Causal Discovery Under the Potential Outcome Framework

Yixiang Luo

December 12, 2021

Abstract

Causal discovery is an important and common problem to the science community. The classical causal discovery approaches under the Directed Acyclic Graph model have essential difficulties in computational feasibility and determining the causal direction. In this project, I propose a model of doing causal discovery under the potential outcome framework and several immature methods created based on existing approaches.

1 Introduction

Gelman (2011) points out that there are two types of causal queries:

1. *Forward causal questions*: the "what if" question, or the inference on the "effects of causes", i.e. what would happen if we do certain interventions?

2. *Reverse causal questions:* the "why" question, or inference on the "causes of effects", i.e. why do certain phenomena happen?

The potential outcome model we learned in most statistical or econometrical classes is mainly used for the forward causal questions. As for the reverse question, a typical workflow under the potential outcome model is made clear by Gelman and Imbens (2013). In particular, the "why" question naturally arises when we observe a strange behavior that cannot be explained by the current model. For example, consider we observe the higher employee have higher salaries in the labor market while standard economic models suggest that wages should be related to productivity, rather than height. Such anomaly would motivate researchers to introduce new variables like health and childhood nutrition into the model according to their domain knowledge and convert the reverse causal question to a forward causal question about the effect of the newly introduced causal variables. However, if we have little domain knowledge to propose reasonable causal variables, can we still tackle the reverse question? In particular, modern data science makes it possible to collect numerous data for a large number of variables that might be causes. Can we identify the causes from the data with a certain guarantee?

Causal discovery under the directed acyclic graph (DAG) framework seems to provide a good toolset for this problem. Most causal discovery approaches can be classified into two categories: the constraint-based approaches and the score-based approaches. The constraint-based methods such as the PC algorithm and the Fast Causal Inference algorithm (Spirtes et al., 2000) learn the undirected graph structure by recursively applying conditional independence test onto possible triples of variable sets in the graph. The causal directions are then partially detected by first identifying the colliders $a \to b \leftarrow c$ via conditional independence test and then using the so-called orientation propagation to impute the other directions. The score-based method such as the Greedy Equivalence Search algorithm (Chickering, 2002) starts with an empty graph and adds the best directed edge at each step which increases fit the most measured by some quasi-Bayesian score.

There are two major difficulties in applying these causal discovery methods. First, as the number of variables increases, learning a complicated causal DAG quickly becomes computationally infeasible and requires a huge amount of data. Second, the causal directions are only partially identifiable up to the so-called Markov equivalent class. To fix the first problem, Wang et al. (2013) and Wang et al. (2014) propose to learn only the local causal graph around a variable of interest; Xie and Geng (2008) and Xie et al. (2006) decompose learning a large DAG into problems of learning many small DAG. However, the causal directions are still only partially identifiable in their methods. For the second difficulty, the newly developed LiNGAM methods (Shimizu et al., 2006, 2011) shed a light on it as it can identify the causal directions except under 5 settings by assuming the noises are far from Gaussian or the causal relationship is nonlinear.

We shouldn't blame the algorithms for the above problems since these difficulties are from the essence of the causal DAG model. Unlike the simple potential outcome model, the DAG model requires knowing a complicated graph to make an inference. Moreover, it is impossible to identify a causal direction if only the values of two variables are observed, which lies in the heart of the direction identifiability problem in DAG causal discovery. However, the potential outcome model naturally doesn't have this issue as it decides the causal direction by domain knowledge. Hence a natural idea is, can we do causal discovery under the potential outcome model and enjoy its advantages? In particular, for the cases where we have domain knowledge of the causal order of the variables, e.g. ordering them by the time when they are measured, but little idea of which variables are potential causes, is there a way to identify the promising causes and their effect based on the observational data? In this project, I'll explore some ideas for it.

The rest of this paper is arranged as follows.

2 Formulation and assumptions

2.1 A motivating example

Let's start with an example that motivates the formulation of the problem. Consider we want to investigate why do some people make more money than others at the age of 50. And we have the following observational data.

1. At birth: the SNPs (genes) of the person, the monthly family income at birth.
2. At age of 12: the height, the graduation exam grades at primary school, the monthly family income when the final exams take place, how good is the environment of the city where the person lives.
3. At age of 18: the university the person enters, the BMI.
4. At age of 30: the hours the person sleeps, the number of friends the person has.
5. The outcome of interest: the income at age of 50.

The key feature of the above variables is that they are measured in a sequence of time. It leads to two consequences:

1. the lower level variables (measured earlier) might be causes of the higher-level variables (measured later) but the reverse is automatically not true. In other words, the causal direction is clear between different levels of variables.

2. the variables of the same level (measured at the same time) may be correlated (have common causes) but have no causal effect on each other.

The second point is a bit subtle. Let me illustrate it with some examples. Consider the SNPs. It is known that some SNPs can have a strong correlation as the corresponding genes are next to each other in the DNA sequence. However, they have no causal effect on each other in the sense that if we edit one gene by, say, CRISPR, the other gene wouldn't change. Another example is the primary school graduate exam grade and the monthly family income when the exam takes place. They might have common causes, e.g. the unemployment of a parent a year ago might affect the study environment of the child thus the exam grades as well as the later family income. But the family income when the exam takes place is expected to have no or at least very little causal effect on the exam grades.

In general, the non-causality between the variables of the same level relies on that they are measured at the same time. This makes the formulation of the problem convenient but weakens the causal conclusion we can draw. For instance, suppose we find that ”the monthly family income when the primary school graduation exam takes place” (”instant income” for short) is a cause and has a causal effect of 10 units. It is dangerous to claim that if we intervene and give a certain amount of money to that family in that month, the child would make 10 units more money at his/her age of 50. In fact, the instant income should be treated as a proxy of ”the family economic status around that time” (”period income” for short). In the causal inference language, the period income is an unobserved common cause of the instant income and the outcome of interest. Importantly, the instant and period income are a ”cause” of each other. Hence the real causal effect should be weaker than we estimate if the latent period variable covers a much longer time period than the instant variable. And in this sense, these instant variables automatically break the unconfoundedness assumption to a certain extent. However, if we insist on being honest to the reality and adopting the period variables, it's natural to see complicated mutual causal effects among many variables, which is beyond the ability of many state-of-the-art methods. I believe using instant variables is a reasonable and mathematically clean way to simplify the real situation.

2.2 Notations and formulation

In this section, I'll formalize what we have seen in the example above.

Let $X^{(l,k)}$ be the possible causal variables we want to investigate, where $l = 1, 2, \ldots, L$ is the index of L levels and $k = 1, 2, \ldots, K_l$ is the index of the K_l variables within each level l. For instance, variable "the height" in the example above is the first variable of level 2. So it is denoted as $X^{(2,1)}$. For simplicity, let

$$j(l,k) = \sum_{s=1}^{l-1} K_s + k$$

be the one dimensional index of the (l,k)-th variable so $X^{(j)} := X^{(l,k)}$ and

$$m = \sum_{s=1}^{l} K_s$$

be the total number of variables. Define $X = \left(X^{(j)}\right)_{j=1,\ldots,m}$ as the vector of all variables and

$$V^{(l,k)} = \left(X^{(s,t)} :\ s \leq l, (s,t) \neq (l,k)\right)$$

as the vector of variables having no higher level than $X^{(l,k)}$ excluding $X^{(l,k)}$, ordered by j.

Denote the outcome of interest as Y. The causal effect is probably easier to think about in the language of the DAG model. But let's write down the potential outcome

$$Y(\mathrm{do}(X^{(l,k)}) = x)$$

as the outcome if we assign treatment $X^{(l,k)} \leftarrow x$ without interfering the pretreatement variables. For simplicity of notation, let $Y^{(l,k)}(x) := Y(\mathrm{do}(X^{(l,k)}) = x)$ be its succinct version. Formally, the potential outcome $Y^{(l,k)}(x)$ is nothing but an "arbitrary" random function of x in the science table. The causal effect of $X^{(l,k)}$ is define as some functional on $Y^{(l,k)}(x)$. For example, if $X^{(l,k)}$ is binary and we employ the ATE, then the causal effect is

$$\tau^{(l,k)} = \mathbb{E}\left[Y^{(l,k)}(1) - Y^{(l,k)}(0)\right].$$

As before, I use j as the one-dimensional alternative for the index (l, k) and define $Y(x) = \left(Y^{(j)}(x_j)\right)_{j=1,\ldots,m}$ as the vector of potential outcomes, where $x = (x_1, x_2, \ldots, x_m) \in \mathcal{X}$ is a vector of the assigned treatment and $\mathcal{X} = \prod_{j=1}^m \mathcal{X}^{(j)}$ is its range.

Suppose we observe

$$(X_i, Y_i(\mathcal{X})) \overset{\text{i.i.d.}}{\sim} (X, Y(\mathcal{X}))$$

for unit $i = 1, 2, \ldots, n$, where $Y(\mathcal{X})$ denotes the joint vector over all values of $x \in \mathcal{X}$. The observed outcome

$$Y_i := Y_i^{(j)}(X_i^{(j)}), \quad \forall j.$$

For simplicity, I write the observed X-variables as a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ with $\boldsymbol{X}_{ij} := X_i^{(j)}$.

The goal is to estimate the treatment effect of each variable $X^{(l,k)}$ on Y and select the causal variables with a certain guarantee. Here I pick the finite sample false discovery rate (FDR) (Benjamini and Hochberg, 1995) control as the target. We can phrase the causes selection problem as hypothesis testing:

$$H_j : \tau^{(j)} = 0.$$

Let $\mathcal{H}_0$ be the set of true null hypothese and $\mathcal{R}$ be the set of hypotheses we reject. Then the false discovery proportion (FDP) is defined as

$$\text{FDP} = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}.$$

And FDR is defined as $\text{FDP} = \mathbb{E}[\text{FDP}]$. The goal is to control $\text{FDR} \leq \alpha$ for some given $\alpha \in (0, 1)$.

2.3 Assumputions

Here I make my assumptions clear.

The first assumption is rephrasing the key feature we see in the motivating example. It is not a component in the potential outcome framework and is informal. I include it here as it makes the PO framework and the other assumptions more plausible.

Assumption 2.1 (causal ordering). *$X^{(l,k)}$ happens no earlier than any $X^{(s,t)}$ for $s \leq l$. And Y happens the last.*

The rest assumptions are basically standard in the potential outcome model.

Assumption 2.2. *Stable Unit Treatment Value Assumption.*

Assumption 2.3 (uniform ignorability)**.**

$$X^{(l,k)} \perp\!\!\!\perp Y^{(l,k)}(x) \mid V^{(l,k)}, \quad \forall x \in \mathcal{X}^{(l,k)}$$

holds for all (l, k)*.*

The uniform ignorability assumption is not hard to understand. Suppose we want to investigate the causal effect of variable $X^{(l,k)}$ (view it as treatment). Then based on the causal ordering assumption, the pre-treatment variables are the elements in $V^{(l,k)}$ and thus can be viewed as covariates. Then the uniform ignorability assumption reduces to the usual ignorability assumption. In the language of the DAG framework, it means any common causes of $X^{(l,k)}$ and Y are blocked by $V^{(l,k)}$.

Remark 2.1. *Is the uniform ignorability assumption plausible? Indeed, it is much stronger than the usual ignorability assumption that applies to a single treatment. But just as in the usual case, the assumption becomes more plausible if we include many more possible causes into* X*. And this is exactly the case we are facing here: we put numerous variables that might be causes of* Y *into* X *and try to figure out the true causes among them. And even if the assumption doesn't hold, we find good proxies for the actual causes which may guide further research.*

3 Proposed methods

In this section, I'll explore some ideas to solve the problem. The first subsection proposes a naive method under the general setting. And the next subsection discusses some weakness of the general method and propose possible solutions under the Gaussian graphical model.

3.1 General setting

In this subsection, I make no model assumption and try to tackle the problem under the general setting. I have little experience in handling treatments of multi-level or continuous value. Hence here when investigating the treatment effect of $X^{(l,k)}$, I dichotomize it and employ the ATE. Formally, this is to define the treatment effect as (suppressing the superscript (l,k) on X, V, Y, $\widetilde{Y}$, w, and τ_d for simplicity)

$$\tau_d := \mathbb{E}\left[\widetilde{Y}(1) - \widetilde{Y}(0)\right],$$

where

$$\widetilde{Y}(z) := \int w(x;z) Y(x),$$

$$w(x;z) := \mathbf{1}_{z=1,\, x \geq \bar{x}} \int d\,P_V(V)\, \frac{d\,P_{X|V}(x)}{P_{X|V}(x \geq \bar{x})} + \mathbf{1}_{z=0,\, x<\bar{x}} \int d\,P_V(V)\, \frac{d\,P_{X|V}(x)}{P_{X|V}(x < \bar{x})},$$

and $\bar{x} = \mathbb{E}[X]$, $P_{X|V}$ is the conditional distribution of X given V, and P_V is the marginal distribution of V. In other words, the dichotomized potential outcome $\widetilde{Y}(x)$ is a weighted average of the original potential outcome $Y(x)$, where the weight is determined based on the proportion of x among all x that falls into the same category.

3.1.1 inference on individual treatment

Consider inferring the treatment effect of a single variable $X^{(l,k)}$. Again, I suppress the superscript (l,k) for simplicity when there is no ambiguity.

As discussed before, the uniform ignorability assumption reduces to the usual ignorability assumption when we view X as treatment and V as covariates. And it's not hard to see the dichotomized potential outcomes still satisfy the ignorability assumption

$$Z \perp\!\!\!\perp \widetilde{Y}(z) \mid V, \quad z = 0, 1, \quad \text{where } Z := \mathbf{1}\{X > \bar{x}\}$$

Hence we can adopt the usual causal inference methods in the potential outcome framework. As a final project, I employ three main approaches for the observational study we learned in the class.

1. *IPW/Hajek estimator.*

$$\hat{\tau}_d^{\text{hajek}} = \frac{\sum_{i=1}^n \frac{Z_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^n \frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^n \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^n \frac{1-Z_i}{1-\hat{e}(X_i)}},$$

   where $\hat{e}$ is the estimated propensity score.

   An estimator of the variance $\hat{V}^{\text{hajek}}$ is obtained by bootstrap.

2. *Doubly robust estimator* with linear model.

$$\hat{\tau}^{\mathrm{dr}} = \hat{\mu}_1^{\mathrm{dr}} - \hat{\mu}_0^{\mathrm{dr}},$$

   where

$$\hat{\mu}_1^{\mathrm{dr}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Z_i \left\{ Y_i - \hat{\mu}_1 \left(X_i \right) \right\}}{\hat{e} \left(X_i \right)} + \hat{\mu}_1 \left(X_i \right) \right],$$

$$\hat{\mu}_0^{\mathrm{dr}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\left(1 - Z_i \right) \left\{ Y_i - \hat{\mu}_0 \left(X_i \right) \right\}}{1 - \hat{e} \left(X_i \right)} + \hat{\mu}_0 \left(X_i \right) \right]$$

   and $\hat{\mu}_1$ is the predicted value of Y by linear regression on the treated group and $\hat{\mu}_0$ is similar but on the control group.

   Again, an estimator of the variance $\hat{V}^{\mathrm{dr}}$ is obtained by bootstrap.

3. *Matching based on propensity score.*

   I adopt the 1-nearest matching measured by the estimated propensity score. The biased corrected estimator $\hat{\tau}^{\mathrm{mbc}}$ for the treatment effect and $\hat{V}^{mbc}$ is employed. The long formulas are omitted here for simplicity. Please see lecture notes section 15 for them.

Remark 3.1. *As we infer the treatment effect for a hight-level variable $X^{(l,k)}$ with large l, the pre-treatment covariates $V^{(l,k)}$ may contain too many variables so that conditioning on $V^{(l,k)}$ may cause the non-overlap issue. A natural idea to solve this is to do the individual inference sequentially from low level to high, and don't condition on those showing no causal effect to the outcome. But this is invalid since*

1. *Due to the nature of hypothesis testing, we can only claim there exists a causal effect with certain confidence but cannot claim the null hypothesis, there is no causal effect, is true even if we don't reject the null.*

2. *What we estimate is the total treatment effect rather than the direct treatment effect. Hence it is possible that $X^{(1,1)}$ has negative direct effect on Y and positive effect via $X^{(1,1)} \to X^{(2,2)} \to Y$ and show zero total effect. But when doing inference on $X^{(2,2)}$, we definitely should condition on $X^{(1,1)}$.*

Fortunately, this problem would be relieved largely if we always use the propensity score as the covariate to condition on. By doing so, the non-overlapping issue happens only if the covariates are too predictive to the treatment variable. If we believe *the causal DAG is sparse, then including many nuisance variables doesn't really make the covariates more predictive of the treatment.*

3.1.2 multiple testing procedures

Once we have an estimator $\hat{\tau}^{(j)}$ for the treatment effect of $X^{(j)}$ and an estimator $\hat{V}^{(j)}$ for the variance of $\hat{\tau}^{(j)}$, we can obtain an asymptotically valid p-value

$$p_j = 2 \cdot \Phi\left(-\left|\frac{\hat{\tau}^{(j)}}{\sqrt{\hat{V}^{(j)}}}\right|\right),$$

by the central limit theorem (if holds), where Φ is the CDF of the standard normal distribution.

Then the causal selection problem becomes the classical multiple testing problem with FDR control. Benjamini and Hochberg (1995) proposed the commonly used BH(α) procedure for this purpose, which rejects

$$\mathcal{R} = \left\{j : p_j \leq \frac{\alpha R}{m}\right\},$$

where

$$R = \max\left\{r : \frac{m p_{(r)}}{r} \leq \alpha\right\}$$

and $p_{(r)}$ is the r-th smallest among all the p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

The BH(α) procedure gives provably FDR control if the p-values are independent or satisfy certain positive dependence called PRDS (Benjamini and Yekutieli, 2001). However, due to the dichotomization and the unknown causal relationship between the X variables, the p-values we get have a complicated and unknown dependency. Hence we need to adopt the more conservative method BY(α) = BH(α/L_m) proposed by Benjamini and Yekutieli (2001), where

$$L_m = \sum_{k=1}^{m} \frac{1}{k} \approx \log m,$$

if we want to control FDR with a theoretical guarantee.

The last concern is that we only have (maybe) asymptotically valid p-values. The following proposition should help.

Proposition 3.1. *Suppose* $\mathbb{P}(p_j \le t) \le t + \varepsilon$ *for all* j *and some* $\varepsilon \ge 0$. *Then the* $\mathrm{BY}(\alpha)$ *procedure based on these p-values control FDR at* $\alpha + mL_m\varepsilon$.

Proof. Let $\mathcal{R}$ denotes the rejection set of $\mathrm{BH}(\alpha)$ and $R = |\mathcal{R}|$ for simplicity of notation. Define

$$S_i = \begin{cases} 1, & p_i \le \frac{\alpha}{m} \\ \frac{1}{2}, & \frac{\alpha}{m} < p_i \le \frac{2\alpha}{m} \\ \vdots & \\ \frac{1}{m}, & \frac{\alpha(m-1)}{m} < p_i \le \alpha \\ 0, & p_i > \alpha \end{cases}.$$

Since $\mathbb{P}(p_j \le t) \le t + \varepsilon$, we have

$$\mathbb{E}\left[S_i\right] \le \left(1 + \frac{1}{2} + \ldots + \frac{1}{m}\right) \cdot \left(\frac{\alpha}{m} + \varepsilon\right) = L_m \cdot \left(\frac{\alpha}{m} + \varepsilon\right).$$

Then we have

$$\mathrm{FDR} = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\left\{p_i \le \frac{\alpha R}{m}\right\}}{R \vee 1}\right] \le \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[S_i\right] \le L_m(\alpha + m\varepsilon),$$

where the first inequality is by

$$p_i \le \frac{\alpha R}{m} \quad \Rightarrow \quad S_i \ge \frac{1}{R} \quad \Rightarrow \quad \frac{\mathbf{1}\left\{p_i \le \frac{\alpha R}{m}\right\}}{R} \le S_i.$$

Finally, noticing $\mathrm{BY}(\alpha) = \mathrm{BH}(\alpha / L_m)$, the propostion is proved. □

It's not hard to see that if the CLT holds and $\hat{V}^{(j)}$ is conservative, then the condition in Proposition 3.1 holds with a large enough n.

3.2 Gaussian graphical model

In this section, I consider a widely adopted model assumption and try a different idea to tackle the problem.

3.3 Formulation

Assume (X, Y) obeys the Gaussian graphical model, i.e they jointly follow a multivariate Gaussian distribution. This allows us to write Y as a Gaussian linear model of any subset of X: for any $\mathcal{L} \subseteq [m]$, we can write

$$Y = \sum_{j \in \mathcal{L}} X^{(j)} \beta_j^{\mathcal{L}} + a^{\mathcal{L}} + \varepsilon^{\mathcal{L}} \tag{1}$$

where $a^{\mathcal{L}}$ is a constant intercept and $\varepsilon^{\mathcal{L}}$ is a Gaussian noise independent of X. The superscript $\mathcal{L}$ indicates these variables may vary as $\mathcal{L}$ varies.

Now come back to the causal inference. Under the potential outcome model with ignorability assumption, formally, the most general definition of existing a causal effect is that conditioning on the pre-treatment variables, the treatment is not independent with the outcome

$$X^{(l,k)} \not\perp Y \mid V^{(l,k)}.$$

With the linear model (1), it's equivalent to have $\beta_j^{\mathcal{L}(j)} \neq 0$ in

$$Y = \sum_{i \in \mathcal{L}(j)} X^{(i)} \beta_i^{\mathcal{L}(j)} + a^{\mathcal{L}(j)} + \varepsilon^{\mathcal{L}(j)}, \tag{2}$$

where

$$\mathcal{L}(j) = \left\{ i : \ i \leq \min \left\{ \sum_{s=1}^{l} K_s : \ \sum_{s=1}^{l} K_s \geq j \right\} \right\}$$

is the set of the one-dimensional variable indeces that has level no higher than $X^{(j)}$. In words, we only need to include $X^{(l,k)}$ and $V^{(l,k)}$ in the linear model when infering the total treatment effect of $X^{(l,k)}$. And this correspond to the regression based method for linear model in the classical causal inference.

Formally, we can rephrase our causal discovery problem as again a multiple hypotheses testing on

$$H_j : \ \beta_j^{\mathcal{L}(j)} = 0, \quad j \in [m],$$

where the m different $\beta_j^{\mathcal{L}(j)}$ come from L different linear models of the form (2).

3.4 Modify the knockoff method

The knockoff method (Barber and Candès, 2015; Candes et al., 2018) is shown to have higher power than the BH method under many practical settings in selecting the nonzero β coefficients in a Gaussian linear model with FDR control. I'll use the fixed-X version of the method which suits our goal the best.

Consider we observe $Y \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ in the model

$$Y = \boldsymbol{X}\beta + \varepsilon.$$

To select the $\beta_j \neq 0$, the knockoff method constructs a so-called knockoff matrix $\tilde{\boldsymbol{X}}$ matrix satisfying

$$\left[\boldsymbol{X}\ \tilde{\boldsymbol{X}}\right]^T \left[\boldsymbol{X}\ \tilde{\boldsymbol{X}}\right] = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top\mathbf{X} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix} := \mathbf{G}$$

for some vector $\boldsymbol{s} \geq 0$ that makes $\boldsymbol{G}$ semi-positive definite. The idea is that the knockoff feature $\tilde{\boldsymbol{X}}_j$ mimics the true feature $\boldsymbol{X}_j$ in its correlation structure with the other features but not the outcome. Hence if $\beta_j = 0$, a feature selection method like lasso should not be able to distinguish $\boldsymbol{X}_j$ with its knockoff $\tilde{\boldsymbol{X}}_j$, which serves as a negative control for FDR.

In particular, knockoff then run Lasso with different penalty level λ in the augmented regression problem $Y \sim [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$ and define

$$\begin{aligned} W_j = {} & (\text{ largest } \lambda \text{ such that } \mathbf{X}_j \text{ or } \tilde{\mathbf{X}}_j \text{ enters Lasso path }) \\ & \times \begin{cases} +1 & \text{if } \mathbf{X}_j \text{ enters before } \tilde{\mathbf{X}}_j \\ -1 & \text{if } \mathbf{X}_j \text{ enters after } \tilde{\mathbf{X}}_j \end{cases} . \end{aligned}$$

Then knockoff reject

$$\mathcal{R} = \{j : W_j \geq T\},$$

where

$$T = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq \alpha\right\}.$$

Our causal selection problem is different from the original feature selection goal of knockoff in the following two aspects.

1. Our coefficient $\beta_j^{\mathcal{L}(j)}$ are from L different model rather than a single linear model.

2. In one particular model (2), only the features/variables having the same level with $X^{(j)}$ should be considered to be selected.

Here I propose a modified knockoff method to tackle our causal selection problem. For each level $l = 1, 2, \ldots, L$, let $\mathcal{L}_l$ be the indeces set of all variables having level no higher than l, e.g. $\mathcal{L}_l = \mathcal{L}(j)$ if $X^{(j)}$ has level l. Let the initial selection set $\mathcal{R} = \emptyset$. Iterate

1. Let $Y = (Y_i)_{i \in [n]}$ be the vector of the observed outcome among n units. Let $matX^{\mathcal{L}_l}$ be the matrix of the observed variables having level no higher than l.

2. Run knockoff at FDR level α/L with a properly chosen $\boldsymbol{s}$ such that $\boldsymbol{s}_j = 0$ for all j such that $X^{(j)}$ has level less than l. By doing so, we will automatically have $W_j = 0$ and hence only the features/variables having the same level as l are considered to be selected. Denote the selection set as $\mathcal{R}_l$.

3. update $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_l$.

And we can show it has provable FDR control.

Theorem 3.1. *The modified knockoff method controls FDR.*

Proof.

$$\begin{aligned}\mathrm{FDR}(\mathcal{R}) = \mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right] &= \sum_{l=1}^{L} \mathbb{E}\left[\frac{|\mathcal{R}_l \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right] \\ &\leq \sum_{l=1}^{L} \mathbb{E}\left[\frac{|\mathcal{R}_l \cap \mathcal{H}_0|}{|\mathcal{R}_l| \vee 1}\right] = \sum_{l=1}^{L} \mathrm{FDR}(\mathcal{R}_l) \leq \sum_{l=1}^{L} \frac{\alpha}{L} = \alpha\end{aligned}$$

□

Remark 3.2. *Knockoff is known to perform well if the true model is sparse and the nominal FDR level is not too small. Otherwise, it might be worse than BH/BY. Hence I expect the modified knockoff method to perform well if we don't have too many levels of variables but have many variables in each level.*

4 Simulation studies

In this section, I examine the proposed methods by numerical simulations. Due to the computational price, e.g. the doubly robust estimator requires $2m$ times model fitting for a single problem, and the time limit of this project, I'm only able to do the experiments on limited settings. But they should reveal typical behaviors of the methods.

Recall we have m variables $X^{(j)}$ for $j = 1, 2, \ldots, m$ with each has its level. I generate the data via the causal DAG model with Gaussian linear model as follows.

1. Create a Erdos-Renyi graph of $m + 1$ nodes with the probability of assigning an edge to be $\pi_1 \in [0, 1]$. Treat the first m nodes as X and the last node as Y.

2. Generate the level of each X variable to make each level has about the same number of variables. For example, if we have $m = 7$ and $L = 3$, I assign $X^{(1)}, X^{(2)}, \dots, X^{(m)}$ to have level

   $$1, 1, 2, 2, 3, 3, 3.$$

   Let Y have the highest level $L + 1$.

3. If there is an edge between two nodes of different levels, then make it an directed causal edge from the lower level to the higher level.

4. Let $pa(v)$ be the set of nodes that eject an edge into v, where v can be either a Y or a X variable. Then from low level to high level, the value of v is determined as

   $$\text{value}(v) \leftarrow \sum_{w \in pa(v)} \mu(w, v) \cdot \text{value}(w) + N(0, 1),$$

   where μ is the signal strength and $N(0, 1)$ is an independent noise.

In short, I adopt the Gaussian graphical model with sparsity level π_1. And the levels of X variables are distributed evenly.

The existence of a causal effect is defined as rejecting

$$H_j : \; X^{(j)} \not\perp\!\!\!\perp Y \mid V^{(j)}.$$

I'll compare the estimated FDR and power, defined as

$$\text{Power} := \mathbb{E}[\frac{|\mathcal{R} \cap \mathcal{H}_1}{|\mathcal{H}_1|}], \quad \mathcal{H}_1 := [m] \setminus \mathcal{H}_0,$$

of the proposed methods.

In the experiments below, I set $m = 100$, $n = 400$ and $\pi_1 = 0.2$. μ is set to have value varies from 0.05 to 0.3. The "Wide", "Moderate", and "Long" structures correspond to set $L = 1, 4, 10$. The code to reproduce the results is available at `https://github.com/yixiangLuo/CDPO`.

Figure 1 shows the estimated FDR and power of the proposed methods as signal strength increases. All methods control FDR empirically (The doubly robust estimator under the "Wide" structure exceeds a little bit But that can be explained by Monte-Carlo error).

Some major takeaways from the figure:

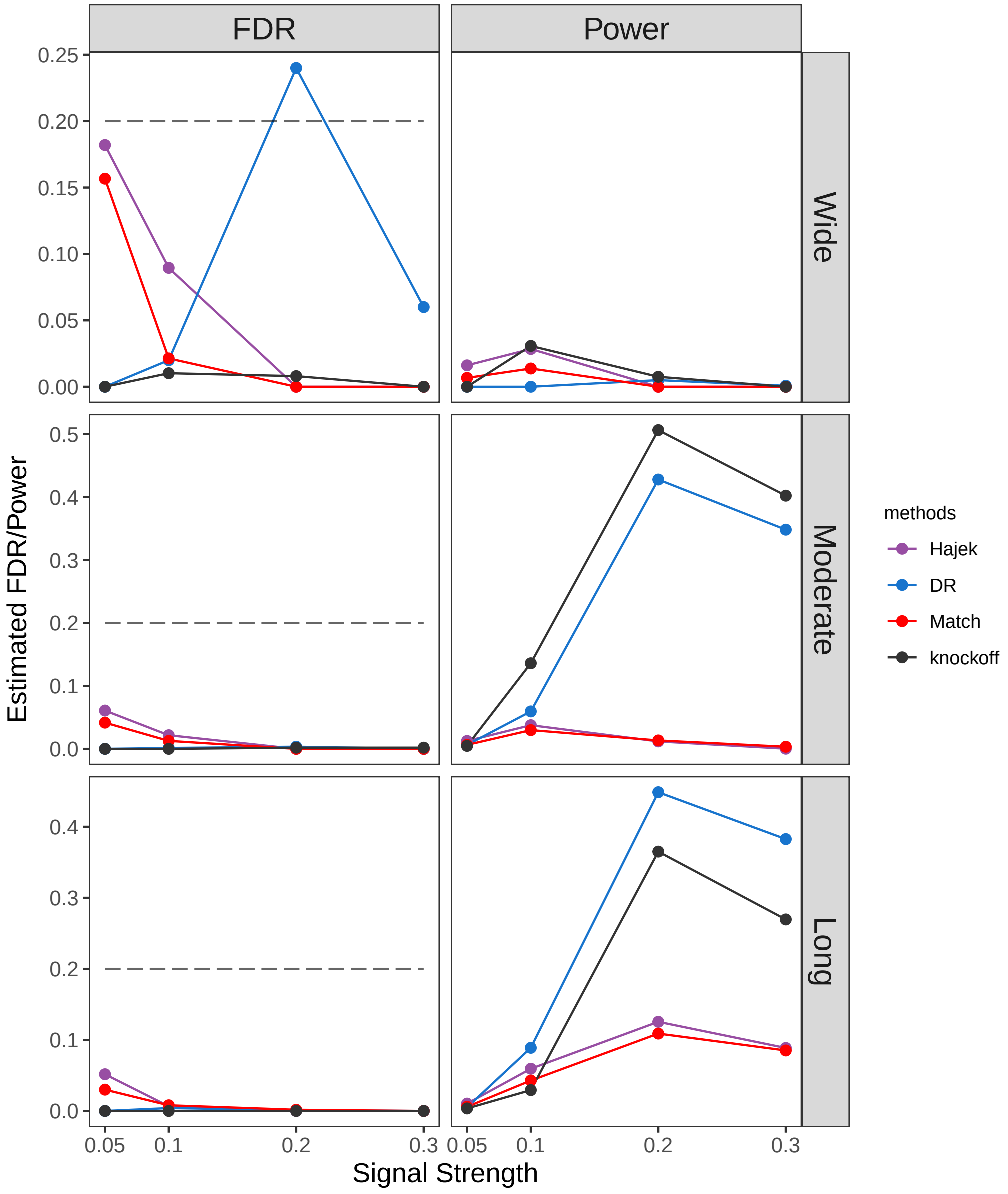

Figure 1: Estimated FDR and power of the proposed methods as signal strength increases. $\alpha = 0.2$. All methods controls FDR empirically. The doubly robust estimator and the modified knockoff are among the best.

1. The doubly robust estimator and the modified knockoff are among the best. In particular, the modified knockoff method is better when the level structure

is "wider" (more X variables in the same level and less total number of levels), which is as expected. The powerless of all methods under the extremely wide setting (the "Wide" in the figure, which has $L = 1$) remains a mystery and needs further investigation.

2. As the signal strength increases after a certain point, e.g. 0.2, the power decreases. For the ATE-based methods (Hajek, DR, and Match), this is because as the signal strength increases, the covariates $V^{(j)}$ become more predictive to the treatment variable $X^{(j)}$. Hence more estimated propensity scores tend to the extreme value 0 or 1 and hence the data point is swiped out. When the number of overlapped data points (propensity score $\in [0.1, 0.9]$) is less than say 10, I enforce not to reject such $X^{(j)}$. For the modified knockoff method, I think it is because that the stronger positively correlated design matrix $\boldsymbol{X}$ introduces a larger whitening noise to knockoff, which is known as an important difficulty of the knockoff method.

5 Discussion

In this project, I propose a new model of layered causal discovery under the potential outcome framework. The model is suitable when we have domain knowledge of the time ordering the variables but little clue of the causes of an outcome of interest.

Several approaches moderately modified from existing methods are proposed to tackle the problem. Numerical simulations show they work okay but not really well. Especially in the sense that as the signal strength increases, none of the methods would attain power 1. This calls for further research on new methods that are really designed for this problem if the whole causal discovery story and framework make sense.

References

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Andrew Gelman. Causality and statistical learning, 2011.

Andrew Gelman and Guido Imbens. Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research, 2013.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Changzhang Wang, You Zhou, and Zhi Geng. Discovering causes and effects of a given node in bayesian networks. *Frontiers of Mathematics in China*, 8(3): 643–663, 2013.

Changzhang Wang, You Zhou, Qiang Zhao, and Zhi Geng. Discovering and orienting the edges connected to a target variable in a dag via a sequential local learning approach. *Computational Statistics & Data Analysis*, 77:252–266, 2014.

Xianchao Xie and Zhi Geng. A recursive method for structural learning of directed acyclic graphs. *The Journal of Machine Learning Research*, 9:459–483, 2008.

Xianchao Xie, Zhi Geng, and Qiang Zhao. Decomposition of structural learning about directed acyclic graphs. *Artificial Intelligence*, 170(4-5):422–439, 2006.