Combining Optimal Adjustment Set Selection and Post Selection Inference in Unknown Causal Graphs

by

Elijah Tamarchenko

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 22, 2023

Contents

List of Figures

List of Tables

Abstract

Covariate selection for backdoor adjustment is often made difficult due to unmeasured confounding; some adjustment sets can lead to bias due to exclusion of relevant confounders, others may be unbiased but statistically inefficient. Rotnitzky et al (2019) propose a graphical criterion for identifying the optimal adjustment set – an unbiased set with minimal asymptotic variance – in settings where the structure of the causal system is known exactly and there are no unobserved common causes. However, in most practical settings, the full causal structure is unknown and likely to exhibit unmeasured confounding. In this case, Entner et al (2013) propose a procedure for identifying an unbiased adjustment set. However, it performs an exponential number of conditional independence tests, which is infeasible in high dimensional settings, and does not consider minimizing variance. We propose a parametric continuous optimization procedure, which performs both covariate selection and effect estimation in a single step. We prove that this procedure identifies the optimal adjustment set in the absence of unmeasured confounders. We further show that under mild assumptions involving an auxiliary variable, if the continuous optimization procedure excludes this auxiliary variable from the covariate selection process, then the effect estimate is provably unbiased even in settings with unmeasured confounders. Further, the procedure often leads to a practical reduction in variance as shown via simulations.

Acknowledgments

There are so many people that I would love to thank, but in attempting to keep my acknowledgements shorter than the rest of my thesis, I must select a smaller subset of the large set of those I would like to thank. I would first of of all like to thank my advisor, Rohit, without whose incredible help and support I would not have able to complete this thesis. Your passion for your students and desire to see them succeed has truly been a 'significant predictor' of my own success in this endeavor. I appreciate immensely all the hours we spent together ruminating over code bugs and unexpected results, and will recall those moments with great joy, even if joy is not the emotion I may have been feeling at the time. I hope you know how much of an impact you have made on my education and my life more broadly, and would love to continue working with and being friends with you into the future. I would also like to thank Katie for agreeing to be my second reader, and for providing an extremely necessary and vital outside perspective to the project. Your curiosity and passion is truly infectious, and I hope you know how much all of your students appreciate the enthusiasm with which you approach every task.

I would also like to thank my family and friends. To my parents, you have managed to instill in me a love for learning and a curiosity for the things around me that will truly last a lifetime, and I cannot imagine being raised by better parents than you. You have supported unquestioningly throughout all of my endeavors, and I would never have reached this point in my life without your constant encouragement. I would like to thank my grandparents for providing me with an example of what it means to be hardworking, and my siblings for providing the example for what it means to chase your dreams and do what you love. I want to thank my friends at Williams, who have commiserated with me over the course of this year and the last four years more generally, and without which my college experience would have been abysmal. I would also like to thank my other professors at Williams who have supported me and taught me almost everything I know, and most of what I still do not.

Finally, to make sure that no-one is left out of my acknowledgements, I would like to thank anyone and everyone who has ever shown me a smile or engaged me in conversation, since life is nothing without the little connections we make with the people around us.

Glossary of Important Terms

$\text{nde}_{\mathcal{G}}(X) \rightarrow$ Non-descendants in graph $\mathcal{G}$

$\text{de}_{\mathcal{G}}(X) \rightarrow$ Descendants of X in graph G

$\text{pa}_{\mathcal{G}}(X) \rightarrow$ Direct parents in G

$\text{mb}_{\mathcal{G}}(X) \rightarrow$ Markov Blanket of X, which is all the variables that lie on a bidirected path with X, and all their parents, minus X

$\text{cn}_{\mathcal{G}}(A, Y) \rightarrow$ The nodes that lie on the direct causal path between A and Y in G

$\text{forb}_{\mathcal{G}}(A, Y) \rightarrow \text{de}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(A, Y))$

$\text{diag}(w_j) \rightarrow$ The diagonal matrix formed using w_j

$C^X_{opt} \rightarrow$ The subset of good predictors for X identified using a consistent variable selection method $\mathcal{M}$

DAG $\rightarrow$ Directed Acyclic Graph

ADMG $\rightarrow$ Acyclic Directed Mixed Graph

$\beta_{ZY.X} \rightarrow$ The coefficient for Z in a regression on Y, given X

Chapter 1

Introduction

Traditional statistical analysis, as taught in most college courses, are focused on identifying and estimating associational relationships between variables. Countless techniques and methodologies have been developed to model increasingly complex relationships between variables, and gleam more insights from data. These methods all come with the caveat however, that causal relationships - relationships of the type 'changes in x *cause* changes in y' - cannot be identified without the use of data from randomized controlled experiments. All statistics students are familiar with the oft-cited mantra "Correlation is not Causation". The advent of causal inference techniques allows analysts to attempt to answer these causal questions without the need for data from experiments, by leveraging previous domain knowledge and the statistical relationships between variables.

To demonstrate this difference in inferential power, let us consider a hypothetical school which is trying to decide how best to improve the achievement of their students (graph of variables shown in Figure 1.1). Their primary hypothesis is that an increase in school funding would help their students, though they require statistical proof of this hypothesis. We can collect data on the average funding per student that schools in the same state have available every year, and we can use standardized test scores as a proxy for student achievement, since quantifying student achievement is a complex issue outside the scope of this thesis. Using classical statistical techniques, we can compute the correlational relationship between school funding and student achievement. This can be done using a regression, a difference of means, or any other available method. A positive linear relationship would imply that as student funding increases, on average we would also expect student achievement to increase. This is however a purely correlational relationship, and does not allow us to make conclusions about the cause of this relationship. From a policy making perspective, this positive correlation may not be enough to warrant any significant changes to the school funding structure, as we would never be sure of their true impact.

Causal inference techniques allow us to take our analysis a step further, and ask the question "Would an increase in school funding *cause* an increase in student achievement?". This amounts to asking a question about a *counterfactual* situation caused by a potential *intervention*. The simplest way of quantifying this is by computing the difference between the average student score in the school at their original funding level a and the average student score of the school after a policy

intervention that increases student funding to level a' (Binarizing the treatment limits the amount of inferential power we have for the question, but we will do so for simplicity's sake). This is formalized in equation 1.1.

$$\mathbb{E}\big[Score(Funding = a') - Score(Funding = a)\big] \tag{1.1}$$

The notation $Score(Funding = a')$ and $Score(Funding = a)$ represent two possible potential outcomes at the two different funding levels. This potential outcome defines $Score$ as a function of $Funding$, and allows us to answer our causal question and compute the effect of school funding on student scores, or the *Average Causal Effect* (ACE). Computing the ACE from equation 1.1 is non-trivial, since it is impossible for us to observe the outcome of two different funding levels at the same time. Simply computing the average of all scores at schools with funding level a and funding level a', and then taking the difference would produce a possibly misleading result, since we know that there are more factors that influence student achievement than merely school funding. These factors are potential *confounding variables* which can bias our result, and thus require inclusion in our model. The choice of which confounding variables to include in our model and which not to is of central focus for this thesis. Traditional causal inference methods usually employ graphical models in order to encode the causal relationships between the variable. An edge in the graphical model represents a potential direct causal link from one node to the next relative to the other nodes in the graph, and the absence of an edge represents the absence of a direct causal relationship between those variable.

As an example, we could encode a simple graph to represent our school funding problem in Figure 1.1[1]. Thus, in our graph, the edge from Parent Income to School Funding represents the fact that Parent Income is a potential cause of School Funding. In fact, we know that in Massachusetts the formula defined by the state for calculating school funding includes the median income of the school district. Conversely, the absence of the edge from Parent Income to Student Scores signifies that Parent Income is not a potential cause of Student Scores. Instead, this causal relationship is mediated by the variable Tutoring. Thus, this causal graph postulates that the causal relationship between Parent Income and Student Scores is fully mediated by School Funding and Tutoring. In this way, the causal graphs encode conditional independence assumptions about the data.

Using our graphical model, we can use the concept of d-separation in order to determine which variables to include in our conditioning set (adjustment set). The definition of d-separation will be formalized and expanded on in Chapter 2, but for now we can state that two nodes in a causal graph are d-connected if there is a path between them which is not blocked by a conditioned variable or a collider (two arrows pointing into the same node). We notice that School Funding has a path to Student Scores through School Resources, through Parent Income, and through Government Funding. Thus, School Funding is d-connected along non-causal paths to Student Scores. In order to run a causal analysis, the variables we are calculating a causal effect on must be d-separated

[1]Note: This is a simplified graph which is used to explain causal reasoning using graphs, and does not represent the full graph needed to encode the given problem

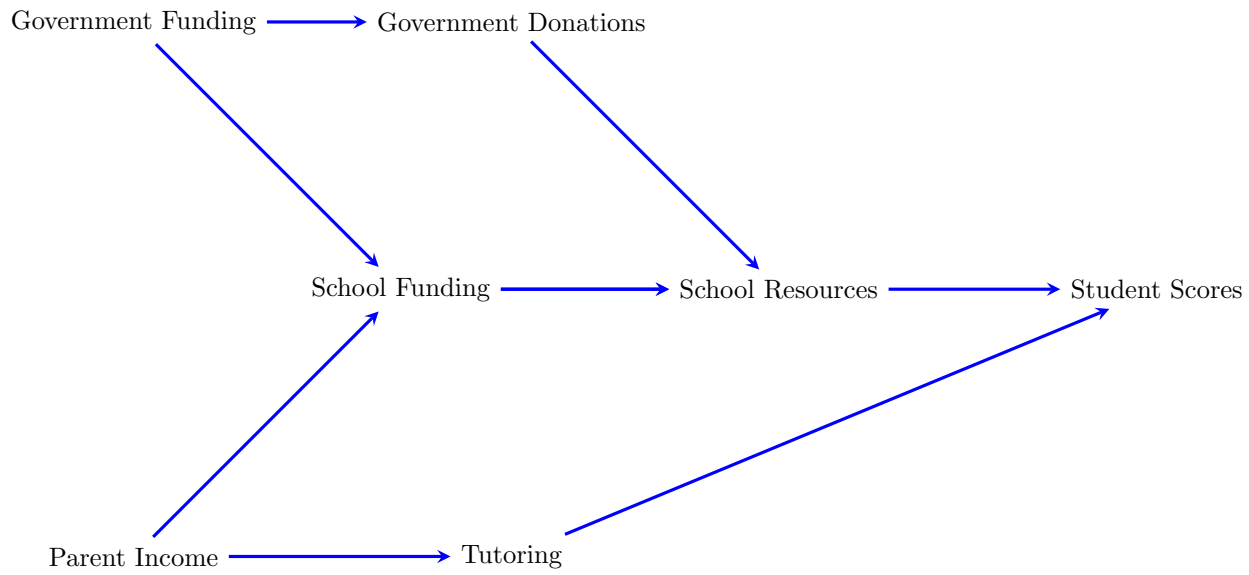


Figure 1.1: An example causal graph for impact of funding on student scores

along non-causal paths, which we can do by conditioning on the variables in the graph. Since the causal effect of School Funding on Student Scores flows through the causal path School Funding $\rightarrow$ School Resources $\rightarrow$ Student Scores, we need to be sure to not block this path, and only block the non-causal paths. Thus, in order to make the two nodes d-seperated, we must condition a node on both paths School Funding $\leftarrow$ Government Funding $\rightarrow$ Government Donation $\rightarrow$ School Resources $\rightarrow$ Student Scores and the path School Funding $\leftarrow$ Parent Income $\rightarrow$ Tutoring $\rightarrow$ Student Scores. We now notice that there are multiple ways in which to block the paths between the variables, or multiple adjustment sets. We can adjust on {Government Funding, Parent Income} or on {Government Donation, Tutoring}, or on many other combinations of these variables. Deciding on which one of these combination of variables to adjust on is a key intuition for this thesis.

When deciding on a set, the two important considerations we want to make are that of low variance and low bias. In order to make sure that our estimate for the causal effect is not biased, we need to condition on a set that d-separates the treatment and the outcome along non-causal paths. Thus, the choice of adjustment set from a pool of valid sets comes down to lower variance. In Section 2.3 we define the formula for computing the average causal effect using the backdoor set, so for now all we need to note is that to make this estimation we run a model on $Y|A, Z$, where Z is the adjustment set. Thus, the better our model $Y|A, Z$ is at predicting Y, the better (lower variance) our final estimate will be. To better predict Y, we should be using variables that are 'closer' to the outcome variable in the graph. The key intuition here is that if we have a path $C_1 \rightarrow C_2 \rightarrow Y$, C_2 will be a better predictor for Y than C_1, since all of the association between C_1 and Y flows through C_2. Direct parents of the outcome (nodes that have an outgoing edge into the outcome) will thus usually be the best predictors for Y, and thus the best variables to use in an adjustment set. We therefore notice in our graph that the adjustment set {Government Donation, Tutoring}

will be the set that is best at predicting Student Scores, and thus should be the adjustment set that provides us with the lowest variance. We now see that using the graph, we can often times easily find the optimal adjustment set, since we just need to find all the nodes closest to the outcome variables which make it d-separated from the treatment along non-causal paths. However, causal inference is often employed in situation where there are hundreds of variables in a system, such as when analyzing proteins or high-dimensional text data, and often times the structure of the graph is not well known. Thus, we need to devise criteria for identifying the best adjustment sets in graphs where the causal structure is unknown.

Chapter 2

Background

This section provides an introduction to the basic concepts of causal inference, as well as more advanced concepts in adjustment set selection and machine learning which will provide the reader with the necessary background knowledge to understand the contributions of the thesis. Section 2.1 introduces the concepts of a graphical model, the basic element which is most often used in causal inference research. Section 2.2 introduces the concept of d-separation, which is a way of conceptualizing independence relationships between variables in the graphical modes. We then apply this concept in Section 2.3 in order to compute the Average Causal Effect, based on a defined backdoor adjustment set. Section 2.4 expands on the concept of the backdoor adjustment set and defines an optimal adjustment set, an adjustment set with some defined optimal property. Finally, Section 2.5 defines the variable selection methods that will be used in the remainder of this thesis.

2.1 Graphical Models and Structural Equations

Causal inference focuses on the study of the discovery, identification, and estimation of causal relationships through data. In order to do this, causal researchers require more information regarding the origins of each variable and the relationships between variables. To be able to compute the causal effect of a treatment on an outcome, we must have information about the causal relationships between all of the variables, in order to avoid introducing bias from a confounding variable into our model. The description of the specific causal relationships between variables is done through the use of *Non-parametric Structural Equation Models* (NPSEM) (Pearl et al., 2000). NPSEMs are a set of generative functions which define the causal origin of a variable in terms of its parents and an error term. The parents of a variable are the set of variables in the dataset which have a potential causal effect on the variable. Note the use of 'potential causal effect' rather than 'causal effect', since the set of parents is found using a combination of literature review and causal discovery, where causal discovery will add as a parent all nodes whose edge to the variable is not probabilistically impossible. These are also often called *NPSEMIE*, with the last two letters signifying *Independent Errors*, since the errors should be assumed to be independent from each other. Errors which are not independent of each other are often a sign of unmeasured confounding, and thus a variable still missing from the

NPSEM. Thus, for a set of variables in $\mathcal{G}$, where each i column in the dataset is a variable $V_i | V_i \in \mathcal{G}$, the structural equation for the vertex will be

$$V_i = f(\text{pa}_{\mathcal{G}}(V_i), \epsilon)$$

where $\text{pa}_{\mathcal{G}}(V_i)$ is the set of all parents of node V_i in the set $\mathcal{G}$, f is an arbitrary function defined over the parents of the vertex, and ϵ_i is the error term for the given vertex. Using NPSEMs as compared to regular *Structural Equation Models* allows us to use an arbitrary function f and arbitrary error term ϵ_i without having to make assumptions regarding the distribution of the errors and the specifics of the function f. Most often, causal effects are calculated using a Linear Gaussian model with additive noise, where the structural equation for a given vertex V_i is more formally defined as a linear combination of its parents and an error term, or

$$V_i = \sum_{V_j \in \text{pa}(V_i)} \theta_{ij} V_j + \epsilon_i$$

where θ_i is the weight vector associated with the vertex V_i.

Instead of relying solely on NPSEMs in order to run causal analysis, researchers have adopted the use of causal graphs in order to represent the same information (Pearl et al., 2000). In a causal graph, a parental relationship between node V_j and V_i (i.e. node V_j is a parent of V_i and thus present in its structural equation) is defined by a directed edge. In a Directed Acyclic Graph (DAG) $\mathcal{G}$ defined over a set of nodes $\mathbf{V}$, nodes are connected to each other solely through the use of directed edges, with the additional requirement that there exist no directed cycles. Because edges in a DAG represent a potential causal relationship between variables, a causal model in a DAG is defined over the absent edges, since these edges imply causal independences between variables. A causal model set over a DAG is a set of distributions that factorize according to the following function:

$$p(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} p(V_i | \text{pa}_{\mathcal{G}}(V_i))$$

where $\text{pa}_{\mathcal{G}}(V_i))$ represents the set of parents of node V_i in the graph $\mathcal{G}$. An acyclic directed mixed graph (ADMG) is a modification of a DAG where we allow for the occurrence of bidirected edges. These bidirected edges represent common causal parent to two variables. Thus, the edge $V_i \leftrightarrow V_j$ means that V_i and V_j share a potential causal parent U_k which is unmeasured and thus not included in the graph.

Using our knowledge of DAGs and NPSEM, we can envision a potential intervention on a variable A in graph $\mathcal{G}$. An intervention on A implies artificially fixing the value of a variable A to some value a_0. These interventions can be formally expressed using do-calculus operators (Pearl and Mackenzie, 2018). The aforementioned intervention using do-calculus would imply modifying the structural equation for A to be a_0, and applying the do operator $do(A = a_0)$ to the graph. This application can be done using a Single World Intervention Graph (SWIG) (Richardson and Robins, 2013). To create a SWIG $\mathcal{G}(a_0)$, we modify a DAG $\mathcal{G}$ by adding a new vertex a_0, removing all outgoing edges from A, and instead turning them into outgoing edges from a_0. Finally, we modify

all descendants of the vertex a_0 into their respective potential outcomes, so that a vertex V_i would become $V_i(a_0)$.

We can see the SWIG in practice as applied to the causal graph of school funding we have been using as an example. We can imagine an intervention on funding, where the state decides that a specific school will now receive per pupil funding of value a. In a DAG, this amounts to the the intervention Sch. Funding(a). Our SWIG for this intervention would be as follows

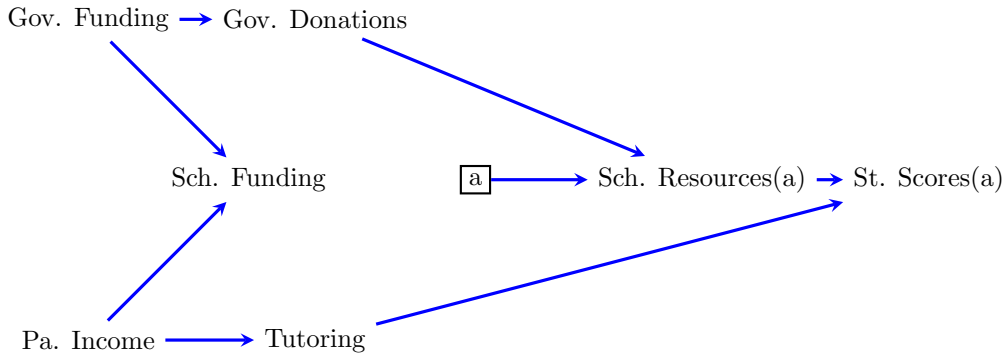
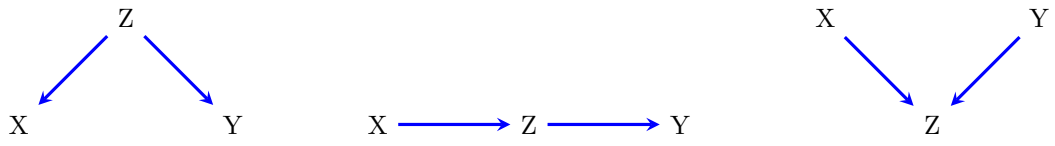


Figure 2.1: An example causal graph for impact of funding on student scores

2.2 D-separation

Since graphical models like DAGs are a graphical way of encoding independence relations between variables, we can leverage the graphical nature in order to create more complex graphical criteria which depict more complex statistical relationships. One such relationship which is very useful in causal inference is the relationship of conditional independence. We already know that two vertices connected by an edge are not statistically independent. However, this is not always the case for two vertices connected by a path of longer than two edges. In order to determine independence, we must first define three types of graphical triplets. The first type of triplet is a fork triplet, shown in figure 2.2i.



(i) A fork triplet (ii) A chain triplet (iii) A collider triplet

Figure 2.2: The three graph triplet components

In the fork triplet, X and Y are correlated with each other by a common cause Z. We note that X and Y are not independent, but when we condition on Z, X and Y are independent. Thus, X and Y are conditionally independent given Z, or $X \perp\!\!\!\perp Y|Z$. Conditioning on Z is said to "block" the path between Z and Y.

Our second structure is a chain, depicted in figure 2.2ii. In a chain triplet, once again $X \not\perp\!\!\!\perp Y$, and just as in the fork triplet $X \perp\!\!\!\perp Y|Z$. Though forks and chains appear different, they have the same statistical properties and are thus statistically indistinguishable from each other - conditioning on Z is said to "block" the path between Z and Y.

Finally, we have the collider triplet depicted in 2.2iii. The collider triplet is caused when two nodes X and Y are both parents of a node Z. In this triplet, $X \perp\!\!\!\perp Y$, while $X \not\perp\!\!\!\perp Y|Z$. Furthermore, when conditioning on any of the descendants of Z, we get the same statistical property that $X \not\perp\!\!\!\perp Y|\mathrm{de}_{\mathcal{G}}(Z)$. The collider is originally blocked, while conditioning on Z or any of its descendants means that the path will be "unblocked".

We can now use these triplets in order to define conditional independence among any two possible vertices in a graph $\mathcal{G}$. Two nodes V_i and V_j are conditionally independent given a set of nodes Z if conditioning on Z blocks all paths in between the two nodes. Paths are considered blocked if there exists a blocked chain or fork, as well as a blocked / unconditioned collider. We can now define d-separation between two vertices as the absence of an unblocked path between the two vertices when conditioning on a set Z. Given the d-separation criterion in the graphical model, we can then connect this to the underlying distribution using the *Global Markov Property* (Pearl et al., 2000):

$$X, Y \text{ d-sep in } \mathcal{G}|Z \Rightarrow X \perp\!\!\!\perp Y|Z \text{ in } p(V) \tag{2.1}$$

The Global Markov Property allows us to ground d-separation in the distribution, and connects the DAG to the NPSEM. D-separation is known to be a both sound and strongly complete algorithm. Using the Global Markov Property, we can define the Local Markov Property on a DAG over its parents, defined as

$$X \perp\!\!\!\perp \mathrm{nde}_{\mathcal{G}}(X) \setminus \mathrm{pa}_{\mathcal{G}}(X)|\mathrm{pa}_{\mathcal{G}}(X) \tag{2.2}$$

where $\mathrm{nde}_{\mathcal{G}}(X)$ represents the non-descendants of the node X in graph $\mathcal{G}$. Thus, a variable is independent of all of its non-descendants (excluding its parents) when it's parents are given. This leads to an important intuition regarding which nodes are useful for predicting other nodes, which is used throughout this thesis.

We can also extend the local Markov Property to ADMGs, as defined in Richardson et al. (2017). This is defined over an ADMG $\mathcal{G}$ as

$$X \perp\!\!\!\perp \mathrm{nde}_{\mathcal{G}}(X) \setminus \mathrm{mb}_{\mathcal{G}}(X)|\mathrm{mb}_{\mathcal{G}}(X) \tag{2.3}$$

where $\mathrm{mb}_{\mathcal{G}}(X)$ is the Ordered Markov Blanket of X, which can be defined as the district of X, and the parents of the district of X, minus X. The district of X is further defined as all the variables that lie on a bidirected path from X. The Local Markov Properties for DAGs and ADMGs tell us that we can isolate a node from the rest of the graph using its parents or its markov blanket.

2.3 Backdoor Adjustment

Using d-separation, we can now formalize a way to compute the Average Causal Effect (ACE). A reminder, the ACE is a measure of how much a change in the treatment causes a change in the outcome. In a Randomized Controlled Trial, the classical experimental setup, all other variables are aimed to be held constant besides the treatment, so that we can just measure the difference in the mean outcome between the treatment and non-treatment groups. However real world data is a lot more messy, which is what motivates us to create the causal DAGs in the first place. In order to make sure the ACE we are computing is not biased by confounding variables, we can utilize d-separation in order to define the backdoor criterion on a graph $\mathcal{G}$, a graphical criterion which tells us which variables to condition on [cite].

The backdoor criterion states that an adjustment set Z is a valid adjustment set if it adheres to 2 criteria:

- None of the vertices Z_i in Z are descendants of the treatment variable A in the graph $\mathcal{G}$.
- A is d-separated from Y given the set Z, i.e. conditioning on all of the nodes in Z blocks all paths from A to Y.

Recall from the introduction that the ACE is defined as

$$ACE = E[Y(a)] - E[Y(a_0)]$$

where a and a_0 are two different treatment options.

The backdoor criterion allows us to define the backdoor adjustment formula, which, given a valid adjustment set Z, defines the Expected value of $Y(a)$ as

$$E[Y(a)] = \sum_Z p(Z) \times E[Y|A = a, Z]$$

Using this formula, we can reformulate the average causal effect to be

$$ACE = \sum_Z p(Z) \times E[Y|A = a, Z] - \sum_Z p(Z) \times E[Y|A = a_0, Z] \tag{2.4}$$

We thus have a valid way for computing the ACE in causal DAGs.

2.4 Optimal Adjustment Sets

Even though we now have a valid formula for computing the ACE, there is still the question of choosing which of the possible valid adjustment sets to use. We notice that our estimation of the ACE relies on a valid estimation of $E[Y|A, Z]$. Thus, the 'best' choice for an adjustment set Z is one which is best at predicting Y. When talking about "best" in statistics, it is often framed in terms of the bias-variance tradeoff. The bias of a given model is the difference between it's prediction of a

value vs the actual true value that we are trying to predict. In terms of the backdoor criterion, this would mean a model that does not fully d-separate the treatment and outcome, so that our prediction of the causal effect is biased. The variance of a model is the amount the estimate changes if the model is re-run, especially on new data. An optimal adjustment set is thus informally defined as the valid adjustment set which is best at predicting the outcome. Best in this case refers to predicting the outcome with lowest amount of variance, while still blocking all spurious correlation such as to retain zero bias. In 2 we prove that the set of optimal predictors for the outcome are the parents of the outcome, and thus the best predictors for the outcome St. Scores would be Sch. Resources and Tutoring. However, if we remember our SWIG, Sch. Resources became Sch. Resources(a), i.e the potential outcome of Sch. Resources. Because of this, it is a mediator and thus cannot be used in our adjustment set. Thus, finding the optimal predictors in causal inference is slightly trickier than in regular statistics. We know that for any set Z we choose, it must fulfill the condition $Y(a) \perp\!\!\!\perp A|Z$, since the set needs to d-separate the outcome from the treatment. We can notice that in a set that is not optimal, i.e. where the parents of the outcome are not being used, there will be a 'better' predictor V_i such that $V_i \not\perp\!\!\!\perp Y|Z$. This is because there is additional information about the outcome that is not being used in the adjustment set. Thus, for an optimal set Z, given all the nodes in the graph $\mathcal{G}$, $\forall V_i \in V \setminus (Z, Y)$, it must be true that $Y \perp\!\!\!\perp V_i|Z$. Rotnitzky and Smucler (2020) defined a graphical criterion to identify an optimal adjustment set O which adheres to these properties, which is defined as follows:

$$\mathbf{O}_{\mathcal{G}}(\mathbf{A}, \mathbf{Y}) = \mathrm{pa}_{\mathcal{G}}(\mathrm{cn}_{\mathcal{G}}(\mathbf{A}, \mathbf{Y})) \setminus \mathrm{forb}_{\mathcal{G}}(\mathbf{A}, \mathbf{Y})$$

Where $\mathbf{O}_{\mathcal{G}}(\mathbf{A}, \mathbf{Y})$ is the optimal adjustment set in graph $\mathcal{G}$, where $\mathrm{cn}_{\mathcal{G}}(\mathbf{A}, \mathbf{Y})$ is the set of nodes that lie on the causal path between A and Y, i.e. the potiential outcomes of the form $V_i(a)$, and $\mathrm{forb}_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}) = \mathrm{de}_{\mathcal{G}}(\mathrm{cn}_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}))$, where $\mathrm{de}_{\mathcal{G}}(X)$ is the set of descendants of the node X in graph $\mathcal{G}$. This means that the optimal adjustment set is defined as the set of all parents of the potential outcomes on the causal path between a and *Y(a)* in a SWIG $\mathcal{G}(a)$.

Thus, in order to find the optimal backdoor adjustment set, we just need to find all of the parents of the potential outcomes on the causal path. In this case, the potential outcomes are Sch. Resources(a) and St. Scores(a), and the set of their parents is {Tutoring, Gov. Donations}. This gives us an easy criterion to use when the DAG is known, though it does not function when the DAG is unknown, since it is reliant on the graph structure.

However, this is often times not the simplest adjustment set, since there are cases when more variables than are needed are present in the optimal adjustment set, such as if we were to add the edge Pa. Income → Sch. Resources, in which case the optimal adjustment set would be {Tutoring, Gov. Donations, Pa. Income}, and we can notice that even if we were to remove Tutoring from this adjustment set, it would still be a valid adjustment set since Sch. Funding would still be d-separated from St. Scores. Thus we can define a *minimal* adjustment set as a set where the removal of any one of the elements would result in a non-valid adjustment set. We can combine a minimal and optimal adjustment set into a minimal optimal adjustment set, which is an optimal adjustment set that cannot be pruned further than it already has.

We notice that the minimal optimal adjustment set is theoretically the set which would reduce the variance of our estimate the most while retaining low bias, since it contains a minimal amount of variables, all of which are the variables that most predict the outcome. The minimal optimal adjustment set is thus most often the goal for causal estimation methods.

2.5 Variable Selection Methods

Crudely put, variable selection methods are methods and algorithms which select a subset of significant predictors for an outcome from a larger set of covariates. This is very popular in machine learning and statistics, especially in fields which deal with high-dimensional datasets, where most covariates are not very useful. We note that the goal of variable selection is very similar to that of adjustment set selection, i.e. selecting a subset of variables that are useful for analysis. In this thesis, we show that using variable selection methods can be extremely useful in identifying an adjustment set. Specifically, we will be focusing on variable selection methods which are 'consistent'. Informally, a consistent variable selection method should recover all the variables that are actually important for prediction of the outcome, and nothing else. More formally, if we have some set of covariates $\mathbf{X}$ and some outcome Y, and some subset $\mathbf{X}' \subseteq \mathbf{X}$ such that $\mathbf{X} \setminus \mathbf{X}' \perp\!\!\!\perp \mathbf{Y} | \mathbf{X}'$, then a consistent variable selection algorithm $\mathcal{M}$ is one which identifies this subset $\mathbf{X}'$ with high probability as the sample size tends to infinity. Obviously in practice we do not have a sample size that tends to infinity, but from empirical observations and simulations the methods are still consistent in practice.

One example of a consistent variable selection method is the stepwise BIC. The formula for the BIC for a given model is

$$BIC = -2l(\theta) + d * \log(n) \tag{2.5}$$

where $l(\theta)$ is the log likelihood of the model, d is the number of parameters used in the model, and n is the sample size. The stepwise BIC process starts by computing the BIC for any initial desired model (usually the empty or full model). Then, we compute the BIC score for all potential modifications to the original model, where either one variable is subtracted from the model, or one variable is added to the model. We choose the model with the lowest BIC score out of these models, and then repeat the process until changing the model would not improve the BIC score. The BIC can be computed over any type of model, but is most often seen in practice for a regression.

Another example of a consistent variable selection procedure is the Lasso penalty. Lasso regression uses shrinkage, where the non-significant coefficients for a regression are shrunk down to zero or near zero. To apply lasso to a regression, we would use the following likelihood function:

$$LASSO = l(\theta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.6}$$

where p is the number of parameters in the model, β_j is the jth coefficient in the model, and λ is a model parameter that has to be decided upon. Both Lasso and BIC are incredibly useful in

determining which variables are important to include in a model. However, since the BIC produces a model over the course of many steps, and Lasso produces a penalized model, we would have to use some form of sample splitting or cross validation in order to compute the final ACE estimate if we were using these variable selection methods in our analysis. Sample splitting would mean splitting our data, and using half of it to find the important covariates, and then the other half to estimate the ACE, which means that we would end up utilizing less data than we have in our estimation. This is a drawback for both Lasso and BIC, which is why in this thesis we use an approximation of the BIC (which they call MIC) defined by Su et al. (2016). This is still a consistent method which both induces sparseness in the covariates set to identify the significant predictors, and produces a final model using this covariates which we can use in our analysis. This allows us to avoid any cross-validation or sample splitting, and lets us utilize the entire dataset in our ACE estimation.

The MIC functions by adding a penalty term meant to approximate the BIC during the maximum likelihood estimation process. The authors define a new term γ_j, and using the tan function $\tanh(\alpha|\beta|^r) = \frac{\exp(2\alpha|\beta|^r)-1}{\exp(2\alpha|\beta|^r)+1}$ they define a weight matrix with terms $w_j = \tanh(\alpha\gamma_j^2)$ such that the coefficients β_j are defined as $\beta_j = \gamma_j * w_j$. The coefficients β_j are the coefficients for each of the variables in the model, and thus γ_j is defined through β_j. Finally, the matrix $\mathbf{W} = \text{diag}(\mathbf{w_j})$, where $\text{diag}(w_j)$ is the diagonal matrix formed using w_j. Using this definition, instead of performing normal MLE to find a regression, we perform a minimization over γ on the function

$$-2l(\mathbf{W}\gamma) + \lambda_\mathbf{0} * \text{tr}(\mathbf{W}) \tag{2.7}$$

where $\text{tr}(\mathbf{W})$ is the trace of the matrix $\mathbf{W}$. This allows us to perform gradient descent on the approximate BIC latent space, and thus only requires us to fit a single model, whose terms should approximate the best BIC model. We use this MIC implementation in the rest of the thesis for all simulations.

Chapter 3

Previous Criteria for Optimal Adjustment Set Selection

The previous section introduced criteria for identifying optimal adjustment sets when the graph is exactly known, such as the criterion in Rotnitzky and Smucler (2020). However, an important area of research in causal inference that does not receive as much attention lies in creating criteria for identifying optimal adjustment sets when the graph is unknown or partially known. Section 3.1 will introduce the concept of a criterion for identifying optimal adjustment sets in this unknown setting, as well as the motivation for expanding this area of research. Sections 3.2, 3.3, and 3.4 introduce and evaluate existing criteria currently in use by both causal inference researchers and the machine learning community at large. I show that currently existing criteria in the unknown setting are either not valid, or rely on background knowledge that may be unavailable to the analyst. This provides context for the research in the remainder of the thesis, which aims to define a truly non-graphical criterion for optimal adjustment set selection.

3.1 Motivation

We have previously seen a robust graphical criterion for identifying optimal adjustment sets, which can be extended to identify minimal optimal adjustment sets (Rotnitzky and Smucler, 2020). In fact, most criteria in the literature are graphical in nature, or require some knowledge of the causal graph for their computation. This knowledge of the causal graph can be either inferred through previous domain knowledge by asking experts, or through the use of causal discovery algorithms. Using previous domain knowledge presupposes a working level of knowledge in the subject matter. Many domains have low level of knowledge, which is partially why they are targeted for causal analysis, and thus will not always be able to provide all the causal relationships between variables in a graph. Furthermore, many fields and questions in causal analysis require the inclusion of hundreds if not thousands of variables in the initial consideration. Examples of such fields are genomics (Bühlmann et al., 2014; Colombo et al., 2014; Maathuis et al., 2010), or natural language processing (Feder

et al., 2022; D'Amour et al., 2021).

The second process mentioned for the uncovering of causal knowledge are methods of causal discovery, such as the Fast Causal Inference (FCI) (Spirtes et al., 2000, 1993) or Fast Greedy Equivalent Search (FGES) (Ramsey et al., 2017) algorithms. These causal discovery algorithms use the conditional independences of the data in order to determine which edges are likely. However, these methods are not perfect, and will not always return the true underlying causal graph, since different algorithms may produce differing graphs (Shen et al., 2020). Thus, the use of causal discovery introduces a further point of failure, as an erroneous causal graph has the potential to bias the rest of the analysis. It is thus useful to develop causal inference methods which do not require any knowledge of the graph, and can function using data from unknown graphs.

We can now look at some of the previously proposed criteria for unknown graphs, and evaluate whether or not they function as intended. For all of these criteria, and for all future chapters, I will use the same exact setup. We will have a treatment A, an outcome Y, and a set of covariates C. Before evaluating the previous criteria, I will lay out to of the main assumptions regarding the setting and algorithms we use in the remainder of the thesis.

Assumption 1. $C \cap \text{de}_{\mathcal{G}}(A) = \emptyset$

where as a reminder, $\text{de}_{\mathcal{G}}(A)$ is the set of descendants of A in $\mathcal{G}$. Thus, Assumption 1 states that all covariates we are working with are non-descendants of the treatment. This is used since in most real world studies there exist some baseline covariates that are correlated with both the treatment and the outcome, and focusing only on pre-treatment covariates allows us to ignore the challenges that come with conditioning on post-treatment variables.

Assumption 2. We use a variable selection technique $\mathcal{M}$ which is a consistent method as defined in Section 2.5.

Assumption 2 is used since our interest is in variable selection and finding subsets of C that exhibit some form of efficiency. Because of this, consistent variable selection methods are an important subclass of algorithms to consider, and these restrictions only improve the results of our analysis. As mentioned earlier, examples of consistent variable selection algorithms include L1 regularization, BIC, and the MIC. For the rest of this thesis, all practical applications will be done using the MIC method from Su et al. (2016) as our consistent variable selection technique $\mathcal{M}$. A python implementation of this method is a sub contribution of this thesis.

3.2 Intersection Criterion

The first method we will review is the intersection criterion, used most recently in a paper by Zeng et al. (2022). In order to identify the backdoor adjustment set, we start with some set of covariates C, then run our consistent variable selection procedure $\mathcal{M}$ to find a subset of good predictors for A, and a subset of good predictors for Y, and finally take the intersection of these two sets. This is defined more formally on some set of covariates $C = \{C_1, C_2, \ldots, C_n\}$. The subset of good predictors for A is defined as $C^A_{opt} \subseteq C$, and the subset of good predictors for Y as $C^Y_{opt} \subseteq C$. We identify

C^A_{opt} by running a variable selection procedure $\mathcal{M}$ for $P(A|C)$, and we identify C^Y_{opt} by running the same variable selection procedure $\mathcal{M}$ over $E[Y|A,C]$[1]. The intersection criterion is thus defined as $C_{Int} = C^A_{opt} \cap C^Y_{opt}$. The main benefit of employing this criterion is that it induces sparseness in the final adjustment set, a desirable and useful outcome in high-dimensional cases where a sizeable conditioning set would make inference difficult. This makes this method and appealing and intuitive solution for machine learning practitioners without extensive experience in causal inference, as their focus is optimizing the bias-variance trade-off. In practice however, I will show that this is not a sound criterion. That it, in certain graphs, the intersection criterion will not provide a valid adjustment set, let alone an optimal one.

Proposition 1. *There exists a DAG $\mathcal{G}$ such that C_{Int} is not a valid backdoor adjustment set with respect to A and Y.*

Proof. Let us assume that $\mathcal{G}$ is a causal DAG with treatment A, outcome Y, and a set of pre-treatment covariates $C = \{C_1, C_2, \dots\}$. Let us further assume that in graph $\mathcal{G}$ there exists some backdoor path between A and Y where every variable in the path is either a parent of Y, a parent of A, or neither, but that no variable in the path is both a parent of Y and A simultaneously. We will also assume that this path is an open backdoor path, i.e. it does not contain any colliders. An example of such a path is provided in Figure 3.1. According to the Local Markov Property for DAGs defined in Equation 2.2, a variable X is independent of the rest of the covariates in the graph given its parents. Since we assume that all variables are pre-treatment variables by Assumption 1, a consistent variable selection technique $\mathcal{M}$ would identify the set $\text{pa}_{\mathcal{G}}(X)$ as the subset of good predictors for X. Thus, $C^A_{opt} = \text{pa}_{\mathcal{G}}(A)$ and $C^Y_{opt} = \text{pa}_{\mathcal{G}}(Y)$. The intersection of $\text{pa}_{\mathcal{G}}(A)$ and $\text{pa}_{\mathcal{G}}(Y)$ will not contain any of the covariates in the aforementioned backdoor path, since none of the variables in the path share both A and Y as parents. Hence, none of the covariates from that path will be included in the conditioning set, and that backdoor path will remain open. This means that C_{Int} will not form a valid backdoor adjustment set, since it will not block all non-causal paths between A and Y. □

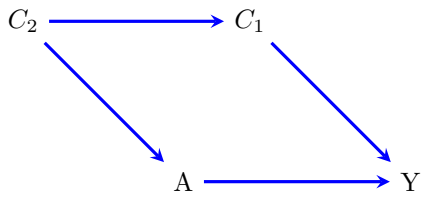


Figure 3.1: A hypothetical counterexample graph for the intersection criterion

An example of such a graph $\mathcal{G}$ is seen in Figure 3.1, and we can quickly see the outcome of the Intersection Criterion in this graph. We see from the logic of the previous proof that the Intersection

[1]Note that if the selection procedure were on $E[Y|C]$, the intersection criterion would produce a valid adjustment set in DAGs, as the intersection would produce the set $\text{pa}_{\mathcal{G}}(Y)$. However, this would not be the optimal set of predictors. Note also that using $E[Y|C]$ would still produce an invalid adjustment set in ADMGs.

Criterion would produce the set {} as the adjustment set, and since there is still a non-causal path of the form $A \leftarrow C_2 \rightarrow C_1 \rightarrow Y$, this adjustment set is not a valid set

To demonstrate the bias incurred when using the intersection criterion, we now simulate data from the data generating process defined in Figure 3.1. In our simulated data, the ACE is specified as 2, and the values for the rest of the edges are shown using the structural equations in Equations 3.1-3.4. We ran 100 trials for each sample size of 50, 100, 250, and 500. We then used the implementation of the MIC to find the set of best predictors for Y and for A[2]. Finally, we took the intersection of these sets to produce the backdoor adjustment set according to the intersection criterion, and compute the ACE using the corresponding backdoor adjustment set formula from Equation 2.4. The resulting graph showing the box plots of ACE estimates for different sample sizes is shown in Figure 3.4.

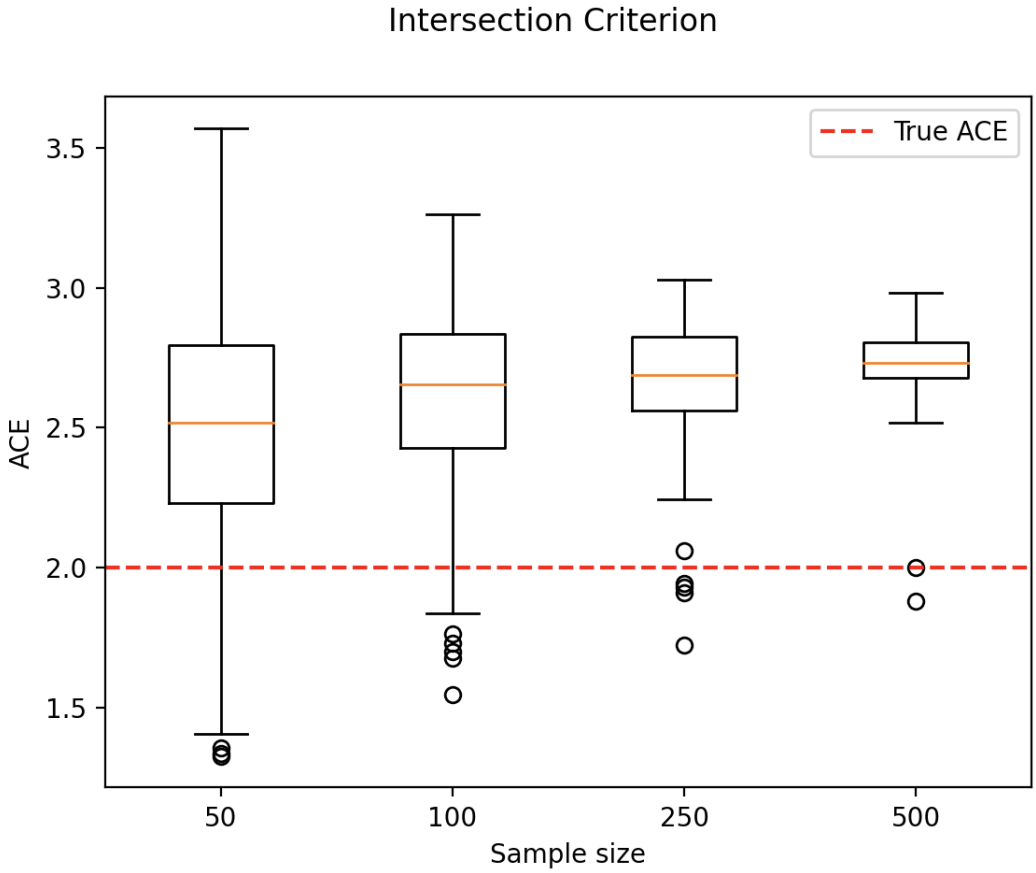


Figure 3.2: Computed ACE for simulated data based on Figure 3.1 at different sample sizes

We see that regardless of the sample size, the estimate for the Average Causal Effect is biased. In fact, as the sample size increases, even though the variance of the estimates decreases, the estimate converges asymptotically to a biased result. Note that this is a low-dimensional setting, where only two covariates are present. The results from this simulation would thus be more heavily pronounced

[2]Note that the MIC identified the theoretically expected sets 90% of the time

in high-dimensional settings and in more complex causal graphs.

$$C_2 = \text{Normal}(0, 1) \tag{3.1}$$
$$p(A = 1 \mid C_2) = \text{expit}(C_2/2) \tag{3.2}$$
$$C_1 \mid C_2 = C_2 + \text{Normal}(0, 1) \tag{3.3}$$
$$Y \mid A, C_1 = 2 + 2 * A + C_1/2 + \text{Normal}(0, 1) \tag{3.4}$$

Thus, despite its intuitiveness and sparsity, this criterion will fail when the assumptions for Proposition 1 are met, and in simulations is shown to produce a biased estimate for the ACE when this is the case. This means that in high-dimensional settings, where the parents of A and Y are less likely to overlap, the criterion is much more likely to produce a biased estimate of the causal effect.

3.3 Union Criterion

We can modify the process for the intersection criterion by taking the union of the two sets of good predictors, rather than the intersection. This would theoretically trade some of the induced sparseness for more soundness. The Union Criterion was recently used in Belloni et al. (2014), and works by taking the union of the sets of good predictors for Y and A, found through the use of the consistent variable selection method $\mathcal{M}$. Formally, we define the Union Criterion as $C_{Union} = C^A_{opt} \cup C^Y_{opt} \setminus A$. We can now prove its soundness in DAGs, where here 'soundness' means that it is theoretically expected to recoever a valid backdoor adjustment set.

Proposition 2. *The Union Criterion is a sound criterion for finding a valid backdoor adjustment set in DAGs*

Proof. By the Local Markov Property defined in Equation 2.2, a consistent variable selection method $\mathcal{M}$ is guaranteed to identify all of the parents of Y for the significant predictor set C^Y_{opt}. It is known thatpa$_\mathcal{G}(Y)$ is a valid backdoor adjustment set. Since taking the union with C^A_{opt} will only add variables to the conditioning set, and adding variables does not change the validity of a conditioning set in the case of a DAG, then the Union of the two sets will still be a valid adjustment set, and thus C_{Union} is a valid adjustment set. □

Thus, the Union is already an improvement over the Intersection Criterion, as it is sound in DAGs. However, as soon as we include unmeasured confounding variables, the soundness falls apart. That is, in certain Acyclic Directed Mixed Graphs (ADMGs), the Union Criterion will not provide a valid adjustment set. We can take a look at a specific example where this is the case, where the data generating process is shown in Figure 3.3.

Our treatment is once again A, the outcome is Y, and the two possible covariates are C_1 and C_2, where mb$_\mathcal{G}(A) = \{C_1, C_2\}$ and mb$_\mathcal{G}(Y) = \{C_1, C_2, A\}$. Thus, we see that the subset of good predictors for A is $C^A_{opt} = \{C_1, C_2\}$, and the subset of good predictors for Y is $C^Y_{opt} = \{C_1, C_2, A\}$, according to the local Markov Property for ADMGs. Thus, the union criterion here will provide us

with the backdoor adjustment set $C_{un} = \{C_1, C_2\}$ (Note also that the intersection criterion would identify the same set in this example). We see that there is still a non-causal path between A and Y of the form $A \leftrightarrow C_2 \leftrightarrow C_1 \leftrightarrow Y$. Thus, the adjustment set provided by the union criterion for the graph in 3.3 is not a valid set.

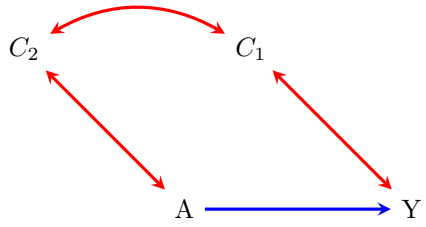
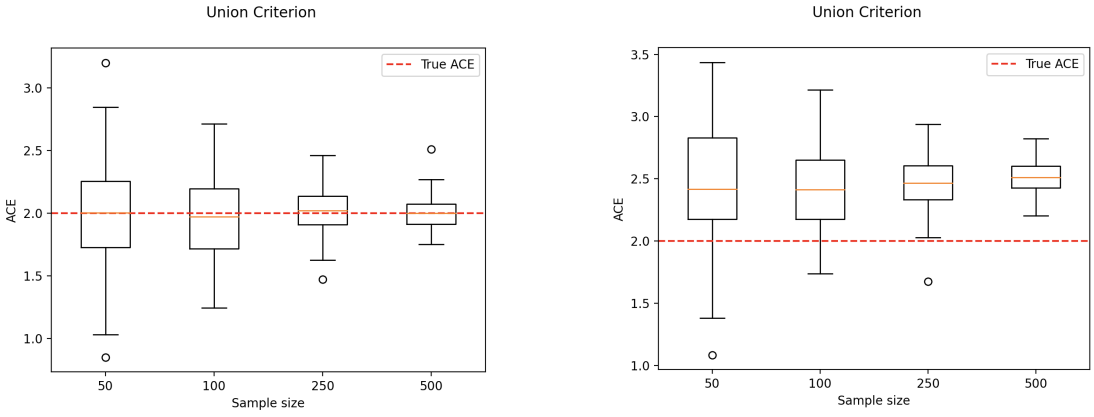


Figure 3.3: A hypothetical counterexample graph for the union criterion

We can once again run simulations to see the outcomes of the union criterion in practice. Here, we simulate data from the data generating process (DGP) defined in both Figure 3.1 and Figure 3.3. Once again, the ACE is set to be 2 in both case, and the structural equations are shown in Equations 3.5-3.11. For each graph, we simulated 100 samples of data for each sample size, then used in the union criterion to find the corresponding backdoor adjustment set, and finally compute the ACE using the backdoor adjustment formula from Equation 2.4. The resulting graphs are shown in Figure 3.4



(i) DGP from Figure 3.1

(ii) DGP from Figure 3.3

Figure 3.4: Computed ACE for simulated data based on DGPs of previous graphs

$$U_1 = \text{Normal}(0, 1) \tag{3.5}$$
$$U_2 = \text{Normal}(0, 1) \tag{3.6}$$
$$U_3 = \text{Normal}(0, 1) \tag{3.7}$$
$$C_2 \mid U_1, U_2 = 2 * U_1 + 2 * U_2 + \text{Normal}(0, 1) \tag{3.8}$$
$$p(A = 1 \mid U_1) = \text{expit}(U_1 + 1) \tag{3.9}$$
$$C_1 \mid U_2, U_3 = 2 * U_2 + 2 * U_3 + \text{Normal}(0, 1) \tag{3.10}$$
$$Y \mid A, U_3 = 1 + 2 * A + U_3 + \text{Normal}(0, 1) \tag{3.11}$$

In Figure 3.4i, we see that the ACE computed using the union criterion for the graph without unmeasured confounders is unbiased, and is centered around the true ACE. As the sample size increases, the variability of the estimate decreases, though it still remains centered around zero. This follows from Proposition 2, as the graph in Figure 3.1 is a DAG. In Figure 3.4ii we still see the decreasing variability of the estimates as the sample size increases, though now the estimates asymptotically converge to a biased estimate. This is because the graph in Figure 3.3 contains unmeasured confounding.

We can now formalize this and prove that for a certain family of ADMGs, the Union Criterion will fail to produce a valid adjustment set.

Proposition 3. *There exists an ADMG $\mathcal{G}$ such that C_{Union} is not a valid backdoor adjustment set with respect to A and Y.*

Proof. In order to prove this, we can use the Local Markov Property for ADMGs as defined in 2.3. A reminder that we are using the Ordered Markov Blanket, which is defined as the district of X, and the parents of the district of X, minus X. The district of X are all of the variables that lie on a bidirected path from X. Our consistent variable selection technique $\mathcal{M}$ would identify the Ordered Markov Blanket of X as the subset of good predictors for X. Thus, $C^A_{opt} = \{\text{pa}_\mathcal{G}(A), \text{mb}_\mathcal{G}(A)\}$ and $C^Y_{opt} = \{\text{pa}_\mathcal{G}(Y), \text{mb}_\mathcal{G}(Y)\}$. Let us now assume that there is a non-causal path of bidirected edges in between A and Y, flowing through some set of covariates C_{bd}. Since all the vertices in C_{bd} are on a bidirected path from both A and Y, C_{bd} will be in both sets of best predictors, and thus will be part of the union criterion set. This means that C_{bd} will be part of the conditioning set. According to the rules of d-separation (2.2), by conditioning on each of the vertices in C_{bd}, we will be opening up the colliders on the bidirected path, and thus the bidirected path itself will be open. Since this is a bidirected path between A and Y, we will have a non-blocked non-causal path between A and Y, and thus our final estimate for the Average Causal Effect would be biased. An example of this type of graph is shown in Figure 3.3, where $C_{bd} = \{C_1, C_2\}$. (Note that this is not the only situation in which case the union criterion produces a biased result, but just one example). □

Thus, in both theory and practice, the union criterion produces a biased estimate of the ACE.

Therefore, despite its soundness in DAGs, the union criterion may fail in the presence of unmeasured confounding, in situations not limited to the one shown in Figure 3.3. This is of course

leads to bias in high-dimensional settings, where there are often a lot of unmeasured confounding variables that may create bi-directed edges. While there are no tests to confirm / deny the existence of unmeasured confounding, in Chapter 4 we define a procedure by which we can confirm of falsify the validity of an adjustment set, which would allow us to identify cases where we might be faced with unmeasured confounding.

3.4 Shpitser-VanderWeele Criterion

We can now move to the last criterion on this list, which is the criterion from VanderWeele and Shpitser (2011). This criterion is based on rules rather than variable selection methods. The backdoor adjustment set identified by this criterion is the union of the set of causes for A and the set of causes of Y. Thus, the backdoor adjustment set for the Shpitser-VanderWeele Criterion (SWC) would be $C_{SWC} = \text{pa}_{\mathcal{G}}(Y) \cup \text{pa}_{\mathcal{G}}(A)$. We notice that for the previous example in 3.3, where both the union and the intersection criterion would have failed, the SWC would identify $C_{SWC} = \{\}$ as the backdoor adjustment set, and thus all of the non-causal paths would remain unblocked, leading to an unbiased estimate. In fact, Shpitser and VanderWeele proved that this criterion is sound for in the presence of both measured and unmeasured confounding. However, using this criterion poses a number of problems.

The first problem is that of usefulness. In most cases, we are answering question using causal inference because of a lack of ability to run randomized controlled trials. This means that often times the setting / domain of the question is one where causation may be difficult to determine. This occurs in fields such as public health or natural language processing. Thus, determining whether a covariate is a cause of the treatment or the outcome may be a question that researchers are unable to answer, thus meaning this criterion would not be usable. This means that this criterion would not be especially useful as a criterion for unknown graphs, since non-graphical criteria are used in cases where there is little background information with which to construct a hypothetical graph. Furthermore, in cases of high-dimensional data, most criterion for known graphs use some form of variable selection and causal discovery in order to learn the graph. Since this causal discovery would not answer all of the questions regarding the parental relationships between variables, using this learned graph would be unreliable. Thus, we would need even more background knowledge to answer both of the SWC questions for every single one of the variables in the graph. Thus, this criterion is only useful in low-dimensional cases with plenty of background knowledge.

The second issue stems from the first, and is that of robustness. In cases where the researchers managed to answer all questions about causal relationships, if one of these questions were answered incorrectly and thus a causal relationship is incorrectly stated, there is no way for the researchers to identify this mistake. They will find an estimate of the causal effect which may be unbiased due to the error, but there is no process in place in order to verify that the backdoor adjustment set that they have selected is valid and thus that their estimate is unbiased. Therefore, even though the criterion is sound, there are no procedures to check the validity of the final adjustment set.

Finally, there is the issue of optimality. Even though this criterion is sound, the set of causes for both Y and A may be a large set depending on the initial set of covariates that we are working

with. Thus, the size of the adjustment set will grow with the size of the covariate set. This means that there could be variables in the adjustment set which are not required in the adjustment set, thus unnecessarily increasing the variability of the final estimate. In cases of high-dimensional data, it is often advantageous to induce sparseness in our adjustment set, since it reduces variance as well as increases the interpretabilty of our final model, which is not something that occurs with the Shpitser-VanderWeele Criterion. Thus, even though the criterion proposed in VanderWeele and Shpitser (2011) is sound for both measured and unmeasured confounders, it may be difficult to actually utilize it in practice.

Chapter 4

New Criterion for Covariate Selection

Chapter 3 explored criteria for identifying a backdoor adjustment set in unknown graphs that have been previously defined in the literature. We saw from Section 3.2 that the Intersection Criterion is not a sound criterion for causal DAGs in the fully observed setting. In Section 3.3 we saw that although the Union Criterion is an improvement as it is sound in causal DAGs, it is still not a sound criterion in ADMGs in the presence on unmeasured confounders. Finally, in Section 3.4 we saw that even though the Shpitser-VanderWeele Criterion is a sound criterion in both DAGs and ADMGs, it requires extensive domain knowledge that is often not available. Thus, we have demonstrated that there is a gap in the literature for sound non-graphical criteria.

This chapter explores a new criterion for identifying a backdoor adjustment set called the Outcome Criterion (OC) which relies solely on variable selection for the outcome variable. Section 4.1 first shows that in the fully observed setting, running variable selection methods on the outcome is sufficient for identifying a valid and optimal backdoor adjustment set. Thus, the backdoor adjustment set will be the set of significant predictors for Y. Section 4.2 then introduces an auxiliary variable in order to falsify or confirm the validity of a proposed backdoor adjustment set in the hidden variable setting. As a result, we can identify a backdoor adjustment set and be confident on whether or not it is a valid set.

4.1 Variable Selection in Fully Observed Settings

As seen in Chapter 3, most current criteria in the unknown graph setting may employ some combination of the significant predictors for the outcome and significant predictors for the treatment. However, this is not always necessary. In fact, if we initially restrict ourselves to the fully observed setting, we notice that using variable selection for the outcome is enough to provide us with a backdoor adjustment set. Not only will it provide a valid adjustment set, but it will provide the optimal adjustment set, as defined in Henckel et al. (2019). We can start with determining what a variable

selection algorithm run for the outcome will produce in a DAG.

Lemma 1. *In a fully observed DAG $\mathcal{G}$, a consistent variable selection procedure on Y will produce the set $\text{pa}_{\mathcal{G}}(Y)$ as the set of best predictors for Y*

Proof. We can use the Local Markov Property from Equation 2.2, and once again note that the outcome is conditionally independent from the rest of the variables in the graph given its parents, since we are working in a pre-outcome graph. We know from Section 2.5 that a consistent variable selection method will asymptotically arrive at the true model of the data. Thus, since Y is independent of all other variables given it's parents, a consistent model selection procedure will remove all variables from the model besides the set $\text{pa}_G(Y)$, since they will be the only significant predictors for Y. □

Lemma 2. *Let O be the optimal adjustment set in $\mathcal{G}$. Then, $\text{pa}_{\mathcal{G}}(Y) = O$*

Proof. We can use the definition of the optimal adjustment set from Henckel et al. (2019), i.e.

$$\mathbf{O}_{\mathcal{G}}(A, Y) = \text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(A, Y,)) \setminus \text{forb}_{\mathcal{G}}(A, Y) \tag{4.1}$$

As a reminder, the set of forbidden nodes is defined as the descendants of all the nodes on the causal path between A and Y in $\mathcal{G}$. Since Assumption 1 specifies that the set of covariates is not composed of any descendants of the outcome, the set of forbidden nodes in our graph will be simply $\{A\}$. (Appendix A addresses the situation in which Assumption 1 is violated, and shows that we can still use the Outcome Criterion to find the optimal adjustment set.) Furthermore, because of Assumption 1, the set $\text{cn}_{\mathcal{G}}(A, Y)$ will just be Y, since there are no nodes on the causal path between A and Y.

Thus, we have that

$$\mathbf{O}_{\mathcal{G}}(A, Y) = \text{pa}_{\mathcal{G}}(Y) \setminus \{A\}$$

□

Using these Lemmas, we can now prove that the set of best predictors for Y is the optimal adjustment set

Theorem 1. *If a consistent variable selection procedure identifies C^Y_{opt} as the best set of predictors for Y in $\mathcal{G}$, then $C^Y_{opt} = \mathbf{O}_{\mathcal{G}}(A, Y)$*

Proof. We know from Lemma 1 that running a consistent variable selection procedure will produce the parents of a variable, and thus $C^Y_{opt} = \text{pa}_{\mathcal{G}}(Y)$. Using Lemma 2, we now see that $C^Y_{opt} = \text{pa}_{\mathcal{G}}(Y) = O$, and thus

$$C^Y_{opt} = \mathbf{O}_{\mathcal{G}}(A, Y) \tag{4.2}$$

□

Thus, when applying variable selection methods to causal inference problems, as least in the fully observed setting, finding the optimal adjustment set is as simple as running variable selection on the outcome. Here we assume that the graph is composed only of non-descendants of the treatment, but check Appendix A for a method on dealing with a graph that contains descendants of A.

4.1.1 On Variable Selection for the Treatment

One great point that could be brought up by researchers, is the question of whether or not using a set of covariates that are predictive of the treatment is a valid adjustment set, and could be used instead of the Outcome Criterion. Indeed, this is the main idea behind IPW and AIPW (Horvitz and Thompson, 1952). Here we must note two important things. Firstly, AIPW is only doubly robust against statistical misspecification, which means that if only one of the AIPW models are misspecified, the estimates should still be accurate. However, if the backdoor set is misspecified, then the estimates will also be biased. The second thing we note is that the set of covariates that are predictive of the treatment is actually a valid adjustment set. However, since we are trying to optimize the bias variance tradeoff, and since the ACE is computed in terms of the outcome, the set of good predictors for the treatment will not be as useful in predicting the outcome as the set of best predictors for the outcome, and thus using a treatment criterion would lead us to have a higher variance. While the set of good predictors for the treatment that would be identified using a consistant variable selection method $\mathcal{M}$ is a valid set, it is not an optimal set as defined in Equation 4.1. We hope that this is reason enough to dissaude researchers from running variable selection for the treatment instead of for the outcome, as it would unnecessarily increase the variance of their estimates.

4.2 Falsification / Confirmation in Hidden Variable Settings

In the previous section I showed that running variable selection on the outcome is sufficient to identify the optimal adjustment set in fully observed settings. However, fully observed settings are rare, and real world applications often involve some form of hidden variables or unmeasured confounders. In the hidden variable setting, the addition of bidirectional edges complicates the identification of the optimal adjustment set. In fact, the Outcome Criterion is not always sound in the presence of hidden variables. This is quite intuitive to see, so we can take a look at a counterexample.

First, a reminder that the Local Markov Property for ADMGs as defined in 2.3 states that

$$X \perp\!\!\!\perp \mathrm{nde}_{\mathcal{G}}(X)/\mathrm{mb}_{\mathcal{G}}(X)|\mathrm{mb}_{\mathcal{G}}(X)$$

where $\mathrm{mb}_{\mathcal{G}}(X)$ is the Markov Blanket of X, which is composed of all of the variables that lie on a bidirected path with X, as well as their parents, excluding X. Due to the Local Markov Property, we know that the consistent variable selection method M would identify the Markov Blanket as the set of significant predictors for a variable.

In Figure 4.2, we can see that a variable selection method for Y would identify the set $\text{mb}_{\mathcal{G}}(Y) = \{C_1, C_2, A\}$ as significant predictors for Y. If we were to use $\{C_1, C_2\}$ as a backdoor adjustment set, then we would open up the collider at C_1 and at C_2, and thus have an unblocked non-causal path between A and Y, causing our final estimate to be biased.

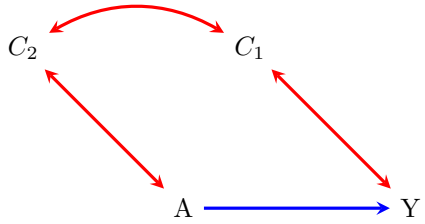


Figure 4.1: A hypothetical counterexample graph for the outcome criterion

Notice that this is the same counterexample as the one in Figure 3.3. Thus, the simple Outcome Criterion is unreliable when dealing with a hidden variable setting.

Before we attempt to modify the Outcome Criterion for it to function in the hidden variable setting, we first must be able to identify when the Outcome Criterion fails in practice. This was one of the pitfalls of the Shpitser-VanderWeele Criterion from Section 3.4, since researchers would be unable to tell whether or not the backdoor adjustment set they have is a valid set. We thus face the issue of falsification / confirmation, or being able to confirm whether a set is a valid set.

In order to be able to test the validity of our backdoor adjustment set, we need to introduce another variable to the problem. We thus define an auxiliary variable Z, which is a pre-treatment variable that shares an edge with the treatment, and potentially the other covariates, but does not share an edge with the outcome or the unmeasured variables. Note that this is a similar setup to that of the instrumental variable used in Instrumental Variable Methods (Wright et al., 1928; Angrist et al., 1996; Angrist and Krueger, 2001).

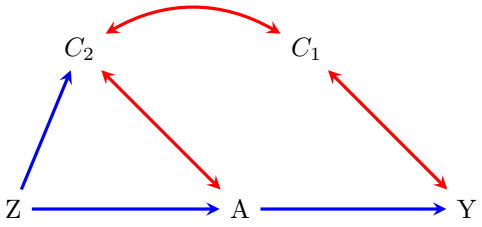


Figure 4.2: A hypothetical counterexample graph for the outcome criterion

Using the presence of a valid auxiliary variable Z in our graph, we can falsify or confirm the validity of the backdoor adjustment set identified using the Outcome Criterion [1]. We first specify the setting we are working in. We have a graph $\mathcal{G}$ with some treatment A and an outcome Y. As a reminder, Assumption 1 states that the $C \cap \text{de}_{\mathcal{G}}(A) = \emptyset$, and thus we have a set of pre-treatment

[1] Note that it is impossible to test whether a meets the definition of an auxiliary variable using real world data. The only time we would be able to do so is when we could use the frontdoor criterion, in which case it would be easiest to just use the frontdoor criterion.

covariates C. Assumption 2 reminds us that we are using a variable selection technique $\mathcal{M}$ which is consistent as defined in 2.5. Finally, we have our new auxiliary variable assumption:

Assumption 3. There exists some variable Z s.t. Z has an an edge into A, and Z does not share any edges with Y or any unmeasured variable U.

Once again, assumption 3 defines the existence of an auxiliary variable in the graph, which allows us to perform more complex computations, and will not only allow us to confirm the validity of a backdoor adjustment set, but will also be used in Chapter 5 to create a doubly robust estimation function which can recover the true causal effect even when the backdoor adjustment set is invalid.

From the definition of the auxiliary variable, there must be an incoming edge into the treatment A from Z in $\mathcal{G}$. Since all the non-causal paths between A and Y must start with an edge incoming into A due to assumption 1, this means that A forms a collider between Z and all non-causal paths between A and Y. Thus, if we were to condition on A, we would open up the path between A and the non-causal paths between the treatment and the outcome, while blocking the causal path between the treatment and the outcome. This means that in order to block any association between Z and Y in the graph where the causal path between A and Y has been removed, we must block the non-causal paths between A and Y, which can be blocked using a valid backdoor adjustment set. Thus, we see that Z is conditionally independent of Y when conditioned on A and a valid backdoor adjustment set. An example of an auxiliary variable that we could use in the initial example from Chapter 1 could be the variable Racial Makeup. In Massachusetts, the proportion of non-white students is a determining factor in the amount of school funding a school receives, thus meaning there is an edge from Racial Makeup to School Funding. If we assume that there is no dierect edge from Racial Makeup to Student Scores (i.e. that all of the correlation between the two is mediated through other variables), then this would be a good auxiliary variable. In general, a good auxiliary variable is one which has a clear and obvious link to the treatment. I now define that an auxiliary variable can be used to falsify or confirm the validity of an adjustment set in Theorem 2

Theorem 2. *Given Assumptions 1, 2, and 3, let $\mathcal{G}$ be an ADMG with treatment A, outcome Y, valid auxiliary variable Z and covariates C. Then, a set C_{adj} is said to be a valid backdoor adjustment set between A and Y iff $Z \perp\!\!\!\perp Y | A, C_{adj}$.*

Proof. $\Rightarrow$ Assume first that C_{adj} is a valid backdoor adjustment set for treatment A and outcome Y. This implies that $A \perp\!\!\!\perp Y | C_{adj}$ in the graph $\mathcal{G}_{null}$ where there is no causal path between A and Y, according to the definition of a backdoor set defined in Section 2.3. This means that all of the non-causal paths between A and Y will be blocked by C_{adj}. Since conditioning on A will block any causal path between Z and Y in $\mathcal{G}$, this means that the only unblocked paths between Z and Y after conditioning on A would be the non-causal paths between A and Y. Since additionally conditioning on C_{adj} would block all non-causal paths between A and Y, and since this would not unblock any causal paths, as C_{adj} is a valid backdoor adjustment set, then conditioning on both A and C_{adj} would block all paths between Z and Y. Thus, if C_{adj} is a valid adjustment set then $Z \perp\!\!\!\perp Y | A, C_{adj}$.

$\Leftarrow$ Now assume that $Z \perp\!\!\!\perp Y | A, C_{adj}$. This implies that there are no unblocked paths between Z and Y. By the definition of the Auxiliary variable, any causal path between Z and Y flows through

A, as Z does not contain any other edges to variables in the graph. This means that conditioning on A blocks all the causal paths between Z and Y in $\mathcal{G}$. Furthermore, conditioning on A opens a collider between Z and all non-causal paths between A and Y. Thus, conditioning on C_{adj} must block all non-causal paths between A and Y in order for the statement $Z \perp\!\!\!\perp Y|A, C_{adj}$ to hold. Since C_{adj} blocks all non-causal paths between A and Y, C_{adj} forms a valid backdoor adjustment set between A and Y. Thus, $Z \perp\!\!\!\perp Y|A, C_{adj}$ implies that C_{adj} is a valid adjustment set for A and Y. □

Using Theorem 2, we can easily see that if we were to run a variable selection method to determine whether Z was a significant predictor for Y given A and C_{adj}, then a consistent variable selection method would determine that Z is a significant predictor of Y iff C_{adj} is not a valid backdoor adjustment set. This is because a variable X should be determined to be a significant predictor of a variable Y given another set C only iff $X \not\perp\!\!\!\perp Y|C$, since we are using a consistent variable selection method $\mathcal{M}$. Thus, by running this final variable selection method on Y, we can determine whether or not C_{adj} is a valid backdoor adjustment set by checking if Z is a significant predictor for Y. This allows us to either falsify or confirm the validity of the backdoor adjustment set determined using the Outcome Criterion defined in Section 4.1. If our backdoor adjustment set is confirmed in this way, we can be sure that our estimate for the ACE using this backdoor adjustment set is an unbiased one. In the case that this method falsifies our backdoor adjustment set, in Chapter 5 I define a doubly robust method for unbiased the estimate for the ACE using this backdoor adjustment set. This method for confirmation and the Outcome Criterion defined in this chapter are empirically evaluated through experimentation using synthetic data in Chapter 6.

While it may seem as though including an auxiliary variable is contrary to the goal of a criterion for 'unknown graphs', this addition is a necessary addition to make in order for us to be more sure of our analysis. Determining an auxiliary variable is a much simpler task than learning an entire causal graph, and thus even though some may see an auxiliary variable as violating the postulate of an 'unknown graph', this single additional piece of information allows us much more leverage in being able to confirm the validity of sets, and as shown in the next chapter, find an unbiased estimate for the causal effect even in the presence of unmeasured confounders.

Chapter 5

Doubly Robust Estimation Using Augmented IV

Now that we have shown a sound method for identifying the optimal backdoor adjustment set in DAGs, and a method for confirming / falsifying the validity of the adjustment set in the case of confounding, we need to explore a way for us to unbias our results in the presence of unmeasured confounding. We will now explore scenarios in which there may not exist any valid backdoor adjustment set. We present a doubly robust functional for estimating the causal effect, which means that it recovers the true causal effect when either there exists a valid backdoor adjustment set which can be identified using the outcome criterion, or there exists and auxiliary variable that can be used to overcome unmeasured confounding by treating it as an instrumental variable (Lousdal, 2018). This is an important property in an estimator since it allows us to recover an effect even when a backdoor set is miss-specified. Note that the work in this chapter and the proofs in this chapter rely on continuous variables, specifically continuity in the treatment. In this situation, the true ACE would thus be the change in Y caused by a unit change in A. A binary treatment would introduce complexities in the relationship between Z and A, and a potential area of future work would be to explore how to reframe the functional to work in a binary case.

5.1 Augmented IV

In Chapter 4, we showed that it is possible for the new outcome criterion to produce an invalid adjustment set if unmeasured confounders are present in the graph. We previously leveraged the relationship between the auxiliary variable and the rest of the graph in order to detect such an occurrence. However, we can continue to leverage the auxiliary variable to recover the true causal effect from our biased estimate. Let us take a look at an example graph, shown in Figure 5.1. In this example, the auxiliary variable as defined in Section 4.2 is Z, our treatment is A, our outcome is Y, our set of covariates is $\{C_1\}$, and U represents an unmeasured variable in our graph.

Before we move on to the rest of the chapter, a quick notational definition. In this chapter, the

notation $\beta_{ZY.A}$ refers to the coefficient for Z in the regression on Y given A. Even though this notation is not widely used in the rest of the field, it's use here is to ease comprehension, since we will be using different coefficients from different regressions. Thus, once again, the first term in the subset β is the term the coefficient is for, the second term is the term the regression is run for, and all the terms after the period are the terms which we are conditioning on in the regression.

We notice how running our consistent variable selection method on Y will produce $\{C_1, A, Z\}$ as the set of significant predictors for Y, meaning that C^Y_{opt} will be the adjustment set identified by the new outcome criterion. Since Z is also included in the significant predictors of Y, we however know that the set identified is not a valid set, and there is still some confounding correlation including in our estimate. If we were to somehow isolate this indirect associational relationship flowing through U, we could remove it from our estimate and thus 'unbias' our result.

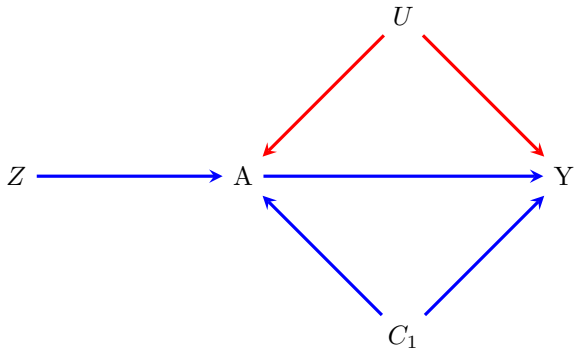


Figure 5.1: An example graph for the unbiasing functional

We first imagine that all the relationships in this figure are given by a linear structural equation model, and we collapse U into a single edge, leaving us with Figure 5.2. We can then label all of the edges on the graph with coefficients. Each coefficient accompanying a direct edge corresponds to the direct causal effect of that variable on its child, ex. δ represents the direct causal effect of A on Y. Each coefficients accompanying a bidirected edge corresponds to the covariance between the error terms of the two vertices. Thus, γ corresponds to the covariance between the error terms ϵ_A and ϵ_Y due to the unmeasured confounder U. From Figure 5.2 we see that the true causal effect here is defined as δ.

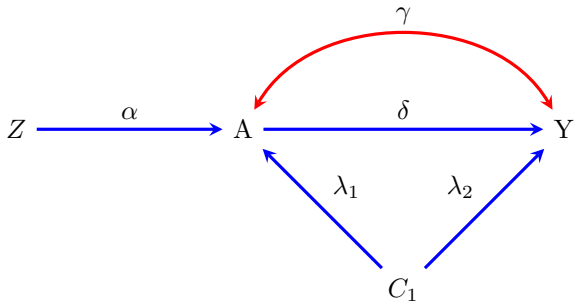


Figure 5.2: An example graph for the unbiasing functional, with corresponding edges labeled

For the purposes of demonstration, under the assumption of linearity, and assuming that Z, A, Y, and C_1 are standardized normal variables, we can now apply Wright's rules of path analysis (Wright et al., 1928) to identify the bias in our backdoor estimate when using C_1 as the adjustment set. The backdoor functional ACE_{bdoor}, which only conditions on C_1, is then the association between A and Y through both the $A \to Y$ edge as well as the $A \leftrightarrow Y$ edge, since conditioning on C_1 blocks association through the $A \leftarrow C_1 \to Y$. Therefore, $ACE_{bdoor} = \delta + \gamma$, that is, it is the true effect plus some confounding U.

In order to isolate γ from δ, we can try estimating Y with Z using a linear regression, while conditioning on the set $\{A, C^Y_{opt}\}$. By conditioning on A, we block the causal path from A to Y, and open up the colliders to the non-causal paths. Conditioning on C_1 blocks the path between C_1 and Y, leaving only the association that flow through A and the bidirected edge. We note that in path analysis, when conditioning on a collider, we multiply the downstream association by a negative, and thus our coefficient for Z in this linear regression would be the relationship between A and Z times the negative association between A and Y through the unmeasured confounder. That is, our coefficient for Z in a linear regression model for $E[Y|A, C^Y_{opt}, Z]$ will be $\beta_{ZY.AC^Y_{opt}} = -\gamma * \alpha$. Therefore, to isolate γ, we need to find the value of α, which is just the relationship between Z and A. We can run another linear regression for A conditioning on C^Y_{opt} and Z, and find an estimate $\beta_{ZA.C^Y_{opt}} = \alpha$, which is the association between Z and A given the adjustment set.

Using all these pieces, we can now 'unbias' our estimate ACE_{bdoor}, by noting that

$$ACE_{bdoor} + \frac{\beta_{ZY.AC^Y_{opt}}}{\beta_{ZA.C^Y_{opt}}} = \gamma + \delta + \frac{-\gamma * \alpha}{\alpha} = \delta$$

.

Here, ACE_{bdoor} is the causal effect estimated using the backdoor adjustment set identified with the outcome criterion, $\beta_{ZY.AC^Y_{opt}}$ is the linear estimate for the coefficient for Z on Y given A and C^Y_{opt}, and $\beta_{ZA.C^Y_{opt}}$ is the linear estimate for the coefficient of Z on A given C^Y_{opt}. We call this formula the 'Augmented IV', since in Section 5.2 we show that it is equivalent to the IV formula in the linear case, but is augmented with backdoor to improve variance.

We walked through a specific example of the unbiasing functional in this section, but in the next section we will prove that this functional is doubly robust functional under assumptions 1-3.

5.2 Double Robustness

Once again, we assume that Assumptions 1-3 hold. Additionally, as mentioned in the introduction to this chapter, we require the treatment to be continuous, and thus we have the following additional assumption:

Assumption 4. The treatment A is a continuous variable over some range $[a, b]$, where $a, b \in \mathbb{R}$

We can now define a theorem over our the Augmented IV from Section 5.1

Theorem 3. *Under Assumptions 1-4, the Augmented IV (shown below) will recover the true causal*

effect in a causal graph $\mathcal{G}$ if C^Y_{opt} is a valid backdoor adjustment set, or if we assume a linear relationship between the variables in the graph.

$$ACE_{adj} = ACE_{bdoor} + \frac{\beta_{ZY.AC^Y_{opt}}}{\beta_{ZA.C^Y_{opt}}} \tag{5.1}$$

Proof. First we examine the statement that the Augmented IV will recover the true causal effect if the identified adjustment set is a valid backdoor adjustment set. We know that if C^Y_{opt} is a valid backdoor adjustment set, then the ACE estimate ACE_{bdoor} computed using this set will be a valid estimate for the true causal effect. This means that our second term in the functional should become 0. We notice that in a graph $\mathcal{G}$, if C^Y_{opt} is a valid adjustment set, then it blocks all non-causal paths between A and Y. When computing our estimate $\beta_{ZY.AC^Y_{opt}}$, we know that conditioning on A will block the causal path between A and Y, and allow association to flow to all of the non-causal paths between A and Y. Since C^Y_{opt} blocks all non-causal paths between A and Y, this means that all paths between A and Y will be blocked. Since we assume that Z is a valid auxiliary variable, it does not share an edge with the outcome, and thus there are no open paths between Z and Y. Thus, the estimate for $\beta_{ZY.AC^Y_{opt}}$ will asymptotically converge to 0. Thus, the entire second term of the functional $\frac{\beta_{ZY.AC^Y_{opt}}}{\beta_{ZA.C^Y_{opt}}}$ will converge to zero, and thus we have the estimate $ACE_{adj} = ACE_{bdoor}$, which we know is a valid estimate for the true causal effect.

Now we can examine the scenario in which the linearity assumption holds. We know that this is one of the assumptions for the Instrumental Variable (Lousdal, 2018), and the second assumption for Instrumental Variables also holds, since Z is a valid auxiliary variable. Thus, since all of the IV assumptions hold, the IV should recover the true causal effect. We can show that the Augmented IV is equivalent to the IV, and thus that the Augmented IV also recovers the true causal effect. We know that the IV functional can be written in terms of covariances (Cameron and Trivedi, 2005), such that

$$IV = \frac{Cov(Y, Z)}{Cov(A, Z)} \tag{5.2}$$

Since we are conditioning on the set C^Y_{opt}, we can use the conditional covariance and use

$$IV = \frac{Cov(Y, Z|C^Y_{opt})}{Cov(A, Z|C^Y_{opt})} \tag{5.3}$$

Firstly, we note that Y, as estimated using out backdoor adjustment set, can be written in the form

$$Y = \beta_0 + \beta_1 * A + \beta_2 * Z + \beta_3 * C^Y_{opt} + \epsilon \tag{5.4}$$

where using our notation we define $\beta_1 = \beta_{AY.ZC^Y_{opt}}$, $\beta_2 = \beta_{ZY.AC^Y_{opt}}$, and $\beta_3 = \beta_{C^Y_{opt}Y.ZA}$.

Taking the covariance of all the terms with respect to the auxiliary variable Z and conditioned on C^Y_{opt} we get

$$\begin{aligned} Cov(Y, Z|C^Y_{opt}) = Cov(\beta_0, Z|C^Y_{opt}) + Cov(\beta_1 * A, Z|C^Y_{opt}) \\ + Cov(\beta_2 * Z, Z|C^Y_{opt}) + Cov(\beta_3 * C^Y_{opt}, Z|C^Y_{opt}) + Cov(\epsilon, Z|C^Y_{opt}) \end{aligned}$$

We can simplify this equation using our knowledge that the error term ϵ will have a covariance of 0 with Z since that is all the noise that is not explained by A, C^Y_{opt}, or Z, we also know that $Cov(\beta_0, Z|C^Y_{opt}) = 0$, that we can move all of the coefficients outside of the covariances, and that $Cov(Z, Z|C^Y_{opt}) = Var(Z|C^Y_{opt})$, and finally that $Cov(C^Y_{opt}, Z|C^Y_{opt})$ from the definition of covariance. Thus, we simplify to get

$$Cov(Y, Z|C^Y_{opt}) = \beta_1 * Cov(A, Z|C^Y_{opt}) + \beta_2 * Var(Z|C^Y_{opt}) \tag{5.5}$$

We can perform a similar operation on A as estimated in our functional using Z and $C^Y_o pt$, and thus

$$A = \beta_3 + \beta_4 * Z + \beta_5 * C^Y_{opt} + \epsilon \tag{5.6}$$

where using our notation once again we define $\beta_4 = \beta_{ZY.C^Y_{opt}}$, $\beta_5 = \beta_{C^Y_{opt}Y.Z}$.

Once again we take the covariance of all terms with respect to Z conditioned on C^Y_{opt}, and simplifying we get

$$Cov(A, Z|C^Y_{opt}) = \beta_4 * Var(Z|C^Y_{opt}) \tag{5.7}$$

We can plug both of these into the IV definition from Equation 5.3 to get

$$IV = \frac{\beta_1 * Cov(A, Z|C^Y_{opt})}{Cov(A, Z|C^Y_{opt}} + \frac{\beta_2 * Var(Z|C^Y_{opt})}{\beta_4 * Var(Z|C^Y_{opt})} \tag{5.8}$$

$$= \beta_1 + \frac{\beta_2}{\beta_4} \tag{5.9}$$

Noting once again that $\beta_1 = \beta_{AY.ZC^Y_{opt}}$, $\beta_2 = \beta_{ZY.AC^Y_{opt}}$ and $\beta_4 = \beta_{ZY.C^Y_{opt}}$, we see that

$$IV = \beta_{AY.ZC^Y_{opt}} + \frac{\beta_{ZY.AC^Y_{opt}}}{\beta_{ZY.C^Y_{opt}}} \tag{5.10}$$

$$= ACE_{bdoor} + \frac{\beta_{ZY.AC^Y_{opt}}}{\beta_{ZY.C^Y_{opt}}} \tag{5.11}$$

We thus see that under linearity and the IV assumptions, the IV functional is equivalent to the Augmented IV, which means the Augmented IV should recover the true causal effect.

Thus, we see that in the general case as well, our Augmented IV will recover the true causal effect if the specified backdoor adjustment set is valid, or if the variables follow a linear relationship. □

We therefore see that our Augmented IV functional is doubly robust, and will recover the true causal effect if either of the assumptions are met. This method differs from the traditional IV method due to its double robustness, as well as the inclusion of a backdoor adjustment set in order to find a first estimate of the causal effect. In Chapter 6 we explore a few simulations using this unbiasing functional, and empirically evaluate the double robustness claim. We also compare the estimates to the previously defined methods, and show that it provides a desirable variance and low bias.

Chapter 6

Simulations

In the previous chapters, we have outlined the issues with currently used criteria in the unknown graph setting, we have proposed our own method that works in the measured confounding case, and a way to unbias the method in the presence of unmeasured confounding. In Chapter 3, we reviewed previous criteria, and run a few simulations to show their limits. In Section 6.1, we will apply our method to the graphs from Chapter 3, and show that our method returns an unbiased estimate of the causal effect using simulated data. In Section 6.2, we test the falsification using an auxiliary variable from Theorem 2, and show that in simulated data it identifies valid backdoor adjustment sets. Finally, in Section 6.3 we evaluate the doubly robust Augmented IV using simulated data, and show that it recovers the true causal effect as intended.

6.1 Evaluating Method on Previous Graphs

We begin our simulations by evaluating the Outcome Criterion and Augmented IV on the graphs that we have previously used to analyze the other criteria currently used in the literature. We can start with examing how the outcome criterion performs on the graph from 3.1, which is reproduced in 6.1 below.

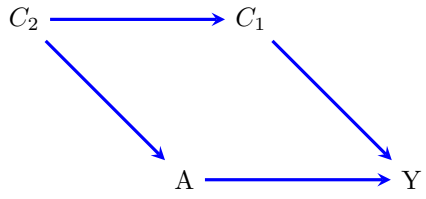


Figure 6.1: The graph from Figure 3.1

Theoretically, the outcome criterion should produce $\{C_1\}$ as the backdoor adjustment set, which in this case is the optimal adjustment set. To implement the Outcome Criterion, we make use of the MIC variable selection method to find the set of best predictors for Y. This is another of the advantages of the Outcome Criterion, since the MIC carries out the variable selection while running

a regression with a penalization term, which means that our MIC model can directly provide us with the estimate for the ACE by using the coefficient for A in the model $Y|A, C$, when assuming a linear relationship. Once again, in our simulated data that ACE is specified as 2, and the values for the rest of the edges in the graph are shown using the structural equations in Equations 3.1-3.4. The resulting graph of the ACE estimates using the Outcome Criterion is shown in Figure 6.2.

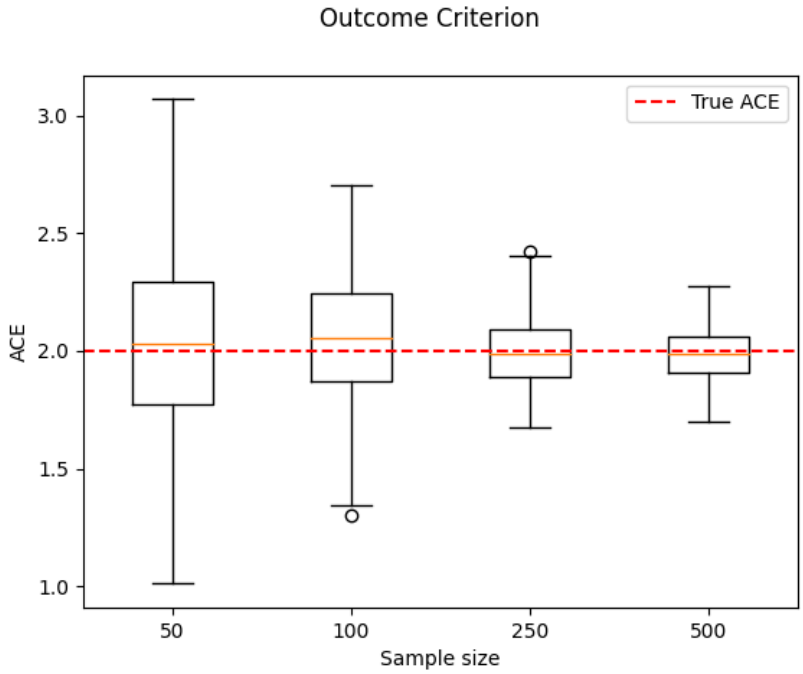


Figure 6.2: Outcome Criterion method for the graph in Figure 6.1

We notice from Figure 6.2 that the estimates are all centered around the true causal effect, meaning that in this situation the Outcome Criterion is unbiased. This is as expected, since we proved that the outcome criterion is unbiased for all DAGs. The boxplots here show the estimates of the ACE for the 100 different trials over that sample size. We notice that the tails on the boxplots for this graph are around the same size as the tails for this graph using the Union Criterion, shown in 3.4i. Although we would expect the outcome criterion to have smaller tails since the conditioning set is smaller, this difference becomes more and more noticeable in larger graphs with more covariates.

We can then move on to testing the Outcome Criterion on the graph from Figure 3.3, which is reproduced again below as Figure 6.3.

Once again, the ACE is set to be 2, the values for rest of the edges are defined in the structural equations in Equations 3.5-3.11, and we are running 100 trials each at sample sizes of 50, 100, 250, and 500. The implementation for the Outcome Criterion is the same as above. In this case, since the graph is a DAG, the Union Criterion would identify the set $\{C_1, C_2\}$ as the adjustment set, which we already know from Section 3.3 is not a valid backdoor adjustment set. We can see the estimates for the ACE using the outcome criterion on this graph in Figure 6.4.

We quickly notice that the estimate for the ACE are slightly biased, and the average computed

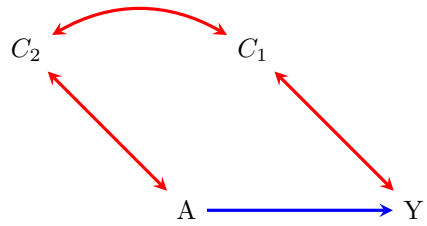


Figure 6.3: The counterexample for the Union Criterion from Figure 3.3

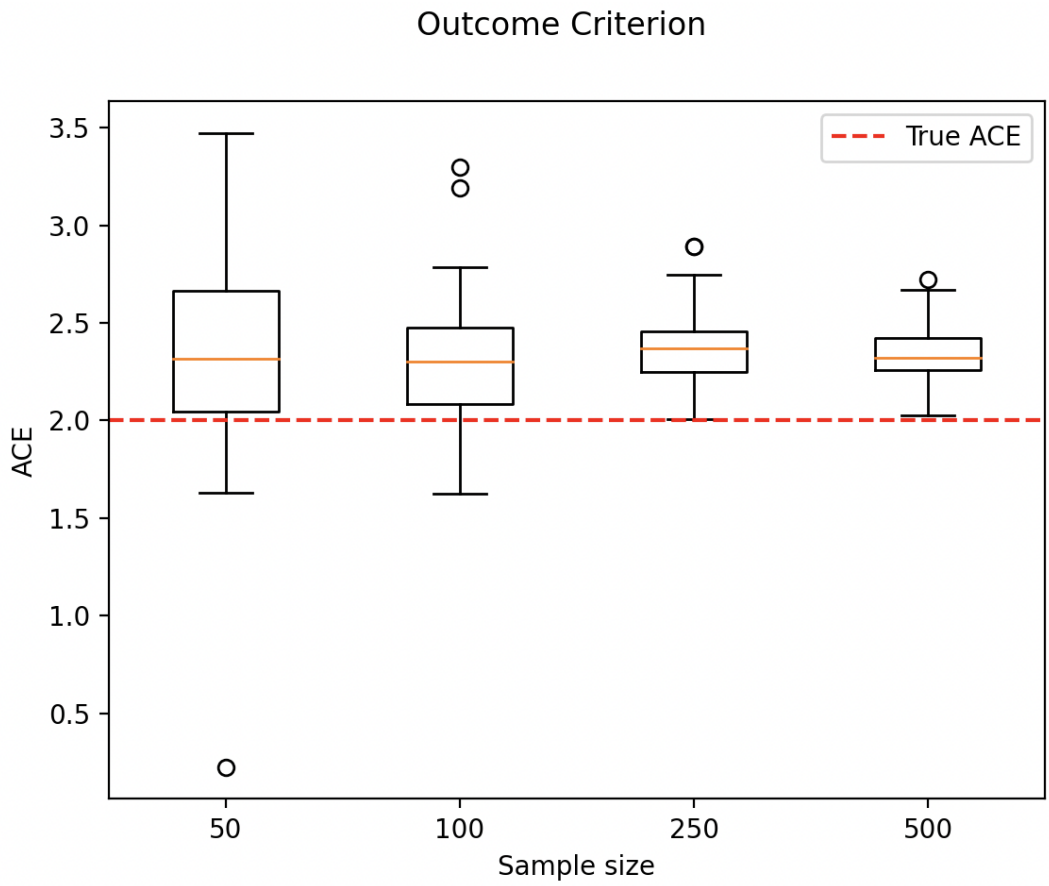


Figure 6.4: Computed ACE for simulated data based on DGPs of Figure 6.3

ACE for all of the sample sizes is about 2.45. However, as compared to the Union Criterion, we can use an auxiliary variable in order to determine whether or not to use our estimate for the ACE, and if the adjustment set is invalid, we can then employ the Augmented IV in order to recover an unbiased estimate. Thus, we add a variable Z into the graph, which for this simulation we will have pointing just into A, and carrying with it a causal effect of 2 on A. We can run a regression including Z, and we notice that the average coefficient for Z in the regression on $Y|A, C, Z$ for each sample size respectively is -0.15, -0.18, -0.19, -.24. We determine the threshold for whether or nto a

variable is non-zero as 0.1, which is derived partially from the literature, and partially from initial testing of the MIC implementation. We thus see that the average coefficient for Z in this graph is significant, signifying that the adjustment set is invalid and our estimates are biased. Note that we just looked at the average Z coefficient here since we evaluate the efficacy of using an auxiliary variable to falsify the Outcome Criterion in Section 6.2, and thus an in depth analysis here is not necessary. Thus, we can employ our Augmented IV. The Augmented IV is implemented by running a regression using the MIC on $Y|A, C, Z$, and extracting the estimate for the biased causal effect using the coefficient for A, and extracting the coefficient for Z as our $\beta_{ZY.AC^Y_{opt}}$. Finally, we will run a regression using on $A|Z, C^Y_{opt}$ to estimate the coefficient $\beta_{Z.AC^Y_{opt}}$, and then plug these coefficients into our Augmented IV from Equation 5.1. We once again apply the Augmented IV to the 100 trials for each sample size, and the resulting estimate for the ACE is shown in Figure 6.5.

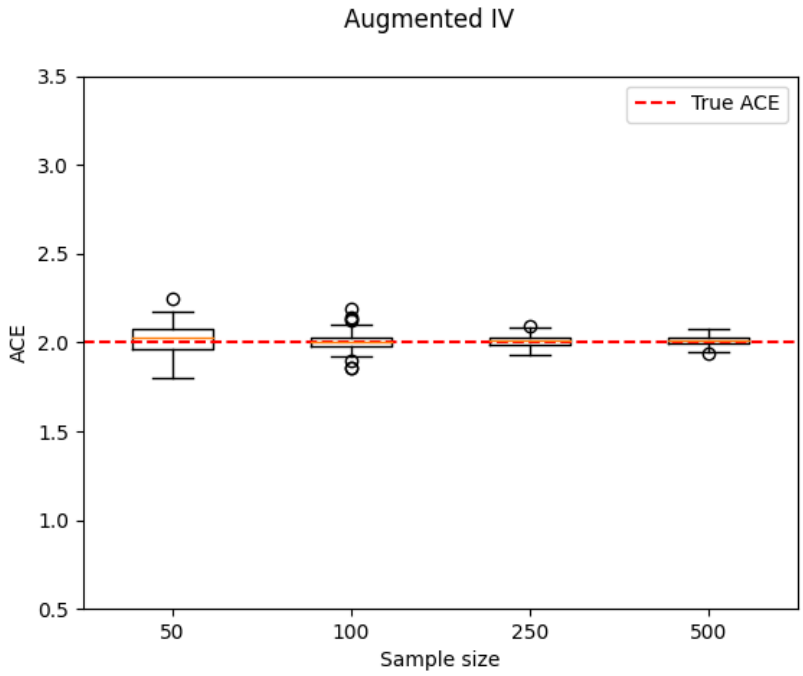
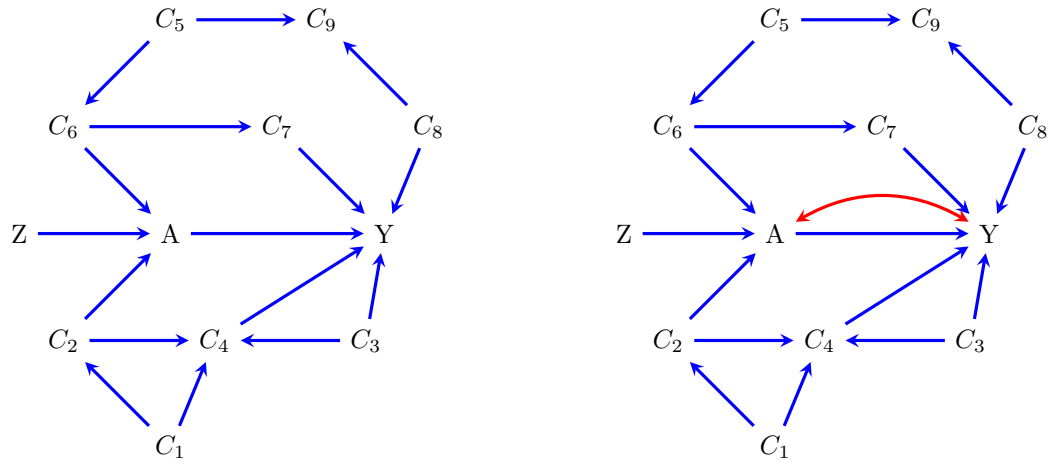


Figure 6.5: Computed ACE using the Augmented IV on the graph from Figure 6.3

We see that the Augmented IV does incredibly well at estimating the true causal effect, and in fact produces very narrow tails on the boxplot, showing that for at least this graph, it performs very accurately. We will do a more in depth evaluation of the Augmented IV in Section 6.3, but for now we notice that it is unbiased in this specific ADMG. Thus, we can already see that our method is an improvement from the previous methods, since using the methodology laid out in this thesis we produce unbiased estimates of the ACE in both graphs, while the Union and Intersection Criterion were both only unbiased in 1 out of the 2 graphs.

6.2 Evaluating Falsification of Method

Having tested the outcome criterion and Augmented IV on our previous graphs, we can now move on to evaluating the effectiveness of using an auxiliary variable to falsify or confirm the backdoor adjustment set identified by the Outcome Criterion. To do this, we can create two similar graphs, such that in one of them the outcome criterion should theoretically produce a valid adjustment set, and then modify it so that in the second graph that outcome criterion will not be able to produce a valid adjustment set. We do this in Figure 6.6, where the only difference between the two graphs is the bidirected edge from A to Y in the second graph.



(i) Graph without unmeasured confounder

(ii) Graph with unmeasured confounder

Figure 6.6: Computed ACE for simulated data based on DGPs of previous graphs

To test the utility of the auxiliary variable, we can run the outcome criterion on each graph over 100 trials, and then see if the auxiliary variable will either falsify or confirm the adjustment set identified by the outcome criterion. We do this over sample size of 50, 100, 250, and 500, and then compute the accuracy, precision, and F1 score of using the auxiliary variable, which is then shown in Table 6.1.

Sample Size	Precision	Recall	F1 Score
50	0.63	0.58	0.6
100	0.77	0.69	0.73
250	0.92	0.92	0.92
500	0.989	0.98	0.98

Table 6.1: Performance of the Auxiliary Variable in determining validity of the adjustment set

From the table above, we see that every single one of the performance measures improves as the sample size increases, which makes sense since our method should asymptotically correctly confirm the adjustment set. We also notice that starting at sample size 100, the F1 score is relatively high, and once we reach 250 it is above 90%. Since there is no previous literature which attempts to

confirm or falsify the validity of a backdoor adjustment set, there is nothing to compare our results to, though noting that a precision and recall of over 90% at 250 samples is a very desirable metric.

6.3 Evaluating Augmented IV

We can now move on to evaluating the Augmented IV introduced in Chapter 5. In order to evaluate the effectiveness of the Augmented IV, we can compare it's performance to regular IV estimation. We will test the two methods on 4 different graphs. We evaluate the methods on a graph where both the theoretical backdoor adjustment set identified by the outcome criterion is valid, and the relationship between the variables is linear. We then change the relationships between the variables from the previous graphs so that the relationships cease to be linear, while the adjustment set should theoretically be still valid. We then look at another graph where the identified backdoor set is invalid, while the relationship between all of the variables remains linear. Finally, we modify this graph to create a graph where both the set is invalid, and the relationship between variables is non-linear[1].

The first graph we will use is shown in Figure 6.7. This graph only contains linear relationships, and contains a valid backdoor adjustment set that should be identified using the outcome criterion.

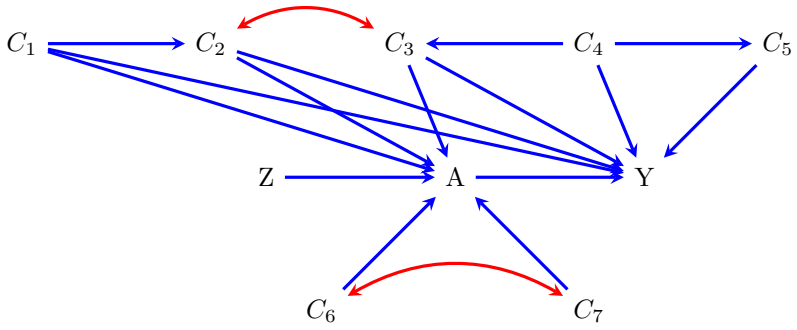


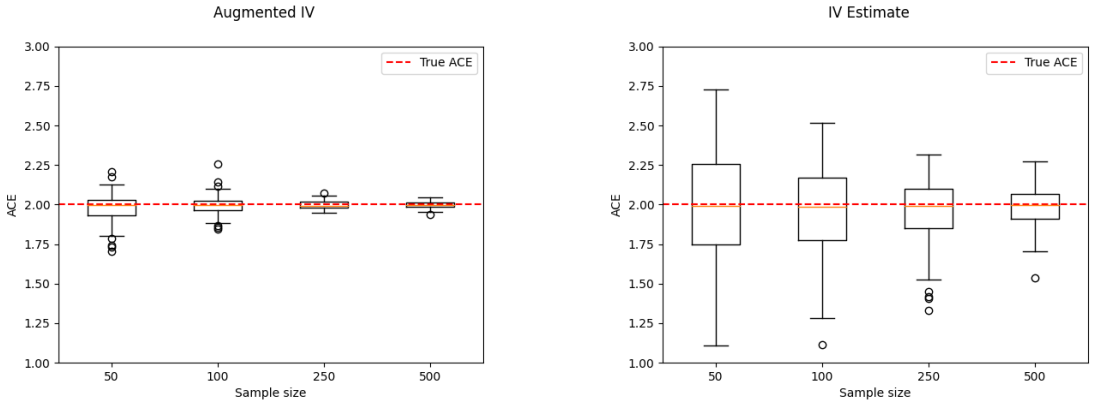
Figure 6.7: Graph with both a valid adjustment set and linearity

Once again, the ACE is set at 2, and we run 100 trials each at sample sizes of 50, 100, 250, 500. The implementation of the Augmented IV is the same as above, and the implementation for IV is done using the covariance definition defined in 5.2. We see the estimates for the ACE for for data generated using the graph from Figure 6.7 in Figure 6.8

We notice that even though both estimates are unbiased, the estimates computed using the Augmented IV variance have a much smaller variance. This follows from the fact that the Augmented IV is calculated by adjusting on a valid adjustment set, which allows our estimate to be much more precise. This is the ideal setting for the Augmented IV.

We can now modify the data generating process so that the relationships between the variables in the graph cease to be linear, while the structure of the graph remains the same. This means that the adjustment set would still be the same, but the linearity assumption would be violated. We

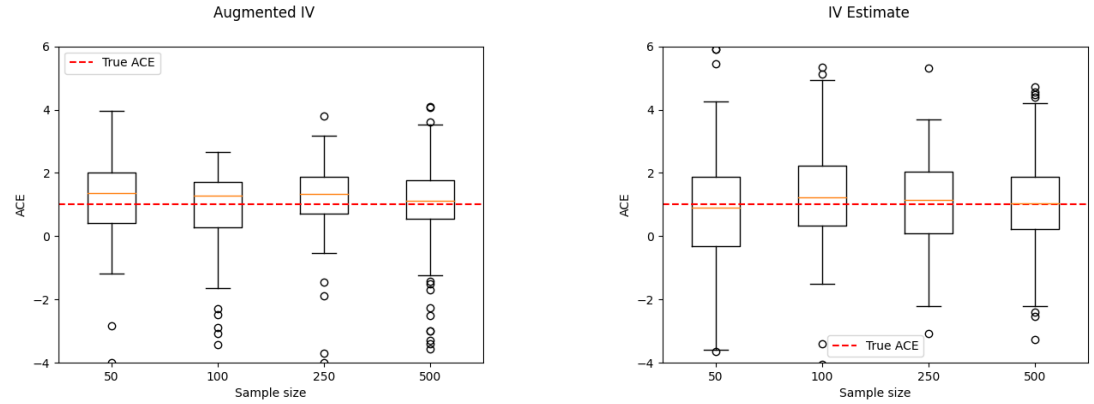
[1]Note that all of the graphs will have Z only pointing to A, since this is a requirement for IV, even though the Augmented IV will work if Z has edges to other covariates in the graph



(i) Estimated ACE using Augmented IV (ii) Estimated ACE using IV

Figure 6.8: Computed ACE for simulated data based on a linear DGPs in Figure 6.7

again simulate 100 trials over sample sizes of 50, 100, 250, and 500. The results are shown in Figure 6.9. Here, because the relationship between the variables is non-linear, the true ACE will be 1, as the formula for Y includes interaction terms between A and the covariates. We see that because there exists a valid backdoor adjustment set, the Augmented IV recovers the true causal effect. We notice though that the IV estimate also recovers the true causal effect, which could be because there is only one interaction term between A and the covariates in the function for Y. However, we see that the estimate using the Augmented IV still have lower variance then the IV estimate.



(i) Estimated ACE using Augmented IV (ii) Estimated ACE using IV

Figure 6.9: Computed ACE for simulated data based on a non-linear DGPs in Figure 6.7

We now move on to the case where there is no valid adjustment set, while the relationship between the variables is linear. This is done in the graph in Figure 6.10, and the resulting ACE estimates are shown in Figure 6.11. Here, since linearity holds, the IV estimate does very well, and since the Augmented IV is just a reformulation of the IV, it also performs well. However, even though there is not a valid adjustment set, including information about the covariates in the model still manages

to improve our estimates, and thus the Augmented IV has a lower variance than the IV estimate.

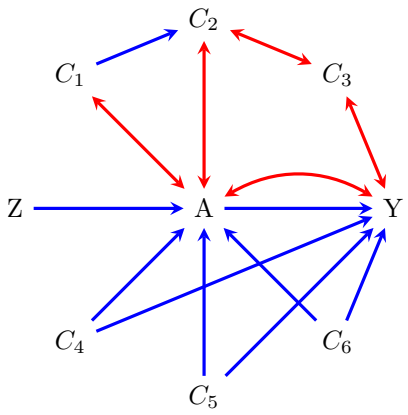
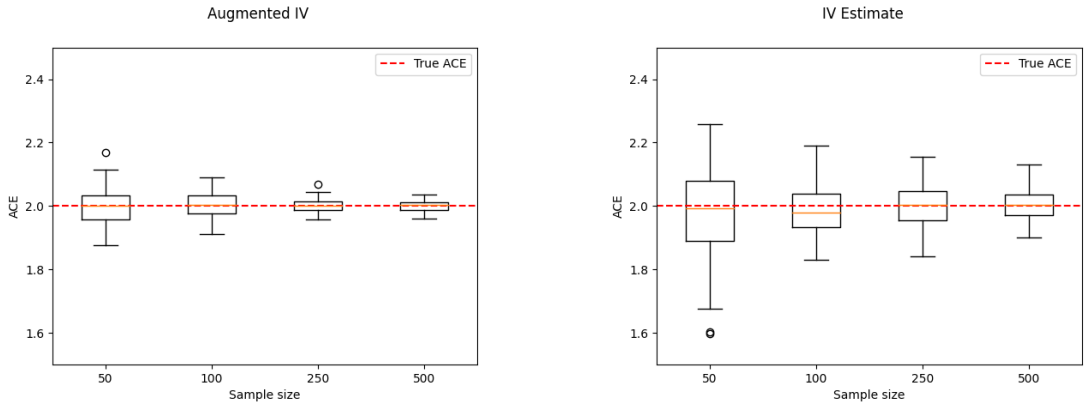


Figure 6.10: Graph without a valid backdoor adjustment set



(i) Estimated ACE using Augmented IV

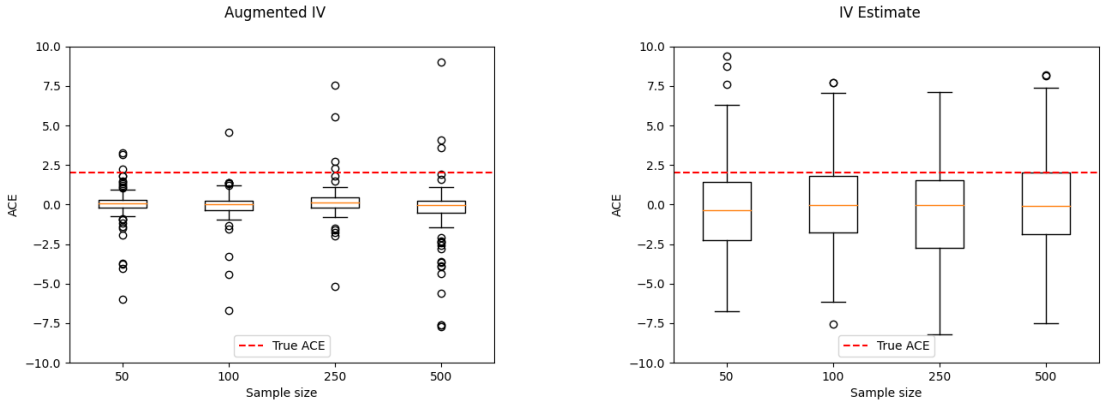
(ii) Estimated ACE using IV

Figure 6.11: Computed ACE for simulated data based on a linear DGPs in Figure 6.10

Finally, we can test the Augmented IV on a graph where there is not valid adjustment set and the data does not follow a linear relationship. We do this by modifying the data generating process from Figure 6.10 to include non-linear terms and non-linear relationships. The resulting ACE estimates are shown in Figure 6.12.

Here, we see that both the IV and Augmented IV are biased, and not centered around the true causal effect. Both the IV and Augmented IV also exhibit a large variance, though the variance of the IV estimate is quite larger than that of the Augmented IV. We note that there are also more outliers for both the Augmented IV and IV that are not pictured in Figure 6.12, as the Augmented IV becomes unstable and can lead to quite drastically large overstimation of the ACE.

From these simulations, we see that the Augmented IV performs as intended, recovering the true casual effect if either there is a valid backdoor adjustment set or the relationship between the variable is linear. We also notice that compared to the regular IV, the Augmented IV exhibits lower



(i) Estimated ACE using Augmented IV

(ii) Estimated ACE using IV

Figure 6.12: Computed ACE for simulated data based on a non-linear DGPs in Figure 6.10

variance in all of the cases explored, showing that reformulating the IV in terms of the coefficients for A and Z and including a covariate set on which to adjust improves the precision of our estimates, and leads to the desired decrease in variance without requiring an increase in bias.

Chapter 7

Conclusions and Future Work

In this thesis, we developed a new criterion for identifying backdoor adjustment sets in the unknown graph setting, as well as a doubly robust Augmented IV functional under the presence of an auxiliary variable that recovers the true causal effect if either the identified backdoor adjustment set is valid, or the relationship between the variables is a linear one. To do so, we first examined previously defined criteria in the literature, and showed that they either fail to perform as expected, or require too much a-priori causal information about the graph. This proves the importance and relevance of this thesis, as these methods are currently in use by many applied researchers, and thus jeopardizing the validity of their results. We then postulate that focusing on significant predictors for the outcome and the treatment is unnecessary, and proved that in DAGs the significant predictors for the outcome constitute an optimal adjustment set. We thus define the new Outcome Criterion, which using a consistent variable selection algorithm $\mathcal{M}$ for the outcome in order to identify the optimal adjustment set. Furthermore, we proved that by using an auxiliary variable that only shares edges with the treatment and covariates, we are able to test the validity of this set and either confirm or falsify it. Finally, in order to recover an unbiased estimate of the true causal effect, we use the auxiliary variable to create an Augmented IV that is doubly robust, and will recover that true causal effect if either the identified backdoor adjustment set is a valid set, or if the relationships between the variables are linear. Finally, we proved that this functional will recover the true causal effect under the assumptions of a continuous treatment variable, a consistent variable selection method $\mathcal{M}$, and a valid auxiliary variable.

Despite these results, there is still plenty of room for future research and modification to the current criterion. The first area of research regards the linearity assumption for the Augmented IV. While it is true that plenty of research is done under the linearity assumptions in fields such as economics and genomics, and that many current methods such as IV still contain the linearity assumption, it is an assumption that is worth exploring. Many problems in causal inference include complex non-linear relationships between variables, and forcing a linearity assumptions limits the uses of the method. This extension can be done through firstly the use of non-linear path analysis (Turner et al., 1961) in order to determine the potential relationships between the variables when the linearity assumption is violated. We can use this analysis to then apply non-parametric mod-

eling techniques and non-parametric variable selection methods to the Augmented IV, and create a reformulation which can recover the true causal effect under double robustness over a less stringent assumption regarding the relationship between the variables. This provides plenty of research and opportunities for expanding the method, as incremental work can be done to reformulate the functional for different relational assumptions.

Using the non-linear path analysis, we can also loosen the restriction on continuous treatments for the Augmented IV. While a lot of causal inference research involves continuous treatments, there is still plenty more which use binary treatments such as economic interventions, policy changes, or population characteristics, and thus our method could be expanded to include these types of problems. This involves exploring the potential relationships between Z and A, and how this relationship is expressed in the associational relationship between variables. We can explore non-parametric modeling techniques, and include these models in the Augmented IV while retaining its doubly robust property.

Finally, another area of potential improvements for the Outcome Criterion is pruning the significant predictor set into a minimal optimal adjustment set. While the current Outcome Criterion is optimal, not all of the significant predictors for Y may be necessary, and thus exploring ways to prune this set through the use of either variable selection methods or other techniques will allow us to shrink the variance of the estimates from the set as much as possible.

Overall, this thesis successfully developed a new criterion for adjustment set selection and a corresponding doubly robust functional for the recovery of the true causal effect in the unknown graph setting. This is an improvement over the previous methods which will not always recover the true causal effect, and thus will allows researchers to be certain in the validity of their results.

Appendix A

Outcome Criterion in Graphs with Forbidden Nodes

This appendix deals with the scenario in which 1 is violated, and thus when there are variables in the covariate set C which are descendants of A. This appendix shows that if we know what these nodes are, we can simply remove them from the graph, and running the outcome criterion on the resulting graph would still form a valid backdoor set in the fully observed setting.

From the Local Markov Property for DAGs defined in Equation 2.2, we see that a variable is conditionally independent of it's non-descendants given its parents. Using Lemma 9 from Tian (2002), we know that the causal effect of Y is identifiable in a graph $\mathcal{G}$ iff that causal effect is identifiable in the graph $\mathcal{G}_{\text{an}}(Y)$. We can thus restrict ourselves to the graph G composed only of the non-descendants of Y, and we can write the Local Markov Property in terms of the outcome, seen in Equation A.1

$$Y \perp\!\!\!\perp \text{nde}_{\mathcal{G}}(Y) \setminus \text{pa}_{\mathcal{G}}(Y) | \text{pa}_{\mathcal{G}}(Y) \tag{A.1}$$

As a reminder, the set of forbidden nodes in the graph $\mathcal{G}$ is defined using a modified version of the definition from Henckel et al. (2019):

$$\text{forb}_{\mathcal{G}}(A, Y) = \text{de}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(A, Y))$$

where $\text{de}_{\mathcal{G}}(X)$ is the set nodes which are descendants of X in $\mathcal{G}$, and $\text{cn}_{\mathcal{G}}(A, Y)$ is the set of nodes which lie on the causal path between A and Y in $\mathcal{G}$.

If we had knowledge of which nodes are a member of the forbidden set, we could just remove them from the graph, and call the resulting graph $\mathcal{G}^*$. Algorithm A.1 shows how the graph would change after the removal of the forbidden nodes. We can use this algorithm to prove that our criterion will still be valid in this resulting graph $\mathcal{G}^*$.

Lemma 3. *If $\mathcal{G}$ is a DAG, then $\mathcal{G}^*$ is also a DAG.*

Algorithm A.1 Remove Forbidden Nodes from $\mathcal{G}$

```
Let τ be a valid topological order for V
Let G* = G
for each node V_i reverse τ do
    if Y ∈ ch_{G*}(V_i) then
        for V_k in pa_{G*}(V_i) do
            Create edge V_k → Y
        end for
    end if
    Remove node V_i from G*
end for
```

Proof. We can move backwards through the topological ordering τ, and apply the algorithm when we encounter a forbidden node V_i. We know that $\text{de}_{\mathcal{G}^*}(V_i) = \emptyset$ or $\{Y\}$, since any descendant of V_i which is not Y would have been removed from $\mathcal{G}^*$ before we consider V_i.

Thus, for each $V_j \in \text{pa}_{\mathcal{G}^*}(V_i)$ we will create an edge $V_j \to Y$. We can now safely remove V_i from G^*, since despite the possibility of the existence of a collider $C_1 \to V_i \leftarrow C_2$, the only descendant of this collider is Y, and thus this collider will never be opened, as is is impossible to control for Y. We therefore safely remove V_i from G^*, and G^* will remains a valid DAG, since there are no bidirected edges.

Therefore, after removing all $V_i \in \text{forb}_{\mathcal{G}}(A, Y,)$, we are left with a graph G^* which is still a valid DAG. □

Lemma 4. *Let O be the optimal adjustment set in $\mathcal{G}$. Then,* $\text{pa}_{\mathcal{G}^*}(Y) = O$

Proof. We can use the definition of the optimal adjustment set from Equation 4.1, which as a reminder is

$$\mathbf{O}_{\mathcal{G}}(A, Y) = \text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(A, Y)) \setminus \text{forb}_{\mathcal{G}}(A, Y)$$

Since $A \in \text{cn}_{\mathcal{G}}(A, Y)$, we know that $\text{cn}_{\mathcal{G}}(A, Y) \setminus A \subset \text{forb}_{\mathcal{G}}(A, Y)$. Following the reverse topological ordering, we once again know that when removing the forbidden nodes using Algorithm A.1, the node V_i will only have Y as a possible child. Thus, when we remove V_i from $\mathcal{G}^*$ using Algorithm A.1, its parents became the parents of Y. Following this procedure for every single forbidden node, we see that $\text{pa}_{\mathcal{G}}(\text{cn}_{\mathcal{G}}(A, Y)) = \text{pa}_{\mathcal{G}^*}(Y)$.

Thus, we have that

$$\mathbf{O}_{\mathcal{G}}(A, Y) = \text{pa}_{\mathcal{G}}(Y) \setminus \text{forb}_{\mathcal{G}}(A, Y)$$

Since we have removed all of the forbidden nodes from $\mathcal{G}$ besides A when creating $\mathcal{G}^*$, we know that $\text{forb}_{\mathcal{G}^*}(A, Y) = \{A\}$, and thus we are left with

$$\mathbf{O}_{\mathcal{G}}(A, Y) = \text{pa}_{\mathcal{G}}(Y) \setminus \{A\}$$

□

Using Lemma 1, we know that running a variable selection procedure $\mathcal{M}$ for Y in the graph $\mathcal{G}^*$

would produce the set $\text{pa}_{\mathcal{G}^*}(Y)$, and thus running our Outcome Criterion would produce the optimal adjustment set in the graph where all of the nodes on the causal path between A and Y are removed.

Therefore, a simple solution to the presence of covariates which are also descendants of A in the graph is to simply remove them from all computations.

Bibliography

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Colombo, D., Maathuis, M. H., et al. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Henckel, L., Perković, E., and Maathuis, M. H. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*.

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.

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging themes in epidemiology*, 15(1):1.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature methods*, 7(4):247–248.

Pearl, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3:121–129.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21(1):7642–7727.

Shen, X., Ma, S., Vemuri, P., and Simon, G. (2020). Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 10(1):2975.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, prediction, and search, volume 81 of. *Lecture notes in statistics*.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Su, X., Wijayasinghe, C. S., Fan, J., and Zhang, Y. (2016). Sparse estimation of cox proportional hazards models via approximated information criteria. *Biometrics*, 72(3):751–759.

Tian, J. (2002). *Studies in causal reasoning and learning*. University of California, Los Angeles.

Turner, M. E., Monroe, R. J., and Lucas, H. L. (1961). Generalized asymptotic regression and non-linear path analysis. *Biometrics*, 17(1):120–143.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413.

Wright, P. G. et al. (1928). Tariff on animal and vegetable oils.

Zeng, J., Gensheimer, M. F., Rubin, D. L., Athey, S., and Shachter, R. D. (2022). Uncovering interpretable potential confounders in electronic medical records. *Nature Communications*, 13(1):1014.