Estimating Causal Effects of Actors on Movie Revenues

December 14, 2019

Abstract

This document provides a detailed report of the independent study undertaken over the semester of Fall 2019 building over Wang and Blei's study "The Blessings of Multiple Causes". This covers the initial motivation for the problem, using related work for a solution, attempting a different approach using Pyro, and evaluating results as claimed in prior work.

Author: Siddhesh Acharekar

Advisors: Dr. Olga Vitek | Dr. Robert Ness

---

1 Summary

Causal Inference from observational data assumes "ignorability": that all confounders are observed. This is a standard, yet untestable assumption. Also many real-world studies involve multiple causes towards one outcome, different variables whose effects are all of interest. This projected attempts to solve one such problem of Estimating Causal Effects of Actors on Movie Revenues by using a factor model to sidestep the search for confounders and does so while maintaining all model assumptions as one data generative process using Pyro, a probabilistic programming framework on Pytorch, enabling us to run valid causal queries on the same model.

2 Introduction

Imagine a simple but lucrative causal inference problem: a movie producer wants to find out how much money a movie will make given a cast of actors he's chosen to for it. Put it another way, the producer is interested in the causal effect of each actor; for example: how much movie will a movie having Brad Pitt, Scarlett Johansson and Ben Affleck make? Or by how much will revenue increase or decrease if Matt Damon is in this movie? So we set out to gather a dataset that can answer this question. We scrape IMDB and obtain cast and revenue data for all movies made after 1980. Then retain only those which made at least USD 1 Million in revenue and retain actors that have made at least 20 movies.

Movie	Cast	Revenue
Avatar	Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ...	$2.7 Billion

Movie	Cast	Revenue
Pirates of the Caribbean: Worlds End	Johnny Depp, Keira Nightly, Orlando Bloom, ...	$963 Million
Spectre	Daniel Craig, Lea Seydoux, Christoph Waltz, ...	$880 Million

We can solve this using a standard machine learning approach and treat it as a regression problem. Let's go with Linear Regression wherein we attribute regression coefficients as causal effects of actors. The DAG we assume here is:

Treating every row $i \in \{1...n\}$ for n movies as a vector of actors $X_i = \{x_1, x_2, ..., x_m\}$ where x_m is a binary variable indicating presence or absence $(1, 0)$ of actor x_m, and revenue as Y_i for movie i; our aim with this Regression model is to estimate $\mathbb{E}[Y|X = x]$ for a new vector of actors x the producer wants to know revenue for. Our dataset is transformed as:

Movie	Adam Sandler	Alec Baldwin	Amy Adams	...	Revenue
Avatar	0	0	0	...	2700000000
Pirates of the Caribbean: Worlds End	0	0	0	...	963000000
Spectre	0	0	1	...	880000000

The problem with this approach is that these estimates may not be accurate. There are always, as with most other causal inference problems, *unobserved confounders*: variables that affect both the causes X_i and outcome Y_i. In the presence of unobserved confounders, causes are correlated with the outcome. For example, with the problem we're trying to address, the *genre* of a movie can be a confounder that affects both its cast and the revenue it generates. An action movie usually makes more money on average as compared to comedy, and tends to cast a different set of actors. When unobserved, the genre produces a statistical dependence between if an actor is in the movie, and its revenue. Eventually, the causal estimates for these actors are biased. $\mathbb{E}[Y_i|X_i = x] \neq \mathbb{E}[Y_i]$.
This dependence is most evident when we test this regression model over actors Cobie Smulders and Judi Dench. Cobie Smulders appears in every *Avengers* movie as "Mariah Hill" and Judi Dench appeared in every *Bond* movie as "M"; two moderately popular actors having minor roles in over a billion dollar revenue movies. When a new movie having either of these two actors is considered, the regression model predicts a very high revenue. Upon further inspection this is because these two actors have very high regression coefficients associated with them.

Now how can we solve this? Suppose we somehow measure all covariates C_i and append it to each data point. $\mathbb{D} = \{(X_i, c_i, Y_i)\}$. Covariates are characteristics of causes in an experiment which can be an independent variable of interest or can be an unwanted confounding variable. If these covariates contain all confounders then we can get unbiased estimates of actor coefficients and revenue, i.e: $\mathbb{E}[Y_i|C_i, X_i = x] = \mathbb{E}[Y_i]$.

This is true when C captures all confounders, or precisely; it's true under the assumption of *ignorability*. Conditional on C, the actors are independent of revenue. $X_i \perp Y_i(X_i = x)|C_i \quad \forall x$. The catch is that this equation must hold for all possible actor configurations x not just for the value of $Y_i(X_i = x)$ for a configuration x. Ignorability implies no unobserved confounders. Herein lies one of the prime challenges of causal inference from observed data: that ignorability is untestable. There is no way to check if we can measure and observe all possible confounders.

3 A causal approach

We now take a look at the *deconfounder*. An algorithm proposed by Wang and Blei to sidestep the search for confounders by exploiting the multiplicity of causes. They propose:

1. Fitting a good latent variable model of the causes (actors). Like Probabilistic PCA or Poissons.
2. Inferring a latent variable Z_i for every configuration of actors X_i.

Finally, this inferred Z_i can be used as a substitute for the unobserved confounders to form causal inferences. This algorithm replaces the unetstable search for possible confounders with the testable goal of building a good factor model over the causes.

3.1 Why the Deconfounder works

The Deconfounder is a factor model that can be viewed as a data generative process that shows certain latent variables Z_i can explain the configuration of causes X_i:

$$Z_i \sim P(\alpha) \quad i = 1, ..., n$$
$$X_{ij}|Z_i \sim P(Z_i, \theta_j) \quad j = 1, ..., m$$

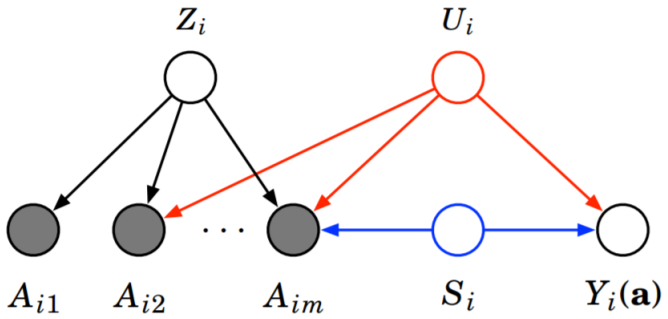
Where:

Z_i is the per movie latent variable (can be multidimensional)
α parameterizes the distribution of Z_i
X_{ij} is the vector representing actors j for movie i
θ_j parameterizes the per-actor distribution of X_{ij}
n is the number of movies
m is the number of actors

The main idea is this: if the factor model captures the distribution of causes-a testable proposition- then we can safely use Z_i as a variable that contains the confounders. If we assume the fitted factor model captures the distribution of causes $P(X_1, X_2, ..., X_m)$. This means that all causes are conditionally independent given the local latent factors,

$$P(X_{i1}, Xi2, ..., X_{im}|Z_i) = \prod_{j=1}^{m} P(X_{ij}|Z_i) \tag{1}$$

Now we make an additional assumption: there are no single-cause confounders, a variable that affects just one of the multiple causes and the outcome. (More precisely, we need to have observed all the single-cause confounders.) With this assumption, the independence statement implies ignorability i.e $X_i \perp Y_i | Z_i$ Ignorability justifies causal inference.

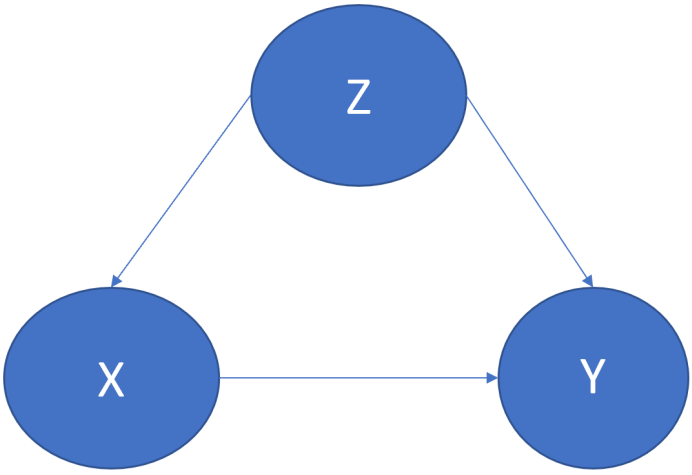


In this image taken from "The Blessings of Multiple Causes", the cause vector X_{ij} is denoted as A_{ij}, Z_i is the substitute confounder, U_i is the multicause confounder and S_i the single cause confounder. The basis for the deconfounder is that if Z_i renders the A_{ij}'s(X_{ij}'s for us) conditionally independent, there cannot be a multicause confounder like U_i by contradiction. The single cause confounder S_i however can not be theoretically ruled out, so we assume there are none.

If we find a factor model that captures the distribution of causes then we have discovered a variable that captures all multi-cause confounders. The reason is that multi-cause confounders induce dependence among the causes, regardless of how they connect to the outcome. Modeling their dependence, for which we have observations, provides a way to estimate variables that capture those confounders.

4 Model Setup and Pyro

We setup our causal approach as a generative model that combines both the factor model and regression into one model over the DAG:



```
def model():
  Z = model_Z()
  X = model_X(Z)
  Y = model_Y(X, Z)
```

4.1 Model Z

The statement Z = *model_Z()* samples our belief of the substitute confounders from a Gaussian. The dimension l of the substitute confounders is a hyperparameter that can be tuned to identify how many confounders best describe our beliefs in the model. Each one of n movies has a Z associated with it so its shape is $n \times l$.

4.2 Model X

The statement X = *model_X(Z)* samples from the distribution $P(X|Z)$. There is a hyperparameter in this distribution: W which acts as a weighted transofmation of Z to X. $P(X|Z)$ is a Bernoulli distribution where $\mathbb{E}[X|Z]$ is a logit function of $Z * W$. When we train, the hyperparameters of W distribution are estimated such that all elements of X (actors) are conditionally independent of each other given Z.

Each one of n movies has m actors associated with it so its shape is $n \times m$.

At this point we have our factor model incorporated to infer latent variables conditional on actors. To recap the process in terms of dimension;

$$
\begin{aligned}
Z.shape &= n \times l \\
W.shape &= l \times m \\
X.shape &= Bernoulli(logit(Z * W)) = n \times m
\end{aligned}
$$

4.3 Model Y

The statement Y = *model_Y(X, Z)* samples the revenue Y from a Gaussian where the mean is a linear combination of X and Z. We setup a Bayesian Regression to estimate the parameters of this linear combination. The predictors in this Regression model are concatenated X and Z where Z is sampled from the inferred distribution of Z and X is the observed data of actors.

4.4 Inference with Pyro

Wang et.al separate the estimation of Z from the estimation of regression parameters. This has to be done because Z by construction renders all causes X (actors) independent of each other. Including the outcome Y (revenue) while learning parameters of Z would make the revenue conditionally independent of X. That violates our primary assumption that actors are a cause of movie rvenue. So they estimate Z, then hard-code it into the regression. This does not fit the one generative model approach we want.

Pyro helps us handle this by running a 2-step training process using their *guide* functions. Guide functions can serve as programmable, data-dependent proposal distributions for importance sampling, rejection sampling, sequential Monte Carlo, MCMC, and independent Metropolis-Hastings, and as variational distributions or inference networks for stochastic variational inference. They are arbitrary stochastic functions used as approximate posterior distributions where we simply mimic the model and assign parameters we want to infer differently from those we don't. We then call a posterior approximating algorithm like SVI on the model and guide and fetch the inferred parameters from the environment.

So for the first training process of optimizing $P(X|Z)$ i.e the parameters Z and W, we use a guide that chooses to optimize hyperparameters of Z and W only, and just samples the rest of the parameters.

For the second training process of optimizing the regression coefficients for $P(Y|X,Z)$ we use a guide that chooses to optimize hyperparameters of the Bayesian Regression setup only, and just samples the rest of the parameters from the learned distributions of the previous training step.

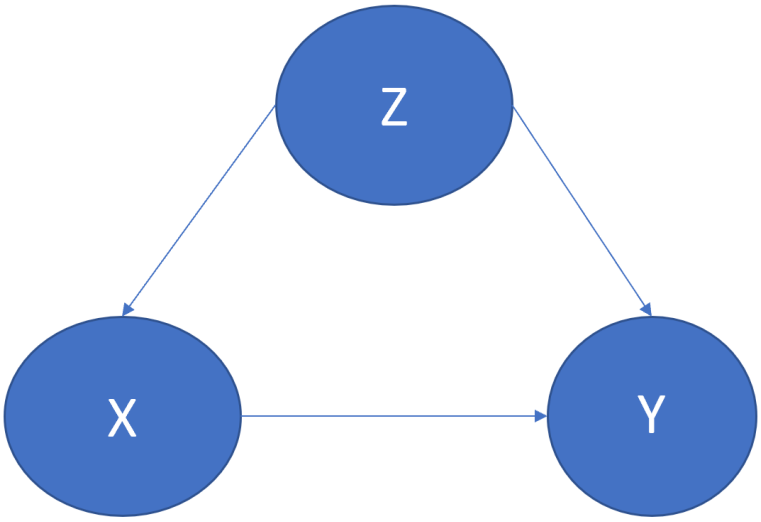
Note how we are still doing regression with X and Z like Wang et.al proposed but with multimodal Z sampled from a distribution and not hardcoded values outside the DAG.

5 Queries

This modelling along with Pyro affords us the freedom to check causal estimates with simple Pyro queries.

5.1 Biased Estimates

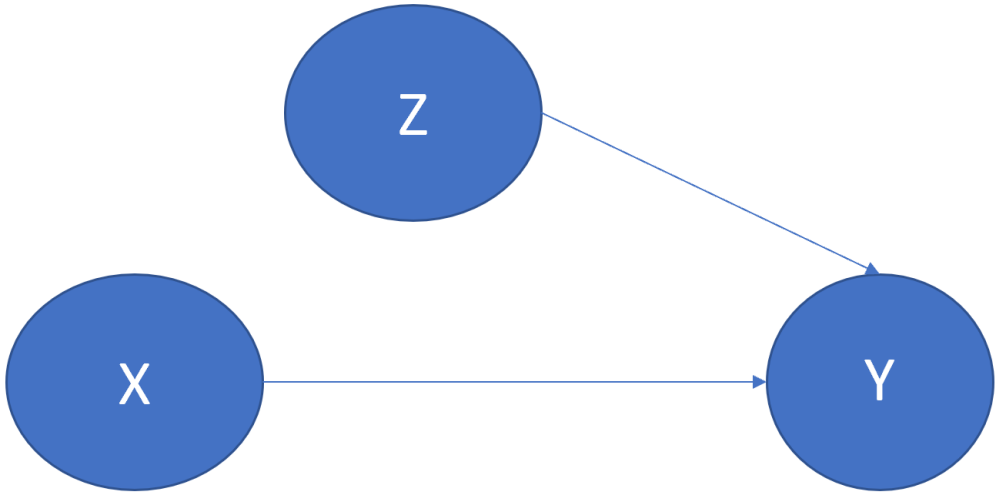
Causal effect of Actor j on Revenue without conditioning for the confounders can be obtained by $\mathbb{E}[Y|X_j = 1] - \mathbb{E}[Y|X_j = 0]$



Sampling from the generative model while Z affects both X and Y is the equivalent of our initial assumption that certain confounders influence both X and Y. So this sampling gives us biased causal estimates that any machine learning model would have given us.

5.2 Unbiased Estimates

Causal effect of Actor j on Revenue while conditioning for the confounders can be obtained by $\mathbb{E}[Y|do(X_j = 1)] - \mathbb{E}[Y|do(X_j = 0)]$



Sampling from the generative model while Z does not affect X but does affect Y is a way to ensure that the effects of X on Y and Z on Y are isolated. So when regression estimates are inferred, the estimates for actors are closer to their true causal effects as the revenue associated to confounders gets attributed to the regression estimates for the substitute confounders.

6 Results

We fit this model over observed data and infer Z, then infer regression parameters.

6.1 Prediction

We run the query $\mathbb{E}[Y|do(X_j = 1)] \quad \forall j \in x$ where x is a cast of actors for the new movie we want to predict revenue Y for.

Method	RMSE
Deconfounder + Regression	178.8
Generative Model	189.43

There is not an improvemment in accuracy, if anything, it gets mildly worse.

6.2 Debiasing Causal Estimates of Actors

We check if our generative model corrects actor biases similarly to those of Wang et.al. We consider actors: Brad Pitt, Anotnio Banderas, Anne Hathaway, Arnold Schwarzenegger, Nicolas Cage, Stanley Tucci, Ben Affleck, and to prove our case: Judi Dench and Colbie Smulders.

An Actor is tagged as Overvalued if their Biased regression estimates are greater than their Unbiased regression estimates and vice versa.

Actor	Deconfounder + Regression	Generative Model
Brad Pitt	Overvalued	Overvalued
Antonio Banderas	Undervalued	Undervalued
Anne Hathaway	Overvalued	Overvalued
Arnold Schwarzenegger	Overvalued	Undervalued
Nicolas Cage	Undervalued	Undervalued
Stanley Tucci	Undervalued	Undervalued
Ben Affleck	Undervalued	Overvalued
Judi Dench	Overvalued	Overvalued
Colbie Smulders	Overvalued	Overvalued

Arnold Schwarzenegger, Ben Affleck, and Brad Pitt(marginally) are some actors for which our model's regression estimates differed from Wang et.al's.

7 Advantages over Wang et.al

This approach, coupled with implementation in Pyro gives us some advantages over the original paper off which it was based.

1) *Z is multi-modal now*. So we don't use hardcoded values of substitute confounders we use values from a distribution for $\mathbb{E}(Z|X)$. With the confounders and regression abstracted, work is now specifying the model parameters and training it.

2) *Pyro* lends a powerful advantage implementation wise because our queries to predict and check causal effects remain the same. If and when a better predictive model comes up, we can simply replace that in our model and continue the same studies with the same queries.

8 Limitations

The causal parameters in a problem (regression params for us) are identified if the causal parameters that could generate the observable data(z and w) are unique. However, if the parameters are not identified, then even if the distribution of the observable data can be estimated perfectly, the causal parameters cannot be estimated consistently. Instead we can only obtain an ignorance region of parameter values to which the observed data give equal support. Even with infinite data, there will be many causal explanations of the observed data that cannot be distinguished on the

basis of the data alone.
Identification is a property of the data-generating process, not a property of the estimation method. Resolving identification issues requires changing the data generating process or adopting assumptions about the data generating process that are untestable in the data alone or choosing different causal parameters to estimate.
Consider covariance matrix for our model:

$$\sigma_{XYZ} = \begin{pmatrix} \sigma_{ZZ} & \sigma_{ZX} & \sigma_{ZY} \\ \sigma_{XZ} & \sigma_{XX} & \sigma_{XY} \\ \sigma_{YZ} & \sigma_{YX} & \sigma_{YY} \end{pmatrix}$$

We're interested in the bottom right entries defined by:

$$\begin{aligned} \Sigma_{XX} &= \alpha\alpha'\sigma_Z^2 + \operatorname{diag}(\sigma_X^2) \\ \Sigma_{XY} &= \Sigma_{XX}\beta + \gamma\sigma_Z^2\alpha \\ \Sigma_{YY} &= (\beta'\alpha + \gamma)^2\sigma_Z^2 + \beta' \operatorname{diag}(\sigma_X^2)\beta + \sigma_Y^2 \end{aligned}$$

LHS is observable but structural params on the right aren't. Goal is to obtain unique values for them. When m(observations)???3, the number of equations exceeds the number of unknowns, but there still exists a family of structural equations with parameters that induce the same observed covariance matrix:

$$(\alpha_1, \beta_1, \gamma_1, \sigma_{Z,1}^2, \sigma_{X,1}^2, \sigma_{Y,1}^2) \neq (\alpha, \beta, \gamma, \sigma_Z^2, \sigma_X^2, \sigma_Y^2)$$

These are observation-equivalent(OE) parameters that cant be distinguished by observed data. It can be shown that there is a class of OE params that allow β to take different values while still explaining observed data. The proof for this is explained in a blog referenced below. This means we can model the data generation params but not the regression params consistently and uniquely.

9 Future Scope

The limited time afforded by a semester, and bottlenecks with understanding the problem and Pyro implementations prevented me from doing a few more things that could be interesting prospects:

1) Using genre as an additional feature in the generative model. This is guaranteed to be a strong confounder.
2) Use the abstraction afforded by the Bayesian Network and Pyro to test other priors, and more sophisticated non-linear regression methods like Neural Networks.
3) Convert this to a Structural Causal model and conduct counterfactual reasoning.

10 References

1. Yixin Wang, David Blei. (2019). "The Blessings of Multiple Causes."
2. Imai, K., Keele, L., and Yamamoto, T. (2010). "Identification, inference and sensitivity analysis for causal mediation effects." Statistical Science, pages 51-71.
3. Alex D'Amour. (2018) (Non-)Identification in Latent Confounder Models.
4. Github Repository for the model.
5. Pyro; Inference with Pyro.
6. Pyro; Bayesian Regression.
7. Blei labs Probabilistic PCA using Edward.