On the Importance of Modeling Unmeasured Confounders in Predictive Inference

Timothy Brathwaite

March 21st, 2020

1 Introduction

In a general prediction problem, the quantity we want to estimate is $P\left(Y^* \mid X^*, X, Y\right)$. Algebraically, this can be written using the rules of probability theory as follows.

$$\begin{aligned} P\left(Y^* \mid X^*, X, Y\right) &= \int P\left(Y^*, \lambda \mid X^*, X, Y\right) \partial\lambda \\ &= \int P\left(Y^* \mid X^*, X, Y, \lambda\right) P\left(\lambda \mid X^*, X, Y\right) \partial\lambda \\ \text{where } Y^* &= \text{The outcome value for a new observation.} \\ X^* &= \text{The explanatory variables for a new observation.} \\ Y &= \text{The historical outcome values.} \\ X &= \text{The historical explanatory values.} \\ \lambda &= \text{Everything we don't know that affects } X \text{ and/or } Y. \end{aligned} \tag{1}$$

At this point, our expression contains two terms. The first term is a model for Y^*, $P\left(Y^* \mid X^*, X, Y, \lambda\right)$. The second is a probability for unknown variables λ, $P\left(\lambda \mid X^*, X, Y\right)$. As written, this expression is not useful because we don't know the form of $P\left(\lambda \mid X^*, X, Y\right)$. To re-express $P\left(\lambda \mid X^*, X, Y\right)$ in terms of quantities we know or can estimate, it is helpful to further specify what λ is. In particular, we can decompose λ as follows. Either:

1. $\lambda = (\tau, \theta)$ or
2. $\lambda = (\tau, \theta, \Omega)$

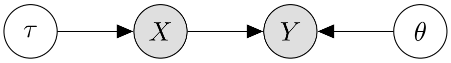
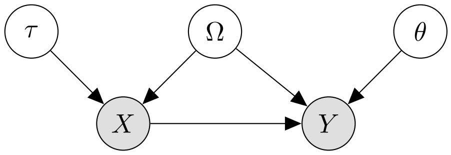
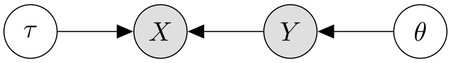
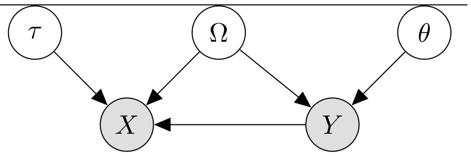
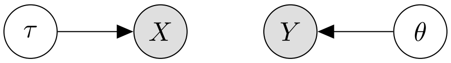
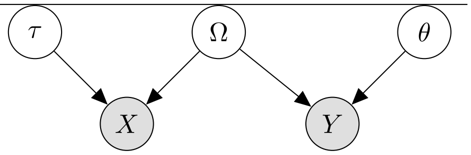
In these scenarios, τ represents unknown variables that only affect X, θ represents unknown variables that affect Y, and Ω represents variables that affect both X and Y.

Likewise, it is useful to specify the type of relationships that can exist between X and Y. We have four types of relationships:

1. X causes Y
2. Y causes X
3. X does not cause Y AND Y does not cause X
4. X causes Y AND Y causes X.

For the moment, we will ignore case four, because it is a mathematically hard case to deal with and because it is uncommon in most practical cases that I deal with. In most problems I face, Y is an outcome of interest that occurs at time t and X is produced at time t' where $t' < t$, so it is not sensible that Y works backwards in time to cause X.

So, combining the three remaining types of relationships between X and Y with the two types of λ, we have the following 6 graphical models.

Case	$\lambda = (\tau, \theta)$	$\lambda = (\tau, \theta, \Omega)$
(i)		
(ii)		
(iii)		

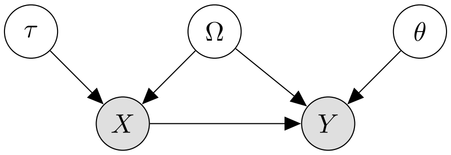
As will be shown in the Section 3, cases where $\lambda = (\tau, \theta)$ correspond to "typical" supervised learning models. Specifically, cases with $\lambda = (\tau, \theta)$ and:

- $X \rightarrow Y$ correspond to discriminative models (e.g. logistic regression).
- $X \leftarrow Y$ correspond to generative models (e.g. naive bayes).
- $X \quad Y$ (e.g. case iii) correspond to "constant" models, where no features are useful for predicting Y.

Scenarios where unobserved confounding is present, i.e. where $\lambda = (\tau, \theta, \Omega)$, are not so commonly discussed in machine learning texts.

2 The Current Scenario: A Discriminative Model with Unobserved Confounding

In particular, the case I want to focus on is depicted by the following graphical model.



In this scenario, there are unmeasured confounders Ω that cause both X and Y, and we also believe that $X \rightarrow Y$. For example, if considering X to be commute travel cost and Y to be transportation mode (e.g. bike, walk, drive, transit), an example of an unobserved confounder might be travel distance if you don't know a user's exact home address. The travel distance will cause one's travel cost and may cause one's travel mode as well (e.g. if a person doesn't want to walk a long distance).

Let's see how we can calculate $P\left(Y^* \mid X^*, X, Y\right)$ in this scenario. Using Equation 1 and the factorization of $P\left(X^*, X, Y \mid \lambda\right)$ implied by the directed acyclic graph (DAG) above, we can continue writing

$$\begin{aligned}
P\left(Y^* \mid X^*, X, Y\right) &= \int P\left(Y^* \mid X^*, X, Y, \lambda\right) P\left(\lambda \mid X^*, X, Y\right) \partial\lambda \\
&= \int P\left(Y^* \mid X^*, \lambda\right) P\left(\lambda \mid X^*, X, Y\right) \partial\lambda \qquad (2) \\
&= \int P\left(Y^* \mid X^*, \Omega, \theta\right) P\left(\lambda \mid X^*, X, Y\right) \partial\lambda
\end{aligned}$$

Next, the term $P(\lambda \mid X^{*}, X, Y)$ can be calculated using Bayes' Rule and by (again) relying on the factorization from the DAG above.

$$\begin{aligned}
P(\lambda \mid X^{*}, X, Y) &= \frac{P(X^{*}, X, Y \mid \lambda)\, P(\lambda)}{P(X^{*}, X, Y)} \\
&= \frac{P(X^{*} \mid \lambda)\, P(X, Y \mid \lambda)\, P(\lambda)}{P(X^{*}, X, Y)} \\
&= \frac{P(X^{*} \mid \lambda)\, P(Y \mid X, \lambda)\, P(X \mid \lambda)\, P(\lambda)}{P(X^{*}, X, Y)} \\
&= \frac{P(X^{*} \mid \tau, \Omega)\, P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)}{P(X^{*}, X, Y)}
\end{aligned} \tag{3}$$

Substituting Equation 3 into Equation 2, we get the following expression.

$$\begin{aligned}
P(Y^{*} \mid X^{*}, X, Y) &= \int \left[\frac{P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)}{P(X^{*}, X, Y)}\right] \partial\lambda \\
&= \frac{\int P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)\, \partial\lambda}{P(X^{*}, X, Y)} \\
&= \frac{\int \left[\frac{P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)}{P(X, Y)}\right] \partial\lambda}{\frac{P(X^{*}, X, Y)}{P(X, Y)}} \\
&= \frac{\int P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(\lambda \mid X, Y)\, \partial\lambda}{\frac{P(X^{*}, X, Y)}{P(X, Y)}} \\
&= \frac{\int P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(\lambda \mid X, Y)\, \partial\lambda}{P(X^{*} \mid X, Y)} \\
&\propto \int P(Y^{*} \mid X^{*}, \theta, \Omega)\, P(X^{*} \mid \tau, \Omega)\, P(\lambda \mid X, Y)\, \partial\lambda \\
&\approx \frac{1}{M} \sum_{i=1}^{M} P(Y^{*} \mid X^{*}, \theta_i, \Omega_i)\, P(X^{*} \mid \tau_i, \Omega_i) \\
\text{where } M &= \text{The number of samples from } P(\lambda \mid X, Y) \\
\lambda_i &= (\tau_i, \theta_i, \Omega_i) = \text{A singe sample from } P(\lambda \mid X, Y) \\
P(\lambda \mid X, Y) &= \frac{P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)}{P(X, Y)}
\end{aligned} \tag{4}$$

The reasoning above suggests that at prediction time, we should use a Monte Carlo estimate of the joint probability of (X^{*}, Y^{*}). We can then normalize across Y^{*} to get a probability in [0, 1].

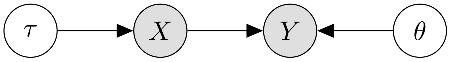
During training, we should estimate the posterior distribution $P(\lambda \mid X, Y)$. Given that Ω is expected to have at least one component per individual and we will only have one or at most a few observations per person in our dataset, we will likely not be in a world where the posterior distribution has concentrated on a single point. Accordingly, it would be best to estimate the entire posterior distribution since there may be wide uncertainty in our posterior beliefs about the value of the unmeasured confounders for some of the observations. However, if one persists in relying on point estimates, one can estimate the maximum a-posterior (MAP) point, which selects λ to maximize the numerator: $P(Y \mid X, \theta, \Omega)\, P(X \mid \tau, \Omega)\, P(\lambda)$.

3 Deriving Known Models

At the end of Section 1, I stated that the situation where $\lambda = (\tau, \theta)$ corresponded to most typically used models such as discriminative and generative models. Below I'll provide a proof of these two points.

3.1 Discriminative Models

Discriminative models can be derived from the following graphical model where $\lambda = (\tau, \theta)$:



The derivation begins with Equation 1 and, using the DAG above, proceeds as follows.

$$\begin{aligned}
P(Y^* \mid X^*, X, Y) &= \int P(Y^* \mid X^*, X, Y, \lambda)\, P(\lambda \mid X^*, X, Y)\, \partial\lambda \\
&= \int P(Y^* \mid X^*, \lambda)\, P(\lambda \mid X^*, X, Y)\, \partial\lambda \\
&= \int P(Y^* \mid X^*, \lambda)\, \frac{P(X^*, X, Y \mid \lambda)\, P(\lambda)}{P(X^*, X, Y)} \partial\lambda \\
&= \int\int P(Y^* \mid X^*, \tau, \theta)\, \frac{P(X^*, X, Y \mid \tau, \theta)\, P(\tau, \theta)}{P(X^*, X, Y)} \partial\tau\partial\theta \\
&= \int\int P(Y^* \mid X^*, \theta)\, \frac{P(X^* \mid \tau)\, P(X, Y \mid \tau, \theta)\, P(\tau, \theta)}{P(X^*, X, Y)} \partial\tau\partial\theta \\
&= \int\int P(Y^* \mid X^*, \theta)\, \frac{P(X^* \mid \tau)\, P(Y \mid X, \theta)\, P(X \mid \tau)\, P(\tau)\, P(\theta)}{P(X^*, X, Y)} \partial\tau\partial\theta \\
&= \int P(Y^* \mid X^*, \theta)\, P(Y \mid X, \theta)\, P(\theta)\, \partial\theta \int \frac{P(X^* \mid \tau)\, P(X \mid \tau)\, P(\tau)}{P(X^*, X, Y)} \partial\tau \\
&= \frac{1}{P(X^*, X, Y)} \int P(Y^* \mid X^*, \theta)\, P(Y \mid X, \theta)\, P(\theta)\, \partial\theta \int P(X^*, X, \tau)\, \partial\tau \\
&= \frac{P(X^*, X)}{P(X^*, X, Y)} \int P(Y^* \mid X^*, \theta)\, P(Y \mid X, \theta)\, P(\theta)\, \partial\theta \\
&= \frac{P(X^*, X)}{\frac{P(X^*, X, Y)}{P(X, Y)}} \int P(Y^* \mid X^*, \theta)\, \frac{P(Y \mid X, \theta)\, P(\theta)}{P(X, Y)} \partial\theta \\
&= \frac{\frac{P(X^*, X)}{P(X)}}{\frac{P(X^*, X, Y)}{P(X, Y)}} \int P(Y^* \mid X^*, \theta)\, \frac{P(Y \mid X, \theta)\, P(X)\, P(\theta)}{P(X, Y)} \partial\theta \\
&= \frac{\frac{P(X^*, X)}{P(X)}}{\frac{P(X^*, X, Y)}{P(X, Y)}} \int P(Y^* \mid X^*, \theta)\, \frac{P(Y \mid X, \theta)\, P(X \mid \theta)\, P(\theta)}{P(X, Y)} \partial\theta \\
&= \frac{P(X^* \mid X)}{P(X^* \mid X, Y)} \int P(Y^* \mid X^*, \theta)\, \frac{P(Y, X, \theta)}{P(X, Y)} \partial\theta \\
&= \frac{P(X^* \mid X)}{P(X^* \mid X, Y)} \int P(Y^* \mid X^*, \theta)\, P(\theta \mid X, Y)\, \partial\theta \\
&= \int P(Y^* \mid X^*, \theta)\, P(\theta \mid X, Y)\, \partial\theta
\end{aligned} \tag{5}$$

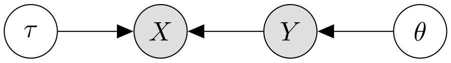
The last line in Equation 5 follows from the following equality:

$$\begin{aligned}
\frac{P\left(X^* \mid X\right)}{P\left(X^* \mid X, Y\right)} &= \frac{\int P\left(X^*, \tau \mid X\right) \partial \tau}{\int P\left(X^*, \tau \mid X, Y\right) \partial \tau} \\
&= \frac{\int P\left(X^* \mid \tau, X\right) P\left(\tau \mid X\right) \partial \tau}{\int P\left(X^* \mid \tau, X, Y\right) P\left(\tau \mid X, Y\right) \partial \tau} \\
&= \frac{\int P\left(X^* \mid \tau\right) P\left(\tau \mid X\right) \partial \tau}{\int P\left(X^* \mid \tau\right) P\left(\tau \mid X\right) \partial \tau} \\
&= 1
\end{aligned} \tag{6}$$

Moreover, the last line in Equation 5 justifies the discriminative learning technique of (1) building a model for Y as a function of X and θ, and (2) taking the expectation of that model's outputs with respect to the posterior distribution of θ. By Equation 5, no other models are required. In particular, no models of X are needed. Note that this contrasts the equivalent scenario with unobserved confounding. There, one needs both a model for Y AND a model for X.

3.2 Generative Models

In this section, I'll show how generative models come from the following graphical model:



First, it is important to note that this graphical model admits the following factorization of the joint probability $P\left(X, Y, \tau, \theta\right)$: $\left\{P\left(X \mid \tau, Y\right), P\left(Y \mid \theta\right), P\left(\tau\right), P\left(\theta\right)\right\}$. As a result, all of our probabilistic expressions for these variables should be written in terms of these basic elements. This requires that we re-express the following probability from Equation 1 as follows:

$$\begin{aligned}
P\left(Y^* \mid X^*, X, Y, \lambda\right) &= P\left(Y^* \mid X^*, \lambda\right) \\
&= \frac{P\left(X^*, Y^* \mid \lambda\right)}{P\left(X^* \mid \lambda\right)} \\
&= \frac{P\left(X^* \mid Y^*, \lambda\right) P\left(Y^* \mid \lambda\right)}{P\left(X^* \mid \lambda\right)} \\
&= \frac{P\left(X^* \mid Y^*, \tau\right) P\left(Y^* \mid \theta\right)}{P\left(X^* \mid \tau, \theta\right)}
\end{aligned} \tag{7}$$

Similarly, we will need the following re-expression:

$$\begin{aligned}
P\left(X^*, X, Y \mid \lambda\right) &= P\left(X^* \mid X, Y, \lambda\right) P\left(X, Y \mid \lambda\right) \\
&= P\left(X^* \mid \lambda\right) P\left(X, Y \mid \lambda\right) \\
&= P\left(X^* \mid \tau, \theta\right) P\left(X, Y \mid \tau, \theta\right) \\
&= P\left(X^* \mid \tau, \theta\right) P\left(X \mid Y, \tau, \theta\right) P\left(Y \mid \tau, \theta\right) \\
&= P\left(X^* \mid \tau, \theta\right) P\left(X \mid Y, \tau\right) P\left(Y \mid \theta\right)
\end{aligned} \tag{8}$$

With these expressions, we can now return to Equation 1 and proceed in the following manner:

$$\begin{aligned}
P\left(Y^{*} \mid X^{*}, X, Y\right) &= \int P\left(Y^{*} \mid X^{*}, X, Y, \lambda\right) P\left(\lambda \mid X^{*}, X, Y\right) \partial\lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right)}{P\left(X^{*} \mid \tau, \theta\right)} P\left(\lambda \mid X^{*}, X, Y\right) \partial\lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right)}{P\left(X^{*} \mid \tau, \theta\right)} \frac{P\left(X^{*}, X, Y \mid \lambda\right) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial\lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right)}{P\left(X^{*} \mid \tau, \theta\right)} \frac{P\left(X^{*} \mid \tau, \theta\right) P\left(X, Y \mid \tau, \theta\right) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial\lambda \\
&= \int P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right) \frac{P\left(X, Y \mid \tau, \theta\right) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial\lambda \\
&= \int P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right) \frac{\frac{P\left(X, Y \mid \tau, \theta\right) P(\lambda)}{P(X, Y)}}{\frac{P\left(X^{*}, X, Y\right)}{P(X, Y)}} \partial\lambda \qquad (9) \\
&= \int P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right) \frac{P\left(\tau, \theta \mid X, Y\right)}{P\left(X^{*} \mid X, Y\right)} \partial\lambda \\
&\propto \int P\left(X^{*} \mid Y^{*}, \tau\right) P\left(Y^{*} \mid \theta\right) P\left(\tau, \theta \mid X, Y\right) \partial\lambda \\
\text{where } P\left(\tau, \theta \mid X, Y\right) &= \frac{P\left(X, Y \mid \tau, \theta\right)}{P(X, Y)} \\
&= \frac{P\left(X \mid Y, \tau\right) P\left(Y \mid \theta\right) P(\lambda)}{P(X, Y)}
\end{aligned}$$

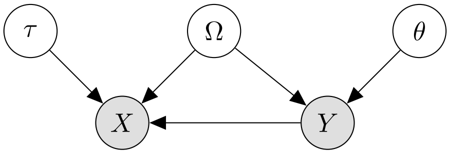
The last three lines of Equation 9 justify the typical generative model. First, models are built for $P(X \mid Y, \tau)$ and $P(Y \mid \theta)$. From these two models we can estimate the posterior distribution of λ: $P(\tau, \theta \mid X, Y)$. Secondly, given samples from $P(\tau, \theta \mid X, Y)$, we can use the two aforementioned models to construct a Monte Carlo estimate of $P(X^{*}, Y^{*} \mid X, Y)$. Finally, we can normalize across Y^{*} to calculate $P(Y^{*} \mid X^{*}, X, Y)$.

Here, as in the case of a discriminative model with unobserved confounding, we need both a model of X AND a model of Y.

4 A Generative Model with Unobserved Confounding

This case is conceptually very similar to the generative model that was just presented. As a result I will proceed by analogy and provide any missing derivations only as requested by others.

First, the graphical model that depicts this scenario is



Secondly, the desired predictive distribution is:

$$
\begin{aligned}
P\left(Y^{*} \mid X^{*}, X, Y\right) &= \int P\left(Y^{*} \mid X^{*}, X, Y, \lambda\right) P\left(\lambda \mid X^{*}, X, Y\right) \partial \lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right)}{P\left(X^{*} \mid \tau, \theta, \Omega\right)} P\left(\lambda \mid X^{*}, X, Y\right) \partial \lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right)}{P\left(X^{*} \mid \tau, \theta, \Omega\right)} \frac{P\left(X^{*}, X, Y \mid \lambda\right) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial \lambda \\
&= \int \frac{P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right)}{P\left(X^{*} \mid \tau, \theta, \Omega\right)} \frac{P\left(X^{*} \mid \tau, \theta, \Omega\right) P(X, Y \mid \tau, \theta, \Omega) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial \lambda \\
&= \int P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right) \frac{P(X, Y \mid \tau, \theta, \Omega) P(\lambda)}{P\left(X^{*}, X, Y\right)} \partial \lambda \\
&= \int P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right) \frac{\frac{P(X, Y \mid \tau, \theta, \Omega) P(\lambda)}{P(X, Y)}}{\frac{P\left(X^{*}, X, Y\right)}{P(X, Y)}} \partial \lambda \\
&= \int P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right) \frac{P(\tau, \theta, \Omega \mid X, Y)}{P\left(X^{*} \mid X, Y\right)} \partial \lambda \\
&\propto \int P\left(X^{*} \mid Y^{*}, \tau, \Omega\right) P\left(Y^{*} \mid \theta, \Omega\right) P(\tau, \theta, \Omega \mid X, Y) \partial \lambda \\
\text{where } P(\tau, \theta, \Omega \mid X, Y) &= \frac{P(X, Y \mid \tau, \theta, \Omega)}{P(X, Y)} \\
&= \frac{P(X \mid Y, \tau, \Omega) P(Y \mid \theta, \Omega) P(\lambda)}{P(X, Y)}
\end{aligned} \tag{10}
$$

As shown by the last three lines of Equation 10 versus the last three lines of Equation 9, there is little mathematical difference between generative models with and without unobserved confounding. Moreover, the prediction equation for a generative model with confounding and a discriminative model with confounding are symmetric: switch X and Y and switch τ and θ.