Some Title

All of us

October 11, 2020

Abstract

1 Introduction

The following chapter is concerned with the existing disconnect between the fields of travel demand modeling and causal inference. More specifically, the chapter is motivated by the current lack of use of methods and findings from the causal inference literature in travel demand modeling.

More often than not, the development of behavioral models in transportation, specifically transportation demand models, is driven by the need to evaluate the impact external interventions, in the form of alternative policies, on a certain outcome of interest. These questions that transportation demand models are built to answer are causal by definition: we are interested in how the system reacts to *external* interventions. Yet, when those models are developed, there is very little consideration given to causality, and when causal concepts are accounted for, the process is done implicitly without a formal framework.

While the field of transportation demand modeling could benefit greatly from incorporating causal inference techniques, there are barriers that have made this integration slower than what one would hope. These barriers stem from the difference between the types of problems transportation demand modelers deal with and those that are typically studied in the causal inference literature. Perhaps the main fundemental difference is that demand modelers are typically trying to forecast the impacts of policies that haven't been implemented or seen before, which requires additional work and a change to the typical causal modeling workflow in order to translate a given policy (treatment) into a set of characteristics and variables that exist in the data and system at hand (please refer to Brathwaite and Walker (2018) for a more thorough discussion of those barriers). While those barriers make the problem of demand modelers harder, there is still a lot to gain from incorporating causal inference techniques where appropriate, and to contribute in turn where the literature lacks.

The relevance of this topic now is motivated by the significant boost in the causal inference literature recently, both in the potential outcomes framework and the causal graphical approach. The goal of this chapter is to formalize a framework for approaching transportation demand modeling problems from a causal perspective. We will draw heavily on the use of directed acyclic graphs (DAGs) formalized by Pearl (2000) as a means of representing the modeler's knowledge and assumptions about a given problem. The chapter will provide an overview of DAGs, the testable implications that come with one's causal representation, the main tests that one could do to falsify/justify a given causal graph, and then how to use a causal graph to estimate the causal relationships of interest. We will demonstrate the use of this framework through simulations, where we clearly show the benefits and implications of this approach as opposed to traditional approaches. The last part of this chapter deals with the more complicated issue of latent confounding, where one variable confounds two or more variables in the causal graph. This type of confounding creates variations in the outcome variable that are not caused by the confounded variables but are correlated by it, which biases the estimated causal effects of those variables if nothing is done to account for the confounding. We focus on a recent technique to deal with latent confounding suggested by Wang and Blei (2019) to address unobsrved confounding in tranportation applications. We describe this problem in more details in section 7.

2 Perils of disregard

Consider the following scenarios: 1) Based on recent input from the public, the Department of Transportation (DOT) of a certain city is considering implementing a new parking policy that discourages parking in the Central Business District and therefore would encourage the use of more active transportation modes such as walking or biking. 2) A certain DOT is considering implementing a Streetcar system to incentivize more transit oriented development in a certain region.

When thinking about these two "policy" proposals (funding the Streetcar system in the second case), it could that the DOT (or any other agency) in question ran analyses (either based on meticulous modeling or anectodal evidence) that allowed them to deduce that implementing such policies would directly result in their desired output or achieve their desired goal. This inherently assumes the causal relationship between the policy and the final output (implementing the new parking policy will increase the share of active transportation modes in CBDs and funding the streetcar system will generate more transit oriented developments). Moreover, these analyses - and conclusions from such analyses - of the impact of the proposed policies are also based on a certain belief of how the the world or the system works. However, in many instances (insert references to policies etc..) such belief of structure of the system is not made clear; we only hear about the policy and its most downstream impact. Such depiction of proposed policies maintains an obscure representation (at least to the knowledge of the average person) of the system and the interactions between its intermediate elements.

Directed Acyclic Graphs (DAGs) otherwise known as causal graphs allow one to clearly represent one's assumptions about the problem at hand. DAGs have been used in fields ranging from epidemiology, scheduling, and network structures and have shown to be extremely useful.

DAGs could show to be very useful in addressing policy questions (in our case transportation policy questions). Brathwaite and Walker (2018) have illustrated an example of how DAGs can be used to answer such questions. Brathwaite and Walker (2018), however, have not showed an empirical application of how using DAGs can result in different model estimates when compared to traditional methods of modeling. In this section, we will show how to structurally approach demand modeling problems while paying close attention to the data generating process and the causal structure of how people make choices. We approach this exercise by building up on work from Brathwaite and Walker (2018) work by going through the motions of empirical exercise exemplifying a simplified problem. We want to note that this example is for illustrative reasons and that it does not reflect the complexities usually accompanying a transportation choice modeling exercise. Let's assume that a company wants to reduce its workforce carbon footprint by moving its employees closer to campus. We would like to model the travel mode choice of these workers based on a dataset from **?**. This dataset comes from the 2012 California Household Travel Survey (CHTS). It contains 4004 home-based school or work tours made by approximately 3850 individuals in the Bay area. Readers interested in a more detailed description of the dataset can refer to **?**. We would like to remind readers that our goal in this section is to only show how one's beliefs and assumptions about the data generation process affect estimates from the outcome model and that controlling for intermediate variables of some policy variable of interest in a causal graph, we do not succeed at recovering the true causal parameters of the variable of interest. For purposes of this exercise, we take the Multinomial Logit model defined in **?** as the true outcome generating process.

The systematic portion of the the multinomial logit model from **?** is specified as follows:

TODO: convert the following equations.

$$U_da = \beta_{time_drive} \times TravelTime + \beta_{cost_per_distance_da} \times CostperDistance_{da} + \beta_{autos} \times NumberofAutos$$

$$U_sr2 = ASC_{sr2} + \beta_{time_drive} \times TravelTime + \beta_{cost_per_distance_sr2} \times CostperDistance_{sr2} + \beta_{autos} \times NumberofAutos + \beta_{cross_bay} \times Cr$$

$$U_sr2 = ASC_{sr3+} + \beta_{time_drive} \times TravelTime + \beta_{cost_per_distance_sr3+} \times CostperDistance_{sr3+} + \beta_{autos} \times NumberofAutos + \beta_{cross_bay} \times$$

$$U_WTW = ASC_{WTW} + \beta_{travel_time_transit} \times TravelTime + \beta_{travel_cost} \times TravelCost$$

$$U_DTW = ASC_{DTW} + \beta_{travel_time_transit} \times TravelTime + \beta_{travel_cost} \times TravelCost$$

$$U_WTD = ASC_{WTD} + \beta_{travel_time_transit} \times TravelTime + \beta_{travel_cost} \times TravelCost$$

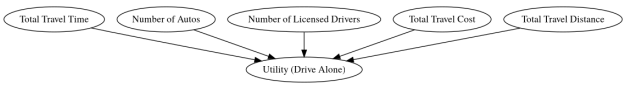

Figure 1: Causal Graph with Independent Covariates

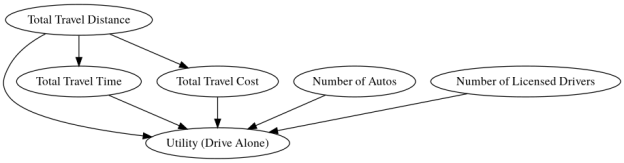

Figure 2: Causal Graph for the Drive Alone Utility Function

$$U_Bike = ASC_{Bike} + \beta_{travel_distance_bike} \times TravelDistance$$
$$U_Walk = ASC_{Walk} + \beta_{travel_distance_walk} \times TravelDistance$$

We achieve our goal through a simulation exercise as follows:

1. We generate a causal graph where we assume that all explanatory variables are observed and that there is no no structural relationship between any of them. We simulate data based on this graph and then simulate outcomes based on parameters from a well-defined mode choice model using a Multinomial Logit (MNL) choice model described in **?**. We will then apply the do-operator Pearl (2000) perturb the travel distance variable to emulate the company's decision to move its employees closer to campus. We will then predict outcomes based on the mode choice model.

2. We generate a different causal graph based on different assumptions of how explanatory variables influence and interact with each other. We estimate the relationships outlined in this causal graph between different nodes based on the data at hand. We simulate "upstream" nodes not affected by any other nodes in the causal graph and use the estimated relationships between the explanatory variables to estimate the remaining explanatory variables. We then use the choice model from step 1 to simulate outcome choices based on this new dataset. We move to apply the do-operator to perturb the variable of interest (in our case travel distance) to replicate the intervention resulting from a policy changing travel distance, and use the estimated relationships between the explanatory variables in the causal graph to simulate all the explanatory variables in the causal graph. We use the estimated choice model to produce outcomes based on the newly estimated data.

We repeat this simulation process numerous times and recover the probabilities that each individual chooses a certain mode. In this exercise, we focus on the probability of choosing car-centric modes (Drive alone, Shared ride with another person, and a shared ride with two or more individuals). We compute the difference in probabilities generated by the different models based on different data generating processes.

Figure 1 shows the causal graph where all explanatory variables are independent of each other for each utility equation.

The variables shown in the graph are assumed to be all of the explanatory variables needed to estimate the outcome and are based on the specification of the MNL model outlined in **?**. We can describe the variables as follows: - Total travel distance: is the total travel distance for individual i and mode j, for all available modes for individual i during trip t of tour l. - Total travel cost: is the travel cost in dollars for individual i and mode j, for all available modes for individual i during trip t of tour l. - Total travel time: is the travel time in minutes for individual i and mode j, for all available modes for individual i during trip t tour l. - Number of autos: is the number of automobiles owned by individual i's Household - Number of licensed drivers: is the number of licensed drives in individual i's household. - Number of kids: is the number of kids (specify age?) in individual i's household. - Cross-bay trip: is a binary variable indicating whether the trip t for individual 1 is a cross-bay trip.

Similarly, Figure 2 through 9 illustrate the causal graphs with "interacting" explanatory variables.

Each of these graphs is based on the utility function of each respective mode in the multinomial logit model estimated in **?**.

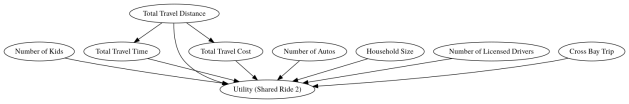

Figure 3: Causal Graph for the Shared Ride 2 Utility Function

Figure 4: Causal Graph for the Shared Ride 3+ Utility Function

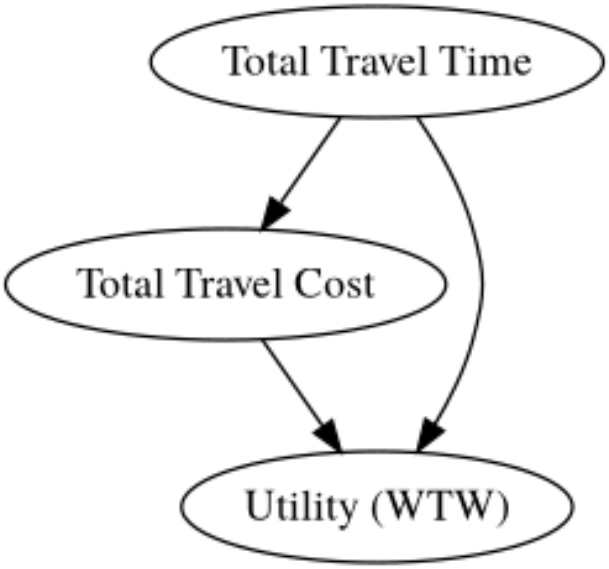

Figure 5: Causal Graph for the Walk-Transit-Walk Utility Function

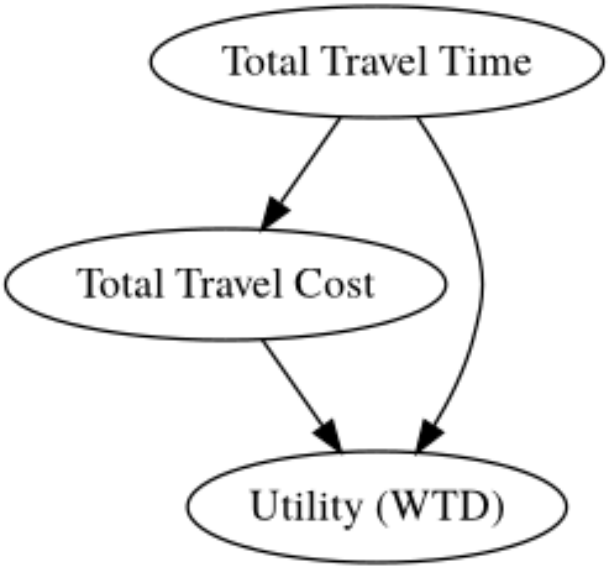

Figure 6: Causal Graph for the Walk-Transit-Drive Utility Function

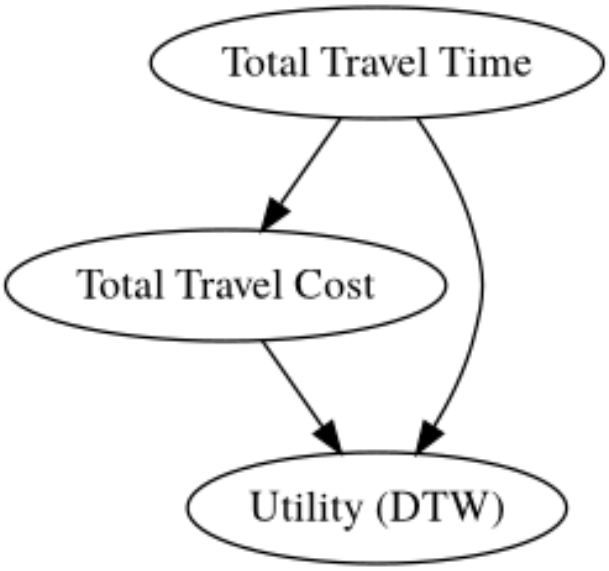

Figure 7: Causal Graph for the Drive-Transit-Walk Utility Function

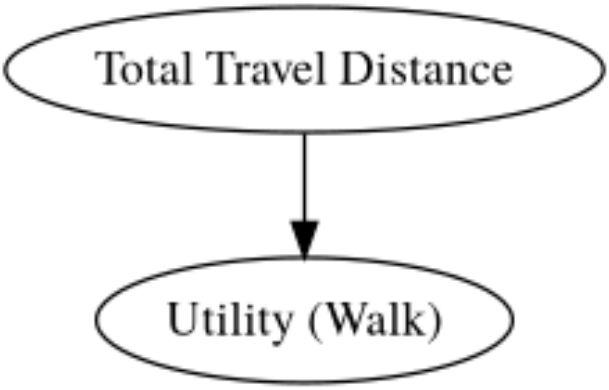

Figure 8: Causal Graph for the Shared Ride 3+ Utility Function

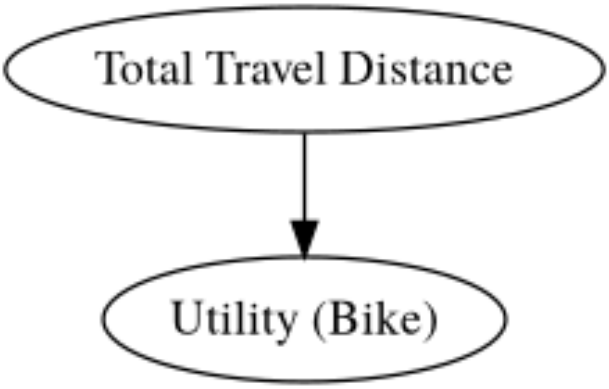

Figure 9: Causal Graph for the Bike Utility Function

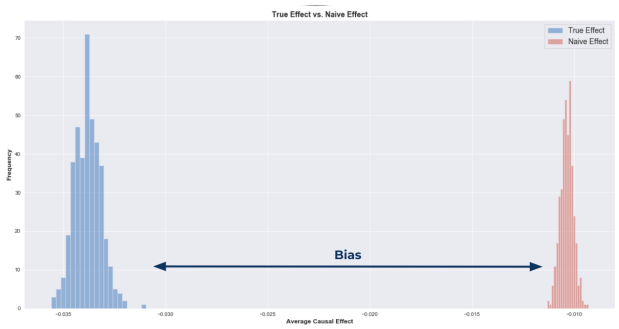

Figure 10: Histograms of the probability of choosing Car Centric Modes under Different Data Generating processes.

We then plot histograms of the computed differences between the average probability of an individual in our sample choosing a car centric mode before and after implementing a policy or intervention aimed at reducing travel distance. These differences are plotted under the two different assumptions about the data generating process illustrated in the causal graphs above. Figure 10 highlights the bias between the estimated probability of an average individual choosing a car centric mode. This difference shows the importance of considering the data generating process when estimating transportation demand models aiming to forecast the impact of proposed policies.

The data generating process might not be easily distinguishable in the majority of situations, mainly due to the complexity of the real world. Therefore, constructing a causal graph that represents the data generating process as much as possible is not an easy task. This chapter explores other topics related to this matter, including some guidance on how to build and test causal graphs representing the researchers beliefs about the data generating process.

3 Overview of Causal Graphs

Causal diagrams and causal graphical models have been introduced by Pearl (2000) as a powerful tool for causal inference, especially in observational studies. Perhaps one of the most important and useful features of causal graphs when dealing with causal inference problems is the clear illustration of the causal relationships between the variables. While a formal introduction to the topic of directed acyclic graphs (DAGs) is beyond the scope of this chapter, here we focus specifically on illustrating the power of DAGs to represent and encode complex causal relationships between variables in an intuitive and clear manner. Interested readers can refer to Pearl (2000) for a thorough introduction.

Consider the causal graph represented in Figure 11. Suppose we're interested in the effect of Z on Y. What the graph in Figure 11 implies is that Z is independent of Y(Z) given X, or in other words, the assignment mechanism is ignorable conditional on the values of X. In such situations, it is sufficient to control for X to obtain an unbiased estimate of the causal effect of Z on Y.

In comparison, note the set of structural equations needed to convey the same assumptions as Figure 11.

$$Z = f_Z(X, \epsilon_Z)$$

$$Y(z) = f_Y(X, z, \epsilon_Y)$$

Now consider the case where there exists another latent confounding variable, U, which also affects both the treatment Z, as well as the outcome, Y. Figure 12 illustrates this assumption in DAG form, and the equations below are the structural equation equivalent:

$$Z = f_Z(X, U, \epsilon_Z)$$

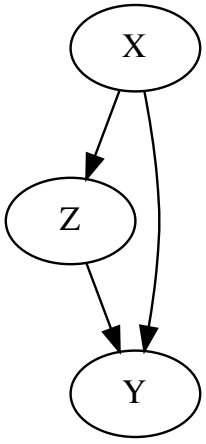

Figure 11: Simple DAG of the causal relationships between X, Y and Z

$$Y(z) = f_Y(X, U, z, \epsilon_Y)$$

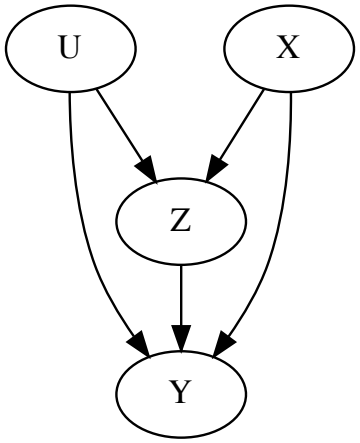

Figure 12: Simple DAG of the causal relationships between X, Y, Z and confounder U

It is clear from Figure 12 that conditioning on X alone does not block all the indirect paths between Z and Y, and thus ommitting it will yield biased results of the causal relationship between Z and Y. The structural equations also show that the ignorability assumption of Z does not hold if we only condition on X, and we risk obtaining biased estimates of the causal effect of Z on Y if we fail to account for U. Even in this very simple example with only three or four variables, one can see the advantage of using a causal

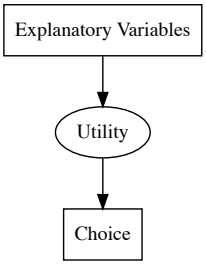

Figure 13: Archetypical RUM causal diagram

graph to encode assumptions. This advantage of DAGs becomes even more apparent in larger problems with many covariates available, and where the causal structure of the data is way more complicated, which are the types of problems typically faced by transportation demand modelers.

Another benefit of DAGs is that they come with a set of testable implications, and incorporating them in any causal analysis adds robustness and defensibility to one's analysis. We discuss those implications in section 5, and illustrate how to use those tests in one's analysis.

Lastly, it is important to note that the graphical approach to causality focuses primarily on issues of identification of causal effects, that is, given a directed acyclic graph (DAG) that encodes an analyst's knowledge and belief about the data generation process of the problem at hand, can a specific causal effect be identified? As such, we emphasize that DAGs are great tools for a modeler to encode their assumptions about a problem, but not necessarily a guide on how to estimate a causal effects of interest.

3.1 Prior Uses of Causal Graphs in Choice Modelling

With the introduction to causal graphical models given above, we (the authors) do not want readers to think that we believe the entire notion of a causal graph is foreign to choice modelling. We do not think this is true. We know that choice modellers have made use of causal graphs in varying forms for years. Here are three examples to illustrate our point. First, consider the case of Random Utility Maximization (RUM) models.

For decades now, choice modellers have used stylized directed graphical models to denote or convey the meaning of their economic framework's notion of causality in the context of their research. As an example, see Figure 13, reproduced from Figure 1 of Ben-Akiva et al. (2002). Here, the authors depict the assumption that explanatory variables cause unobserved utilities which collectively cause the choice via utility maximization. The reason I call such diagrams stylized is because of their lack of detail of the relationships amongst the explanatory variables. When speaking of these covariates collectively, one is unable to distinguish whether one explanatory variable causes another. Such intra-covariate relationships are important for estimating the causal effects of a given policy, as demonstrated in Section 2.

Another area of relation with prior choice modelling literature is in Integrated Choice and Latent Variable (ICLV) models. These models make inference on one or more latent variables, alongside the typical parameters in one's model. An example of such ICLV models is in Figure 14, based on Figure 5 of Ben-Akiva et al. (2002). As with the causal graphs of RUM models, ICLV causal graphs depict a particular set of structural modeling assumptions about the data generating process. Here concern typically centers around unobserved mediators, i.e., variables caused by one's measured covariates that also influence one's outcome. Omitting these variables leads to inconsistent estimates for one's remaining parameters, thus leading researchers to care deeply about ICLV models that can overcome such problems.

Finally, consider activity based travel demand models. These models often come with a causal diagram that depicts the interrelations between the outcomes in the model (e.g., household location choice, destination choice, travel mode choice, departure time choice, route choice etc.) For example, see Figure 15, from Bradley et al. (2010, Fig.1). The purpose of such graphs is to explain the structure of the entire system

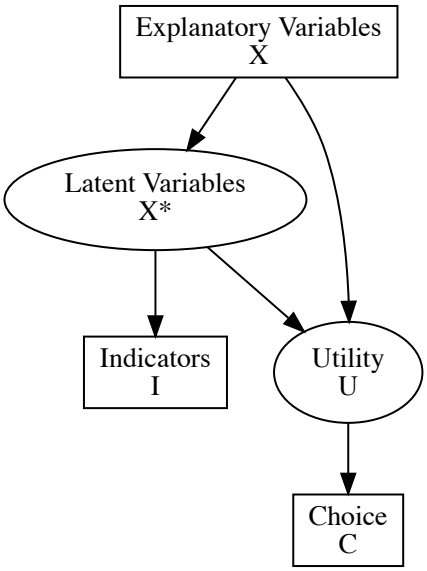

Figure 14: Archetypical ICLV causal diagram

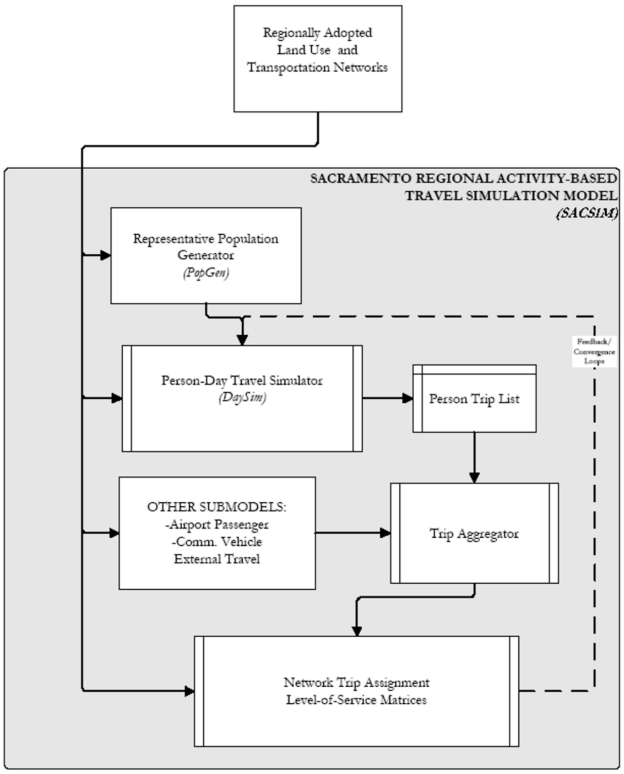

Figure 15: Archetypical causal diagram for activity-based models

of outcome models. In particular, these graphs depict the researcher's assumed causal ordering of which outcomes temporally precede other outcomes. Diagrams of activity-based models also show the modeller's assumptions about how considerations of future choices affect temporally-preceding decisions.

In RUM, ICLV, and activity-based model diagrams, the main distinction between the causal graphs of Pearl (1995) and the causal diagrams in econometrics are two two-fold. First, causal graphs in RUM, ICLV, and activity-based models ignore relations between the explanatory variables. Typically, choice modelling causal graphs depict all explanatory variables together as if they are in independently generated groups that then affect one's utilities and choices. In the language used by a large group of causal inference scholars, econometric causal diagrams ignore the treatment assignment mechanism.

Secondly, causal graphs in choice modelling papers are purely didactic. They convey how choice modellers perceive the world and the choice generation process. However, they are seldom treated as a fully fledged model that has empirical implications that merit verification (e.g. conditional independence implications). In this sense, choice modellers ignore the efforts from causal inference researchers in computer science. There, researchers spend much effort testing their causal graphs to see if the data supports their graphs implications.

In conclusion, choice modellers have long made use of causal graphs in select contexts to convey causal assumptions about choice processes. Thus far, however, we have underutilized these tools. We seldom use causal graphs to encode assumptions about how our explanatory variables came to be in choice situations, and we do not routinely test causal graphs against empirical choice data. These two issues represent opportunities for the field of choice modelling to gain from the insights and work of those who study causal inference. The rest of the chapter will focus on how we can construct causal diagrams that pay attention to both explanatory and outcome variables, how we can test a given graph against one's data, and how one can deal with "real-world" graphs which frequently contain some sort of unobserved confounding.

4 (Initial) Causal Graph Construction

Construction of an initial causal graph typically proceeds as follows. At a high level, we

1. adopt a population perspective,
2. brainstorm all variables that we think affect the system that generates our observations,
3. remove any variables that would cause bias in our causal inferences,
4. connect all variables in our graph according to our a-priori beliefs about causal relations amongst these variables,
5. consider how the graph structure may differ between individuals and subgroups within the population.

The following paragraphs will describe these steps in detail.

To begin, we adopt the position of a researcher concerned about population level relationships. This means we will think through what is a likely generative model for all individuals. Later on, we will devote time to thinking about how subgroups and individual heterogeneity may affect our causal graphs.

Now, we add our first variables to our graph, the outcome variable(s) of interest. Then, we note all the variables that we believes to influence the variable(s) of interest. We refer to these variables to as our initial explanatory variables.

Next, we iterate through the initial explanatory variables. For each current explanatory variable in the iteration, we think of variables that may modify the effect of the current explanatory variable on the outcome(s) of interest. We refer to these variables as effect modifiers. Note that some effect modifiers may be a part of our initial explanatory variables. For any effect modifiers that one thinks of, outside of the initial explanatory variables, add them to our causal graph.

After adding explanatory and effect modifying variables to the graph, we turn our attention to mediating variables. Again, we iterate through each explanatory variable. On each iteration, we brainstorm all the variables through which our current explanatory variable influences the outcome. For instance, consider how the presence of a bike lane influences bicycle mode choice. We hypothesize that an individual's subjective perception of safety is be the primary (or sole) variable through which bicycle lane presence influences mode choice. Accordingly, we add subjective perception of safety to our causal graph for travel mode choice.

Coming to our second to last category of variables, we think of confounding variables. The process is similar to how we generated effect modifying variables. We iterate through each of the explanatory, mediating, and effect modifying variables, thinking specifically of any variables that both cause the current variable in the iteration and cause the outcome variable(s). We call these variables, which cause our outcome and current variables in the iteration, confounding variables. We will add each of them to our causal graph.

For the last set of variables, we should explicitly consider the role of time, even in research that may be cross-sectional due to the data that is available to us. In reality, how do we think our system evolves over time? If we consider multiple observations of a given decision maker, how do variables observed of that decision maker at time t partially cause future variables important to the context or outcomes for that decision maker at time $t+1$? How do the actions of a decision maker i at time t partially cause the future context or outcomes of a decision maker j? We should add explicit nodes to our graph, subscripted or denoted by time, to show the cross-time causal relationships in our system.

At this point, we have added all the outcome, explanatory, effect modifying, mediating, confounding, and time-indexed variables that we can think of to our causal graph. They are all disconnected nodes, singletons, in the graph. We now focus on pruning nodes from this graph, before drawing our final hypothesized connections. In particular, we focus on pruning "post-outcome" variables that are not part of the causal graph for future time periods or other observations. The reason for this is that conditioning on such post-outcome variables biases one's causal effect estimates.

To remove the problematic variables, we iterate through each of the non-outcome variables in our graph, and we assess whether the variable is actually a result of the outcome (perhaps in combination with other variables in our graph). These post-outcome variables temporally follow the outcome variable(s) but do not cause variables in the causal graph for other observations. We remove all such post-outcome variables from our graph.

Now is a good time to step back and consider what other researchers have thought. Specifically, we should conduct a literature review to see how other researchers have conceptualized the problem that we are working on. Have they included variables that we have not? Were those variables related our outcomes of interest? If so, should we add these variables to our causal graph, and how? Do the included variables of other researchers point suggest the existence of confounders in their work that we should include in our graph? Have other researchers ascribed differing roles to our graph's current variables than we have? For example, have other researchers judged a variable to be confounder, when we solely thought of the variable as an affect modifier? Critically examine the evidence for these alternative decisions to see if we should also reconsider how we're judging our variables.

Finally, we need to connect the variables in our graph.

1. Draw direct arrows from our explanatory variables, confounders, and effect modifiers to the outcomes.
2. Draw arrows from the explanatory variables to the mediators, and then draw arrows from the mediators to the outcomes.
3. Draw arrows from the confounders to the explanatory variables and mediators that they may cause.
4. Draw arrows from the variables in time t to the variables that they cause in time $t+1$.

After drawing in all arrows, we should now have a fully connected causal graph. Take a moment to look over the graph to ensure there are no remaining singletons and that we have not drawn any spurious connections. Also, take a moment to celebrate. Drawing one's first causal graph is hard work!

After celebrating, take a moment to pursue the following graph editing exercises. First, think about how the graph might differ across sub-populations. What sub-populations exist in your population of interest? Are there any causal relationships that should not exist for a given sub-population? For instance, are the outcomes in some sub-populations independent of a given explanatory variable? Can you think of any inverted causal relationships in this sub-population? (I.e., for a given sub-population $B \rightarrow A$ instead of $A \rightarrow B$?) Consider adding these sub-population indices to one's initial causal graph, or if this is not clear enough, draw amended causal graphs for each sub-population. Now, one can actually relax. This concludes the "purely mental" drafting of one's causal graph. In the next section, we'll look at testing this graph against data, and making any edits deemed empirically necessary.

5 Testing of Causal Graphs

5.1 Description

5.1.1 Testing observable assumptions

In the last section, we reviewed a process for creating an initial causal graph using expert opinion. Critically, after drafting a causal graph, we should immediately test it against empirical data. This is important because, if our graph captures inaccurate assumptions about the data generating process, then we have no reason to think that our conclusions from using the graph will be accurate.

To test our causal graphs against data, we will first test the implications of our graph that involve observable variables only. We will defer the task of testing implications that involve unobserved / latent variables to later in this subsection. For now, recall our discussion in Section 3 about the two basic implications of causal graphs: marginal independence and conditional independence. In both cases, direct testing of marginal or conditional independence amongst nodes in the causal graph may be difficult. Indeed, there are no direct tests of conditional independence that can detect all types of dependence, especially for continuous variables (Bergsma, 2004; Shah et al., 2020).

As a result of this hardness, there are a myriad of research efforts aimed at testing conditional independence of two variables X and Y, given a third variable Z and any number of assumptions about the variables or the test statistic itself. Some researchers proceed under the assumption that one has access to an approximation of the conditional distribution of $X \mid Z$ (Candès et al., 2018; Berrett et al., 2019). Other researchers designed conditional independence tests for general cases, assuming smoothness of the underlying data distributions and assuming accurate estimation of the distribution of the test statistic under the null hypothesis of conditional independence (e.g. Zhang et al. (2012); Strobl et al. (2019)).

Different from (but not excluding) these approaches, we will take an easier and less decisive route. If a pair of variables have conditionally or marginally independent distributions, then their statistical moments will also be conditionally or marginally independent. Instead of testing for marginal or conditional independence in distribution, we will perform a more tractable test for marginal or conditional independence in means. If the variables in question are not conditionally or marginally independent in their means, then we know they are not independent in their distributions. Conversely, even if a set of variables are marginally or conditionally independent in their means, this **does not** imply that the variables are independent in distribution.

This approach of indirectly assessing distributional independence by testing mean independence is not new. The following papers have all proposed and implemented such an idea: Burkart and Király (2017); Chalupka et al. (2018); Inácio et al. (2019). For conditional independences, the crux of the approach is to predict Y based on X and Z, then compare against a prediction of Y based on a resampled value of X and the original Z. If Y is mean-independent of X given Z, i.e. $E[Y \mid X, Z] = E[Y \mid Z]$, then the predictive power of a model with resampled X should resemble the predictive power of a model with the original X. After all, in both cases, the conditional expectation of Y is independent of one's X values (real or resampled). When assessing marginal independencies, one removes Z from the models for the expectation of Y and proceeds as described.

Note that as with the case of testing distributional independence, testing mean independence still requires researchers to make choices. We have to select models for $E[Y \mid X, Z]$ and $E[Y \mid Z]$, respectively, for testing conditional and marginal mean-independence. We also have to choose the performance statistic (e.g. R^2, log-likelihood, etc.) to compare these models. Lastly, we also have to select a resampling method. In particular, how (if at all) will our resampling strategy account for the possible dependence between X and Z?

For our demonstration, we made the following choices. First, we used linear regressions to model $E[Y \mid X, Z]$ and $E[Y \mid Z]$. Second, we chose to use R^2 as our test statistic for judging the predictive performance. Third, we have chosen to resample the X vectors without replacement, keeping the length of the resampled vector equal to the length of the original vector. In other words, we permute X. And finally, we have chosen to visualize these tests by plotting a vertical line to display our observed test-statistic and by plotting the distribution of our test statistics based on the permutations of Y.

Our rationale for these choices are as follows. In our dataset, most of our explanatory variables were continuous (at least in theory). Accordingly, R^2 seemed a sensible performance metric for a model of the conditional expectation of a continuous random variable.

In contrast to our choice of performance metric, we chose our conditional expectation models and resampling methods based on empirical testing. In particular, we created simulations to assess our mean-independence testing procedure. We assessed the performance of our mean-independence testing procedures using simulations where $Y \leftarrow Z \rightarrow X$ and X either did or did not cause Y.

Of particular importance were our simulations under the null hypothesis where X was conditionally independent of Y. Our initial simulations used random forests as our conditional expectation models and permutations as resampling methods. Random forests are a non-parametric method that would allow us to have less fear of model misspecification, and permutations are easy to implement. However, under the null hypothesis, we discovered that the tests based on the random forest models did not result in p-values that were uniformly distributed. Given that we planned to use these procedures in a manner akin to hypothesis-testing, we hoped that our test statistics would be U-statistics. When we switched from the combination of random forests and permutations to linear regressions and permutations, our p-values were indeed empirically, uniformly distributed. Moreover, we still retained high power.

We do not claim that these choices for assessing mean independence will always be appropriate. Indeed, one should assess one's tests on simulated data that resembles one's real data. For our illustrative purposes though, the combination of linear regressions, permutations, and R^2 resulted in adequate tests of marginal and conditional mean-independence.

5.1.2 Testing assumptions involving latent variables

Now that we have described testing with observable variables, we can more easily describe conditional independence tests that involve unobserved (i.e., latent) variables. Indeed, when dealing with observational data, we will frequently find ourselves not having observed all variables that are of interest. Nevertheless, we still wish to test whether our data contradicts our graph. One way to directly extend our conditional independence testing to account for latent variables is to adopt a missing data perspective and impute the latent variables from a prior distribution. In particular, we can generalize our previous tests as follows.

First, we can consider that instead of performing one test with a set of observed X, observed Y and observed Z, we instead perform tests of observed X, observed Y, and imputed Z. This recasts the randomness underlying the null-distribution in our original test statistic of R^2 as a function of our permutation of Y and our imputation of Z. Here, we impute Z by sampling from the prior or posterior distribution of Z, depending on whether we're testing independencies before or after performing inference on our model's parameters. Moreover, we'll now compute the p-value of this test statistic by marginalizing over the permutations and imputations. Specifically, our p-value will be

$$E_{\text{samples, permutations}}\left[\mathbb{I}\left\{R^2\left(X, Y, Z_{\text{sampled}}\right) < R^2\left(X_{\text{sampled}}, Y^{\text{permuted}}_{\text{sampled}}, Z_{\text{sampled}}\right)\right\}\right] \tag{1}$$

where $\mathbb{I}$ represents the indicator function that equals one if the condition inside its braces is true and zero otherwise. For reference, this is the same as the p-value for test statistics (or discrepancies) defined in Gelman et al. (1996, Eq. 7).

Lastly, because our test statistic is itself a random variable that depends on the imputed values of Z, we must change our visualization method. Instead of plotting a single line versus a distribution, we now plot two distributions against one another. We first plot the distribution of "observed" test statistics that we computed using the observed X, observed Y and imputed Z values. Then, we plot the distribution of "sampled" test statistics using prior samples of (X, Y, Z) as a reference. Here, we have one value per imputed vector Z in both the observed and sampled distributions. However, for the distribution of "sampled" test statistics, we marginalize over permutations of Y since this distribution represents the null hypothesis of conditional or marginal independence.

That last paragraph may have been confusing, so we'll walk through an example. Figure 16 shows the causal graph implied by applying the deconfounder algorithm of Wang and Blei (2019) to our dataset. We describe this application more thoroughly in Section 7. For now, we present the graph to highlight the assumptions that we will test. Specifically, we'll examine the assumption that the observed number of licensed drivers in a household is independent of the observed number of automobiles in that household, conditional on the latent variable X^*.

Figure 17 shows the result of following the aforementioned testing procedures for assumptions involving latent variables. From the Figure, we see that the distribution of "observed" test statistics clusters around

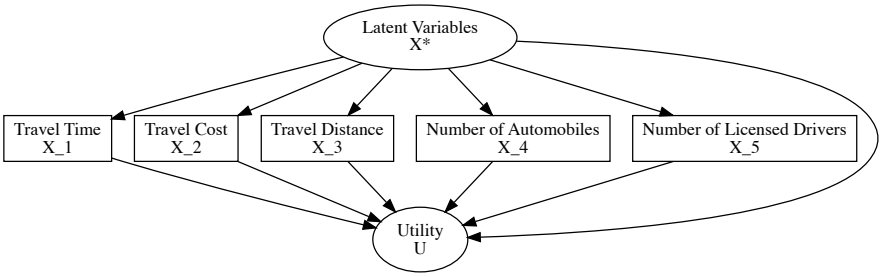

Figure 16: Causal graph from applying the deconfounder algorithm (Wang and Blei, 2019) to our dataset.

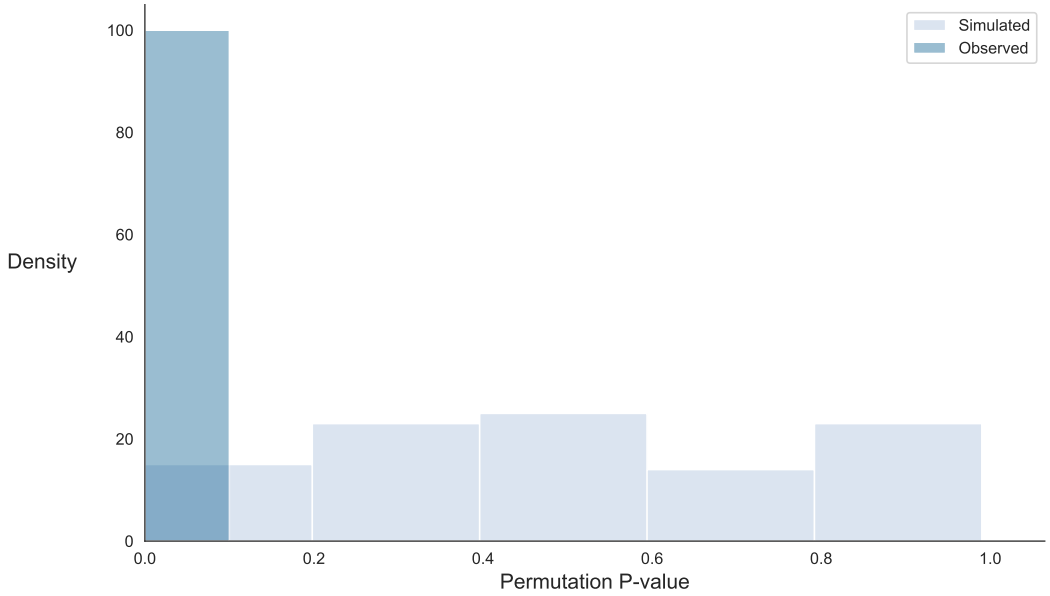

Figure 17: Results of testing that the number of drivers is independent of the number of automobiles in the household, conditional on the latent variable.

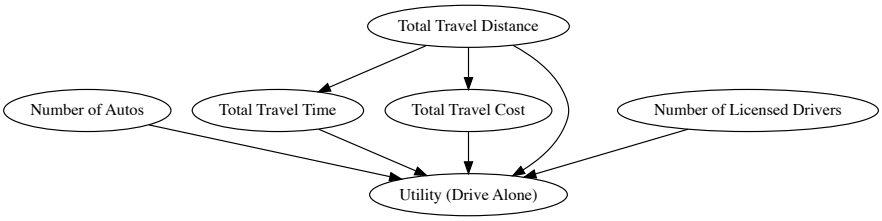

Figure 18: Expository causal graph of drive alone utility

zero while the distribution of "sampled" test statistics is closer to uniform. This result highlights the fact that (generally) there is "no free lunch": our approach to testing assumptions involving latent variables has its drawbacks. In particular, these tests are sensitive to assumptions about the joint prior distribution, $P_{\text{prior}}(X, Y, Z)$. If, as in this case[1], the observed data (X, Y) is unlikely under the joint prior distribution $P_{\text{prior}}(X, Y, Z)$ then the conditional independence test is likely to fail. As always, the failing test indicates that the observed data is unlike the data simulated and used to construct the reference distribution. Unfortunately, we are unsure of how much the dissimilarity comes from violations of conditional independence properties versus other sources. This highlights the need for extensive prior predictive checking of one's assumed joint prior, $P_{\text{prior}}(X, Y, Z)$, *before* using conditional independence tests on causal graphs with latent variables.

5.2 Demonstration

To demonstrate the testing procedures described above, we used the causal graph in Figure 18. This causal graph shows a set of hypothesized causal relationships between variables thought to contribute to the utility of the drive-alone travel alternative in our dataset. As drawn, this graph encodes multiple marginal and conditional independence assumptions. Tools such as Daggity (Textor et al., 2016) can be used to automatically infer all independencies based on one's graph. For didactic purposes, however, we focused our attention on two particular independence assumptions.

First, we tested the assumption of marginal independence between the number of licensed drivers and the number of automobiles in a household. A-priori, we may expect this independence to be have low-probability since the number of automobiles may generally be positively related to the number of licensed drivers in a household. Secondly, we tested the assumption of that travel cost was independent of travel time, conditional on travel distance. Unlike the first independence that we focus on, this conditional independence assumption is a-priori more credible. Despite their differing levels of a-priori credibility, we will see how we can test both using the procedures described in the previous subsection.

In particular, Figure 19 shows the results of using permutation, linear regression, and R^2 to test the hypothesis of marginal independence between the number of automobiles and the number of licensed drivers in a household. The empirical p-value of 0 confirms that the observed data is unlikely given the null-hypothesis of marginal, mean-independence. More specifically, when regressing the number of licensed drivers in a household on the number of cars in that household, one achieves an R^2 near 0.4. In contrast, when permuting the number of cars in the household and re-estimating the regression, the distribution of p-values concentrates around 0. This plot visualizes the fact that—through the lens of our chosen test statistic (R^2), linear regression model, and permutation-based resampling strategy—data generated under an assumption of marginal mean-independence does not "look like" the observed data. Accordingly, we should consider

[1] Details of the prior predictive checks that show prior-data mismatch are not shown due to space constraints. Please see `https://github.com/hassan-obeid/tr_b_causal_2020/blob/master/notebooks/final/_04-tb-testing-your-causal-graph.ipynb`

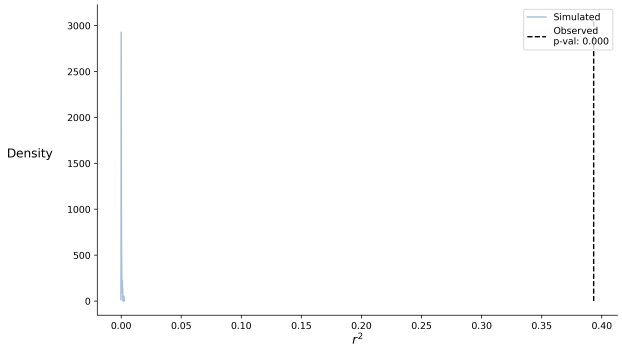

Figure 19: Marginal independence test results for the number of cars and licensed drivers in a household

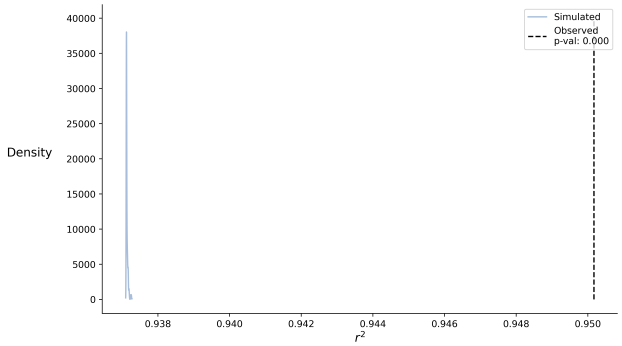

Figure 20: Conditional independence test results for travel time and travel cost given travel distance

the weaker assumption of marginal dependence, where we can make data-generating assumptions that offer greater realism and concordance with our observations.

In Figure 20, we have an analogous visualization of a conditional independence test. Here, we test the hypothesis that travel time is mean-independent of travel cost, conditional on travel distance. To execute this test, we model the conditional expectation of travel time as a linear function of travel cost and travel distance. As with the marginal independence test results, the R^2 of the model using the observed values of travel cost are greater than the model's R^2 using any simulated datasets. Again, this means that the R^2 using observed data is unlikely given our method of sampling from the null distribution of R^2 given conditional, mean-independence of travel time and travel cost.

As before, we should respond to this result by considering conditional dependence between our variables of interest. In particular, we should investigate the idea that travel time is associated with travel cost, conditional on travel distance. Why might this be the case? Does travel time cause travel cost when driving alone? Does travel cost cause travel time while driving alone? Does some other set of variables (potentially unmeasured) cause both travel time and travel cost?

Thinking through these questions, we can immediately think of latent variables that cause both travel time, travel cost, and the choice, even after conditioning on one's travel distance. For example, consider whether one drives alone over the San Francisco Bay Bridge. If one crosses the bridge, then traffic delays will likely increase one's travel time. Moreover, if one crosses the bridge, then one's travel cost is higher due to tolls that one must pay. And finally, if one takes a toll lane to get across the bridge faster, then one also pays a higher price. Overall, we can think of intuitive explanations for why there might be a dependence between travel time and travel cost, even after conditioning on travel distance. These explanations suggest particular relationships to analyze and particular variables, such as bay bridge crossings, to include in one's mode choice (i.e. outcome) model.

5.3 Additional techniques

The methods presented in this section have focused on testing independence assertions using one's dataset. While this approach may be the most accessible strategy for testing one's causal graph, there are other applicable techniques. For instance, Pitchforth and Mengersen (2013) put forth a detailed checklist of qualitative questions for one's causal graph. Answering these questions should increase the trustworthiness of one's graph. Alternatively, there are other quantitative tests of one's causal graph that were not explored in this section.

For instance, a causal graph makes assumptions about the number of independent variables in one's dataset: i.e., the nodes in one's graph without parent-nodes. One can test this assumption by estimating the "intrinsic dimension" of one's data, and testing whether the intrinsic dimension equals the number of independent variables implied by one's graph. For more information on estimating the intrinsic dimension of a dataset, see Camastra and Staiano (2016); Song et al. (2019). See Chenwei et al. (2019) for an extension of this idea for the situation where one suspects or cannot rule out unobserved confounding.

Another empirical implication of one's causal graph is the existence of so-called "vanishing tetrads" (Spearman, 1904). This term signifies that the difference between the product of two particular pairs of covariances must be zero. As stated, this implication of one's causal graph is hard to intuitively understand. However, one can graphically determine the existence of vanishing tetrads and determine which variables are part of these tetrads. Once one has determined the tetrads that should vanish according to one's graph, one can estimate the corresponding covariances and test to see if their difference of products is indeed unlikely to be zero. Such a test is yet another way to empirically determine whether one's graph is incompatible with one's dataset. For the original theorems proving that tetrads can be graphically identified and characterized, see Shafer et al. (1996) and references therein. For a more detailed and intuitive explanation of the graphical criterion for vanishing tetrads, see Thoemmes et al. (2018).

Finally, we note that there are still a whole host of other techniques for testing one's causal graphs. Many of these remaining techniques are useful when one's causal graph contains unobserved (i.e., latent) variables. On one hand, we can use "triad constraint" tests that test independence between "pseudo-residual" values and one's explanatory variables (Cai et al., 2019). Results from these tests are particularly useful for judging assumptions about how unobserved variables in our graph relate to each other and to our observed variables.

Relatedly, one can make use of constraints on entropies of our observed variables instead of independencies. The basic idea is that differing assumptions about the structure of the unobserved variables in our graph imply differing amounts of entropy in the variables that we do observe, so we should test for these differences in entropy. For more information and examples, see the literature about:

- inequality constraints, e.g. Tian and Pearl (2002); Kang and Tian (2006); Ver Steeg and Galstyan (2011)
- information inequalities, e.g. Chaves et al. (2014a)
- entropic inequalities, e.g. Chaves et al. (2014b)
- the inflation technique, e.g. Wolfe et al. (2019); Navascués and Wolfe (2020)

6 Causal Discovery

In the previous section, we detailed one method for testing the marginal and conditional independence assumptions of one's causal graph. As presented, this strategy requires one to first have a causal graph. Indeed, this requirement is why Section 4 provides instruction on how to construct an initial causal graph using expert opinion.

The unstated presumption is that one will test one's working causal graph, and if any of tests fail, one will then revise the graph until all assumptions appear plausible. Essentially, one proceeds in a loop of postulating, testing, and editing one's causal graph until it is not blatantly contradicted by one's data. If this sounds tedious, one can consider making one's computer work for you. Specifically, by exchanging the human-in-the-loop for a purely algorithmic endeavor, one arrives at the practice of causal discovery: inferring one's causal graph from one's data.

This section will describe why causal discovery is important, provide a brief overview of the main concepts in causal discovery, and show the results of using causal discovery algorithms on the dataset described in Section 2.

6.1 Why use causal discovery?

As a topic of study, causal discovery is important for numerous reasons. Three of these reasons are the following. First, causal discovery promotes robustness and understanding of our causal effect estimates. By comparing against our causal effect estimates when using a data-driven causal graph, we can begin to understand how our prior beliefs affect our estimates. Secondly, causal discovery helps inspire creation and editing of expert-opinion graphs. Indeed, it's typically easier to edit than to create from scratch. Graphs found via causal discovery can spark ideas for one's own graph if they feature different causal relations than are present in one's own graph, and they provide a starting point for graph revision. Thirdly, causal discovery algorithms can aid characterization of posterior uncertainty in one's causal graph and causal effect estimates. Each causal graph discovered from one's data represents an alternative way of understanding the world, and we can quantify the probability of each of these causal models representing our data generating process.

Let's begin with robustness. Here, one way of reducing the probability of incorrect inferences is to give oneself multiple chances to be correct. In particular, the Section 5 tested, expert-opinion based graph was never meant to be the stopping point in one's exploration of possible causal relationships. Instead, we advocate using multiple graphs to help us understand our causal effect estimates.

Specifically, consider that a causal graph is equivalent to a square matrix with one row for each variable, and binary entries that denote whether the variable for the row causes the variable on the column (Glymour et al., 2019, p. 3). A prior that expresses maximum ignorance over the graphs would be a prior where each matrix entry followed its own Bernoulli distribution with $p = 0.5$. Conversely, a prior expressing maximum certainty over the graph could be a prior with 100% certainty that our expert-opinion graph is the true causal graph. These two extremes represent reference priors that we can use in sensitivity analyses. By observing the distribution of causal effect estimates under each prior, we can reflect the uncertainty from not knowing the "true" causal graph, and we can begin to understand the ways that our estimates are sensitive our prior beliefs.

Beyond increasing our understanding of our estimates, causal discovery algorithms can help us create causal graphs based on expert-opinion. In particular, criticizing causal graphs is easy.

As noted by Pearl (1995, p. 708), "every pair of nodes in the graph waves its own warning flag in front of the modeller's eyes: 'Have you neglected an arrow or a dashed arc?" Additionally, the presence of directed causal relations is vividly placed before one's eyes for immediate criticism (e.g., is $X \rightarrow Y$ plausible?). Similarly, undirected or bi-directed edges between variables highlight causal ambiguity, thereby inviting analysts to resolve the question of directionality in the relationship. And conversely, we may learn from causal links (i.e. arrows) that are present in the graphs output from our causal discovery algorithm that we initially overlooked. By contrasting and criticizing alternative graphs, we clarify the strengths and deficiencies of our own point of view. Then, once we've identified elements that we think should or should not be present in a causal graph for our dataset, we can amend our hand-crafted graph to meet these requirements.

Lastly, causal discovery can help us characterize our posterior uncertainty about the data-generating causal graph. They enable approximation of the posterior distribution over causal graphs, in at least two ways. One approach is to use a weighted likelihood bootstrap approximation (Newton and Raftery, 1994) to the posterior distribution over graphs. In this approach, one would first sample a vector of weights for the likelihood terms. Then, one would use those weights in one's causal discovery algorithm to produce a single 'sample' from the posterior approximation.

Alternatively, we could use yet another randomize-then-optimize (Bardsley et al., 2014; Orabona et al., 2014) approach to sampling from an approximate posterior distribution. Here, one would sample from a prior on entries of the matrix representation of a causal graph. Semantically, we sample constraints such as "$X \rightarrow Y$ (MUST | MUST NOT) be in the causal graph". We "sample" from the approximate posterior by running the causal discovery algorithm on the original dataset, with the inclusion of these randomly generated constraints. And, of course, we can consider hybrids of these two posterior approximation schemes.

With these posterior approximation methods, we can generate causal graphs for all the purposes mentioned above:

- for criticism and inspiration,
- for distributional analyses of one's causal effect estimates conditional on each sampled graph, and
- for distributional analyses of the causal graphs themselves.
  I.e., how certain are we of any one causal graph?

6.2 Overview of causal discovery algorithms

The following subsection presents a non-exhaustive overview of causal discovery algorithms. For conciseness, an exhaustive review of causal discovery algorithms is out of the scope of the article. Crucially, we rely heavily on review papers such as Glymour et al. (2019) and Spirtes and Zhang (2016) to fill in our gaps.

Overall, there are three classes of causal discovery algorithms. One class attempts to directly infer the marginal and conditional independences in one's causal system. That is, this class of algorithms identifies a so-called Markov Equivalence Class (MEC) of graphs. Considering the discussion of independence testing in Section 5, this class of algorithms is perhaps best understood as repeated and systematic independence tests. These causal discovery algorithms are constraint-based algorithms, because the observed independencies represent constraints that define the space of plausible causal graphs for the dataset. Common constraint-based algorithms include the Peter-Clarke (PC) algorithm and the Fast Causal Inference (FCI) algorithm (Glymour et al., 2001). The PC algorithm assumes no unobserved confounding variables, whereas the FCI allows for (and sometimes infers) the presence of such unobserved confounders.

The second class of algorithms are known as score-based algorithms. They proceed sequentially by pairs of variables, making changes with a constraint on how their chosen score is affected by this change. First, we start from a fully disconnected graph, and we add directed causal relationships so as to increase the score of the generative model that corresponds to our graph. Then, once we have found the graph structure that maximizes the score on our data, we prune as many directed causal relationships as can be done without harming our score. In the end, the algorithms return the resulting set of graphs that retain maximal score. Common score-based algorithms include the Greedy Equivalence Search (GES) algorithm (Chickering, 2002). Typically these methods operate under stricter assumptions than constrain-based methods, but combinations of the two ideas have often outperformed methods that are strictly in either class (Glymour et al., 2019).

Lastly, the third class of algorithms relies on hypothesized Functional Causal Models (FCM) (Goudet et al., 2018) for each of the variables in one's system, as a function of their parents. The defining characteristic in such algorithms is that they exploit asymmetries in the residuals of models for the hypothesized relationships of $X \to Y$, $Y \to X$. In particular, the residuals will be independent of the hypothesized cause in only one of the two potential models. As noted in the description of such algorithms, they are especially useful for discovering causal relationships between pairs of variables. In other words, discovery methods based on FCMs enable orientation of the undirected or bi-directed edges that represent ambiguity over whether $X \to Y$ or $Y \to X$. This orienting capability can be usefully combined with the previous causal discovery algorithms that infer classes graphs with equivalent marginal and conditional independencies. And of course, extensions of these techniques exist for inference in the presence of unobserved confounding (Goudet et al., 2018, Sec. 6) and for learning an entire causal graph as opposed to only dealing with variable pairs (Zheng et al., 2020).

6.3 An application of causal discovery

To demonstrate the methods in this section, we chose to use the simplest causal discovery algorithm: the PC algorithm (Glymour et al., 2001). For practitioners who may reasonably start with the simplest method they can find and increase methodological complexity as needed, this should be an illuminating starting point. Moreover, as noted above, the PC algorithm assumes that one has no unobserved confounders. As noted in Section 5.2, we expect that our causal graph for the drive alone utility of our dataset likely contains unobserved confounders. Accordingly, using the PC algorithm on our example allows us to observe how the algorithm behaves under violation of its assumptions. Does the algorithm fail gracefully and point towards violations of its assumptions, or does it return incorrect results with certainty?

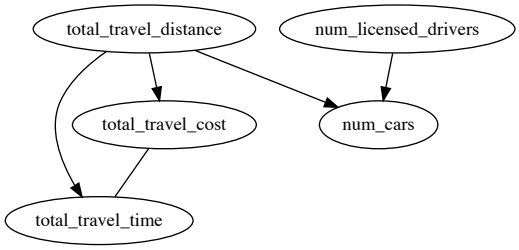

Figure 21: Result of using the PC algorithm on variables in the drive-alone utility

To begin, it helps to understand the general procedure of the PC algorithm. The algorithm uses all our observed variables to infer a skeleton of a causal graph. That is, the algorithm attempts to infer an undirected graph that denotes which variables relate to which other variables, ignoring the directionality of causation between them. Then, after inferring a skeleton, the algorithm attempts to orient the undirected edges as much as possible. Optimistically, one ends up with a fully directed acyclic graph. In other cases, one recovers a mix of directed and undirected edges, denoting a MEC of graphs with the same independence properties. In the final case, one recovers an undirected causal graph. This corresponds to the situation where one cannot infer which objects cause which other objects. One merely knows that given sets of objects relate to each other, not why this relationship exists.

With these basics understood, we can now present the results of using the PC algorithm to infer the causal graph for the variables present in Figure 18. For reference, the variables (travel time, travel cost, travel distance, number of automobiles in one's household, and the number of licensed drivers in one's household) are all thought to influence the utility of commuting by driving-alone. We care about the causal relationships between those explanatory variables because intervening on any one of these variables may cause downstream impacts on another explanatory variable. To make accurate causal inferences, we need to know and account for these differing pathways of influence on one's utility and mode choice when estimating the causal effect of interventions or treatments. Causal discovery helps us discover these differing pathways of influence by generating plausible causal graphs for our datasets.

Figure 21 shows the final result of applying the PC algorithm to the variables in our drive-alone utility function. Two main differences exist between this graph and the example graph shown in Figure 18. First, the number of licensed drivers is not independent of the number of automobiles in the household. The graph discovered via the PC algorithm depicts the number of automobiles in one's household as being a function of the number of licensed drivers that one has in one's household and as being a function of how far one has to travel to work or to school. The second major difference is the presence of an undirected edge between travel time and travel cost. The graph discovered by the PC algorithm states that travel time and travel cost are dependent, even after conditioning on travel distance.

As described in Section 5.2, independence testing foreshadows (indeed, determines) both of these results. Marginal independence testing showed us that the number of licensed drivers in a household and the number of automobiles in that household are not independent. Likewise, conditional independence testing showed us that travel cost is not independent of travel time, conditional on travel distance. Far from being a surprising congruence, this alignment is to be expected: conditional and marginal independence testing results are, in fact, used to generate the graph returned by the PC algorithm.

Relative to manual independence testing, what do we gain from using causal discovery algorithms? Critically, we gain at least three benefits from causal discovery algorithms. First, we gain insight into the dependence structure of our variables. For instance, we discover that the number of licensed drivers in a household may cause the number of automobiles in that household, as opposed to the number of automobiles causing the number of licensed drivers. Secondly, we gain a greater sense of uncertainty in our causal graphs. For instance, we can bootstrap our data and look at the distribution of inferred causal graphs. Moreover, even our discovery of graphs with undirected edges is an expression of uncertainty about the direction

of particular causal relationships. In our experience, modellers seldom express and quantitatively address uncertainty about the direction of causal relationships. Lastly, we gain comprehensiveness from performing a greater amount of tests than we might otherwise perform if only testing the assumptions of our currently-hypothesized causal graph.

7 Latent Confounding

This section focuses on the more complicated and realistic case that is typically faced by demand modelers, where we have latent confounders that affect the treatment assignment of our causal variables of interest. We first go over a few examples of confounding in transportation analysese, explain the challenges that come with such cases, and present a few approaches to dealing with this issue, specifically the recent de-confounder technique of Wang and Blei (2019). Our application will show how directed acyclic graphs can help increase transparency about one's reasoning regarding the number of confounders, assumptions for which variables are confounded, and the models needed to estimate the causal effects. In addition to the application above, we use simplified simulation scenarios to further investigate the usefulness and pitfalls of this approach for generating accurate model estimates.

7.1 Examples of confounding

Confounding occurs when a certain (confounding) variable induces variations in both the outcome as well as the treatment (policy) variables of interest, creating correlation between the treatment and the outcome that is not caused by the treatment variable itself. When the confounding variable (or variables) is observed, we can control for the confounding effect, and there exists many methods in the literature for how to do that, including post-stratification, multiple regression, propensity score methods, etc. It is when the confounding variable is unobserved that the problem becomes significantly more challenging. For an illustration, imagine we're interested [TODO: insert relevant transportation example with confounding]. Without accounting for the latent confounder, we risk getting very biased estimates of the effects of our treatment variable of interest, created by the induced variations by the confounder in both the treatment and the outcome variable.

The problem of latent confounding and ommitted variable bias is widely acknowledged in the transportation literature, and demand modelers can draw on a list of method to account for confouding in some specific circumstances. Integrated choice and latent variable (ICLV) models are a way to account for the effect of unobserved attitudinal variables which may affect the selection of some individuals into some treatment level, as well as an outcome of interest. For example, a person's unobserved beliefs and attitudes towards being environmentelly friendly may both affect whether she chooses to bike to work, as well as whether she lives close to a bike infrastructure, which may bias our observational study if we're interested in the effect adding a bikelane has on the mode share of bicycles. Those methods, however, rely on collecting attitudinal indicators typically obtained by conducting (usually) expensive surveys with a sample of the people. This may not be an option in many cases [TODO: list some examples where it's challenging to conduct surveys with people to collect indicator questions].

7.2 The deconfounder algorithm

One recent method that has been proposed to deal with the problem of latent confounding without the need to collect additional data is the deconfounder algorithm by Wang and Blei (2019). The method attempts to control for the confouding variable by estimating a "subsitute confounder", a set of variables that once controlled for, renders all variation in the treatment variables of interest exogenous. The process of applying the method is actually quite straightforward and simple. It proceeds as follows:

- First, estimate the substitute confounder using any good latent variable model the modeler chooses. The authors suggest estimating a factor model with k factors on the set of covariates the modeler is interested in.
- Second, check the factor model's accuracy using posterior predictive checks (add details).

- Once a sufficiently accurate latent variable model is recovered, use it to estimate an expected value of the latent variable for each observation, and control for this value in the outcome model, alongside the treatment variables and other covariates of interest.

One main assumption of the deconfounder algorithm is that the data at hand should have multi-cause confounders, thus the title of the paper, "The blessings of multiple causes". In other words, this method works when the unobserved confounders affect multiple of the observed causes (or treatment variables) of interest, alongside the outcome. This assumption is weaker than the one required for ignorability to hold, which requires the absence of both single cause and multi-cause confounders for accurate causal inferences.

7.3 Case study: Simulation

The purpose of this section is to investigate the effectiveness of the deconfounder algorithm (Wang and Blei, 2019) in adjusting for unobserved confounding. We use a simulated mode choice data where travel distance linearly confounds both travel time and travel cost. We then mask the travel distance data and treat it as an unobserved variable.

We estimate three models:

- Model 1: A multinomial logit with the correct original specification, EXCEPT we ommit the travel distance variable in the specification without trying to adjust for it. This model represents the worst case scenario where a modeler ignores, or is unaware of, unobserved confounding.
- Model 2: We use the deconfounder algorithm to try to recover the confounder (travel distance). In this method, we use all the variables in each mode's utility to recover that mode's confounder. This is in line with the approach taken in Wang and Blei (2019), where they use all the set of observed variables in the factor model to recover a substitute confounder.
- Model 3: We use the deconfounder algorithm to try to recover the confounder (travel distance), but this time, we only use travel time and cost in the factor model, instead of all the variables in the utility specification of each mode. By only using what we know are confounded variables to recover the substitute confounder, our goal is to analyze whether we can improve the accuracy of this approach by adopting a more thoughtful approach to hypothesizing on the confounder variables. This can be in the form of building and testing candidate causal graphs that try to illustrate this confounding.

We compare both the estimates of the coefficients on travel time and cost from each of those three models to the true estimates used in the simulation, as well as the distribution of the recovered substitute confounder under each of models 2 and 3 to the true confounder. The main findings of this exercise are the following:

Using the true variables believed to be confounded (i.e. method 3 where only travel time and cost are used to recover the confounder) leads to a better recovery of the true confounder. Figure 22 and 23 show a QQ plot of the true and recovered confounders under models 2 and 3 respectively. Looking at those figures, one can see that the distribution of the recovered substitute confounder under method 3 is closer to that of the true confounder than method 2. This suggests that it may be better to run the deconfounder algorithm based on a hypothesized causal graph, rather than just running it on all the observed covariates. Please refer to Section 4 for how to build plausible causal graphs.

Additionally, and perhaps most importantly, the effectiveness of the deconfounder algorithm is very sensitive to small errors and misfits in the recovered confounder. Although method 3 returns a relatively good fit of the true confounder (based on the QQ plot), the adjusted coefficients on travel time and cost do not exhibit any reduction in the bias resulting from ommitting the true confounder, and the coefficients on the recovered confounder are highly insignificant. This raises questions about the usefulness of the deconfounder algorithm in practice. While this limitation may raise questions on the usefulness of the deconfounder algorithm, it is important to point out that it is not only a by-product of the deconfounder algorithm itself. In fact, it is one of the built in characteristic of ommitted variable bias, and perhaps more broadly, the problem of error-in-variables in regression. To illustrate this, suppose we actually observe the true confounder variable, but with some white, random guassian noise. Figure 24 shows how the bias in the parameter of interest increases quickly as a function of the standard deviation of the random noise. This emphasizes the difficulty of recovering unbiased estimates in the presence latent confounders, and highlights potential limitations with methods that attempt to recover a substitute confounder to control for.

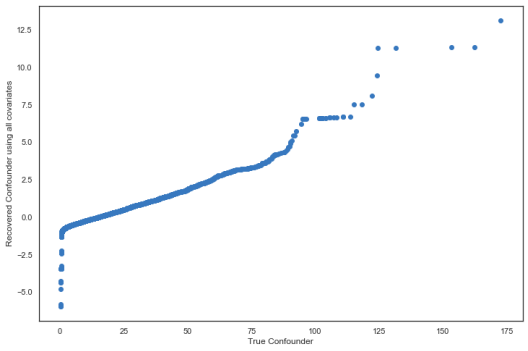

Figure 22: QQ plot of true confounder against recovered confounder using all covariates

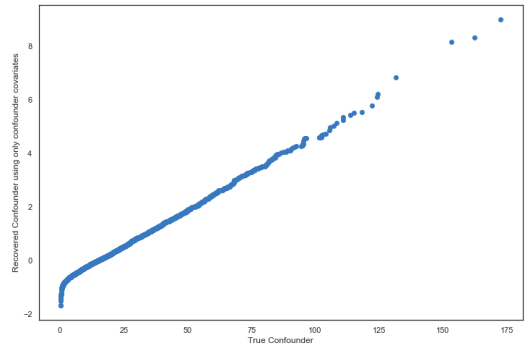

Figure 23: QQ plot of true confounder against recovered confounder using all covariates

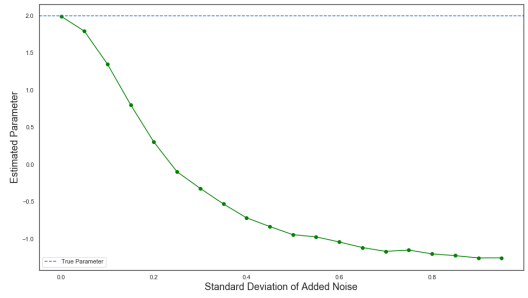

Figure 24: Sensitivity of causal estimates to random errors in the confounder variable

8 Conclusion

In this chapter, we presented a short simulation aimed at highlighting the importance of using causal graphs and causal inference methods in transportation demand modeling endeavors. This chapter was motivated by the existing disconnect between the fields of travel demand modeling and causal inference. Given the nature of transportation demand modeling efforts, one would expect that causal inference methods have already been used in the field. To our surprise, there has been very little mention of causal inference in the field of travel demand modeling. Brathwaite and Walker (2018) have documented this disconnect and presented an initial framework for addressing it.

Following up on their paper, this paper presents a numerical exercising aiming to show the importance of the data generating process in the estimation of effects of interest resulting from external "interventions". We showed, using the same utility specification model, that two data generating processes lead to bias in estimates of the effect of a policy intervention.

Data generating processes can be illustrated effectively using DAGs as shown by Judea Pearl. We have shown that DAGs help the researcher clearly represent their assumptions about the causal relationships within the data generating process. We presented a detailed process allowing researchers to construct causal graphs based on their expert opinion of the problem at hand, more specifically the variables of interest and their governing relationships.

However, researchers should not assume that the causal graphs they constructed are "correct" by default. These graphs carry many of the researchers beliefs on how the world operates. These relationships are represented through the connections between the coded nodes in the causal graph.

Before researchers start with their modeling efforts on tangible problems, they should ensure the assumptions encoded in their proposed causal graphs are valid. We present a method for testing the implications of the researcher's causal graph, both observed and latent.

Beyond testing the implications and assumptions encoded in a researcher's causal graph, we describe why causal discovery is important in discerning relationships between covariates within the data at hand and shortly present several causal discovery algorithms used by other researchers in different fields. We then move to highlighting the use of one of these causal discovery algorithms on a simple example. We showed that causal discovery methods help us test independence within our data, get a clearer picture about the uncertainty within the variables shown in the causal graph, and expand the level and amount of tests performed on our data when compared to the tests implicated by the structure of the researcher's proposed causal graph.

We then summarize one of the existing methods aimed at dealing with more realistic cases than the one illustrated in the selection on observables simulation. We present examples of latent confounding within the transportaion demand modeling field and highlight some examples of current methods aiming at addressing problems resulting from confouding in certain situations. We present a recently developed algorithm aiming to address the problems not addressed by previously developed methods dealing with confouding. We use this model on the same example illustrated in the selection on observables simulations and highlight the instances where it leads to a better recovery of a confounder in one's causal graph.

We then show that while the deconfounder algorithm might be a promising step in the right direction, it still has some notable deficiencies. We show that the deconfounder algorithms shows to be sensitive to small errors in one's data.

References

Johnathan M Bardsley, Antti Solonen, Heikki Haario, and Marko Laine. Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1895–A1910, 2014.

Moshe Ben-Akiva, Joan Walker, Adriana T Bernardino, Dinesh A Gopinath, Taka Morikawa, and Amalia Polydoropoulou. Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges*, pages 431–470, 2002.

Wicher Pieter Bergsma. *Testing conditional independence for continuous random variables*. Eurandom, 2004.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

Mark Bradley, John L Bowman, and Bruce Griesenbeck. Sacsim: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3(1):5–31, 2010.

Timothy Brathwaite and Joan L Walker. Causal inference in travel demand modeling (and the lack thereof). *Journal of choice modelling*, 26:1–18, 2018.

Samuel Burkart and Franz J Király. Predictive independence testing, predictive conditional independence testing, and predictive graphical modelling. *arXiv preprint arXiv:1711.05869*, 2017.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, pages 12883–12892, 2019.

Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

Emmanuel Candès, 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.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

R Chaves, L Luft, TO Maciel, D Gross, D Janzing, and B Schölkopf. Inferring latent structures via information inequalities. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 112–121, 2014a.

Rafael Chaves, Lukas Luft, and David Gross. Causal structures from entropic information: geometry and novel scenarios. *New Journal of Physics*, 16(4):043001, 2014b.

DING Chenwei, Mingming Gong, Kun Zhang, and Dacheng Tao. Likelihood-free overcomplete ica and applications in causal discovery. In *Advances in Neural Information Processing Systems*, pages 6883–6893, 2019.

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

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

Clark Glymour, Richard Scheines, and Peter Spirtes. *Causation, prediction, and search*. MIT Press, 2001.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

Marco Henrique de Almeida Inácio, Rafael Izbicki, and Rafael Bassi Stern. Conditional independence testing: a predictive perspective. *arXiv preprint arXiv:1908.00105*, 2019.

Changsung Kang and Jin Tian. Inequality constraints in causal models with hidden variables. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 233–240, 2006.

Miguel Navascués and Elie Wolfe. The inflation technique completely solves the causal compatibility problem. *Journal of Causal Inference*, 8(1):70–91, 2020.

Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.

Francesco Orabona, Tamir Hazan, Anand Sarwate, and Tommi Jaakkola. On measure concentration of random maximum a-posteriori perturbations. In *International Conference on Machine Learning*, pages 432–440, 2014.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Jegar Pitchforth and Kerrie Mengersen. A proposed validation framework for expert elicited bayesian networks. *Expert Systems with Applications*, 40(1):162–167, 2013.

Glenn Shafer, Alexander Kogan, and Peter Spirtes. Vanishing tetrad differences and model structure. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 4(03):209–224, 1996.

Rajen D Shah, Jonas Peters, et al. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

Jing Song, Satoshi Oyama, and Masahito Kurihara. Identification of possible common causes by intrinsic dimension estimation. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8. IEEE, 2019.

C Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. Springer, 2016.

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.

Felix Thoemmes, Yves Rosseel, and Johannes Textor. Local fit evaluation of structural equation models using graphical criteria. *Psychological methods*, 23(1):27, 2018.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527, 2002.

Greg Ver Steeg and Aram Galstyan. A sequence of relaxations constraining hidden variable models. *arXiv*, pages arXiv–1106, 2011.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Elie Wolfe, Robert W Spekkens, and Tobias Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425, 2020.