Some Title

All of us

December 1, 2020

Abstract

In this chapter, we focus on using causal graphs to make causal inferences in choice modelling contexts. We address a longstanding disconnect whereby choice modellers have not adopted techniques from causal inference researchers, even when trying to make causal inferences. To bridge this disconnect, we conduct simulation studies to first demonstrate the need for paying close attention to causal graphs in choice modelling. We then present new guidelines, methods, and perspectives for the construction and validation of causal graphs, complete with empirical examples. Next we give examples and direction for using one's causal graphs to make causal inferences in common choice modelling scenarios, including those with latent confounding. Finally, we also provide extensive literature reviews of testing and discovery methods for causal graphs. At the chapter's conclusion, choice modellers should have a much clearer understanding of (or references to find out about) why they should use causal graphs, what graphs to use for their dataset, and how they can use a given causal graph to complete their causal inferences.

1 Introduction

In transportation, we often build models that are used for policy evaluation. Specifically, we often use behavioral models to evaluate the impact of external interventions on travel outcomes. These evaluations are excplicitly causal: we are interested in how the system reacts to *interventions*. Yet, when these models are developed, causality is often not considered. Furthermore, when causal concepts are accounted for, the process is done implicitly without a formal framework.

This chapter is concerned with filling this gap. Specifically, we focus on addressing the existing disconnect between the fields of travel demand modeling and causal inference. The chapter is motivated by the current lack of use of methods and findings from the causal inference literature in travel demand modeling.

While the field of transportation demand modeling could benefit greatly from incorporating causal inference techniques, there are barriers that have slowed this integration. 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 fundemental difference is that demand modelers are typically trying to forecast the impacts of policies that haven't been implemented or seen before. Forecasting the effects of unseen interventions requires additional work and a change to the typical causal modeling workflow. In particular, we must translate a given policy (treatment) into a set of characteristics and variables that exist in the data and system at hand. (For a more thorough discussion of this and other barriers, please refer to Brathwaite and Walker (2018a).)

While these barriers complicate the efforts of demand modelers, there is still a lot to gain from incorporating causal inference techniques where appropriate and from contributing to the causal inference literature where it's lacking.

The relevance of this topic now is motivated by the significant boost in the causal inference literature recently, both in the potential outcomes and the causal graphical modeling frameworks. The goal of this chapter is to formalize a workflow 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 or 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 with 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 unobserved confounding when collecting additional data is not feasible. We describe this problem in more details in section 7.

2 Perils of disregard

As stated in 1, the development of transportation demand models aims at evaluating the impact of policies on a certain transportation demand related outcome. As an example, consider the proposals from the following fictitious scenarios: - Based on input from the public, the Department of Transportation (DOT) of is considering implementing a new parking policy. This proposed policy would put in place pricing changes and parking restrictions that would discourage individuals from parking in the central business district. The DOT (and its constituents) believes that this policy would encourage people to use more active transportation modes (e.g: walking, biking, etc..). - A certain DOT is considering constructing a streetcar (trolley, tram) line in a low/mid-income area in its jurisdiction. Citing examples from other cities and countries, the DOT claims that the new streetcar line will create more transit oriented developments and increase economic activity in the areas surrounding the proposed project.

Thinking more closely about these proposals, it is clear that they assume a causal relationship between the proposed project/policy and the desired goals. The DOT in question analyzed data, concluding that such a policy or project would cause the desired output and achieve the desired goal. In the presented scenarios, the DOT claims that implementing the new parking policy will cause an increase the share of active transportation modes in the central business district and that constructing the proposed streetcar line will cause more transit oriented developments and economic activity.

Policymakers base their analyses and conclusions on hypotheses or beliefs of how the world operates. In other words, the data analysis is based on a certain belief about the data generating process. However, policymakers often do not present their beliefs about such data generating process. As a result, these proposals maintain an obscure representation of how the policy or project will achieve their desired goals.

DAGs allow one to clearly encode their assumptions about the data generating process and the problem at hand. Researchers and practitioners have made use of DAGs in fields ranging from medicine and epidemiology (**?**, **?**) to economics (**?**) and have found them to be practical.

Likewise, as in the fields presented above, DAGs could prove useful in addressing transportation policy questions, similar to the ones presented above. Brathwaite and Walker (2018a) have proposed a framework illustrating how practitioners and researchers can use DAGs to answer such transportation modeling questions in a causal context. However, Brathwaite and Walker (2018a) have not shown an empirical application of their framework and how it results in different conclusions when compared to traditional modeling approaches.

In this section, we will present an example illustrating the importance of using DAGs in transportation demand modeling efforts. Specifically, we will illustrate how different assumptions about the data generating process result in different conclusions. To make our point, we build upon Brathwaite and Walker (2018a). Here, we present an empirical exercise using a simplified transportation modeling problem. Before going any further, we note that this example is illustrative, and it does not reflect all complexities in a typical transportation choice modeling problem. Indeed, our goal in this section is not to recover the causal effect of the proposed intervention. Instead, we aim to only show how different DAGs would result in different conclusions about the effect of the proposed intervention.

Let us assume that a company wants to reduce its workforce carbon footprint by moving its employees closer to their campus. We would like to forecast how such an intervention would change the share of employees driving to work. We model this travel mode choice problem based on a dataset from Brathwaite and Walker (2018b). This dataset is based on the 2012 California Household Travel Survey, and it contains

approximately 4000 home-based school or work tours made by approximately 3850 individuals in the California Bay area. Readers interested in a more detailed description of the dataset can refer to Brathwaite and Walker (2018b). For purposes of this exercise, we consider the Multinomial Logit model defined in Brathwaite and Walker (2018b) as the true outcome generating process.

The systematic utility equations of the adopted multinomial logit model are specified as follows:

$$
\begin{aligned}
U_{\text{da}} &= \beta_{\text{travel\_time}} \times \text{Travel\_Time} + \beta_{\text{cost\_per\_distance\_da}} \times \text{Cost\_per\_Distance}_{\text{da}} \\
&\quad + \beta_{\text{autos}} \times \text{Number\_of\_Autos} \\
U_{\text{sr2}} &= ASC_{\text{sr2}} + \beta_{\text{time\_drive}} \times \text{Travel\_Time} \\
&\quad + \beta_{\text{cost\_per\_distance\_sr2}} \times \text{Cost\_per\_Distance}_{\text{sr2}} + \beta_{\text{autos}} \times \text{Number\_of\_Autos} \\
&\quad + \beta_{\text{cross\_bay}} \times \text{Cross\_Bay} + \beta_{\text{hh\_size}} \times \text{HH\_Size} \\
&\quad + \beta_{\text{n\_kids\_hh}} \times \text{Number\_of\_kids} \\
U_{\text{sr3+}} &= ASC_{\text{sr3+}} + \beta_{\text{time\_drive}} \times \text{Travel\_Time} \\
&\quad + \beta_{\text{cost\_per\_distance\_sr3+}} \times \text{Cost\_per\_Distance}_{\text{sr3+}} + \beta_{\text{autos}} \times \text{Number\_of\_Autos} \\
&\quad + \beta_{\text{cross\_bay}} \times \text{Cross\_Bay} + \beta_{\text{hh\_size}} \times \text{HH\_Size} \\
&\quad + \beta_{\text{n\_kids\_hh}} \times \text{Number\_of\_kids} \\
U_{\text{WTW}} &= ASC_{\text{WTW}} + \beta_{\text{travel\_time\_transit}} \times \text{Travel\_Time} \\
&\quad + \beta_{\text{travel\_cost}} \times \text{Travel\_Cost} \\
U_{\text{DTW}} &= ASC_{\text{DTW}} + \beta_{\text{travel\_time\_transit}} \times \text{Travel\_Time} \\
&\quad + \beta_{\text{travel\_cost}} \times \text{Travel\_Cost} \\
U_{\text{WTD}} &= ASC_{\text{WTD}} + \beta_{\text{travel\_time\_transit}} \times \text{Travel\_Time} \\
&\quad + \beta_{\text{travel\_cost}} \times \text{Travel\_Cost} \\
U_{\text{Bike}} &= ASC_{\text{Bike}} + \beta_{\text{travel\_distance\_bike}} \times \text{Travel\_Distance} \\
U_{\text{Walk}} &= ASC_{\text{Walk}} + \beta_{\text{travel\_distance\_walk}} \times \text{Travel\_Distance}
\end{aligned}
$$

For our purposes, we assume that there are no latent variables. The variables were chosen based on the specification of the MNL model outlined by the equations above. 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 in individual i's household.
- Cross-bay trip: is a binary variable indicating whether the trip t in tour l for individual i is a cross-bay trip.

Figure 1 illustrates the DAG where all explanatory variables are independent of each other for each utility equation. 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 mode in the multinomial logit model specified in Brathwaite and Walker (2018b).

Using the setup described above, we conduct our simulation exercise as follows:

Figure 1: Causal Graph with Independent Covariates

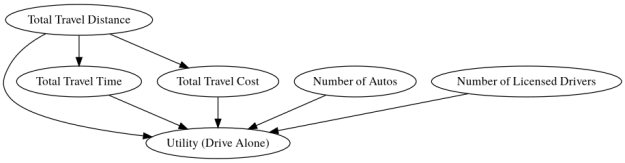

Figure 2: Causal Graph for the Drive Alone Utility Function

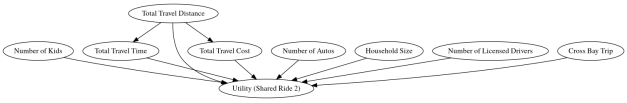

Figure 3: Causal Graph for the Shared Ride 2 Utility Function

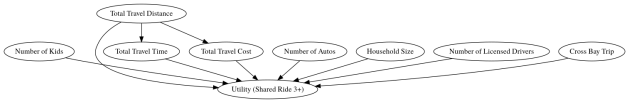

Figure 4: Causal Graph for the Shared Ride 3+ Utility Function

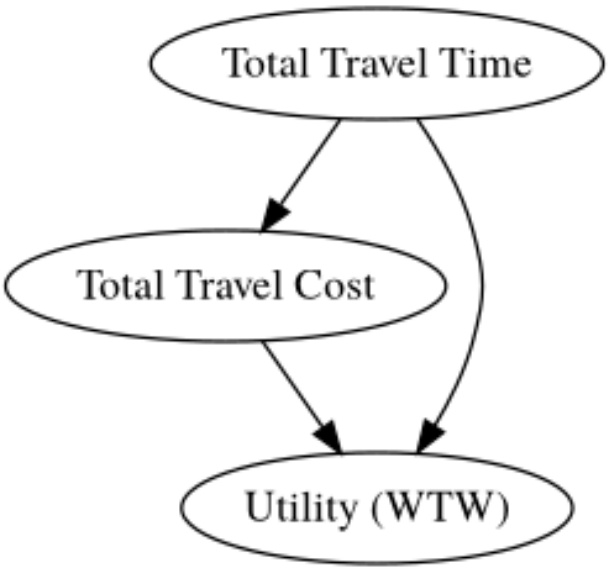

Figure 5: Causal Graph for the Walk-Transit-Walk Utility Function

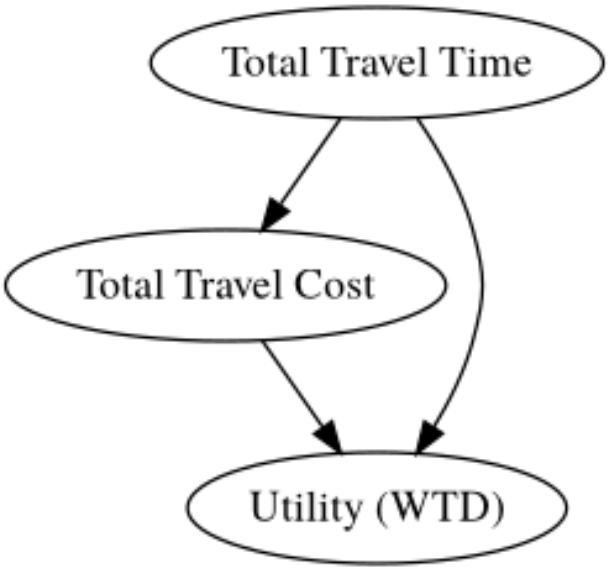

Figure 6: Causal Graph for the Walk-Transit-Drive Utility Function

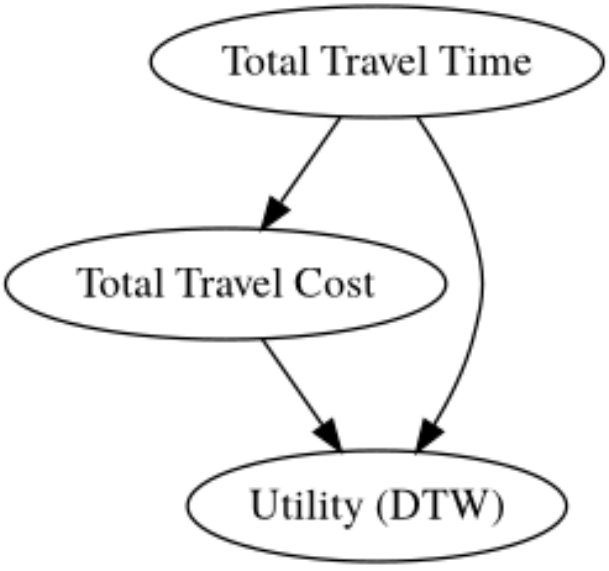

Figure 7: Causal Graph for the Drive-Transit-Walk Utility Function

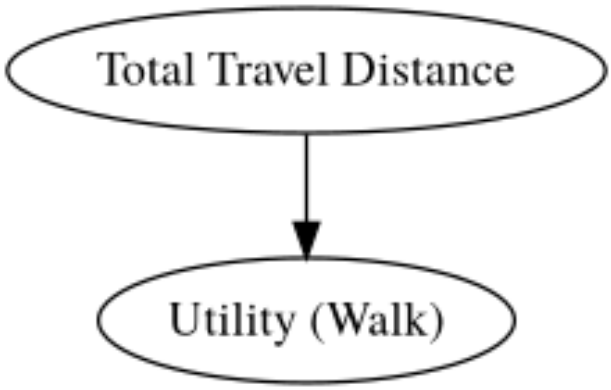

Figure 8: Causal Graph for the Shared Ride 3+ Utility Function

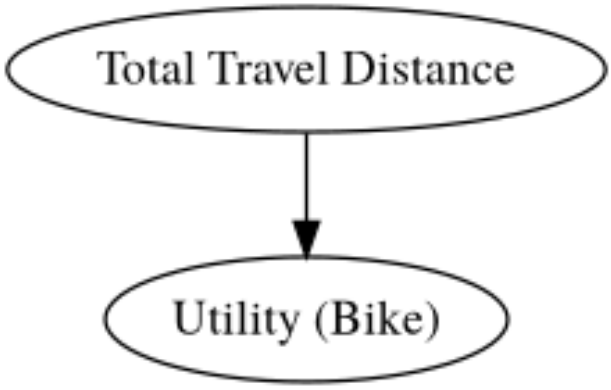

Figure 9: Causal Graph for the Bike Utility Function

1. We build a DAG (Figure 1) where we assume all explanatory variables are independent of each other.
2. We simulate data based on this graph and generate outcomes based on a chosen "true" multinomial logit model.
3. We predict outcomes based on the mode choice model.
4. We apply the do-operator Pearl (2000) to our simulated data to reduce the travel distance for all individuals in the dataset. This emulates a company's decision to move its employees closer to campus.
5. We predict new outcomes based on the mode choice model.
6. We build a different DAG (Figure 2 through Figure 9) for each utility function based on assumptions of how explanatory variables interact and influence the outcome.
7. We use the real dataset to estimate the relationships between different variables outlined in Figure 2 through Figure 9.
8. We simulate data for variables without parent nodes in Figure Figure 2 through Figure 9 and we use the estimated relationships from step 6 to simulate the remaining explanatory variables.
9. We then use the choice model to simulate outcome choices based on data simulated for Figure 2 through Figure 9.
10. We apply the do-operator, as above, to reduce the travel distance for all individuals in the dataset and to emulate a company's decision to move its employees closer to campus.
11. We use the estimated relationships for this causal graph to simulate all the explanatory variables in Figure 2 through Figure 9(namely the variables affected by travel distance).
12. We use the estimated choice model to produce outcomes based on the simulated data from the previous step.

We repeat this simulation process several times and recover the choice probability of all modes for all individuals. For our purposes, we focus on car-centric modes (Drive alone, Shared ride with another person, and a shared ride with two or more individuals). We compute the difference in mode choice probabilities from different models based on the two constructed DAGs.

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. To prepare readers to create causal graphs, the next section will provide a brief overview of them and their history in choice modelling. Then, Section 4 explores this topic and includes some guidance on how to build causal graphs representing the researchers beliefs about the data generating process. Section 5 follows up with guidance on how to test one's causal graphs against one's data.

3 Overview of Causal Graphs

The previous section, 2, motivated why the knowledge encoded in one's DAG is important. This section gives a high level overview of what DAGs are and provides references for more thorough readings on the subject. Causal diagrams and causal graphical models have been introduced by **?** as a powerful tool for

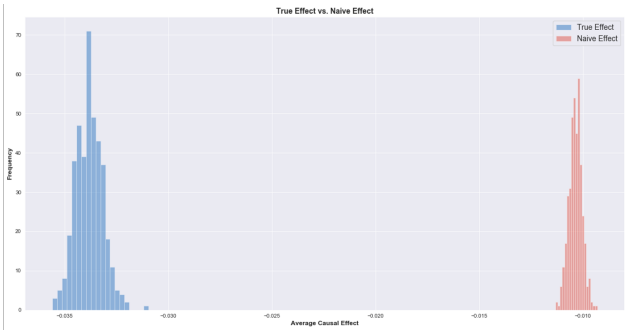

Figure 10: Histograms of the probability of choosing Car Centric Modes under Different Data Generating processes.

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 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 a treatment Z on Y. What the graph in Figure 11 implies is that Z is independent of Y(Z) given X. In causal jargon, the mechanism by which the treatment Z was assigned is ignorable, once we control for covariates X. In such situations, it is sufficient to control for X to obtain an unbiased estimate of the causal effect of Z on Y.

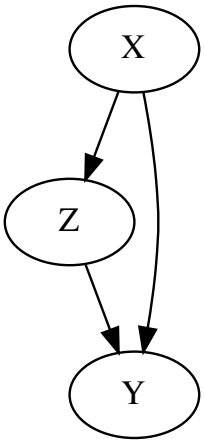

Figure 11: Simple DAG of the causal relationships between X, Y and Z

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:

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

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

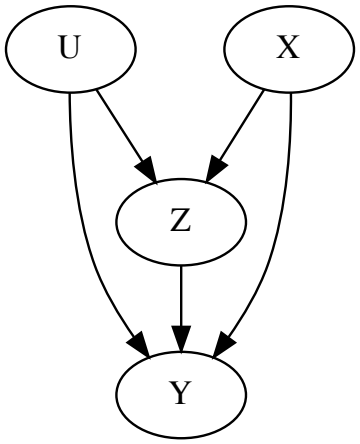

Figure 12: Simple DAG of the causal relationships between X, Y, Z and confounder U

Figure 12 shows an indirect connection between Z and Y that goes through U. Therefore, even if we condition on X, omitting U will yield biased results of the relationship between Z and Y, due to the variation in both Z and Y caused by U. The structural equations also show that the ignorability assumption of Z does not hold if we only control for X, and we risk obtaining biased estimates of the causal effect of Z on Y if we fail to account for U, sometimes referred to as a common cause. Even in this very simple example with only three or four variables, we gain an advantage in parsimony and expressiveness by using a causal graph to encode assumptions. The conciseness of DAGs becomes even more useful in real-world applications. Here, applied problems may have many available covariates and a complex causal structure. In such problems, fully writing out the structural equations for the system may be tedious, and readers may still find it hard to understand.

Another benefit of DAGs is that they come with testable implications. Incorporating these tests in any causal analysis adds robustness and defensibility to one's study. In section 5, we discuss these implications further, and we illustrate how to tests for them 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, even though we will use more familiar statistical tools to actually infer/estimate a causal effects of interest. This separation of identification and inference into two separate problems is often advocated for by Judea Pearl himself, emphasizing that whether a causal effect of interest is identifiable does not

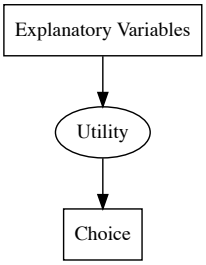

Figure 13: Archetypical RUM causal diagram

3.1 Prior Uses of Causal Graphs in Choice Modelling

In our last subsection, we reviewed the basics of causal graphs: what are they and why are they useful? We targeted that subsection at choice modellers who are unfamiliar with these tools. However, readers should be aware of the history of causal graphs in choice modelling. Indeed, choice modellers have used causal graphs (in limited fashion) for years. Here are three examples to illustrate our point. First, consider the case of Random Utility Maximization (RUM) models.

For decades, choice modellers have used stylized DAGs to depict RUM models. In particular, these diagrams illustrate an assumed choice-making (i.e., causal) process. As an example, see Figure 13, reproduced from Figure 1 of Ben-Akiva et al. (2002). Ben-Akiva et al. draw a two-part process. First, they assume that explanatory variables cause unobserved utilities. Then, they assume that the unobserved utilities cause the choice.

Note that although RUM diagrams adequately show the choice process, we still call them stylized. Specifically, these diagrams lack detail about the relationships between explanatory variables. When speaking collectively, we cannot tell if one explanatory variable causes another. Unfortunately, as shown in Section 2, such knowledge is crucial. Without more detailed causal knowledge, our inferences may be inconsistent and arbitrarily bad.

Besides RUM models, causal graphs appear in the literature on Integrated Choice and Latent Variable (ICLV) models. An example of an ICLV model is in Figure 14, based on Figure 5 of Ben-Akiva et al. (2002). As with RUM models, ICLV causal graphs represent assumptions about the choice process. Here concern typically centers around unobserved mediators. These unobserved variables cause the outcome and are caused by one's observed covariates. Omitting such mediators leads to models that misrepresent the assumed choice process. As a result, researchers care deeply about ICLV models that avoid these behavioral misrepresentations.

Finally, consider activity based travel demand models (ABM). 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. 2). The purpose of these graphs is to explain the structure of the entire system of outcome models.

In particular, ABM diagrams highlight two sets of researcher assumptions. They detail the researcher's beliefs about which choices, i.e. outcomes, precede others. For instance, work location choice preceding auto-ownership choice. ABM diagrams also detail which downstream choices partially cause which preceding ones. For example, considerations about one's travel mode choice influences one's daily activity choices. This happens even though the activity choices precede the travel mode choices.

Two main features distinguish RUM, ICLV, and ABM graphs from the causal graphs of Pearl (1995). First, causal graphs in choice modelling traditionally ignore relations between the explanatory variables. Typically, choice modelling causal graphs show all explanatory variables together as a monolith. The relationships between explanatory variables is often not specified. Even worse, researchers may tacitly treat the variables as if they are jointly independent. In the language used by a large group of causal inference

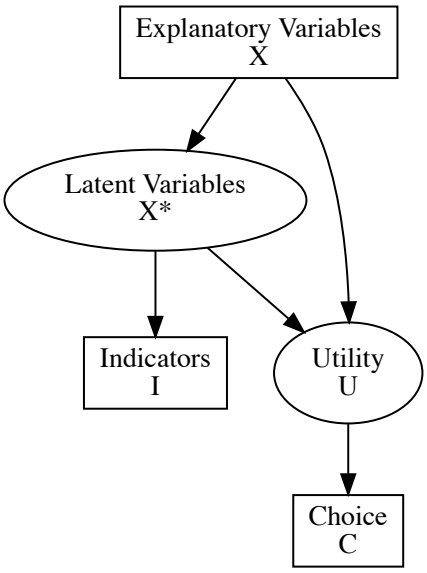

Figure 14: Archetypical ICLV causal diagram

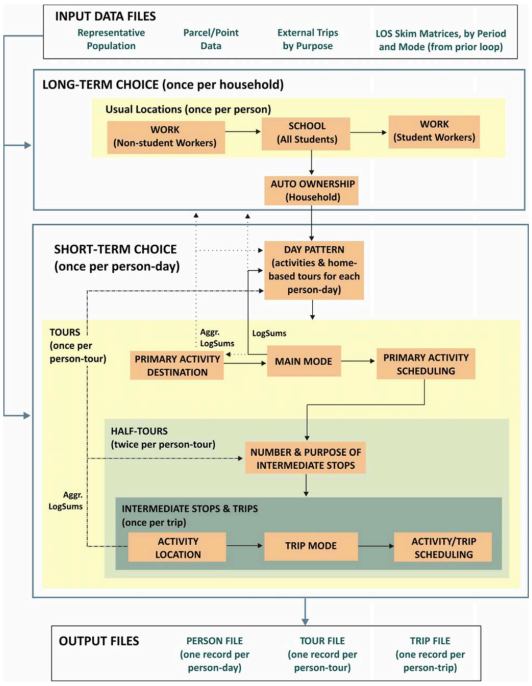

Figure 15: Archetypical causal diagram for activity-based models

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 model themselves. In particular, choice modellers rarely test the predictions of causal graphs. In doing so, we underuse causal implications such as statistical independence. Worse, we possibly violate these implications with great frequency. Such actions ignore the efforts from causal inference in computer science. There, researchers stress using their data to test the implications of their causal graphs.

In conclusion, choice modellers have long made use of causal graphs. In select contexts, causal graphs convey causal assumptions about choice processes. Thus far, however, our field has underutilized these tools. We seldom use causal graphs to encode our assumptions about how our explanatory variables came to be. Moreover, we do not routinely test causal graphs against empirical choice data.

These two issues are opportunities for choice modelling to gain from the insights and methods of causal inference scholars. The rest of the chapter will focus on the following three topics. How we can construct causal diagrams that pay attention to explanatory and outcome variables? How we can test a given causal graph against one's data? How can we deal with "real-world" graphs featuring 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 could cause bias in our causal inferences,
4. connect all variables in our graph according to our a-priori beliefs about causal relations amongst them
5. consider how the graph structure may differ between individuals and subgroups within the population.

The following paragraphs 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 in this section, we will devote time to thinking about how subgroups and individual heterogeneity may affect our causal graphs.

Now, we add our first variable(s) to our graph, the outcome variable(s) of interest in our problem. Then, we list all the variables we believe to influence the outcome(s). We refer to these variables as our initial explanatory variables.

Next, we iterate through these 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[1]. Note that some effect modifiers may be a part of our list of initial explanatory variables. For any effect modifiers that we think of, outside of the list of initial explanatory variables, we add them to our causal graph.

Overall, modifiers are important because our treatment effects systematically vary with them. Accordingly, if better understand when our treatments will be effective, then we can better target them. For instance, imagine that a region-wide lockdown reduces the 14-day rolling average of new COVID-19 cases by X% (on average). Of course, we know that a lockdown's effectiveness is modified by the percentage of workers who must continue going out to work. If most residents in an area are essential workers, then a lockdown will be less effective there, as compared with other locales. We might wish to target other interventions for that region, as a replacement or supplement for the lockdown. Targeting aside, knowledge of modifiers is also crucial to generalizing treatment effect inferences from one population to another. To credibly transport our inferences, we must know what variables cause the treatment effects to differ between populations, and we must know how the distributions of those variables differs across populations (Pearl and Bareinboim, 2014).

After adding explanatory and effect modifying variables to the graph, we turn our attention to mediating variables. A mediating variable is one through which an explanatory variable influences our outcome(s) of

[1]Note, effect modifiers and confounders are easily confused. Both variables cause the outcome. The difference is that effect modifiers do not cause the explanatory / treatment variables. Confounders do.

interest. Such variables have multiple uses. Under certain instances of confounding, mediators enable the "front-door" criterion to identify one's causal effect (Glynn and Kashin, 2018; Bellemare and Bloem, 2019; Gupta et al., 2020). Similarly, subject to particular causal assumptions, mediating variables permit inference on long-term outcomes of a selected intervention, given only its short-term proxies (Athey et al., 2019; Yang et al., 2020).

To find these mediators, we again iterate through each explanatory variable. On each iteration, we brainstorm variables along paths of influence from our explanatory variable to our 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 the primary (or sole) mediator 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. As an example, consider a person's attitude towards environmental conservation. This attitude may cause both that individual's observed distance to their workplace (another explanatory variable) and that individual's choice of travel mode. Both in this example and in general, we should add such confounding variables 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 or due to the problem itself. In reality, how do we think our system evolves over time? If we consider multiple observations of a given decision maker, how does that decision maker's observed variables at time t partially cause future variables important to the context or outcome(s) for that decision maker at time $t' > t$? 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 to our causal graph all the outcome, explanatory, effect modifying, mediating, confounding, and time-indexed variables that we believe are relevant for our problem. However, they are all disconnected nodes, i.e., 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 would bias our causal effect estimates.

To remove the problematic variables, we iterate through each of the non-outcome variables in our graph, and we assess whether each 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 about our problem. Specifically, we should conduct a literature review to see how other researchers have conceptualized the topic 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? How should these variables enter our graph? Do the included variables of other researchers 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 a confounder, when we solely thought of the variable as an effect modifier? As we answer these questions, we should 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. Pause. Take a moment to look over the graph to ensure there are no remaining singletons and that we have not drawn any spurious connections. Then, take a moment to celebrate. Drawing a project'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, if any, exist in your population of interest? Are there any causal relationships that should, or should not, 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 that are specific to this sub-population? (I.e., for a given sub-population, does $B \rightarrow A$ 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 modified causal graphs for each sub-population of interest. 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 test it against our available data. Doing so will add robustness and credibility to our final conclusions. 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. These efforts rely on additional assumptions about the variables or the test statistic itself. Some researchers create tests 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)).

In this chapter, we will take an easier and less decisive route to testing independence. If a pair of variables have conditionally or marginally independent distributions, then their statistical moments will also be conditionally or marginally independent. Accordingly, we will not test for marginal or conditional independence in distribution. We will instead 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. Mean independence simply provides justification for placing greater belief in the variables being distributionally independent.

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

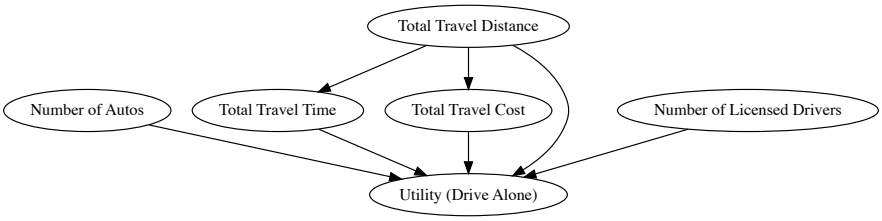

Figure 16: Expository causal graph of drive alone utility

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. First, we have to select models for $E[Y \mid X, Z]$ and $E[Y \mid Z]$, respectively, for testing conditional and marginal mean-independence. Second, we 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 used R^2 as our test statistic for judging the regressions' predictive performances. Third, we resampled X without replacement, keeping the length of the resampled vector equal to the length of the original vector. In other words, we permuted X. Finally, we visualized our tests by encoding our observed test-statistic as a vertical line and by plotting the kernel density estimate of our test statistic's distribution.

Our rationales 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 through 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, the random forest-based tests resulted in non-uniform p-values. Such non-uniform p-values makes it harder to a-priori reject a false model (Gelman et al., 2013). When we switched from the combination of random forests and permutations to linear regressions and permutations, our p-values turned out to be empirically, uniformly distributed. Moreover, we still retained high statistical 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 dataset and simulations though, the combination of linear regressions, permutations, and R^2 resulted in adequate tests of marginal and conditional mean-independence.

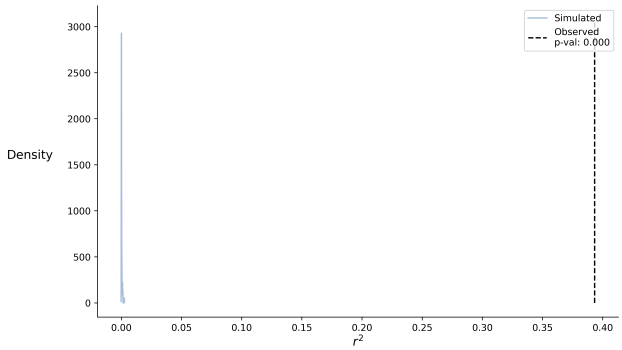

Figure 17: Marginal independence test results for the number of cars and licensed drivers in a household

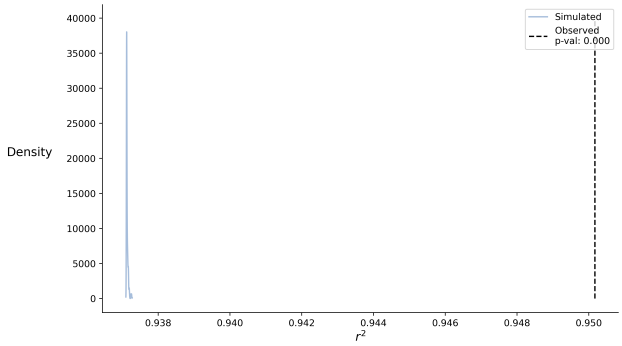

Figure 18: Conditional independence test results for travel time and travel cost given travel distance

5.2 Demonstration

To demonstrate the testing procedures described above, we used the causal graph in Figure 16. 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. Luckily, tools such as Daggity (Textor et al., 2016) can 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 assign low probability to this independence. Generally, we expect the number of automobiles to be positively related to the number of licensed drivers in a household. Secondly, we tested the assumption that travel cost was independent of travel time, conditional on travel distance. Unlike the previous independence assertion, this conditional independence is a-priori more credible. In both cases, we will test our assumptions using the previous subsection's procedures.

In particular, Figure 17 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 the weaker assumption of marginal dependence. This relaxation may contain data-generating assumptions that better reflect our observations.

In Figure 18, 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, these results suggest conditional dependence between our variables of interest. In particular, we should investigate how travel time and travel cost relate, 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, intuitive explanations exist for conditional dependence between travel time and travel cost. 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.

5.2.1 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 expanding our test. Instead of performing one test with a set of observed X, observed Y and observed Z, we 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 this test's p-value by averaging 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, note that our "observed" test statistic is itself a random variable: it depends on the imputed values of Z. Because we now have a distribution of observed test statistics, we 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. Here, we apply the deconfounder algorithm of Wang and Blei (2019) to our data. In particular, Figure 19 shows the deconfounder's assumed causal graph. Note, 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^*.

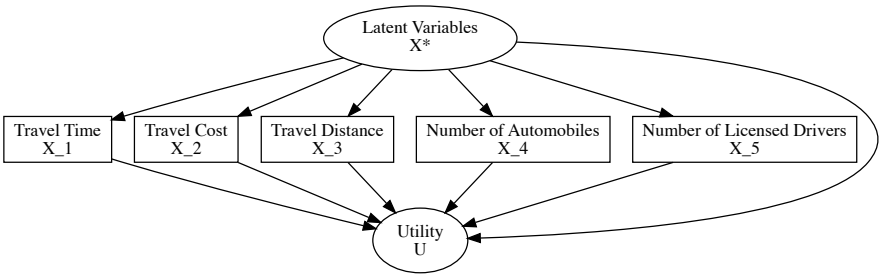

Figure 19: Causal graph from applying the deconfounder algorithm (Wang and Blei, 2019) to our dataset.

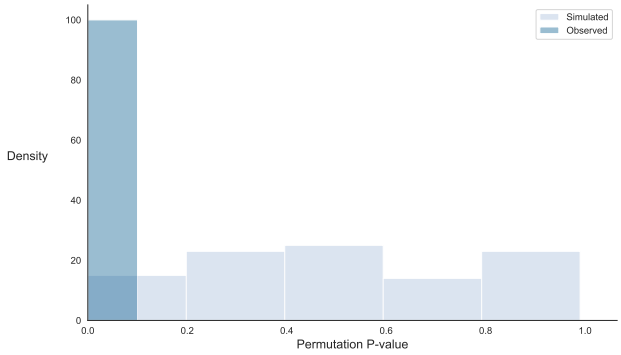

Figure 20: Results of testing that the number of drivers is independent of the number of automobiles in the household, conditional on the latent variable.

Figure 20 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 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[2], 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 simulated data used to make the reference distribution. Unfortunately, we are unsure of how much dissimilarity comes from conditional independence violations. The observed data can differ distributionally from the simulated data in many ways. 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. Specifically, we want our marginal priors $P_{\text{prior}}(X^*) = \int P_{\text{prior}}(X^* \mid Z^*) P(Z^*) \partial Z^*$ to reasonably well represent the observed X (and the same for $P_{\text{prior}}(Y^*)$ and Y). Then, any remaining discrepancies between our observed data and our simulated data can be mainly attributed to their differing conditional independence properties.

5.3 Additional techniques

The methods presented in this section test independence assertions using one's dataset. While perhaps the most accessible strategy for testing one's causal graph, other techniques apply as well. For instance, Pitchforth and Mengersen (2013) proposed a 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, causal graphs encode assumptions about the number of independent variables in one's data. The independent variables are the parentless-nodes in one's graph. Crucially, each graph assumes a particular number of such parentless-nodes. To test this assumption we first estimate our data's "intrinsic dimension." Then, we test whether the intrinsic dimension equals the number of parentless-nodes in our graph. For more information on estimating the intrinsic dimension of a dataset, see Camastra and Staiano (2016); Song et al. (2019). Additionally, see Chenwei et al. (2019) for an extension of this idea when we 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. Then, one can estimate the necessary 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 useful for judging 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 idea is that differing latent variable graphs imply differing entropies in our observed variables. Accordingly, we test for these entropies and constraints. 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)

[2]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`

- 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

Our previous section detailed a method for testing 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 instructions on how to construct an initial causal graph using expert opinion.

Our unstated presumption is that we will test our proposed causal graph. If any of our tests fail, we will then revise the graph until its assumptions appear defensible. Essentially, we postulate, test, and edit our causal graph until the data conform to the graph's assumptions. If this iterative discovery process sounds tedious, that is because it can be!

In this section, we will discuss how we can avoid such repetitive, manual graph editing and testing. Specifically, this section describes causal discovery: algorithmically inferring our causal graph from data. Here, we detail why causal discovery is important; we provide a brief overview of the main concepts in causal discovery; and we show the results of using causal discovery algorithms on the dataset described in Section 2. At the end, we provide references to some further topics at the intersection of causal discovery and experimentation.

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. For example, we can use causal discovery to understand how our prior beliefs affect our inferences. To do so, we could compare our inferred causal effects under expert-opinion versus data-driven causal graphs.

Secondly, causal discovery helps inspire the creation and editing of expert-opinion graphs. If we have not already created our own causal graphs, then we will find that it's typically easier to edit discovered graphs than to create from scratch. Alternatively, graphs found via causal discovery can spark revisions if we have already created our own graphs. This is especially likely when the discovered graphs feature different causal relations than are present in our own graph. 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, our effect estimates are dependent on our causal graphs. To assess dependence strength, we can compare our effect estimates under different causal graphs. For instance, we can use a discovered graph versus our expert-opinion graph. Any graph-induced differences reflect and characterize structural sensitivity in our causal effect inferences.

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 easier than creating them.

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 $X \to Y$ plausible?). Similarly, undirected or bi-directed edges between variables highlight causal ambiguity. Such call-outs invite 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 "$X \to Y$ (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 the causal graphs, and
- for distributional analyses of one's causal effect estimates conditional on each sampled graph. I.e., how certain are we of any one causal graph?

6.2 Overview of causal discovery algorithms

This 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 score-based algorithms. These techniques proceed sequentially, in pairs of variables. For each considered pair, an edge may be added, removed, or reversed in direction. These changes are made greedily, so long as our scoring criterion improves. Most commonly, the scoring criterion is the Bayesian Information Criterion for the joint prediction of all variables in our dataset (Malinsky and Danks, 2018).

To begin using a score-based algorithm, we start with a fully disconnected graph. To this graph, we add directed causal relationships to increase the score of the generative model that corresponds to our graph. Once we cannot improve the score this way anymore, we have found the graph structure that maximizes the score on our data. From this graph, we then prune as many directed causal relationships as possible 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 constraint-based methods. Nonetheless, combinations of score and constraint-based ideas have outperformed methods from either class alone (Glymour et al., 2019).

Lastly, the third class of algorithms estimates Functional Causal Models (FCM) (Goudet et al., 2018) for each of the variables in our system. 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,

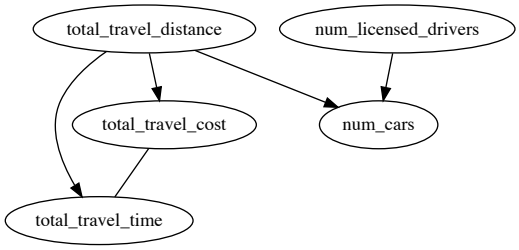

Figure 21: Result of using the PC algorithm on variables in the drive-alone utility

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 a graph's undirected or bi-directed edges. By orienting these edges, FCMs reduce 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 of 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 dealing solely 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 available and increase methodological complexity as needed, this should be an illuminating starting point. As noted in Section 5.2, we suspect that our causal graph for the drive alone utility of our dataset contains unobserved confounding. Given that the PC algorithm assumes that we are free of unobserved confounding, 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?

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, we end up with a fully directed acyclic graph. In other cases, we recover a mix of directed and undirected edges, denoting a MEC of graphs with the same independence properties. In the final case, we recover an undirected causal graph. This corresponds to the situation where we cannot infer which objects cause which other objects. We merely know 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 16. 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 16. 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 a function of two variables: the number of licensed drivers that one has in one's household and 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 discovered graph labels travel time and travel cost as 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 revealed that travel cost is not independent of travel time, conditional on travel distance. Far from being a surprising congruence, we should expect this alignment: conditional and marginal independence testing results are 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 recover that the number of licensed drivers causes the number of automobiles, as opposed to the reverse. 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, discovered graphs with undirected edges express uncertainty about what-causes-what. Disclosing and marginalizing over structural uncertainty of this kind is rare in choice modelling, as far as we know. Lastly, we gain confidence in our results. Through automated tests, we ensure our results are not based on untested or implausible causal assumptions.

6.4 Experimental extensions

As just demonstrated, causal discovery algorithms take in data and output causal graphs. So far, we described causal discovery algorithms in the context of observational data, where there is no random assignment to treatment. However, this setting has its limitations. Indeed, limitations on a variable's observability lead to causal discovery algorithms sometimes returning graphs with undirected edges, representing latent confounding. To overcome issues of confounding, there are causal discovery algorithms that make use of experimental data. In particular, one should think of three kinds of experiments: planned, unplanned, and natural.

Planned experiments are what we typically think of as experiments. These include A/B tests in technology companies, medical trials for drug certification, randomized controlled trials by economists and social scientists, etc. Unplanned experiments are instances of true randomization that were not pre-meditated for experimentation. For instance, technology companies commonly use random allocation (e.g., 'np.random.choice' in Python) throughout their code bases, outside of formal experiments. These are often placeholders or temporary measures, but each such instance is an experiment that allocates some individuals to one treatment and some to another.

Finally, natural experiments are instances where our system is subject to non-random interventions. Sometimes, these natural experiments take the form of a policy that we apply uniformly across our system. For example, imagine a country-wide change in immigration policy that affects the supply of employees into the country, when one is analyzing firm success. In other instances, the interventions affect specific individuals. For instance, rounding prices up or down to the nearest dollar allocates people with a "true price" of $2.51 to a 3.00 treatment, and it allocates people with a "true price" of $2.49 to a 2.00 treatment. This treatment allocation is not random, but perhaps there are ways it is "as good as random" thus permitting us to treat it similarly to a formal experiment.

In making use of planned, unplanned, and natural experiments, there are roughly two types of causal discovery algorithms. One class of causal discovery algorithm takes a causal graph with unresolved ambiguities and outputs an experimentation plan whose resulting data will permit resolution of the causal graph. For algorithms of this kind, see work by Ghassami et al. (2018), the work cited therein, and related literature. The second class of algorithm takes available experimental and observational data and tries to best combine these sources of information to construct a causal graph that is valid for all datasets. For references and guidance in this vein, see work such as Tian and Pearl (2013); Peters et al. (2016); Ghassami et al. (2017);

Zhang et al. (2017); Kallus et al. (2018).

7 Latent Confounding

This section focuses on latent confounding in causal inference problems. As mentioned in Section 4, latent confounders affect the treatment assignment of our causal variables of interest, as well as the outcome. While more complicated, this setup is the more realistic and typical case faced by demand modelers. We first go over a few examples of confounding in transportation analyses, 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). We will show how directed acyclic graphs can help clarify one's reasoning regarding the number of confounders, the assumptions for which variables are confounded, and the models needed to estimate the causal effects. We do that using a simplified simulation scenario to investigate the usefulness and pitfalls of the deconfounder 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 in the effect of adding a bike on the mode share of bicycling in a given neighborhood. To estimate this effect, one strategy that a modeler may use is to develop a disaggregate mode choice model with a dummy variable for whether a bikelane exists between an individual's trip origin and destination, and then treating the coefficient on this variable as the causal effect of adding a bikelane on the log-odds of choosing to bike. The problem with such strategy is that individuals may self-select to live in an area where bicycle infrastructure exists because they have a preference for commuting by bike. In other words, if we define an additional variable to encode a person's latent inherent preference for biking, then this variable determines both a person's likelihood for living in an area with an existing bicycle infrastructure, as well as whether that person choose to bike. This variable is subsequently expected not to lack "balance" across the treatment and control groups; those who live near a bicycle lane are expected to have higher preference for biking than those who don't. Without accounting for this 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 additional surveys with a sample of the people. This may not be an option in many cases; for example, we may not be able to reach the individuals to conduct additional surveys with them.

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 confounding variable by estimating a "substitute 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.
- 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 only have multi-cause confounders, thus the title of the paper, "The blessings of multiple causes". In other words, this method works when all 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 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 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 stronger prior on which variables are confounded. This can be in the form of building and testing candidate causal graphs that try to illustrate this confounding.

We compare both the coefficient estimates on travel time and cost from each of those three models to the true estimates used in the simulation. We also compare 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, we 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. Moreover, the coefficients on the recovered confounder are highly insignificant. This raises questions about the usefulness

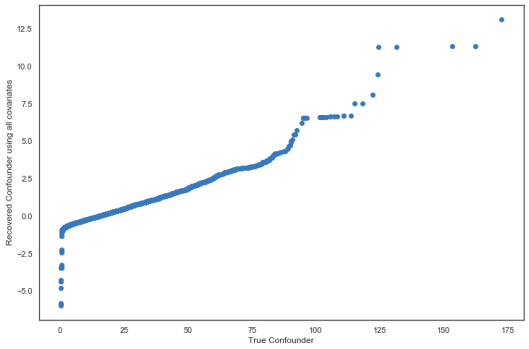

Figure 22: QQ plot of true confounder against recovered confounder using all covariates

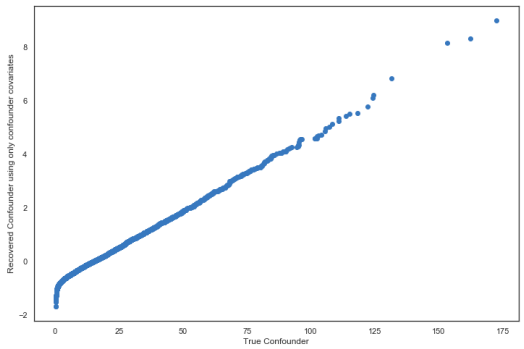

Figure 23: QQ plot of true confounder against recovered confounder using all covariates

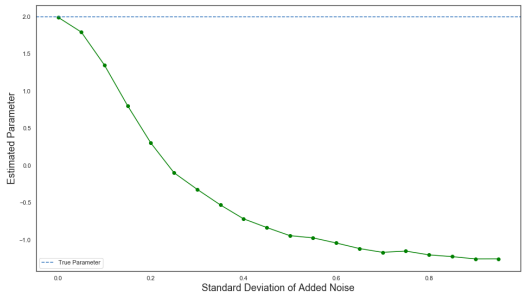

Figure 24: Sensitivity of causal estimates to random errors in the confounder variable

of the deconfounder algorithm in practice. Limitations notwithstanding, it is important to point out that sensitivity to small errors and misfits 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.

8 Discussion

In this chapter, we've focused on why causal graphs are important, how to create them, how to test them, and on how to use them in applied problems with latent confounding. Our latent confounding example showed that statistical inference of our models' parameters may still be a challenge, even with a correct causal graph. Such challenges lead us to the following set of post-graph-construction topics:

- model estimation
- model checking
- experimental design
- experiment analysis
- decision analysis

To us, each item above is important for getting credible results from our analysis and for maximizing our interventions' benefits. Accordingly, even though we will neglect details due to space and time constraints, we briefly discuss these topics below.

Perhaps most obviously, after creating and testing a causal graph, we will use it to estimate the effects of our interventions. To compute our effect estimates, we will need to evaluate statements such as the probability of a particular node taking a given value, conditional on the values of that node's parents. These probabilities will come from our estimated models, so model estimation is critically important to our effect estimation. Thankfully, this is the part of causal inference that choice *modellers* are most familiar with. For instance, Kostic et al. (2020) first use the PC and GES causal discovery algorithms to generate a causal graph, then they test the graph qualitatively, and finally they estimate models corresponding to this causal graph. As another example, Garrido et al. (2020) start from where we end: at a known (or selected) causal graph. They then use neural network density estimators to model the necessary probabilities for estimating one's causal effects of interest.

Next, after estimating the models for our causal graph but before interpreting or using our results, we should check our entire system of models. Here, there are multiple, non-competing ways of performing these

diagnostics. We can check our models separately, jointly, or in subsets. For thoroughness' sake, we can even perform all these checks instead of one kind.

From a disaggregate perspective, we can consider a sequential application of model checking exercises, one per estimated model. Ideally, each model checking process will include the use of visual diagnostics, as (for example) described in Brathwaite (2018). Alternatively, we can check subsets of models together instead of checking one model at a time. For example, Tran et al. (2016) jointly check their models for all variables that their intervention will set (i.e. the treatment assignment variables), and then they separately check their outcome variable models. Finally, we can check all our models jointly by defining global diagnostic measures over all nodes in our causal graph. See Williamson et al. (2013) for a demonstration.

Following model diagnostics, we are ready to use our models and causal graph to inform real interventions. These interventions can come in two kind: an experiment or a "full-scale" implementation of one's policy. If our intervention is experimental, then we are likely interested in one of two aims. We either want to decide between one or more treatment options, or we want to learn about our system, though not necessarily to make a decision. In both cases, however, we pay great attention to the design of our experiments.

When experimenting to make decisions, such as whether to launch a given treatment or not, we pay extra attention to the size of our experiment. Specifically, we want our experiment's sample size to be large enough such that after we update our beliefs using the experimental data, that we have at least our minimum desired probability of making the correct decision. The decision can be to declare the effect of a treatment statistically different from zero, but more frequently, the decision will be more fundamental such as "implement treatment A." For thorough explanations of how to conceptualize and design experiments in a bayesian, model-based setting, see Chaloner and Verdinelli (1995) and Wang et al. (2002). For examples and guidance on how to use one's causal graph structure to guide the general design of one's experiment, beyond sample size, see Madrigal et al. (2007). There, the structure of one's causal graph is used to inform general design decisions such as the clustered allocation of individuals to treatment, and the experimental design is itself analyzed graphically. Additionally, note the relations to reinforcement learning where an agent has to perform experiments in order to discover the action/intervention that will maximize her expected, counterfactual reward. In this context, Lee and Bareinboim (2018) have shown that designing our experiments without guidance from one's causal model is generally suboptimal, and that we can achieve optimality by leveraging our causal graph to design our experimentation plan.

Now, let's transition from experimentation for decision making to consider experimentation for learning. Imagine that we are at a transportation network company and that we are running a pricing experiment to learn about price elasticities of our customers. Here, there is no immediate decision being made, but we learn about an edge in our causal graph: the edge from price of a trip (treatment) to purchase of the ride (outcome). In other cases, we may experiment to learn not just about the strength of an edge, but about the presence of edges and the structure of the graph more generally. For example, we may wish to remove residual ambiguity from a causal discovery process that outputs a Markov Equivalence Class of graphs instead of a single causal DAG. In these situations, we are interested in optimally designing an experiment (or series of experiments) to learn a causal graph (or its properties). We are further interested in how we can leverage potentially multiple experimental datasets to improve our causal graphs. For a review of the literature on experimentation for learning and construction/refinement of a preliminary causal graph, see Kalisch and Bühlmann (2014, Sec. 3.1.2). For more recent approaches in this vein, see works such as Triantafillou and Tsamardinos (2015); Kocaoglu et al. (2017); Brouillard et al. (2020); Rantanen et al. (2020).

Finally, after we have run any experiments that we are interested in, we still must decide how to intervene in our population. Three major types of questions come to mind immediately:

1. Should we launch the treatment(s) at all?
2. Should we launch the treatments to everyone?
3. How should the treatments be dispersed/implemented?

After updating our posterior beliefs with the experimental data, we'll want to analyze and come to a conclusion about launching our treatment(s). In particular, we will wish to determine the expected distribution of impacts under each alternative decision. Here, Manski (2019) should provide the basic idioms of thought and pointers to the larger literature on treatment choice from a decision theoretic perspective.

Next, if we've decided to launch the treatments, we come to the question of who should receive the treatment(s)? Everyone? A select few? Are there certain subgroups that should receive the treatment but not others? These questions fundamentally revolve around the level and nature of heterogeneity in treatment effects. We will, with good reason, want to search for evidence of heterogeneity and characterize it if found. In doing so, we should consult articles such as Pearl (2017) and Webster-Clark et al. (2020) for guidance on how to perform one's subgroup analysis in light of one's causal graph. This should help us avoid drawing incorrect conclusions or misinterpreting our analyses.

Thirdly, we will need to answer logistical questions about the levels and the frequency of treatment. With regard to choosing the levels of (possibly continuous and multiple) treatments, recent work on causal bayesian optimization represents the state of the art in this area (Aglietti et al., 2020). Moreover, the entire field of reinforcement learning focuses on running experiments and learning from past observations to determine the treatment arms/levels that will maximize one's reward (however we define it). Accordingly, we stand to gain much by consulting the work on and principles from causal reinforcement learning (c.f. Bareinboim et al. (2015)) when choosing our optimal treatment plan.

9 Conclusion

Travel demand problems aim at forecasting the impact of proposed project and policies. These problems are causal in nature. To our surprise, there has been very little mention of causal inference in the field of travel demand modeling. Brathwaite and Walker (2018a) have documented this disconnect and presented an initial framework for addressing it.

In this chapter, we build up on Brathwaite and Walker (2018a) and highlight the importance of using causal graphs and causal inference methods in transportation demand modeling efforts.

We presented a numerical exercise 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 outcome model, that different assumptions about the data generating processes could lead to bias in estimates of the effect of interest.

Data generating processes can be illustrated effectively using DAGs as shown by Pearl (1995). We have shown that DAGs help researchers and practitioners represent their assumptions about the data generating process.

We presented a process allowing researchers and practitioners to construct causal graphs based on their expert opinion of the problem at hand. We also presented methods for testing the implications encoded in the the practitioners proposed causal graph. These implications might be either observed, latent, or both.

Beyond testing the implications and assumptions encoded in a one's causal graph, we describe why causal discovery is important in discerning relationships between covariates one's data. We shortly presented several causal discovery algorithms used by other researchers in different fields. We highlighted 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, develop 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 presented examples of latent confounding within the transportation demand modeling field and highlight some examples of current methods aiming at addressing problems resulting from confounding in certain situations.

We present a recently developed algorithm by Wang and Blei (2019) aiming to address the problems not addressed by previously developed methods dealing with confounding. 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. More specifically, we show that the deconfounder algorithm shows to be sensitive to small errors in one's data.

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326v3*, 2019.

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.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

Marc F Bellemare and Jeffrey R Bloem. The paper of how: Estimating treatment effects using the front-door criterion. Technical report, Working Paper, 2019.

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. Check yourself before you wreck yourself: Assessing discrete choice models through predictive simulations. *arXiv preprint arXiv:1806.02307*, 2018.

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

Timothy Brathwaite and Joan L. Walker. Asymmetric, closed-form, finite-parameter models of multinomial choice. *Journal of Choice Modelling*, 29:78–112, 2018b.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33, 2020.

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-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

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.

Sergio Garrido, Stanislav S Borysov, Jeppe Rich, and Francisco C Pereira. Estimating causal effects with the neural autoregressive density estimator. *arXiv preprint arXiv:2008.07283*, 2020.

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

Andrew Gelman et al. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7:2595–2602, 2013.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733. PMLR, 2018.

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.

Adam N Glynn and Konstantin Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, 113(523):1040–1049, 2018.

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.

Shantanu Gupta, Zachary C Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. *arXiv preprint arXiv:2003.11991*, 2020.

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

Markus Kalisch and Peter Bühlmann. Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 11(1):3–21, 2014.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in neural information processing systems*, pages 10888–10897, 2018.

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.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7018–7028, 2017.

Bojan Kostic, Stephane Hess, Joachim Scheiner, Christian Holz-Rau, Francisco C Pereira, et al. Uncovering life-course patterns with causal discovery and survival analysis. *arXiv preprint arXiv:2001.11399*, 2020.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018.

Ana Maria Madrigal et al. Cluster allocation design networks. *Bayesian Analysis*, 2(3):557–589, 2007.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

Charles F Manski. Treatment choice with trial data: statistical decision theory should supplant hypothesis testing. *The American Statistician*, 73(sup1):296–304, 2019.

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.

Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 5(78):947–1012, 2016.

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

Kari Rantanen, Antti Hyttinen, and Matti Järvisalo. Learning optimal cyclic causal graphs from interventional data. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models (PGM 2020)*. Journal of Machine Learning Research, 2020.

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.

Jin Tian and Judea Pearl. Causal discovery from changes. *arXiv preprint arXiv:1301.2312*, 2013.

Dustin Tran, Francisco JR Ruiz, Susan Athey, and David M Blei. Model criticism for bayesian causal inference. *arXiv preprint arXiv:1610.09037*, 2016.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

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

Fei Wang, Alan E Gelfand, et al. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2):193–208, 2002.

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

Michael Webster-Clark, John A Baron, Michele Jonsson Funk, and Daniel Westreich. How subgroup analyses can miss the trees for the forest plots: A simulation study. *Journal of clinical epidemiology*, 126:65–70, 2020.

David M Williamson, Russell Almond, and Robert Mislevy. Model criticism of bayesian networks with latent variables. *arXiv preprint arXiv:1301.3902*, 2013.

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

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020.

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.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from non-stationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access, 2017.

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.