On Associative Confounder Bias

Priyantha WIJAYATUNGA

Abstract. Conditioning on some set of confounders that causally affect both treatment and outcome variables can be sufficient for eliminating bias introduced by all such confounders when estimating causal effect of the treatment on the outcome from observational data. It is done by including them in propensity score model in so-called potential outcome framework for causal inference whereas in causal graphical modeling framework usual conditioning on them is done. However in the former framework, it is confusing when modeler finds a variable that is non-causally associated with both the treatment and the outcome. Some argue that such variables should also be included in the analysis for removing bias. But others argue that they introduce no bias so they should be excluded and conditioning on them introduces spurious dependence between the treatment and the outcome, thus resulting extra bias in the estimation. We show that there may be errors in both the arguments in different contexts. When such a variable is found neither of the actions may give the correct causal effect estimate. Selecting one action over the other is needed in order to be less wrong. We discuss how to select the better action.

Keywords. causal effect estimation, confounder selection, graphical models.

1. Introduction

For making causal inferences from observational data (see [1] and [2]) it is important to find, ideally all the potential pretreatment confounders of the given causal relation between the cause (treatment variable) and the effect (outcome variable), in order to obtain unbiased causal effect estimate of the former on the latter. Let \( Z \) denote the treatment received by subjects, taking values from the set \( \mathcal{Z} = \{0, 1\} \) and let \( Y \) denote the outcome, taking values from the set \( \mathcal{Y} = \{0, 1\} \) where 0 denotes failure and 1 denotes success. In potential outcome causal model \([1]\) it is accepted existence of pair of potential outcomes \((Y_1, Y_0)\) for each subject, where \( Y_i \) is the outcome that would have been observed had the treatment been \( Z = i \) for \( i = 0, 1 \). It is assumed that the pair is independent of the treatment assignment, written as \((Y_0, Y_1) \perp \!\!\!\!\perp Z\) when the treatment assignments are randomized as in case of a randomized experiment. However in observational studies, the treatment assignments are not randomized. Then, useful assumption for causal inference is that the potential outcomes are conditionally independent of the treatment assignment given the pretreatment covariates, say, multivariate \( X \). Ideally, \( X \) denotes ‘all’ the potential pre-treatment confounders of \( Z \) and \( Y \) and then it is written as \((Y_1, Y_2) \perp \!\!\!\!\perp Z \mid X \). That is, to estimate the causal effect of \( Z \) on \( Y \), we need to condition on (control for) \( X \). However,
it is not necessary to consider all the pretreatment confounders but any 'sufficient' subset of them. Finding such a sufficient set of confounders is somewhat problematic and the potential outcome framework offers no clear way to do it. However, causal graphical modeling framework [2] offers one way that is called 'back door criterion'. It shows how to choose a subset of covariates in order to identify the causal effect (to estimate it without bias). When a causal graphical model is identified on $Z\), $Y$ and all their causal factors the criterion can find a sufficient subset and such a set is called an 'admissible' or a 'deconfounding' set in the literature. Considering some covariates as confounders by ignoring such a criterion can sometimes introduce further bias (p. 351 of [2]). However, the back-door criterion is not complete [2]; there exist causal graphical models where the criterion fails for some sets of covariates though adjusting for them results in valid causal effect estimates.

So, the problem of confounder selection is important in casual inference. In the potential outcome causal model, when the analyst has found all the confounders then he/she uses them either directly or indirectly (in so-called propensity score models [3], [4]) for removing induced bias from them. However, any factor that is causing both the treatment and the outcome could be identified relatively easily as pretreatment confounders with subject domain knowledge. For that it is important to decide causal directions among the variables. But there may be other factors such as the ones that are non-causally related with both the treatment and the outcome, for e.g., those with associations. It seems that some researchers tend to use them for conditioning too, for e.g., including them in the propensity score model assuming that it removes the bias due to them. However generally, the causal graphical modelers do not consider them as confounders. Recently there was a debate (see [5] [6], [7] and [8]) on this issue; if it is necessary to condition on a variable that is not causally related with both the treatment and the outcome but associated with both. In the debate, Rubin argues for and Pearl and his colleagues argue against saying that it will only introduce extra bias. Our goal here is to analyze these arguments a little more deeper and to understand when we should condition on them. We use graphical modeling framework to estimate causal effects therefore, we begin by giving some details of it. We argue that in some cases, it is desirable to condition whereas in others, it is not. Mostly the decision should be taken considering strengths of associations of the potential confounder with the treatment and the outcome.

2. Covariate Selection for Adjustment of Confounding

We use concept of intervention in causal graphical models (also called do-calculus) described in [2] and [9] for the causal effect estimation. This approach is equivalent to the potential outcome model (see Ch. 7 of [2] and [10]). To recall the reader with this calculus, first define the probability distribution of a random variable with conditioning by intervention or action on another variable. For an observed random data sample on a vector of random variables, say, $X = (X_1, ..., X_\alpha)$, we can find the joint probability distribution of them, say, $p(X = x) = p(x)$. We can have a factorization of $p(x)$; let it be $p(x) = \prod_i p(x_i | pa_i)$ where $PA_i \subseteq \{X_1, ..., X_{i-1}\}$ with the exception of $PA_1 = \emptyset$ (empty set) using some conditional independence assumptions within $X$. Note that here we denote random variables (or sets of them) by uppercase letters/expressions (such as $X, PA$, etc.) and their values by relevant lowercase expressions ($x, pa$, respectively). For a causal
structure on $X$ one can use, for e.g., time order of happening to index the variables such that cause variables have higher indices than those of effect variables. For any $i$, such that $2 \leq i \leq n$ if $p(x_i|pa_i) \neq p(x_i)$ then the probability distribution of vector of random variables without $X_i$, say, $X_{-i} = \{X_1,\ldots,X_n\} \{X_i\}$ when $X_i$ is intervened to a particular value of it, say, $x_i$, written as $do(X_i = x_i)$, denoted by $p(x_{-i}|do(X_i = x_i))$ is defined as follows:

$$p(x_{-i}|do(X_i = x_i)) = \frac{p(x_i)}{p(x_i|pa_i)} = \prod_{k=1,k \neq i}^{n} p(x_i|pa_k)$$

where the last expression is corresponding conditional probability distribution when we have observed $X_i = x_i$. That is, generally two probability distributions differ.

Now, consider two different causal relationships between $X$, $Y$ and $Z$: the first one is such that $X$ is a cause of both $Z$ and $Y$, and $Z$ is a cause of $Y$ which is represented as causal network model $p(y,z,x) = p(x)p(z|x)p(y|x,z)$ shown by left hand side diagram and the second one is such that $X$ and $Z$ are causes of $Y$ which is represented as a causal network model $p(y,z,x) = p(x)p(z)p(y|x,z)$ shown by right hand side diagram in the Figure 1. And if we intervene on $Z$ as $do(Z = z)$ for $z = 0,1$, then marginal intervention distribution of $Y$ for the first causal model is $p(y|do(Z = z)) = \sum_x p(x)p(z|x)p(y|x,z)/p(z|x) = \sum_x p(y|x,z)p(z|x) p(x)$ whereas that for the second causal model is $p(y|do(Z = z)) = \sum_x p(x)p(z)p(y|x,z)/p(z) = \sum_x p(y|x,z)p(x) = p(y|x)$, since $X \perp Z$ in latter case. And the causal effect of the treatment option $Z = z_1$ compared to the control option $Z = z_0$ is defined as $\sum_x y p(y|do(Z = z_1)) - \sum_x y p(y|do(Z = z_0))$. It is identifiable if $\sum_x p(y|x,z)p(x)$ is a valid functional for $z = z_0,z_1$. Then we see that the estimates for the two cases are different.

The above observation can be shown for a more general causal model. Let $X = (X_1,\ldots,X_n)$ be according to time order and $X_0$ represent a set variables that causally affect $X_p$ but we are not sure about chronological order of the elements of $X_0$ with $X_i$ for $1 \leq i < p \leq n$. Let parents (causes) of $X_j$ in $X$ be $PA_j$ and that in $(X_0,X)$ be $PA_j^+$, so $PA_j = PA_j^+$ for $j \leq p - 1$ and $PA_p^+ = PA_p \cup X_0$. Then, the joint probability distribution of $(X,X_0)$ is $p(x_0,x) = p(x_0) \prod_{j=1}^{n} p(x_j|pa_j^+)$ and the intervention (on $X_i$) distribution is

$$p(y,z,x) = p(x)p(z|x)p(y|x,z)$$

$$p(y,z,x) = p(x)p(z)p(y|x,z)$$

Figure 1. Two Bayesian network causal models.
\[
p(x_p|do(X_i = x_i)) = \sum_{x_0, x_1, \ldots, x_i, x_{i+1}, \ldots, x_n} p(x_0) \prod_{j=1}^{n} p(x_j|pa_j^+) \\
= \sum_{x_0, x_1, \ldots, x_i, x_{i+1}, \ldots, x_n} p(x_0) \prod_{j=1}^{p-1} p(x_j|D_j)p(x_p|D_p, x_0) \\
= \sum_{x_0, x_1, \ldots, x_i, x_{i+1}, \ldots, x_p-1} \sum_{x_{i+1}, \ldots, x_n} p(x_0)p(x_1, \ldots, x_{i-1})p(x_{i+1}, \ldots, x_p|D_{i+1}, x_0) \\
= \sum_{x_0, pa_p} p(x_0)p(pa_p \setminus \{x_i\})p(x_p|pa_p, x_0) \\
= \sum_{x_0, pa_p} p(x_p|x_i, pa_p \setminus \{x_i\}, x_0)p(pa_p \setminus \{x_i\}, x_0)
\]

where \(D_j = \{X_1, \ldots, X_{j-1}\}\) such that \(D_1 = \emptyset\) and \(X_i \in PA_p\). This is of the form of \(p(y|do(z)) = \sum_z p(y|z, x)p(x)\) where \(Z\) and \(X\) affects \(Y\) directly and so is \(X\) on \(Z\). If we assume that some of the variables in \(X_0\) are associated with some of the variables in the vector \(\{X_1, \ldots, X_{i-1}\}\) or, causally related or associated with variables in \(\{X_{i+1}, \ldots, X_n\}\) then the above result holds.

\[
p(x_p|do(X_i = x_i)) = \sum_{x_0, pa_p} \frac{p(x_p, x_i, pa_p \setminus \{x_i\}, x_0)}{p(x_i|pa_p \setminus \{x_i\}, x_0)}p(x_i|pa_p) \\
= \sum_{x_0, pa_p} \frac{p(x_p, x_i, pa_p \setminus \{x_i\}, x_0)}{p(x_i|pa_p \setminus \{x_i\}, x_0)} = \sum_{x_0, pa_p} \frac{p(x_p, x_i, pa_p \setminus \{x_i\}, x_0)}{p(x_i|pa_p)}p(x_i|pa_p) \\
= \sum_{x_0, pa_p} \frac{p(x_p, x_i, pa_p)}{p(x_i|pa_p)} = \sum_{x_0, pa_p} p(x_p|x_i, pa_p) p(pa_p)
\]

where \(PA_p^- = PA_p \cap PA_i\). Again, this is in the form of \(p(y|do(z)) = \sum_z p(y|z, x)p(x)\) where \(X\) represents all the direct causal variables common to both \(Z\) and \(Y\). And above simplifications show that we can select the confounding variable set as follows.

**Proposition 1** Let \(X''\) denote the set of all potential causal variables of \(Y\) except for \(Z\) and let \(P(Y|Z, X'') = P(Y|Z, X')\) where \(X'\) is the smallest subset of \(X''\) in some sense. Then the smallest subset of \(X'\) of \(X''\) in a similar sense such that \(P(Z|X') = P(Z|X)\) is a sufficient set of confounders for estimating \(P(Y|do(Z))\).

Here the smallest subset \(A\) of \(X\) can be a set of variables whose sum of their configurations is the smallest. This rule gives a simple way to select covariates for removing confounding bias. We avoid the proof of this rule but it is clear from the above discussion. Recall that the back-door criterion is known to be incomplete (see Ch. 11 of [2]) meaning that the criterion fails for some sets of covariates but adjusting for them is sufficient for removing confounding bias. Above rule avoids inclusion of covariates such as instrumental variables, especially for building propensity score models. In fact, in literature sufficient confounder set is selected such that, firstly each confounder in it is a cause of the treatment, and then it is a cause of the outcome [11]. However, it should be
done in the other way round; a confounder should be predictive of the outcome first and then it should also predictive of the treatment. Following this order we do not miss any important confounders, since any confounder should be related to the outcome at the first place. For e.g., consider a causal model for estimating causal effect of teacher’s instructional practice (Z) on student’s reading comprehension achievement (Y) as discussed in [12]. It is assumed that the teacher’s reading knowledge X is a causal confounder such that it affects directly both Z and Y. Furthermore, it is assumed that the teacher’s professional development in reading U affects directly to Z and X and the teacher’s general knowledge (W) affects directly to X and Y. The causal diagram is shown as Model 1 in Figure 2. Then it is easy to see that \( p(y|do(z)) = \sum_x p(y|z, x)p(x) \). And if we believe that U and W are dependent, for e.g., through a common cause, then we get that \( p(y|do(z)) = \sum_{x, w} p(y|z, x, w)p(x, w) \) when \( Z \perp W \), that is reasonable to assume.

3. Associative Confounders

There is a controversy among the research community about kinds of variables that should be considered as confounders for including, especially in the propensity score models in the potential outcome causal model, since therein the causal diagrams showing the causal structure are often not used. In fact, initially the propensity score concept came into light to describe the treatment allocation process [4], [3]. In the current practice some authors argue that all the variables related to outcome should be included in the propensity score model [13] (there can be some redundancy then) whereas others argue that all the variables related to both the treatment and the outcome should be included [14]. However, the problems occur when one finds variables that have non-causal (associative) relationships with the treatment or the outcome. Researchers usually replace any such association between two variables with a causal fork using so-called common cause principle. This is to replace an association with causal relations [7]. Simply, the principle says that a non-causal association between two variables can be replaced by a third variable that is causally affecting the both. For e.g., such an association between two variables X and Y with a model, say, \( M_1: X \rightarrow Y \) can be replaced by a model, say, \( M_2: X \leftarrow U \rightarrow Y \) where arrows indicate causal relations. Then, \( U \) is said to be a common cause of X and Y. Note that we omit the possibility of having feedback causal relations.

Now, let we observe a covariate X that is non-causally associated with both Z and Y, which is the topic of Rubin and Pearl debate. It can be assumed that the non-causal as-
association structures $Z \rightarrow X \rightarrow Y$, is embedded in the context and therefore, apply a causal fork to each of the two associations separately. In fact, the argument of [7] and [6] is based on applying two causal forks for the two non-causal associations, one for each, thus making $X$ a so-called M-collider [12]. Their model of discussion is the Model A in Figure 3 but the argument is based on the model $Z \leftarrow U \rightarrow X \leftarrow W \rightarrow Y$ that is called an M-structure due to its shape. Here $U$ and $W$ are taken to be independent. An example of this model is given in [13]: measuring causal effect of low education ($Z$) on later diabetes risk ($Y$) where it is assumed that mother’s previous diabetes status ($X$) is an associative covariate. A medical opinion is that family income during the childhood ($U$) is a cause of $X$ and $Z$, and mother’s genetic risk of diabetes ($W$) is a cause of $X$ and $Y$.

Though $U$ and $W$ can be independent, it is appropriate to think that it is a special case and generally, there is some dependence between them. In fact, one can just assume it but here we investigate how and when such cases arise and discuss which actions are appropriate then. For Model B of Figure 3 we can write the joint probability distribution of all the variables as $p(y|x,u,w,z) = p(u)p(w)p(x|u,w)p(z|u)p(y|z,w)$, so with intervention $Z = do(z)$, we get $p(y|x,u,w|do(z)) = p(u)p(w)p(x|u,w)p(y|z,w)$. Then,

$$p(y|do(z)) = \sum_{u,w,x} p(u)p(w)p(x|u,w)p(y|z,w) = \sum_{w} p(w)p(y|z,w) = \sum_{w} p(w|z)p(y|z,w)$$

$$= p(y|z) = \sum_{x} p(y|z,x)p(x|z) \neq \sum_{x} p(y|z)$$

if $W \perp Z$ and, since $Z \perp X$. Note that we have $W \perp Z$ whenever $U \perp W$. That is, the true probability of $Y$ when $Z$ is intervened is different from that obtained by conditioning on $X$. And ignoring $X$ gives the true intervened probability. So, when assuming $U \perp W$, conditioning on $X$ may result in a biased causal effect estimate; above inequality shows that the biasness may have caused due to the dependence between $X$ and $Z$, since $p(x) \neq p(x|z)$, i.e., when $X$ and $Z$ are weakly dependent the biasness is small. Note that if an error is occurred in the estimate of $p(y|do(z))$ for $z = 0, 1$ then it may not necessarily result in an error of same magnitude in causal effect estimate that is $\sum_{x} yp(y|do(Z = 1)) - \sum_{x} yp(y|do(Z = 0))$, i.e., two errors may result in a different error. Resultant error (bias) due to conditioning on $X$ is $\sum_{x} yp(y|Z = 1, x)[p(x) - p(x|Z = 1)] - \sum_{x} yp(y|Z = 0, x)[p(x) - p(x|Z = 0)]$. For simplicity, we concentrate on errors that can occur in estimation of $p(y|do(z))$ for $z = 0, 1$. Note that, above discussion is valid when some of other confounders, say, $X_1$ are present where $X \perp X_1$. And in the above analysis we made a strong assumption that $W \perp Z$, but this may not sometimes be true in reality. In the following section we show this possibility.

3.1. Dependence of $X$ with $Z$ and $Y$

It is natural to consider the cases when $W \perp Z$ then it may be that $p(y|do(z)) \neq p(y|z)$ even if $U \perp W$. However, since $W$ is hidden it is unclear how to consider this case. In fact, for Model B in Figure 3 we have $W \perp Z$ [16]. Let us assume the case that $X$ and $Y$ are strongly dependent. We use a geometric figure that is used to visualize the Simpson’s paradox [17] to explore this possibility. Let the association between $Y$ and $X$ be such that $p(x|y) < p(x|y')$. Then, for some $T$ we have that $p(t'|y)p(x|y,t') + p(t|y)p(x|y,t) < p(t'|y')p(x|y',t') + p(t|y')p(x|y',t)$. Note that there are infinitely many such $T$ but they can be artificial unless given some meaningful interpretation, ideally to few of them. Now
consider the case of $p(x|y, t') < p(x|y, t)$. It is important to note that the value $p(x \mid y)$ dissects positive length $p(x|y, t) - p(x|y, t')$ according to ratio $p(t|y): p(t'|y)$:

$$\frac{p(t \mid y) + p(t' \mid y)}{p(t \mid y) + p(t' \mid y)p(x \mid y) + p(t \mid y)p(x \mid y, t) - p(t \mid y)p(x \mid y, t')}{p(t \mid y)\{p(x \mid y) - p(x \mid y, t)} = p(t' \mid y)\{p(x \mid y) - p(x \mid y, t')} \right)$$

Now if $T$ is a common cause of $X$ and $Y$ association then we should have $p(x|y, t) = p(x|y, t') = p(x|t)$ and $p(x|y, t') = p(x|y', t') = p(x|y')$. Therefore, in Figure 4 those ratios are marked with braces. Since the selection of $T$ is restricted by the strength of the dependence between $X$ and $Y$, for a higher value of it, we can have a higher dependence between $Y$ and $T$. And if the strength of the dependence between $X$ and $Y$ is characterized by $p(y|x)$ then that between $X$ and $T$ should be also higher. And the other case is similar, i.e., taking $p(x|y, t') > p(x|y, t)$.

If $T$ is a common cause of the association between $X$ and $Y$, then the dependences between $X$ and $W$, and $Y$ and $W$ can be strong given that the dependence between $X$ and $Y$ is strong. Similarly, a strong association between $Z$ and $X$ implies that those between $Z$ and $U$, and $U$ and $X$ can be strong. With similar arguments, these imply that $U$ and $W$ can be dependent. An alternative way to see that $U$ and $W$ are not independent when the associations between $X$ and $Z$, and $X$ and $Y$ are strong is to use correlations. In Figure 5 it is shown that for any three random variables, say, $A, B$ and $C$ the correlation coefficients among them satisfy the relationship $\rho_{AC}^2 + \rho_{BC}^2 + \rho_{AB}^2 \leq 1 + 2\rho_{AC}\rho_{BC}\rho_{AB}$. If, for e.g., when $\rho_{XZ} = 0.8$ and $\rho_{XY} = 0.7$ then we cannot have $U$ and $W$ such that $\rho_{UW} = 0$. Therefore, when the dependences between $Z$ and $X$, and $Y$ and $X$ are strong it may be that the introduced two common causes for those associations are dependent. Furthermore, there can be another possibility for these two associations; both associations may be due one cause, i.e., both $U$ and $W$ refer to the same hidden variable ($V$ in the Model C in Figure 5).

However, current studies are often done without considering these possibilities. But some researchers have shown that conditioning on associated covariates introduces only
Figure 4. A common cause variable \( T \) for the negative correlation between \( X \) and \( Y \), \( p(x|y) < p(x'|y') \). For the probabilities \( p \) and \( q \) where \( p + q = 1 \) the expression \( \{ p(t|y) \} : \{ p(t'|y) \} \) means that the lengths of two line segments on which \( p \) and \( q \) appear are according to the ratio \( p : q \).

Figure 5. Common cause models

a small bias. Their claims may be due to these contexts. Sometimes it is advised \cite{8} to control for all the pretreatment covariates but the graphical causal model researchers reject this idea. Therefore, in the next section we take a look at different possibilities of associative covariates and try to understand when the biasness can be amplified.

3.2. Deciding on Conditioning

Consider the case of two dependent hidden causes, i.e., \( U \perp\!\!\!\!\perp W \) (such as Model D in Figure 5) where the dependence is causal or non-causal. Then,
\[ p(y|do(z)) = \sum_{w,z} p(u,w)p(x|u,w)p(y|z,w) = \sum_w p(w)p(y|z,w) = \sum_{w,x} p(x)p(w|x)p(y|w,z,x) \]

\[ \neq \sum_{w,x} p(x)p(w|x)p(y|w,z,x) = \sum_x p(y|x)p(x) \]

if \( p(w|x) \neq p(w|z,x) \) i.e., \( W \perp \perp Z|X \), conditioning on \( X \) does not give the correct probability estimate that is \( \sum_w p(y|z,w)p(w) \). And ignoring \( X \) also does not give the correct estimate, since then we get \( p(y|do(z)) = p(y|z) = \sum_w p(y|z,w)p(w|z) \neq \sum_w p(y|z,w)p(w) \), i.e., we need to assume \( Z \perp \perp W \) in order to have the correct probability for the case but we know that \( Z \nmid W \), especially when associations between \( X \) and \( Y \), and \( Z \) and \( Y \) are strong. That is, to condition on \( X \) we should have \( W \perp \perp Z|X \) and to ignore \( X \) we should have \( W \perp Z \). So, the question remains is that which statement should be accepted in order to be more correct against the other; either \( W \perp Z|X \) or \( W \perp Z \). Accepting the former (rejecting the latter) is to condition on \( X \) and vice versa. But none of the conditions can be tested, since they involve unobservable \( W \).

However, with some subject domain knowledge if one can assume meaningful \( U \) and \( W \) and then recognize their dependences with \( X \) (based on those between \( Z \) and \( X \), and \( Y \) and \( X \)) it may be possible to decide which option can be better. For e.g., if those dependences are not strong and causation of \( U \) and \( W \) on \( X \) is mostly based on explaining away phenomenon \([19]\), then it may not be desirable to condition of \( X \). Note that the explaining away phenomenon is that when we see \( X = 1 \) then observing \( U = 1 \) makes \( P(W = 1) \) lower and vice versa. If conditioned on \( X \) in this case, then comparative strata of data sample in the causal effect calculation may have imbalances in the causal variables \( U \) and \( W \). This can cause biased causal effect estimates. And when the dependences of \( U \) and \( W \) with \( X \) is assumed to be high then it is less likely that there is an explaining away phenomenon, i.e., most probably the causation is monotonic (when we see \( X = 1 \) then observing \( U = 1 \) makes \( P(W = 1) \) higher and vice versa) then conditioning on \( X \) can be beneficial because it results in imbalances in the causal variables \( U \) and \( W \). Though one can reason about the actions to be taken as done above, it requires extensive simulation studies to confirm them.

Now consider the case of single hidden cause, say, \( V \) (Model C in Figure 5). Then

\[ p(y|do(z)) = \sum_{x,v} p(v)p(x|v)p(y|z,v) = \sum_v p(y|z,v)p(v) = \sum_{x,v} p(x)p(v|x)p(y|z,v,x) \]

\[ \neq \sum_{x,v} p(x)p(v|x)p(y|z,v,x) = \sum_x p(y|z,x)p(x) \]

Therefore, here also conditioning on \( X \) does not gives the correct probability estimate that is \( \sum_v p(y|z,v)p(v) \) if \( p(v|x) \neq p(v|z,x) \), i.e., \( V \perp \perp Z|X \). But ignoring \( X \) also does not give the correct estimate as \( p(y|do(z)) \neq p(y|z) \) in this case. Since \( p(y|z) = \sum_v p(y|z,v)p(v|z) \), ignoring \( X \) means assuming \( V \perp Z \), but we know that \( Z \) and \( V \) should be dependent. So, similar to the above case, the question remains is that which should be accepted against the other in order to be more correct; either \( V \perp Z|X \) or \( V \perp Z \). Accepting the former is to condition on \( X \) and vice versa. But similar to the above case where the dependences are higher, assuming \( V \perp Z|X \) can be better than assuming \( V \perp Z \), therefore conditioning on \( X \). If the subject domain knowledge shows that there is a single common cause \( V \) then it is beneficial to condition on \( X \).
4. Conclusion

Causal effect estimation tasks from observational data need to consider confounders of the causal relation of interest for controlling for (conditioning on). However, it is not necessary that all of them are considered but a “sufficient” subset of them. Often the current practice is to select them according to their predictive ability of the treatment firstly and then the outcome. But it should be done otherwise; firstly they should be predictive of the outcome and then the treatment. And we show how to handle associative confounders (those are not causing both the treatment and outcome but associated with them) where currently there is no clear consensus about using them. It is often beneficial to condition on associative confounders when they are strongly dependent with both the treatment and outcome whereas it is not so for weakly dependent ones.

Acknowledgments: Financial support for this research is from the Swedish Research Council for Health, Working Life and Welfare (FORTE) and SIMSAM at Umeå. And the author is thankful to Slawomir Nowaczyk and anonymous referees for their comments.

References

[1] D. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association 100(469) (2005), 322–331.
[2] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, New York, 2009.
[3] P. R. Rosenbaum and D. B. Rubin, The central role of the propensity score in observational studies for causal effects. Biometrika 70(1) (1983), 41–55.
[4] P. R. Rosenbaum and D. B. Rubin, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association 79(387) (1984), 516–524.
[5] I. Shrier, Letter to the Editor, Statistics in Medicine 28 (2009), 1315–1318.
[6] J. Pearl, Letter to the Editor, Statistics in Medicine 28 (2009), 1415–1416.
[7] A. Sjölander, Letter to the Editor, Statistics in Medicine 28 (2009), 1416–1420.
[8] D. Rubin, Author’s Reply, Statistics in Medicine 28 (2009), 1420–1423.
[9] S. L. Lauritzen and T. S. Richardson, Chain graph models and their causal interpretations. Journal of Royal Statistical Society. Series B 64(3) (2002), 321–361.
[10] P. Wijayatunga, Causal Effect Estimation Methods. Journal of Statistical and Econometric Methods 3(2) (2014), 153–170.
[11] T. J. VanderWeele and I. Shpitser, A New Criterion for Confounder Selection. Biometrics 67 (2011), 1406–1413.
[12] B. Kelcey and J. Carlisle, The Threshold of Embedded M Collider Bias and Confounding Bias. Society for Research on Educational Effectiveness Conference, (2011). http://files.eric.ed.gov/fulltext/ED519118.pdf
[13] D. B. Rubin, Matching using estimated propensity scores: related theory and practice. Biometrics 52 (1996), 249–264.
[14] S. M. Perkins, W. Tu, M. G. Underhill, X. H. Zhou and M. D. Murray, M. D. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Safety 9 (2000), 93–101.
[15] L. Dallolio, R. Bellocco, L. Richiardi and M. P. Fantini, M.P. Using directed acyclic graphs to understand confounding in observational studies. Biomedical Statistics and Clinical Epidemiology 3(2) (2010), 89–96.
[16] Lauritzen, S. L., Dawid, A. P., B. N. Larsen, and H.-G. Leimer, Independent Properties of Directed Markov Fields Networks, 20 (1990), 491–505.
[17] P. Wijayatunga, Viewing Simpson’s paradox. Statistica & Applicazioni, XII(2) (2014), 225–235.
[18] E. Langford, N. Schwertman and M. Owens, Is the property of being positively correlated transitive? The American Statistician 55(4) (2001), 322–324.
[19] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Second Edition), Morgan Kauffmann, San Mateo, CA., 1988.