Causal Inference with Outcome-Dependent Missingness And Self-Censoring

by

Jacob M. Chen

Professor Rohit Bhattacharya, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Computer Science

Williams College
Williamstown, Massachusetts

May 10, 2023

Contents

List of Figures

List of Tables

Abstract

Missing data is often unavoidable in fields that rely on observational data such as epidemiology, economics, political science, and certain branches of social science. We consider missingness in the context of causal inference when the outcome of interest may be missing. If the outcome directly affects its own missingness status, i.e., it is "self-censoring", this may lead to severely biased causal effect estimates. Miao et al. (2015) proposed the shadow variable method to correct for bias due to self-censoring, however, verifying the required model assumptions can be difficult. Here, we propose a test based on a randomized incentive variable offered to encourage reporting of the outcome that can be used to verify identification assumptions that are sufficient to correct for both self-censoring and confounding bias. Concretely, the test confirms whether a given set of pre-treatment covariates are sufficient to block all backdoor paths between the treatment and outcome as well as all paths between the treatment and missingness indicator after conditioning on the outcome. We show that under these conditions, it is possible to obtain unbiased estimates of the causal effect by using the treatment as a shadow variable. We show that this leads to an intuitive inverse probability weighting estimating equation that uses a product of the treatment and response weights. We evaluate the efficacy of our test and downstream estimator via simulations.

Acknowledgments

I would not be where I am today if it were not for the multitudes of people working from the shadows to support me throughout the years I have been at Williams and far before. First and foremost, I would like to thank Rohit, my thesis advisor. Rohit has supported me since a year before this thesis began as I explored causal inference and missing data as a research area before I even realized how deeply engaging and meaningful I would come to find it. Throughout the two years I have known Rohit, he has become not just a mentor, but also a friend with whom I have shared many laughs over humorous complaints of silly paper reviews, getting ghosted by famous researchers, and potential oracles whispering into the ears of researchers who can magically come up with estimating equations. With the help and support of Rohit, I attended my first computer science conference and submitted my first academic paper. Rohit has been nothing short of instrumental in jump starting my journey into research and giving me the confidence that I am capable. I hope to be an educator who is as compassionate, humble, and intelligent as Rohit one day. Here's to many more years of collaboration, laughter, and sharing durians together.

Dan Malinsky has also made a significant contribution to our work here. He has been a springboard for new ideas and sparked discussion when Rohit and I reach obstacles in our thinking. As a biostatistician, Dan helped us ground our methods in practicality. Dan also showed us that our initial proof was wrong, but that just pushed me and Rohit to think about the problem in new and creative ways! Thank you, Dan!

I would also like to extend my deep gratitude towards Mark, my second reader, who provided me with insightful comments and made sure that this thesis, especially the first couple of chapters, is as precise and accessible as possible. I am also extremely grateful for the entire computer science department staff and faculty. The staff and faculty in computer science at Williams have so much rapport and camaraderie, and it definitely rubs off on us and makes us even more excited to learn. Thank you for being an inspiration to us all.

I would also like to thank my brother and his parents who have supported me in the pursuit of a thesis and given me numerous encouragements[1]. There is an immeasurable amount of friends at Williams and from home who have been rooting for me from the very beginning, but it would be impossible to enumerate all of you here[2]. I would not be here today without all of you; you all know who you are. The data may be missing, but you all are fully observed in my heart and mind.

[1] Yes, I know that my brother's parents are my parents, too. Do it, I won't? What's it to ya??

[2] In fact, the process of listing all of you would be NP-hard and definitely not win any awards for efficiency.

Chapter 1

Motivating Missing Data and Confounding Bias

Missing data and confounding bias are often unavoidable in observational studies. For example, when researchers are interested in investigating sensitive topics such as drug use or risky sexual behavior, participants in the study may opt out of answering some of these questions, resulting in missing data. In addition, there exists the possibility that confounders – variables that affect both the treatment an individual is assigned and the outcome of interest for that individual – can obscure estimations of the causal effect. In this thesis, we discuss the most difficult kind of missing data bias, which arises from *self-censoring*. Self-censoring is a phenomenon that occurs when the value of a variable directly affects whether or not that value is reported. Take, for example, asking respondents in a survey about whether or not they smoke. Due to the social stigma associated with smoking, those who do smoke are less likely to respond to the survey; this is one classic example of self-censoring missing data.

In order to understand why self-censoring is the most difficult form of missingness to correct for, we describe here previous work in missing data theory that has established a hierarchy of missingness mechanisms (Mohan and Pearl, 2021). The most simple form of missingness is known as *missing completely at random* (MCAR), which occurs when data are missing in a manner independent of any other variables in the dataset. Next, *missing at random* (MAR) occurs when data are missing dependent only on other variables in the dataset that are fully observed, i.e. variables that do not have missing values. When data are missing dependent on variables that are partially observed, i.e. variables that have missing values, this is known as *missing not at random* (MNAR) data.

In general, a precondition to unbiased estimates when working with missing data is *identification*. Identification describes the precise conditions under which a target parameter defined in terms of potential outcomes can be recovered in an unbiased manner from the observed data (Rubin, 1976). Under MCAR and MAR missingness, identification is always possible (Rubin, 1976). However, identification is much more difficult under MNAR settings because we do not have full access to the variables that are causing the missingness. Because of this, MNAR missingness is regarded as one of the most difficult type of missingness to recover from. Despite these challenges, progress has been made by representing the target distribution and the missingness mechanism as a causal graph, which has led to identification for particular kinds of MNAR models (Daniel et al., 2012). For example, Bhattacharya et al. (2019); Nabi et al. (2020) have shown that identification is possible for MNAR data as long as self-censoring and certain graphical structures are absent from the graph. In addition, many works in missing data specifically exclude the possibility of self-censoring in their methods (Tu et al., 2019; Malinsky et al., 2021). Hence, self-censoring – a unique but common case of MNAR missingness – is widely regarded as the most difficult form of missingness to correct for.

In this thesis, we discuss methods for dealing with self-censoring in conjunction with confounding bias. By extending previous work in controlling for confounding bias and overcoming self-censoring, we present identification results for the causal effect under a self-censoring outcome of interest (Entner et al., 2013; Miao et al., 2015). To verify whether or not identifying assumptions hold for our estimation strategy, we propose empirically testable conditions under which our identifying assumptions hold. In the rest of this chapter, we present a hypothetical study that suffers from missing data and confounding bias to demonstrate the issues that arise when we ignore either one of these biases when estimating statistical or causal quantities.

1.1 Motivating Example

Imagine a group of researchers who are interested in evaluating the effect of a sexual education course on condom use. The researchers were not able to randomly assign individuals to take the sexual education course; hence, there exists the possibility of confounders – variables that affect both enrollment in the course and condom use habits – that bias the data if they are not controlled for. Furthermore, there is no way for the researchers to measure condom use other than by asking respondents directly. Due to the sensitive nature of the question of whether respondents use condoms, the researchers have reason to believe that individuals who use condoms less frequently are more likely to opt-out of reporting their condom use status. Hence, we have a missing data problem in this study as well.

In an effort to increase the response rate of whether or not individuals use condoms, the researchers randomly assign certain individuals to receive interviews from the Telephone Audio Computer-Assisted Self-Interviewing (TACASI) program (Turner et al., 2009). Instead of having a human interviewer call respondents, a computer program with a recording of a human voice is used instead. Turner et al. (2009) have shown that the TACASI program increases the response rate when querying sensitive topics such as condom use. Such a method is known as using an *incentive for response*. Some other common incentives for response may include gift cards. While increasing the response rate increases the proportion of people who respond to the survey, this does not completely eliminate the missing data problem.

Let us formalize our observational study and our missing data problem. Let A, the treatment, represent whether or not an individual enrolls in the sexual education course. Let Y, the outcome, represent *observed* values of whether an individual uses condoms *frequently* or *infrequently*; hence, Y is a binary variable. Let R_Y represent whether or not an individual responds to the survey with a value of 1 representing that the individual did respond to the survey and a value of 0 representing that the individual did not respond to the survey. The variable R_Y is known as the *missingness indicator*. Finally, let $Y^{(R_Y=1)}$ represent the *potential outcome* of condom use as if the researchers had hypothetically been able to observe the condom use status for each individual in the dataset. For brevity, we use the notation $Y^{(1)}$ from here on instead of $Y^{(R_Y=1)}$ to represent the potential outcome of Y if we had hypothetically been able to observe it.[1]

The variable Y is determined by $Y^{(1)}$ and R_Y according to the following equations:

$$Y = \begin{cases} Y^{(1)} & \text{if } R_Y = 1 \\ ? & \text{if } R_Y = 0 \end{cases}$$

Figure 1.1 shows a simple graph depicting the missing data problem described above. The true condom use status of each individual, represented by $Y^{(1)}$, affects whether or not an individual reports their condom use status, represented by R_Y. The variable Y, which represents values of condom use status that researchers actually get to observe, is determined by both $Y^{(1)}$ and R_Y. In this thesis, we make the crucial assumption that an individual reports their condom use status truthfully whenever they do decide to report it. Due to the deterministic nature of the relationship between the variable Y and the variables $Y^{(1)}$ and R_Y, we will omit Y from graphs moving forward.

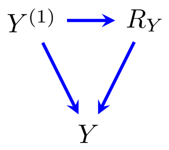


Figure 1.1: Graph depicting the missing data problem.

Note that, in this motivating example, we have described exactly a situation where the outcome exhibits self-censoring. In this thesis, we will be focusing specifically on situations where the outcome is self-censoring, as often happens in real-world scenarios like our motivating example.

First, let us attempt to do something simple with our dataset: estimating the mean of frequent condom use. Table 1.1 shows a hypothetical dataset with 10 individuals. Only 6 out of 10 individuals report their true condom use status, and so the rows of data where $R_Y = 0$ shows a question mark for the value of Y. If we calculate the mean of condom use by ignoring missing data, we will get that $\frac{5}{6} \approx 0.83$ of the individuals in the dataset use condoms frequently. However, if we calculate the mean of condom use status using the column $Y^{(1)}$ where we contrary to fact observe the condom use status of every individual in the dataset, we will see that only $\frac{6}{10} = 0.6$ of the individuals in the dataset use condoms frequently. As the reader will notice, ignoring missing data will result in biased estimates for the average. Hence, attempting to estimate anything more complicated, such as causal effects, will also result in bias because any such estimations require calculating the mean as a subroutine.

[1] Similar notation for representing potential outcomes under missingness has been used in Nabi et al. (2022).

	$Y^{(1)}$	Y	R_Y
1	0	?	0
2	1	1	1
3	1	1	1
4	0	?	0
5	0	?	0
6	1	?	0
7	1	1	1
8	0	0	1
9	1	1	1
10	1	1	1

Table 1.1: Table representing a hypothetical dataset of 10 individuals with the potential outcome $Y^{(1)}$ shown.

In this example, we have access to values of $Y^{(1)}$ as if an oracle was able to give that information to us in secret. However, in real-world datasets, we only have access to the columns of data Y and R_Y. In other words, there is no "ground truth" average that we can compare our estimate to in a real-world scenario. This makes working with missing data even more challenging. By giving a concrete example using a table and comparing observed and potential outcomes, we have shown that missing data is non-ignorable if we would like to infer unbiased estimates.

Next, let us consider how we might recover the *average causal effect* (ACE) of the sexual education course on condom use as if there were no missing data. Because there is no missing data, we use Y instead of $Y^{(1)}$. In fact, in the case where is no missing data, $Y = Y^{(1)}$ for each individual in the dataset. Here, we have an observational study, which means that there are possibly confounders in the dataset that cause both whether or not an individual enrolls in the sexual education course and whether or not they use condoms frequently. For example, let Z_1 represent whether or not an individual has adequate access to public transportation. If an individual has access to public transportation, then they are more likely to be able to attend the sexual education course and access places like pharmacies where they may purchase condoms. On the other hand, if an individual does not have access to public transportation, then they are less likely to both enroll in the sexual education course and have fair access to condoms. The reader may imagine other possible confounders, and let us call them Z_2 and Z_3. Each confounder can have varying causal relationships with A and Y. In this thesis, we allow for some of these confounders to be completely unobserved, which further complicates the process of controlling for confounding bias.

Figure 1.2 gives a graphical representation of the confounding bias problem. Intuitively, non-causal paths – paths that are not directed paths from A to Y – between the treatment and outcome may "leak" information that will end up biasing the estimate for the causal effect. In order to obtain an unbiased measure for the average causal effect, then, we target potential outcomes, just like when we are adjusting for missing data.

Rather than the potential outcome of Y if we had hypothetically been able to observe its value,

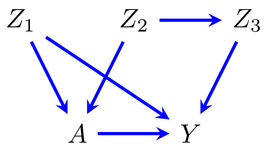


Figure 1.2: Graph depicting the confounding bias problem.

though, here we attempt to recover the potential outcome of Y if we had hypothetically been able to set the value of the treatment of a particular individual to 0 or 1. In notation, $Y^{(A=0)}$ represents the potential outcome of setting the treatment to 0. Here, we use the uppercase letter to represent a hypothetical intervention. Similarly, $Y^{(A=1)}$ represents the potential outcome of setting the treatment to 1. If the treatment for a particular individual had in reality been 0, then the observed value of Y is the same $Y^{(A=0)}$ for that individual. However, we do not know what the value of $Y^{(A=1)}$ is for that individual because we cannot go back in time and reassign the treatment. This relationship between observed Y and the potential outcomes $Y^{(A=0)}$ and $Y^{(A=1)}$ is known as *consistency* and is summarized in the following equations (Jerzy, 1923):

$$Y = Y^{(A=0)} \text{ if } A = 0$$
$$Y = Y^{(A=1)} \text{ if } A = 1$$

The potential outcomes $Y^{(A=0)}$ and $Y^{(A=1)}$ exist for each individual in the dataset, but, for any given individual, one of the potential outcomes is missing. Table 1.2 shows a hypothetical dataset with 10 individuals with the potential outcomes $Y^{(A=0)}$ and $Y^{(A=1)}$ augmented onto it with question marks to represent the missing values for the potential outcomes. When adjusting for confounding bias, the target parameter of interest is the average causal effect, which is formally defined as

$$\text{ACE} = \mathbb{E}[Y^{(A=1)}] - \mathbb{E}[Y^{(A=0)}] \tag{1.1}$$

In words, equation 1.1 states that the ACE is equal to the average of the outcome if we had hypothetically assigned a treatment value of 1 to each individual in the dataset subtracted by the average of the outcome if we had hypothetically assigned a treatment value of 0 to each individual in the dataset.

Just like in missing data, we do not have access to the values of the potential outcomes when we are adjusting for confounding bias. Furthermore, in most cases, $\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] \neq \mathbb{E}[Y^{(A=1)}] - \mathbb{E}[Y^{(A=0)}]$, where $\mathbb{E}[Y \mid A = 1]$ is the average of the observed values of Y for individuals where $A = 1$ and $\mathbb{E}[Y \mid A = 0]$ is the average of the observed values of Y for individuals where $A = 0$. Hence, an analysis using simple averages will yield biased estimates, and, like missing data, confounding bias is non-ignorable.

	A	Y	$Y^{(A=0)}$	$Y^{(A=1)}$
1	1	0	?	0
2	1	1	?	1
3	0	1	1	?
4	0	0	0	?
5	1	0	?	0
6	1	1	?	0
7	0	1	1	?
8	0	0	0	?
9	0	1	1	?
10	0	1	1	?

Table 1.2: Table representing a hypothetical dataset of 10 individuals with the potential outcomes $Y^{(A=0)}$ and $Y^{(A=1)}$ shown.

1.2 Thesis Overview

Missing data and confounding bias separately complicate the process of making unbiased estimates for means and causal effects. When both of these challenges are present simultaneously, recovering the causal effect may seem insurmountable. Despite the difficulty of this recovery task, however, the simultaneous presence of both issues is common in observational studies, especially in those attempting to measure outcomes related to socially sensitive topics.

In this thesis, we show that it is possible to overcome self-censoring missing data and confounding bias by employing a simple series of statistical tests that verify a set of identifying conditions for the ACE under a separate set of standard assumptions. If the statistical tests are satisfied, then we show how we may use an inverse probability weighted identification strategy to recover the average causal effect. The rest of this thesis is structured as follows. Chapter 2 will formally introduce graphical models for interpreting missing data and causal inference. Next, Chapter 4 will define *shadow variables*, auxiliary variables that allow us to recover from self-censoring under a set of conditions. Chapter 5 will describe the statistical tests that allow us to verify whether or not the necessary conditions for our identification strategy are fulfilled. After describing the theory, Chapter 6 will describe the practical procedure for executing the statistical tests and recovering the ACE. Chapter 7 will present results from our simulation study to demonstrate the accuracy and validity of our method. Finally, Chapter 8 will conclude the thesis and describe possible areas for future work.

Chapter 2

Graphical Models for Missing Data and Causal Inference

In this chapter, we formalize the concepts of missing data and confounding bias that we introduced in Chapter 1. Specifically, we discuss these topics in the context of directed acyclic graphs (DAGs) that visually display central assumptions about the missing data and confounding bias problems. We also give specific identification conditions for when it is possible to make unbiased estimates for target parameters of interest in the presence of missing data and confounding bias as well as equations for making those estimates.

This chapter lays the groundwork for more complex discussions about overcoming missing data by using *shadow variables* in Chapter 4 and our identification strategy for outcome-dependent self-censoring in Chapter 5. We start by formally introducing directed acyclic graphs. Then, we discuss the setup for the missing data problem, the missing data hierarchy, and how to recover from missing data. Finally, we show how confounding bias is related to missing data and how to make unbiased estimates in the presence of confounders.

2.1 Directed Acyclic Graphs and their Statistical Models

Directed acyclic graphs are defined by a set of vertices $\mathbf{V}$ and a set of edges $\mathbf{E}$. An edge E connects two vertices unidirectionally; that is, an edge from a vertex A to a vertex Y is not equivalent to an edge from Y to A. A *directed path* in a DAG is defined by a sequence of k unique vertices $V_1, V_2, \ldots, V_k$ where, for any two vertices V_i and V_{i+1}, there exists a directed edge from V_i to V_{i+1}. A *path*, on the other hand, is defined by a sequence of k unique vertices $V_1, V_2, \ldots, V_k$ where, for any two vertices V_i and V_{i+1}, there exists a directed edge from V_i to V_{i+1} or a directed edge from V_{i+1} to V_i. A *directed cycle* is a sequence of directed edges $V_1 \rightarrow V_2 \rightarrow \ldots \rightarrow V_k$ such that $V_1 = V_k$. By definition, a DAG is not allowed to have directed cycles. Additionally, we define the *parents* of a vertex V in a DAG $\mathcal{G}$ – denoted by $\text{pa}_{\mathcal{G}}(V)$ – as the set of vertices that have a directed edge to V. Similarly, the *children* of a vertex V is the set of vertices that V has a direct edge to. Figure 2.1

shows a simple DAG.

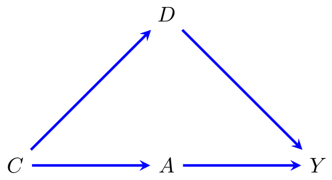


Figure 2.1: A simple DAG with vertex set $\mathbf{V} = \{A, C, D, Y\}$ and edge set $\mathbf{E} = \{C \rightarrow A, C \rightarrow D, A \rightarrow Y, D \rightarrow Y\}$.

Statistical models are interpretations of DAGs that are defined by a set of variables $\mathbf{V}$ and a *probability distribution* $p(\mathbf{V})$ that describes conditional independence relations between the variables in $\mathbf{V}$. *Conditional probability distributions* are denoted as $p(\mathbf{A} \mid \mathbf{B})$, where we interpret this notation as the probability distribution of the variables in the set $\mathbf{A}$ after observing the values of the variables in the set $\mathbf{B}$. If $\mathbf{B}$ is the empty set $\emptyset$, then we have a *marginal probability distribution* over the set of variables in $\mathbf{A}$. A joint distribution from a statistical model represented by $p(V_1, V_2, \ldots, V_n)$ can be decomposed into a product of conditional distributions via the *chain rule of probability* as follows:

$$p(V_1, V_2, \ldots, V_n) = p(V_1) \times p(V_2 \mid V_1) \times p(V_3 \mid V_1, V_2) \times \cdots \times p(V_n \mid V_1, \ldots, V_{n-1}).$$

Given a DAG $\mathcal{G}$, the *statistical model* for $\mathcal{G}$ is defined as follows (Pearl, 1988):

$$p(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)). \tag{2.1}$$

We refer to equation 2.1 as the *factorization* of a DAG. DAG factorizations are important because they tell us how observed data will behave if it is generated according to a certain DAG. They also allow us to infer conditional independencies between variables in the underlying dataset. To demonstrate this, let us consider observed data generated according to the DAG in Figure 2.1. The factorization of this DAG according to equation 2.1 is

$$p(C) \times p(D \mid C) \times p(A \mid C) \times p(Y \mid A, D).$$

Meanwhile, the chain rule factorization of the joint distribution $p(C, D, A, Y)$ is

$$p(C) \times p(D \mid C) \times p(A \mid C, D) \times p(Y \mid A, D, C).$$

From the two expressions above, we deduce that $p(Y \mid A, D) = p(Y \mid A, D, C)$ and that $p(A \mid C) = p(A \mid C, D)$. Because it does not matter whether C is in the conditioning set for the probability distribution of Y, Y is independent of C given its parents A and D. For the same reason, A is independent of D given its parent C. When two variables X and Y are independent given a set of variables $\mathbf{Z}$, this means that the variable Y does not provide any predictive information about X if we have already observed the variables in the set $\mathbf{Z}$ and vice versa.

The list of independencies generated by the statistical interpretation of the DAG, such as the one just given, is known as the *local Markov property.* More generally, the local Markov property states that each V_i in the DAG is independent of all the variables before it in any topological order given its parents.

2.2 Causal Models of a DAG

Aside from a statistical interpretation, DAGs also have a causal interpretation defined by two properties. The first property states that the presence of an edge X to Y denotes that X is a potential direct cause of Y relative to other variables in the graph. The second property is that any variable that is a common cause of two or more observed variables in the DAG is present in the graph (Spirtes et al., 2000).

We may use the *g-formula*, as shown in equation 2.2, to find the effect of *intervening* on a set of variables in the DAG, which is represented by a truncated factorization of the original factorization of the DAG according to equation 2.1. An intervention on a set of variables corresponds to inducing a data distribution representing data where we "intervene" and set the value of a set of variables to some arbitrary value. To generalize the intervention operation, we often use as the superscript a lowercase letter to represent intervention on that variable. For example, $\mathbf{Y}^{(\mathbf{a})}$ represents the potential outcome of $\mathbf{Y}$ had we intervened on $\mathbf{A}$ and set it to values of $\mathbf{a}$, possibly contrary to fact. In addition, the vertical line on the right-hand side of the expression in equation 2.2 denotes that we are evaluating the expression only at values of the subscript. Note that when you have $\mathbf{A} = \emptyset$ in equation 2.2, that represents not intervening on any variables, and you recover the factorization of a DAG according to its statistical interpretation.

$$p((\mathbf{V} \setminus \mathbf{A})^{(\mathbf{a})}) = \left. \frac{\prod_{V_i \in \mathbf{V}} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i))}{\prod_{A_i \in \mathbf{A}} p(A_i \mid \mathrm{pa}_{\mathcal{G}}(A_i))} \right|_{\mathbf{A}=\mathbf{a}} \quad \textit{(g-formula)} \tag{2.2}$$

To demonstrate the g-formula, let us say that we would like to intervene on the variable A at values of $A = a$ in Figure 2.1. To do so, we apply the g-formula as follows:

$$\begin{aligned} p((C, D, Y)^{(a)}) &= \left. \frac{p(C) \times p(D \mid C) \times p(A \mid C) \times p(Y \mid A, D)}{p(A \mid C)} \right|_{A=a} \\ &= p(C) \times p(D \mid C) \times p(Y \mid A, D)|_{A=a} \\ &= p(C) \times p(D \mid C) \times p(Y \mid A = a, D) \end{aligned}$$

The DAG that represents the post-intervention distribution above is shown in Figure 2.2. Because we have intervened on A and set its value to some arbitrary value, it no longer has any parents. Notice that the conditional distributions of C and D are not modified when intervening on A, that is $C^{(a)} = C$ and $D^{(a)} = D$. However, Y is a descendant of A, so the conditional distribution of Y is indeed modified when intervening on A. Therefore, we have a distinct node depicting the potential outcome $Y^{(a)}$ in Figure 2.2.

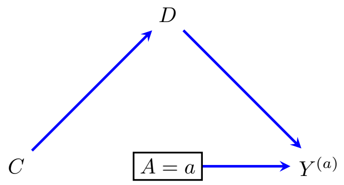


Figure 2.2: A DAG where we intervene on A and set it to the value a.

2.3 D-Separation

In practice, it is difficult to use the local Markov property to find conditional independencies for arbitrary pairs of variables in a DAG. Therefore, we use a concept called *d-separation* to help us determine conditional independencies between variables in the data distribution $p(\mathbf{V})$ – where $\mathbf{V}$ is the set of all variables in the dataset – of a DAG (Pearl, 1988).

The global Markov property states that when X and Y are d-separated given a set of variables $\mathbf{Z}$, this implies that X and Y are conditionally independent in $p(\mathbf{V})$ given $\mathbf{Z}$. Intuitively, Y being d-separated from X given $\mathbf{Z}$ means that Y gives us no useful information on X when we already have information on variables in $\mathbf{Z}$ and vice versa. The reader will notice that this intuitive interpretation of d-separation is identical to the meaning of independence in statistical models of DAGs. This is not a coincidence; the exact purpose of d-separation is to help us reason about independence relations between variables directly from a DAG using its graphical properties. Now, we will discuss the exact graphical properties that define d-separation.

D-separation is defined in terms of chains, forks, and colliders, each of which are shown below in order. Chains and forks are said to be *blocked* if we condition on Z and *unblocked* if we do not condition on Z. Colliders, on the other hand, are blocked when we do not condition on Z but unblocked when we do condition on Z. A path between two variables X and Y given a conditioning set of variables $\mathbf{Z}$ is considered to be blocked if it contains at least one blocked chain, fork, or collider when conditioning on the variables in $\mathbf{Z}$. X and Y are said to be *d-separated* given a set of variables $\mathbf{Z}$ if all paths between X and Y are blocked when conditioning on $\mathbf{Z}$. If X and Y are d-separated when conditioning on $\mathbf{Z}$, we denote this in mathematical notation with $X \perp\!\!\!\perp_d Y \mid \mathbf{Z}$. In this thesis, an alternative way for denoting conditioning on a set of variables $\mathbf{Z}$ is by saying controlling for a set of variables $\mathbf{Z}$.

1. $X \rightarrow Z \rightarrow Y$ *(chain)*
2. $X \leftarrow Z \rightarrow Y$ *(fork)*
3. $X \rightarrow Z \leftarrow Y$ *(collider)*

Now, we use Figure 2.1 as an example to demonstrate d-separation. First, let us consider whether or not the variables C and Y are d-separated. Recall that we must block all paths between C and Y in order for the two variables to be considered d-separated. There are two paths between C and Y: $C \rightarrow D \rightarrow Y$ and $C \rightarrow A \rightarrow Y$. Both of these paths are chains, so we may block them both by controlling for D and A, respectively. Hence, C and Y are d-separated when controlling for D and

A. In notation, $C \perp\!\!\!\perp_d Y \mid D, A$. Next, let us consider the variables D and A. Again, there are two paths between D and A: $D \leftarrow C \rightarrow A$ and $D \rightarrow Y \leftarrow A$. The first path is a fork, so we may control for C to block the path. The second path, however, is a collider, so the path is already blocked. Therefore, we control for only C to d-separate D and A. In notation, $D \perp\!\!\!\perp_d A \mid C$.

When there are colliders in a DAG, we must also take into account whether or not we are controlling for the *descendant* of a collider. A descendant V_d of a variable V_a is a variable such that there is a directed path from V_a to V_d. The descendant of a collider is a variable that is the descendant of the variable in the middle of a collider. Figure 2.3 shows a collider with a variable D that is a descendant of the collider. According to d-separation, controlling for the descendant of a collider D unblocks the collider between $A \rightarrow B \leftarrow C$. That is $A \not\perp\!\!\!\perp_d C \mid D$ in Figure 2.3. If we do not control for the descendant of the collider, though, the collider will remain blocked.

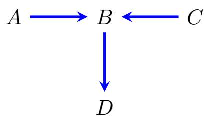


Figure 2.3: An example of a collider with a descendant.

For the rest of this thesis, we will be making the standard *faithfulness assumption* in causal discovery (Spirtes et al., 2000). This assumption states that any independencies in the full data law distribution $p(\mathbf{V})$ must correspond to d-separation statements in $\mathcal{G}$. Formally, $A \perp\!\!\!\perp B \mid \mathbf{C}$ in $p(\mathbf{V})$ if and only if $A \perp\!\!\!\perp_d B \mid \mathbf{C}$ in $\mathcal{G}$. Intuitively, this assumption states that independencies that are present in the observed data will be reflected by d-separation statements in the DAG that represents the causal structure of the data. There are, however, scenarios where the faithfulness assumption does not hold, such as when the effect of A on B is equal and opposite to the effect of B on C. In the data distribution $p(\mathbf{V})$, then, A will be uncorrelated with C. However, the DAG representation of the relationships between the variables is given by $A \rightarrow B \rightarrow C$ where A and C are clearly not d-separated. Here, independence in the data distribution does not correspond to d-separation. Situations like these, however, are rare enough that it is reasonable to assume that faithfulness holds at all times.

We make one final note about d-separation when working with sets of variables. We say that the set of variables $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given a set of variables $\mathbf{Z}$ if, for each possible pair of variables $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $X \perp\!\!\!\perp_d Y \mid \mathbf{Z}$.

2.4 Recoverability of Probability Distributions Under Missing Data

As described in the motivating example in Chapter 1, we use the variables Y, $Y^{(1)}$, and R_Y to represent missingness in the variable Y. Y represents observed values of the variable, which can be a value or a question mark '?' if the value was unreported for an individual; $Y^{(1)}$ represents the potential outcomes of the value of Y had we hypothetically been able to observe Y for all individuals in the dataset; and R_Y – known as the missingness indicator for Y – represents whether or not we observed a value or a question mark for Y. Here, we build off of previous work that use DAGs to represent substantive assumptions about causal relations among variables, including indicators of missingness (Daniel et al., 2012; Mohan et al., 2013; Mohan and Pearl, 2021). We start by formally defining MCAR, MAR, and MNAR missingness in the context of DAGs. We then define probability distributions of missing data and discuss the conditions under which they are identifiable.

First, the *missingness mechanism* for a dataset defines how the indicators of missingness are causally related to other variables in the dataset. When using DAGs to represent missingness mechanisms, we may classify them into one of three types: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). Let $\mathbf{O}$ represent the set of variables in a dataset that are fully observed, i.e. no individual in the dataset has a value of '?' in the dataset for all values of $\mathbf{O}$. Let $\mathbf{S}$ represent the set of variables in the dataset that are partially observed, i.e. have some amount of missingness in them. Let $\mathbf{S^{(1)}}$ represent the set of variables that represent the potential outcomes of the variables in $\mathbf{S}$. Finally, let $\mathbf{R}$ represent the set of variables that represent the indicators of missingness for all variables in the set $\mathbf{S}$. In addition, a restriction that we impose on missing data DAGs is that variables in $\mathbf{R}$ cannot be parents of variables in the sets $\mathbf{O}$ and $\mathbf{S^{(1)}}$. By definition, variables in $\mathbf{R}$ will always be a parent of variables in $\mathbf{S}$.

It is important to note that there are missingness mechanisms that cannot be represented graphically and are therefore excluded from this discussion. One example of such a missingness mechanism is the discrete choice model proposed by Tchetgen et al. (2018). Another missingness mechanism that cannot be represented graphically is the MAR model proposed in the seminal work Rubin (1976) where the authors provide a sound and complete criterion for when it is acceptable to ignore the process that causes missing data. Despite these limitations, the framework of graphically representing missing data is still useful in understanding the assumptions encoded in the problem and in deriving theoretical guarantees regarding the behavior of the data. Hence, we focus on graphical representations of missing data in this thesis. Note that, although missingness mechanisms can have the same names in non-graphical and graphical contexts, they represent distinct sets of assumptions in the two contexts. In this thesis, we specifically work with graphical criteria for missingness.

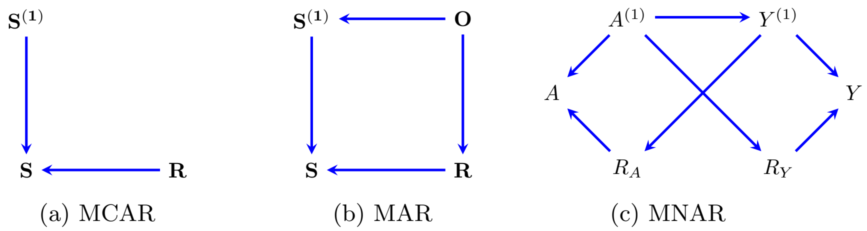
If $\mathbf{R} \perp\!\!\!\perp (\mathbf{S^{(1)}}, \mathbf{O})$, then we say that the missingness mechanism is MCAR. Figure 2.4a shows a DAG that generates an MCAR missingness mechanism. Here, the set $\mathbf{O}$ is an empty set, and the independence for the MCAR criterion holds because the vertices, in order, $\mathbf{S^{(1)}}$, $\mathbf{S}$, and $\mathbf{R}$, form a collider. Here, we give an example of a missingness process that is MCAR. Imagine if researchers start out with a complete dataset about condom use without missing data. However, a research assistant accidentally spills coffee on the machine storing the data, corrupting each entry of data with a probability of 0.5[1]. This would be an example of when the missingness mechanism of the data is MCAR because the missingness mechanism is independent of variables in the dataset.

Next, a missingness mechanism is MAR if $\mathbf{R} \perp\!\!\!\perp \mathbf{S^{(1)}} \mid \mathbf{O}$ holds in the dataset. Intuitively, data is MAR when the variables exhibiting missingness are independent of whether or not they are missing when controlling for variables that are fully observed. Figure 2.4b shows a DAG where the MAR criterion holds. There is a fork between the sets of variables $\mathbf{S^{(1)}}$ and $\mathbf{R}$, but conditioning on the set

[1] This example is inspired by a real-life event that transpired to the authors.

$\mathbf{O}$ is sufficient to d-separate the two sets of variables. To make the concept more concrete, imagine that $\mathbf{O}$ consists of a variable income that affects both whether or not an individual uses condoms and whether or not they end up participating in the survey. Like in Figure 2.4b, this creates a fork between the vertices $\mathbf{S^{(1)}}$ and $\mathbf{R}$ with $\mathbf{O}$. Much of the missing data literature thus far has focused on methods that are only applicable to MAR scenarios.

Finally, we consider a missingness mechanism to be MNAR when neither the MCAR nor MAR conditions hold for a dataset[2]. In other words, whether or not partially observed variables are missing is dependent on other partially observed variables in an MNAR missingness mechanism. Figure 2.4c shows an example of an MNAR missingness mechanism where $\mathbf{S^{(1)}} = \{A^{(1)}, Y^{(1)}\}$, $\mathbf{S} = \{A, Y\}$, $\mathbf{R} = \{R_A, R_Y\}$, and $\mathbf{O} = \emptyset$. The MCAR condition $\mathbf{R} \perp\!\!\!\perp (\mathbf{S^{(1)}}, \mathbf{O})$ does not hold here because there are edges directly between variables in $\mathbf{R}$ and variables in $\mathbf{S^{(1)}}$. Since $\mathbf{O}$ is the empty set in this example, the MAR condition does not hold either. In general, MNAR missingness mechanisms are expected to occur in real-life observational studies despite it being the most difficult form of missingness to correct for.



(a) MCAR (b) MAR (c) MNAR

Figure 2.4: DAGs representing MCAR, MAR, and MNAR missingness mechanisms.

Regardless of the type of missingness that we observe in the dataset, we may induce a statistical model from the DAG representation of the missing data problem like we would a typical DAG. According to equation 2.1, the DAG shown in Figure 2.4b factorizes as

$$p(\mathbf{S^{(1)}}, \mathbf{S}, \mathbf{R}, \mathbf{O}) = p(\mathbf{O}) \times p(\mathbf{S^{(1)}} \mid \mathbf{O}) \times p(\mathbf{R} \mid \mathbf{O}) \times p(\mathbf{S} \mid \mathbf{S^{(1)}}, \mathbf{R}).$$

Now, we introduce concepts related to identification under missing data. The distribution $p(\mathbf{S^{(1)}}, \mathbf{R}, \mathbf{O})$, which includes only the non-deterministic parts of the missing data DAG, is known as the *full law*. We say that variables in $\mathbf{S}$ are *deterministic* because they depend fully on variables in $\mathbf{S^{(1)}}$ and $\mathbf{R}$ and do not have any notions of error or randomness associated with them. Because we are working with missing data, we only have access to the probability distribution $p(\mathbf{S}, \mathbf{R}, \mathbf{O})$. This is known as the *observed data distribution*. We define the *target law* of the missing data problem to be $p(\mathbf{S^{(1)}}, \mathbf{O})$, the probability distribution of fully observed variables and partially observed variables if the partially observed variables had hypothetically been fully observed (Nabi et al., 2020). We say that a missing data problem is *recoverable* if it is possible to non-parametrically – meaning not assuming anything about the underlying structure and shape of the data – define the target law as a function of the observed data distribution. When a missing data problem is *recoverable*, it is by

[2] When the MAR condition does not hold, this implies that the MCAR condition also does not hold. Therefore, just the MAR assumption not holding is sufficient to imply MNAR.

definition also identified because the distributions of potential outcomes of the variables exhibiting missingness are included in the target law.

Here, we show how all MCAR and MAR missingness mechanisms are recoverable (Rubin, 1976). To demonstrate this, we use a property of conditional probability distributions known as *conditional ignorability*. This property states that, if $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, then

$$p(X \mid \mathbf{Z}) = p(X \mid \mathbf{Z}, Y) \tag{2.3}$$

Intuitively, conditional ignorability states that controlling for the variable Y does not change anything about the probability distribution for the variable X when conditioning on the set of variables $\mathbf{Z}$ if X and Y are independent when controlling for $\mathbf{Z}$. Going the other way, we are also allowed to drop Y from the conditioning bar if independence between X and Y given $\mathbf{Z}$ holds. Next, the equations that define the relationship between the variables Y, $Y^{(1)}$, and R_Y are reproduced below:

$$Y = \begin{cases} Y^{(1)} & \text{if } R_Y = 1 \\ ? & \text{if } R_Y = 0 \end{cases} \tag{2.4}$$

Equation 2.4 is known as *missing data consistency*, which states that when $\mathbf{R} = 1$, then $\mathbf{S^{(1)}} = \mathbf{S}$. We also use this property to show how to recover the target law in the MCAR and MAR scenarios.

Our recovery strategy of the target law relies on the chain rule of probability – defined as $p(A, B) = p(A \mid B) \times p(B)$. First, if we evaluate the observed data distribution at only where $\mathbf{R} = 1$, this is equivalent to the probability distribution $p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})$. We may prove this property as follows:

$$\begin{aligned} p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S}) &= p(\mathbf{O}, \mathbf{S} \mid \mathbf{R} = 1) \times p(\mathbf{R} = 1) \text{ \textit{(by chain rule of probability)}} \\ &= p(\mathbf{O}, \mathbf{S^{(1)}} \mid \mathbf{R} = 1) \times p(\mathbf{R} = 1) \text{ \textit{(by missing data consistency)}} \\ &= p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}}) \text{ \textit{(by chain rule of probability)}} \end{aligned}$$

Next, the chain rule of probability allows us to recover the target law as shown in equation 2.5.

$$\begin{aligned} p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}}) &= p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O}) \times p(\mathbf{S^{(1)}}, \mathbf{O}) \text{ \textit{(by chain rule of probability)}} \\ p(\mathbf{S^{(1)}}, \mathbf{O}) &= \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})} \end{aligned} \tag{2.5}$$

The numerator on the right-hand side of equation 2.5 is simply the observed data distribution restricted to only observed rows of data. The denominator $p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})$, on the other hand, is known as the *propensity score* of the missingness indicators because it represents the propensity that a piece of data will be observed. Recovery of the target law is hence dependent on whether or not we are able to identify the propensity score as a function of the observed data distribution. To obtain the data distribution $p(\mathbf{S^{(1)}}, \mathbf{O})$ in practice, we may "divide" the data distribution $p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})$ by assigning each individual in the dataset a weight of $\frac{1}{p(\mathbf{R}=1 \mid \mathbf{S^{(1)}}, \mathbf{O})}$. These weights are known as the *inverse probability weights*. Then, we resample the dataset representing $p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})$ with replacement using the inverse probability weights. The resampled dataset will be representative of the data distribution $p(\mathbf{S^{(1)}}, \mathbf{O})$.

Using this recovery strategy along with conditional ignorability, we show that it is always possible to recover the target law under MCAR and MAR missingness mechanisms. First we show how the target law is recovered under an MCAR missingness mechanism. Recall that $\mathbf{R} \perp\!\!\!\perp (\mathbf{S^{(1)}}, \mathbf{O})$ when MCAR holds.

$$p(\mathbf{S^{(1)}}, \mathbf{O}) = \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})} \textit{ (by equation 2.5)}$$

$$p(\mathbf{S^{(1)}}, \mathbf{O}) = \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1)} \textit{ (by conditional ignorability)}$$

Hence, we may recover the target law under MCAR missingness by identifying the marginal propensity of missingness for each partially observed variable.

Next, we show how to recover the target law under a MAR missingness mechanism. Recall that $\mathbf{R} \perp\!\!\!\perp \mathbf{S^{(1)}} \mid \mathbf{O}$ under a MAR missingness mechanism.

$$p(\mathbf{S^{(1)}}, \mathbf{O}) = \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})} \textit{ (by equation 2.5)}$$

$$p(\mathbf{S^{(1)}}, \mathbf{O}) = \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1 \mid \mathbf{O})} \textit{ (by conditional ignorability)}$$

To recover the target law under MAR missingness, then, we need to find the propensity score conditional on the set of fully observed variables $\mathbf{O}$. Because both the set of variables $\mathbf{R}$ and $\mathbf{O}$ are fully observed, this propensity score is identified from the observed data distribution by calculating or creating a model for the probability that $\mathbf{R} = 1$ as a function of variables in $\mathbf{O}$. Therefore, the target law is always recoverable under a MAR missingness mechanism as well.

For MNAR missingness mechanisms, the target law is not guaranteed to be recoverable, and recovery sometimes involves using advanced techniques. Amazingly, though, the target law is recoverable for the MNAR missingness mechanism shown in Figure 2.4c by just using conditional ignorability. As discussed above, recovery of the target law is possible whenever the propensity score for each missingness indicator in $\mathbf{R}$ is recoverable. Hence, our goal is to recover $p(R_A = 1 \mid Y^{(1)}, A^{(1)})$ and $p(R_Y = 1 \mid Y^{(1)}, A^{(1)})$. Using the graph shown in Figure 2.4c, we may deduce the two following d-separation independencies: $R_A \perp\!\!\!\perp A^{(1)} \mid Y^{(1)}$ and $R_Y \perp\!\!\!\perp Y^{(1)} \mid A^{(1)}$. Conditional ignorability then allows us to conclude that $p(R_A = 1 \mid Y^{(1)}, A^{(1)}) = p(R_A = 1 \mid Y^{(1)})$ and $p(R_Y = 1 \mid Y^{(1)}, A^{(1)}) = p(R_Y = 1 \mid A^{(1)})$.

Next, we observe two further d-separation independencies – $R_A \perp\!\!\!\perp R_Y \mid Y^{(1)}$ and $R_Y \perp\!\!\!\perp R_A \mid A^{(1)}$. We once again use conditional ignorability; we insert R_Y past the conditioning bar for the propensity score of R_A and R_A past the conditioning bar for the propensity score of R_Y as follows: $p(R_A = 1 \mid Y^{(1)}) = p(R_A = 1 \mid Y^{(1)}, R_Y)$ and $p(R_Y = 1 \mid A^{(1)}) = p(R_Y = 1 \mid A^{(1)}, R_A)$. From the definition of conditional probability, we may consider the propensity score of R_A at only values where $R_Y = 1$ and the propensity score of R_Y at only values where $R_A = 1$. From missing data consistency, then, the potential outcomes $Y^{(1)}$ and $A^{(1)}$ will both reduce to the observed data variables Y and

A in both propensity scores. Therefore, the two propensity scores we need to recover the target law – $p(R_A = 1 \mid Y, R_Y = 1)$ and $p(R_Y = 1 \mid A, R_A = 1)$ – are both non-parametrically defined by the observed data law. Hence, the target law is recoverable despite the challenges posed by an MNAR missingness mechanism.

As discussed above, however, not all MNAR missingness mechanisms are recoverable by using simple conditional ignorability arguments. Consider Figure 2.5 that shows a unique type of MNAR missingness known as *self-censoring.* This type of missingness occurs when the value of a variable directly affects whether or not that value is reported. Because of the direct edge from $Y^{(1)}$ to R_Y, we may not use conditional ignorability, and it is not immediately clear how we might be able to define $p(R_Y = 1 \mid Y^{(1)}, Z_1)$ in terms of the observed data law. Because of this difficulty, self-censoring is widely regarded as the most difficult form of missingness to adjust for.

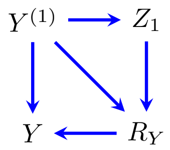


Figure 2.5: A self-censoring missingness mechanism where it is difficult to recover the target law.

Despite the difficulties of recoverability under MNAR data, work from Bhattacharya et al. (2019) and Nabi et al. (2020) has proved a sound and complete criterion for when it is possible to recover the full law – which we can use to recover the target law – of a missing data DAG model in the presence of MNAR missingness. Their criterion, however, specifically excludes self-censoring in addition to other graphical structures in the DAG. In other words, self-censoring prevents non-parametric identification of the full law unless additional (non-structural) assumptions are made about the data generating process. This does not mean that parameters of interest, such as average causal effects, are completely unrecoverable, though. In this thesis, we consider methods specifically for recovering from outcome-dependent self-censoring in spite of the challenges of working with self-censored data.

It is also worth noting here the similarities between the g-formula in equation 2.2 and the missing data identification equation 2.5. Equation 2.5 is actually a direct application of the g-formula on the observed data distribution when intervening on the missingness indicators to have a value of 1 as shown below. Note that $\mathbf{O^{(1)}} = \mathbf{O}$ because the value of variables in $\mathbf{O}$ do not depend on values in $\mathbf{R}$. Furthermore, below, we shorten the notation $(\mathbf{R} = 1)$ in the superscript to just (1) to save space.

$$\begin{aligned}
p((\mathbf{O}, \mathbf{S})^{(1)}) &= p(\mathbf{O}, \mathbf{S^{(1)}}) \quad \textit{(}\mathbf{O^{(1)}} = \mathbf{O}\textit{)} \\
&= \left.\frac{p(\mathbf{O}, \mathbf{R}, \mathbf{S})}{p(\mathbf{R} \mid \mathbf{S}, \mathbf{O})}\right|_{\mathbf{R}=1} \quad \textit{(by equation 2.2)} \\
&= \left.\frac{p(\mathbf{O}, \mathbf{R}, \mathbf{S^{(1)}})}{p(\mathbf{R} \mid \mathbf{S^{(1)}}, \mathbf{O})}\right|_{\mathbf{R}=1} \quad \textit{(by missing data consistency)} \\
&= \frac{p(\mathbf{O}, \mathbf{R} = 1, \mathbf{S^{(1)}})}{p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})} \quad \textit{(same as equation 2.5)}
\end{aligned}$$

As a final note regarding how we will represent missing data DAGs for the rest of this thesis, we will omit all variables representing just observed data, such as Y, from all future DAGs because of the deterministic nature of such a variable on its potential outcome and missingness indicator.

2.5 Inverse Probability Weighted Estimator for Recovering from Missing Data

In the previous section, we considered situations under which the target law is recoverable. In this section, we consider how we may be able to practically recover the expected value of the potential outcome $\mathbb{E}[Y^{(1)}]$ when the target law is recoverable. Let us consider a case of self-censoring as shown in Figure 2.6 where the full data law is provably unrecoverable (Bhattacharya et al., 2019; Nabi et al., 2020). Here, the variable Y contains missing data and causes its own missingness; this is represented by the edge from $Y^{(1)}$ to R_Y. Furthermore, we have a set of covariates $\mathbf{Z}$ that are also parents of R_Y.

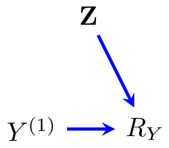


Figure 2.6: A case of self-censoring that will demonstrate recovery of $\mathbb{E}[Y^{(1)}]$.

In order to recover the expected value of $Y^{(1)}$, we may use an *inverse probability weighted* (IPW) estimator that follows directly from the identification results that we have discussed in section 2.4. Intuitively, this method finds the likelihood for researchers to observe a value for Y for each individual in the dataset. Then, for each individual we do get to observe, we *re-weight* by their inverse probability of response. For individuals who are very unlikely to respond, we give their response a higher weight and vice versa. On average, then, we will obtain the expected value of $Y^{(1)}$. Equation 2.6 shows the formula for using the IPW estimator for recovering from missing data. The usability of this equation is dependent on our ability to recover the propensity score $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$, and the reader will notice that this is a similar condition to that of equation 2.5. In equation 2.5, identification of the target law relies on defining the propensity score $p(\mathbf{R} = 1 \mid \mathbf{S^{(1)}}, \mathbf{O})$ in terms of the observed data distribution. Here, we require defining $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ in terms of the observed data distribution.

$$\mathbb{E}[Y^{(1)}] = \mathbb{E}[\frac{Y \times R_Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})}] \tag{2.6}$$

Despite self-censoring preventing us from recovering the full data law, we are still able to recover the expected value of the potential outcome despite missing data and self-censoring on the condition that the propensity score is known using equation 2.6. The strategy for identifying the propensity score from the observed data distribution, however, is not always obvious, especially when a variable is self-censoring. We will discuss identification strategies of the propensity score under self-censoring settings in Chapter 4.

Here, we give the proof for equation 2.6. In the proof, we use the *law of the unconscious statistician*, which states that $\mathbb{E}[g(\mathbf{V})] = \sum_{\mathbf{v}} g(\mathbf{v}) \times p(\mathbf{V})$, where $p(\mathbf{V})$ is the probability distribution of the set of variables $\mathbf{V}$.

Proof.

$$
\begin{aligned}
\mathbb{E}[\frac{Y \times R_Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})}] &=^{(1)} \sum_{R_Y, Y^{(1)}, \mathbf{Z}, Y} p(R_Y, Y^{(1)}, \mathbf{Z}, Y) \times \frac{Y \times R_Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})} \\
&=^{(2)} \sum_{Y^{(1)}, \mathbf{Z}} p(R_Y = 1, Y^{(1)}, \mathbf{Z}) \times \frac{Y^{(1)}}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})} \\
&=^{(3)} \sum_{Y^{(1)}, \mathbf{Z}} p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(Y^{(1)} \mid \mathbf{Z}) \times p(\mathbf{Z}) \times \frac{Y^{(1)}}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})} \\
&=^{(4)} \sum_{Y^{(1)}, \mathbf{Z}} p(Y^{(1)}) \times p(\mathbf{Z}) \times Y^{(1)} \\
&=^{(5)} \sum_{Y^{(1)}} p(Y^{(1)}) \times Y^{(1)} \\
&=^{(6)} \mathbb{E}[Y^{(1)}]
\end{aligned}
$$

□

In (1), we apply the law of the unconscious statistician; in (2), we evaluate the sum over R_Y and apply missing data consistency; in (3), we apply the chain rule of probability; in (4), we cancel out the term $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ from the numerator and denominator and use conditional ignorability from the d-separation $Y^{(1)} \perp\!\!\!\perp \mathbf{Z}$ in Figure 2.6; in (5), we sum over the set of variables $\mathbf{Z}$, which evaluates to 1; and in (6), we apply the definition of expectation.

2.6 Recovering Causal Effects (Adjusting for Confounding Bias)

Now that we have considered how we may recover from missing data in an observational study, let us consider how we may be able to recover from confounding bias if there were no missing data in our dataset. Because there is no missing data in this scenario, we simply use the variable Y to represent the outcome instead of the potential outcome $Y^{(1)}$. Much like missing data, we use DAGs to represent problems where we must adjust for confounding bias. Figure 2.7 shows one such DAG. Let us continue to use our motivating example where A represents whether or not an individual received the sexual education course and Y represents whether or not they use condoms frequently. Here, we also have a set of *confounders* $\mathbf{Z}$. A confounder is any variable that lies on a non-causal path between the treatment and outcome, where a causal path between the treatment and outcome is a directed path from the treatment to the outcome.

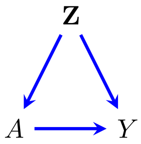


Figure 2.7: DAG showing a confounding bias problem.

Intuitively, when we do not control, or adjust, for the variables in **Z** in the DAG, then non-causal information will leak from the treatment to the outcome through the *backdoor path* $A \leftarrow \mathbf{Z} \rightarrow Y$. Backdoor paths are any paths between the treatment and outcome where the path starts with an edge pointing into the treatment. If we are able to block all backdoor paths by controlling for different variables in the DAG, then we will have successfully adjusted for the confounding bias. We may use d-separation to help us determine which variables we must control for to block all backdoor paths.

Let us now define the problem more formally. Just like in missing data, we target potential outcomes. However, instead of the potential outcome had we hypothetically been able to observe it for all individuals like in missing data, we are interested in the potential outcome had we hypothetically been able to assign a specific treatment to every individual in the dataset. The potential outcome $Y^{(A=0)}$ represents the value of the outcome had we hypothetically assigned a treatment of $A = 0$ to the individual. Similarly, the potential outcome $Y^{(A=1)}$ represents the value of the outcome had we hypothetically assigned a treatment of $A = 1$ to the individual. A value for both potential outcomes exists regardless of which treatment was actually assigned to an individual, but we only get to observe one of them. This relationship is known as *consistency* and is summarized by the equations

$$Y = \begin{cases} Y^{(A=0)} & \text{if } A = 0 \\ Y^{(A=1)} & \text{if } A = 1 \end{cases}$$

The reader will notice that in both adjusting for missing data and adjusting for confounding bias, we are concerned with recovering potential outcomes. In fact, the research areas of missing data and causal inference are often considered two sides of the same coin. In a way, adjusting for confounding bias is the same as adjusting for missing data because we have missing data for the potential outcomes of interest $Y^{(A=0)}$ and $Y^{(A=1)}$.

Using the equations defined above, we may add the potential outcomes $Y^{(A=0)}$ and $Y^{(A=1)}$ to our DAG to represent a confounding bias problem as shown in Figure 2.8[3]. As these quantities represent values of the outcome where the treatment is hypothetically set to some predefined value, the treatment A is not a cause of the potential outcomes. Similar to missing data, the values of the outcome variable Y are deterministically defined by the rules of consistency.

In general, the target parameter of inference when adjusting for confounding bias is the *average*

[3] This figure makes clear the relationship between missing data and causal inference; however, a general theory for the relationship between causal graphs and missing data graphs is still an open question. See Nabi et al. (2022) for details.

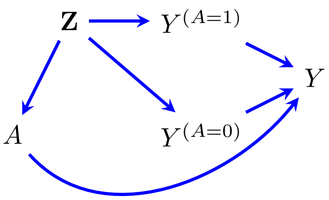


Figure 2.8: DAG with potential outcomes appended.

causal effect (ACE), which is defined as the expected value of the difference between $Y^{(A=1)}$ and $Y^{(A=0)}$. That is, ACE $= \mathbb{E}[Y^{(A=1)} - Y^{(A=0)}] = \mathbb{E}[Y^{(A=1)}] - \mathbb{E}[Y^{(A=0)}]$. In practice, we compute $\mathbb{E}[Y^{(A=1)}]$ and $\mathbb{E}[Y^{(A=0)}]$ separately and take their difference to find the average causal effect, a quantity that is adjusted for confounding bias.

It is impossible, however, to recover what the values of the potential outcomes are for each individual in the dataset. That is, the individual level causal effect is not identified in general. However, it may be possible to compute the average causal effect with only observed data using a technique called *backdoor adjustment* (Pearl, 1995, 2009). As described above, we focus on recovering the quantity $\mathbb{E}[Y^{(A=1)}]$; the procedure for recovering $\mathbb{E}[Y^{(A=0)}]$ follows similarly. At a high level, backdoor adjustment provides a set of graphical criteria that finds a set of variables to control for that blocks all non-causal flows of information from the treatment to the outcome. Then, it uses the *backdoor adjustment formula* to calculate the expected value of the potential outcomes. A set of variables $\mathbf{Z}$ is said to satisfy the *backdoor adjustment criterion* relative to a treatment A and outcome Y if

(B1) $\mathbf{Z}$ does not contain any variables that are causal descendants of A, and

(B2) A and Y are d-separated when conditioning on $\mathbf{Z}$ in a modified graph where all outgoing edges from A are deleted.

If a set $\mathbf{Z}$ is a valid backdoor adjustment set, then we have the following backdoor adjustment formula:

$$\mathbb{E}[Y^{(A=1)}] = \sum_{\mathbf{Z}} \mathbb{E}[Y \mid A = 1, \mathbf{Z}] \times p(\mathbf{Z}) \tag{2.7}$$

Below, we also give a proof for equation 2.7 in the general case. In (1), we apply the law of the unconscious statistician. In (2), we use the law of total probability to sum over the set of variables $\mathbf{Z}$ and union $\mathbf{Z}$ with the probability distribution of $Y^{(A=1)}$. In (3), we use the chain rule of probability. In (4), we know that the potential outcome is d-separated from the treatment when conditioning on the set of variables $\mathbf{Z}$ – $Y^{(A=1)} \perp\!\!\!\perp A \mid \mathbf{Z}$ – from the structure of the graph and condition (B2). All paths that start with an outgoing edge from A are blocked from the potential outcomes by the collider at Y. Since the potential outcomes are defined to be parents of the observed outcome via a deterministic relationship, if A is d-separated from the observed outcome Y conditional on $\mathbf{Z}$ according to (B2), then A must also be d-separated from the potential outcomes conditional on

$\mathbf{Z}$. Using this independence and conditional ignorability, we insert A past the conditioning bar of $p(Y^{(A=1)} \mid \mathbf{Z})$ in (4). Next, by the definition of conditional independence, $Y^{(A=1)}$ is independent of A when we consider only individuals where $A = 1$ because $Y^{(A=1)}$ is independent of A for all values of A. Hence, in (5), we use the definition of conditional independence to consider only individuals where $A = 1$. According to consistency, $Y = Y^{(A=1)}$ when $A = 1$. Therefore, we may also substitute Y into the equation in step (5). In (6), we use the law of the unconscious statistician to convert the probability distribution of Y back into an expected value.

Proof.

$$\begin{aligned}
\mathbb{E}[Y^{(A=1)}] &=^{(1)} \sum_{Y^{(A=1)}} p(Y^{(A=1)}) \times Y^{(A=1)} \\
&=^{(2)} \sum_{Y^{(A=1)},\mathbf{Z}} p(Y^{(A=1)}, \mathbf{Z}) \times Y^{(A=1)} \\
&=^{(3)} \sum_{Y^{(A=1)},\mathbf{Z}} p(Y^{(A=1)} \mid \mathbf{Z}) \times p(\mathbf{Z}) \times Y^{(A=1)} \\
&=^{(4)} \sum_{Y^{(A=1)},\mathbf{Z}} p(Y^{(A=1)} \mid A, \mathbf{Z}) \times p(\mathbf{Z}) \times Y^{(A=1)} \\
&=^{(5)} \sum_{Y,\mathbf{Z}} p(Y \mid A = 1, \mathbf{Z}) \times p(\mathbf{Z}) \times Y \\
&=^{(6)} \sum_{\mathbf{Z}} \mathbb{E}[Y \mid A = 1, \mathbf{Z}] \times p(\mathbf{Z})
\end{aligned}$$

□

2.7 Inverse Probability Weighted Estimator for Causal Effects

An alternate method for estimating the expected value of the potential outcome $\mathbb{E}[Y^{(a)}]$ from the backdoor adjustment formula is to use the IPW method, which is given by equation 2.8 where $\mathbf{Z}$ is a valid backdoor adjustment set (Horvitz and Thompson, 1952). The IPW estimating equation for recovering from confounding bias is directly analagous to that for missing data given in equation 2.6.

$$\mathbb{E}[Y^{(a)}] = \mathbb{E}\Big[\frac{\mathbb{1}(A = a) \times Y}{p(A = a \mid \mathbf{Z})}\Big] \tag{2.8}$$

The function $\mathbb{1}(A = a)$ is known as an *indicator function*; this indicator function has a value of 1 when $A = a$ and 0 otherwise. The indicator function ensures that we are only considering individuals with the value of the treatment that we would like to find the potential outcome for. This is analogous to the missing data IPW estimator considering only individuals where $R_Y = 1$. Similar to equation 2.6, we divide each individual by their inverse probability of treatment. This ensures that individuals who have a high probability of being assigned a particular treatment are given less weight while individuals that have a low probability of being assigned a particular treatment are given more weight. Finally, we take the empirical average of all the values of the outcome after the inverse probability reweighting to recover the expected value of the potential outcome. Note that individuals that are assigned a treatment $A \neq a$ contribute a value of 0 to the empirical average.

Here, we also give the proof showing that the IPW functional for recovering potential outcomes is equivalent to the backdoor adjustment formula. At a high level, we use the law of the unconscious statistician and the chain rule of probability to show equivalence between the two estimating equations. As this proof is similar to previous proofs in this chapter, we omit a detailed explanation.

Proof.

$$
\begin{aligned}
\mathbb{E}\Big[\frac{\mathbb{1}(A=a)\times Y}{p(A=a \mid \mathbf{Z})}\Big] &= \sum_{Y,A,\mathbf{Z}} p(Y,A,\mathbf{Z}) \times \frac{\mathbb{1}(A=a)\times Y}{p(A=a \mid \mathbf{Z})} \\
&= \sum_{Y,\mathbf{Z}} p(Y,A=a,\mathbf{Z}) \times \frac{Y}{p(A=a \mid \mathbf{Z})} \\
&= \sum_{Y,\mathbf{Z}} p(Y \mid A=a,\mathbf{Z}) \times p(A=a \mid \mathbf{Z}) \times p(\mathbf{Z}) \times \frac{Y}{p(A=a \mid \mathbf{Z})} \\
&= \sum_{Y,\mathbf{Z}} p(Y \mid A=a,\mathbf{Z}) \times p(\mathbf{Z}) \times Y \\
&= \sum_{\mathbf{Z}} \mathbb{E}[Y \mid A=a,\mathbf{Z}] \times p(\mathbf{Z})
\end{aligned}
$$

□

Chapter 3

Previous Work

In this chapter, we give a brief overview of previous work on missing data and methods for searching for valid covariate adjustment sets. We show how previous work thus far has made significant progress but is still lacking for addressing missing data problems with outcome-dependent self-censoring. At the end of this chapter, we introduce work from Entner et al. (2013) for covariate selection that we will extend in this thesis for identification of the causal effect.

3.1 Previous Work on Missing Data

As discussed in Chapter 2, we focus on graphical representations of missing data. Further, it is always possible to recover from MCAR and MAR missingness mechanisms (Daniel et al., 2012; Rubin, 1976). MNAR missingness mechanisms, specifically that of self-censoring, though, remain relatively unexplored (Mohan and Pearl, 2021). Recent work from Bhattacharya et al. (2019) and Nabi et al. (2020) have shown that it is always possible to recover the full data law as long as self-censoring and certain graphical structures are not present in the DAG. However, their criterion specifically excludes self-censoring, which is the focus of this thesis. In this section, we first provide a brief overview of previous work on self-censoring. We then discuss previous work on MNAR missingness in general.

This paragraph gives a quick overview on related work in self-censoring. Sportisse et al. (2020) propose an imputation method for self-censored data that assumes factorization according to a certain latent variable DAG and parametric models for the missingness process. Mohan et al. (2018) propose methods for recovery of the full data law when all variables are discrete in linear systems and when certain matrices corresponding to conditional probability tables are invertible. Duarte et al. (2021) propose an algorithm for computing bounds on the causal effect in the discrete setting, which may converge to point identification in certain cases. d'Haultfoeuille (2010) and Tchetgen Tchetgen and Wirth (2017) propose instrumental variable methods that place homogeneity restrictions on the missingness process in addition to requiring the presence of a valid instrument. These methods provide advances in the literature, but they do not consider the missing data problem in conjunction with the confounding bias problem.

Next, we discuss works that provide methods for recovering causal effects under MNAR missingness mechanisms. Saadati and Tian (2019) propose a sound graphical criterion for when the causal effect is recoverable if any variable in the dataset may be MNAR. At a high level, their criterion checks if $Y^{(a)} \perp\!\!\!\perp \mathbf{R}$ holds in the DAG representing the problem, where $Y^{(a)}$ is the potential outcome for the outcome of interest at some treatment a and $\mathbf{R}$ is the set of all indicators of missingness. If this d-separation is true, then applying conditional ignorability to the backdoor adjustment formula allows us to recover the ACE. Their criterion is intuitive to check for and understand theoretically, but it does not allow for self-censoring on the outcome. Another limitation of their method is that we must already know the true structure of the DAG in order to use it.

In Yang et al. (2019), the authors similarly address the challenge of recovering causal effects under MNAR data. The limiting assumption of their work, however, is that the missingness mechanism must be outcome-independent. That is, $Y \perp\!\!\!\perp \mathbf{R}$ in the DAG representing the problem must be true, where Y is the outcome and $\mathbf{R}$ is the set of all missingness indicators. Under this assumption, the authors identify the causal effect by solving an integral equation, and they also propose a non-parametric two-stage least squares estimator. While the assumption of outcome-independent missingness is plausible if the outcome is measured after all the covariates and missingness indicators, there are many real-world situations where this assumption will not hold.

While significant progress has been made in recovering from self-censoring and causal effects under MNAR missingness mechanisms, recovering causal effects under outcome-dependent missingness and self-censoring without knowing the true structure of the DAG has not been explored thus far to our knowledge.

3.2 Known Covariate Selection Methods

In this section we discuss known methods for selecting a set of covariates to use for adjustment. Even when there is no missing data, finding a valid set $\mathbf{Z}$ that satisfies the backdoor adjustment criteria (B1) and (B2) for a given treatment and outcome is a challenge. Controlling for too little variables leaves backdoor paths open; on the other hand, controlling for too many variables may open colliders and open backdoor paths between the treatment and outcome. Since Pearl proposed the backdoor adjustment criterion and formula, researchers have devoted much effort into finding different methods for selecting a valid set of covariates for backdoor adjustment. Here, we review progress in the literature thus far.

Shpitser et al. (2012) give a complete graphical criterion for when covariate adjustment can identify the causal effect when the full structure of the DAG is known. Their criteria also takes into account covariate adjustment methods besides backdoor adjustment. Using their criterion, however, requires knowing the full structure of the graph, which, in practice, is often not available. One way to ameliorate this is to first use a *structural learning* algorithms to identify the full structure of the graph. Then, we may apply graphical criteria, such as that proposed by Shpitser et al. (2012), to recover the causal effect.

Structural learning algorithms, however, may not always be precise. For example, a challenge that structural learning algorithms must overcome is that multiple causal DAGs can imply the same set of conditional independencies or d-separations. DAGs that imply the same set of conditional independencies are known as *markov equivalent* (Verma and Pearl, 2022). Given a set of variables $\mathbf{V}$, the set of graphs that are all markov equivalent with each other is known as the *markov equivalence class* for $\mathbf{V}$. Often times, structural learning algorithms can only identify the markov equivalence class that includes the true structure of the graph for a given problem.

There are two main algorithms for structural learning – the PC algorithm and the greedy equivalence search (GES) algorithm (Spirtes et al., 2000; Chickering, 2002). The PC algorithm evaluates conditional independencies shown by the data to deduce undirected edges and colliders using the faithfulness assumption. It then uses a set of orientation rules to orient as many undirected edges as possible to arrive at a graph with undirected edges that represents a markov equivalence class. The GES algorithm, on the other hand, is a score-based structural learning algorithm that attempts to minimize a metric known as the Bayesian Information Criterion (BIC) score, a value that is inversely proportional to the number of conditional independencies that match between the true and proposed causal graph. GES starts with a forward phase that adds as many edges to a DAG as possible until the BIC score no longer improves with the addition of more edges. It then finishes with a backward phase that attempts to delete and reverse edges until the BIC score no longer improves with more edge reversals and deletions. As described above, a popular method in causal inference for covariate selection is to use a structural learning algorithm to estimate the structure of the DAG then using graphical criteria to find a valid set of covariates to adjust for.

In many scenarios, however, it is not necessary to recover the full structure of the DAG. For example, individual causal relationships between variables may not be relevant, and the researchers are only interested in finding a valid covariate adjustment set to recover the ACE. To address situations such as these, Entner et al. (2013) propose a sound method for covariate selection between a treatment A and outcome Y that does not require any prior knowledge of the true graph. Given that the set $\mathbf{W}$ represents the set of all covariates in the dataset, their method enumerates all possible assignments of a $W_i \in \mathbf{W}$ and $\mathbf{Z} \subseteq \{\mathbf{W} \setminus W_i\}$ such that

(i) $W_i \not\perp\!\!\!\perp Y \mid \mathbf{Z}$

(ii) $W_i \perp\!\!\!\perp Y \mid A, \mathbf{Z}$.

When an assignment of W_i and $\mathbf{Z}$ satisfies both (i) and (ii), then the set $\mathbf{Z}$ satisfies the backdoor criterion (B1) and (B2) with respect to a treatment A and outcome Y. Although this method does not recover the full structure of the DAG, it still shares a similarity with the PC algorithm in that it also performs a long series of conditional independence tests.

In this thesis, we extend the method for covariate selection from Entner et al. (2013) for use in missing data settings where the outcome suffers from self-censoring. As shown in section 3.1, however, recovering from MNAR missingness, and especially self-censoring, is particularly challenging. To address this, we use *shadow variables*, proposed by Miao et al. (2015), which is the topic of the next chapter.

Chapter 4

Shadow Variables

Shadow variables are auxiliary variables that help us recover the data distribution as well as the propensity score for the missingness indicator of a self-censoring variable (Miao et al., 2015). The propensity score for the missingness indicator of a self-censoring variable is of interest because it is key for applying the inverse probability weighted estimator shown in equation 2.6 for recovering the average causal effect.

A fully observed variable is only a valid shadow variable, however, if certain conditions that are generally untestable are met. Specifically, there is a test that requires having access to the potential outcome of the self-censoring variable, which is exactly one of the target parameters that the missingness problem does not have access to. Intuitively, shadow variables are variables that have some non-zero effect on a self-censoring variable, which means they contain some useful information about the self-censoring variable. The shadow variable is also independent of the missingness mechanism, which makes it possible to use the shadow variable to infer information about the self-censoring variable without being biased by missing data.

There are two methods in which we may use a shadow variable to recover from self-censoring: the matrix inversion method and the odds ratio factorization of the propensity score method. In this chapter, we give an overview of the assumptions and conditions of a shadow variable and how to use the two aforementioned methods for recovering from self-censoring.

4.1 Assumptions and Conditions

In this section, we introduce assumptions and conditions necessary for a variable to be considered as a valid shadow variable. We then give examples for when valid shadow variables exist in different DAGs. Let $Y^{(1)}$ be the potential outcome of a self-censoring variable Y and R_Y be the missingness indicator for this variable. Further, let $\mathbf{Z}$ be a set of fully observed covariates in the graph. A variable S is a valid shadow variable if it is a fully observed variable that satisfies the following independence relations (Miao et al., 2015):

(S1) $S \not\perp\!\!\!\perp Y^{(1)} \mid R_Y = 1, \mathbf{Z}$, and

(S2) $S \perp\!\!\!\perp R_Y \mid Y^{(1)}, \mathbf{Z}$.

When a variable S satisfies (S1) and (S2) for a set of variables $\mathbf{Z}$, then we refer to $\mathbf{Z}$ as the *shadow variable adjustment set*. For the rest of this thesis, we also assume an extension of faithfulness – as discussed in section 2.3 – used in causal discovery procedures for missing data settings known as *missing data faithfulness* (Tu et al., 2019). This assumption states that any independencies that exist conditional on R_Y in a data distribution $p(\mathbf{V})$ must also hold in the observed data, i.e., conditional on $R_Y = 1$. Formally, we assume that $A \perp\!\!\!\perp B \mid \mathbf{C}, R_Y$ in $p(\mathbf{V})$ if and only if $A \perp\!\!\!\perp B \mid \mathbf{C}, R_Y = 1$ in $p(\mathbf{V})$, when $Y^{(1)}$ is one of A, B, or an element of the conditioning set $\mathbf{C}$.

From missing data faithfulness and missing data consistency, condition (S1) is in general testable from observed data. However, condition (S2) is in general untestable as it requires conditioning on $Y^{(1)}$, the very variable we do not have access to in a self-censoring scenario. Despite this challenge in empirically confirming the validity of a shadow variable, we discuss theoretical foundations for when we are able to use and how we use shadow variables in this chapter. Figure 4.1 gives a graphical presentation for situations under which S is a valid shadow variable.

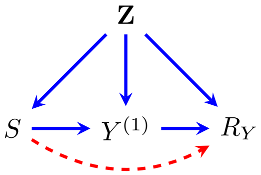


Figure 4.1: S is a valid shadow variable when the red dashed edge is absent.

Let us first consider the scenario where the red dashed edge is absent from the graph. By using d-separation, we can see that both (S1) and (S2) are satisfied. First, there is an edge directly between S and $Y^{(1)}$, which satisfies condition (S1). Intuitively, this condition is necessary as the shadow variable S must have some non-zero relationship with the self-censoring variable $Y^{(1)}$. Otherwise, having information about S would give us no useful information about $Y^{(1)}$. Next, there are three possible paths between S and R_Y: $S \rightarrow Y^{(1)} \rightarrow R_Y$, $S \leftarrow \mathbf{Z} \rightarrow R_Y$, and $S \leftarrow \mathbf{Z} \rightarrow Y^{(1)} \rightarrow R_Y$. According to d-separation, all three of these paths will be blocked when we control for $Y^{(1)}$ and $\mathbf{Z}$ as we do in (S2). Hence, (S2) is satisfied as well, and S is a valid shadow variable for overcoming self-censoring on the self-censoring variable Y. Intuitively, we need (S2) in order to make sure that S is not affected by missing data and can provide unbiased information on the self-censoring variable.

Next, let us consider the scenario where the red dashed edge is present in the graph. When such an edge exists, it is not possible to d-separate S and R_Y, so (S2) can never be fulfilled. Hence, S can never be a valid shadow variable under this circumstance. In addition to a direct edge from S to R_Y that can prevent S from being a valid shadow variable, the possibility for *unmeasured confounding* can also invalidate S from fulfilling (S1) and (S2). An *unmeasured confounder*, also known as *latent variable*, U is a variable in the data distribution $p(\mathbf{V})$ that is entirely unobserved; hence, we are not able to control for it. When unmeasured confounders exist in a DAG, we say that unmeasured confounding is present. The existence of unmeasured confounders complicates the process of finding a valid shadow variable because it leaves open flows of information between S and

R_Y that we cannot control for.

Consider Figure 4.2 where we have two unmeasured confounders U_1 and U_2. U_1 is a common cause between the self-censoring variable and its missingness indicator while U_2 is a common cause between the potential shadow variable S and the missingness indicator of the self-censoring variable R_Y. If the outgoing edges from U_1 were present in the graph, then (S2) can never be fulfilled because the path $S \rightarrow Y^{(1)} \leftarrow U_1 \rightarrow R_Y$ will be opened when we control for $Y^{(1)}$ and open the collider. On the other hand, if the outgoing edges from U_2 were present in the graph, then (S2) can also never be fulfilled; it would not be possible to block the path $S \leftarrow U_2 \rightarrow R_Y$ because we cannot control for U_2.

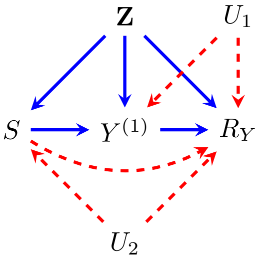


Figure 4.2: S as a valid shadow variable when the red dashed edges are absent.

In addition, we require that the data distribution $p(Y^{(1)} \mid R_Y = 1, S, \mathbf{Z})$ satisfy a widely used completeness condition in order for S to be considered a valid shadow variable. The completeness condition requires that for all square-integrable functions $h(S, Y^{(1)})$, $\mathbb{E}[h(S, Y^{(1)}) \mid R_Y = 1, S, \mathbf{Z}] = 0$ almost surely if and only if $h(S, Y^{(1)}) = 0$ almost surely. The reader may refer to Newey and Powell (2003) or Section 3 of Miao et al. (2015) for more details.

When a variable S fulfills (S1) and (S2) and the completeness condition, then we may use it to recover the probability distribution of the self-censoring variable with the matrix inversion method and the propensity score of the missingness indicator associated with the self-censoring variable with the odds ratio factorization method. We discuss the two methods in section 4.2 and 4.3, respectively.

Here, we give a note on using shorthand to represent unmeasured confounders. Often, we use bidirected edges between two variables to represent the existence of unmeasured confounding between them. For example, $Y^{(1)} \leftrightarrow R_Y$ is shorthand for $Y^{(1)} \leftarrow U_1 \rightarrow R_Y$. DAGs with bidirected edges added in such a manner are known as *acyclic directed mixed graphs* (ADMGs) (Richardson et al., 2023). Instead of d-separation, a similar concept known as *m-separation* applies to ADMGs (Richardson, 2003). Making independence statements between variables in m-separation is identical to that of d-separation except that the notion of colliders is extended to the following structures as well:

1. $X \rightarrow Z \leftrightarrow Y$
2. $X \leftrightarrow Z \leftarrow Y$
3. $X \leftrightarrow Z \leftrightarrow Y$

Figure 4.2 represented as an ADMG is shown in Figure 4.3.

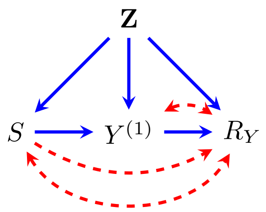


Figure 4.3: Figure 4.2 represented as an ADMG.

4.2 Matrix Inversion

The matrix inversion method allows us to recover the marginal probability distribution of the self-censoring variable $p(Y^{(1)})$ (Mohan and Pearl, 2021). To demonstrate this method, we consider S and $Y^{(1)}$ to be binary variables, but this method also generalizes to discrete S and $Y^{(1)}$ as long as S has the same or more levels than $Y^{(1)}$. Further, we do not include the shadow variable adjustment set $\mathbf{Z}$ in this discussion, but inserting it past the conditioning bar for each probability distribution is trivial and does not influence the correctness of the derivations. In the following demonstration of the method, the lower case variables s and $\bar{s}$ represent the two possible values for the variable S. The analog is also true for the variable $Y^{(1)}$. The method works as follows:

$$\begin{aligned} p(S) &=^{(1)} \sum_{Y^{(1)}} p(S \mid Y^{(1)}) \times p(Y^{(1)}) \\ \begin{pmatrix} p(\bar{s}) \\ p(s) \end{pmatrix} &=^{(2)} \begin{pmatrix} p(\bar{s} \mid \bar{y}^{(1)}) & p(\bar{s} \mid y^{(1)}) \\ p(s \mid \bar{y}^{(1)}) & p(s \mid y^{(1)}) \end{pmatrix} \begin{pmatrix} p(\bar{y}^{(1)}) \\ p(y^{(1)}) \end{pmatrix} \\ \begin{pmatrix} p(\bar{s} \mid \bar{y}^{(1)}) & p(\bar{s} \mid y^{(1)}) \\ p(s \mid \bar{y}^{(1)}) & p(s \mid y^{(1)}) \end{pmatrix}^{-1} \begin{pmatrix} p(\bar{s}) \\ p(s) \end{pmatrix} &=^{(3)} \begin{pmatrix} p(\bar{y}^{(1)}) \\ p(y^{(1)}) \end{pmatrix} \\ &= p(Y^{(1)}) \end{aligned}$$

In step (1), we apply the total law of probability to sum over $Y^{(1)}$ and the chain rule of probability to separate $p(S, Y^{(1)})$ into two data distributions. In step (2), we simply convert the previous equation into a matrix format. In step (3), we multiply the inverse of the square matrix to both sides of the equation to recover the data distribution $p(Y^{(1)})$ and conclude the matrix inversion method.

The matrix inversion method depends on the identification of the data distributions $p(S)$ and $p(S \mid Y^{(1)})$. First, $p(S)$ is identified because S is a fully observed variable. Next, $p(S \mid Y^{(1)})$ is identified as follows:

$$\begin{aligned} p(S \mid Y^{(1)}) &= p(S \mid Y^{(1)}, R_Y) \text{ (by (S2) and conditional ignorability)} \\ &= p(S \mid Y^{(1)}, R_Y = 1) \text{ (by definition of conditional probability)} \\ &= p(S \mid Y, R_Y = 1) \text{ (by missing data consistency)} \end{aligned}$$

The definition of conditional probability states that, if X is independent of Y for all values of Z, then X is independent of Y for a subset of possible values of Z. Therefore, both data distributions $p(S)$ and $p(S \mid Y^{(1)})$ are identified, and we can successfully recover the data distribution $p(Y^{(1)})$ with matrix inversion despite self-censoring.

4.3 Recovering Propensity Scores with the Odds Ratio Factorization

The odds ratio factorization of the propensity score, originally proposed by Yun Chen (2007) and used in Miao et al. (2015), allows us to identify the propensity score $p(R_Y = 1 \mid S, Y^{(1)}, \mathbf{Z})$. An *odds ratio* between two variables is a statistic that measures the level of association between two variables X and Y. It is defined in terms of probabilities as

$$OR(X, Y \mid Z) = \frac{p(X \mid Y, Z)}{p(X = x_0 \mid Y, Z)} \times \frac{p(X = x_0 \mid Y = y_0, Z)}{p(X \mid Y = y_0, Z)}$$

where x_0 is the *reference value* of X and y_0 is the *reference value* of Y. Reference values are arbitrary values within the domain of X and Y that we choose to define the odds ratio over. From the above definition, we can see that the odds ratio is 1 whenever $X = x_0$ or $Y = y_0$ (either variable is at its reference value). Further, the odds ratio is 1 for *all* values of X, Y, Z if and only if $X \perp\!\!\!\perp Y \mid Z$. This characterizes all settings where the OR can attain a value of 1. The odds ratio is also symmetric, i.e., $\text{OR}(X, Y \mid Z) = \text{OR}(Y, X \mid Z)$. The range of the odds ratio function is $(0, \infty)$.

Following Miao et al. (2015), we use an odds ratio factorization of the propensity score to perform the identification of $p(R_Y = 1 \mid S, Y^{(1)}, \mathbf{Z})$ where S is a valid shadow variable and $\mathbf{Z}$ is a valid shadow variable adjustment set for the self-censoring variable Y. From shadow variable condition (S2), we may use conditional ignorability to deduce that $p(R_Y = 1 \mid S, Y^{(1)}, \mathbf{Z}) = p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$. Hence, we focus on recovering the quantity $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ in this section. Without loss of generality we pick any arbitrary value y_0 to be the reference value for the outcome and $R_Y = 1$ to be the reference value of the missingness indicator. As discussed previously, the odds ratio between the outcome $Y^{(1)}$ and the missingness indicator R_Y is 1 whenever either variable is at its reference value. Let $\pi_0(\mathbf{Z})$ denote the propensity score for R_Y at the reference value y_0, i.e., $\pi_0(\mathbf{Z}) := p(R_Y = 1 \mid Y^{(1)} = y_0, \mathbf{Z})$. Let $\eta(Y^{(1)}, \mathbf{Z})$ denote the conditional odds ratio function relating $Y^{(1)}$ and R_Y at values where $R_Y = 0$, i.e., $\eta(Y^{(1)}, \mathbf{Z}) := \text{OR}(Y^{(1)}, R_Y = 0 \mid \mathbf{Z})$. Then the odds ratio factorization of the propensity score can be written as,

$$p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) = \frac{\pi_0(\mathbf{Z})}{\pi_0(\mathbf{Z}) + \eta(Y^{(1)}, \mathbf{Z})(1 - \pi_0(\mathbf{Z}))} \tag{4.1}$$

whenever $Y^{(1)} = y_0$, $\eta(Y^{(1)}, \mathbf{Z}) = 1$, and we have that $p(R_Y = 1 \mid Y^{(1)} = y_0, \mathbf{Z}) = \pi_0(\mathbf{Z})$. At any other value of $Y^{(1)}$, $\eta(Y^{(1)}, \mathbf{Z}) \neq 1$, and the odds ratio factorization of the propensity score will return a different value.

The derivation for 4.1 is as follows. First, according to Yun Chen (2007), the odds ratio factorization of the probability distribution $p(R_Y, Y^{(1)} \mid \mathbf{Z})$ is defined as

$$p(R_Y, Y^{(1)} \mid \mathbf{Z}) = \frac{p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})}{\sum_{R_Y, Y^{(1)}} p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})} \tag{4.2}$$

where y_0 is the reference value for the potential outcome of the self-censoring variable $Y^{(1)}$ and 1 is the reference value for the missingness indicator R_Y. Let $\psi = p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})$ and $\psi_1 = \psi|_{R_Y=1}$. The denominator of equation 4.2 is known as the *normalization term*. Below, we prove equation 4.1.

Proof.

$$\begin{aligned}
p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) &=_{(1)} \frac{p(R_Y = 1, Y^{(1)} \mid \mathbf{Z})}{p(Y^{(1)} \mid \mathbf{Z})} \\
&=_{(2)} \frac{p(R_Y = 1, Y^{(1)} \mid \mathbf{Z})}{\sum_{R_Y} p(R_Y, Y^{(1)} \mid \mathbf{Z})} \\
&=_{(3)} \frac{\frac{\psi_1}{\sum_{R_Y, Y^{(1)}} \psi}}{\frac{\sum_{R_Y} \psi}{\sum_{R_Y, Y^{(1)}} \psi}} \\
&=_{(4)} \frac{\psi_1}{\sum_{R_Y} \psi} \\
&=_{(5)} \frac{p(R_Y = 1 \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y = 1 \mid \mathbf{Z})}{\sum_{R_Y} p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})} \\
&=_{(6)} \frac{p(R_Y = 1 \mid Y^{(1)} = y_0, \mathbf{Z}) \times p(Y^{(1)} \mid R_Y = 1, \mathbf{Z})}{p(Y^{(1)} \mid R_Y = 1, \mathbf{Z}) \times \sum_{R_Y} p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})} \\
&=_{(7)} \frac{p(R_Y = 1 \mid Y^{(1)} = y_0, \mathbf{Z})}{\sum_{R_Y} p(R_Y \mid Y^{(1)} = y_0, \mathbf{Z}) \times \mathrm{OR}(Y^{(1)}, R_Y \mid \mathbf{Z})} \\
&=_{(8)} \frac{\pi_0(\mathbf{Z})}{\pi_0(\mathbf{Z}) + \eta(Y^{(1)}, \mathbf{Z})(1 - \pi_0(\mathbf{Z}))}
\end{aligned}$$

□

In (1), we apply the chain rule of probability. In (2), we apply the total law of probability to sum over R_Y. In (3), we apply equation 4.2 to both the numerator and the denominator. When applying equation 4.2 to $p(R_Y = 1, Y^{(1)} \mid \mathbf{Z})$ in the numerator, the normalization term is a sum over all values of R_Y. Further, in the denominator, the normalization term is not a function of R_Y, which allows us to move it out of the sum over R_Y. In (4), we cancel out like terms in both the numerator and denominator. In (5), we simply expand out ψ_1 and ψ according to our previous definitions of these two terms. In (6), we note that $\mathrm{OR}(Y^{(1)}, R_Y = 1 \mid \mathbf{Z})$ has R_Y at its reference value of 1; hence, it is equal to 1. Furthermore, we move the term $p(Y^{(1)} \mid R_Y = 1, \mathbf{Z})$ outside of the sum in the denominator because this term is not a function of R_Y. In (7), we cancel out like terms from the numerator and denominator. Finally, in (8), we sum out R_Y, which has only two possible values. When $R_Y = 1$, R_Y is at its reference value in the odds ratio, so the odds ratio term disappears, and we are just left with $\pi_0(\mathbf{Z})$. When $R_Y = 0$, we know that $p(R_Y = 0 \mid Y^{(1)} = y_0, \mathbf{Z}) = 1 - \pi_0(\mathbf{Z})$ and that neither $Y^{(1)}$ nor R_Y are at their reference values in the odds ratio term. Therefore, the odds ratio term remains.

Using the odds ratio factorization described in equation 4.1, identification of the propensity score reduces to defining $\pi_0(\mathbf{Z})$ and $\eta(Y^{(1)}, \mathbf{Z})$ in terms of the observed data distribution. At a high level, we may parameterize $\pi_0(\mathbf{Z}) = \text{expit}(\beta_1 Z_1 + \beta_2 Z_2 \ldots \beta_k Z_k)$ and $\eta(Y^{(1)}, \mathbf{Z}) = \exp(\gamma Y^{(1)})$, where k is the cardinality of the set $\mathbf{Z}$, $\exp(x) = e^x$ and $\text{expit}(x) = \frac{1}{1+\exp(-x)}$. When S is a valid shadow variable, we may use a system of estimating equations to recover the parameters $\beta_1, \beta_2, \ldots, \beta_k, \gamma$ in terms of the observed data distribution. Hence, we recover the propensity score of the missingness indicator $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ despite self-censoring. The practical procedure for estimating the propensity score will be discussed further in Chapter 6.

4.4 Equivalence of Matrix Inversion and Odds Ratio Factorization

Although the two methods for using a shadow variable to overcome self-censoring were proposed by different researchers and may appear fundamentally different, the underlying mathematical identity of both methods is actually the same. As discussed in section 4.3, we use a system of estimating equations to recover the parameters in the parameterization of $\pi_0(\mathbf{Z})$ and $\eta(Y^{(1)}, \mathbf{Z})$. The general form of the estimating equation is

$$\mathbb{E}\left[\frac{R_Y \times V}{p(R_Y=1 \mid Y^{(1)}, S)} - V\right] = 0,$$

where V is a variable or a constant. We now prove equivalence between the matrix inversion and odds ratio factorization methods by showing how the estimating equation above when using $V = S$ where S is the shadow variable can also recover the probability distribution $p(Y^{(1)})$ in the binary case. Like the matrix inversion method, this proof may be generalized to discrete S and $Y^{(1)}$ as long as S has the same or more levels than $Y^{(1)}$.

In (1), we use the law of the unconscious statistician. In (2), we multiply the probability distribution $p(R_Y, Y^{(1)}, S)$ into both terms inside the parentheses and add both sides of the equation by the term $\sum_{R_Y, Y^{(1)}, S} p(R_Y, Y^{(1)}, S) \times S$. On the left-hand side in (3), we evaluate the sum at values of $R_Y = 1$. On the right-hand side in (3), we sum over all values of R_Y and $Y^{(1)}$. R_Y and $Y^{(1)}$ only appear in the probability distribution, so they evaluate to 1 when we sum over all of their possible values. In (4), we use the chain rule of probability to expand the probability distribution $p(R_Y = 1, Y^{(1)}, S)$. In (5), we cancel out the like term $p(R_Y = 1 \mid Y^{(1)}, S)$ in the numerator and denominator on the left-hand side of the equation. In (6), we evaluate both sides of the equation at all values of S. We are considering the binary case, so values at $S = 0$ contribute nothing to the sum. In (7), we expand the left-hand side of the equation by explicitly summing over all possible values of $Y^{(1)}$. As we are considering the binary case, $Y^{(1)}$ has only values of 0 and 1. In (8), we substitute the expression $p(Y^{(1)} = 0)$ with the expression $1 - p(Y^{(1)} = 1)$; this substitution works because we are considering the binary case.

The terms $p(S=1), p(S=1 \mid Y^{(1)}=0)$, and $p(S=1 \mid Y^{(1)}=1)$ are all identified because S is a fully observed variable and because we may use conditional ignorability in conjunction with missing data consistency to identify the conditional probability distributions in terms of only observed data. We now have one equation with one unknown, which we can use to recover $p(Y^{(1)}=1)$ and, by extension, $p(Y^{(1)})$.

Proof.

$$\mathbb{E}\left[\tfrac{R_Y \times S}{p(R_Y=1|Y^{(1)},S)} - S\right] = 0$$

$$\sum_{R_Y,Y^{(1)},S} p(R_Y, Y^{(1)}, S) \times \left(\frac{R_Y \times S}{p(R_Y=1 \mid Y^{(1)},S)} - S\right) =^{(1)} 0$$

$$\sum_{R_Y,Y^{(1)},S} p(R_Y, Y^{(1)}, S) \times \frac{R_Y \times S}{p(R_Y=1 \mid Y^{(1)},S)} =^{(2)} \sum_{R_Y,Y^{(1)},S} p(R_Y, Y^{(1)}, S) \times S$$

$$\sum_{Y^{(1)},S} p(R_Y=1, Y^{(1)}, S) \times \frac{S}{p(R_Y=1 \mid Y^{(1)},S)} =^{(3)} \sum_{S} p(S) \times S$$

$$\sum_{Y^{(1)},S} p(R_Y=1 \mid Y^{(1)}, S) \times p(S \mid Y^{(1)}) \times p(Y^{(1)}) \times \frac{S}{p(R_Y=1 \mid Y^{(1)},S)} =^{(4)} \sum_{S} p(S) \times S$$

$$\sum_{Y^{(1)},S} p(S \mid Y^{(1)}) \times p(Y^{(1)}) \times S =^{(5)} \sum_{S} p(S) \times S$$

$$\sum_{Y^{(1)}} p(S=1 \mid Y^{(1)}) \times p(Y^{(1)}) =^{(6)} p(S=1)$$

$$p(S=1 \mid Y^{(1)}=0) \times p(Y^{(1)}=0) + p(S=1 \mid Y^{(1)}=1) \times p(Y^{(1)}=1) =^{(7)} p(S=1)$$

$$p(S=1 \mid Y^{(1)}=0) \times (1 - p(Y^{(1)}=1)) + p(S=1 \mid Y^{(1)}=1) \times p(Y^{(1)}=1) =^{(8)} p(S=1)$$

□

Chapter 5

Tests for Identifying Conditions of the ACE

In this chapter, we present our tests for identifying conditions of the ACE under outcome-dependent self-censoring and confounding bias. Furthermore, we allow for the possibility of unmeasured confounders. We start this chapter by formalizing our problem assumptions in the context of DAGs and connecting it back to the motivating example described in Chapter 1. We then present our identification strategy and estimating equation for the ACE. Afterwards, we present empirical tests to verify the conditions under which our identification strategy is valid. Finally, we discuss some of the limitations of our method.

5.1 Problem Setup and Assumptions

In this section, we define a generic setup for the missing data and confounding bias problem. We assume the causal structure of the system is represented via a directed acyclic graph $\mathcal{G}$ defined over a set of vertices $\mathbf{V} = \{A, Y^{(1)}, R_Y, Y, I\} \cup \mathbf{W} \cup \mathbf{U}$, where A represents the treatment variable, $Y^{(1)}$ represents the potential outcome of the outcome variable of interest, R_Y represents the corresponding binary missingness indicator for the outcome, Y represents the factual observed outcome – which may be either a numeric value or "?" if the observation is missing, I represents the incentive variable, $\mathbf{W}$ represents an observed set of pre-treatment covariates, and $\mathbf{U}$ denotes a set of unmeasured pre-treatment covariates. The entire set $\mathbf{V}$ is assumed to be causally sufficient, i.e., there are no additional unmeasured common causes of any two variables in $\mathbf{V}$. The incentive variable I is any variable that affects whether or not an individual responds to the outcome. One possibility for such a variable is the TACASI program discussed in section 1.1.

As discussed in Chapter 2, we make the faithfulness assumption, which states that any independencies in the distribution $p(\mathbf{V})$ must correspond to d-separation statements in $\mathcal{G}$ (Spirtes et al., 2000). For tests that involve conditioning on the missing outcome $Y^{(1)}$, we also assume an extension of faithfulness used in causal discovery procedures for missing data settings (Tu et al., 2019). This assumption states that any independencies that exist conditional on R_Y in $p(\mathbf{V})$ must also hold in the observed data, i.e., conditional on $R_Y = 1$. Formally, we assume $A \perp\!\!\!\perp B \mid \mathbf{C}, R_Y$ in $p(\mathbf{V})$ if and only if $A \perp\!\!\!\perp B \mid \mathbf{C}, R_Y = 1$ in $p(\mathbf{V})$, when $Y^{(1)}$ is one of A, B, or an element of the conditioning set $\mathbf{C}$.

The above assumptions are commonly used across most graphical model selection procedures (Spirtes et al., 2000). We now list and provide brief justification for additional structural assumptions important for our method. Let $\text{pa}_{\mathcal{G}}(V)$ denote the parents of V in the graph $\mathcal{G}$. We assume that the data are generated from a distribution $p(\mathbf{V})$ that is Markov and faithful with respect to a DAG $\mathcal{G}$ that satisfies the structural assumptions (M1)-(M4).

(M1) The only parents of Y are $Y^{(1)}$ and R_Y, i.e., $\text{pa}_{\mathcal{G}}(Y) = \{Y^{(1)}, R_Y\}$.

(M2) There is an edge from A to $Y^{(1)}$ and an edge from $Y^{(1)}$ to R_Y (i.e. the causal path $A \rightarrow Y^{(1)} \rightarrow R_Y$ exists in the graph).

(M3) The incentive I is randomly assigned (i.e. $\text{pa}_{\mathcal{G}}(I) = \emptyset$) and may only be a parent of the missingness indicator R_Y.

(M4) Y is not a parent of any variables in $\mathbf{V}$ and does not have any children, R_Y is not a parent of any variables in $\mathbf{V} \setminus \{Y\}$, A is not a parent of any variables in $\mathbf{W} \cup \mathbf{U}$, and $Y^{(1)}$ cannot be a parent of A nor any variables in $\mathbf{W} \cup \mathbf{U}$. That is, we have an ordering, $\{I\} \cup \mathbf{W} \cup \mathbf{U} < A < Y^{(1)} < R_Y < Y$.

Assumption (M1) and disallowing R_Y from having any children aside from Y in assumption (M4) are standard restrictions in missing data DAG models (Mohan et al., 2013). We require assumption (M2) as it simplifies our empirical tests. However, (M2) is a relatively mild assumption as the existence of these edges is the primary motivation for applying our method. Assumption (M3) makes sure that I is a valid proxy variable for designing indirect tests about the validity of the treatment as a shadow variable. Finally, assumption (M4) states that $\mathbf{W} \cup \mathbf{U}$ are all pre-treatment variables.

Figure 5.1 graphically displays assumptions (M1)-(M4). The red edges are assumed to exist whereas the blue edges may or not exist. We draw Y here to illustrate (M1), but we will omit it and the red dashed edges in all figures going forward due to the deterministic nature of its relation with $Y^{(1)}$ and R_Y.

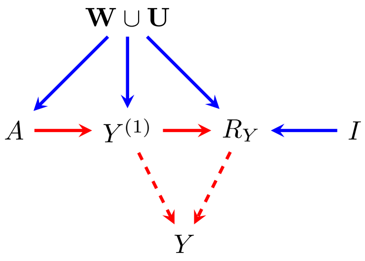


Figure 5.1: Graph depicting assumptions (M1)-(M4).

We may connect this graph back to the problem described in the motivating example in Chapter 1. A represents whether or not an individual participated in the sexual education course. $Y^{(1)}$

represents the potential outcome of whether or not individuals use condoms. R_Y is a binary variable that can have values of either 0, representing that an individual did not respond to whether or not they use condoms, or 1, representing an individual did respond to whether or not they use condoms. Y represents observed values of whether or not individuals use condoms; it has a value equal to $Y^{(1)}$ whenever R_Y has a value of 1, and it has a value equal to "?" whenever R_Y has a value of 0. The set $\mathbf{W}$ represents all fully observed confounders, and the set $\mathbf{U}$ represents all latent confounders. Finally, I represents the incentive variable, which, in this case, is whether or not an individual received a phone interview by a person or by the TACASI program.

Target of Inference

Under missingness, a causal effect of the treatment on the outcome corresponds to a contrast between the potential outcomes $Y^{(A=a,R_Y=1)}$ and $Y^{(A=a',R_Y=1)}$, where $Y^{(A=a,R_Y=1)}$ denotes the value of the outcome had the treatment been set to some value a via intervention and had the outcome been observed. Moving forward, we use $Y^{(a,1)}$ and $Y^{(a',1)}$ for brevity. Our target of inference is the average causal effect (ACE), i.e., the mean difference $\mathbb{E}[Y^{(a,1)} - Y^{(a',1)}]$.[1] In the next section we derive an identification formula for the ACE in terms of the observed data distribution $p(A, Y, R_Y, \mathbf{W})$ based on the backdoor adjustment formula from Pearl (1995) and the shadow variable method proposed by Miao et al. (2015).

5.2 Identification Strategy

In this section we demonstrate identification of the ACE assuming we are given a valid backdoor adjustment set and shadow variable adjustment set $\mathbf{Z} \subset \mathbf{W}$. Note that we use a strict subset relation as we will use at least one of the remaining pre-treatment covariates for verification of the identifying assumptions. We can also ignore the incentive variable I here because it does not play a role in identification, only in the verification of the assumptions surrounding $\mathbf{Z}$ later on.

In the following, we focus on identification of $\mathbb{E}[Y^{(a,1)}]$; identification of $\mathbb{E}[Y^{(a',1)}]$ follows similarly. We perform identification in two steps. In the first step, we assume we have access to the underlying counterfactual $Y^{(1)}$. When $\mathbf{Z}$ is a valid backdoor adjustment set relative to the treatment A and outcome $Y^{(1)}$ and satisfies (B1) and (B2), we may use the backdoor adjustment formula – equation 2.7 – to identify $\mathbb{E}[Y^{(a,1)}]$ as follows:

$$\mathbb{E}[Y^{(a,1)}] = \sum_{\mathbf{Z}} \mathbb{E}[Y^{(1)} \mid A = a, \mathbf{Z}] \times p(\mathbf{Z}) \tag{5.1}$$

However, due to missingness, we have access to only the marginal $p(A, Y, R_Y, \mathbf{Z})$, which does not include $Y^{(1)}$. Under self-censoring, the expression $\mathbb{E}[Y^{(1)} \mid A = a, \mathbf{Z}]$ is not equal to the observed expression $\mathbb{E}[Y \mid A = a, \mathbf{Z}, R_Y = 1]$ since $Y^{(1)} \not\perp\!\!\!\perp R_Y \mid A, \mathbf{Z}$ from assumption (M2), i.e., due to self-censoring.

[1] Similar notation for representing potential outcomes under missingness has been used in Nabi et al. (2022). This can also be expressed using do-notation as in Saadati and Tian (2019).

To identify (5.1), it is sufficient to identify the joint distribution $p(A, Y^{(1)}, \mathbf{Z})$. Following equation 2.5, we have the following identifying equation for the target law:

$$p(A, Y^{(1)}, \mathbf{Z}) = \frac{p(A, Y^{(1)}, \mathbf{Z}, R_Y = 1)}{p(R_Y = 1 \mid A, Y^{(1)}, \mathbf{Z})}. \tag{5.2}$$

The numerator is a function of observed data due to missing data consistency. Identification of this joint then reduces to identification of the propensity score $p(R_Y = 1 \mid A, Y^{(1)}, \mathbf{Z})$. If we are able to verify that A is a valid shadow variable, then the independence $A \perp\!\!\!\perp R_Y \mid Y^{(1)}, \mathbf{Z}$ will hold from (S2). Hence, $p(R_Y = 1 \mid A, Y^{(1)}, \mathbf{Z}) = p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$. To recover the propensity score, we use use the odds ratio factorization of the propensity score discussed in section 4.3. Hence, we recover the target law as specified in equation 5.2. In summary, we recover the propensity score for the missingness indicator R_Y, and this probability distribution allows us to apply backdoor adjustment and recover the ACE.

Consider Figure 5.2, which demonstrates how certain sets of pre-treatment covariates might be sufficient to adjust for confounding but not missingness or vice versa. In this DAG, the set $\{W_1\}$ satisfies (B1) and (B2)[2]. However, it does not satisfy (S2) when considering the treatment A as the shadow variable because of the open collider at $Y^{(1)}$ on the path $A \rightarrow Y^{(1)} \leftarrow W_2 \rightarrow R_Y$. On the other hand, the set $\{W_2\}$ satisfies (S2) when considering the treatment A as the shadow variable, but the backdoor path through W_1 remains open[3]. Only the set $\mathbf{Z} = \{W_1, W_2\}$ satisfies all of (B1), (B2), (S1), and (S2) simultaneously. We now present our main identification result that relies on a set of covariates that satisfy all these conditions simultaneously.

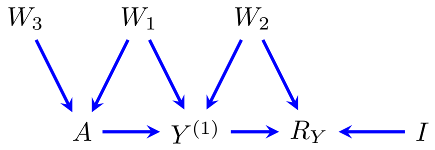


Figure 5.2: Graph demonstrating how different subsets of $\mathbf{W}$ satisfy the identifying assumptions (B1), (B2), (S1), and (S2).

Theorem 1. *Under structural assumptions (M1)-(M4) and completeness of $p(Y^{(1)} \mid R_Y = 1, A, \mathbf{Z})$, if $\mathbf{Z}$ satisfies (B1) and (B2) and A is a valid shadow variable satisfying (S1) and (S2) conditional on $\mathbf{Z}$, then the expected value of the counterfactual outcome $\mathbb{E}[Y^{(a,1)}]$ is identified from the observed data distribution $p(A, Y, \mathbf{Z}, R_Y)$ as follows:*

$$\mathbb{E}[Y^{(a,1)}] = \mathbb{E}\left[\frac{R_Y \times \mathbb{I}(A = a) \times Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A = a \mid \mathbf{Z})}\right] \tag{5.3}$$

Here, we give a proof for theorem 1. We first note that R_Y in the numerator ensures that we only use observed rows of data. Further, the propensity scores in the denominator are identified:

[2] The reader may refer to Chapter 2 for definitions of the backdoor adjustment conditions (B1) and (B2).

[3] The reader may refer to Chapter 4 for definitions of the shadow variable conditions (S1) and (S2).

$p(A = a \mid \mathbf{Z})$ only depends on observed quantities, and $p(R_Y = 1 \mid A, Y^{(1)}, \mathbf{Z}) = p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ is identified using (S1), (S2), the completeness condition, and the treatment variable A as the shadow variable. We now prove that the proposed identifying functional (5.3) is equal to the backdoor adjustment functional under the full data law.

Proof.

$$
\begin{aligned}
&\mathbb{E}\Bigg[\frac{R_Y \times \mathbb{I}(A = a) \times Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A = a \mid \mathbf{Z})}\Bigg] \\
&=^{(1)} \sum_{R_Y, Y^{(1)}, A, \mathbf{Z}, Y} p(R_Y, Y^{(1)}, A, \mathbf{Z}, Y) \times \frac{R_Y \times \mathbb{I}(A = a) \times Y}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A = a \mid \mathbf{Z})} \\
&=^{(2)} \sum_{Y^{(1)}, A, \mathbf{Z}} p(R_Y = 1, Y^{(1)}, A, \mathbf{Z}) \times \frac{\mathbb{I}(A = a) \times Y^{(1)}}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A \mid \mathbf{Z})} \\
&=^{(3)} \sum_{Y^{(1)}, A, \mathbf{Z}} p(R_Y = 1 \mid Y^{(1)}, A, \mathbf{Z}) p(Y^{(1)} \mid A, \mathbf{Z}) p(A \mid \mathbf{Z}) p(\mathbf{Z}) \times \frac{\mathbb{I}(A = a) \times Y^{(1)}}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A = a \mid \mathbf{Z})} \\
&=^{(4)} \sum_{Y^{(1)}, \mathbf{Z}} p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) p(Y^{(1)} \mid A = a, \mathbf{Z}) p(A = a \mid \mathbf{Z}) p(\mathbf{Z}) \times \frac{Y^{(1)}}{p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z}) \times p(A = a \mid \mathbf{Z})} \\
&=^{(5)} \sum_{Y^{(1)}, \mathbf{Z}} p(Y^{(1)} \mid A = a, \mathbf{Z}) \times p(\mathbf{Z}) \times Y^{(1)} \\
&=^{(6)} \sum_{\mathbf{Z}} \mathbb{E}[Y^{(1)} \mid A = a, \mathbf{Z}] \times p(\mathbf{Z}) =^{(7)} \mathbb{E}[Y^{(a,1)}].
\end{aligned}
$$

In (1) we apply the law of the unconscious statistician; in (2) we evaluate the sum over R_Y and use missing data consistency; in (3) we apply the chain rule of probability; in (4) we evaluate the sum over A and drop A from the conditional probability distribution $p(R_Y = 1 \mid Y^{(1)}, A, \mathbf{Z})$ by using conditional ignorability and condition (S2); (5) follows from cancellation of common terms in the numerator and denominator; (6) follows from the definition of expectation; the last step (7) follows from the fact that $\mathbf{Z}$ satisfies the backdoor conditions (B1) and (B2). □

The identification argument in theorem 1 and its proof relies on the absence of certain edges from the graph $\mathcal{G}$ in order to satisfy conditions (S1), (S2), (B1), and (B2). Figure 5.3 shows the edges that, if present, preclude identification. It is possible for the treatment A and a set $\mathbf{Z} \subset \mathbf{W}$ to satisfy the shadow variable conditions (S1) and (S2) whenever the red dashed edges are absent. Similarly, it is possible for $\mathbf{Z} \subset \mathbf{W}$ to satisfy the backdoor conditions (B1) and (B2) whenever the green dashed edge is absent. In essence, our method in the next section tests for the absence of these edges.

5.3 Tests for Identifying Conditions

As described in Chapter 3, the two-stage method described in Entner et al. (2013) searches for a $W \in \mathbf{W}$ and $\mathbf{Z} \subseteq \mathbf{W} \setminus \{W\}$ such that (i) $W \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ and (ii) $W \perp\!\!\!\perp Y \mid \mathbf{Z}, A$. They assume the same partial order of variables as in assumption (M4) except that the variables I and R_Y are not

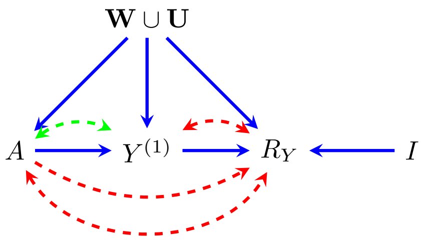


Figure 5.3: Red and green dashed edges impede identification.

present in the graph. In addition, Y is a fully observed variable. Entner et al. (2013) proved that, when conditions (i) and (ii) hold, $\mathbf{Z}$ is a valid backdoor adjustment set for the causal effect of A on Y.

For our method, we propose to test if

(C1) $I \not\perp\!\!\!\perp R_Y$

and then search for some $W \in \mathbf{W}$ and $\mathbf{Z} \subseteq \mathbf{W} \setminus \{W\}$ such that

(C2) $A \perp\!\!\!\perp I \mid Y^{(1)}, R_Y = 1, \mathbf{Z}$

(C3) $W \not\perp\!\!\!\perp R_Y \mid \mathbf{Z}$

(C4) $W \perp\!\!\!\perp R_Y \mid A, \mathbf{Z}$

Conditions (C1) and (C2) are used to confirm the absence of the red dashed edges in Figure 5.3 that would prevent us from using A as a valid shadow variable; meanwhile, conditions (C3) and (C4) are a modification of the tests from Entner et al. (2013) that indirectly confirm the validity of $\mathbf{Z}$ as a backdoor adjustment set for the potential outcome by using the missingness indicator to formulate the tests instead. This is formalized in the theorem below.

Theorem 2. *When (C1)-(C4) hold in a distribution $p(\mathbf{V})$ that is Markov and faithful w.r.t a causal DAG $\mathcal{G}$ satisfying assumptions (M1)-(M4), $\mathbf{Z}$ is a valid backdoor adjustment set for the causal effect of A on $Y^{(1)}$, and A is a valid shadow variable with $\mathbf{Z}$ as a shadow variable adjustment set.*

Proof. From Entner et al. (2013), if (C3) and (C4) hold, $\mathbf{Z}$ is a valid backdoor adjustment set for computing the causal effect of A on R_Y. From assumptions (M2) and (M4), all backdoor paths between A and $Y^{(1)}$, i.e., $A \leftarrow \cdots \rightarrow Y^{(1)}$, also extend into a backdoor path $A \leftarrow \cdots \rightarrow Y^{(1)} \rightarrow R_Y$ from A to R_Y. Thus, if $\mathbf{Z}$ blocks all backdoor paths between A and R_Y, it also blocks all backdoor paths between A and $Y^{(1)}$, making $\mathbf{Z}$ a valid adjustment set for the effect of A on $Y^{(1)}$.

Next, we prove that $\mathbf{Z}$ can be used for shadow variable adjustment using A as the shadow variable. Condition (S1) holds trivially as we have $A \rightarrow Y^{(1)}$ according to assumption (M2). We now prove that (S2) also holds: that $A \perp\!\!\!\perp R_Y \mid Y^{(1)}, \mathbf{Z}$. First, under assumption (M2), (C1) ensures that the randomized incentive I has a directed edge to R_Y and no other outgoing edges. In order for A to be a valid shadow variable, all paths between A and R_Y conditional on $Y^{(1)}$ and $\mathbf{Z}$ must be blocked. The following 4 cases cover all possible paths between A and R_Y given our assumptions.

1. Causal paths from A to R_Y. One possible causal path $A \rightarrow Y^{(1)} \rightarrow R_Y$ exists by assumption (M4), but it is blocked by conditioning on $Y^{(1)}$. The second possibility – $A \rightarrow R_Y$ – cannot exist because it would imply an open path between I and A conditional on $Y^{(1)}, R_Y = 1$, and $\mathbf{Z}$, contradicting (C2).

2. Paths of the form $A \rightarrow Y^{(1)} \leftarrow \ldots \rightarrow R_Y$. Any open paths of this form contradict (C2), as it implies the dependence $A \not\perp\!\!\!\perp I \mid Y^{(1)}, R_Y = 1, \mathbf{Z}$.

3. Backdoor paths of the form $A \leftarrow \ldots \rightarrow R_Y$ that do not contain $Y^{(1)}$ as a collider. All such paths are backdoor paths between A and R_Y, and, as previously noted, are blocked by $\mathbf{Z}$ based on conditions (C3) and (C4).

4. Backdoor paths containing $Y^{(1)}$ as a collider, i.e., $A \leftarrow \ldots \rightarrow Y^{(1)} \leftarrow \ldots \rightarrow R_Y$. All such paths are still blocked despite conditioning on $Y^{(1)}$ as $\mathbf{Z}$ blocks all backdoor paths between A and $Y^{(1)}$.

Therefore, $\mathbf{Z}$ is a valid backdoor adjustment set for the causal effect of A on $Y^{(1)}$, fulfilling (B1) and (B2), and it is also a valid shadow variable adjustment set for A to be a valid shadow variable, fulfilling (S1) and (S2) for A. □

We use Figure 5.4 to illustrate an example of how our tests proceed. Let us begin by only considering the solid blue edges. First, we confirm that (C1) holds. Next, none of the conditions (C2)-(C4) can be satisfied by using $\mathbf{Z} = \emptyset$ and $W = W_i$ for $i = 1, 2, 3$. When considering singleton adjustment sets, we get that, for $W = W_3$ and $\mathbf{Z} = \{W_1\}$, condition (C2) is satisfied. However, this set does not satisfy (C3) and (C4) because of the open collider at W_1 that introduces an open backdoor path between A and $Y^{(1)}$. Finally for adjustment sets of size 2, we get that, when $W = W_3$ and $\mathbf{Z} = \{W_1, W_2\}$, all conditions are satisfied, providing the correct conclusion that the effect is identified via (5.3) using $\mathbf{Z} = \{W_1, W_2\}$.

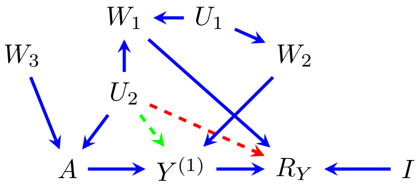


Figure 5.4: $W = W_3$, $\mathbf{Z} = \{W_1, W_2\}$

Next, consider the same DAG in Figure 5.4 with the green dashed edge being present. Because there is unmeasured confounding between the treatment and outcome variables, there is no possible $\mathbf{Z}$ that can be a valid backdoor adjustment set. Therefore, (C3) and (C4) can never be satisfied. Next, let us consider a DAG where the red dashed edge is present. The unmeasured confounding between A and R_Y violates (S2) when using A as a shadow variable, and it also creates a path between A and I that ensures that (C2) will never be satisfied. In both cases, our method correctly concludes that no adjustment set is possible if we are using the identification strategy presented in Theorem 1.

5.4 Limitations

There exist DAGs where our identification strategy works but where our method is not able to detect the existence of a valid backdoor and shadow variable adjustment set. Consider Figure 5.5 where W_3 is now a confounder between the treatment and outcome. Despite $\mathbf{Z} = \{W_1, W_2, W_3\}$ fulfilling all the critical assumptions set up in Theorem 1, our method will incorrectly conclude that there is no valid adjustment set because there exists no $W \in \mathbf{W}$ that can be used as an auxiliary variable to test (C3) and (C4). This same limitation exists, however, in the method proposed by Entner et al. (2013) in settings without missing data.

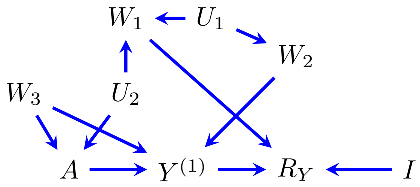


Figure 5.5: Setting where identification is possible using $\mathbf{Z} = \{W_1, W_2, W_3\}$, but there are no observed independence constraints that our method can use to verify this.

Furthermore, the incentive variable I is assumed to be a randomized variable in our method. In existing observational studies, it may not always be possible to identify such a variable, which would make applying our method more challenging. However, it may be possible to relax this condition such that I is conditionally randomized based on a set of pre-treatment covariates. This could potentially allow for more flexibility in the identification of a valid incentive variable, but we do not pursue this line of inquiry here.

Chapter 6

Practical Procedure

In this chapter, we give a practical procedure for verifying the identifying conditions described in section 5.3 that allow us to identify the average causal effect under a self-censoring outcome. The first step in the practical procedure is to verify conditions (C1) through (C4). To do so, we provide a simple algorithm that searches through all possible subsets of the set of confounders $\mathbf{W}$ to see if there exists some $W \in \mathbf{W}$ and set $\mathbf{Z} \subset \mathbf{W}$ satisfying all of the identifying conditions. Next, we describe how we may recover the propensity scores $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$ and $p(A = a \mid \mathbf{Z})$ from the observed data distribution. As a part of the practical estimation procedure, we use specific parameterizations of the propensity scores and the odds ratios. However, our identification results are non-parametric, and the reader may use any model that they prefer to estimate the probability distributions.

6.1 Independence Tests

The first step in the practical estimation procedure is to verify (C1). If this test fails, then we are unable to empirically verify the validity of candidate adjustment sets using the given incentive I, and the practical procedure ends. If the test for (C1) succeeds, Algorithm 6.1 then searches over all possible assignments for W and $\mathbf{Z}$ to see if there is a combination of assignments that fulfills (C2)-(C4). The search in Algorithm 6.1 proceeds in a style similar to the PC algorithm (Spirtes et al., 2000) for causal discovery (and the order of tests described for Figure 5.4) where conditioning sets of size 0 are tested first, followed by sets of size 1, and so on. Such a search is exponential in time complexity, but searching for valid adjustment sets is NP-hard in general.[1]

We use likelihood ratio tests for all tests in Algorithm 6.1; however, suitable non-parametric tests such as kernel conditional independence tests can also be applied (Zhang et al., 2011). If Algorithm 6.1 returns a set $\mathbf{Z}$, then we have found a set satisfying (B1), (B2), (S1), and (S2), and we may use A as a valid shadow variable and $\mathbf{Z}$ as a valid backdoor adjustment set. Otherwise, we conclude that no adjustment set could be validated using our tests. Note that, even if Algorithm 6.1

[1] If desired, however, one may also perform a search for candidate adjustment sets using only a subset of all possible subsets of $\mathbf{W}$.

is unable to find an adjustment set, this does not eliminate the possibility that the causal effect is identified through other strategies. Interestingly, whenever Algorithm 6.1 does not find a valid adjustment set, it may also still be possible to identify the causal effect using Theorem 1. Whether or not Algorithm 6.1 is able to find a valid adjustment set is contingent on the existence of a valid auxiliary variable $W \in \mathbf{W}$ to test for conditions (C3) and (C4).

Algorithm 6.1 for finding a valid adjustment set $\mathbf{Z}$.

```
for each $W \in \mathbf{W}$ do
    $\mathbf{Z}_f \leftarrow \mathbf{W} \setminus \{W\}$
    for $i$ from 0 to $|\mathbf{Z}_f|$ do
        $\mathbb{Z}_s \leftarrow$ all possible subsets of $\mathbf{Z}_f$ with size $i$
        for each $\mathbf{Z} \in \mathbb{Z}_s$ do
            if (C2)-(C4) are true using $W$ and $\mathbf{Z}$ then
                return $\mathbf{Z}$
            end if
        end for
    end for
end for
return “no adjustment set found”
```

The complexity of Algorithm 6.1 is exponential because it requires enumerating through the power set of $\mathbf{W}$. Hence, it definitely will not win any awards for efficiency. However, in practice, the complexity may not necessarily pose any significant challenges to applying this method. According to a meta study that evaluated studies in applied health research using DAGs, most studies in such a context only use DAGs with an average of twelve variables (Tennant et al., 2021). DAGs of this size and those of similar sizes should pose no computational issues for our method. For high-dimensional settings, researchers typically rely on some sparsity assumptions, e.g., limiting the maximum size of the conditioning set (as in causal discovery applications) or, alternatively, finding a low-dimensional representation of the high-dimensional confounders that is sufficient for adjustment (Ma et al., 2019). It is possible to apply such methods in conjunction with ours to deal with self-censoring in the high-dimensional case.

6.2 Causal Effect Estimation

If we find a valid set $\mathbf{Z}$ from Algorithm 6.1, we then start by estimating the distribution $p(A = a \mid \mathbf{Z})$ using the observed data distributions because all of the variables in this distribution are fully observed. As the treatment variable is typically binary, we use a linear logistic regression model where all the variables in the set $\mathbf{Z}$ are predictors of $A = a$. More flexible models, such as generalized additive models or random forests, are also possible depending on the sample size.

Next, we estimate the propensity score $p(R_Y = 1 \mid Y^{(1)}, \mathbf{Z})$. Let the cardinality of $\mathbf{Z}$ be $|\mathbf{Z}| = k$. As discussed in Chapter 4, we use the odds ratio factorization of the propensity score defined in (4.1), and we use the parameterizations $\pi_0(\mathbf{Z}) = \text{expit}(\beta_1 Z_1 + \beta_2 Z_2 \dots \beta_k Z_k)$ and $\eta(Y^{(1)}, \mathbf{Z}) = \exp(\gamma Y^{(1)})$. In total, we need to estimate $k + 1$ parameters to recover the propensity score of R_Y; hence, we need

$k+1$ zero-mean estimating equations of the form $\mathbb{E}[(\frac{R_Y}{p(R_Y=1|Y^{(1)},\mathbf{Z};\boldsymbol{\beta},\gamma)}-1)\times h(A,\mathbf{Z})]$. A simple choice for the first k equations is to use $h(A,\mathbf{Z})=Z_i$ for each $Z_i\in\mathbf{Z}$. For the final equation, we use $h(A,\mathbf{Z})=\overline{A}$, where $\overline{A}$ is the mean of the variable A. We use A here because it is the variable we are using as the shadow variable. This gives us the following system of equations:

$$\mathbb{E}\left[\left(\frac{R_Y}{p(R_Y=1|Y^{(1)},\mathbf{Z};\boldsymbol{\beta},\gamma)}-1\right)\begin{bmatrix}Z_1\\Z_2\\\dots\\Z_k\\\overline{A}\end{bmatrix}\right]=0. \tag{6.1}$$

Because the expression $\frac{R_Y}{p(R_Y=1|Y^{(1)},\mathbf{Z};\boldsymbol{\beta},\gamma)}$ is 0 for each individual in the dataset where $R_Y=0$, the system of estimating equations only uses observed rows of $Y^{(1)}$. To find estimates for each of the $k+1$ parameters, we may use any root finding algorithm. Different models for the propensity score $\pi_0(\mathbf{Z})$ and the odds ratio $\eta(Y^{(1)},\mathbf{Z})$ are also possible in theory as long as there are $k+1$ or fewer parameters that the system of equations (6.1) needs to recover.

Using the models for $p(A=a\mid\mathbf{Z})$ and $p(R_Y=1\mid Y^{(1)},\mathbf{Z})$, we get estimates of both propensity scores for each row of data. Since inverse probability weighting estimators can be unstable with small and large weights, we *clip* these propensity scores to be values between $p_{low}=0.01$ and $p_{high}=0.99$ (Hernán and Robins, 2010; Crump et al., 2009). To clip propensity scores, we round propensity scores lower than 0.01 up to 0.01 and round propensity scores greater than 0.99 down to 0.99. Finally, we estimate the causal effect by taking the empirical average for the expectation shown in the identifying functional in (5.3) for treatment values a and a'.

A summary of our procedure is as follows: (i) Test (C1), if it holds, proceed to (ii), else, terminate the algorithm. (ii) Test (C2)-(C4) using Algorithm 6.1, if it returns a set $\mathbf{Z}$, proceed to (iii), else terminate the algorithm. (iii) Estimate the propensity scores for A and R_Y and plug them into an inverse probability weighted estimator for the ACE based on the identifying functional in (5.3).

6.3 Python Implementation of the Practical Procedure

We also provide a Python implementation of the practical procedure, which may be found at this link: `https://github.com/jacobmchen/mnar-recoverability`. The two files containing the implementations described in sections 6.1 and 6.2 are `shadow_covariate_selection.py` and `shadow_recovery.py`, respectively.

`shadow_covariate_selection.py` contains code defining a class `ShadowCovariateSelection` that implements Algorithm 6.1. At a high level, the class defines a constructor that takes as input a pandas dataframe and the variables A (the treatment variable), Y (the observed outcome), R_Y (the missingness indicator for the outcome), and I (the incentive variable). It also reads a parameter for α for the independence tests. If the user does not specify a value for α, a value of 0.05 is used. To make independence tests between different variables in the dataset, we use a weighted likelihood ratio test implemented in `ratio_test.py`. Finally, we have a function in the class that implements

Algorithm 6.1 where the user can specify the maximum size of the adjustment set $\mathbf{Z}$. If no maximum size is specified, the set $\mathbf{Z}$ will be allowed to grow to the size of the set $\mathbf{W}$ minus one.

`shadow_recovery.py` contains code defining a class `ShadowRecovery` that implements the procedure described in section 6.2 that uses a system of equations to recover the propensity score. The class defines a constructor that takes as input a pandas dataframe and the variables A (the treatment), Y (the observed outcome), R_Y (the missingness indicator), and $\mathbf{Z}$ (an adjustment set satisfying (C1)-(C4) identified with Algorithm 6.1). As discussed above, we use a linear logistic regression to estimate the propensity score for $A = a$. To find the propensity score for $R_Y = 1$, we call a method that finds the $k+1$ roots for the estimating equations in (6.1) by using the root optimizing function provided in the scipy library. All of the pieces required for estimating the average causal effect are now recovered, so we calculate and return it. We also implement a clipping function that takes as input a numpy array of propensity score estimates and clips them using $p_{low} = 0.01$ and $p_{high} = 0.99$. Finally, we provide a method for calculating the confidence intervals the estimate of the ACE that bootstraps the original dataset 200 times and estimates the average causal effect for each bootstrapped dataset. It then finds the 2.5th and 97.5th percentile of the distribution of average causal effects to return as the confidence interval.

Chapter 7

Simulation Study

In this chapter, we describe a simulation study that evaluates the effectiveness of the practical procedure given in Chapter 6 using a synthetic dataset. While it would be ideal to use a real-world dataset, it is difficult to find real-world data with an accompanying randomized controlled trial (RCT) that can be used as ground-truth. In fact, most datasets with ground-truth in causal inference turn out to be synthetic or semi-synthetic. We start by describing the data generating process that we will use to create the synthetic dataset to test our methods on. Then, we describe the methods that we run and the empirical results. Overall, our simulation study shows promise in our method's ability to accurately identify adjustment sets and make unbiased estimates for the average causal effect.

7.1 Data Generating Process

For our simulations, we generate data according to the graph shown in Figure 7.1 and modifications of it that violate the shadow variable or backdoor conditions. We generate the pre-treatment covariates $\mathbf{W}$ from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} =$

$$\begin{bmatrix} 1.2 & 0 & 0 & 0 \\ 0 & 1 & 0.4 & 0.4 \\ 0 & 0.4 & 1 & 0.3 \\ 0 & 0.4 & 0.3 & 1 \end{bmatrix}.$$

The above data generating process is equivalent to a structural equation model with correlated errors due to unmeasured confounders between the pairs (W_2, W_3), (W_2, W_4), and (W_3, W_4). We generate $A, Y^{(1)}, I$, and R_Y according to structural equation models following edges in Figure 7.1. Note that we also clip all probabilities to be between the ranges of 0.01 and 0.99. This specifically applies to the variables $A, Y^{(1)}$, and R_Y. We generate A as a binary variable with the following probabilities:

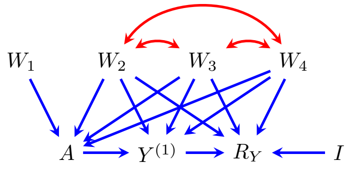


Figure 7.1: Graph used in our simulations.

$$p(A = 1 \mid W_1, W_2, W_3, W_4) = \text{expit}(0.52 + 2 * W_1 + 2 * W_2 + 2 * W_3 + 2 * W_4)$$
$$p(A = 0 \mid W_1, W_2, W_3, W_4) = 1 - p(A = 1 \mid W_1, W_2, W_3, W_4)$$

Next, $Y^{(1)}$ is generated similarly with the following probabilities:

$$p(Y^{(1)} = 1 \mid A, W_2, W_3, W_4) = \text{expit}(3 * A + 2 * W_2 + 2 * W_3 + 2 * W_4)$$
$$p(Y^{(1)} = 0 \mid A, W_2, W_3, W_4) = 1 - p(Y^{(1)} = 1 \mid A, W_2, W_3, W_4)$$

The variable I is simply generated as a random normal variable with mean 0 and variance 2, i.e. $I \sim \mathcal{N}(0, 2)$.

We use the odds ratio parameterization to generate R_Y with the following two probabilities. We first specify $p(R_Y = 1 \mid Y^{(1)} = 0, \mathbf{W} \setminus \{W_1\}, I)$, which represents the probability of $R_Y = 1$ when $Y^{(1)}$ is at our chosen reference value of 0. Recall that the odds ratio $\eta(Y^{(1)}, \mathbf{Z})$ will have a value of 1 whenever $Y^{(1)}$ is at its reference value. We then use that probability to generate $p(R_Y = 1 \mid Y^{(1)}, \mathbf{W} \setminus \{W_1\}, I)$ at all values of $Y^{(1)}$. To save space in the equations below, we use the symbol π_0 to denote $p(R_Y = 1 \mid Y^{(1)} = 0, \mathbf{W} \setminus \{W_1\}, I)$.

$$p(R_Y = 1 \mid Y^{(1)} = 0, \mathbf{W} \setminus \{W_1\}, I) = \text{expit}(W_2 + W_3 + W_4 + 0.5 * I) = \pi_0$$

$$p(R_Y = 1 \mid Y^{(1)}, \mathbf{W} \setminus \{W_1\}, I) = \frac{\pi_0}{\pi_0 + \exp(-1.5 * Y^{(1)}) \times (1 - \pi_0)}.$$

7.2 Simulations and Results

Our first set of experiments focuses on evaluating the effectiveness of Algorithm 6.1 for finding a valid adjustment set when such a set exists and for correctly identifying that no valid adjustment set exists when no such set exists. Before describing the experiments, we define key terms for evaluating accuracy. A *true positive* (TP) occurs when an adjustment set exists and the algorithm identifies the correct set of covariates. If the algorithm does not find an adjustment set or returns an incorrect one when an adjustment set exists, then this is a *false negative* (FN). A *true negative* (TN) occurs when no possible adjustment set exists and the algorithm correctly finds no adjustment set. If the algorithm detects an adjustment set when no such set exists, this is considered a *false positive* (FP). *Sensitivity* is defined as $\frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}}$, and *specificity* is defined as $\frac{\#\text{TN}}{\#\text{TN}+\#\text{FP}}$.

The experiment proceeds as follows. We first run 200 trials with data generated according to Figure 7.1 where we expect Algorithm 6.1 to return $\mathbf{Z} = \{W_2, W_3, W_4\}$ as that is the correct adjustment set. We then run 200 trials where we expect the algorithm to return “no adjustment set found” because we use data generated either according to a DAG where we add the edge $A \rightarrow R_Y$ to Figure 7.1 with probability 0.5 or where we treat W_4 as a latent variable with probability 0.5. The size of the dataset for each of these trials ranges from 500 to 10,000, and we use significance levels of $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.1$ to conclude dependence between two variables for our independence tests. Note that the effective sample size for some tests is roughly 60% of the full sample size due to the missingness of the outcome. Table 7.1 summarizes the results for $\alpha = 0.01$, Table 7.2 summarizes the results for $\alpha = 0.05$, and Table 7.3 summarizes the results for $\alpha = 0.1$.

Sample Size	**Sensitivity**	**Specificity**
500	0.0	0.394
2500	0.269	0.383
5000	0.690	0.717
10000	0.828	0.952

Table 7.1: Results of covariate search experiment when using a significance level of $\alpha = 0.01$

Sample Size	**Sensitivity**	**Specificity**
500	0.011	0.350
2500	0.577	0.556
5000	0.812	0.818
10000	0.930	0.925

Table 7.2: Results of covariate search experiment when using a significance level of $\alpha = 0.05$.

Sample Size	**Sensitivity**	**Specificity**
500	0.022	0.363
2500	0.671	0.630
5000	0.837	0.770
10000	0.949	0.857

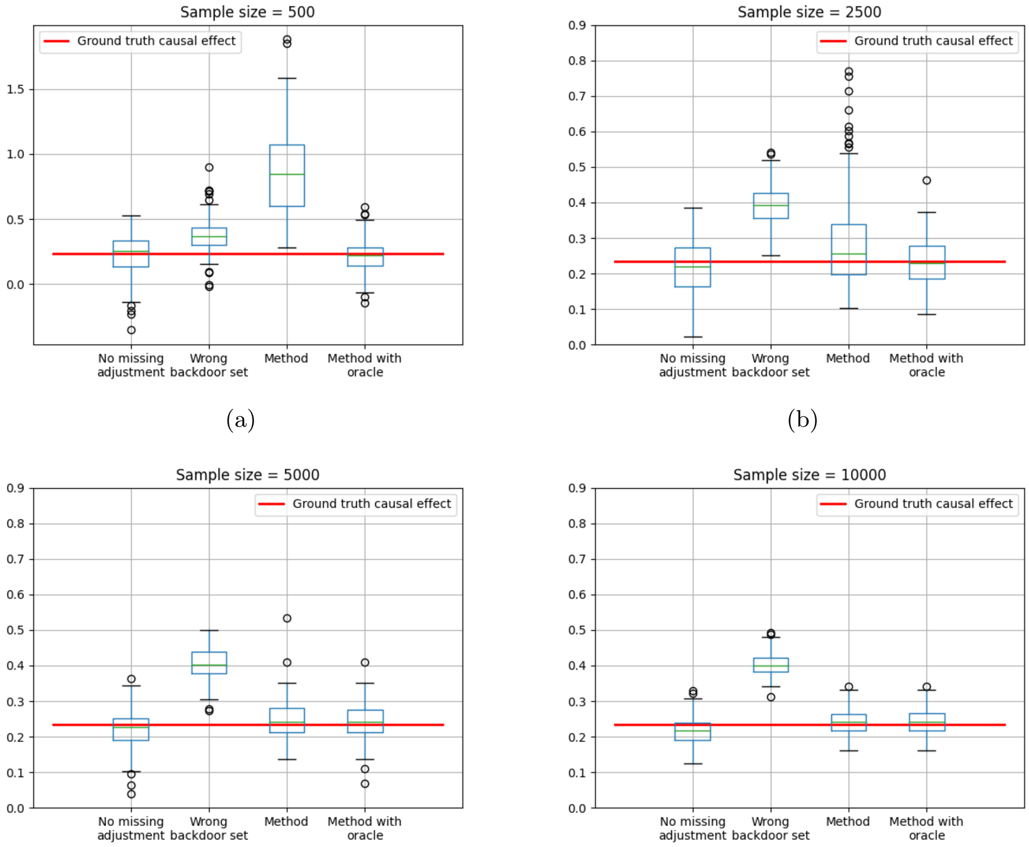
Table 7.3: Tables showing the accuracy of tests for different sample sizes and $\alpha = 0.1$.

As α increases, the accuracy of the tests for correctly predicting an adjustment set when one is possible – achieving a true positive – increases. On the other hand, the accuracy of the tests for correctly identifying that there is no possible adjustment set when no such set exists – achieving a true negative – decreases as p-value increases. For all p-values, the accuracy of the tests in general increases as the sample size increases. A happy medium for achieving a reasonable sensitivity and specificity seems to be using $\alpha = 0.05$.

The second set of experiments uses data from Figure 7.1 to evaluate the effectiveness of our method for downstream causal effect estimation. We compare these estimates to the bias introduced by either failing to adjust for missing data but using the correct backdoor adjustment set or by failing to use a valid backdoor adjustment set but correctly adjusting for missing data. To estimate the

ACE without adjusting for missing data, we restrict the dataset to only rows of data where the outcome is observed and use a standard inverse probability weighted estimator. To estimate the ACE while using an invalid backdoor adjustment set, we use a subset of the correct adjustment set – $\{W_2, W_3\}$ – after correctly adjusting for missing data. We further compare these two estimates with one obtained by running our full procedure described in Chapter 6 with $\alpha = 0.05$. To emphasize the importance of reliable conditional independence tests, we also generate estimates obtained using an *independence oracle* in Algorithm 6.1. An independence oracle always gives the true independence or dependence statement between two variables regardless of the observed dataset. We run 200 trials each for the sample sizes 500, 2,500, 5,000, and 10,000.

The results of the experiments for all 4 sample sizes are shown in Figure 7.2. The red line in the figures shows the ground truth causal effect. At sample size 10,000, the search algorithm correctly identifies the adjustment set for 93% of the trials, so the estimates from the full practical estimation procedure are nearly identical to the estimates from using the independence oracle. As the sample size increases, estimation from failing to adjust for missing data and confounding bias converge to biased values for the ACE. The method that ignores missingness produces reasonable estimates at low sample sizes, but it shows asymptotic convergence to a biased estimate. As expected, estimates obtained by using our full pipeline converge to the ground truth causal effect as sample size increases. The Python implementation linked in Chapter 6 also contains the implementation for the data generating process and experiments in the file `experiment.py`.



(a) (b)

(c) (d)

Figure 7.2: Simulation results for estimating the average causal effect using different methods and sample sizes at $\alpha = 0.05$.

Chapter 8

Conclusion and Future Work

In this thesis, we discussed methods for causal effect estimation with a self-censoring outcome and when the underlying causal structure of the graph is unknown. We proved that when a set $\mathbf{Z}$ satisfies both the backdoor adjustment set and the shadow variable adjustment set criteria, then identification of the average causal effect is possible through an inverse probability weighting functional. We further described a series of tests that can be used to empirically identify such a valid set $\mathbf{Z}$ using only observed data. We presented a simple search algorithm that uses our tests as a subroutine. We also gave a practical estimation strategy for the average causal effect using the aforementioned inverse probability weighting estimator if a valid set $\mathbf{Z}$ is found by using the odds ratio factorization of the propensity score. Finally, we showed through experiments based on synthetic data the accuracy of our search algorithm and estimation procedure.

Although recent work has made significant progress in identifying full data laws and structural learning algorithms under MNAR missingness mechanisms, methods for these tasks under self-censoring is still relatively underdeveloped (Bhattacharya et al., 2019; Nabi et al., 2020; Malinsky et al., 2021; Tu et al., 2019; Liu and Constantinou, 2022). In fact, all of the previously cited works provide results that are only theoretically sound in the absence of self-censoring from the data generating process. Despite the relatively little attention that the self-censoring problem has received thus far in the literature, self-censoring is still expected to occur in observational studies, specifically those that ask socially sensitive questions to respondents. In this thesis, we make incremental progress on estimating average causal effects and identifying a valid backdoor adjustment set while allowing for the possibility of self-censoring on the outcome. To the best of our knowledge, no previous work in the literature thus far has proposed a full causal inference pipeline with a covariate search and estimation strategy in self-censoring settings. In addition to our theoretical contributions, we also completely implement the practical estimation strategy described in Chapter 6 in Python for use by researchers. We make a substantial contribution to the fields of causal inference and missing data.

As always, some open questions remain. In our method, we use a randomized incentive variable to help us test if the treatment satisfies the shadow variable assumptions (S1) and (S2). However, such a randomized incentive variable may not always exist in observational studies. To what extent can we relax the assumption that the incentive variable be randomized? Such an advance would allow our method to be applied to a wider range of existing observational studies.

Furthermore, the inverse probability weighted estimators that we propose in our identifying equation are subject to instability when the propensity scores are close to 0 or close to 1. In such scenarios, we divide by very small and large numbers, which can pull the empirical average away from the ground truth. To address for this, we apply clipping in our synthetic data generating process and estimation strategy by setting minimum and maximum values for the propensity scores. Can we design semiparametric estimation strategies for the tests and final estimation piece that exhibit desirable statistical properties, such as robustness to model mispecification and lower asymptotic variance? In addition, can we find doubly robust estimation strategies for the propensity score of response and odds ratio as discussed in Tchetgen Tchetgen et al. (2010)? Such advances would make our practical estimation strategy more reliable under smaller sample sizes.

Next, the causal effect may also be identified by other methods such as the frontdoor criterion when the backdoor criterion does not apply. However, research in testing for the frontdoor criterion is relatively recent, and testing for the frontdoor criterion can be considerably more complex (Bhattacharya and Nabi, 2022). Using the frontdoor criterion to recover from outcome-dependent self-censoring would also involve verifying the frontdoor assumptions and the shadow variable assumptions simultaneously under missing data. Furthermore, if the outcome exhibits self-censoring, then it is also plausible that the mediator variable also exhibits self-censoring. These considerations will present challenges in both the tests for identifying assumptions and downstream causal effect estimation. This is an interesting line of inquiry, though, and identifying different ways of applying known causal effect estimation strategies to self-censoring settings will allow researchers working with observational data more flexibility.

Another interesting line of inquiry based on our work is its applicability to structural learning algorithms. In particular, work on shadow variables from Miao et al. (2015) may be helpful in extending work on *Y structures* (Mani et al., 2012). Y structures are DAGs with four variables such that no other DAG with four variables imply the exact same conditional independencies as the Y structure. Given this property, a structural learning algorithm will always asymptotically recover the Y structure from observed data if it is the true data generating process. In addition, a Y structure is able to rule out unmeasured confounding between its last two variables in topological order. Figure 8.1a shows a Y structure. At a high level, structural learning algorithms are able to rule out unmeasured confounding between A and Y because adding a bidirected edge between these two variables creates two colliders – $W_1 \rightarrow A \leftrightarrow Y$ and $W_2 \rightarrow A \leftrightarrow Y$. The addition of these two colliders create independencies not present in the observed dataset; therefore, the bidirected edge cannot exist. In practice, as long as we can identify two causes of A that are independent of each other, any structural learning algorithm will be able to detect a causal relationship between A and Y while ruling out the possibility of unmeasured confounding. Figure 8.1b shows a partial Y structure, which is the same as a Y structure except we also append an edge from W_1 to Y. All properties discussed for Y structures above also apply to partial Y structures.

In this thesis, we consider recovering the average causal effect from a self-censoring outcome. Y structures may provide a framework for working with situations where both the treatment and

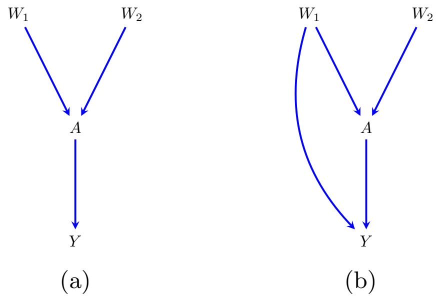


Figure 8.1: A Y structure and a partial Y structure, repsectively.

outcome are self-censoring. Intuitively, it may be possible to use one of the variables W_1 or W_2 as a shadow variable. Such an advance would allow us to rule out unmeasured confounding between a self-censoring treatment and a self-censoring outcome. It would also lay the foundation for potential structural learning methods that allow for self-censoring missingness mechanisms in general. A major obstacle for advancement in this work, however, is how we might be able to test for the validity of potential shadow variables.

In this chapter, we have considered the contributions of this thesis. We have also discussed potential avenues for future work on both the method we have proposed as well as new methods in different areas of causal inference. Missing data, especially MNAR missingness and self-censoring, is inevitable in observational studies. It is important that we do not shy away from working with even the hardest forms of missingness even though the challenges imposed by such missingness mechanisms may sometimes seem insurmountable. As researchers in causal inference and missing data, our work paves the way for our colleagues who work more closely with data to report unbiased results that will allow for informed decision-making and engender positive changes in our lives.

Bibliography

Bhattacharya, R. and Nabi, R. (2022). On testability of the front-door model via verma constraints. In *Uncertainty in Artificial Intelligence*, pages 202–212. PMLR.

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. M. (2019). Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

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

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. (2021). An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471*.

d'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15.

Entner, D., Hoyer, P., and Spirtes, P. (2013). Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial Intelligence and Statistics*, pages 256–264. PMLR.

Hernán, M. A. and Robins, J. M. (2010). *Causal Inference: What If*. CRC Boca Raton, FL.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Jerzy, N. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science*, 5:463–472.

Liu, Y. and Constantinou, A. C. (2022). Greedy structure learning from data that contain systematic missing values. *Machine Learning*, 111(10):3867–3896.

Ma, S., Zhu, L., Zhang, Z., Tsai, C.-L., and Carroll, R. J. (2019). A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Annals of statistics*, 47(3):1505.

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. (2021). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9.

Mani, S., Spirtes, P. L., and Cooper, G. F. (2012). A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*.

Miao, W., Liu, L., Tchetgen, E. T., and Geng, Z. (2015). Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in neural information processing systems*, 26.

Mohan, K., Thoemmes, F., and Pearl, J. (2018). Estimation with incomplete data: The linear case. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7153–7163. PMLR.

Nabi, R., Bhattacharya, R., Shpitser, I., and Robins, J. (2022). Causal and counterfactual views of missing data models. *arXiv preprint arXiv:2210.05558*.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

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

Pearl, J. (2009). *Causality*. Cambridge University Press.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2023). Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Saadati, M. and Tian, J. (2019). Adjustment criteria for recovering causal effects from missing data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 561–577. Springer.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2012). On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*.

Spirtes, P. L., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Sportisse, A., Boyer, C., and Josse, J. (2020). Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077.

Tchetgen, E. J. T., Wang, L., and Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Tchetgen Tchetgen, E. J. and Wirth, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131.

Tennant, P. W., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., et al. (2021). Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2):620–632.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. (2019). Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR.

Turner, C. F., Al-Tayyib, A., Rogers, S. M., Eggleston, E., Villarroel, M. A., Roman, A. M., Chromy, J. R., and Cooley, P. C. (2009). Improving epidemiological surveys of sexual behaviour conducted by telephone. *International Journal of Epidemiology*, 38(4):1118–1127.

Verma, T. S. and Pearl, J. (2022). Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236.

Yang, S., Wang, L., and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888.

Yun Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813.