Causal Inference with Instrumental Variables

Kun Kuang
Zhejiang University
https://kunkuang.github.io/
An example of decision making

• Does predictive models guide decision making?
• When should the System change algorithm from A to B?
• Is the new algorithm B better?
• Say algorithm that provides promotion or discount link to a different customers
An example of decision making

- Measure success rate (SR)

<table>
<thead>
<tr>
<th></th>
<th>Old Algorithm (A)</th>
<th>New Algorithm (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50/1000 (5%)</td>
<td>54/1000 (5.4%)</td>
<td></td>
</tr>
</tbody>
</table>

New algorithm increases overall success rate, so it is better?

<table>
<thead>
<tr>
<th></th>
<th>Old Algorithm (A)</th>
<th>New Algorithm (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-income Users</td>
<td>10/400 (2.5%)</td>
<td>4/200 (2%)</td>
</tr>
<tr>
<td>High-income Users</td>
<td>40/600 (6.6%)</td>
<td>50/800 (6.2%)</td>
</tr>
<tr>
<td>Overall</td>
<td>50/1000 (5%)</td>
<td>54/1000 (5.4%)</td>
</tr>
</tbody>
</table>

Which is better?
Decision Making with Causality

• Causal Effect Estimation is necessary for decision making!

Causal effect estimation plays an important role on decision making!
A practical definition

Definition: T causes Y if and only if changing T leads to a change in Y, keep everything else constant.

Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Two key points: changing T, keeping everything else constant

*Interventionist definition* [http://plato.stanford.edu/entries/causation-mani/]
Treatment Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Individual Treatment Effect (ITE)
  \[ ITE(i) = Y_i(T_i = 1) - Y_i(T_i = 0) \]
- Average Treatment Effect (ATE):
  \[ ATE = E[Y(T = 1) - Y(T = 0)] \]

Counterfactual problem: $Y(T = 1)$ or $Y(T = 0)$
Randomized Experiments are the “Gold Standard”

• Drawbacks of randomized experiments:
  • Cost
  • Unethical

Two key points: changing T, keeping everything else constant
Causal Inference with Observational Data

- Definition of ATE: \( ATE = E[Y(T = 1) - Y(T = 0)] \)
- In observational data, we have units with different \( T \):
  \[ E[Y(T = 1)] \text{ or } E[Y(T = 0)] \]
- Can we estimate ATE by directly comparing the average outcome between groups with \( T=1 \) and \( T=0 \)?
  - \textbf{No}, because confounders \( X \) might not be constant
- Two key points:
  - Changing \( T \) (\( T=1 \) and \( T=0 \))
  - Keeping everything else (Confounder \( X \)) constant
Causal Inference with Observational Data

• Counterfactual problem: \( Y(T = 1) \) or \( Y(T = 0) \)
• In observational data, we have units with different \( T \):
  \[
  E[Y(T = 1)] \text{ or } E[Y(T = 0)]
  \]
• Can we estimate ATE by directly comparing the average outcome between groups with \( T=1 \) and \( T=0 \)?
  • No, because confounders \( X \) might not be constant

• Two key points:

Balancing Confounders’ Distribution
Related Work

- **Matching Methods**
  - *Exactly Matching, Coarse Matching*
  - Poor performance in high dimensional settings

- **Propensity Score based Methods**
  - Propensity score \( e(X) = p(T = 1|X) \)
  - *Matching, Weighting, Doubly Robust*
  - Treat all observed variables as confounders, and ignore the non-confounders
  - Mainly designed for binary treatment

(a) Previous Causal Framework.
Related Work

• Representation Learning based Methods
  • Similar representation between treatment groups.
  • Accurate prediction on factual/counterfactual outcome

\[
\begin{align*}
\min_{h, \Phi} & \quad \frac{1}{n} \sum_{i=1}^{n} w_i \cdot L(h(\Phi(x_i), t_i), y_i) + \lambda \cdot \mathcal{R}(h) \\
& \quad + \alpha \cdot \text{IPM}_G \left( \{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1} \right), \\
\text{with} & \quad w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}, \quad \text{where} \quad u = \frac{1}{n} \sum_{i=1}^{n} t_i \\
\text{and} & \quad \mathcal{R} \text{ is a model complexity term.}
\end{align*}
\]

• Confounder differentiation, binary treatment, might ignore confounders

New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D²VD, DeR-CFR)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV, GIV)
- Continuous/Complex treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD, CRNet)
New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition ($D^2VD$, DeR-CFR)

- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV, GIV)

- Continuous/Complex treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD, CRNet)
Instrumental Variable Regression

Conditions of IV (instrumental variable)
- Relevance: $P(T|Z) \neq P(T)$
- Exclusion: $P(Y|Z,T,U) \neq P(Y|T,U)$
- Unconfounded: $Z \perp U$

2SLS:
- Stage 1: regressing $T$ on $Z$
  $$\hat{T} = \hat{g}(Z)$$
- Stage 2: regressing $Y$ on $\hat{T}$
  $$\hat{Y} = \hat{f}(\hat{T})$$

Assuming the additive separability of noise $U$ and limited to linear setting.
Non-linear Instrumental Variable Regression

Stage 1: regressing $T$ on $Z$ and $X$
\[ \hat{T} = \hat{g}(Z, X) \]

Stage 2: regressing $Y$ on $\hat{T}$ and $X$
\[ \hat{Y} = \hat{f}(\hat{T}, X) \]

Stage 1 regression brings confounding bias in stage 2
CB-IV (Confounder Balanced IV regression):

Stage 1 (Treatment regression): regressing $T$ on $Z$ and $X$

$$\hat{T} = \hat{g}(Z, X)$$

Confounder balancing: learning a balanced confounder representation $\phi(X)$ such that $\hat{T} \perp \phi(X)$

Stage 2 (Outcome regression): regressing $Y$ on $\hat{T}$ and $\phi(X)$

$$\hat{Y} = \hat{f}(\hat{T}, \phi(X))$$

### Confounder Balanced Instrumental Variable Regression

**Table 2**: The bias (mean ± std) of ATE estimation on real-world data ($Data - m_Z - m_X - m_U$)

<table>
<thead>
<tr>
<th>Method</th>
<th>IHDP-2-6-0</th>
<th>IHDP-2-4-2</th>
<th>Twins-5-8-0</th>
<th>Twins-5-5-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepIV-LOG</td>
<td>2.8736 ± 0.0577</td>
<td>2.6227 ± 0.0651</td>
<td>0.0135 ± 0.0215</td>
<td>0.0237 ± 0.0111</td>
</tr>
<tr>
<td>DeepIV-GMM</td>
<td>3.7760 ± 0.0316</td>
<td>3.7396 ± 0.0402</td>
<td>0.0194 ± 0.0047</td>
<td>0.0221 ± 0.0041</td>
</tr>
<tr>
<td>OneSV</td>
<td>1.7249 ± 0.3752</td>
<td>1.7411 ± 0.3422</td>
<td>0.0083 ± 0.0191</td>
<td>0.0080 ± 0.0167</td>
</tr>
<tr>
<td>DFIV</td>
<td>3.5543 ± 0.0891</td>
<td>3.6218 ± 0.1038</td>
<td>0.0268 ± 0.0005</td>
<td>0.0265 ± 0.0003</td>
</tr>
<tr>
<td>DFL</td>
<td>3.2018 ± 0.0496</td>
<td>3.1991 ± 0.0374</td>
<td>0.0624 ± 0.0586</td>
<td>0.0847 ± 0.0049</td>
</tr>
<tr>
<td>DirectRep</td>
<td>0.0675 ± 0.0562</td>
<td>0.4600 ± 0.0711</td>
<td>0.0167 ± 0.0171</td>
<td>0.0193 ± 0.0251</td>
</tr>
<tr>
<td>CFR</td>
<td>0.0854 ± 0.0579</td>
<td>0.4826 ± 0.0642</td>
<td>0.0115 ± 0.0167</td>
<td>0.0223 ± 0.0176</td>
</tr>
<tr>
<td>DRCFR</td>
<td>0.0553 ± 0.0644</td>
<td>0.4336 ± 0.0692</td>
<td>0.0114 ± 0.0221</td>
<td>0.0118 ± 0.0173</td>
</tr>
<tr>
<td>CB-IV</td>
<td>0.0117 ± 0.3882</td>
<td>0.1601 ± 0.2499</td>
<td>0.0067 ± 0.0271</td>
<td>0.0014 ± 0.0249</td>
</tr>
</tbody>
</table>

**Out-of-Sample**

<table>
<thead>
<tr>
<th>Method</th>
<th>IHDP-2-6-0</th>
<th>IHDP-2-4-2</th>
<th>Twins-5-8-0</th>
<th>Twins-5-5-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepIV-LOG</td>
<td>2.8760 ± 0.0553</td>
<td>2.6226 ± 0.0692</td>
<td>0.0140 ± 0.0208</td>
<td>0.0238 ± 0.0111</td>
</tr>
<tr>
<td>DRCFR</td>
<td>0.0450 ± 0.0953</td>
<td>0.4321 ± 0.0673</td>
<td>0.0113 ± 0.0219</td>
<td>0.0118 ± 0.0174</td>
</tr>
<tr>
<td>CB-IV</td>
<td>0.0150 ± 0.3927</td>
<td>0.1578 ± 0.2540</td>
<td>0.0065 ± 0.0270</td>
<td>0.0015 ± 0.0247</td>
</tr>
</tbody>
</table>

**IV regression based methods**

**Confounder balancing based methods**

*Wu A, Kuang K, Li B, et al. Instrumental Variable Regression with Confounder Balancing, ICML 2022*

Require a well predefined valid IV
No-predefined IV: Confounded IVs

Confounded IVs:
- Violation on Unconfounded Instrument: $Z_i$ correlates to $E$ conditioning on $X$
- Often happens in real cases and leads to failure of all IV methods.

Our setting:
- $\{Z_i\}_{i=1}^m$ represents candidates for IV
- A subset of $\{Z_i\}_{i=1}^m$ are valid, while others are confounded IVs

Our Goal:
- Estimating Individual Causal Effect

Conditional Independence Criteria:

- Generating $E'$ such that $Y \perp (Z_1, Z_2, ..., Z_m) \mid E', T, X$
- $E'$ captures the info of confounders from $E$ rather than recovers $E$
CVAE-IV: Constructing substitute for confounders

- Conditional Independence Criteria:
  
  - Generating $E'$ such that $Y \perp (Z_1, Z_2, \ldots, Z_m) \mid E', T, X$
  
  - $E'$ captures $E$ rather than recovers $E$

- Constructing the conditional variational autoencoder model (CVAE)
  
  - Variational inference:
    
    $\log p(Y, Z \mid T, X) \\ \geq \mathbb{E}[\log p_\theta(Y, Z \mid E', T, X)] - D_K L(q_\phi(E' \mid Y, Z, X, T) \parallel p(E' \mid T, X))$

  - Separating construction of outcome from that of IVs:
    
    $\log p_\theta(Y, Z \mid E', T, X) = \log p_\theta(Y \mid E', T, X) + \log p_\theta(Z \mid E', T, X)$

- Constructing the IVs using Cholesky approximation:
  
  - Isotropic Gaussian eliminates the dependence among $\{Z_i\}_{i=1}^m$

  - Estimating covariance matrix of $\{Z_i\}_{i=1}^m$ using Cholesky decomposition
    
    $L_{Chol} = \log |\Sigma_Z(E')| + (Z - \mu(E'))^T \Sigma_Z(E')^{-1}(Z - \mu(E'))$

  - Reconstruction of $\{Z_i\}_{i=1}^m$
    
    $L_{Chol} = -2 \sum_{i=1}^m (\log L(E_{i\mu}) + (\bar{Z} - \mu(E'))^T L(E')L(E')^{-1}(\bar{Z} - \mu(E'))$

- Constructing the IVs using Cholesky approximation:
  
    $L_{rec} = \log p_\theta(Y \mid E', T, X) \approx \frac{1}{\sigma_{Y'}} \| Y - Y'(E', T, X) \|^2_2$
<table>
<thead>
<tr>
<th>Scenario</th>
<th>Function</th>
<th>Dim</th>
<th>DirectNN</th>
<th>2SLS-Ploy</th>
<th>KernelIV</th>
<th>DeepIV</th>
<th>CEVAE</th>
<th>ModeIV</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁ :</td>
<td>Linear</td>
<td>Low</td>
<td>2.091</td>
<td>55.437</td>
<td>8.565</td>
<td>1.977</td>
<td>9.959</td>
<td>3.120</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>2.069</td>
<td>56.476</td>
<td>9.220</td>
<td>2.261</td>
<td>10.283</td>
<td>4.426</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>Abs</td>
<td>Low</td>
<td>1.874</td>
<td>44.144</td>
<td>6.893</td>
<td>2.426</td>
<td>8.481</td>
<td>1.988</td>
<td>0.697</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1.671</td>
<td>41.731</td>
<td>7.998</td>
<td>1.675</td>
<td>9.921</td>
<td>2.088</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>Square</td>
<td>Low</td>
<td>1.414</td>
<td>52.872</td>
<td>8.521</td>
<td>1.163</td>
<td>6.569</td>
<td>2.423</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1.602</td>
<td>41.731</td>
<td>9.988</td>
<td>1.545</td>
<td>7.039</td>
<td>2.018</td>
<td>0.791</td>
</tr>
<tr>
<td>S₂ :</td>
<td>Linear</td>
<td>Low</td>
<td>1.788</td>
<td>45.287</td>
<td>8.922</td>
<td>1.910</td>
<td>7.621</td>
<td>3.314</td>
<td>0.706</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>2.152</td>
<td>56.484</td>
<td>9.215</td>
<td>1.788</td>
<td>8.109</td>
<td>2.040</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td>Abs</td>
<td>Low</td>
<td>1.595</td>
<td>34.214</td>
<td>5.864</td>
<td>1.224</td>
<td>7.178</td>
<td>2.064</td>
<td>0.562</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.836</td>
<td>42.181</td>
<td>4.030</td>
<td>0.835</td>
<td>8.015</td>
<td>2.134</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>Square</td>
<td>Low</td>
<td>1.168</td>
<td>41.108</td>
<td>5.749</td>
<td>1.064</td>
<td>10.385</td>
<td>1.565</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1.650</td>
<td>51.129</td>
<td>6.030</td>
<td>1.568</td>
<td>9.015</td>
<td>2.031</td>
<td>0.257</td>
</tr>
<tr>
<td>S₃ :</td>
<td>Linear</td>
<td>Low</td>
<td>1.650</td>
<td>41.009</td>
<td>6.617</td>
<td>2.023</td>
<td>7.234</td>
<td>3.638</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1.821</td>
<td>41.374</td>
<td>7.656</td>
<td>1.729</td>
<td>7.203</td>
<td>4.134</td>
<td>0.608</td>
</tr>
<tr>
<td></td>
<td>Abs</td>
<td>Low</td>
<td>2.095</td>
<td>44.471</td>
<td>4.867</td>
<td>1.148</td>
<td>9.789</td>
<td>1.916</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>1.590</td>
<td>41.132</td>
<td>6.330</td>
<td>1.484</td>
<td>7.293</td>
<td>1.972</td>
<td>0.504</td>
</tr>
<tr>
<td></td>
<td>Square</td>
<td>Low</td>
<td>1.111</td>
<td>41.199</td>
<td>7.826</td>
<td>1.019</td>
<td>8.362</td>
<td>1.580</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>2.179</td>
<td>41.915</td>
<td>6.407</td>
<td>1.442</td>
<td>8.825</td>
<td>1.709</td>
<td>0.941</td>
</tr>
</tbody>
</table>

---

Haotian Wang, Wenjing Yang, Longqi Yang, Anpeng Wu, Liyang Xu, Jing Ren, Fei Wu, Kun Kuang.
Estimating Individualized Causal Effect with Confounded Instruments. KDD 2022
AutoIV: Counterfactual Learning with Unobserved Confounders via Automatically generating IVs

Conditions of IV
- Relevance: \( P(T|Z) \neq P(T) \)
- Exclusion: \( P(Y|Z, T, C) \neq P(Y|T, C) \)
- Unconfounded: \( Z \perp C \)

Mutual Information
Representation Learning

But exclusion might not be satisfied

AutoIV: Counterfactual Learning with Unobserved Confounders via Automatically generating IVs


Figure 2: Response function prediction in low-dimensional scenarios.
IVs in Causal Inference and Machine Learning

Instrumental Variables in Causal Inference and Machine Learning: A Survey

Anpeng Wu, Kun Kuang*, Ruoxuan Xiong, Fei Wu, Senior Member, IEEE

Abstract—Causal inference, which refers to the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect, is crucial for stable learning and decision making by understanding the mechanism underlying the data. How to precisely and unbiasedly estimate the treatment effect from observational data with unobserved confounders is becoming an appealing research direction in both causal inference and machine learning communities. Instrumental Variables (IV) plays a critical role to draw causal inference from the settings where the treatment of interest cannot be randomly assigned and even with unobserved confounders. In recent years, IV methods have attracted considerable attention in the literature of both causal inference and machine learning, and various IV-based methods have sprung up. This paper serves as the first effort to systematically and comprehensively introduce and discuss the IV methods and their applications in both causal inference and machine learning. Firstly, we provide the formal definition of IVs and discuss the identification problem of IV regression methods under different assumptions. Secondly, we categorize the existing work on IV methods into three groups according to the focus on the proposed method, including two-stage least squares based IV methods, control function based IV methods and evaluation on IVs. For each category, the main advances of both traditional statistical methods and machine learning enhanced methods are presented and discussed. Then, we introduce a variety of applications of IV methods in real world scenarios and summarize the available datasets and algorithms. Finally, we summarize the whole literature, discuss the open problems and suggest promising future research directions for IV method and its application. We also develop a toolkit of IVs methods reviewed in this survey at https://github.com/causal-machine-learning-lab/mtlv.
Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Fei Wu, Instrumental Variables in Causal Inference and Machine Learning, working paper.
Thank You!

Kun Kuang
kunkuang@zju.edu.cn
Homepage: https://kunkuang.github.io/