beta_1 + \epsilon, \quad \epsilon = \text{ability} \times \beta_2 + u. \]

  - If “years of education” is correlated with “ability,”
    \[ \mathbb{E}[\text{“years of education”} \times \epsilon] \neq 0 \]

  \[ \rightarrow \text{We cannot consistently estimate } \beta_1 \text{ with OLS.} \]
Instrumental Variable (IV) Method

- By using IVs, we can estimate the parameter $\beta$.
- The IV is a random variable $Z$ satisfying the following conditions:
  1. Uncorrelated to the error term: $\mathbb{E}[Z^T \varepsilon] = 0$.
  2. Correlated with the endogeneous variable $X$.

\[ Z(IV) \xrightarrow{} X \ (\text{years of education}) \xrightarrow{\beta} Y(\text{wage}) \]

\[ U \ (\text{ability}) \]
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)):
  \[ Y = f^*(X) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] \neq 0. \]

- Want to estimate the structural function \( f^* \).

- \( \mathbb{E}[\varepsilon|X] \neq 0 \) → least-squares 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 function.

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 $\mathbb{E}_Z \left[ (\mathbb{E}_{Y,X}[\varepsilon|Z])^2 \right]$:

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

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

\[
\begin{align*}
    r^* &= \arg \min_r \mathbb{E}_z \left[ \mathbb{E}_{y,x} \left[ (r^*(y, x|z) - r(y, X|Z))^2 \right] \right] \\
    &= \arg \min_r \mathbb{E}_z \left[ \mathbb{E}_{y,x} \left[ (r^*(y, X|Z))^2 - 2r^*(y, X|Z)r(Y, X|Z) + r^2(Y, X|Z) \right] \right] \\
    &= \arg \min_r \mathbb{E}_z \left[ \mathbb{E}_{y,x} \left[ -2r^*(y, X|Z)r(Y, X|Z) + r^2(Y, X|Z) \right] \right] \\
    &= \arg \min_r -2\mathbb{E}_z \left[ \mathbb{E}_{y,x} \left[ r(Y, X|Z) \right] \right] + \mathbb{E}_{y,x,z} \left[ r^2(Y, X|Z) \right].
\end{align*}
\]

Ex. least-Squares Importance Fitting (LSIF, Kanamori et al. (2009))
Our goal is to show the upper bound of

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

Under appropriate conditions, we have

$$\mathbb{E}_0 \mathbb{E}[r^*(Y_1, X_1 | Z_1) - \hat{r}(Y_1, X_1 | Z_1)] = O_p(n^{-1/(2+\gamma)}),$$

where $\gamma$ corresponds to the smoothness of $r^*$. 

**Lemma:** estimation error of the conditional density ratio
Estimation Error Analysis

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

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

Theorem: mean squared error of the conditional moment restrictions

Under appropriate conditions, we have

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


- 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)).
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.


Hansen, B. E. (2022), Econometrics. 2022