Causal Discovery and Causal Inference

A review from the perspective of causal effect estimation and time series

Zhen Zhou

Southeast University

School of Transportation

zzhou602@seu.edu.cn

CONTENTS

1. The Definition of Causal Effect

Why we need causal inference? Imagining an experiment to assess the causal effect between smoking and lung cancer. Let T denotes the treatment variable and Y denotes the outcome variable, in this experiment, Y is the case of getting lung cancer. $T=1$ means smoking and $T=0$ means no smoking. Taking an individual as an example, the individual treatment effect (ITE) for this person is:

$$ITE_i = Y_i(T=1) - Y_i(T=0) \tag{1.1}$$

This represents the treatment effect of smoking for a single individual. But for a group, we need to use average treatment effect (ATE) for measurement, which can be denoted as:

$$ATE = E[Y(T=1) - Y(T=0)] = E[Y(T=1)] - E[Y(T=0)] \tag{1.2}$$

Here is the problem, that is, we will never know the actual value of $Y(T=1)$ and $Y(T=0)$, we can only observe Y after the selection of T. This is REALLY important, cause Y will always be there, and will never be changed by changing T, this is one of the most important assumptions in causal inference. Under this assumption, we can have:

$$
\begin{aligned}
ATE &= E[Y(T=1)] - E[Y(T=0)] \\
&= E[Y(T=1) \mid T=1] - E[Y(T=0) \mid T=0] \\
&= E[Y \mid T=1] - E[Y \mid T=0]
\end{aligned} \tag{1.3}
$$

From the perspective of Bayes' theorem, equation (1.3) holds only when T is independent of Y, which can also be denoted as $Y \perp T$.

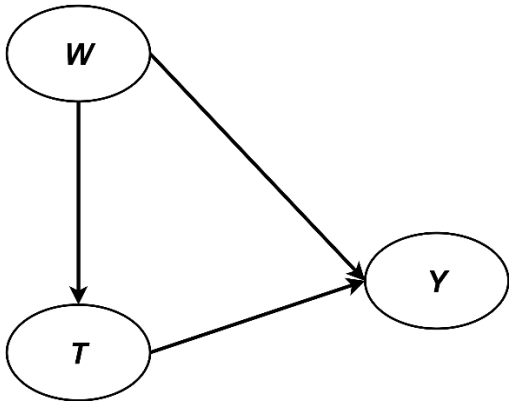

Figure 1 Basic causal structure with confounding

However, in real life, T is usually not independent of Y. For example, in Figure 1, confounding occurs because covariate W is associated with both the treatment variable T and outcome variable Y. When confounding occurs, it can be difficult to determine whether the observed relationship between the treatment T and outcome variable Y is truly causal, or whether it is due to the influence of the confounding variable.

The conditional unconfoundedness assumption can be made to ensure the independence of Y and T given covariates W, which is $Y \perp T \mid W$. By using conditional unconfoundedness assumption, we can introduce conditional average treatment effect (CATE) to causal effect estimation, which is:

$$
\begin{aligned}
CATE &= E[Y(T=1) - Y(T=0) \mid W] \\
&= E[Y(T=1) \mid T=1, W] - E[Y(T=0) \mid T=0, W] \\
&= E[Y \mid T=1, W] - E[Y \mid T=0, W]
\end{aligned} \tag{1.4}
$$

According to iterated expectations:

$$
E_X(X) = E_y[E_X(X \mid y)] \tag{1.5}
$$

We can finally have:

$$
\begin{aligned}
ATE &= E[Y(T=1) - Y(T=0)] \\
&= E_W[E(Y(T=1) - Y(T=0) \mid W)] \\
&= E_W[E(Y \mid T=1, W) - E(Y \mid T=0, W)]
\end{aligned} \tag{1.6}
$$

It should be noted that ITE is different from CATE. ITE focuses on the treatment effect for an individual while CATE focuses on ATE for a subgroup *w*. Hence, CATE can be generally called heterogeneous treatment effect.

2. Uplift Model

Why we need CATE? The answer is heterogeneous. When populations are heterogeneous, different subgroups have different ATE, which will Leads to false inferences if we still use ATE. Uplift modeling (Wilson Pok, 2020) refers to models that predict the incremental change in behavior caused by an intervention. Figure 2 provides four subgroups in a population in lift model.

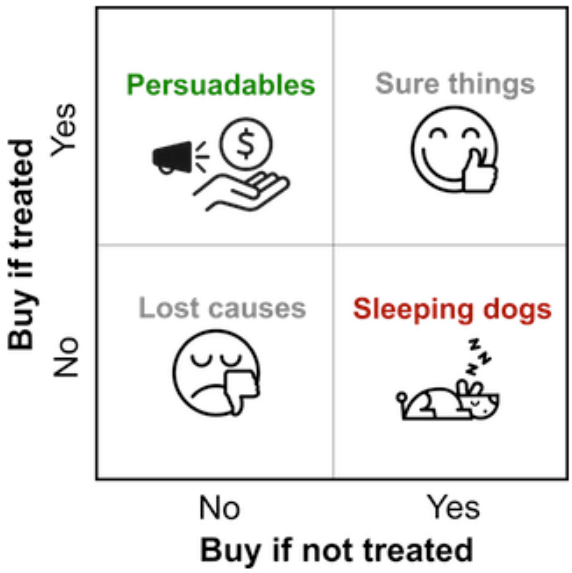

Figure 2 An example of uplift model (Wilson Pok, 2020)

The definition of these four subgroups can be identified as:

- *"Sure things" are customers who were always going to buy the product, regardless of whether or not they received a marketing intervention (zero uplift).*
- *Persuadables are customers who became more likely to buy the product because they received a marketing intervention (positive uplift).*
- *Sleeping dogs are customers who became less likely to buy the product because they received a marketing intervention (negative uplift).*

- *Lost causes are customers who were never going to buy the product, regardless of whether or not they received a marketing intervention (zero uplift).*

Uplift model only focuses on the subgroup "Persuadables", because incremental change results from the treatment of receiving a market intervention. In causal inference, uplift model is proposed to estimate heterogeneous treatment effect in different subgroups, which is CATE in equation (1.4).

3. Estimation of Heterogeneous Treatment Effect for Continuous Variable

In order to generalize to continuous variables, the original average binary treatment effect problem can be converted to the problem of estimating heterogeneous continuous treatment effect (can also be named as CATE), which can be formulated as:

$$HCTE = E\left(Y \mid W, T = T_1\right) - E\left(Y \mid W, T = T_0\right) \tag{3.1}$$

HTE only focuses on the treatment effect of subgroups. Regression, forest etc. method can be applied to estimate continuous treatment effect. Below is a simple example about the estimation of HTE by using regression method.

We use Y as outcome and T as the treatment, as well as the confounders W as predictors. The estimation function can be denoted as:

$$E\left(Y \mid W, T\right) = \beta_0 + \beta_t T + \beta_w W = \beta_0 + \beta_t T + \sum_{i=1}^{n} \beta_i W_i \tag{3.2}$$

We define a new observation c^j , which mean this observation just change treatment variable and confounders W remain unchanged

$$\begin{aligned} HTE &= E\left(Y \mid W, T = T_1\right) - E\left(Y \mid W, T = T_0\right) \\ &= \beta_0 + \beta_t T_1 + \beta_w W - \beta_0 + \beta_t T_0 + \beta_w W \\ &= \beta_t (T_1 - T_0) \end{aligned} \tag{3.3}$$

If treatment variable is binary, which means $T_0 = 1$ and $T_1 = 0$

$$HTE = \beta_t (T_1 - T_0) = \beta_t \tag{3.4}$$

In equation (3.4), the average treatment effect is β_t , which can be estimated by using

regression and the estimation term can be denoted as β_t .

If treatment variable T is a continuous variable, and covariates W hold constant. let $T_1 = T_0 + 1$, equation (3.4) can be re-formulated as:

$$\beta_t = \beta_t(T_0 + 1 - T_0) \tag{3.5}$$

This would imply that the treatment effect of increasing T by one is β_t . However, the example above is inaccurate, it is really hard to estimate HTE because of the high dimensional covariates W, so we need to use some techniques to estimate HTE more accurately.

3.1 Generalized propensity score (GPS) method

Recall equation (1.6), the assumption $Y \perp T \mid W$ need to be satisfied, which means that the treatment variables T are independent of the potential outcomes variable Y given covariates W. Why we need propensity score (PS)? Firstly, the introduction of propensity score is to reduce the impact of confounding variables W on the estimated treatment effect. Secondly, W is a multivariate, which is very hard to be controlled to keep the independence between Y and T. Hence, PS is proposed to replace W and ensure the unconfoundedness of the model.

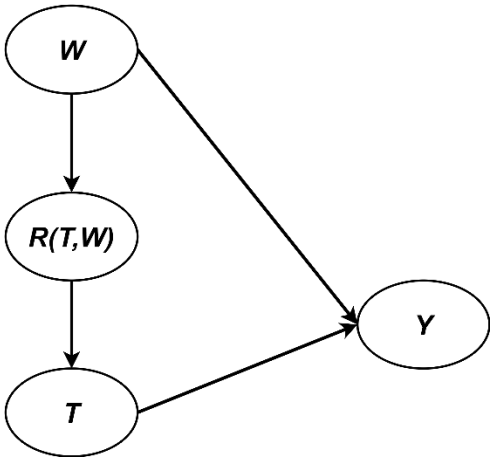

Figure 3 PS blocks the backdoor path that W blocks

The basic structure of causal relationship of *Y, T, W* and $R(T,W)$ can be found in Figure 3. The propensity score $R(T,W)$ is the probability of receiving a treatment T given a set of observed covariates W. According to strongly ignorable treatment assignment proposed by Rosenbaum and Rubin (ROSENBAUM & RUBIN, 1983), if all confounders have been observed, and

unconfoundedness is satisfied, the original assumption can be converted to a weak assumption ("weak unconfoundedness"), which can be denoted as $Y \perp \mathrm{T} \mid R(T,W)$.

$$Y \perp T \mid W \rightarrow Y \perp W \mid R(T,W) \tag{3.6}$$

After the construction of $Y \perp T \mid R(T,W)$, we just need to control a single variable $R(T,W)$ to ensure the independence between treatment variable and outcome variable.

It is really important to figure out why we need to "ensure the independence between treatment variable and outcome variable", this related to the basic definition of causal inference, and has been illustrated in Section 1.

We can introduce propensity score into the estimation of treatment effect. In binary case, the propensity score means the probability of adding treatment given covariates *W*, which can be denoted as:

$$R_i(T,W) = P(T_i = 1 \mid W_i) \tag{3.7}$$

It is easy to calculate PS if treatment variable is binary, but when treatment variable is continuous, it will need a technique to calculate PS.

In continuous case, we use generalized propensity score (GPS) (Hirano & Imbens, 2004) to estimate treatment effect for continuous variable. In order to remove confounding factors, the original function can be formulated as

$$E\left(Y \mid W,T\right) = \beta_0 + \beta_t F(T) + \beta_w R(T,W) \tag{3.8}$$

Where $R(T,W)$ is generalized propensity score and $F(T)$ is an artificially constructed function on *T*. Assuming $T_i \mid W_i \sim N(\alpha_0 + \alpha_1 W_i, \sigma^2)$, $(\alpha_0, \alpha_1, \sigma^2)$ can be estimated by using maximum likelihood estimation. And $R_i(T,W) = P(T_i = 1 \mid W_i)$ can be calculated as

$$R_i(T_i,W_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(T_i - \alpha_0 - \alpha_1 W_i)) \tag{3.9}$$

Y_i can be predicted by setting T_i and R_i as independent variable. $(\beta_0, \beta_t, \beta_w)$ can be estimated by using regression model.

Finally, by changing treatment variable *T, the* heterogeneous continuous treatment effect can be estimated. Also, average treatment effect for specific $T = t$ can also be estimated.

The function of $E\left(Y \mid W,T\right)$ can be generalized to improve the performance of estimation. In

previous research, Hirano & Imbens provided a case where $E(Y|W,T)$ can be formulated as

$$\begin{aligned} E(Y|W,T) &= \beta_0 + \beta_t T + \beta_w R(T,W) \\ &= \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 R + \beta_4 R^2 + \beta_5 TR \end{aligned} \tag{3.10}$$

And by using regression model, we can estimate $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. The treatment effect of sample $(T = T_i, W = W_i)$ can finally be estimated as:

$$\begin{aligned} &\hat{Y}(T = T_i, W = W_i) \\ &= \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 R_i(T_i, W_i) + \beta_4 R_i(T_i, W_i)^2 + \beta_5 T R_i(T_i, W_i) \end{aligned} \tag{3.11}$$

Back to (3.4), when $T_1 = T_0 + 1$, the heterogeneous continuous treatment effect can be denoted as:

$$HTE_{T_0 \to T_1} = \hat{Y}(T = T_1, W = W_i) - \hat{Y}(T = T_0, W = W_i) \tag{3.12}$$

It means when $T = T_0$ increase 1 unit, with W hold constant, the heterogeneous continuous treatment effect would be $HTE_{T_0 \to T_1}$, this can be used to quantify the effect of a change in T on Y.

We can also estimate average treatment effect at a specific $T = t$, which is:

$$ATE_{T=t} = E_W[\hat{Y}(T = t, W)] \tag{3.13}$$

3.2 Linear Double Machine Learning

The idea of linear double machine learning (LinearDML) originates from partially linear regression (Robinson, 1988) and can be formulated as:

$$\begin{aligned} Y &= \theta(W)T + g(W) + \epsilon \\ T &= f(W) + \eta \end{aligned} \tag{3.14}$$

Where $E(\epsilon|T,W) = 0$ and $E(\eta|W) = 0$. A navie modeling of *Y*, *W* and *T* will result in a biased estimate of $\theta(W)$ which comes from the estimation of $g(W)$ and overfitting.

LinearDML method (Chernozhukov et al., 2018) is a two-stage method proposed by Chernozhukov et al. and can be applied in continuous case. LinearDML applies two machine learning models to fit residual models, residual models *M1* and *M2* can be denoted as:

$$\begin{aligned} M1 &: \tilde{Y} = Y - \hat{Y}(W) \\ M2 &: \tilde{T} = T - \hat{T}(W) \end{aligned} \tag{3.15}$$

Where $\hat{Y}(W) = E(Y|W)$ and $\hat{T}(W) = E(T|W)$. $\tilde{Y}$ and $\tilde{T}$ are fitted residuals from the application of two machine learning models $\hat{Y}(W)$ and $\hat{T}(W)$ which can be fitted using any kinds of machine learning models. The inputs of these two machine learning models are covariates *W*.

We apply DML to estimate ATE and HTE, hence, a simple linear regression model can be constructed to estimate ATE $\theta(W)$, which is $\tilde{Y} = a + \theta(W)\tilde{T} + \epsilon$. $\theta(W)$ can be either parametric or non-parametric. The parametric model θ can be fitted directly, while the non-parametric model $\theta(W)$ needs to be transformed, and it is a function related to covariates *W*. It should be noted that DML uses cross-fitting to ensure the unbiased estimation of parameters. The basic idea of cross-fitting is to split data set into two samples, and then the model uses sample one to estimate residuals while uses sample two to estimate θ . The model further use sample two to estimate residuals while uses sample one to estimate θ . θ can be taken an average to get the final estimation.

If least squares method is applied to estimate $\theta(W)$, the objective function would be $\min_{\theta(W)}(\tilde{Y} - \theta(W)\tilde{T})^2$, and finally the estimation would be $\hat{\theta}(W) = \frac{\tilde{Y}}{\tilde{T}}$. $\tilde{Y}$ and $\tilde{T}$ are also related to *W*, hence, $\frac{\tilde{Y}}{\tilde{T}}$ can be denoted as the target term to estimate $\theta(W)$, which is $\theta(W) = \alpha W = \frac{\tilde{Y}}{\tilde{T}}$.

$$M3: \tilde{Y} = a + \alpha\tilde{T}W + \epsilon \tag{3.16}$$

According to equation (3.15), we can infer:

$$\begin{aligned} \tilde{Y} &= \hat{a} + \hat{\theta}(W)\tilde{T} \\ Y - \hat{Y}(W) &= \hat{a} + \hat{\theta}(W)(T - \hat{T}(W)) \\ Y &= \hat{a} + \hat{Y}(W) + \hat{\theta}(W)(T - \hat{T}(W)) \\ Y &= \hat{a} + \hat{Y}(W) + \hat{\alpha}W(T - \hat{T}(W)) \end{aligned} \tag{3.17}$$

Given a sample pair (W_i, T_i) , the prediction *Y* is:

$$Y = \hat{a} + \hat{Y}(W_i) + \hat{\alpha}W_i(T_i - \hat{T}(W_i)) \tag{3.18}$$

Right now, only simple linear regression is applied to estimate ATE, we can apply any other regression or machine models to estimated ATE. It should be noted that if we use the residuals of *T* to regress the residuals of *Y*, the estimated parameters are the ATE.

How to estimate HTE by using LinearDML? Assuming θ is a parametric model, we can add an interaction term in the model, which is:

$$\tilde{Y} = a + \theta_1 \tilde{T} + \theta_2 W \tilde{T} + \epsilon \tag{3.19}$$

Considering equation (3.12), the $HTE_{T_0 \to T_1}$ estimated by LinearDML would be:

$$\begin{aligned} HTE_{T_0 \to T_1} &= \tilde{Y}(T = T_1, W = W_i) - \tilde{Y}(T = T_0, W = W_i) \\ &= \hat{\theta}_1 \tilde{T}_1 + \hat{\theta}_2 W_i \tilde{T}_1 - \hat{\theta}_1 \tilde{T}_0 + \hat{\theta}_2 W_i \tilde{T}_0 \end{aligned} \tag{3.20}$$

3.3 Generalized random forest

Generalized Random Forest (GRF) (Athey et al., 2016) is a tree-based method proposed by Athey et al.

The first step of GRF is the same with DML, which constructs two machine learning models to fit residual models.

$$\begin{aligned} M1 &: \tilde{Y} = Y - \hat{Y}(W) \\ M2 &: \tilde{T} = T - \hat{T}(W) \end{aligned} \tag{3.21}$$

4. Transfer Entropy for Causal Discovery

First, let us recall the definition of joint Shannon entropy $H(X,Y)$ of two variables *X* and *Y*.

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P_{X,Y}(x,y) \log P_{X,Y}(x,y) \tag{4.1}$$

And the conditional Shannon entropy $H(Y \mid X)$ is

$$H(Y \mid X) = H(X,Y) - H(X) \tag{4.2}$$

The definition of transfer entropy $TE(X \to Y)$ can be formulated as the difference between the conditional entropy. The first term can be interpreted as the uncertainty of forward outcome variable *Y* under the condition of past outcome variable *Y*. The second term can be interpreted as the uncertainty of forward outcome variable *Y* under the condition of past outcome variable *Y* and past independent variable X. We can further define “forward” and “past”. We introduce a lag variable *L*, the past variable start from $t-1$ to $t-L$, the timestamp of forward variable is current time *t*. By using the equation of conditional Shannon entropy, transfer entropy can be rewritten as:

$$
\begin{aligned}
TE(X \to Y) &= H(Y^F \mid Y^P) - H(Y^F \mid X^P, Y^P) \\
&= H(Y^t \mid Y^{t-1:t-L}) - H(Y^t \mid X^{t-1:t-L}, Y^{t-1:t-L}) \\
&= H(X^{t-1:t-L}, Y^{t-1:t-L}) - H(Y^t, X^{t-1:t-L}, Y^{t-1:t-L}) \\
&\quad + H(Y^t, Y^{t-1:t-L}) - H(Y^{t-1:t-L})
\end{aligned} \tag{4.3}
$$

It is clear that transfer entropy has direction, which mean $TE(Y \to X)$ may not equal to $TE(X \to Y)$.

We can define net information outflow to quantify the direction and value of transfer entropy, which is:

$$
TE(X \to Y) = TE(X \to Y) - TE(Y \to X) \tag{4.4}
$$

If $TE(X \to Y) > 0$, the information from X to Y take the main role. Barnett (Barnett et al., 2009) proved that the linear granger-causality and transfer entropy are equivalent if all variables are joint Gaussian distribution.

In conclusion, the directionality and quantality of transfer entropy can be used for causal discovery in multivariate time series.

5. Reference

Athey, S., Tibshirani, J., & Wager, S. (2016). *Generalized Random Forests*. http://arxiv.org/abs/1610.01271

Barnett, L., Barrett, A., & Seth, A. (2009). Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters*, *103*, 238701. https://doi.org/10.1103/PhysRevLett.103.238701

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Hirano, K., & Imbens, G. W. (2004). The Propensity Score with Continuous Treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (pp. 73–84). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0470090456.ch7

Robinson, P. M. (1988). ROOT-N-CONSISTENT SEMIPARAMETRIC REGRESSION. *Econometrica*, *56*, 931–954.

ROSENBAUM, P. R., & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Wilson Pok. (2020, August 10). *How uplift modeling works*. Https://Ambiata.Com/Blog/2020-07-07-Uplift-Modeling/.