# The Supplementary of "Learning Conditional Instrumental Variable Representation for Causal Effect Estimation"

Debo Cheng[1*](✉), Ziqi Xu[1*], Jiuyong Li[1], Lin Liu[1],
Thuc Duy Le[1], and Jixue Liu[1]

[1]University of South Australia, Adelaide, Australia
{Debo.Cheng,Ziqi.Xu}@mymail.unisa.edu.au,
{Jiuyong.Li,Lin.Liu,Thuc.Le,Jixue.Liu}@unisa.edu.au

In this Appendix, we provide additional graphical notations and definitions, details of synthetic and real-world datasets, and more experimental results.

## A Preliminaries

*Paths.* In a graph $\mathcal{G}$, a path $\pi$ between $V_1$ and $V_p$ consists of a sequence of distinct nodes $\langle V_1, \ldots, V_p \rangle$ with every pair of successive nodes being adjacent. A node $V$ lies on the path $\pi$ if $V$ belongs to the sequence $\langle V_1, \ldots, V_p \rangle$. A path $\pi$ is a directed or causal path if all edges along it are directed such as $V_1 \to \ldots \to V_p$.

*Ancestral relationships.* In a DAG $\mathcal{G}$, $V_i$ is a parent of $V_j$ (and $V_j$ is a child of $V_i$) if $V_i \to V_j$ appears in this graph. In a directed path $\pi$, $V_i$ is an ancestor of $V_j$ and $V_j$ is a descendant of $V_i$ if all arrows along $\pi$ point to $V_j$.

In graphical causal modelling, the assumptions of Markov property, faithfulness and causal sufficiency are often involved to discuss the relationship between the causal graph and the distribution of the data.

**Definition 1 (Markov property [6]).** *Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and the joint probability distribution of $\mathbf{V}$ ($prob(\mathbf{V})$), $\mathcal{G}$ satisfies the Markov property if for $\forall V_i \in \mathbf{V}$, $V_i$ is probabilistically independent of all of its non-descendants, given the set of parents $V_i$.*

**Definition 2 (Faithfulness [7]).** *A DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is faithful to a joint distribution $prob(\mathbf{V})$ over the set of variables $\mathbf{V}$ if and only if every independence present in $prob(\mathbf{V})$ is entailed by $\mathcal{G}$ and satisfies the Markov property. A joint distribution $prob(\mathbf{V})$ over the set of variables $\mathbf{V}$ is faithful to the DAG $\mathcal{G}$ if and only if the DAG $\mathcal{G}$ is faithful to the joint distribution $prob(\mathbf{V})$.*

**Definition 3 (Causal sufficiency [7]).** *A given dataset satisfies causal sufficiency if in the dataset for every pair of observed variables, all their common causes are observed.*

* These authors contributed equally.

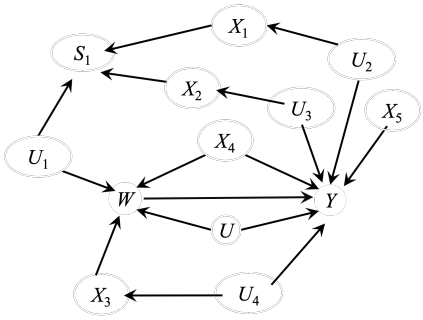


**Fig. 1.** The true causal DAG with a latent confounder $U$ between $W$ and $Y$ is utilised to generate the synthetic datasets in Simulation Study. $\mathbf{X} = \{S, X_1, X_2, X_3, X_4, X_5\}$ are pretreatment variables, and $\mathbf{U} = \{U, U_1, U_2, U_3, U_4\}$ are five latent confounders. Note that $S$ is a CIV conditioning on $\{X_1, X_2\}$.

In a DAG, d-separation is a graphical criterion that enables the identification of conditional independence between variables entailed in the DAG when the Markov property, faithfulness and causal sufficiency are satisfied [6, 7].

**Definition 4 (d-separation [6]).** *A path $\pi$ in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is said to be d-separated (or blocked) by a set of nodes $\mathbf{Z}$ if and only if (i) $\pi$ contains a chain $V_i \to V_k \to V_j$ or a fork $V_i \leftarrow V_k \to V_j$ such that the middle node $V_k$ is in $\mathbf{Z}$, or (ii) $\pi$ contains a collider $V_k$ such that $V_k$ is not in $\mathbf{Z}$ and no descendant of $V_k$ is in $\mathbf{Z}$. A set $\mathbf{Z}$ is said to d-separate $V_i$ from $V_j$ ($V_i \perp\!\!\!\perp_d V_j \mid \mathbf{Z}$) if and only if $\mathbf{Z}$ blocks every path between $V_i$ to $V_j$. otherwise they are said to be d-connected by $\mathbf{Z}$, denoted as $V_i \not\perp\!\!\!\perp_d V_j \mid \mathbf{Z}$.*

## B Experiments

### B.1 Simulation Study

The simulated datasets are generated from the true DAG in Fig. 1, and the specifications of the data generation are as follows: $U, U_1, U_2, U_3, U_4 \sim N(0, 1)$ and $\epsilon_{1,2,3,s} \sim N(0, 0.5)$, where $N(,)$ denotes the normal distribution. $X_1 \sim N(0, 1) + 0.5 * U_2 + \epsilon_1$, $X_2 \sim N(0, 1) + 0.5 * U_3 + \epsilon_2$, $X_3 \sim N(0, 1) + 0.5 * U_4 + \epsilon_3$, $S \sim N(0, 1) + 2 * U_1 + 1.5 * X_1 + 1.5 * X_2 + \epsilon_s$, $X_4 \sim N(1, 1)$, and $X_5 \sim N(3, 1)$.

The treatment assignment $W$ is generated from $n$ ($n$ denotes the sample size) Bernoulli trials by using the assignment probability $P(W = 1 \mid U, U_1, X_3) = [1 + exp\{2 - 1 * U - 1 * U_1 - 1 * X_3 - 1 * X_4\}]$. The potential outcome is generated from $Y_W = 2 + 2 * W + 2 * U + 2 * U_3 + 2 * U_4 + 1 * X_4 + 1 * X_5 + \epsilon_W$ where $\epsilon_W \sim N(0, 1)$. Note that true $ACE$ is fixed to 2 on all synthetic datasets.

The experimental results of $\varepsilon_{ACE}$ and $\sqrt{\varepsilon_{PEHE}}$ on within-samples are reported in Tables 1 and 2, respectively.

*Results.* Tables 1 and 2 support the same conclusion drawn in the main text.

**Table 1.** The within-sample absolute error $\varepsilon_{ACE}$ (mean±std) over 30 synthetic datasets. The best results are highlighted in boldface and the runner-up results are underlined.

| Samples | | 2k | 6k | 10k | 20k |
|---|---|---|---|---|---|
| Estimators | | $\varepsilon_{ACE}$ | $\varepsilon_{ACE}$ | $\varepsilon_{ACE}$ | $\varepsilon_{ACE}$ |
| ML-based | DML | 5.507±0.387 | 5.624±0.182 | 5.619±0.122 | 5.633±0.096 |
| | DRL | 5.746±0.404 | 5.833±0.186 | 5.825±0.156 | 5.860±0.106 |
| tree-based | BART | 3.890±0.368 | 3.999±0.156 | 4.014±0.152 | 4.046±0.106 |
| | CF | 3.218±0.325 | 3.255±0.140 | 3.277±0.131 | 3.306±0.077 |
| VAE-based | CEVAE | 5.558 ±0.407 | 5.698±0.194 | 5.640±0.172 | 5.706±0.112 |
| | TEDVAE | 5.671±0.399 | 5.583±0.194 | 5.644±0.167 | 5.674±0.100 |
| IV-based | OrthIV | 2.212±1.260 | 1.952±0.585 | 1.792±0.607 | 1.974±0.419 |
| | DMLIV | 2.170±1.189 | 1.888±0.572 | 1.790±0.626 | 1.971±0.432 |
| | DeepIV | **0.352±0.180** | 0.632±0.245 | 0.727±0.315 | 0.757±0.354 |
| | CFIVR | 1.217±0.924 | **0.514±0.369** | **0.552±0.461** | **0.416±0.296** |
| DRVAE.CIV | | 0.612±0.090 | 0.588±0.055 | 0.536±0.085 | 0.512±0.091 |

### B.2 Experiments on Three Real-World Datasets

*SchoolingReturns.* The dataset is from the national longitudinal survey of youth (NLSY), a well-known dataset of US young employees, aged range from 24 to 34 [2]. The treatment is the education of employees, and the outcome is raw wages in 1976 (in cents per hour). The data contains 3,010 individuals and 19 covariates. The covariates include experience (Years of labour market experience), ethnicity, resident information of an individual, age, nearcollege (whether an individual grew up near a 4-year college?), marital status, Father's educational attainment, Mother's educational attainment, and so on. A goal of the studies on this dataset is to investigate the causal effect of education on earnings. Card [2] used geographical proximity to a college, i.e. the covariate *nearcollege* as an instrument variable. We take $ACE = 0.1329$ with 95% conditional interval (0.0484, 0.2175) from [8] as the reference causal effect.

*Cattaneo.* The Cattaneo ( [3]) is usually used to study the ACE of maternal smoking status during pregnancy ($W$) on a baby's birth weight (in grams). Cattaneo2 consists of the birth weights of 4,642 singleton births in Pennsylvania, USA ( [1, 3]). Cattaneo contains 864 smoking mothers ($W$=1) and 3,778 non-smoking mothers ($W$=0). The dataset contains several covariates: mother's age, mother's marital status, an indicator for the previous infant where the newborn died, mother's race, mother's education, father's education, number of prenatal care visits, months since last birth, an indicator of firstborn infant and indicator of alcohol consumption during pregnancy. The authors ( [1]) found a strong negative effect of maternal smoking on the weights of babies, that is, about $200g$ to $250g$ lower for a baby with a mother smoking during pregnancy than for a baby without by statistical analysis on all covariates.

http://www.stata-press.com/data/r13/cattaneo2.dta

**Table 2.** The within-sample $\sqrt{\varepsilon_{PEHE}}$ (mean±std) over 30 synthetic datasets. The lowest $\sqrt{\varepsilon_{PEHE}}$ are highlighted in boldface and the runner-up results are underlined.

| Samples | | 2k | 6k | 10k | 20k |
|---|---|---|---|---|---|
| Estimators | | $\sqrt{\varepsilon_{PEHE}}$ | $\sqrt{\varepsilon_{PEHE}}$ | $\sqrt{\varepsilon_{PEHE}}$ | $\sqrt{\varepsilon_{PEHE}}$ |
| ML-based | DML | 5.455±0.353 | 5.596±0.174 | 5.587±0.115 | 5.588±0.090 |
| | DRL | 5.671±0.370 | 5.786±0.182 | 5.774±0.144 | 5.794±0.101 |
| tree-based | BART | 4.185±344 | 4.227±0.149 | 4.234±0.146 | 4.253±0.106 |
| | CF | 3.475±0.301 | 3.504±0.129 | 3.522±0.121 | 3.547±0.072 |
| VAE-based | CEVAE | 6.061±0.352 | 6.149±0.178 | 6.115±0.149 | 6.173±0.101 |
| | TEDVAE | 6.076±0.337 | 6.149±0.175 | 6.119±0.148 | 6.147±0.091 |
| IV-based | OrthIV | 3.050±0.700 | 2.804±0.303 | 2.736±0.255 | 2.784±0.213 |
| | DMLIV | 3.009±0.664 | 2.772±0.280 | 2.738±0.268 | 2.784±0.216 |
| | DeepIV | **2.403±0.036** | **2.408±0.038** | 2.418±0.062 | 2.425±0.065 |
| | CFIVR | 3.048±0.649 | 2.457±0.252 | **2.432±0.355** | **2.328±0.144** |
| DRVAE.CIV | | 2.460±0.041 | 2.454±0.029 | 2.449±0.027 | 2.448±0.015 |

*Right heart catheterization (RHC).* RHC is a real-world dataset obtained from an observational study regarding a diagnostic procedure for the management of critically ill patients [4]. The RHC dataset can be downloaded from the **R** package *Hmisc*. RHC contains information on hospitalised adult patients from five medical centres in the USA. These hospitalised adult patients participated in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). Treatment $W$ indicates whether a patient received an RHC within 24 hours of admission. The outcome $Y$ is whether a patient died at any time up to 180 days after admission. The original RHC dataset has 5,735 samples with 73 covariates. We preprocess the original data, as suggested by Loh et al. [5], and the final dataset contains 2,707 samples with 72 covariates. Note that the empirical conclusion is that applying RHC leads to higher mortality within 180 days than not applying RHC [4].

## References

1. Almond, D., Chay, K.Y., Lee, D.S.: The costs of low birth weight. The Quarterly Journal of Economics **120**(3), 1031–1083 (2005)
2. Card, D.: Using geographic variation in college proximity to estimate the return to schooling (1993)
3. Cattaneo, M.D.: Efficient semiparametric estimation of multi-valued treatment effects under ignorability. Journal of Econometrics **155**(2), 138–154 (2010)
4. Connors, A.F., Speroff, T., et al.: The effectiveness of right heart catheterization in the initial care of critically iii patients. Journal of the American Medical Association **276**(11), 889–897 (1996)
5. Loh, W.W., Vansteelandt, S.: Confounder selection strategies targeting stable treatment effect estimators. Statistics in Medicine **40**(3), 607–630 (2021)
6. Pearl, J.: Causality. Cambridge university press (2009)

https://CRAN.R-project.org/package=Hmisc

7. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
8. Verbeek, M.: A guide to modern econometrics. John Wiley & Sons (2008)