A short note on the adjustment sets

Yen-Chi Chen
University of Washington
April 24, 2020

Adjustment sets are useful concepts in causal inference. It appears when a distribution factorizes with respect to a DAG (directed acyclic graph) $G = (V,E)$, where V is the vertex set and E is the edge set. For simplicity, consider a multivariate distribution with variables X,Y,Z such that $X,Y \in \mathbb{R}$ and $Z \in \mathbb{R}^d$. Here we use X as the intervention variable and Y as the outcome variable.

Let $p(x,y,z)$ be the underlying density function. When the density/distribution function factorizes with respect to G, the vertex set $V = (X,Y,Z)$. Using the edges in E, we can easily define parents of each node. Let PA_X be the parentd of node X. The graph factorization implies the following decomposition:

$$p(x,y,z) = p(x|\mathsf{PA}_X)p(y|\mathsf{PA}_Y)\prod_j p(z_j|\mathsf{PA}_{Z_j}).$$

In causal inference, we are often interested in the effect on an outcome variable Y from an intervention of $X = x$. Using the g-formula (a.k.a. truncation factorization), the intervention of $X = x$ can be described as[1]

$$p(y,z|\mathbf{do}(x)) = p(y|\mathsf{PA}_Y)\prod_j p(z_j|\mathsf{PA}_{Z_j}).$$

Namely, we remove the term $p(x|\mathsf{PA}_X)$ in the joint distribution.

Thus, the marginal effect on y will be

$$p(y|\mathbf{do}(x)) = \int p(y|\mathsf{PA}_Y)\prod_j p(z_j|\mathsf{PA}_{Z_j})dz.$$

It is easy to see that in general

$$p(y|\mathbf{do}(x)) \neq p(y|x).$$

However, as one may expect, some Z_j in the integral $\int p(y|\mathsf{PA}_Y)\prod_j p(z_j|\mathsf{PA}_{Z_j})dz$ may not be used. So there might be a subset $W \subset Z$ such that

$$p(y|\mathbf{do}(x)) = \int p(y|w,x)p(w)dw. \tag{1}$$

If W satisfies the above equality, then W is called an **adjustment set**. In a sense, the adjustment set describes the variables that we need to adjust for computing the causal effect from X on Y.

The adjustment set is generally not unique. There can be many adjustment sets for the same outcome variable and the same intervention variable. So this leads to a natural question: how do we compare different adjustment sets? will an adjustment set with less variables be better than a set with more variable?

[1] an introduction can be found in Section 12.2 of http://faculty.washington.edu/yenchic/19A_stat535/Lec12_causal_missing.pdf

1 Doubly robust estimator

To answer these questions, we consider a specific quantity of interest: the average of Y under the intervention $X = x$. To simplify the problem, we assume that X is binary. In this case, the parameter of interest is

$$\mu_x = \mathbb{E}(Y|\mathbf{do}(X=x)) = \int yp(y|w,x)p(w)dydw.$$

One can easily show the following two equivalent expressions of μ_x:

$$\begin{aligned}\mu_x &= \mathbb{E}\left(\frac{YI(X=x)}{P(X|W)}\right)\\ &= \mathbb{E}(m(x,W)),\end{aligned}$$

where $P(x|w) = P(X=x|W=w)$ is known as the propensity score and $m(x,w) = \mathbb{E}(Y|X=x,W=w)$ is the regression function. The first equality leads to the the inverse probability weighting (IPW) estimator and the second equality is associated with the regression adjustment (a.k.a. g-computation estimator).

Given observations

$$(X_1,Y_1,Z_1),\cdots,(X_n,Y_n,Z_n),$$

the IPW estimator is

$$\widehat{\mu}_{x,IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_iI(X_i=x)}{P(X_i|W_i)}$$

and the the regression adjustment (RA) estimator is

$$\widehat{\mu}_{x,RA} = \frac{1}{n}\sum_{i=1}^{n} m(x,W_i).$$

In fact, there is an estimator the combines both estimators. This is known as the *doubly robust (DR)* estimator

$$\widehat{\mu}_{x,DR} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(X_i=x)}{P(x|W_i)}(Y_i - m(x,W_i)) + m(x,W_i).$$

Each of the above estimator has a closed-form *influence function* that

$$\psi_{IPW}(X,Y,W) = \frac{YI(X=x)}{P(x|W)}, \quad \psi_{RA}(X,Y,W) = m(x,W), \quad \psi_{DR}(X,Y,W) = \frac{I(X=x)}{P(x|W)}(Y - m(x,W)) + m(x,W).$$

And one can easily see that

$$n\mathsf{Var}(\widehat{\mu}_{x,DR}) = \mathsf{Var}(\psi_{DR}(X,Y,W))$$

and similarly for RA and DR estimators. Thus, the variance of each estimator is characterized by the influence function.

2 Comparing nested adjustment sets

Now we focus on the DR estimator. Consider a case where we have nested adjustment sets W and C such that $W = (B, C)$. Then

- If $X \perp B | C$, then $W = (B, C)$ is better than C.
- If $Y \perp B | C, X$, then C is better than $W = (B, C)$.

This result was proved in the following paper (under a slightly different setting):

> Smucler, E.; Sapienza, Facundo; Rotnitzky, A. (2020). Efficient adjustment sets in causal graphical models with hidden variables. arXiv preprint arXiv:2004.10521.

Here we give the formal descriptions of what we meant.

Proposition 1 *Suppose that $W = (B, C)$ such that*

$$(G1) \quad X \perp B | C$$

and C is also another adjustment set. Then

$$\mathsf{Var}(\psi_{DR}(X, Y, C)) \geq \mathsf{Var}(\psi_{DR}(X, Y, B, C)) = \mathsf{Var}(\psi_{DR}(X, Y, W)).$$

Namely, although C is an adjustment set with less variables, its variance may be higher than the adjustment set $W = (B, C)$.

Here is another interesting result–sometimes, an adjustment set with less variables may indeed have a lower variance.

Proposition 2 *Suppose that $W = (B, C)$ such that*

$$(G2) \quad Y \perp B | C, X$$

and C is also another adjustment set. Then

$$\mathsf{Var}(\psi_{DR}(X, Y, C)) \leq \mathsf{Var}(\psi_{DR}(X, Y, B, C)) = \mathsf{Var}(\psi_{DR}(X, Y, W)).$$

Namely, C is a better adjustment set than $W = (B, C)$.

Therefore, depending on the underlying assumptions, we may use different adjustment sets to obtain an estimator with a smaller variance. Note that the proofs of the two propositions are not long so we provide them in Section 4.

3 Remarks

- **Information adjustment interpretation.** The assumption (G1) $X \perp B|C$ that leads to a larger adjustment set was not too surprising. This is the case where B becomes irrelevant to X if we know C. However, B could still be influencing Y. Thus, adjusting for B will stabilize the the estimate of the causal effect from X to Y.

  On the other hand, the assumption (G2) $Y \perp B|X,C$ means that when we adjust for X,C, B has no direct effect onto Y. Therefore, including B will not improve the estimate of the effect from X on Y.

- **Case of IPW and RA estimators.** Interestingly, (G1) and (G2) assumptions do not change the efficiency (variance) of the RA estimator so even with any of these two assumptions, we still cannot tell if an adjustment set with more/less variables will be better (but note that if we have to estimate the regression function, the one with less variable is preferred).

  For the IPW estimator, (G2) does imply that we should be using the estimator with less variables. This is due to the fact that (using a similar derivation as the proof of Proposition 2)

  $$\mathbb{E}(\psi_{IPW}(X,Y,B,C)|X,Y,C) = \psi_{IPW}(X,Y,C) \Rightarrow \mathsf{Var}(\psi_{IPW}(X,Y,B,C)) \geq \mathsf{Var}(\psi_{IPW}(X,Y,C)).$$

  (G1) does not tell us which estimator is better since under (G1), $\psi_{IPW}(X,Y,B,C) = \psi_{IPW}(X,Y,C)$.

- **The regression function and/or the propensity score has to be known.** The above analysis assumes that the propensity score and the regression function are both known. In practice, they are often unknown and has to be estimated. It is known that when we include more variables, estimation is often harder. In a parametric model setting, the rate is the same but the variance could be inflated due to the use of more variables. In a nonparametric setting, the rate is often drastically slowed down by the number of variables (an usual rate is $O(n^{-\frac{2}{4+d}})$) so including more variables may not help at all.

4 Proofs

Proof.[Proof of Proposition 1] Using $W = (B,C)$, we can write

$$\psi_{DR}(X,Y,W) = \psi_{DR}(X,Y,B,C) = \frac{I(X=x)}{P(x|B,C)}(Y - m(x,B,C)) + m(x,B,C)$$

and

$$\psi_{DR}(X,Y,C) = \frac{I(X=x)}{P(x|C)}(Y - m(x,C)) + m(x,C).$$

Assumption (G1) $X \perp B|C$ implies that $P(x|B,C) = P(x|C)$ so

$$\begin{aligned}\psi_{DR}(X,Y,C) &= \frac{I(X=x)}{P(x|C)}(Y - m(x,C)) + m(x,C)\\ &= \frac{I(X=x)}{P(x|B,C)}(Y - m(x,C)) + m(x,C)\\ &= \psi_{DR}(X,Y,B,C) - (m(x,C) - m(x,B,C))\left(\frac{I(X=x)}{P(x|C)} - 1\right).\end{aligned}$$

Here are two interesting facts. First, the mean of the second term is

$$\begin{aligned}
\mathbb{E}\left((m(x,C)-m(x,B,C))\left(\frac{I(X=x)}{P(x|C)}-1\right)\right) &= \mathbb{E}\left(m(x,C)-m(x,B,C))\mathbb{E}\left(\frac{I(X=x)}{P(x|C)}-1|B,C\right)\right) \\
&= \mathbb{E}\left(m(x,C)-m(x,B,C))\left(\frac{P(x|B,C)}{P(x|C)}-1|B,C\right)\right) \\
&= 0.
\end{aligned}$$

Second, the product of the two terms is

$$\begin{aligned}
&\mathbb{E}\left(\psi_{DR}(X,Y,B,C)(m(x,C)-m(x,B,C))\left(\frac{I(X=x)}{P(x|C)}-1\right)\right) \\
&\quad = \mathbb{E}\left(\mathbb{E}(\psi_{DR}(X,Y,B,C)|X,B,C)(m(x,C)-m(x,B,C))\left(\frac{I(X=x)}{P(x|C)}-1\right)\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}(\psi_{DR}(X,Y,B,C)|X,B,C) &= \left(\frac{I(X=x)}{p(x|B,C)}(m(X,B,C)-m(x,B,C))\right)+m(x,B,C) \\
&= m(x,B,C).
\end{aligned}$$

So the product of the two terms equals

$$\begin{aligned}
&\mathbb{E}\left(m(x,B,C)(m(x,C)-m(x,B,C))\left(\frac{I(X=x)}{P(x|C)}-1\right)\right) \\
&\quad = \mathbb{E}\left(m(x,B,C)(m(x,C)-m(x,B,C))\mathbb{E}\left(\frac{I(X=x)}{P(x|C)}-1|B,C\right)\right) \\
&\quad = \mathbb{E}\left(m(x,B,C)(m(x,C)-m(x,B,C))\left(\frac{P(x|B,C)}{P(x|C)}-1\right)\right) \\
&\quad = 0.
\end{aligned}$$

This two fact implies that the covariance

$$\mathsf{Cov}\left(\psi_{DR}(X,Y,B,C),(m(x,C)-m(x,B,C))\left(\frac{I(X=x)}{P(x|C)}-1\right)\right)=0.$$

As a result, we conclude that

$$\mathsf{Var}(\psi_{DR}(X,Y,C)) \geq \mathsf{Var}(\psi_{DR}(X,Y,B,C)) = \mathsf{Var}(\psi_{DR}(X,Y,W))$$

□

Proof.[Proof of Proposition 2]

Note that (G2) $Y \perp B|C,X$ implies that $p(b|x,y,c) = p(b|x,c)$. Thus, one can easily see that

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{P(X|B,C)}|X,Y,C\right) &= \int \frac{p(b|X,Y,C)}{p(X|b,C)}db \\
&= \int \frac{p(b|X,C)}{p(X|b,C)}db \\
&= \int \frac{p(b,C)}{p(X,C)}db \\
&= \frac{1}{p(X|C)}.
\end{aligned}$$

Also, $m(x,b,c) = m(x,c)$ due to (G2) so the influence function can be equivalently written as

$$\psi_{RA}(Y,X,B,C) = m(x,C)\left(1 - \frac{I(X=x)}{p(x|B,C)}\right) + \frac{I(X=x)}{p(x|B,C)}Y$$

As a result,

$$\begin{aligned}
\mathbb{E}(\psi_{DR}(X,Y,B,C)|X,Y,C) &= m(x,C)\mathbb{E}\left(1 - \frac{I(X=x)}{p(X|B,C)}|X,Y,C\right) + Y\mathbb{E}\left(\frac{I(X=x)}{p(X|B,C)}|X,Y,C\right) \\
&= m(x,C)\left(1 - \frac{I(X=x)}{p(x|C)}\right) + \frac{I(X=x)}{p(x|C)}Y \\
&= \psi_{RA}(Y,X,C).
\end{aligned}$$

By the law of total variance, for any random variables S,T, $\mathsf{Var}(S) = \mathsf{Var}(\mathbb{E}(S|T)) + \mathbb{E}(\mathsf{Var}(S|T))$, which implies

$$\mathsf{Var}(S) \geq \mathsf{Var}(\mathbb{E}(S|T)).$$

We conclude that

$$\mathsf{Var}(\psi_{DR}(X,Y,B,C)) \geq \mathsf{Var}(\mathbb{E}(\psi_{DR}(X,Y,B,C)|X,Y,C)) = \mathsf{Var}(\psi_{DR}(X,Y,C)).$$

□