Causation and Intervention

Frederick Eberhardt

Abstract

Accounts of causal discovery have traditionally split into approaches based on passive observational data and approaches based on experimental interventions that take control of (the distribution of) one or more variables. The former includes a vast number of techniques for the inference to causal structure on the basis of statistical features of data, while the latter provides in addition a methodology of how an experiment should be performed, in order to be informative about causal structure. In this thesis, the causal Bayes net framework is used to integrate these two approaches and general guidelines are provided not only of *how* experiments should be performed but also *which* experiments should be performed to discover the causal structure among a potentially large number of random variables. In that sense this thesis aims to extend considerations found in experimental design from single experiments to sequences of experiments. To do so, the thesis provides a precise account of what constitutes an intervention that allows for, but does not necessessitate, a role of agency in interventions. We describe a space of interventions that is broader than standard randomized controlled trials, and explore what implications follow for discovery when different types of interventions are used. Results pertaining to the methodology of causal discovery, its limits, the efficiency of its search strategies and the meta-analysis of experimental results are presented. This thesis analyses the combinatorics of sequences of experiments for causal discovery, ties the discovery problem into a game-theoretic framework and points to some of the (many) difficulties that remain open research questions.

Contents

Thank you!

In 2004 Richard Scheines called me into his office and posed the very simple question to me that had occurred to him in one of his classes: Just how many interventions are necessary and sufficient to discover the causal graph among a set of variables? As far as I know, no-one knew the answer. This thesis is an answer three years later. And it is only partial. The small question developed into a research project that will keep me busy for years to come. This project would not have started without Richard, nor would it have continued without his support and enthusiasm. I have been extraordinarily lucky to have had Richard as an advisor: imagine an advisor of which no better can be imagined – and he is real. No less am I indebted to Clark Glymour. With his relentless criticisms and acute sense of where I just did not have it quite straight yet, Clark pointed me in directions and towards results I do not think I would have reached on my own. In a similar sense, Teddy Seidenfeld contributed enormously to where this thesis stands now. I do not have words to express my gratitude and feel extremely honoured to have had the opportunity to have shared time with all three. They have profoundly shaped my way of thinking about many areas of research. Their knowledgability and their extraordinary sense of the right question to ask will remain a reminder to me of just how much more there is to learn and how much better I will have to become. But the academic side only makes half a PhD. It is the openness, the serious consideration of developing ideas, the shared excitement about results and the heckling and being heckled that I will remember as the good times. I found the Carnegie Mellon Philosophy Department to be a very congenial place, where there was a genuine sense of collaboration and a refusal to adhere to borders of disciplines. Consequently, many other people have been in different ways a direct cause of this work. Among those, I would particularly like to thank (in alphabetical order) Brent Bryan, David Danks, Carlos Guestrin, Kevin Kelly, Joseph Ramsey, Peter Spirtes and Jiji Zhang.

Throughout the thesis I write in the first person plural. This “we” is not royal, but rather a tribute to the very collaborative nature of the work in this thesis. There are many other people, who throughout my life have been indirect causes of this thesis and have supported me in my academic endeavours, in particular (in chronological order) Christian Böttcher, Robert Close and family, Ian Clark, Angela Horwitz and John Worrall. Three other people deserve a mention in this place even though a few lines in the acknowledgements do not do justice to the love and support I have received from them: My parents Ulrike Eberhardt and Steven Shemeld, and my girlfriend Minoli Ratnatunga, the best outcome of my time at university.

The research for this thesis was funded in part by a grant from the McDonnell Foundation.

THANK YOU.

Chapter 1

Introduction

In the most general sense, this thesis attempts to contribute to the literature on causal discovery: Given a set of variables, how should one proceed to discover the causal relations between the variables? The aim is to provide a methodology of discovery that is sensitive to the particular features of causal relations and that is able to identify these features efficiently. Our focus will be primarily on the discovery of causal *structure*, i.e. the qualitative relation that specifies which variable is a direct cause of which other variable. While this separation of structure search and quantitative estimation of the dependencies of effects on their causes (the parameterization) is common, the two aspects are by no means separate endeavours. Knowledge of the true causal structure simplifies the estimation of parameters, while under certain assumptions only differences in the parameters can help discover the causal structure.

The project builds on the representation of causal relations within the framework of causal Bayes nets [43, 32]. In a causal Bayes net the causal structure is represented by a directed acyclic graph and a joint probability distribution over the variables accounts for the dependencies among the variables. Two assumptions, the causal Markov and the causal Faithfulness conditions provide the foundation of this representation. Together they imply a correspondence between the causal separation relations implied by the graph and the indepedence relations present in the joint probability distribution over the graph. As is common in most causal structure search, and certainly within the causal Bayes net framework, this thesis assumes a well defined set of causal variables. This is not to say, that *latent* (unmeasured) variables are assumed away, but rather that the set of given variables are taken to be meaningful causal variables with at least a hypothetical notion of how they can be subject to an intervention.[1]

More specifically, the thesis focuses on *active* search for causal structure. Given a set of causal variables, one can distinguish two ways to learn about causal relations between the variables: One can either obtain so-called *passive-observational* measurements of the variables or one can subject a subset of the variables to an experimental intervention. In the first case, the values of the variables are recorded as they "naturally occur", as, for example, in logitudinal or cross-sectional studies. In the second case, one intervenes on the "natural process" by controlling the (distribution over the values of) some of the variables and records the values of the others. This use of interventions for causal discovery is found in randomized experiments in clinical trials, but in general the notion of interventions on a causal system is much broader than the simple randomization of one or more variables. There are many ways to manipulate a set of variables and in principle there is no reason to think that it takes some agent to do so. The inferences that can be drawn about the causal structure differ between passive observational and interventional data, since the approaches imply different constraints on the set of variables. In both cases however, the causal structure may still be underdetermined, i.e. there may be several different structures that – even in the sample limit – cannot be distinguished. In such cases, a sequence of experiments is required to uniquely identify the true causal structure.

To obtain information in each experiment that can be combined to recover the true causal structure, sequences of experiments have to be designed carefully. Not every sequence of experiments will recover the true causal structure and some sequences of experiments might ultimately recover the causal structure, but could never be performed because they are too long, too complex, too expensive or require experiments that cannot be performed easily. Consequently, rather than just discovery in principle, this thesis aims at a better understanding of (at least one sense of) *efficient* discovery. What is the best sequence of experiments given a set of assumptions and constraints?

The thesis proceeds by providing background information on causal discovery in the first two chapters, defines interventions in the third and then provides several different results on discovery using interventions in Chapter 4 and 5. Chapter 6 provides algorithms that instantiate the results of the previous two

[1] Discovery or determination of causal variables or an account of what it takes to be a causal variable is something that is not addressed here. It is a separate (but not independent) important problem, that inherits many of the difficulties found in the analysis of natural kinds.

chapters and Chapter 7 discusses the particularly tricky issue of conflict resolution that arises when experiments that are subject to statistical errors are combined. Chapter 8 describes a set of simulation results using some of the algorithms and results of earlier chapters, and lastly, Chapter 9 summarizes the results briefly and discusses several open problems.

1.1 History of Methodology for Causal Discovery

Galileo's astronomical observations and experiments with inclined planes illustrate the contrast between passive observational and experimental discovery. In the case of the astronomical observations Galileo was able to record the positions of some of Jupiter's moons and noted that their periodic appearance and disappearance was consistent with an orbit around Jupiter. Galileo was able to develop hypotheses about the orbits and derive testable predictions but was limited to observing and making inferences about a system he had no control over.

In contrast, in his experiments with inclined planes Galileo was able to carefully control size, weight and initial velocity for each object on the plane, as well as the length and inclination angle of the plane. By fixing all the other variables to particular values he could test whether a change in weight resulted in a change in the acceleration of the object on the inclined plane (as suggested by Aristotle). In modern terminology we would say that Galileo fixed or clamped all but one potential cause variable, varied the one remaining potential cause variable and measured the difference in the outcome variable. While the strategy of this procedure is evident, Galileo does not give an explicit account of the methodological role or the advantages of the interventions (clamping or varying variables) to the aim of discovery. The implicit argument is that changes in the outcome could only have arisen due to changes in the causal influence of the varied variable, since all other variables are held constant. While Galileo's experiments are surely not the first instantiation of such a method, they are the earliest we have good records for.

First accounts of a methodology of causal discovery appear in Bacon [1]. Bacon suggested that in order to find the cause of a particular phenomenon one should construct two lists, one of positive and one of negative instances. The list of positive instances should be ordered by increasing degree of the occurrence of the phenomenon. The cause of the phenomenon is then the set of properties present in all the positive instances and absent in all the negative instances and the intensity of the properties increases according to the ordering in the list of positive instances. Bacon does not distinguish between observing the instances and bringing particular circumstances about artificially, so it seems as if his method restricts itself to passive observational discovery.

Much later Mill develops a very similar methodology for causal discovery with his experimental methods of agreement and difference [28]. Mill is aware of the distinction in implementation of passive observation vs. intervention but is ambiguous about the difference in the epistemic access the two approaches provide.[2] Mill puts forth five canons:

First Canon (Method of Agreement): If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all instances agree is the cause (or effect) of the given phenomenon.

Second Canon (Method of Difference): If an instance in which the phenomenon under investigation occurs and an instance in which it does not occur have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

Third Canon: If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

Fourth Canon (Method of Residues): Subduct from any phenomenon such

[2] "For the purpose of varying the circumstances, we may have recourse (according to a distinction commonly made) either to observation or to experiment; we may either find an instance in nature suited to our purposes or, by an artificial arrangement of circumstances, make one. [...] There is, in short, no difference in kind, no logical distinction, between the two processes of investigation. There are, however, practical distinctions to which it is of considerable importance to advert." ([28], p. 211) But contrast this quote with: "But if we cannot artificially produce the phenomenon A, the conclusion that it is the cause of A remains subject to very considerable doubt. [...] Unfortunately, it is hardly ever possible to ascertain all the antecedents unless the phenomenon is one which we can produce artificially." ([28], p. 213)

part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents.

Fifth Canon (Method of Concomitant Variations): Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.

Mill indicates that the first two Canons essentially embody Bacon's methodology[3], but he takes the method of difference to embody more explicitly the controlled experimental design that Galileo used with the inclined planes, although he does not refer to Galileo in particular.[4] While Mill speaks of artificial experiments there is no explicit discussion of whether other potential cause variables should be *clamped* (i.e. held fixed) at some particular value or whether one should aim through careful experimental design to increase the likelihood of obtaining corresponding samples where the other variables *incidentally happen* to have the same values. This distinction is now known as the difference between *statistically conditioning* on (or statistically controlling for) a variable as opposed to *clamping* (or experimentally controlling for) the variable. It is doubtful whether Mill was aware of the difference.

If Mill could assume that the experimenter is dealing with a causally sufficient set of variables (i.e. there are no unmeasured common causes) and if he could ensure that his matching of samples does not amount to conditioning on a common effect of two variables, then the epistemological difference between clamping and statistically conditioning disappears anyway, since both can be used to isolate the causal connection between one potential cause-effect pair. However, using statistical conditioning alone may still pose a significant data collection problem.[5] It does not seem plausible that Mill was aware of these aspects. In a discussion of the limitations of his methodology Mill hints at the problem of causally insufficient sets of variables.[6] But he does not discuss the

[3] [28], p. 216.

[4] "Of these methods, that of difference is more particularly a method of artificial experiment, while that of agreement is more especially the resource employed where experimentation is impossible." [28], p. 216.

[5] Some constellations of variable values might be very rare if one does not force them by clamping.

[6] "In other cases, when we intend to try an experiment, we do not reckon it enough that there be no circumstance in the case the presence of which is unknown to us. We require, also, that none of the circumstances which we do know shall have effects susceptible of being confounded with those of the agents whose properties we wish to study." [28], p. 251.

methodological value of the assumption of causal sufficiency and provides no principled method to ensure either that causal sufficiency is satisfied or that conditioning on common effects is avoided, or what one ought to do if either assumption fails or is not known to be satisfied.

In 1935 R.A. Fisher laid out in detail the methodological aspects of discovery using randomization in *The Design of Experiments* [13]. Fisher had realized that if one randomized the values of the purported cause variable, the *treatment*, one could break any correlation due to latent common causes of the treatment and outcome variables, thereby removing spurious correlations.[7] In addition, randomization provided a reference distribution over the treatment variable that together with an assumption about the functional form of the distribution of the outcome variable allowed the estimation of statistical parameters representing the strength of the causal influence of the treatment on the outcome. Together with design principles that guided the randomization within particular subgroups (blocks), randomization could be used to perform experiments that were sensitive to interactive effects between different treatment variables.[8] In all cases, randomization of the treatment is supposed to ensure that any observed correlation between treatment and outcome could be attributed to the effect of the treatment on the outcome (instead of some unmeasured common cause). Fisher's insight led to a vast development in experimental designs involving randomization. Fisher's basic idea of one treatment and one outcome variable is extended to sets of treatment and outcome variables in Factor experiments and variations of Latin-square designs. On the basis of these developments – and despite their limitations, some of which are mentioned below – randomized trials have become the golden standard to test the efficacy of new treatments in medical research.

The methodology underlying Fisher's approach has not gone without criticism. The main complaint, repeated in various versions [50, 21, 16] concerns the claim that the randomization makes the randomized variable *independent* or its normal causes. If the values of the intervened variable are determined by some random process, then it may happen – by an unfortunate coincidence – that the randomized sample is correlated with some cause of the variable. For example, a random sample from a population of a city might just pick out a

[7] [13], p. 19-21.

[8] Two variables have an interactive effect on some third variable if their causal influence depends on the state of the other variable. For example, the battery has no effect on the motor starting if the fuel tank is empty.

sample containing only women (despite the fact that men and women live in the city), and if gender is an unknown common cause of the treatment and outcome, then the inference about the causal relation between treatment and outcome based on this sample may incorrectly postulate a direct cause where none exists. Randomization only breaks the correlation between the intervened variable and its normal causes in the large sample limit. For finite samples the methodological advantage randomization provides is not guaranteed. This has led to large discussion of how to handle “unfortunate” random samples.

In cases, where the causes of the treatment variable are known, this difficulty with random samples can be avoided by assigning the treatment such that the distribution of the treatment variable is balanced with regard to the value assignments of its causes (see balanced, matched or stratified experimental designs). In the 1970s Rubin and Holland ([37, 38, 39]) developed an approach to causal discovery that built on these insights. They viewed causal discovery as a massive missing data problem. In their view, the problem with causal discovery is that for any individual token, one only ever observes one instantiation of the variable values. Causal claims, however, are supposed to support counterfactual inferences, i.e. what would have happened if certain variable settings had been otherwise. Consequently, for them, the problem of discovery is to determine the values of individual tokens in circumstances that did not actually occur, e.g. for a particular treated participant, the crucial missing information is how this participant would have reacted if she had not been subject to treatment. The states of variables that did not occur for a particular token are treated as missing data. In order to recover this “missing data” the aim of their design and analysis is to match each individual token (e.g. a participant in a medical experiment) with one (or a group) that resembles it as closely as possible, but is not in the same treatment group. The causal effect of the treatment is then measured in terms of the average difference in outcome between the matched pair (of groups). Rubin’s approach to causal discovery still follows the randomized design closely. Conceptually it is presented as a missing data problem, tokens are matched in subgroups (strata) but treatment is randomized within these strata. This randomization within strata is an attempt to break any dependence on unknown common causes, for which such active balancing is impossible.

A different approach to tackling the problems with finite random samples is put forward in Bayesian experimental designs [4]. These designs attempt to remedy the concerns raised with regard to spurious correlations in random samples by taking the actual outcome of the randomization into account explicitly in the analysis of the data. On a high level, Bayesian designs assume that a prior distribution is given over the potential causal structures, which is updated, using Bayes' formula, with the actual data observed in an experiment. Such a design is more robust with regard to correlated samples, since the independence of the random sample is not presumed in the analysis and the outcome of the random sample enters the updating process. Intuitively, the analysis is sensitive to correlated samples and can be appropriately conservative in its conclusions when they occur. Furthermore, Bayesian experimental designs allow for an explicit representation of the experimental cost, and the trade-off between experimental cost (e.g. ethical concerns or difficulty of experiment) and expected knowledge gain about the causal structure [46, 20]. We will discuss Bayesian approaches to the problem of discovery of causal structure in more detail in Chapter 3.

All these experimental designs assume a bipartite separation of the variables into a set of treatment and outcome variables (and possibly co-variates) and do not – at least not in any principled manner – address cases that do not fit such a framework. That is, the experimenter is supposed to know in advance, which variables are the potential causes (treatments) and which the potential effects (outcomes). In cases where no such separation is known a priori or where such a bipartite separation is not possible (e.g. network structures), the principles underlying these experimental design do not provide any guidance on how to proceed. Consequently, in the case of causal structure search, these methods are only informative about a very narrow set of structures. The aim of an experiment is to determine whether or not the treatment variable has an effect on the outcome variable, and if so, how strong it is. No guidance is given on how to choose the set of treatment and outcome variables in the first place. This may be due to the fact that a bipartite separation of the set of variables is supported in many fields (e.g. due to time ordering information or limited possibility of interventions), but one can certainly imagine many cases where such restrictions are unwarranted (e.g. when there are several causal pathways between the treatment and outcome variables).

A more general representational framework of causal structure was developed in a branch of computer science and philosophy: causal Bayes nets [43, 32]. In contrast to the bipartite structures assumed to underlie the models in traditional experimental design, causal Bayes nets consider general graphical networks (although we will restrict the discussion to acyclic ones here). Causal Bayes nets represent the causal relations among a set of variables by a graphical structure and a probability distribution over the variables. The framework allows for the representation of interventions (as in the case of randomized experiments) and the computation of their effects. This enables us to model the various experimental designs discussed above. In this thesis we will explore how this framework can be used to provide informative guidelines on how to choose different experimental interventions given a set of variables and assumptions about the search space.

1.2 Bayes Nets and Formal Definitions

Causal Bayes nets[9] [43, 32] provide a framework that connects a causal structure over a set of variables with a probability distribution over those variables. Formally, causal Bayes nets are represented by a directed acyclic graph (DAG) $G = (\mathbf{V}, \mathbf{E})$ over a set of variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ with a set of directed edges $\mathbf{E}$, and a probability distribution $P(\mathbf{V})$ over the graph.

To reference particular graphical relations, genealogical terms are used in the obvious ways: The parents of a node X, $pa(X)$, in G are the nodes Y with $Y \rightarrow X$, the children of X, $ch(X)$, are all the nodes Z with $X \rightarrow Z$. The ancestors of a node X, $anc(X)$, are all nodes with a directed path into X (and X itself), the descendents, $desc(X)$, are all nodes for which there exists a directed path from X to that node (X is a descendent of itself). The neighbors of X, $neigh(X)$, are the union of the parents and the children.
Two assumptions connect the graph to the probability distribution.

Assumption 1.2.1: Causal Markov Condition
A directed acyclic graph G over $\mathbf{V}$ and a probability distribution $P(\mathbf{V})$ satisfy the Markov condition if and only if for every W in $\mathbf{V}$, W is independent of $\mathbf{V} \setminus (desc(W) \cup pa(W))$ given $pa(W)$.

Assumption 1.2.2: Causal Faithfulness Condition
The probability distribution $P(\mathbf{V})$ is faithful to the graph G if all and only the independence relations true in $P(\mathbf{V})$ are entailed by the Markov condition applied to G.

These two assumptions enable a correspondence of a graphical separation criterium (see *d-separation* below) with conditional independence relations. The causal Markov assumption is a generalization of the ideas of “screening off” underlying Reichenbach’s Principle of Common Cause [34]. Intuitively, the Markov

[9] The formal defintions provided here follow Spirtes et al. [43] as closely as possible. Many definitions and theorems are taken verbatum from their text.

condition states that if two variables are distributionally dependent, then they are causally connected, while the faithfulness condition claims the convers: If two variables are distributionally independent, then they are causally disconnected.[10]

We will use the following graph terminology:

Definition 1.2.3: Graphical Terms

adjacent: Vertices V_1 and V_2 are adjacent in G if there is an edge between them, with these vertices as endpoints.

path: A path between two vertices V_1 and V_2 in a graph is a sequence of vertices starting with V_1 and ending with V_2 such that for each pair of consequent vertices in the sequence, there is an edge in graph G.

undirected path: An undirected path between two vertices V_1 and V_2 in a graph is a sequence of vertices starting with V_1 and ending with V_2 such that for each pair of consequent vertices in the sequence, there is an undirected edge in graph G.

acyclic path: A path is acyclic if it contains no vertex more than once, otherwise it is cyclic.

directed path: A directed path between two vertices V_1 and V_2 is a sequence of vertices starting with V_1 and ending with V_2 such that for each pair of consequent vertices X_1, X_2, occurring in that order in the sequence, there is an edge $E_{X_1,X_2} = X_1 \rightarrow X_2$ in graph G.

source of path: The source of a path is the starting vertex of a directed path.

sink of path: The sink of a path is the end vertex of a directed path.

undirected graph: An undirected graph is a graph that only contains undirected edges.

directed graph: A directed graph is a graph that only contains directed edges.

directed acyclic graph (DAG): A DAG is a directed graph where all paths are acyclic.

[10] Note that Markov and faithfulness relate causal features with distributional features, not sample features.

complete graph: A graph is complete if there is an edge between every pair of vertices.

connected graph: A graph is connected if, when all edges are made undirected, there is an undirected path between every pair of vertices.

subgraph: $G' = (\mathbf{V}', \mathbf{E}')$ is a subgraph of $G = (\mathbf{V}, \mathbf{E})$ if $\mathbf{V}' \subseteq \mathbf{V}$ and $\mathbf{E}' \subseteq \mathbf{E}$.

clique: A clique is a set of vertices in $\mathbf{V}$ over which there is a complete (sub)graph, i.e. there is an edge between every pair of vertices.

tier ordering: $T(\mathbf{V}, \succ_T)$ A tier ordering is an ordering of mutually exclusive and exhaustive sets $T_1 \succ_T \ldots \succ_T T_m$ of the variables $X_1, \ldots X_N \in \mathbf{V}$ in a directed acyclic graph such that T_1 contains all the variables with no parents in $\mathbf{V}$, and for any T_j with $j > 1$, we have $X_i \in T_j$ if and only if there is a $Y \in T_{j-1}$ with $Y \in pa(X_i)$ and there does not exist a $W \in T_k$ where $k \geq j$ with $W \in anc(X_i) \setminus \{X_i\}$.[11]

mediator: Vertex Y is a mediator on a directed path between V_1 and V_2 if it is on the path but not the source or the sink. Note that Y may be a mediator relative to one path, while it is not relative to another.

common cause: Vertex Y is a common cause of V_1 and V_2 if there is a directed path from Y to V_1 and a directed path from Y to V_2 with no vertex shared on the paths except Y. Note that Y may be a common cause relative to one path, while it is not relative to another.

collider: Vertex Y is a collider on a path between V_1 and V_2 if the substructure $X \rightarrow Y \leftarrow Z$ is contained in the path, for some variables X, Z on the path. Note that Y may be a collider relative to one path, while it is not relative to another.

non-collider: Vertex Y is a non-collider on a path between V_1 and V_2 if it is on the path and a mediator or a common cause.

unshielded collider: Vertex Y is an unshielded collider in a DAG G if G contains the substructure $V_1 \rightarrow Y \leftarrow V_2$ and there does not exist an edge $V_1 \rightarrow V_2$ or $V_1 \leftarrow V_2$. An unshielded collider implies particular

[11] The definition implies that all the roots (and unconnected variables) of a graph are in the first tier, and that any variable in tier k is connected to a root by a directed path of length $k-1$. There may be shorter connections, too, but at least one of length $k-1$.

independence relations that can be discovered.[12] An unshielded collider is also sometimes referred to as a v-structure.

active vertex: A vertex V is active on a path relative to a set of vertices $\mathbf{W}$ just in case either (i) V is a collider, and V or any of its descendents is in $\mathbf{W}$, or (ii) V is a non-collider and is not in $\mathbf{W}$.

active path: A path U is active relative to a set of vertices $\mathbf{W}$ just in case every vertex on U is active relative to $\mathbf{W}$.

latent variable: A latent variable is a variable that does not form part of the variables under consideration, i.e. it is unmeasured and not in $\mathbf{V}$, but it is causally connected to the variables in $\mathbf{V}$.

confounder: A confounder is a latent common cause of two variables in $\mathbf{V}$.

causal sufficiency: A set of variables $\mathbf{V}$ is causally sufficient if there are no confounders.

exogeneity: A variable X is *exogenous* to a set of variables $\mathbf{V}$ if there does not exist a variable $Y \in \mathbf{V}$ such that Y is a cause of X.

Definition 1.2.4: d-separation
For a graph G, if X and Y are vertices in G, $X \neq Y$, and $\mathbf{W}$ is a set of vertices in G not containing X or Y, then X and Y are *d-separated* given $\mathbf{W}$ in G if and only if there exists no undirected path U between X and Y, such that

1. every collider in U has a descendent in $\mathbf{W}$
2. no other vertex in U is in $\mathbf{W}$.

We say that if $X \neq Y$, and X and Y are not in $\mathbf{W}$, then X and Y are *d-connected* given set $\mathbf{W}$ if and only if they are not d-separated given $\mathbf{W}$. If $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are disjoint sets of vertices in G and $\mathbf{U}$ and $\mathbf{V}$ are not empty then we say that $\mathbf{U}$ and $\mathbf{V}$ are d-separated given $\mathbf{W}$ if and only if every pair $< U, V >$ in the cartesian product of $\mathbf{U}$ and $\mathbf{V}$ is d-separated given $\mathbf{W}$. If $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are disjoint sets of vertices in G and $\mathbf{U}$ and $\mathbf{V}$ are not empty then we say that $\mathbf{U}$ and $\mathbf{V}$ are d-connected given $\mathbf{W}$ if and only if $\mathbf{U}$ and $\mathbf{V}$ are not d-separated.

[12] Independence relations of an unshielded collider when all three variables are passively observed: (i) X and Y are dependent, (ii) Y and Z are dependent, (iii) X and Z are independent, and (iv) X and Z are dependent conditional on Y. These independence relations imply that Y is a common effect of X and Z, i.e. $X \rightarrow Y$ and $Z \rightarrow Y$. The constraints can be appropriately extended when there are more than three variables.

1.3 Causal Discovery based on Passive Observation

Causal discovery depends on distinguishing different causal structures by the different probabilistic features they imply in particular circumstances given a set of assumptions. Which causal structures can be distinguished depends in part on which techniques of differentiation are available and what assumptions one is willing to make about the nature of the probability distribution and the causal structure. Domain knowledge or information given by the data collection process (such as time order) might further inform the search for causal structure. Without additional domain knowledge and without interventions, one is left with techniques that distinguish sets of (passive observational) probability distributions that are Markov and faithful to different causal graphs. Two main approaches can be distinguished: score based techniques and constraint based techniques. Score based techniques compute a numerical score for the goodness of fit of each possible causal model given the data. There are many different scores based on the likelihood, priors, model complexity or information theoretic measures. The score is supposed to pick out those models that are most likely or most plausible to have generated the data. In contrast, constraint based search procedures test for different probabilistic constraints in the data, such as (conditional) independencies or differences in some statistic, that restrict the possible causal models. Causal models are selected depending on whether they satisfy the constraints found in the data. For any particular set of constraints, there may be equivalence classes of causal structures, such that all structures in one equivalence class are different, but indistinguishable even in the large sample limit. For example, if one considers passive observational data only, then the classes of causal structures that imply the same (conditional) independencies, are referred to as observational Markov equivalence classes (OMEs). All graphs in a Markov equivalence class imply the same conditional independencies among the observed variables. Concretely, assuming causal sufficiency, the following three causal structures form one OME. They all imply only one independence constraint, namely X is independent of Z given Y ($X \perp\!\!\!\perp Z | Y$):

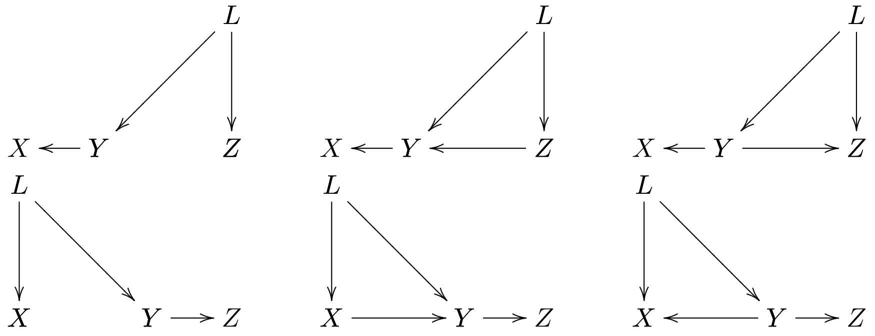
1. $X \leftarrow Y \rightarrow Z$
2. $X \rightarrow Y \rightarrow Z$
3. $X \leftarrow Y \leftarrow Z$

The OME can be represented by a pattern [47]:

$$\begin{tikzcd} X \arrow[r, dash] & Y \arrow[r, dash] & Z \end{tikzcd}$$

An undirected edge between two variables X and Y indicates that each graph in the OME has an edge between X and Y but that the edge points in different directions for different graphs.

The class of structures indistinguishable by independence constraints becomes substantially larger, when the assumption of causal sufficiency is dropped. If L is a latent variable, then in addition to the previous three structures we have:



Algorithms that use independence constraints to discover causal structure are – with passive observational data – limited in how much can be discovered. There are numerous ways to search for causal structure in passive observational data and not all algorithms are limited to observational Markov equivalence classes. Here I will give an overview with specific focus on algorithms that will be integrated (in modified form) into the search procedures in Chapter 6.

1.3.1 PC-Algorithm

The PC-algorithm[13] is a constraint based causal structure search algorithm. It uses independence constraints to infer causal structure between variables under the assumption that the set of variables is causally sufficient. The PC-algorithm proceeds in two stages, the first searches for adjacencies between variables, while the second resolves – where possible – orientations of edges. The PC-algorithm is initialized with a complete undirected graph over the set of variables. It proceeds by testing for independencies in order of size of the conditioning set. It considers

[13] [43], p. 84.

all pairwise independencies first and where an independency is found the edge between the variables is removed. As edges are removed only a subset of all possible higher-order independence tests have to be considered. At the end of this first stage, all adjacencies in the graph have been resolved. In a second stage all unshielded triples X, Y, Z of vertices[14] are subjected to a collider test: If Y is not contained in the conditioning set that was used to remove the XZ-edge, then the unshielded triple is oriented as $X \rightarrow Y \leftarrow Z$, otherwise it is left unoriented. Lastly, a set of rules (Meek-Rules, [23]) are employed that orient edges, whose orientation is implied by the existing edges and and known orientations. For example, if there is a $X \rightarrow Y \rightarrow Z$-path and an undirected edge from X to Z, then that is oriented $X \rightarrow Z$ (to preserve acyclicity of the graph). The output of the PC-algorithm is a pattern that represents the Markov equivalence class of graphs that all imply the same independence constraints that were found in the data.

The PC-algorithm is provably correct in the sense that given an independence oracle that supplies the independence constraints true in the distribution that generated the data, the PC-algorithm recovers as much information about the true causal structure as is possible with the available independence facts. However, it only guarantees pointwise consistency to the true model, which implies that there is no one fixed sample size that bounds the error probabilities simultaneously for all probability distributions Markov and faithful to a DAG over the given variables [36].

The algorithm can be supplemented with background knowledge that forces the existence or prohibits the presence of particular edges with and without orientations. Extensions of the PC-algorithm to causal structure search in sets of variables that are causally insufficient are implemented in the FCI-algorithm [43].

1.3.2 Convservative PC-Algorithm

The conservative PC-algorithm [33] is a slight modification of the original PC-algorithm. One of the sources of errors in the PC-algorithm is the orientation of unshielded triples as colliders. For any unshielded triple X, Y, Z of vertices the PC-algorithm only checks the conditioning set $\mathbf{C}$ that was used to remove the XZ-edge. If $\mathbf{C}$ does not contain Y, then the unshielded triple is oriented as a

[14] A triple X, Y, Z of vertices is unshielded if X and Y are adjacent and Y and Z are adjacent, but X and Z are non-adjacent.

collider. Instead of just considering $\mathbf{C}$, the cPC-algorithm checks all subsets of potential parents of X and Z: If Y is not contained in any such set $\mathbf{C}'$ such that $X \perp\!\!\!\perp Z | \mathbf{C}'$, then the triple is oriented as unshielded collider: $X \rightarrow Y \leftarrow Z$. If Y is contained in all such sets $\mathbf{C}'$ such that $X \perp\!\!\!\perp Z | \mathbf{C}'$, then the triple is left as it is, connected by two undirected edges. For any other case the triple is marked as "unfaithful", which indicates that the algorithm is unable to tell whether there is an unshielded collider, whether Y is a common cause or whether there is a chain from X to Z or vice versa. The output of the cPC-algorithm is an augmented Markov equivalence class that reflects the uncertainty for the triples that are marked as unfaithful colliders, by including the graphs that place a collider at the unfaithful collider vertex. Simulation studies in [33] show that the cPC-is much less error-prone.

The cPC-algorithm inherits the correctness results of the PC-algorithm, but in addition, the cPC-algorithm is uniformly consistent, i.e. there is one sample size that bounds the error probabilities for all possible probability distributions Markov and faithful to a DAG over the given variables.

1.3.3 GES-Algorithm

The Greedy Equivalence Search (GES) algorithm [24, 5] is in contrast to the previous two a score based algorithm. It uses a score to identify the true causal structure. GES places a prior over all possible directed acyclic graphs over the given N variables. It is initialized with an empty graph and proceeds in two stages, a forward and a backward stage. In the forward stage edges are added to the graph until there is no improvement in the score. In the following backward stage edges are removed again until there is no improvement in the score. The forward stage is a greedy search over equivalence classes that differ from the current equivalence class by having one additional edge.[15] The search proceeds until a local maximum is reached. The following backward stage is similar, only that equivalence classes are considered that match the current equivalence class except for one missing edge. The Bayesian scoring criterion is used to evaluate the score of any step in the search.

The GES-algorithm, like the PC-algorithm is provably correct. In the large sample limit it returns the Markov equivalence class containing the true causal graph.

[15] For details see [5], p. 521-522.

1.3.4 Independent Component Analysis - LiNGAM

More recently, an entirely different approach has been put forward for causal structure search for linear non-Gaussian models: LiNGAM [41]. The approach makes use of a statistical tool to identify linear models: Independent Component Analysis (ICA). In a linear model the variables $\mathbf{x} = (X_1, \ldots, X_N)$ are linear functions of their parents and the error terms, here represented by $\mathbf{e}$, which are assumed to be non-Gaussian and independent. In the LiNGAM-algorithm the causal model is represented my a matrix equation:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

where the matrix $\mathbf{B}$ is an $N \times N$ matrix representing the edge coefficients on edges connecting the variables. If the columns of $\mathbf{B}$ correspond to the hierarchical order of the variables in the graph, then $\mathbf{B}$ is lower triangular. In a linear model the variables are linear functions of their parents and the error terms, here represented by $\mathbf{e}$, which are assumed to be non-Gaussian and independent. The basic idea of LiNGAM is to recover the $\mathbf{B}$-matrix from a data matrix: Solving the above equation for $\mathbf{x}$ we get

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$

This equation together with the above assumptions fits the ICA framework, and it follows that $(\mathbf{I} - \mathbf{B})$ is identifiable.[16] The upshot is that the non-Gaussian error assumption enables discovery of the exact true causal structure, as opposed to only the Markov equivalence classes. Efforts have been made that extend the LiNGAM-approach to latent variables [18].

Given these algorithms it becomes clear that any account of the limits of discovery must be sensitive to the particular algorithms employed, the assumptions they make about the functional form of the model and the criteria that are used for the search.

1.3.5 Limitations of Search by Passive Observation

The various versions of the PC-algorithm and the GES-algorithm are limited even in the large sample limit to discovering the Markov equivalence class con-

[16] For more detail see [41].

taining the true graph. If the search method is limited to independence constraints, then in general, interventions are required to uniquely identify the true graph in a Markov equivalence class.

The results based on the LiNGAM-algorithm suggest that there are no limitations to structure search in passive observational data – at least not in principle – when sample size is not the concern. For non-normally distributed data any structure among the observed variables, and with the extensions of the more recent results, even the presence of latent variables can be discovered. The limitation that the results do not hold for normally distributed data is a rather minor weakness. However, the results do not extend to discrete variables. Much of the following, but not all, will provide methods and results that apply to structure search for discrete models as well.

The advantage of interventions is that they expand the set of criteria that can be used for identification of causal structure. Interventions provide in addition to the passive obervational distribution over the variables, a manipulated distribution. Different graphs, that appear equivalent given passive observational data only, can be distinguished by their different manipulated distributions. For example, consider the set of three graphs discussed at the beginning of this section that form the following OME:

$$\begin{array}{c}
X \longleftarrow Y \longrightarrow Z \\
X \longrightarrow Y \longrightarrow Z \\
X \longleftarrow Y \longleftarrow Z
\end{array}$$

A randomization of Y would make Y independent of its causes, i.e. it would leave X and Z unconditionally dependent in the first case – the common cause – but independent (due to the randomization) in the second and third, the two chain graphs. Further, for the second graph, where X is the source of the chain, X and Y would also be independent, whereas X and Z would not, and in the third, where Z is the source of the chain, X and Z would be independent but X and Y would not. The three equivalent structures become distinguishable even for simple independence tests by their different manipulated distributions, represented by the following three manipulated graphs in which the causal influences that get destroyed by the intervention on Y are removed:

1. $X \leftarrow Y \rightarrow Z$
2. $X \qquad Y \rightarrow Z$

3. $X \leftarrow Y \qquad Z$

Consequently, interventions are not only of interest to make discovery possible at all, but also to enable discovery with means that would not suffice if only passive observational data were available. In addition, there are various considerations relating to the estimation of parameters that make interventions a useful tool, but we will not go into any detail of this aspect of interventions in this thesis.

How interventions should be placed in order to distinguish different causal structures and under which circumstances different types of interventions distinguish different causal structures is the subject of this thesis. There are four cases to consider: First, a single intervention may not be sufficient to distinguish causal structures. Second, under certain assumptions no interventions are needed to distinguish causal structures. Third, sometimes causal structures can be distinguished without interventions, but the use of interventions allows a weakening of other assumptions while maintaining the same level of identifiability. And fourth, there are cases for which interventions are the only known way to make progress.

Chapter 2

Intervention

In this chapter we lay out the space of interventions. We review in some detail the three most comprehensive accounts of interventions, by Pearl [32], by Spirtes et al. [43], and by Woodward [49]. We follow all three accounts in their key features, but attempt to give a more general minimal description of interventions in the second section that includes interventions of different strength and interventions that may not be entirely exogenous. In a third section we discuss some of the aspects of our definition of interventions that may be considered controversial on philosophical grounds. In the last remaining sections we tighten the general definition of interventions to specify particular types of interventions. Since this thesis is about discovery, we focus the discussion of interventions on epistemological issues. We leave a discussion of the metaphysics of interventions aside, since that issue is closely connected with the more general issue of what constitutes a causal variable. A broader discussion of such issues can be found in Woodward's *Making things happen* [49].

2.1 Interventions and Causes

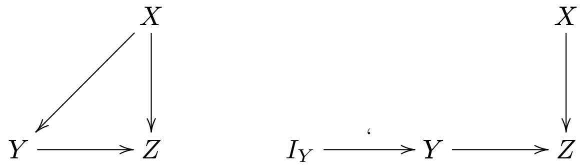
Pearl [32] and Sprites et al. [43] focus on the formal constraints an intervention imposes on a system of variables, while Woodward gives a more general metaphysical analysis of the aspects of intervention.

Pearl considers interventions relative to a structural equation model that describes the causal relations over a set of variables. For Pearl an intervention is *atomic* if the intervention "amounts to lifting [the intervened variable] X_i

from the old functional mechanism $x_i = f(pa_i, u_i)$ [where pa_i are the graphical parents of X_i and u_i are the unobserved influences on X_i] and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed."[1] In a causal Bayes net $G = (\mathbf{V}, \mathbf{E})$, Pearl represents an intervention on a variable $X \in \mathbf{V}$ by an *intervention variable* I that is added to the causal structure G with a direct link $I \rightarrow X$ into the intervened variable. For Pearl, the intervention variable can take values in $\{idle, do(x_i)\}$ where x_i ranges over the values of X_i. In contrast to Fisher's randomizations [13] that determine a *distribution* over the intervened variable, Pearl's atomic intervention forces (clamps) the intervened variable to one particular value. But like a randomization, the intervention breaks the causal influence of the normal causes on the intervened variable, which is reflected in the resulting "manipulated" distribution of X_i conditional on its graphical parents (normal causes).[2]

$$P(X_i = x_i | pa_i, I_i) = \begin{cases} P(X_i = x_i | pa_i) & \text{if } I_i = idle \\ 0 & \text{if } I_i = do(x_i') \text{ and } x_i \neq x_i' \\ 1 & \text{if } I_i = do(x_i') \text{ and } x_i = x_i' \end{cases}$$

Since the intervention breaks the causal influence of the normal causes on the intervened variable, the effect of this type of intervention can be represented by a manipulated graph, in which the edges into the intervened variable (other than from the intervention variable) are removed, as shown in the following figure.



If the true graph is shown on the left, then an intervention on Y results in the manipulated graph, where the causal influence of X on Y is destroyed, on the right.

On the basis of this account of an intervention on a single variable, Pearl develops the *do*-calculus[3] to compute the effect of interventions on the probability distribution over the set of variables $\mathbf{V}$:

[1] [32], p. 70.
[2] [32], p. 71.
[3] [32], p. 85.

"Let X, Y and Z be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$ [...]. Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that X is held constant at x and that (under this condition) $Z = z$ is observed.

Algorithm 2.1.1: Pearl's Rules of *do*-Calculus
Let G be the directed acyclic graph associated with a causal model and let $P(.)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z and W, we have the following rules.

Rule 1: (insertion/deletion of observations)

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

Rule 2: (action/ observation exchange)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

Rule 3: (insertion/ deletion of actions)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$."

Pearl uses this calculus of interventions to specify conditions for the identifiability of causal connections:[4]

Theorem 2.1.2: Pearl: Identifiability
"A causal effect $q = P(y_1, \ldots, y_k|\hat{x}_1, \ldots, \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in the theorem above, that reduces q into a standard (i.e. "hat"-free) probability expression involving observed quantities."

[4] [32], p. 86.

It has since been shown that the *do*-calculus is complete [19].

The underlying idea is that by specifying appropriate formal conditions on the intervention, namely, in Pearl's case the clamping of the variable to a particular value, and the exogeneity of the intervention, one can characterize conditions in which the true causal structure can be discovered. Furthermore, some of these conditions allow for discovery, when the causal structure would otherwise have not been identifiable. Pearl's *do*-calculus consequently gives an epistemological account of a particular type of intervention.

Spirtes et al. [43] largely share Pearl's representation of interventions in causal Bayes nets, although they permit an intervention to force non-degenerate distributions over the intervened variable. Spirtes et al. provide a simpler and more general theorem for the computation of the effects of interventions:

Theorem 2.1.3: SGS: Manipulation Theorem
"Let $G = \{\mathbf{V}, \mathbf{E}\}$ be a directed acyclic graph and let $\mathbf{I}$ be the set of variables in $\mathbf{V}$ that are subject to an intervention. Then G_{unman} is the unmanipulated graph corresponding to the unmanipulated distribution $P_{unman}(\mathbf{V})$ and G_{man} is the manipulated graph, in which for each variable $X \in \mathbf{I}$ the edges incident on X are removed and an intervention variable $I_{s(X)} \rightarrow X$ is added. A variable $X \in \mathbf{V}$ is in $man(\mathbf{I})$ if it is subject to an intervention, i.e. if it is a direct child of an intervention variable $I_{s(X)}$. Then

$$P_{unman(\mathbf{I})}(\mathbf{V}) = \prod_{X \in \mathbf{V}} P_{unman(\mathbf{I})}(X|pa(G_{unman}, X))$$

$$\begin{aligned} P_{man(\mathbf{I})}(\mathbf{V}) \quad = & \prod_{X \in man(\mathbf{I})} P_{man(\mathbf{I})}(X|I_{s(X)} = 1) \times \\ & \prod_{X \in \mathbf{V} \setminus man(\mathbf{I})} P_{unman(\mathbf{I})}(X|pa(G_{unman}, X)) \end{aligned}$$

for all values of $\mathbf{V}$ for which each of the conditional distributions is defined."

This theorem specifies how an intervention manipulates the causal structure and the probability distribution over the set of variables $\mathbf{V}$. The theorem can be used both to predict the effect of interventions on a known causal structure and to derive statistical features that can be used to discover the causal structure by using interventions.

While these rules and theorems specify formal constraints resulting from interventions, they do not tell us what an intervention or a causal effect is. One of the difficulties in providing a more detailed metaphysical account of an intervention is the interdependence between the notion of causal effect and the notion of intervention. Pearl uses his description of an intervention to define a causal effect:[5]

Definition 2.1.4: Pearl: Causal Effect
"Given two disjoint sets of variables, $\mathbf{X}$ and $\mathbf{Y}$, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$, denoted either as $P(y|\hat{x})$ or as $P(y|do(x))$, is a function from $\mathbf{X}$ to the space of probability distributions on $\mathbf{Y}$. For each realization x of $\mathbf{X}$, $P(y|\hat{x})$ gives the probability of $\mathbf{Y} = y$ induced by deleting from the model all equations corresponding to variables in $\mathbf{X}$ and substituting $\mathbf{X} = x$ in the remaining equations."

So, for Pearl, for two variables to stand in the relationship of cause and effect amounts to a functional relation between the manipulation of variables in set $\mathbf{X}$ and the probability distribution over a set of variables $\mathbf{Y}$. However, this function is only defined in terms of a model and the appropriate changes of equations in a model. Pearl's account of intervention does not tell us what constitutes an intervention until a model has been specified.

Woodward also defines a direct cause in terms of interventions, but his strategy is different:[6]

Definition 2.1.5: Woodward: Direct Cause
"A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set $\mathbf{V}$ is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in $\mathbf{V}$ besides X and Y are held fixed at some value by interventions."

Woodward's intervention does not depend on a specified model of the causal relations connecting the variables. Instead, he appears to go for an arguably circular definition. He defines an intervention in terms of influences that control (the distribution over) the values of the intervened variable:

Definition 2.1.6: Woodward: Intervention Variable
"I is an intervention variable for X with respect to Y if and only if I meets the following conditions:

1. I causes X.

[5] [32], p. 70.

[6] [49], p. 55 and p. 98. Korb [22] largely follows Woodward in this approach.

2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of the other variables that cause X and instead only depends on the value taken by I.

3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y, if any, that are built into the $I-X-Y$ connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., the variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.

4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X."

Definition 2.1.7: Woodward: Intervention
"I's assuming some value $I = z_i$ is an intervention on X with respect to Y if and only if I is an intervention variable for X with respect to Y and $I = z_i$ is an actual cause of the value taken by X."[7]

Both Pearl and Woodward give an account of causes in terms of interventions. The main difference between Pearl and Woodward is that Woodward does not require the reference to a fully specified model or a set of equations when defining an intervention and he commits explicitly to causal terminology in defining an intervention. Woodward is fully aware of the circularity this appears to entail. If causal claims are to be understood in terms of (hypothetical) interventions, it suggests that the notion of an intervention is more fundamental than that of a cause. However, as Woodward's definition explicitly acknowledges, the definition of interventions quite obviously involves causal terms, since probabilistic features are insufficient to adequately distinguish interventions from conditionalization.[8] Woodward[9] and Hitchcock [17] discuss this apparent circularity in the definition of causal relations and interventions at length, but they do not consider it vicious. Woodward argues that the causal

[7] One variable is an actual cause of another variable if it is a cause at the individual (token) level as opposed to the population level. Woodward provides a definition of what constitutes an actual cause, which we will not discuss here, apart from indicating that the resort to token causation does not reduce the definitional problems of type causation and certainly does not refer the problems to grounds that are supported by broad agreement among philosophers.

[8] See [25], [32], Chapter 3 and many others.

[9] [49], p. 104-107.

claim in the definition of an intervention depends only on an understanding of the causal effect of an intervention on the intervened variable, but does not depend on an account of causal relationships between ordinary causal variables (variables in $\mathbf{V}$). There is for Woodward, so to speak, a difference between the causal relation between an intervention variable and an intervened variable on the one hand, and the causal relation between two ordinary causal variables (whatever it may take to be one) on the other. The former is needed for an understanding of the concept of intervention, while the concept of an intervention is required to understand the latter. Woodward may be understood as trying to draw a line between causation and causality: The former, involving interventions, is required for a definition of the latter, which makes no explicit claim about interventions. To Woodward, this non-vicious circularity between the definition of intervention and that of direct cause is simply an indicator that the notions of cause and intervention cannot be reduced to more primitive notions.

Spirtes et al. [43] take a different approach: A cause is a primitive and interventions are additional exogenous causes that augment the causal graph over the variables under consideration. These interventions imply particular additional ("context specific") independencies in the graph when the interventions are active. But while taking a cause to be a primitive provides an elegant way to escape the problem, it seems like a rather simple way of disposing of a problem that could keep many philosophers in business.[10]

Both Pearl and Woodward only consider interventions that make the intervened variable (causally) independent of its normal causes – by clamping or randomization – they do not consider weaker forms of interventions. They also both only appear to consider interventions as variables that are uncaused, although there is no explicit statement to that effect. Both Pearl and Woodward appear to consider the intervention variable to be a mere representational artefact of the model, rather than corresponding to any real variable. Neither is committed to a distribution over the intervention variable. Pearl quite explicitly only considers distributions *conditional* on values of the intervention variable. But also, neither restricts themselves to intervention variables as decision points. If the intervention variable were a decision point, then there might be some debate as to whether it has a marginal distribution, and an intervention would involve some form of agency or free will to determine the state of the

[10] Given much of the philosophical literature, this appears to be an end in itself.

intervention variable. Woodward explictly emphasizes that his understanding of interventions does not require a notion of agency.[11]

2.2 Interventions as Discovery Tool

Fisher [13] develops the theory of randomized controlled trials for their particular value for discovery. He describes a randomization as an intervention that assigns treatment at random after all disturbing causes are determined (in the sense of "set", not in the sense of "discovered"). Ideally, treatments should be "the last in time of the stages in the physical history of the objects which might affect their experimental reaction."[12] Interventions are used to create circumstances that support the inference from a particular observed association between the treatment and the outcome to the causal influence of the treatment on the outcome. The randomization is supposed to ensure that no influences other than the intervention determine the state of the intervened variable and that consequently any confounding due to known or unknown common causes can be eliminated.

This feature will be one of the guiding considerations in specifying interventions as discovery tools. Our aim will be to provide criteria that are weaker than Fisher's. We follow Pearl in many representational aspects and Woodward in rejecting agency as a necessary component for an intervention. But in doing so, we have to give a precise account of an intervention variable and how an intervention differs from an ordinary causal relation. We will first provide a minimal definition of an intervention (for discovery) and then turn to more specific common forms of intervention.

Definition 2.2.1: Intervention – discrete model
Given a discrete causal Bayes net $G = (\mathbf{V}, \mathbf{E})$, with probability distribution $P(\mathbf{V})$, where each $X_i \in \mathbf{V}$ has k_i values, an intervention I on a subset $\mathbf{S} \subseteq \mathbf{V}$ satisfies the following criteria:

1. $I \notin \mathbf{V}$ is a variable with $1 + \prod_{X_i \in \mathbf{S}} k_i$ states, i.e. I has a state for each combination of values of the variables in $\mathbf{S}$ and one additional *idle* or

[11] See section on *Nonanthropomorphism* in [49].
[12] [13], p. 20.

0-state:[13]

$$\begin{aligned}\{idle, \quad & [do(x_{s(1),1}), do(x_{s(2),1}), \ldots, do(x_{s(s),1})], \\ & \ldots, \\ & [do(x_{s(1),k_{s(1)}}), do(x_{s(2),k_{s(1)}}), \ldots, do(x_{s(s),k_{s(s)}})]\}\end{aligned}$$

2. I is a direct cause of each variable $X \in \mathbf{S}$, i.e. $I \rightarrow X$. That is, there is at least one value $X = x$ such that $P(X = x|I = 0) \neq P(X = x|I = k \neq 0)$.[14]

3. There is a joint distribution $P(\mathbf{V}, I)$ over the variables in $\mathbf{V}$ and I and consequently I has a – possibly degenerate – marginal distribution over its values. If there are several simultaneous interventions, there is a joint distribution over $\mathbf{V}$ and the set of simultaneous interventions $\mathbf{I}$.

4. The probabilistic effects of causes of I, if any, are known. That is, if there is a variable C, with $C \rightarrow I$ and possibly $C \in \mathbf{V}$, then $P(I, \mathbf{V} \setminus \{C\}|C)$ is assumed to be known. This is most relevant when $C \in \mathbf{V}$ or when C is a (latent) common cause of I and some variable in $\mathbf{V}$.

5. If there is a variable $C \in \mathbf{V}$ that is a cause of I, then there exists at least one other cause of I that is not in $\mathbf{V}$.[15]

6. When $I = idle$ $(I = 0)$, the passive observational distribution over $\mathbf{V}$ obtains, i.e.

$$\begin{aligned} P(\mathbf{V}|I = 0) &= P(\mathbf{V}) \\ &= \prod_{V_i \in \mathbf{V}} P(V_i|pa(V_i)) \\ &= P(\mathbf{S}|pa(\mathbf{S})) \prod_{V_i \in \mathbf{V} \setminus \mathbf{S}} P(V_i|pa(V_i)) \end{aligned}$$

[13] The idle state represents circumstances in which the intervention is ineffectual, i.e. the state of the intervened variable is functionally independent of the state of the intervention variable. This feature is not essential in the further discussion, but is an attempt to capture the notion that an intervention can be turned off and that there are circumstances in which one can speak of a passive observational distribution over the variables. Woodward assumes implicitly that there is such a state and Pearl accounts for it in terms of a functional independence of the structural equations of the intervened variable on an idle intervention variable. This state space is a direct extension of the states of Pearl's intervention variable to an intervention on a set of variables.

[14] This is sometimes referred to as a test pair condition, that a direct cause must satisfy.

[15] See discussion below on the difference between causes and interventions for a motivation of this assumption.

7. When $I = k \neq 0$, the conditional distribution over $\mathbf{S}$ is manipulated, i.e.

$$P(\mathbf{V}|I = k) = P(\mathbf{S}|pa(\mathbf{S}), I = k) \prod_{V_i \in \mathbf{V} \setminus \mathbf{S}} P(V_i|pa(V_i))$$

where

$$P(\mathbf{S}|pa(\mathbf{S}), I = k) = \prod_{X \in \mathbf{S}} P^*(X|pa(X), I = k)$$

and for each $X \in \mathbf{S}$ we have

$$P^*(X|pa(X), I = k) \neq P(X|pa(X), I = 0)$$

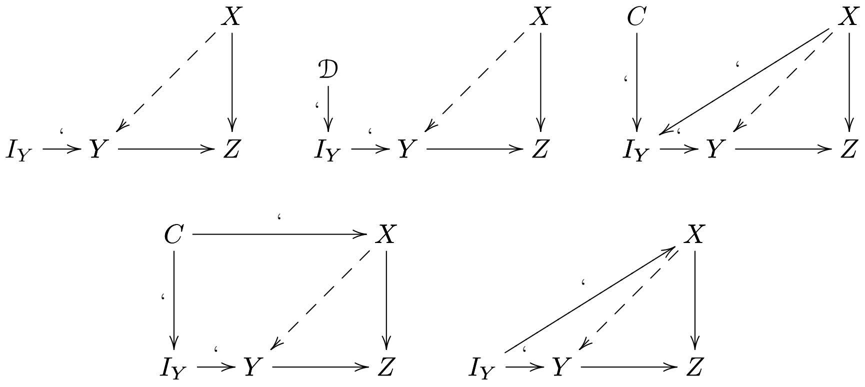
8. If I is an intervention that involves a *decision* to intervene, then a further decision node $\mathcal{D}$ is associated with I. $\mathcal{D}$ is a direct cause of I, has several states (depending on the state space of the decision), but no marginal distribution, i.e. $P(\mathbf{V}, I)$ is only defined given a specific choice of the value of $\mathcal{D}$).

An intervention is said to be *confounding* if $\mathbf{S}$ is larger than a singleton set, i.e. if an intervention manipulates more than one variable.[16] An intervention is *confounded* if there is a common cause C of the intervention variable and some other intervention variable or of some variable in $\mathbf{V}$.

If the variables in $\mathbf{V}$ are continuous, then so is I (though higher dimensional if I is confounding). The state space of I must cover the space of all combinations of states of variables that I intervenes on. The distribution over I may, of course, not be continuous. The state space of the decision node $\mathcal{D}$ is a space of probability distributions representing the possible intervention distributions.
Below we illustrate the main constellations that relate an intervention variable I to other variables in the system. Suppose the set of variables under consideration is $\mathbf{V} = \{X, Y, Z\}$. Let $\mathcal{D}$ be a decision node and let C be a variable that is not

[16] A confounding intervention is different from multiple simultaneous interventions. For a confounding intervention the intervention variable is a common cause of the variables it intervenes on and consequently the variables are correlated as a result of the intervention. Multiple simultaneous interventions influence the intervened variables independently.

in **V**.



The first graph shows a standard intervention on Y. Whether or not the intervention breaks the causal effect of X on Y is left open by the current definition. The second graph shows that a decision might influence the distribution over the intervention variable. For example, one may decide to determine the treatment of a patient by flipping a fair coin, or, instead, one may decide to flip a weighted coin. Both are possible intervention distributions (distributions over the values of the intervention variable), which one can decide between. A decision point is always separated from the variables under consideration (i.e **V**) by an intervention variable (**I**) specifying the intervention distribution. If there is a variable in **V**, such as X in the third graph, that influences the intervention distribution, then there must be another cause of I, in this case C, that also influences the distribution over I, otherwise I is not an intervention. C could, of course, also be a decision point (as e.g. for conditional interventions). In the second row, the first graph shows a confounded intervention, while the second graph shows a confounding intervention.

The definition of an intervention specifies the key features relevant for an intervention as discovery tool. For an intervention to be useful to discovery, it must place constraints on the system it intervenes on and bring exogenous influences into the system of variables under consideration. The former is achieved by the manipulation of the marginal distribution of the intervened variables and by assumptions about parts of the causal structure surrounding the intervention variable. The latter is achieved by ensuring that the intervention is either exogenous to **V** or has at least one cause that is not in **V**. Both aspects can be used for causal discovery since they distinguish causal structures which might otherwise appear indistinguishable.

The definition is minimal in the sense that it does not restrict how many variables are influenced by an intervention, there is no restriction on the effect the intervention has on the intervened variable (other than that the intervention has *some* effect on the distribution of the intervened variable given its causal parents). In particular, there is no requirement that the intervention should make the intervened variable independent of its normal causes. The definition leaves open the possibility of considering interventions that are dependent on variables in the system or that may be confounded by other variables. The distribution over the intervention variable allows for an explicit representation of the intended intervention distribution which may or may not be successfully conferred on the intervened variable.

2.2.1 Discussion

Difference between Interventions and Causes

An intervention variable is similar to an ordinary causal variable in that it has states, a marginal distribution and may have causal ancestors. Lack of causal ancestors is not a necessary requirement for interventions and it does not generally restrict discovery much, if they are causal ancestors of the intervention variable alone. With the minimal definition we can easily make sense of interventions which have further causes. We need not even restrict ourselves to exogeneity. There are many cases, where it is quite plausible to speak of the variables under investigation as being a cause of the intervention and discovery procedures can be adapted accordingly. For some interventions the state of the variables that are subject to study have an explicit influence on the intervention and its distribution, as, for example, in sequential experiments. More generally, many scientists are led to particular experiments on variables, because those variables produced curious data in the past. Hence, at least informally, one would speak of the variables being a cause of the intervention.

One may argue that the variables that influence the intervention are different to the variables intervened upon. If a time dimension were included, they would have a different time stamp. But this distinction does not help, since one would still refer to all variables at all time instances as “the variables under investigation”, i.e. as forming part of $\mathbf{V}$. Exogeneity, would therefore not be satisfied, and one would not want to definitionally enforce exogeneity by excluding variables prior to an intervention from the set $\mathbf{V}$: First, for most other aspects of the search procedure, these variables contain relevant information and second, such a restriction would exclude variables from $\mathbf{V}$ that need not be excluded, e.g. variables that are causally disconnected from the intervention (which ones those are, may not be known from the outset). The basic point is simple: Making exogeneity a necessary requirement of an intervention in general is excessively restrictive. Analyses of interventional data can be adapted to a whole variety of possible constellations. Interventions should be thought of as ordinary causes, like any other cause, potentially fully connected in a network. It therefore follows quite obviously that some interventions influence more than one variable simultaneously. When such a confounding intervention occurs unintentionally, it is often referred to as a "fat hand" intervention, since it is similar to someone trying to manipulate some intricate object with insufficiently slim fingers.

But an intervention differs from normal causal variables in (i) that it can assume a large variety of different marginal distributions, (ii) that it can be affected by decision points that do not have a marginal distribution, (iii) that it need not correspond to some real variable and can in most cases be considered a representational artefact of the model. Our specification of the distributional constraints is just in terms of a distribution over the set of variables $\mathbf{V}$ *conditional* on states of the intervention I and for the most part that will be sufficient. However, the marginal distribution over the intervention variable has several functions: (a) it distinguishes the intervention variable from a decision point and therefore clearly distinguishes the intervention distribution from the place where agency may enter the process; (b) it can be used to explicitly represent non-degenerate intervention distributions, as they are used in randomized trials; and (c) it can be used to represent different intervention distributions, e.g. whether a randomization is uniform over the values of the intervened variable or not.

Interventions also differ from ordinary causal variables in that they require some knowledge about local causal structure. In our definition of an intervention, the direct connection between an intervention variable and an intervened variable is known, and any confounding of the intervention is assumed to be known, as is any causal influence on the intervention variable from variables in $\mathbf{V}$. This partial knowledge of causal structure is necessary for an intervention, if it is going to be used as a discovery tool. Such knowledge places appropriate constraints on the variables under investigation. Causal background knowledge about particular causal paths also supplies similar constraints that can be used for causal learning, but the difference between background knowledge and interventions is that an intervention introduces an *external* influence into the system.

This aspect is preserved in particular by point 5 of the definition. If the intervention variable were completely (distributionally) determined by variables in $\mathbf{V}$, then it could be marginalized out and would amount to background knowledge regarding one specific pathway internal to $\mathbf{V}$. The external influence on the intervention variable prevents a full determination of when the intervention occurs as a function of the variables under investigation. Intuitively, requiring that an intervention introduces an external influence into the system ensures that the variables under investigation cannot "switch the intervention off", when effects of interest would occur. For example, if the variable that influences the intervention always sets the intervention on Y to "idle" when some other variable X has the value $X = 1$, then an interactive effect between X and Y on variable Z may not or may only very rarely be observed, since the intervention is prevented from forcing the value of Y to the state that creates the interactive effect when $X = 1$. An external influence on I creates an additional trigger for the intervention. Of course, higher level causes of the intervention or interactive causes of the intervention may still prevent the efficacy of an intervention at the "interesting times", but now we are splitting hairs.

Agency

Following Woodward, we distinguish the role of agency (in form of decision points) from that of an intervention. The underlying motivation is that we do not want an account of interventions to depend on an account of agency or free will. It seems quite unncessary to require for an understanding of interventions an account of what constitutes free will or agency. The crucial aspect of interventions as discovery tools is that they place known constraints on a system of variables. Agency is unnecessary. A machine can perform interventions and discover causal structure from the resulting data.

Interventions are points at which agency *can* impact the system of variables under consideration, but interventions can also be understood without reference to agency. Without any component of agency, interventions are viewed as marking the borderline of the set of variables under consideration. For a particular set of variables $\mathbf{V}$, an intervention introduces an external influence into the system, but if the system of variables were expanded, the intervention variable may become an ordinary causal variable, caused by a set of other variables that come into consideration through the expansion.

If the intervention is associated with a decision point for which no marginal distribution can be given, then an intervention remains an intervention variable, no matter how far one expands the system of variables, since no joint distribution over the intervention and its causal parents (in this case including a decision variable) can be given; only a conditional distribution of the intervention variable given a state of the decision variable can be specified. Decisions therefore remain as sources of causal influence that are external to the system, no matter how far the system is expanded.

On our definition, decision points are two levels removed from the variables in $\mathbf{V}$. This enables us to represent decisions as choices of intervention distributions, which may or may not be conferred upon the intervened variable appropriately. So there are three levels: The choice of one particular intervention distribution at the decision point, the intervention variable, which follows the intervention distribution, and the intervened variable that may or may not adhere to the intervention distribution. We thus have a very natural way to model failure to comply with treatment. Furthermore, the choice, for which one may not be able to provide a distribution, is a decision about which experiment to perform, i.e. which intervention distribution to pick, and not about which treatment to assign to which individual.

In the following two subsections we refine the definition of interventions further to consider two particular types of interventions that lie at opposite ends of a spectrum of harder to softer interventions.

2.2.2 Structural Interventions

Interventions which make the intervened variable independent of its normal causes are sometimes referred to as *randomizations* (following Fisher), *surgical* interventions (following Pearl), *ideal* interventions (following Spirtes et al.) or *independent* interventions (following Korb). I will refer to them as *structural* interventions, because they manipulate the causal structure among the variables.

Definition 2.2.2: Structural Intervention
Given a set of measured variables $\mathbf{V}$, a *structural* intervention I_s on a subset $\mathbf{S} \subseteq \mathbf{V}$ is an intervention on $\mathbf{S}$ that satisfies the following additional constraints:

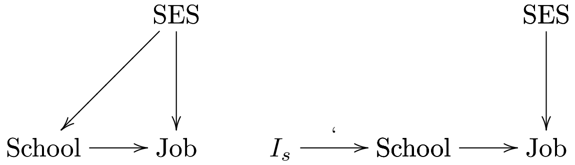
1. There is no common cause of I_s and any variable in $\mathbf{V}$ and no variable in $\mathbf{V}$ is a cause of I_s.

2. When $I_s = k$ with $k \neq 0$, I_s makes every variable in $\mathbf{S}$ independent of

its causes (breaks the edges that are incident on the variables in $\mathbf{S}$). I_s *determines* the distribution of $\mathbf{S}$, that is, in the factored joint distribution $P(\mathbf{V})$, the term $P(\mathbf{S}|pa(\mathbf{S}))$ is replaced with the term $P(\mathbf{S}|I_s = k)$, all other terms are unchanged.

The first restriction on the causal structure surrounding the intervention variable could be dropped. As a result confounded structural interventions would be possible. The limits of what can be learned about the causal structure then depend on what this restriction is weakened to. Simple causes of interventions are unproblematic, but confounding causes can weaken the discovery procedure. We will not pursue this line.

The definition of a structural intervention implies that the causal structure (as opposed to just the parameterization) is manipulated, since any causal influence on the intervened variable (other than from the intervention) is destroyed. For example, if the allocation of children to particular schools is randomized, then the social economic status of the family, which would under normal circumstances influence which school district a child lives in, is made independent of the school allocation. If social economic status is also a cause of job prospects, then the randomization destroys any confounding of school attendance and job prospects:



The manipulated causal structure is referred to as the *post-manipulation graph.*

Definition 2.2.3: Post-Manipulation Graph
Given a graph G and a set $\mathbf{S}$ of variables subject to a structural intervention, the post-manipulation graph is the graph where all the edges incident on any intervened variable ($X \in \mathbf{S}$) are removed.

The change in causal structure goes along with a change in the joint probability distribution over the variables that is specified by the Manipulation theorem (Theorem 2.1.3) given earlier. A structural intervention may manipulate multiple variables simultaneously in a correlated manner (confounding interventions). In order to achieve the surgical effect on the causal structure, a structural intervention requires that the distribution over the intervened variable is fully determined by the intervention. This can be achieved in many ways: The intervened variable can be *clamped* to a particular value (a degenerate intervention distribution), or a randomization device can be used to impose some non-degenerate distribution over the values of the intervened variable. While a randomization breaks the *causal* dependence between the intervened variable and its normal causes, probabilistic independence is only guanranteed in the large sample limit. For any finite sample the randomized distribution over the intervened variable may still turn out to be associated by coincidence with the distribution over the causes of the intervened variable. When this occurs mistakes can occur in the inference to causal structure between the intervened and other variables. How this type of situation should be resolved and what, given that samples are only ever finite, should be made of randomization in experimental practice, has led to much debate in the literature on experimental design. One solution is to balance the distribution of samples with regard to the causes of the intervened variable. If only some of the causes are known, one can balance with regard to those and randomize the intervention variable within those blocks.[17]

Fisher-type randomized trials, as they are found in medical research, and controlled experiments in which variables are fixed to particular values can be modeled as structural interventions. The advantage of a non-degenerate intervention distribution (i.e. not clamping) is that one can explore interactive causes more easily: If X only has an effect on Y when Z is in a particular state $Z = z_1$, then it is of no use to clamp $Z = z_2$. However, if we have a non-degenerate distribution over the values of Z, then the interactive cause can be discovered.

2.2.3 Parametric Interventions

Structural interventions take full control of the intervened variable, but an intervention need not be that strong. To qualify, an intervention only needs to influence the conditional distribution. This weaker form of an intervention is captured in the notion of a *parametric* intervention, also sometimes referred to as a *partial*, *soft*, *conditional* or *dependent* intervention.

Definition 2.2.4: Parametric Intervention
Given a set of measured variables $\mathbf{V}$, a *parametric* intervention I_p on a subset $\mathbf{S} \subseteq \mathbf{V}$ is an intervention on $\mathbf{S}$ that satisfies the following additional constraints:

[17] Care needs to be taken with balanced designs. Balancing does not break the causal effect, and therefore balancing with respect to e.g. a common effect can lead to erroneous inferences. While balanced designs allow for the same inferences as structural interventions in many cases, they are not identical.

1. There is no common cause of I_p and any variable in $\mathbf{V}$ and no variable in $\mathbf{V}$ is a cause of I_p.

2. When $I_p = k$ with $k \neq 0$, I_p does not make the variables in $\mathbf{S}$ independent of their causes in $\mathbf{V}$ (it does not break any edges that are incident on variables in $\mathbf{S}$).[18] In the factored joint distribution $P(\mathbf{V})$, the term $P(\mathbf{S}|pa(\mathbf{S}))$ is replaced with the term $P^*(\mathbf{S}|pa(X), I_p = k)$, where

   $$P^*(\mathbf{S}|pa(X), I_p = k) \neq P(\mathbf{S}|pa(X), I_p = 0).$$

   Otherwise all terms remain unchanged.

As with structural interventions, the first constraint (the lack of causes of the intervention variable) is not essential, but can impact discovery strategies. And as before, multiple variables can be subject to a correlated parametric intervention, and again the constraints on intervention distributions are minimal.

Although a parametric intervention does not imply any structural changes among the variables in $\mathbf{V}$ and the post-manipulation graph is only changed by the addition of the intervention variables, its influence is evident in the manipulated probability distribution. The manipulation theorem 2.1.3 can be adapted accordingly.

Theorem 2.2.5: Manipulation Theorem for Parametric Interventions Let $G = \{\mathbf{V}, \mathbf{E}\}$ be a directed acyclic graph and let $\mathbf{S}$ be the set of variables in $\mathbf{V}$ that are subject to a parametric intervention. Then G_{unman} is the unmanipulated graph corresponding to the unmanipulated distribution $P_{unman}(\mathbf{V})$ and G_{man} is the manipulated graph, in which for each variable $X \in \mathbf{S}$ an intervention variable $I_{p(X)}$ is added with $I_{p(X)} \rightarrow X$. A variable $X \in \mathbf{V}$ is in $man(\mathbf{S})$ if it is subject to an intervention, i.e. if it is a direct child of an intervention variable $I_{p(X)}$. Then

$$P_{unman}(\mathbf{V}) = \prod_{X \in \mathbf{V}} P_{unman}(X|pa(G_{unman}, X))$$

$$P_{man}(\mathbf{V}) \quad = \quad \prod_{X \in \mathbf{S}} P_{man}(X|pa(G_{unman}, X), I_{p(X)} = k) \times$$

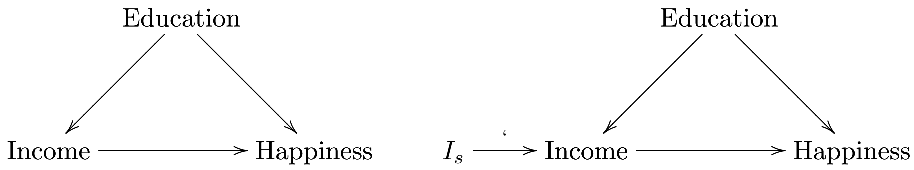
[18] Note, that we restrict parametric interventions to those types of interventions that do not break any structure but instead *only* influence parameters.

$$\prod_{X \in \mathbf{V} \setminus \mathbf{S}} P_{unman}(X|pa(G_{unman}, X))$$

for all values of $\mathbf{V}$ for which each of the conditional distributions is defined.

Since I_p does not make the variables in $\mathbf{S}$ independent of their causes (parents in the graph) I_p is not a structural intervention. Instead, I_p changes (and increases the number of) the parameters in the conditional distribution of the intervened variable on its parents.

A simple example of a parametric intervention is an intervention on the income of participants in an experiment: Rather than setting their income according to an independent probability distribution, thereby determining it completely, a parametric intervention increases their income by, say, \$1,000. This would have the effect that people with high incomes would still have high incomes, determined largely by the original causes for their high income, but we would have changed the conditional probability distribution, due to the influence of the additional money.



It is not necessary for a parametric intervention to consist of adding a constant to the value of the intervened variable. It is possible to perform a parametric intervention on binary variables as well – all that is required is that there is a change in the conditional probability distribution from the passive observational case such that:

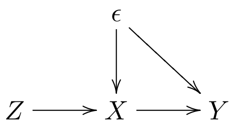
$$P(X|pa(X), I_{p(X)} = 0) = P(X|pa(X)) \neq P(X|pa(X), I_{p(X)} = 1)$$

This constraint is theoretically trivial to satisfy, since the addition of the intervention variable doubles the size of the conditional probability table of the intervened variable X.

The difficulty of a parametric intervention, however, is how to perform it when nothing is known about the causal structure among variables. If the normal causes are not known, how can one ensure in performing an intervention that some of the causal influence from the normal causes on the intervened variable is preserved? If there is reason to believe that the causal relations are additive, then there might also be reason to believe that an intervention is a further additive cause in the model. For the finite discrete case, the situation is not as clear. We do not have an account of how to *guarantee* that an intervention is parametric. We take this to be one of the major limitations of this type of intervention.

Parametric Interventions and Instrumental Variables

The technique of using parametric interventions for causal discovery is closely related to the theory of instrumental variables in economics. Suppose there are two variables, X and Y with $Y = \beta X + \epsilon$, where ϵ is an error term. A problem for the estimation of β arises when X is correlated with ϵ, the error. In such a case a consistent estimator of β can be found if there is a variable Z (called an instrument) that is correlated with X, but independent of ϵ and correlated with Y only through X, i.e. $Z \perp\!\!\!\perp Y | \{X, \epsilon\}$. Graphically, this can be represented by assuming that ϵ is a latent common cause of X and Y, as shown in the following figure:



This structure mirrors the set-up for parametric interventions: If Z were an intervention variable of X, the same independence relations would apply: The instrument is independent of ϵ and correlated with Y only through X, which is the same as requiring that the intervention is on X only and exogenous (and uncaused) with respect to the set $\mathbf{V} = \{\epsilon, X, Y\}$. The independence relations implied by this particular constellation of variables make both instrumental variables and parametric interventions a powerful discovery tool even for causally insufficient sets of variables. We return to this point in the following chapters.

The difference between the two is mainly in the semantics: Instrumental variables are generally taken to be real variables, corresponding to some causally relevant feature in the real world, whereas intervention variables can just be a feature of the model. As with instrumental variables, we assume here that there is a well-defined marginal distribution over the intervention variables, however there is generally no *requirement* for instrumental variables to have an "idle"-state for which they have no effect on the set of variables under investigation.[19]

[19] But note the comment earlier, that this idle state is a conceptual feature of interventions

Unlike intervention variables, all of their states can have an active influence on the set of variables.

2.2.4 Contrast of Structural and Parametric Interventions

Structural and parametric interventions are the two extremes on a continuum of weaker to harder interventions. The structural intervention makes the intervened variable independent of its causes whereas a parametric intervention is only an intervention on the parameterization of the causal model. There are all sorts of other interventions that have a weaker or stronger effect on the structure or parameterization, making the intervened variable independent of more or fewer of its causes. The distinction of the two extreme forms of interventions as I have presented it here can be found in work by Korb [22] and the need for a weaker version than just the structural intervention is described with various examples by Campbell [3]. Korb also discusses the possibility of mixed interventions. In that case the manipulated distribution is a mixture of two manipulated distributions, one structurally manipulated and one parametrically manipulated. As he notes, these mixtures can be represented by manipulated distributions that are somewhere between structurally and parametrically manipulated ones. It shows that there is a wide variety of additional modeling assumptions one can make about the particular nature of the manipulated distribution. The effect (and problems) with regard to causal discovery in light of the two extreme types of interventions are discussed in detail in the next chapters.

The interventions we discuss here are all designed for static models, i.e. they do not work without adjustment for time series models or dynamic Bayes nets. In dynamic models one has to account for how fast and for how long the effect of interventions percolates through a dynamic system and at what time intervals the system is being sampled. Further, one has to distinguish between an intervention at one time instance and a continuously occurring intervention. In general it is not guaranteed that the effect of an intervention will fade away, since one might have chaotic effects in a dynamic system. Here we just provide a short (and probably incomplete) list of problems that a full account of interventions on time series needs to consider.

1. What is the nature of an intervention on a time series?

and formally fairly inessential for discovery procedures.

2. Is there a single intervention at one time tick or is the intervention repeated at every time tick?

3. In order to draw inferences from data, does the data sampling have to be synchronized with the interventions?

4. Does the dynamic system return to an equilibrium state after an intervention? How is this ensured?

5. If the dynamic system is non-linear, its development might be sensitive to initial conditions and hence predictions of interventions may be impossible. How does the model accommodate this?

Similar considerations apply to interventions on cyclic or undirected models.[20] An account needs to describe how the effect of the intervention percolates through a cyclic or undirected structure, how feedback is handled and how the intervention affects cycles. We will not pursue this here.

[20] For more detail see [35].

Chapter 3

Search with Interventions: Pure Search Strategies

The following two chapters explore the possibilities and limitations of causal structure search using interventions. Both chapters consider sequences of experiments. How experiments in a sequence are chosen is specified by search strategies. The chapters are divided according to two types of search strategies that correspond in game-theoretic terms to pure and mixed strategies. Pure strategies specify one particular experiment (with probability 1) for each possible scenario, i.e. for each history of experiments already performed and each possible current state. We distinguish two types of pure strategies: fixed and adaptive ones. A fixed search strategy determines one particular sequence of experiments *before* any data is collected. No adjustments and no early stops are permitted after that, the full sequence of experiments must be performed. An adaptive strategy also announces a sequence of experiments prior to the first experiment, but the specification of the sequence of experiments can be *contingent* on experimental outcomes. That is, an adaptive strategy can specify in advance how it will adapt in light of particular outcomes of some experiment. Chapter 3 covers fixed and adaptive strategies. It is divided into four main sections: An introduction to search for causal structure that covers work that is closely related to this thesis, a section that describes the space of assumptions that will become relevant throughout this chapter and the next, and a section each on fixed and adaptive strategies. In Chapter 4 we cover mixed strategies. A mixed strategy can specify for each possible scenario (history of experiments and current state) a distribution over the possible experiments. A random sample from that distribution then determines the experiment in the sequence. For each type of search strategy we analyze how many experiments are necessary and sufficient to discover the causal structure under various assumptions. We consider structural and parametric interventions, and causally sufficient and insufficient sets of variables. We provide bounds on the number of experiments and search strategies that implement sequences of experiments that respect the bounds

3.1 Bayesian Searches with Interventions

Results in this thesis are most closely related to work by Tong & Koller [44] and Murphy [29] on selecting the best next experiment to perform when searching for causal structure. Both use a Bayesian approach to structure learning and both use information theoretic measures to identify the optimal next experiment. Their basic framework assumes a prior distribution $P(\mathcal{G})$ over the space of directed acyclic graphs over a given set of N variables, and a prior over the parameterization for each graph $P(\theta_G|\mathcal{G} = G)$, both of which are updated given data D from an experiment $\mathcal{E}$ that manipulates a subset of the variables $\mathbf{V}$ in the graph. Given data D from experiment $\mathcal{E}$, the distribution over graphs is updated by Bayes Theorem for each $G \in \mathcal{G}$:

$$P(G|\mathcal{E}, \theta_G, D) = \frac{P(\theta_G|G, \mathcal{E}, D)P(G)}{\sum_{G' \in \mathcal{G}} P(\theta_{G'}|G', \mathcal{E}, D)P(G')}$$

Concretely, this means that for each possible graph $G \in \mathcal{G}$, the manipulated graph $G_{\mathcal{E}}$ is computed given the intervention set of experiment $\mathcal{E}$. Then, for each graph $G_{\mathcal{E}}$ the parameters of $G_{\mathcal{E}}$ that are not manipulated by the experiment are updated given the data, and lastly, with the updated parameters in place, the distribution over the graphs can be updated. Technically, it is an update of the distribution of manipulated graphs, but the update of the manipulated graphs reflects back on the unmanipulated graphs, as does the update on the parameters, for those parameters that were not affected by the manipulation. Two graphs that have the same manipulated graph receive the same boost from the data, since the likelihood of the data is the same for both graphs in the experiment (assuming the relevant parameters are the same). The posterior distribution over graphs after one experiment becomes the prior distribution for the next experiment and the process iterates.

Both Tong & Koller and Murphy use an information theoretic measure to determine the next experiment. They perform the experiment that minimizes a form of entropy in the posterior distribution over possible graphs. Tong & Koller minimize the average of the entropy for each pair of variables, while Murphy minimizes the entropy of the posterior over graphs directly.

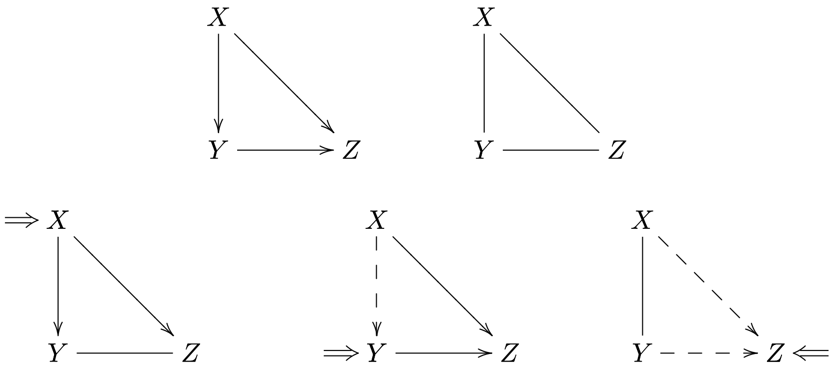
The computation involved in the update is enormous: For the exact computation of the posterior one has to compute the posterior value for each possible DAG over the variables. To do so, one has to integrate out the model parameters for each DAG. The integrals are simple, but large in number, and the number of DAGs grows super-exponentially in the number of vertices. For the choice of the best next experiment one theoretically has to consider each possible intervention (of which there are 2^N, where N is the number of variables) and determine its expected impact on the entropy.

Both Tong & Koller and Murphy do not perform the exact computation. Tong & Koller restrict themselves to networks that have a total ordering over the variables. Murphy uses sampling techniques to estimate the result. Both parties only consider structural interventions that clamp the intervened variables to a particular value and only consider intervention sets of up to 4 variables. Their choice of experiment is the best next experiment, which need not be the best choice for a sequence of experiments.

Tong & Koller and Murphy both provide a fully Bayesian account of how active search for causal structure using sequences of experiments can be performed. Our approach considers the constraint based counterpart.

3.2 Search Space Assumptions

By imposing extra constraints on the causal system interventions provide additional leverage for causal structure search. As indicated in the previous chapter, an intervention can help distinguish causal structures that appear equivalent in passive observational data. In general, it is not the case that a single intervention on one variable is sufficient to uniquely identify the one true causal structure in an observational Markov equivalence class (OME). For example, all the complete graphs over three variables imply the same independence constraints under passive observation and consequently form an OME. If the first graph in the top row is the true graph, then only its OME (represented by the pattern on the right in the first row) can be discovered by passive observation using independence tests. A single intervention on one variable will help discover some edges, but it does not uniquely identify the true causal structure in all cases. The three graphs in the second row below show the post-manipulation equivalence classes for a single structural intervention on X, Y and Z, respectively (the intervention variables are ommitted for simplicity):



Restricting ourselves to independence tests only, in the case of a structural intervention on X, the XY-edge and the XZ-edge can be determined, but only the adjacency, not the orientation of the YZ-edge can be discovered. For a structural intervention on Y, the YZ-edge can be determined and the XZ-edge can be determined ($X - Z - Y$ form an unshielded collider in the post-manipulation graph since the structural intervention on Y destroys the XY-edge). However, for the XY-edge, it is only known that either there is an incoming edge or there is no edge at all (represented by the dashed arrow), since X and Y appear independent in the manipulated distribution. Similarly, for an intervention on Z, only disjunctive information is discovered for the XZ- and YZ-edges, and only the adjacency is recovered for the XY-edge. If it is known that the graph is complete, then only in the case of an intervention on Y are we able to identify the graph uniquely. In all other cases a further intervention is necessary. It turns out that any pair of experiments involving two different single interventions is sufficient and two experiments are in the worst case necessary to identify the graph uniquely. To analyze sequences of experiments formally, we use the following definition of an experiment:

Definition 3.2.1: Experiment

An experiment $\mathcal{E}_i$ on a set of variables $\mathbf{V}$ is represented by a triple of sets $\mathcal{E}_i = (\mathbf{S_i}, \mathbf{U_i}, \mathbf{Pol_i})$, where $\mathbf{S_i}$ represents the subset of $\mathbf{V}$ that is subject to an intervention in $\mathcal{E}_i$, $\mathbf{Pol_i}$ is the corresponding set of intervention variables, and $\mathbf{U_i}$ contains the remaining passively observed variables. $\mathbf{U_i} \cup \mathbf{S_i} = \mathbf{V}$, $\mathbf{U_i} \cap \mathbf{S_i} = \emptyset$ and $\mathbf{V} \cap \mathbf{Pol_i} = \emptyset$.

According to this definition a passive observation is considered to be one experiment (with an empty intervention set). Furthermore, it is possible that an experiment involves multiple independent, but simultaneous interventions on the set of variables. We say that an experiment is a single intervention experiment if $\mathbf{S}$ is a singleton set. We refer to a multiple intervention experiment if $\mathbf{S}$ contains several variables and each has its own corresponding intervention variable in $\mathbf{Pol}$.

The results that we present in the following two sections and in the next chapter are sensitive to assumptions that are made about the search space. Since we consider sequenes of experiments with a variety of search strategies, with different types of interventions and different numbers of intervention variables, assumptions can be combined in many ways. Many of the assumptions are not independent, their effect on the search depends on what other assumptions are made. Throughout this thesis we take four assumptions about the causal structure to be fundamental: Causal Markov, Causal Faithfulness, acyclicity and knowledge of the distribution family. The first two were presented in Section 1.2, so we just add the third and fourth here:

Assumption 3.2.2: Acyclicity of Causal Structure
The true causal structure over a set of variables is acyclic.

Assumption 3.2.3: Distribution Family
We assume that the true model is a discrete binary model or a linear model with normal errors, and that it is known, which of the two it is.

We take the causal Markov assumption (Assumption 1.2.1) to be a core characteristic of causal processes. That is, if a causal system violates the Markov assumption, then there is something fundamentally wrong in trying to describe the system as causal, e.g. the variables are misspecified, the value space of the variables is inappropriate or the process is not causal. Causal faithfulness (Assumption 1.2.2) is not essential for all results. In particular, there are results similar to ours by Nyberg and Korb [31], that do not assume faithfulness. But unless stated otherwise, it is assumed. Acyclicity is, of course, violated in many real cases. We make this assumption for simplicity, since the difficulties arising from cyclic causal structures and the problems for discovery procedures involving interventions on cyclic structures go beyond the scope of this thesis.

We consider two types of models. For the discrete case we analyze models where each variable has two states, for the continuous case we consider linear models with normal errors, i.e. the value of each variable is determined by a linear sum of the values of its parents plus an (independent) error term that is normally distributed with mean zero: For each variable $X \in \mathbf{V}$, $X = \sum_i c_i Y_i + e$ with $Y_i \in pa(X)$. Many of our results extend to other types of models and distributions as well, but we do not consider them explicitly here. In particular, most results also hold for discrete models with several (as opposed to just two) states and the results on linear normal models extend to continuous additive models as well.

Assuming Markov and faithfulness, probabilistic independencies can be used to identify d-separation relations. In particular, the independence of two variables X and Y for some conditioning set $\mathbf{C}$ implies that there is no edge $X \rightarrow Y$ or $Y \rightarrow X$ in graph G. However, if one cannot find a conditioning set that makes the two variables independent, it follows that $X \rightarrow Y$ or $Y \rightarrow X$ *only if* one also assumes causal sufficiency, i.e. that there are no latent common causes. If there is an unmeasured variable L with $X \leftarrow L \rightarrow Y$ then X and Y are correlated despite the fact that there is no conditioning set (involving observed variables only) that makes them independent. The assumption of causal sufficiency has significant impact on what can or cannot be discovered about a particular causal structure. Causal sufficiency belongs to the set of assumptions we switch on and off: We present results that rely on these assumptions and results that do not depend on their satisfaction.

Assumption 3.2.4: Causal Sufficiency
There are no latent common causes of the set of variables $\mathbf{V}$.

Assumption 3.2.5: Oracle
The experiment returns the independence relations (or, in Section 3.3.3, correlation values) true in the manipulated population distribution.

Assumption 3.2.6: Independence Tests
(Conditional) independence tests are the only admissible means to identify causal structure given a distribution over the variables.

Assumption 3.2.5 allows an analysis of the discovery problem independent of statistical variability. In this way the combinatorical problems arising from the combination of different experiments can be separated from sampling issues. Assumption 3.2.6 restricts the search methods to conditional independence tests. The main appeal of conditional independence tests is that they are distribution free and relate most directly to the qualitative nature of the causal structure. There are, of course, other – even distribution free – tests to search for causal structure given a particular – possibly manipulated – distribution over a set of variables, and we will return to these in the cases where independence tests turn out to be insufficient. No assumptions (other than acyclicity) are made about the nature of the causal structure and no background knowledge or time order information is presumed.

For different sets of assumptions and different types of interventions we give bounds on the number of experiments necessary and sufficient to discover the causal structure among N variables. These bounds are worst case bounds on different types of search strategies. A search strategy determines the sequence of experiments, i.e. which set of variables is subject to an intervention at which stage in the sequence.

Definition 3.2.7: Search Strategy
A search strategy is a complete plan of which experiment will be performed next at any point in the sequence of experiment and for any history of experimental outcomes that may occur.

We consider three general families of strategies: fixed, adaptive and, in the next chapter, mixed strategies. Strategies differ in how experiments are chosen given the available information about the true underlying causal structure from earlier experiments. For any set of search space assumptions, we describe three aspects: the bound on the number of experiments sufficient and (where possible) in the worst case necessary to discover the causal structure given this set of assumptions, a strategy that specifies a sequence of experiments that stays within the bound, and (in Chapter 6) an algorithm that combines the information from the different experiments to determine the causal structure. The strategies we specify respect the bounds, but are not always unique, they do not necessarily minimize the number of variables subject to intervention, nor do they always minimize the size of the largest intervention set in the sequence of experiments. In particular, if one is willing to perform more experiments, then in most cases one can do better on these other measures.

3.3 Fixed Search Strategies

We first consider fixed search strategies. Fixed search strategies specify one particular sequence of experiments. Which experiment is performed only depends on how many variables there are and what the current index in the sequence of experiments is. One can think of a fixed strategy as announcing the sequence of experiments before any experiments are performed. Later experiments cannot be adapted to take the results from earlier ones into account. Fixed strategies can be used to identify the longest sequence of experiments that may be required to discover any particular graph. Such a mini-max guarantee provides a worst-case bound for a search procedure: Since fixed strategies are independent of particular experimental outcomes, there is for any number N of variables a fixed sequence of experiments that guarantees that no matter what the true graph is, it will always be determined uniquely within the number of experiments specified by the fixed strategy. To make the worst case bound tight, the fixed strategy should further guarantee that for any fixed sequence of experiments that is shorter there is some graph which that sequence will not uniquely determine. Importantly, the fixed strategy may not depend on a "lucky" or adaptive choice of intervention set.

In the following five subsections we consider structure search with fixed strategies under a variety of different combinations of assumptions and different types of interventions. We first consider fixed strategies using structural interventions and then fixed strategies using parametric interventions. Within each of these cases we consider causally sufficient and insufficient sets of variables and strategies with single and strategies with multiple simultaneous interventions in each experiment. In a third subsection we consider search strategies based on tests of differences in correlations instead of independence tests, in the fourth section we assume that particular background knowledge is available and in the fifth section we briefly consider some restrictions that the search strategy may be subject to. Unless specified otherwise, all results hold for both discrete and linear models.

3.3.1 Fixed Strategies with Structural Interventions

Under assumptions 1.2.1-3.2.6 above and allowing only a single structural interventions on one variable per experiment, we get the following bound on the number of experiments necessary and sufficient to learn the causal structure among N variables.[1]

Theorem 3.3.1: (fixed strategy) Single Structural Interventions, Causally Sufficient
$N-1$ experiments are sufficient and in the worst case necessary to determine the causal graph among $N > 2$ variables[2] when only a single structural intervention is allowed in each experiment.

This bound implies that no matter what the true underlying causal structure is, there is a fixed sequence of experiments that guarantees that this graph can be uniquely identified. Furthermore, for any shorter sequence of experiments involving single interventions only, there is a graph – in this case a complete graph – the sequence cannot identify uniquely. The theorem provides a worst case bound in the sense that it uniquely identifies any possible (acyclic) causal structure over N variables independently of how the selection of variables for intervention relate to the causal structure. The fixed sequence of experiments of length $N-1$ that guarantees the bound is specified by strategy 3.3.2:

Strategy 3.3.2: (fixed) Single Structural Intervention, Causally Sufficient
Given N causally sufficient variables $X_1, \ldots, X_n$, let the sequence of experiments $\mathcal{E}_1, \ldots, \mathcal{E}_{N-1}$ be such that $\mathcal{E}_i = (\mathbf{S_i}, \mathbf{U_i}, \mathbf{Pol_i})$ with $\mathbf{S_i} = \{X_i\}$ and where $I_{s(X_i)} \in \mathbf{Pol_i}$ is a structural intervention.

The strategy is not sensitive to the order of the variables subject to intervention, nor does it matter which particular variable X_i is the variable X_n that is not subject to an intervention. The strategy is unique (up to re-ordering and re-naming of variables) in the sense that it is the only fixed strategy that guarantees to recover every causal structure within the bound.

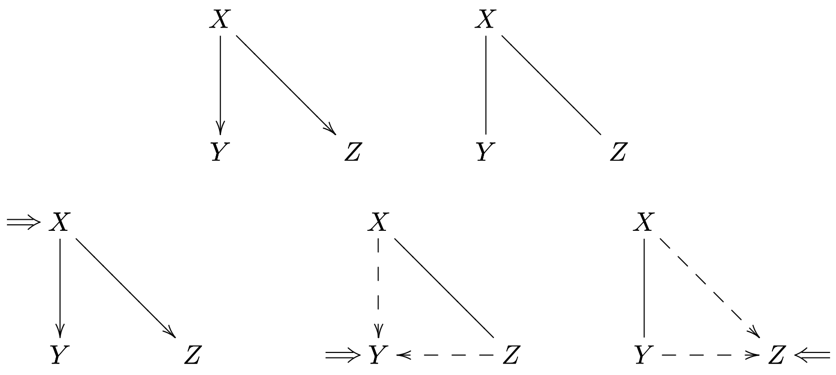
In this type of worst case analysis there is no advantage to passive observation. Any fixed sequence of experiments involving a passive observational experiment does not recover every graph or it exceeds the bound in the worst

[1] All proofs are given in the appendix. This result was first presented in [11].

[2] For $N = 2$, two experiments are sufficient and in the worst case necessary.

case.

We illustrate the result with a simple example over three variables:

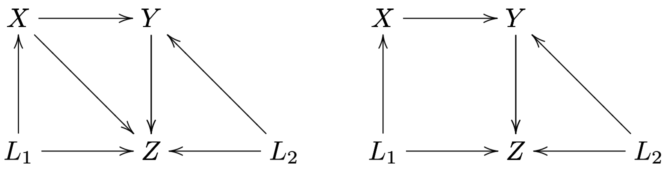


Suppose the true graph is the one shown on the top left. The remaining four graphs show what knowledge can be obtained about the true graph under passive observation (top right) and intervention on X, Y and Z, respectively, on the bottom row. Only an intervention on X identifies the true graph uniquely. All other experiments require a second experiment. In all cases, a second experiment consisting of an intervention on X would (of course) suffice. For the middle graph in the second row a passive observation or an intervention on Z would also suffice; for the graph on the bottom right a passive observation, or an intervention on either X or Y would suffice to discover the true causal graph. If no background knowledge is available, one cannot guarantee a priori that an intervention on X is sufficient (since Y or Z may be the common cause instead), so since $N = 3$, $N - 1 = 2$ experiments are sufficient and in the worst case necessary.

If assumption 3.2.4 (causal sufficiency) is dropped, then, in the worst case, discovery is no longer possible:

Theorem 3.3.3: (fixed strategy) Single Structural Intervention, Causally Insufficient
Given a causally insufficient set of variables, no sequence of experiments is sufficient to determine the worst case causal graph among N variables when only a single structural intervention is permitted in each experiment.

The negative result derives directly from the fact that in the case of a causally insufficient set of variables we cannot use the lack of a conditioning set that makes two variables conditionally independent, to establish an adjacency between the variables: two variables might be non-adjacent, but dependent for every conditioning set due to a latent common cause. Consider the following two graphs over three variables with two latent variables L_1 and L_2:



No sequence of single intervention experiments distinguishes these two graphs.

If, in contrast, causal sufficiency is maintained, but we allow the possibility that multiple variables can be subject to an intervention simultaneously and independently, we can reduce the bound on the number of experiments significantly:[3]

Theorem 3.3.4: (fixed strategy) Multiple Structural Interventions, Causally Sufficient
$\lfloor \log_2(N) \rfloor + 1$ experiments are sufficient and in the worst case necessary to determine the causal graph among N variables when multiple simultaneous and independent structural interventions are allowed in each experiment.

The worst case graph is again a complete graph over the variables, but even that is discovered within the limits of the bound by a strategy of the type of Strategy 3.3.5:

Strategy 3.3.5: (fixed) Multiple Structural Interventions, Causally Sufficient
Given N causally sufficient variables $X_1, \ldots, X_N$, let the sequence of experiments consist of $k = \lfloor \log_2(N) \rfloor + 1$ experiments $\mathcal{E}_1, \ldots, \mathcal{E}_k$ with $\mathcal{E}_i = (\mathbf{S_i}, \mathbf{U_i}, \mathbf{Pol_i})$ such that for each pair of variables X, Y in $\mathbf{V}$ one of the following holds:

1. There is an experiment $\mathcal{E}_i$ such that $X \in \mathbf{S_i}$ and $Y \in \mathbf{U}_i$ and an experiment $\mathcal{E}_j$ such that $X \in \mathbf{U}_j$ and $Y \in \mathbf{S_j}$. That is, one experiment where X is subject to an intervention and Y is not, and one where Y is subject to an intervention and X is not.

[3] This result was first presented in [10].

2. There is an experiment $\mathcal{E}_i$ such that $X \in \mathbf{S_i}$ and $Y \in \mathbf{U}_i$ and an experiment $\mathcal{E}_j$ such that $X \in \mathbf{U}_j$ and $Y \in \mathbf{U}_j$. That is, one experiment where X is subject to an intervention and Y is not, and one where both X and Y are passively observed.

We do not specify a particular sequence of experiments in the strategy, since for different values of N there can be several ways of satisfying the constraints. One way of satisfying these constraints is that each experiment intervenes on $\lfloor N/2 \rfloor$ variables simultaneously, with different combinations each time (see Figure 3.1 for examples for $N = 8$ and $N = 7$).

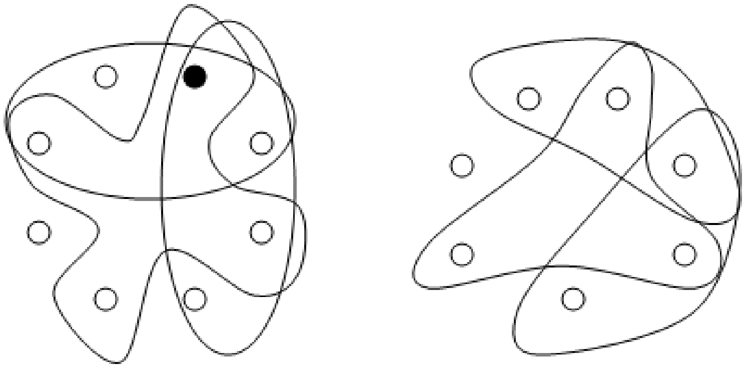


Figure 3.1: *Intervention sets for Strategy 3.3.5 for $N = 8$ and $N = 7$. For $N = 8$ only three of the four intervention sets are shown, but one variable (marked in black) is contained in all of them. So there are in this case many choices for the fourth intervention set, as long as it does not contain the black vertex.*

In general this approach can be formalized by using Cantor sets: For N variables, the first intervention set is $\mathbf{S}_1 = \{X_i | 0 < i \leq \lfloor \frac{N}{2} \rfloor\}$. The kth intervention set ($k > 2$) is determined by selecting those X_i for intervention, whose i are covered by the mapping of the interval $(0; N/2]$ by $k - 1$ iterated applications of the functions

$$\begin{aligned} f_1(x) &= 0.5x \\ f_2(x) &= 0.5x + 0.5N \end{aligned}$$

The result are intervention sets that follow the construction of a particular type of Cantor set (see the first example with $N = 7$ in the following table). When N is a power of 2, an additional passive observation is required (see the first example with $N = 8$ in the table). But this approach is not unique. Depending on the number of variables, there can be quite a bit of flexibility in the sequence of experiments, as the following examples in the table show.

Variables	Experiments	Intervention Sets
$N = 7$	3	$\mathbf{I}_1 = \{X_1, X_2, X_3\}$ $\mathbf{I}_2 = \{X_1, X_4, X_5\}$ $\mathbf{I}_3 = \{X_2, X_4, X_6\}$
$N = 7$	3	$\mathbf{I}_1 = \{X_1, X_2, X_3\}$ $\mathbf{I}_2 = \{X_2, X_4, X_5\}$ $\mathbf{I}_3 = \{X_3, X_4, X_6\}$
$N = 8$	4	$\mathbf{I}_1 = \{X_1, X_2, X_3, X_4\}$ $\mathbf{I}_2 = \{X_1, X_2, X_5, X_6\}$ $\mathbf{I}_3 = \{X_1, X_3, X_5, X_7\}$ $\mathbf{I}_4 = \emptyset$
$N = 8$	4	$\mathbf{I}_1 = \{X_1, X_2, X_3\}$ $\mathbf{I}_2 = \{X_1, X_4, X_5\}$ $\mathbf{I}_3 = \{X_2, X_4, X_6\}$ $\mathbf{I}_4 = \{X_7\}$
$N = 15$	4	$\mathbf{I}_1 = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$ $\mathbf{I}_2 = \{X_1, X_2, X_3, X_8, X_9, X_{10}, X_{11}\}$ $\mathbf{I}_3 = \{X_1, X_4, X_5, X_8, X_9, X_{12}, X_{13}\}$ $\mathbf{I}_4 = \{X_2, X_4, X_6, X_8, X_{10}, X_{12}, X_{14}\}$

There are three forms of flexibility in arranging the intervention sets: First, the intervention sets are insensitive to renaming of the variables. It does not matter which variable is X_1 and which is X_2, as long as different indices refer to different variables. Second, if we keep the size of the intervention set in each experiment constant, then there is *generally* some flexibility in exactly which combinations of variables are subject to interventions. We can see an example of this in the first two examples with $N = 7$: The intervention sets are the same size in both cases (3,3,3), but they contain different variables in the sense that the sets are not equal up to renaming. Third, there *sometimes* is flexibility in the size of the intervention sets. This is evident in the two examples with $N = 8$. The distribution over the intervention sets is different: (4,4,4,0) vs. (3,3,3,1). This flexibility does not exist for $N = 7$. We cannot have an intervention set of size two or four without exceeding the bound on the number of experiments.

The underlying intuition is: the smaller $2^m - N$ is, where m is the smallest integer such that $2^m > N$, the more flexibility there is in the distribution of the size of the intervention sets. When N is a power of 2, then the number of experiments increases by one. So, the closer $2^m - 1$ and N are, the closer we approach the bound of what can at best be learned from m experiments, and hence flexibility decreases.

These features could become extremely relevant when considering cost functions. If the cost function is over the number of variables intervened (rather than, say, number of experiments, or sample size), then it is desirable to keep the total size of the intervention sets low (as in the second example with $N = 8$).

If we now drop assumption 3.2.4 (causal sufficiency) while keeping multiple simultaneous interventions, then, unlike for the single intervention case, causal discovery remains possible:

Theorem 3.3.6: (fixed strategy) Multiple Structural Interventions, Causally Insufficient
Given a causally insufficient set of variables, N experiments are sufficient and in the worst case necessary to discover the causal structure among the N observed variables if multiple variables can be subject to a structural intervention simultaneously and independently in each experiment.

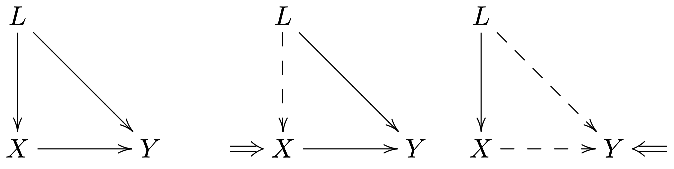
In this case, only the causal structure among the *observed* variables is discovered. Given a set of variables, the structure among the observed variables is the subgraph that only contains vertices that are measured and edges that connect two measured vertices. In general, the location or presence of latent common causes cannot be discovered on the basis of independence tests alone.[4] The fixed strategy corresponding to the bound is:

Strategy 3.3.7: (fixed) Multiple Structural Interventions, Causally Insufficient
Given N causally insufficient variables $X_1, \ldots, X_N$, let the sequence of experiments $\mathcal{E}_1, \ldots, \mathcal{E}_N$ be such that each $\mathcal{E}_i$ consists of a simultaneous structural intervention on $N - 1$ variables, leaving out a different one each time.

We conjecture that this strategy of experiments is the only one that guarantees attainment of the bound for every possible DAG over N variables. It

[4] Note that we only consider latent common causes, we do not consider latent embedded variables, even though there might be some circumstances in which they can be detected. In general, discovery of latent embedded variables is still largely an open problem.

is possible to discover the graph with other fixed strategies that do not require an intervention on $N-1$ variables for *each* experiment. But for such fixed strategies more than N experiments are required and at least one of them must still intervene on all but one variable. We illustrate the bound for the simplest possible case of two variables, where L is a latent common cause:



Suppose the true graph is the one on the left, then the remaining two graphs show the manipulated structures for the two possible (informative) interventions. An intervention on Y is insufficient and a second experiment must be performed. After the second experiment, only the structure among the observed variables (X and Y) is guaranteed to be discovered.

3.3.2 Discovery with Parametric Interventions

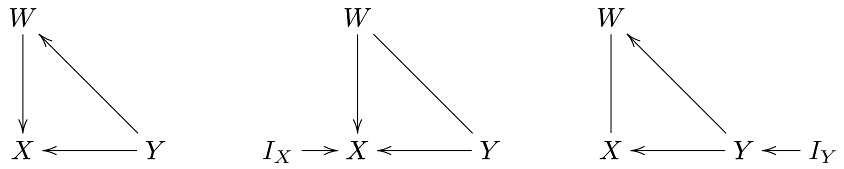
Parametric interventions do not destroy causal structure and therefore can be used (and combined) more efficiently than structural interventions. In particular, when two variables are both subject to a structural intervention, then all information about the causal structure between them is lost. This is not the case for parametric interventions and consequently, if they can be performed, parametric interventions result in quite different bounds on the number of experiments. Under assumptions 1.2.1-3.2.6, but now with parametric interventions, we get:[5]

Theorem 3.3.8: (fixed strategy) Single Parametric Intervention, Causally Sufficient
$N-1$ experiments are sufficient and in the worst case necessary to determine the causal graph among N variables when only a single parametric intervention is allowed in each experiment.

For single parametric interventions the bound is no different to Theorem 3.3.1 (single structural interventions) and the search strategy is the same as Strategy 3.3.2, just with parametric instead of structural interventions. The

[5] These results were first presented in [12].

strategy remains insensitive to order and which variable is not subject to an intervention. We give an example with three variables:



The first graph is assumed to be the true graph among the variables. Since the parametric interventions do not destroy any causal structure, no edges are missing in the manipulated graphs shown for a parametric intervention on X and Y, respectively. Consequently, adjacencies can be determined by any standard structure search method for passive observational data. Orientations can be determined by a collider test: In the first experiment I_X, X and W form an unshielded triple, since I_X and X are adjacent, and X and W are adjancent, but I_X and W are (by construction) non-adjacent. Since we find that $I_X \perp\!\!\!\perp W$, but $I_X \not\perp\!\!\!\perp W|X$, we can orient the triple as a collider. Similarly for I_X, X and Y. We cannot orient the WY-edge. However, in the second experiment I_Y, Y and W also form an unshielded triple, but in this case we find $I_Y \perp\!\!\!\perp W|Y$. Consequently, the triple is not a collider, and since we know that $I_Y \rightarrow Y$, we can orient $Y \rightarrow W$. Combined, the experiments resolve the causal structure uniquely.

In the way we have described the example, the collider tests depend on a marginal distribution on the intervention variables. But in principle that is not necessary, since the same information can be obtained by considering differences in conditional distributions. For example, in the first experiment we know all the adjancencies. If we find that $P(Y|I_X = k_1) \neq P(Y|I_X = k_2)$, then the X, Y-edge can be oriented from X to Y, if the quantities are equal, then one can conclude $X \leftarrow Y$ (which would be true in the specific case of this example).

So far, we have not seen any difference in the number of experiments between parametric and structural interventions. But if we consider multiple simultaneous parametric interventions, we do much better than before:

Theorem 3.3.9: (fixed strategy) Multiple Parametric Interventions, Causally Sufficient

One experiment is necessary and sufficient to determine the causal graph among N variables when multiple simultaneous parametric interventions are allowed in each experiment.

The corresponding search strategy is trivial:

Strategy 3.3.10: (fixed) Multiple Parametric, Causally Sufficient
Given N causally sufficient variables $X_1, \ldots, X_n$, let the sequence of experiments consist of one experiments $\mathcal{E}_1$ such that all but one variable is subject to a parametric intervention.

It does not matter which variable is not subject to an intervention. Under the given assumptions one may also subject all variables to parametric interventions simultaneously (instead of just $N-1$), but it is not necessary. The huge reduction in the number of experiments results from the fact that parametric interventions can be combined independently of each other, as they do not destroy causal structure. The reduction in the number of experiments comes at a price: In comparison to the structural intervention strategies, substantially more conditional independence tests may be needed to perform all the collider tests. In the one experiment of the above strategy all variables may have to be tested for colliders on the basis of the data from a single experiment. The collider tests generally involve higher order independence tests, which may not be as reliable.

If we drop assumption 3.2.4 (causal sufficiency), the case for parametric interventions becomes substantially more complicated. In general, just given assumptions 3.2.5 and 3.2.6 above, parametric interventions are insufficient to uniquely identify the worst case possible graph, no matter whether we consider single or multiple parametric interventions per experiment. So, for completeness, we can state:

Theorem 3.3.11: (fixed strategy) Parametric Interventions, Causally Insufficient
No sequence of experiments is sufficient to determine the worst case causal graph among N causally insufficient variables if only parametric interventions (single or multiple) are allowed in the experiments.

However, it is not the case that all the power of identifying causal connections is suddenly lost. Two example cases will illustrate the problem. Consider the first two graphs in Figure 3.2.

There are no independence constraints among the observed variables (and intervention variables) that distinguish the first two graphs. However, the third graph in the same figure can be distinguished from both of the first two. $I_1 \perp\!\!\!\perp X_2$ and $I_1 \perp\!\!\!\perp I_2 | X_2$ hold for the third graph but do not in the first two graphs.[6] That is, parametric interventions imply their own equivalence classes, which are different from the equivalence classes resulting from passive observation or structural interventions. Theorem 3.3.11 states that no sequence of experiments involving parametric interventions results in a singleton equivalence class if all graphs over N variables are considered possible.

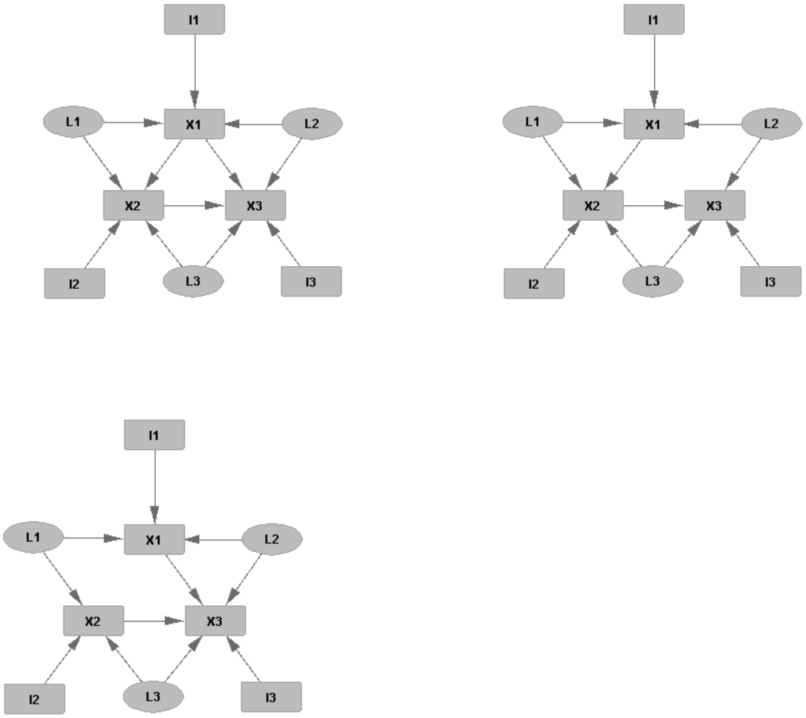


Figure 3.2: *No independence constraints among the observed variables (and intervention variables) distinguish the first two graphs, but there exist independence constraints among the observed variables (and intervention variables) that distinguish the third graph from both of the first graphs.*

[6] The number of independence constraints that had to be checked is enormous. We checked them automatically by using a feature of the Causality Lab program (http://www.phil.cmu.edu/projects/causality-lab/).

The conditions for recovering the causal structure among causally insufficient variables uniquely when only parametric interventions are available, depends on a result by Verma and Pearl [47] on inducing paths.[7]

Definition 3.3.12: Inducing Path
If G is a directed acyclic graph over a set of variables $\mathbf{V}$, $\mathbf{O}$ is a subset of $\mathbf{V}$ containing X and Y, and $X \neq Y$, then an undirected path p between X and Y is an inducing path relative to $\mathbf{O}$ if and only if every member of $\mathbf{O}$ on p except for the endpoints is a collider on p, and every collider on p is an ancestor of either X or Y.

If we can determine adjacencies among a set of causally insufficient variables, then orientations can be determined by collider tests. Adjacencies can be determined when there are no inducing paths (Theorem 6.1 in [43]). So we can characterize the cases when parametric interventions are sufficient for causal discovery on causally insufficient sets of variables:

Theorem 3.3.13: Parametric Interventions and Inducing Paths
Let G be a graph over a set of variables $\mathbf{V}$ and let $\mathbf{O}$ be a subset of $\mathbf{V}$ containing the observed variables. Let G_{man} be the graph G where each variable $X \in \mathbf{O}$ is extended with an intervention variable $I_X \rightarrow X$. The subgraph $G_{\mathbf{O}}$ of G over the observed variables can be uniquely determined by parametric interventions on each variable in $\mathbf{O}$ if and only if for each pair of variables $X, Y \in \mathbf{O}$ that are non-adjacent in G, there is no inducing path between I_X and Y and no inducing path between I_Y and X relative to $\mathbf{V} \cup \{I_X | X \in \mathbf{O}\}$ in G_{man}.

Needless to say, this theorem only provides the conditions, it does not give any indication of how one can ensure that these conditions are satisfied.

We can thus summarize our results on the number of experiments sufficient and in the worst case necessary when **tests of independence** are used to discover causal structure:

Type of Experiment	Causal Sufficiency	Structural	Parametric
Single	Yes	$N-1$	$N-1$
Multiple	Yes	$\log_2(N)+1$	1
Single	No	impossible	impossible
Multiple	No	N	impossible

[7] The definition is taken form [43].

For each of these strategies, the maximum number of interventions necessary in some experiment in a strategy that satisfies the corresponding above bound is given in the following table:

Type of Experiment	Causal Sufficiency	Structural	Parametric
Single	Yes	1	1
Multiple	Yes	$N/2$	$N-1$
Single	No	–	–
Multiple	No	N-1	–

3.3.3 Tests of Differences in Correlation

The previous section was based on Assumption 3.2.6, i.e. that tests of independence are the only permissible means to search for causal structure in the available data. Independence tests do poorly when the set of variables is not causally sufficient. We either are unable to uniquely discover the causal structure or we have to intervene on all but one variable in each experiment. If instead of Assumption 3.2.6 we allow tests that check for differences in correlation (in the case of linear models) and use that information to differentiate causal structures, then results turn out differently again. In the causally sufficient cases (Theorems 3.3.1, 3.3.4, 3.3.8 and 3.3.9), this change of assumptions makes no difference to the number of experiments needed for causal discovery *in the worst case* – independence tests are all one needs. Of course, there may be particular causal structures that imply certain constraints on correlations that can be used to identify the causal structure, where independence tests fail. But in the worst case, for a fixed strategy, differences in correlations do not provide any benefits in terms of the length of the sequences of experiments, when the set of variables is causally sufficient. However, in the case of causal insufficiency, the negative results of Theorem 3.3.3 can be avoided – at least for linear models:

Theorem 3.3.14: (fixed strategy) Single Structural Intervention, Correlation-Test, Causally Insufficient
Given a set of N causally insufficient variables, N experiments are sufficient and in the worst case necessary to determine the causal graph among the observed variables when only a single structural intervention is allowed in each

experiment and the model is linear.

Furthermore, correlation tests combined with these experiments enable us to recover the presence and location of latent common causes:

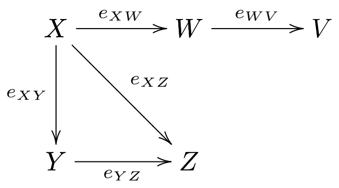
Theorem 3.3.15: Search for Latent Common Causes: Single Structural Interventions
Given a set of N causally insufficient variables and assuming the model is linear, N experiments, with a single structural intervention only per experiment, are sufficient and in the worst case necessary to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

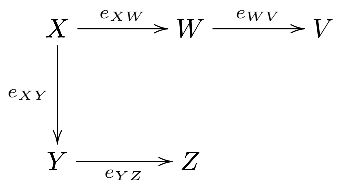
In Chapter 6 (Algorithms) we provide an algorithm that is able to identify latent variables. By integrating these results with algorithms that search for structure among latent variables, we conjecture that one is able to discover the structure among observed variables, the presence and location of latent variables *and* the structure *among* latent variables.

The bound signals a significant increase in the information that can – in theory – be obtained from a sequence of single structural interventions. Not only does this bound imply that it is possible to discover the causal structure among the observed variables with single interventions despite the fact that the set of variables is causally insufficient, one can even discover the location of latent variables. However, to do so, requires a substantial amount of analysis of the data, as will become evident from the presentation of the algorithms in Chapter 6. The sequence of experiments for a search strategy that respects the previous two theorems is straightforward:

Strategy 3.3.16: (fixed) Single Structural Intervention, Correlation-Test, Causally Insufficient
Given N causally insufficient variables $X_1, \ldots, X_N$, let the sequence of experiments be such that each experiment $\mathcal{E}_i$ is a structural intervention on a different X_i.

The strategy essentially matches Strategy 3.3.2 with one extra experiment. We illustrate the result with one very simple example over five variables. More detail is given in the corresponding section in Chapter 6 and the Appendix.





Suppose the true graph over V, W, X, Y, Z with its edge-coefficients is given on the left, and assume that each pair of variables is confounded by a latent common cause not shown in the diagram. After performing $N = 5$ experiments, each involving a structural intervention on one variable, we can define a partial ordering ($\succ$) over the variables, such that $X_i \succ X_j$ if and only if $X_i \not\perp X_j$ in $\mathcal{E}_i$, the experiment in which X_i is subject to a structural intervention. In our example we have $X \succ W \succ V$ and $X \succ Y \succ Z$. From this partial order we can construct a partial order graph (POG)[8], which contains a direct edge whenever two variables follow each other directly in the partial ordering. The POG for the partial order here is shown as the second graph. The correlations due to the direct edges between two variables can be determined in the experiment in which the cause variable is subject to a structural intervention. The POG is a subgraph of the true graph. Now, by considering connections in the POG in an order over the paths that ensures that connections of variables closer in the graph hierarchy are determined before variables are considered that are further apart in the graph hierarchy[9], we can test for other direct edges between any pair of variables X_i, X_j with $X_i \succ X_j$ by comparing the total correlation between the two variables in the experiment in which X_i is subject to a structural intervention with the correlation due to the paths between the variables that are already known. Since the structural intervention breaks any influence of latent variables on X_i, we are only comparing correlations due to paths among the observable variables. If we start with "closer" connections in the graph, we can ensure that all other direct connections are known before we consider direct connections from vertices high in the graph to ones close to the sink. Concretely, in our example, there are two paths of length greater than one in the POG: $X \rightarrow W \rightarrow V$ and $X \rightarrow Y \rightarrow Z$. In this case, it does nor matter which path we consider first. So assume we start with the XWV-path. We test whether the correlation ρ_{XV} between X and V in the experiment $\mathcal{E}_X$, in

[8] See definition in the Appendix.

[9] Details of this order are given in Chapter 6 and the Appendix.

which X was subject to a structural intervention, is equal to the correlation due to the known path, i.e. whether $\rho_{XV} = e_{XW}e_{WV}$. Since this is the case, no direct edge is added from X to V. However, in the case of the XYZ-path, $\rho_{XZ} \neq e_{XY}e_{YZ}$ and hence an edge $X \rightarrow Z$ is added and the residual correlation $e_{XZ} = \rho_{XZ} - e_{XY}e_{YZ}$ is associated with the direct edge. If there were other paths in the POG, they would be considered next, but for this example we are done, we have discovered the structure (and correlations due to that structure) over the observed variables.

We can now search for latent variables by comparing the correlation due to the structure over the observed variables with the passively observed correlation between variables. If they are not equal, there must be latent variables. In this case we start from the root of the graph. If there are several roots, we must check whether any pair of roots is confounded before we consider other pairs of variables. Here we first consider whether any direct edge from the root is confounded by a latent variable. Concretely, we test whether the active correlation $\rho_{XW} = e_{XW}$ is equal to the passively observed correlation τ_{XW}. If it is, there is no latent common cause between X and W. If it is not equal, then we have discovered a latent common cause and can associate the residual correlation with the latent common cause. Next we would consider any paths of length two from the root. By considering potential common causes in a particular top-down order (described in more detail in the chapter on algorithms), we can ensure that we discover all confounders among variables higher in the graph before we consider confounders of variables closer to the sink. By subtracting the correlations due to all known pathways from the passive observational correlations between two variables, we can identify all latent common causes between two variables.

As this example shows, the bounds of Theorems 3.3.14 and 3.3.15 depend on being able to establish a partial order among the variables in the least number of experiments. Using multiple simultaneous interventions, this can be done in even fewer experiments, improving the case for search in causally insufficient sets of variables substantially from the N-experiment bound in Theorem 3.3.6:

Theorem 3.3.17: (fixed strategy) Multiple Structural Interventions, Correlation-Test, Causally Insufficient

$2\lceil \log_2(N) \rceil$ experiments are sufficient to determine the causal graph among N causally insufficient variables when multiple simultaneous structural interventions can be performed in each experiment and the model is linear.

Again this result extends to the discovery of latent variables:

Theorem 3.3.18: Search for Latent Common Causes: Multiple Structural Interventions
Given a set of N causally insufficient variables and assuming the model is linear, $2\lceil\log_2(N)\rceil + 1$ experiments, with multiple simltaneous interventions per experiment, are sufficient to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

For these two theorems the bounds are not tight. Depending on N, the bounds can be made tighter. The one experiment difference is just to ensure that all pairs of variables are passively observed at some point in the sequence of experiments. The strategy specifies the conditions that the sequence of experiments needs to satisfy:

Strategy 3.3.19: (fixed) Multiple Structural Intervention, Correlation-Test, Causally Insufficient
Given N causally insufficient variables $X_1, \ldots, X_N$, let the sequence of experiments consists of experiments such that for each pair of variables X, Y there is an experiment $\mathcal{E}_i = (\mathbf{S_i}, \mathbf{U_i}, \mathbf{Pol}_i)$ with $X \in \mathbf{S_i}$ and $Y \in \mathbf{U}_i$ and there is an experiment $\mathcal{E}_j$ with $X \in \mathbf{U_j}$ and $Y \in \mathbf{S_j}$ and all interventions are structural. For discovery of latent variables there must also be an experiment $\mathcal{E}_k$ with $\{X, Y\} \subseteq \mathbf{U_k}$.

The conditions for the structure among the observed variables are easily satisfied by intervening on different sets of $N/2$ variables (like the Cantor set construction for Strategy 3.3.5) and their complements. However, one can do better by combining the interventions more optimally. For example, for $N = 6$ the requirements can be satisfied with four (instead of six) experiments by using the intervention sets:

$$\{X_1, X_2, X_3\}, \{X_3, X_4, X_5\}, \{X_5, X_6, X_1\}, \{X_2, X_4, X_6\}$$

It is an open question whether there is a tight bound that can be given independently as a simple function of N, and whether it is always necessary to extend the tight bound by one extra experiment to determine the passive observational correlations.

The discovery procedure in the multiple intervention case is, apart from how the partial order is determined, essentially analogous to the case of a single structural intervention per experiment. Care only needs to be taken that the correlation tests are appropriately adjusted to those paths that are active in the relevant experiment, since the multiple simultaneous interventions may break causal connections.

It might appear that the structural aspect of the structural interventions is doing the work for these results, i.e. that the results would not be possible if the incoming edges from the latent common causes were not destroyed by an intervention. This is not entirely true. The crucial aspect of these results is access to unconfounded variables. Structural interventions are one way of producing unconfounded variables, but for parametric interventions, the *intervention* variables are also unconfounded. Consequently, the negative result for parametric interventions in Theorem 3.3.11 can – for linear models – also be reversed when correlation tests are used:

Theorem 3.3.20: (fixed strategy) Single Parametric Interventions, Correlation-Test, Causally Insufficient
N experiments are sufficient and in the worst case necessary to determine the causal graph among N causally insufficient variables when only a single parametric intervention can be performed in each experiment and the model is linear.

With regard to discovery of latent variables, Theorem 3.3.15 applies here for parametric interventions as well and the strategy also does not differ from the structural intervention case (Strategy 3.3.16), with N experiments intervening on a different variable each time, just using parametric interventions. Since parametric interventions can be combined independently without interfering, the previous result extends straightforwardly to multiple simultaneous parametric interventions:

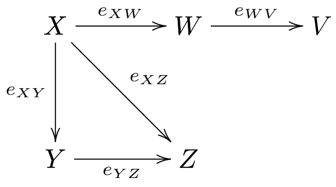
Theorem 3.3.21: (fixed strategy) Multiple Parametric Interventions, Correlation-Test, Causally Insufficient
One experiment is sufficient and in the worst case necessary to determine the causal graph among N causally insufficient variables when multiple simultaneous parametric interventions can be performed in each experiment and the model is linear.

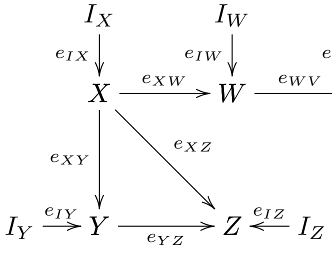
Theorem 3.3.22: Search for Latent Common Causes: Multiple Parametric Interventions
Given a set of N causally insufficient variables and assuming the model is linear,

one experiment, with multiple simultaneous interventions per experiment, is sufficient and in the worst case necessary to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

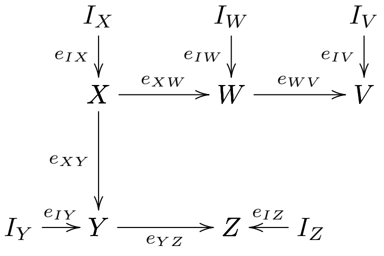
Strategy 3.3.23: (fixed) Multiple Parametric Interventions, Correlation-Test, Causally Insufficient
Given N causally insufficient variables $X_1, \ldots, X_N$, let the sequence of experiments consist of a single experiment $\mathcal{E}_1$ such that each X_i is subject to a parametric intervention.

The algorithm in the case of parametric interventions is very similar to the structural intervention case with one small tweak. We will illustrate the tweak with the same graph we used before, shown again on the left. Again we assume that there is a latent common cause for each pair of variables in the true graph.





The same graph on the right includes all the intervention variables. We can still construct a partial order and a POG, only this time $X \succ Y$ if and only if $I_X \not\perp\!\!\!\perp Y$ in the experiment in which X is subject to a parametric intervention. Hence, the POG is the same as before:



Since one can determine the correlations between an intervention variable and its intervened variable directly from the data (since the interventions are assumed to be unconfounded), we can compute the correlations of the direct edges in the POG. For example, $e_{XW} = \rho_{I_X,W}/e_{IX}$, where $\rho_{I_X,W}$ is the correlation between

I_X and W when X is subject to a parametric intervention. The procedure is otherwise exactly the same as before. Instead of determining the active correlation ρ in terms of the correlation between the intervened variable and some other variable, it is determined between the *intervention variable* and some other variable, and then appropriately reduced by the effect of the intervention variable on the intervened variable (i.e. by division by $e_{I\bullet}$) to yield the correlations for individual edges.

We can thus now summarize the results for search procedures using tests for **differences in correlation**:

Type of Experiment	Causal Sufficiency	Structural	Parametric
Single	Yes	$N-1$	$N-1$
Multiple	Yes	$\log_2(N)+1$	1
Single	No	N	N
Multiple	No	$2\log_2(N)+1$	1

For each of these strategies, the maximum number of interventions that occurs for some experiment in a sequence that satisfies the corresponding bound above is given in the following table:

Type of Experiment	Causal Sufficiency	Structural	Parametric
Single	Yes	1	1
Multiple	Yes	$N/2$	$N-1$
Single	No	1	1
Multiple	No	$N/2$	N

The results cannot be extended easily to the discrete case. As our example showed, the results rely on the ability to compute the effect of causal pathways individually, to compute the correlation due to particular subsets of pathways and to be able to compare correlations according to different pathways. This is not always possible for discrete models. Discrete models may contain interactive effects, which prevent an account of a causal effect that can be associated with individual pathways. Presumably, however, these results would hold for discrete models that do not allow interactive causes, such as, for example, noisy-or models.[10] We therefore conjecture that linearity, or at least some form of additivity in the functional form of the model, is a necessary assumption.

[10] See [30] for details on noisy-or models.

3.3.4 Search given Structural Knowledge

We have so far considered fixed strategies with single and multiple interventions per experiment, with parametric and structural interventions, on causally sufficient and insufficient sets, using independence tests and tests of differences in correlations. This is how far we are going to go in analyzing the impact of changes in the background assumptions until we relax the assumption of an oracle in the simulations. We now return to the full set of initial assumptions: causally sufficient sets of variables, independence tests only and we will restrict ourselves to structural interventions. In addition to these assumptions we will now *add* further assumptions and investigate their impact. The results described so far make no assumptions about any prior knowledge regarding restrictions on the possible causal strutures (other than acyclicity). There are many different ways in which some knowledge about the causal structure may already be available. To represent compactly all the knowledge that is already available or has been gathered during a sequence of experiments about the causal structure underlying a set of variables, we define a knowledge graph:

Definition 3.3.24: Knowledge Graph
A knowledge graph is a mixed graph over a set of variables $\mathbf{V}$ such that any two variables are connected by at most one of the following edge-types:

direct cause: A directed edge represents the knowledge that one variable, the start, is a direct cause of the other, the end (relative to $\mathbf{V}$): $X \longrightarrow Y$

non-adjacency: The absence of an edge represents the knowledge that neither variable is a direct cause of the other: $X \qquad Y$

adjacency: An undirected edge represents a direct causal connection between the two variables, whose orientation is not known: $X \;\text{——}\; Y$

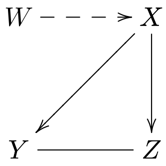
semi-directed: A semi-directed edge from variable X to Y represents the knowledge that either neither variable is a direct cause of the other or that X is a direct cause of Y: $X \dashrightarrow Y$

no knowledge: A no-knowledge edge represents that nothing is known about the direct connection between the two variables: $X \text{ —?— } Y$

An edge in a knowledge graph is considered *known* if it is of one of the first two edge-types, otherwise it is unknown. A knowledge graph is said to represent a *causal structure* uniquely when each of its edges is *known* and the structure is acyclic.

A knowledge graph, like a pattern for an observational Markov equivalence class (OME), can be used to represent an equivalence class of graphs that imply the same independence constraints. The main difference to a pattern is that a knowledge graph can represent information about independence relations resulting from structural interventions. Clearly, an OME of a causally sufficient set of variables can be represented as a knowledge graph, since an OME only requires the first three edge-types. The fourth edge-type is needed when only one variable of a pair is subject to a structural intervention and the pair appears non-adjacent in the post-manipulation graph. The fifth edge-type is needed when two variables are subject to structural interventions simultaneously, since no information about the connection between them is gained. Not all knowledge graphs represent equivalence classes of graphs that can be obtained from sequences of experiments (see below).

A knowledge graph can be used to represent prior knowledge about the causal structure or to summarize knowledge that has been gained from a sequence of experiments so far. Given a non-trivial knowledge graph, the choice of variables to subject to an intervention in an experiment should be sensitive to the information already present in the knowlegde graph. For example, suppose that for a set of four variables W, X, Y, Z, the following knowledge graph is known:



Clearly, the next intervention should be an intervention on Y or Z (and possibly, but not necessarily W). The OPTINTER algorithm computes intervention sets for knowledge graphs.

Algorithm 3.3.25: OPTINTER: Intervention Set Selection
Given a knowledge graph over a set of vertices $\mathbf{V}$, each vertex in $\mathbf{V}$ can be determined to be *admissible* or *inadmissible* and each vertex has a counter (of clique memberships). Let *maxInter* be the maximum size of the intervention set $\mathbf{S}$ for the next experiment.

1. Mark all vertices as *admissible* and set the counters for each vertex to 0.

2. Initialize the intervention set $\mathbf{S}$ to be the empty set.

3. Find all maximal cliques of vertices connected by *unknown* edges and order them $C_{\max}$ to $C_{\min}$ by the number of vertices they contain.[11] (No need to resolve ties.)

4. Each clique can be either *resolved* or *unresolved.* Mark all maximal cliques as *unresolved.*

5. Compute $h = 2^{\lceil \log_2(|C_{\max}|) \rceil - 1}$ (the closest power of 2 with $2^n < |C_{\max}|$).

6. Let the *relevant* cliques $C_1, ..., C_k$ be the cliques with $|C_i| > h$.

7. Sort all *relevant* cliques in order of size, place among equal sized cliques the ones with the most *inadmissible* nodes first.

8. Let C_{curr} be the first (largest) *unresolved* clique in the list of *relevant* cliques.

9. For each vertex $U \in C_{curr}$, set its counter to the number of *unresolved relevant* cliques C_i it is part of.

10. While $(|\mathbf{S}| < \text{maxInter})\&\&(|C_{curr} \cap \mathbf{S}| < |C_{curr}| - h)$, select vertex $V \in C_{curr}$ such that V is *admissible* and has the highest count; select randomly among ties.[12] Place it in $\mathbf{S}$.

    (a) [13] For any *relevant* clique C_i, if $|C_i \cap \mathbf{S}| = |C_i| - h$, then mark C_i as *resolved.*

    (b) For any *relevant* clique C_i, if $|C_i \cap \mathbf{S}| = h$, mark its vertices as *inadmissible.*

11. Return to 7 and start over until all *relevant* cliques are *resolved* or when no further *relevant* cliques can be resolved.

[11] Note, that if the knowledge graph results from a sequence of experiments, then there cannot be cliques of semi-drected edges. All cliques of *unknown* edges are necessarily cliques of no-knowledge edges or cliques of undirected edges.

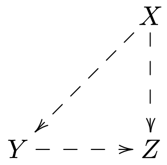
[12] In a fixed strategy some deterministic rule can be used to select among tied variables.

[13] The following two constraints on *admissible* nodes and *resolved* cliques are appropriate for a fixed strategy. In a mixed search strategy different lower limits apply for the optimal number of variables that should be included in an intervention set.

12. (Post Process: While possible with regard to the constraints (a) and (b) of step 10, add vertices to the intervention set to resolve additional maximal cliques.)[14]

13. Return the intervention set.

OPTINTER attempts to find intervention sets in knowledge graphs that are close to optimal: if OPTINTER is called before each new experiment, the shortest sequence of experiments necessary for the worst case graph consistent with the initial knowledge graph is determined. OPTINTER is not exactly optimal for arbitrary knowledge graphs, since they might represent knowledge about the causal structure that cannot be obtained from sequences of experiments. For example, for a knowledge graph over three variables where each edge is a semi-directed edge, a passive observation would be one optimal solution, resulting in a single experiment. Such a knowledge graph cannot be obtained by any sequence of experiments.



OPTINTER returns an intervention set that contains one variable, and – depending on which variable it is, e.g. Y or Z – it is not optimal, since two experiments, instead of one are required. However, we conjecture that for certain types of knowledge graphs, OPTINTER is optimal, when multiple simultaneous interventions are permitted in each experiment.

One particular type of knowledge graphs are patterns representing passive observational Markov equivalence classes (OMEs), which can be determined from samples where no intervention was performed. Assuming causal sufficiency, if the OME of the true graph is given, we conjecture the following bound on the number of experiments required to uniquely identify the true causal structure:

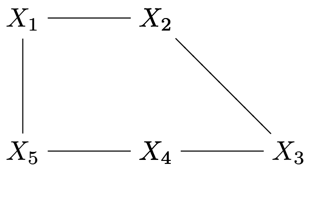
Conjecture 3.3.26: Structural Interventions on OME
Given an OME of the true graph, if multiple simultaneous and independent structural interventions can be performed in each experiment, then $\lceil \log_2(|C_{\max}|) \rceil$ experiments are sufficient and in the worst case necessary to recover the true causal graph, where $C_{\max}$ is the largest clique of undirected edges in the true OME.

[14] This step is not invoked for fixed strategies.

Here the worst case consists of the most unfortunate choice of variables for intervention relative to the graphs in the OME that one cannot guarantee to avoid.

The conjecture is based on the realization that cliques are the main obstacle to discovery using structural interventions. Cliques provide shields for colliders, which increase the combinations of how edges could be oriented. Hence, to discover the causal graph in the least number of experiments the main aim is to determine the orientation of edges in cliques (the adjacencies are already given by the OME). For a clique of size $|C|$, an experiment that intervenes on k variables in the clique determines the orientation of $k(|C| - k)$ of the clique's edges. This value is maximized for $k = 1/2|C|$. An algorithm that reduces by half the size of all cliques of undirected edges for which no edge-orientation is known clearly satisfies the conjectured bound. In general, the requirement is a little weaker: If all undirected cliques $C_{>h}$, that are larger than h variables, where $h = 2^{\lceil \log_2(|C_{\max}|)\rceil - 1}$ (the closest power of 2 below $|C_{\max}|$), are reduced to cliques of size h in each experiment, then the conjectured bound is satisfied. So for any such clique C, $|C| - h$ variables have to be subject to an intervention simultaneously.

The conjecture currently remains without proof because it is not entirely clear whether it is possible to find intervention sets for any OME that break down appropriately all such cliques $C_{>h}$ in the graph. Since cliques may overlap, a proof of the conjecture must guarantee that intervention sets can always be found that resolve all of the overlapping cliques at once. For example, if the equivalence class is given by X_1 —— X_2 —— X_3, then the conjecture claims that the causal structure can be resolved in a single experiment. There are two maximal cliques, $\{X_1, X_2\}$ and $\{X_2, X_3\}$, that must be resolved in this single experiment. Clearly we cannot select for each clique independently, which variables to intervene on, since that may result in a choice of intervening on X_1 and X_2 simultaneously, which obviously would not resolve the graph. In this case the optimal choice of intervention set – $\{X_2\}$ – is obvious, but in general this is not as simple. In particular, consider an undirected five-cycle:



Here we have four overlapping maximal cliques of size 2. Again, one experiment should be sufficient to recover the causal structure, but in this case it is impossible to find an intervention set that resolves all four cliques simultaneously. Any selection would result in one clique for which both or no variable is contained in the intervention set. However, the five-cycle is not a counterexample to the conjecture since an undirected five-cycle (or any other cordless cycle greater than three) cannot occur in any OME – there would always be an unshielded collider.[15] Hence, the conjecture certainly does not hold for arbitrary knowledge graphs, but is dependent on the assumption that the graph is an OME or a knowledge graph that is consistent with some sequence of experiments.

For arbitrary knowledge graphs there is a general negative graph theoretic result due to Folkman [14] that prevents a general version of the conjecture even for knowledge graphs of one edge-type. It relates the problem of intervention set selection to coloring theorems:

Theorem 3.3.27: Folkman Clique Theorem – paraphrased
For any clique-size $c \geq 3$, there is a graph G, whose largest clique has size c and for which every edge two-coloring has a clique of size c in one color.

Considering only the undirected edges of an OME, let an edge be colored red in experiment $\mathcal{E}$ if it connects an intervened and a non-intervened variable, and blue if it connects variables that are both subject to an intervention or both passively observed. Folkman's theorem implies that for any integer $c \geq 3$, there is a graph G whose largest clique has c members. Any coloring, so in particular colorings of the type we defined, would result in a clique of size c of blue or red edges only. Due to the way our coloring is defined, it is impossible for the clique to be among red edges, since three variables cannot be fully connected by red edges. Consequently, the clique is among blue edges, and since the intervened variables are separated from the non-intervened ones (by red edges), the clique must be either among the intervened variables only or among the unintervened variables only. That is, after the experiment, we are left with a clique of the same size as we started off with. Since there is no way to reduce cliques of *unknown* edges by more than half, the conjectured bound does not hold for general knowledge graphs, not even for general adjacency graphs. However, there is hope – as with the five-cycle – that the graphs that satisfy Folkman's theorem are not OMEs and not derivable from OMEs by

[15] It turns out that for the undirected 5-cycle knowledge graph, OPTINTER recovers the structure in the minimum number of experiments *for that knowledge graph*, namely two.

sequences of experiments. If that fails, an argument is needed that such graphs are sufficiently rare so as not to be of practical worry.[16]

Computing the appropriate intervention set given a knowledge graph is closely related to the MAX-CUT problem of the subgraph of the knowledge graph containing *unknown* edges only, which is in general NP-complete. There are approximation algorithms, with the best offering a 0.878-approximation [15]. The approximation MAX-CUT algorithm is, of course, not designed, with the specific aim to orient edges in cliques. Hence, an approximate MAX-CUT might not be sufficient to guarantee the conjectured bound (even if true). The OPTINTER algorithm is a greedy algorithm that selects the intervention set specifically in light of the conjectured bound. In simulations it always successfully found an appropriate intervention set resulting in a sequence of experiments satisfying the bound, but that is no proof of correctness, nor really much of a plausibility argument, since the space of graphs is large and our simulations can only cover a small area. OPTINTER is computationally expensive, so for large graphs a MAX-CUT approximation algorithm is probably a better choice, even if it results in additional experiments.

If we assume that Conjecture 3.3.26 is true, we can specify a fixed strategy for sequences of experiments on OMEs:

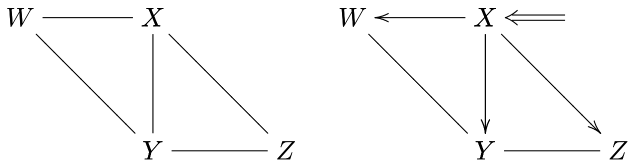
Strategy 3.3.28: (fixed) Multiple Structural Interventions, Causally Sufficient - OME
Given an OME over N causally sufficient variables $X_1, \ldots, X_N$, let the sequence of experiments consist of $k = \lceil \log_2(|C_{\max}|) \rceil$ experiments, such that the intervention set $\mathbf{S}$ of each experiment is determined by the OPTINTER algorithm.

In this fixed strategy, OPTINTER determines the intervention set of the first experiment on the basis of the OME. For any later experiment $\mathcal{E}_j$, OPTINTER computes the intervention set based on the original OME, where any edge between variables X and Y such that $X \in \mathbf{S_i}$ and $Y \in \mathbf{U}_i$ for some $i < j$ is considered *known*. That is, edges that are guaranteed to be determined by earlier experiments (sample issues aside) are considered known for the determination of the intervention set of subsequent experiments. These edges can be determined before any experiments are performed.

For example, if the initial OME is the graph on the left, then OPTINTER

[16] I am very grateful to Oleg Pikhurko for pointing me to Folkman's Theorem. If my rendition of Folkman's Theorem is incorrect, then that is entirely my fault.

will compute a priori that the intervention set in the first experiment is $\{X\}$ (assuming some appropriate rule that resolves the tie between X and Y, e.g. in this case just lexicographic order). Given that the first experiment will consist of an intervention on X, a resolution of the orientation of at least the XW-, XY- and XZ-edges is guaranteed. So for the computation of the intervention set for the second experiment, these edges can be considered known, therefore leaving from the original OME only W —— Y —— Z, which obviously implies the intervention set $\{Y\}$ for the second experiment. One possible outcome of the first experiment that shows that two experiments are in the worst case necessary, is shown on the right.



The second experiment of the pre-determined fixed strategy intervening on Y resolves the remaining unknown orientations.

So far we only considered experiments involving *multiple simultaneous* interventions on an OME. The difficulty in specifying a bound on the number of experiments given an OME, when only single interventions are permitted, is due to the interdependence of orientation of two adjacent edges. For example, if we know from passive observation that the OME of the true causal graph is a chain of undirected edges,

$$X_1 \text{ —— } X_2 \text{ ——}\cdots\text{—— } X_N$$

then we know that this chain cannot contain any unshielded colliders, since they would have been discovered in the passive observation. But it could contain a common cause at any vertex (except the ends). Consequently, if only a single structural intervention can be performed per experiment, the most efficient strategy would be to intervene on the middle vertex, resulting in a sequence of $\log_2(N)$ experiments in the worst case to recover the causal structure. With multiple simultaneous interventions a single experiment with a structural intervention on every other variable would solve the problem (as implied by the conjecture). Similar arguments apply to tree structures and particular planar networks made up of triangles or diamond shapes. Ultimately, if there are many dependencies between the directions of edges, then these can be exploited to improve the efficiency of discovery. In general the number of experiments necessary and sufficient to discover the causal graph given an OME when only single interventions can be performed per experiment, is bounded by the $\sum_i(|C_i| - 1)$, where the C_i are non-overlapping maximal cliques. But we have no results on how tight this bound is, nor do we have a method, other than brute force, to compute the appropriate intervention sets. For similar reasons, the intervention sets computed by OPTINTER are not minimal with regard to the number of variables subject to an intervention in one experiment or over the entire sequence of experiments.

3.3.5 Restrictions

Lastly, in this section on fixed strategies, we will consider some restrictions that might limit the discovery procedure. Actual experimental conditions might place a whole variety of additional restrictions on the experimental procedure that would undercut some of the assumptions made in the theorems of this section. Here we just list a few results that are adjusted for some interesting restrictions on the set of assumptions.

Limited Intervention Sets

Several of the search strategies that satisfy the bounds for multiple simultaneous interventions require large intervention sets: up to $N/2$ for the $\log(N) + 1$ bound of Theorem 3.3.4 for structural interventions, and up to $N - 1$ for the one experiment bound of Theorem 3.3.9 for parametric interventions, where N is the number of variables in $\mathbf{V}$. If we are not able to handle such a large intervention set, but can only intervene on $k < N/2$ (or $k < N-1$) variables at once, then – under assumptions 1.2.1-3.2.6 – the bounds increase as follows:[17]

Theorem 3.3.29: Limited Structural Intervention Set, Causally Sufficient

Given N causally sufficient variables, if the number of simultaneous structural interventions is limited by $k_{max} < \frac{N}{2}$ in any one experiment, then

$$(\frac{N}{k_{max}} - 1) + \frac{N}{2k_{max}} \log_2(k_{max})$$

[17] This result for structural interventions was first presented in [10], but there we incorrectly claimed that the bound was in the worst case necessary. Unfortunately, it is just sufficient.

experiments are sufficient to discover the causal graph.

The result is essentially a mixture of the bounds for single and multiple structural interventions. In the case of parametric interventions we only need to ensure to subject each variable to a parametric intervention. Since they can be combined independently, the result is trivial:

Theorem 3.3.30: Limited Parametric Intervention Set, Causally Sufficient
Given N causally sufficient variables, if the number of simultaneous parametric interventions is limited by $k_{\max} < N - 1$ in any one experiment, $\lceil \frac{N-1}{k_{\max}} \rceil$ experiments are sufficient and in the worst case necessary to discover the causal graph.

Limited Conditioning Sets

The bounds on the numbers of experiments in the case of a causally sufficient set of variables depend on the ability to – in the worst case – consider conditional independence tests with conditioning sets of size $N - 2$, where N is the number of variables in $\mathbf{V}$. If it is not possible or ill-advised (e.g. due to lack of data) to consider such large conditioning sets and we are limited to conditioning sets of size $k < N - 2$, then we are unable to distinguish a direct causal link between two variables from a causal connection that – in the worst case – consists of up to $N - 2 - k$ variables (all of which must be non-colliders). One can compensate for this problem by intervening on more variables simultaneously, breaking all possible paths, other than the direct one, by additional interventions. However, in general we face for the worst case analysis a similar problem to the causally insufficienct case (Theorem 3.3.6). That is, one either has to perform a large number of simutaneous interventions or one is left only with some partial order information on the variables, which may be sufficient if the graph is sparse, but of no use to resolve the worst case, if only independence tests are considered.

Other Knowledge

Conjecture 3.3.26, if true, implies a whole set of corollaries: If it is known that the true graph has at most k edges, then the worst case graph for discovery is the graph where those edges are arranged into the largest possible clique. The number of experiments necessary in this case and sufficient for all other graphs with k edges is given by the log of the resulting clique-size. In the case of k

edges, the bound on the number of experiments is approximately $\sqrt{2k}$. If the maximum number of parents of any node is limited by k, then the largest clique one can construct has size $k+1$, so the bound on the number of experiments is approximately $\log_2(k+1)$. The strategy is always the same: Determine the largest clique consistent with the constraints. The worst case number of experiments is then given by the log of the clique-size.

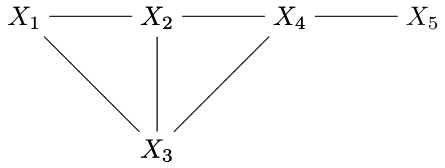
Nyberg and Korb [31] consider a case of causal structure search with interventions when faithfulness is not assumed. They prove that even without assuming faithfulness, but assuming causal sufficiency, the causal structure among a set of variables can be recovered, both with structural and parametric interventions. Like us, they do not give an account of how parametric interventions – especially on an unfaithful structure – should be implemented. Their work suggests a more general analysis of the impact of the faithfulness assumption on search strategies using interventions.

Eaton and Murphy [7], inspired by biological cases, consider interventions in which the target of the intervention is not known. In their case, no edge from an intervention variable to some intervened variable, if it exists, is known. All that is known is that the intervention variable cannot be a descendent of any vertex in the system under consideration. This knowledge that the intervention variable is exogenous, even though it is not known how the variable is connected to other variables in the system, helps search for causal structure.

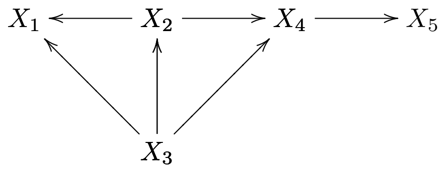
We would model their set-up differently. In their description the variable that is manipulated – a chemical – is known, as is its distribution, but it is not known which variables, if any, it affects, and how it does so. It is known that the chemical variable is not a descendent of any other variable. Instead of considering the chemical variable as an intervention variable, for which the intervened variables are unknown, we would model the chemical as an explicit causal variable that is subject to an intervention. Background knowledge can be added to the search algorithm that ensures that no other variable in the system affects this variable. Any dependence would be recognized as an outgoing causal connection. The problem would not be framed as a "blind" intervention, but as a structure search with background knowledge.

3.4 Adaptive Strategies

Unlike fixed stratgies that announce one particular sequence of experiments prior to the first experiment, adaptive strategies can adjust in light of the information gained as each experiment is performed. That is, adaptive strategies specify for each stage in a sequence of experiments and for each history of information gained so far, a specific choice for the next experiment. However, the choice at any one stage may – unlike the case for fixed strategies – depend on the information that has been gained in previous experiments. The adaptation becomes particularly relevant if one discovers that one is not up against the worst case. If one finds that there are independencies (implying non-adjacencies), that in turn imply dependencies in the orientation of particular edges, then one can in many cases discover the true causal graph in fewer experiments than specified by the worst case bound. For example, consider the following observational Markov equivalence class:



One can use several different fixed strategies, but the most appropriate among the ones seen so far clearly seems to be the one specified by Strategy 3.3.28. The corresponding bound (Conjecture 3.3.26) implies that two experiments should be sufficient to recover the causal structure (and necessary in the worst case). Consequently, the fixed strategy might suggest two experiments with intervention sets $\mathbf{S}_1 = \{X_2\}$ and $\mathbf{S}_2 = \{X_1, X_4\}$. These two experiments would, in fact, recover the causal structure no matter which of the graphs in the OME is true. However, if the true underlying graph is

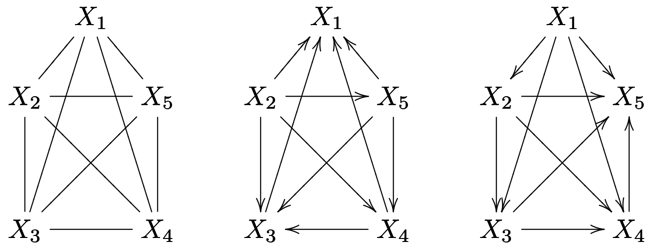


then the first experiment already supplies sufficient constraints (non-existence of collider and acyclicity) to determine the graph uniquely. The second experiment is superfluous. An adaptive strategy would adjust accordingly and stop the search. Similar adaptations occur when there are dependencies in how edges are oriented. Consider an OME that is a chain, for example:

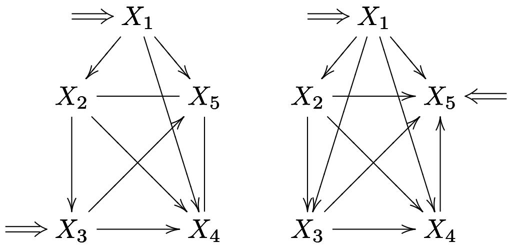
$$X_1 \text{ —— } X_2 \text{ ——}\cdots\text{—— } X_n$$

If the first experiment happens to be an intervention on X_1 (for whatever reason) and it turns out that there is an edge $X_1 \rightarrow X_2$, then that automatically resolves the orientations of the entire chain (namely from X_1 to X_n) and no further experiment is needed.

A different form of adaptation may occur, when information becomes available to select the intervention set for the next experiment more optimally, and thereby shorten the overall sequence of experiments. For example, suppose it is known that the OME of the true graph is a complete graph over 5 variables, shown on the left below. A fixed strategy might consist of the following three intervention sets for three experiments that guarantee to discover the graph within the bound: $\{X_1, X_3\}, \{X_1, X_4\}$ and $\{X_5\}$. All three experiments would be necessary if the true graph were the graph shown in the middle below. But suppose, in fact, the true graph is the graph on the right.



After the first experiment intervening on X_1 and X_3 we get the knowledge graph shown on the left below (semi-directed edges are resolved since adjacencies are known from the OME, the no-knowledge edge between X_1 and X_3 is omitted).



The X_3X_1-connection and the $X_2X_5X_4$-connection remain unresolved. An adaptive strategy can under these circumstances adapt and select $\{X_1, X_5\}$ as the intervention set for the second experiment, which resolves the true causal graph in two, instead of in three experiments, shown on the right above.

There are cases, where there are several choices of intervention sets, and there may be no reason to prefer one over the other. For example, if there are two variables X and Y, then – assuming causal sufficiency – there are three possible causal structures. Suppose the true causal structure is $X \rightarrow Y$. If one were to intervene on X, the causal structure would be discovered in one experiment. However, in the case of a passive observation or an intervention on Y, two experiments would be required. Since any of the three structures might occur, and the situation is symmetric with regard to the three possible experiments, one cannot ensure a priori against a sequence of two experiments. The worst case analyses in the previous section considered the longest of these possible sequences of experiments, i.e. the least fortunate but not avoidable selection of variables to subject to an intervention in each experiment in the sequence. In the case of a complete graph such a worst case sequence of experiments intervenes on each variable in the hierarchical order of the graph starting from the graph's root or sink.

It turns out, that against a worst case graph, there is no adaptive strategy that can guarantee to do better than the fixed strategy bounds in the previous section. That is, if the true graph is the worst case graph, then it is impossible to ensure that an adaptive strategy does better than the number of experiments specified by the fixed strategy bound. Or, in other words, there is no way to guarantee that one performs experiments that provide information about the true underlying graphical structure that could be used to adapt the search. Of course, an adaptive strategy may be *lucky* to select intervention sets that improve the bound, but there is no guarantee.

Theorem 3.4.1: Adaptive vs. Fixed Strategies
Under the same assumptions, no adaptive strategy can improve on the worst case bounds of theorems 3.3.1, 3.3.4, 3.3.6, 3.3.8 and 3.3.9 if the true graph is a worst case graph.

The implication of this theorem is not (!) that adaptive strategies are useless, but rather that they do not change the worst case analysis. In any particular search procedure, one does, of course, want to adapt in light of the findings of previous experiments, since one may not be facing the worst case structure or one may be lucky in how intervention sets turned out. In fact, the worst case might be quite rare.

If adjacency information is known, as in the case of an OME, then the orientation dependencies can (in principle) be computed and the intervention set can be selected accordingly. But if no adjacency information is available, then it is not so clear how interventions should be placed in an adaptive strategy. One does not want to lose adjacency information unnecessarily between variables by subjecting too many variables to simultaneous structural interventions. But one also does not want to intervene on too few and then discover a dense adjacency graph. This raises the question of what the first experiment in an adaptive strategy should be, when no adjacency information is available. Since the fixed strategies presented in the previous section are designed explicitly with the worst case in mind, they do not always provide an optimal starting point if one is not up against the worst case. For example, in the comments on Theorem 3.3.2 we indicated that passive observation was of no value against the worst case graph. However, even if the graph is slightly less than complete, information gained from passive observation (the OME) might imply a significant reduction in the number of experiments necessary to discover the causal structure (if Conjecture 3.3.26 is true, the reduction would be from a function of the total number of variables to a function of the size of the largest clique). Similarly, if the complete graph is an extremely unlikely structure, then it may not be a good idea to start with an intervention on $N/2$ variables (as suggested in some instantiations of Strategy 3.3.5), nor is it necessary to intervene on $N-1$ variables simultaneously (as suggested in Strategy 3.3.7) if there is reason to believe that not every pair of variables is subject to confounding. In general little can be said about an adaptive first experiment – all a first experiment can be adapted to are specific assumptions if available (e.g. distributions over possible graphs).

While an adaptive strategy is no better than the fixed strategy in the worst case, it can be designed to weakly dominate the fixed strategy, i.e. to stay within the bounds described by the fixed strategies, never do worse and do better where possible. Under these considerations, the cases where there is the kind of flexbility of adjusting intervention sets (as in the examples of Strategy 3.3.5 with 8 variables), the first experiment should intervene on as few variables as possible, so that non-adjacencies are discovered early. However, as the case for the same strategy with 7 variables shows, this flexibility is not always available. Without that flexibility, one can only hope for a "lucky" intervention set or some non-adjacency to appear early. An adaptive strategy that weakly dominates the fixed strategy 3.3.5 is essentially a version of Strategy 3.3.28 with a slight modification of the OPTINTER algorithm, and where the OPTINTER algorithm is called after each experiment, when as much as possible about the causal structure has already been determined:

Strategy 3.4.2: Adaptive: Structural Interventions, Causally Sufficient
Given a causally sufficient set of N variables $X_1, \ldots, X_N$, let the first experiment intervene on as few variables as possible (determined by the flexibility described for Strategy 3.3.5) and perform each subsequent experiment $\mathcal{E}_i$ with $\mathbf{S_i}$ determined by OPTINTER including step 12. OPTINTER is called on the basis of the current total knowledge graph obtained after each experiment.

For causally insufficient sets of variables, the adaptive search remains an open problem and appropriate changes have yet to be made to the OPTINTER algorithm.

Chapter 4

Search with Interventions: Mixed Search Strategies

The bounds presented in the previous chapter for fixed and adaptive strategies are worst case results. In general the worst case might be very rare and one would like to have a sense of the expected case. The expected performance of an algorithm depends heavily on what constitutes the space of possibilities and the form of the distribution over these possibilities. One interesting case can be analyzed straightforwardly and it is supported by a game-theoretic interpretation of the discovery problem: the worst case expected performance, i.e. the upper bound on the expected length of sequences of experiments necessary and sufficient to discover the causal structure, no matter what the distribution over the set of directed acyclic graphs is. That is, for each distribution $P(\mathcal{G})$ over the set $\mathcal{G}$ of directed acyclic graphs over N variables, take the expectation $E_{P(.)}$ of the number of experiments $\#ex(.)$ necessary and sufficient to discover the true causal graph G uniquely; then take the upper bound – the supremum – of those expectations. Or formally:

$$\forall P(\mathcal{G}) \quad \sup_{E}[E_P(\#ex(G))]$$

The key to computing this quantity is the specification of $\#ex(G)$ for some causal structure G. To specify this quantity we need to specify how experiments are chosen. But how experiments are chosen affects which causal structures are difficult to learn, and thereby affects the supremum. For example: If the first experiment always consists of an intervention on variable X, then causal structures in which X is independent of, or an effect (but not a cause!) of the other variables, are more difficult to discover because any incoming causal influence on X is destroyed by the intervention, and so the structure cannot be distinguished from ones in which X is causally independent of the other variables. Consequently, a distribution that puts more weight on those graphs will be a candidate for the maximum expectation. But if the first variable subject to intervention is determined by a flip of a (N-sided) coin to determine the variable subject to the first intervention, then $1/N$ of the time – when the coin determines X – the Scientist will do poorly, but in some of the $(N-1)/N$ other times, she will benefit from intervening on the causes of X, thereby improving the expectation.

A restriction to fixed choices of experiments (given a particular set of evidence so far in the sequence of experiments) therefore appears artificial, and could even be detrimental. Consequently, $\#ex(G)$ is computed as the number of experiments necessary and sufficient to discover the causal graph given a strategy S, where S specifies for each possible choice of experiments (and history of evidence) a probability distribution over experiments such that for every alternative strategy, S', the supremum is higher (or equal) to the supremum for S. Formally, $\#ex(G)$ is computed given a strategy S such that

$$\forall S' \neq S \quad \sup_E E_P(\#ex_{S'}(G)) \geq \sup_E E_P(\#ex_S(G))$$

Given any possible history of causal relations discovered in the data generated in the sequence of experiments so far, S specifies a distribution over the choices of the next experiment. As the formal definitions indicate, there is an interdependence between the appropriate choice (or distribution over choices) of the next experiment and the underlying distribution over causal structures. But this is a problem that we can get control over in a game theoretic framework.

In considering distributions over graphs *and* distributions over possible experiments we broaden the space of search strategies from fixed and adaptive strategies to include mixed strategies. The next experiment in a sequence is determined on the basis of a random sample from an appropriately weighted distribution over the options. The choice for the experimenter is to pick the appropriate distributions over experiments. Fixed and adaptive strategies, which correspond to pure strategies in game-theoretic terms, must commit to a particular choice of experiment for any circumstance. Mixed strategies consider a (weighted) random selection between the options.

In principle, the success of a mixed strategy can be measured in several ways, but given our interest in a more representative measure of the quality of the search strategy (in constrast to the worst case results) we will focus on the *expected* number of experiments. Such an analysis is also supported by more theoretical reasons: The focus on worst case expectation allows us to build on Nash theory, which generalizes beyond two-person zero sum games. An analysis of the strategy in terms of a minimax considerations would not lend itself to such generalization. Since there are many possible distributions over graphs and many possible distributions over experiments, we will here only consider the upper bound on the expectation, i.e. the distribution over graphs that results in the highest expected number of experiments, against a mixed search strategy that aims to minimize the number of experiments.

We start by recasting the causal discovery problem using interventions as a sequential game that can be analyzed using game-theoretic techniques. In a second section we then give results on the worst case bounds for the number of experiments, when mixed search strategies are used. As in the previous chapter, we describe strategies that instantiate sequences of experiments that respect the given bounds, and again we consider both single and multiple intervention strategies. In a last section we tie together this chapter with the previous one by comparing the results within a general game-theoretic framework.

4.1 Discovery as a Game

Our approach to causal discovery can be viewed as a game between Nature and the Scientist, similar to Wald [48]. The Scientist attempts to discover the true causal structure and Nature tries to make discovery as difficult as possible – in our case, in terms of the number of experiments. Nature gets to decide what the truth is, but then has to stick with it, while the Scientist performs her experiments. The game can be seen as a zero-sum game, in the sense that Nature wins when there are more experiments, and the Scientist loses at the same rate. Nature gets to select the true graph but may not change or adapt the graph after its first move. Hence, Nature's pure strategies are all the directed acyclic graphs over N variables, but Nature may play mixed strategies by selecting the true graph on the basis of a random sample from some distribution over the pure strategies. The Scientist performs experiments to determine the true graph. After each experiment, the equivalence class of graphs that contains the true graph and is consistent with the sequence of experiments so far, is revealed. The pure strategies for the Scientist are all possible sequences of experiments. The Scientist may end the game after any sequence of experiments by declaring one of the graphs remaining in his information set (the equivalence class of graphs containing the true graph consistent with the sequence of experiments) as true. If the Scientist is correct, her pay-off is the negative number of experiments that were performed (the fewer experiments, the better). If the Scientist is incorrect, the pay-off is $-\infty$.[1] The Scientist may also play a mixed strategy over the possible experiments.

In this thesis we only consider the case for structural interventions, we assume that the set of variables is causally sufficient and that we have access to an independence oracle.

Since, by Theorem 3.3.1 and Theorem 3.3.4, the worst case bound on the number of experiments is $N-1$ for single interventions and $\log_2(N)+1$ for multiple simultaneous interventions, we do not need to consider pure strategies for the Scientist that are longer than these bounds, i.e. the theorems give upper bounds on the worst case loss for the Scientist (assuming she is paying attention to the game). To illustrate how the discovery problem is mapped onto a game, we consider one simple example, and then a more elaborate one.

Example: Two and Three Variables

We consider two examples, one with two and one with three variables. For two causally sufficient variables X and Y, there are three possible causal structures $\mathbf{Sa}, \mathbf{Sb}$ and $\mathbf{Sc}$:

$$\mathbf{Sa}: X \quad Y \qquad \mathbf{Sb}: X \rightarrow Y \qquad \mathbf{Sc}: X \leftarrow Y$$

Two experiments involving structural interventions are sufficient and in the worst case necessary to discover the causal structure uniquely. The full game of Nature against Scientist is given in Figure 4.1.

[1] Of course, one could integrate a notion into the pay-off structure, that accounts for how wrong the Scientist is, but we leave this for future consideration. A pay-off of $-\infty$ also ensures that if at all possible, the Scientist will not end the game by guessing, but will continue playing (searching) while it is possible to guarantee discovery of the true graph. We thereby force the Scientist to be able to justify her response in the sense that it is provably the unique correct response.

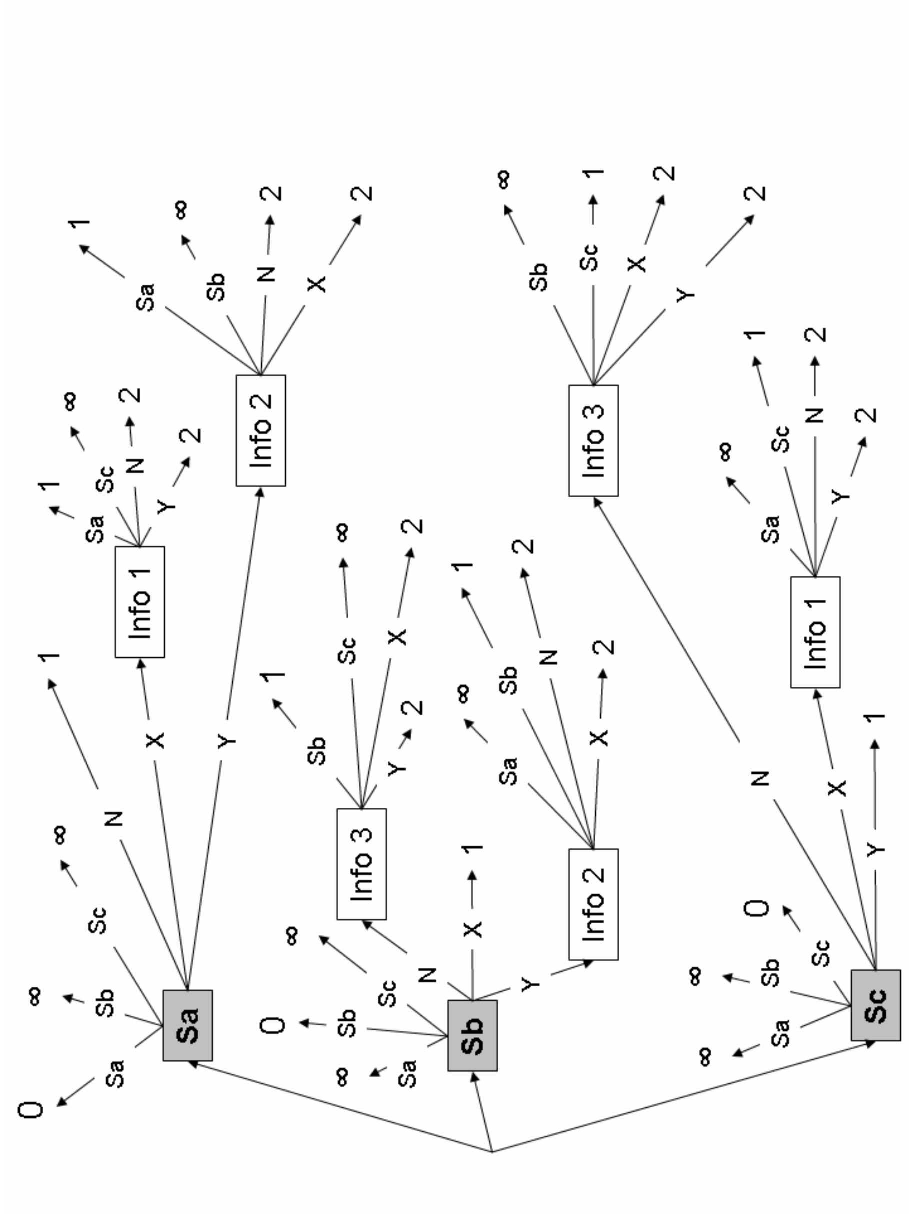


Figure 4.1: *Discovery of Causal Structure as a game of Nature against the Scientist, here for two causally sufficient variables. See description of this example in section 4.1.*

Nature can select among the three structures (grey boxes) $\mathbf{Sa}, \mathbf{Sb}$ and **Sc**. The Scientist does not know which structure is selected, so $\mathbf{Sa}, \mathbf{Sb}$ and **Sc** form an Information set, Info 0 (not shown in the figure). The Scientist makes the next move and can end the game by guessing one of the structures without collecting any data (represented by the three arrows leaving each grey box upwards with Sa, Sb or Sc and the respective pay-offs to Nature of 0 when the choice was correct and ∞ when incorrect). Alternatively, the Scientist can perform a passive observation (N), an intervention on the first variable (X), or an intervention on the second variable (Y). Depending on the choice and the true underlying graph, the game is either resolved because the graph can be uniquely identified (pay-offs are indicated), or a new information set is returned, corresponding to one of three knowledge graphs:

Info 1: X was subject to intervention and Y did not covary, so the edge is either into X or there is no edge between X and Y: $X \dashleftarrow Y$

Info 2: Y was subject to intervention and X did not covary, so the edge is either into Y or there is no edge between X and Y: $X \dashrightarrow Y$

Info 3: X and Y were passively observed and covaried, so there is an edge between X and Y, but the direction is unknown: X —— Y

Again, the Scientist can end the game at this point with a guess, or can continue with a further experiment. There is no need to consider strategies of more than two experiments, since they uniquely resolve the true causal graph.

In tabular form, the game is represented as follows. We split the table into three tables corresponding to sequences with 0,1 and 2 experiments. $A(\mathbf{Sb})$ means that the Scientist guesses structure **Sb** and thereby ends the game early. "*" indicates that this state does not occur:

Zero experiments, i.e. the game is ended immediately with a guess by the Scientist (columns). Depending on the true graph (rows), the pay-offs to Nature are indicated. If the guess was correct, the pay-off is 0, if it is incorrect, it is ∞.

	$A(\mathbf{Sa})$	$A(\mathbf{Sb})$	$A(\mathbf{Sc})$
Sa	0	∞	∞
Sb	∞	0	∞
Sc	∞	∞	0

One experiment, i.e. an experiment is performed (the first three columns show the passive observation N; columns 4-6 show the intervention on X; and the last three columns show the pay-off for an intervention on Y). The experiment either resolves the graph immediately (columns 1,6 and 9) or it is followed by a guess indicated after the comma in the column header (only guesses of graphs that have not already been excluded by the experiment are shown). Again only pay-offs to Nature are shown. Combinations with a "*" do not occur. For example, for the first row, second column we have the true graph as empty and the strategy is to perform a passive observation (N) followed by a guess of structure **Sb**. But if the true structure is **Sa**, the empty graph, then the structure would be uniquely determined by the passive observation already (as column 1 shows) and there would be no need for a guess (in this case the guess is for **Sb**). Hence, that combination does not occur and no pay-off is given.

	N	N, A(**Sb**)	N,A(**Sc**)	X, A(**Sa**)	X,A(**Sc**)	X	Y, A(**Sa**)	Y, A(**Sb**)	Y
Sa	1	*	*	1	∞	*	1	∞	*
Sb	*	1	∞	*	*	1	∞	1	*
Sc	*	∞	1	∞	1	*	*	*	1

Two Experiments: The two experiments performed are shown in their order in the column headers (only relevant combinations of experiments are shown). Since two experiments always resolve the causal structure among two variables uniquely, no guesses are needed and the pay-off is always two. Again some combinations do not occur since the first experiment would already resolve the graph uniquely and then no further experiment is required, e.g. row 1, column 1 is already resolved by the passive observation (see row 1, column 1 of the previous table).

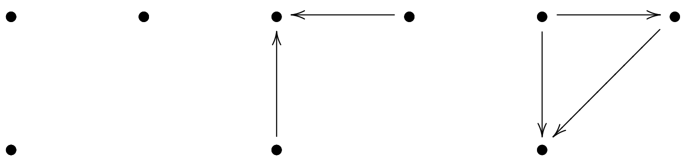
	N,X	N,Y	X,N	X,Y	Y,N	Y,X
Sa	*	*	2	2	2	2
Sb	2	2	*	*	2	2
Sc	2	2	2	2	*	*

An analysis of the game shows that the Nash equilibrium is given by a mixed strategy that is uniform over the three possible structures for Nature, and a mixed strategy for the Scientist that is uniform over passive observation, an intervention on X and an intervention on Y for the first experiment, and indifferent between possible (relevant) experiments for the second experiment, if a second experiment is necessary. There is no Nash-equilibrium over pure strategies and returning with a guess at any point is (obviously, given the infinities in the pay-off structure) not Nash. That is, if Nature "selects" the true causal structure among the two variables on the basis of flipping a fair three-sided coin, then Nature is making the discovery task maximally difficult (in the earlier described sense) for the Scientist. On the other side, by rolling a three-sided die to choose whether to intervene on X, intervene on Y or just passively observe in the first experiment, the Scientist is doing the best she can to discover Nature's secrets efficiently, given that Nature is an adversarial player. Any other strategy, even mixed, will do no better and may well be worse (or will allow Nature to adapt accordingly to make things worse). As a Nash equilibrium, this pair of strategies for Nature and the Scientist characterizes a state in which a unilateral move by Nature or by the Scientist does not improve their individual score.

The solution of the game is given by the value of the Nash equilibrium. It represents the expected pay-off to Nature (and loss to the Scientist) when playing the mixed strategy that is Nash. Since the game is zero-sum, the Nash equilibrium also corresponds to the mini-max solution, i.e. the Nash equilibrium gives us an upper bound on the expectation of the number of experiments necessary and sufficient to discover the causal structure. For this two variable game it is 5/3 experiments, slightly better than the fixed strategy bound of 2. The Scientist's strategy is in this case not an equalizer, since some graphs are resolved in one experiments and others in two. The Nash equilibrium is, if we ignore the indifference for the second experiment, unique. There is no Nash-equilibrium over pure strategies, i.e. there is no Nash equilibrium if Nature selects one particular causal structure with probability 1, since then the Scientist could adapt to do better by selecting one particular experiment with probability 1. But given the adapted behavior of the Scientist, Nature could improve her score by choosing a different causal structure with probability 1 – and the problem goes around in circles, without equilibrium.

Guessing (ending the game early) only becomes a viable option, when Nature is restricted to playing a subset of the possible structures. The mixed strategy of the Scientist is a Bayes solution, since it is a best response to the uniform distribution over structures. This is not the case for a fixed strategy with two experiments: There is no distribution over structures (degenerate or otherwise) which Nature might play such that the two-experiment-fixed-strategy (single interventions) constitutes the best response for the Scientist. The Scientist must use a mixed strategy to be Bayes. Or, from Nature's perspective: Nature cannot force the Scientist to the fixed strategy bound in expectation.

Interestingly, this last point does not apply in the case of three variable graphs. In the case of three variables, the game is substantially more complicated. There are 25 pure strategies for Nature (all DAGs over three variables) and well over 100 pure strategies for the Scientist (including all the early stops by guessing). We computed a Nash equilibrium that determines the solution of 2 for the game: The worst case expected number of experiments necessary and sufficient to determine the causal graph over three variables is two. That is, in the case of three variables, Nature can force the Scientist to the fixed strategy bound ($N - 1 = 2$) even in expectation, by placing a uniform distribution over the set of 10 graphs represented by the following three types of structures: one empty graph, three graphs with colliders and six complete graphs.



Intuitively one can see the reason for this result: If the Scientist does not intervene at all in the first experiment, then she will require between two and three experiments for the complete graphs (depending on where the intervention occurs in the second experiment) and only one for the common effects and the empty graph. In total the average is two experiments. If she intervenes on one variable in the first experiment, then – depending on the true graph – one of three possibilities occurs: (i) she cannot distinguish the common effect graph from the empty graph or (ii) she cannot distinguish the common effect graph from a complete graph or (iii) she cannot distinguish two complete graphs. In any case, she needs a second experiment. Intervening on two variables simultaneously would necessarily at least require two experiments, since the causal relation between the two intervened variables cannot be determined in the first experiment. We know from the results on fixed strategies that any combination of two different experiments involving single interventions resolves the graph over three variables. This distribution of graphs implies that these fixed (pure) strategies are best responses. Clearly, since they are fixed, and necessarily resolve the graph, they are also equalizers (same pay-off, no matter which graph is true). If we consider mixed strategies, they will not fare any better against this distribution by Nature (otherwise the pure strategies would not be best responses), but they are not all equalizers. For example, if the passive observation is included as a possible first experiment in the mixed strategy, then the mixed strategy is no longer an equalizer: 2/5 of the time it finds the graph in one experiment and 3/5 of the time it requires 8/3 experiments (i.e. two experiments on average overall). If the mixed strategy is restriced to experiments with interventions, then the mixed strategy is also an equalizer.

The Nash equilibrium is not unique even with regard to the mixtures Nature can play. But we conjecture that on three variables every mixed strategy for Nature that forms a Nash equilibrium, has support over at least some of the complete graphs. Obviously, no pure strategy for Nature is Nash, since guessing would then be optimal for the Scientist.

The two examples illustrate how the discovery problem is framed as a sequential game and how an analysis can proceed to obtain a worst case expected number of experiments. We proceed with more general results for mixed strategies on sets of variables with arbitrary size.

4.2 General Mixed Strategy Results

4.2.1 Single Interventions per Experiment

For single interventions per experiment, the three variable game is unique: For no other number of variables can Nature force the Scientist to the fixed strategy bound. In other words, only for three variables does the fixed strategy bound provide a Bayes solution (or a rationalizable strategy). The general result for mixed strategies using single interventions per experiment is given by the following theorem. It specifies the worst case expected length of the sequences of experiments necessary and sufficient to discover the causal structure among a set of N causally sufficient variables. It shows that single-intervention mixed strategies provide substantial improvement on their pure (fixed or adaptive) strategy counterparts.

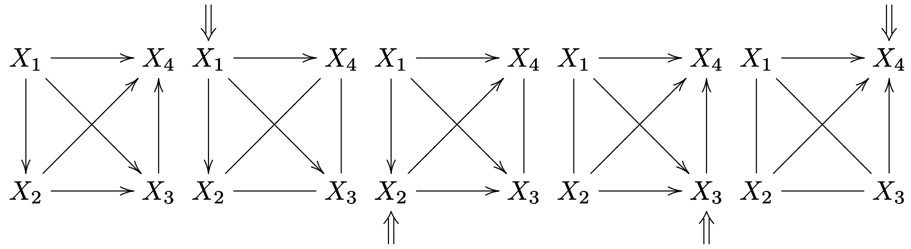
Theorem 4.2.1: (mixed strategy) Single Structural Interventions, Causally Sufficient
Given a set of $N > 3$ causally sufficient variables, the worst case expected number of experiments necessary and sufficient to discover the causal structure is

$\frac{2}{3}N - \frac{1}{3}$ experiments if only one variable can be subject to a structural intervention per experiment.

This bound is the value of a Nash equilibrium of the game: Nature plays a mixed strategy that is uniform over the complete graphs over N variables, and the Scientist uses the following strategy:

Strategy 4.2.2: (mixed) Single Structural Interventions, Causally Sufficient
Given N causally sufficient variables $X_1, \ldots, X_n$, let each experiment $\mathcal{E}_i$ in the sequence have $\mathbf{S_i} = \{X_j\}$, where X_j is selected uniformly from the variables that have not yet been subject to an intervention.

For any single intervention on a vertex in a complete graph, the intervention determines all outgoing edges from the intervened variable, it determines all edges incident on the intervened variable (since, given that the graph is complete, any independence with the intervened variable is an indication of an incoming edge), and it determines all edges that go between any ancestor and any descendent of the intervened variable (since the intervention breakes incoming edges, it creates unshielded colliders that can be discovered). Intuitively then, any single variable intervention cuts an undirected clique into two undirected sub-cliques by resolving all the edges that go between the two sub-cliques. Depending on which variable is subject to intervention, the two sub-cliques have different sizes. For example, suppose the true graph over four variables is given on the left, then the following four graphs are the post-manipulation graphs for every possible single intervention.



Depending on the location of the single intervention, the remaining indirected cliques vary in size between two and three. Since any variable is equally likely to be subject to an intervention, all cuts are equally likely. The bound for the mixed strategy is therefore based on a recursion of these cuts, assuming each is equally likely.

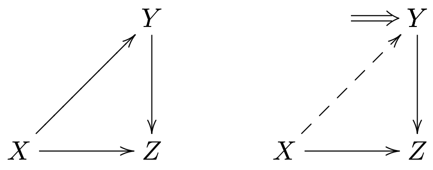
From Nature's perspective, there is no advantage in considering incomplete causal structures, since for $N > 3$ variables, two single interventions have to be performed anyway, and in those two experiments any missing edge would be detected.

4.2.2 Multiple Simultaneous Interventions per Experiment

For multiple simultaneous interventions the case is more complicated. Incomplete graphs do become relevant again and as a result the problem of determining the intervention set for the first experiment (see discussion of this problem in the section on adaptive strategies) returns. Furthermore, the space of possibilities that needs to be considered now pushes the boundaries of what becomes computationally feasible and so we resort to simulations to estimate the precise answer.

In the case of single interventions per experiment, for $N > 3$, at least two experiments are necessary anyway just to determine orientation information. Consequently, adjancency information can be established "for free" along the way, and hence there is no benefit to Nature in having any support over non-complete graphs (since a non-adjacency would amount to giving an orientation problem away for free). However, in the case of multiple simultaneous interventions, lack of support on the non-complete graphs is relevant to the Scientist's strategy. If Nature only had support over complete graphs, then the Scientist is able to infer the existence of an incoming edge into the intervened variable from an independence between an intervened and a non-intervened variable. For multiple simultaneous interventions adjacency information is lost between two intervened variables. So any multiple intervention search strategy must ensure it performs sufficient experiments to determine the adjanceny information. From Nature's perspective, to preserve maximum uncertainty about adjacency for the Scientist, Nature's graph distribution must have support over both the complete and some incomplete graphs.

A simple example will illustrate the point and be useful for the further discussion. Suppose the first of the graphs below over three variables is the true graph. The second graph shows the knowledge graph after a structural intervention on Y:



If it is known that the distribution of possible graphs *only* has support on complete graphs, or if it is known that it has *no* support on complete graphs, then the post-manipulation graph after an intervention on Y is sufficient to resolve the structure uniquely, to a complete graph or an unshielded collider, respectively. If there is support both over the graph that has an unshielded collider at Z and the graph that is complete, then a further experiment must be performed to distinguish the two.

If the true graph can be incomplete, then, as in the adaptive case, it is not clear from the outset, what the optimal size for the first intervention set is. But in comparison to the adaptive case this problem is aggravated, since in addition to the fact that the true causal structure may not be complete, the analysis of mixed strategies occurs in terms of the worst case *expected* number of experiments, and hence any bound on the number of experiments must take "lucky" choices of the intervention set into account. (The absolute worst case number of experiments is, of course, for mixed strategies the same as for fixed strategies.) Consequently, the choice of the size of the first intervention set is very sensitive to the distribution over possible graphs. If the distribution contains many or mainly sparse graphs, small intervention sets are better, since then non-adjacencies can be located and determined quickly; if graphs are near complete, larger intervention sets are better to efficiently resolve orientations. In general, the only heuristic one can recommend is that interventions sets should be slightly smaller than for fixed strategies, since not all graphs will be worst case graphs.

We do not know a closed form for the optimal size for the first intervention set for any distribution over graphs, and the problem is computationally not feasible using standard Nash equilibrium solvers. So we approach the problem by giving upper and lower bounds. Below we provide two sizes for the intervention set in a first experiment that are optimal for two scenarios that are relevant to determining the worst case expected number of experiments: If the true graph over N variables is complete and that information is *known* (to the

Scientist) and *used* to infer causal structure then the first row applies: MixCompleteKnown. If the true graph is some complete graph over the N variables, but its completeness is *not known* and *not used* in the inferences of the mixed strategy, the second row gives the optimal size (with regard to the expected number of experiments in the sequence) of the intervention set for the first experiment: mixCompleteUnknown. In some cases, as for $N = 7$, the difference between starting with two and starting with three variables in the first experiment is minimal. We only list the one best value, without making any claims about how detrimental a different choice would be to the expected number of experiments.

All the results are established on the basis of simulations sampling complete graphs ($1,000$ graphs for $N < 10$, 100 graphs for $10 \leq N \leq 12$ and 10 graphs above). The structure search in the simulations is performed using an independence oracle. We contrast the values for the mixed strategies with the *minimum* size of the intervention set of the first experiment of a fixed strategy that would still satisfy the $\log_2(N) + 1$ bound (Theorem 3.3.4) for multiple simultaneous interventions (third row):

N	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mixed Complete Known	1	2	2	2	2	2	2	2	2	2	2	2	2	2
Mixed Complete Unknown	1	1	1	2	3	3	4	4	3	3	3	3	3	3
Fixed Strategy	1	0	1	2	3	0	1	2	3	4	5	6	7	0

On the basis of these sizes for the first intervention set, we can specify a mixed search strategy for complete graphs:

Strategy 4.2.3: (mixed) Multiple Structural Interventions, Causally Sufficient

Given N causally sufficient variables, let the first experiment be a structural intervention on k variables selected uniformly from the N variables, where k corresponds to the appropriate entry in the above table. For every further experiment, the intervention set is selected by OPTINTER, where step 10 is a random sample among *admissible* vertices that are tied with regard to the number of their clique memberships, and step 12 is included.

The k variables can be sampled uniformly, since the space of complete graphs is symmetric with regard to every variable.

Despite the specification of a strategy (albeit mixed) for the Scientist, and thereby making the problem (from Naure's perspective) more decision theoretic than game theoretic, we do not yet have any closed form solutions to report on the worst case expected number of experiments when multiple simultaneous interventions are permitted in each experiment. The space of possibilities for Nature is still too large to compute with standard machinery. So even if we hypothesize that the above strategy is the Nash strategy for the Scientist, we are unable to compute the exact Nash equilibrium even for a particular $N > 3$. Instead, we compute upper and lower bounds of the exact worst case expected number of experiments and do so for N up to 13.

The upper and lower bounds on the exact worst case expectation are based on a conjecture that (a) the above strategy is the Nash strategy for the Scientist, and (b) that graphs sampled from a uniform distribution over complete graphs maximize the expectation of the length of the sequence of experiments, when *no* distributional information and no information other than Assumptions 1.2.1 to 3.2.6 and the knowledge graphs resulting from each of the experiments in a sequence are used to infer the true causal structure. That is discovery requires the most experiments if inferences are prohibited from an independence between X and Y when X is subject to an intervention and Y is not, to a direct cause $Y \rightarrow X$.

Intuitively, the upper and lower bound are constructed by not using all information (upper), or by using more information than is available in the exact game (lower). Complete graphs are the most difficult to resolve in terms of orientation, but to ensure that the determination of adjacency information is non-trivial, incomplete graphs must occur with some positive probability. To determine a lower bound on the exact solution, we use a uniform distribution that is restricted to complete graphs. Such a distribution is still going to supply a relatively high expected number of experiments, since all the orientation information must be resolved, but it is a strict lower bound since there is no support on incomplete graphs and hence the adjacency information is given by default. Consequently, some orientations can be resolved more easily (thanks to the inference from an independence between X and Y if X is subject to an intervention and Y is not, to a direct cause $Y \rightarrow X$).

In constrast, for the upper bound we take the opposite approach: We use the same distribution of graphs (uniform on complete graphs), but allow no distributional information into the structure search process, i.e. an adjacency may only be determined if the appropriate experiment(s) have been performed. This will provide an upper bound on the true worst case expectation since the entire adjacency information must be determined as if it were a distribution over all possible graphs, even though the actual distribution is restricted. In other words, Nature gets to play the most difficult graphs for search, but does not have to pay the price for having such narrow support. The Scientist, on the other hand, is up against the worst case graphs and does not get the occasional gain of a simpler graph, that would normally have to occur to prevent her from fast-tracking inferences to causal structure.

In order to compute both the upper and lower bound, we use the appropriate row from the earlier table (before Strategy 4.2.3) for the size of the first intervention set, and then simulate the search with an independence oracle. A full enumeration of all the possible ways that a complete graph could be subject to all the possible sequences of interventions is computationally not feasible even for fairly small N. We simulate the bounds by sampling for each N random complete graphs over N variables and then test the mixed strategy 4.2.3 given above, once with and once without the extra distributional information. For the lower bound the mixed strategy is supplemented with the information that the graph is complete by initializing it with a complete undirected knowledge graph, for the upper bound the mixed strategy is run without such information, i.e. initializing it with a knowledge graph containing only no-knowledge-edges.

For each $N < 10$ we performed 1,000 iterations, for $10 \leq N \leq 12$ we performed 100 iterations and for $N = 13$ only 10 iterations returned in reasonable time. We computed the average number of experiments over the iterations.

Figure 4.2 shows the $\log_2(N)+1$ bound of fixed Strategy 3.3.5 with multiple simultaneous interventions as dots. The two lines show the simulated upper and lower bounds on the worst case expected number of experiments for a mixed strategy using multiple simultaneous interventions per experiment. The figure shows that for multiple simultaneous interventions the fixed strategy is remarkably close to the worst case expected number of experiments, assuming that the simulated bounds are indicative of the true value. It appears that against the worst case distribution over graphs, mixed strategies cannot be expected to do much better than fixed strategies (and they are much harder to compute). The lines are not entirely smooth since there are various effects resulting from the discrete nature of the problem (discrete intervention sets, discrete numbers of experiments, etc.).

While the difference in the number of experiments between fixed and mixed strategies might not be so great, the table (before Strategy 4.2.3) of interven-

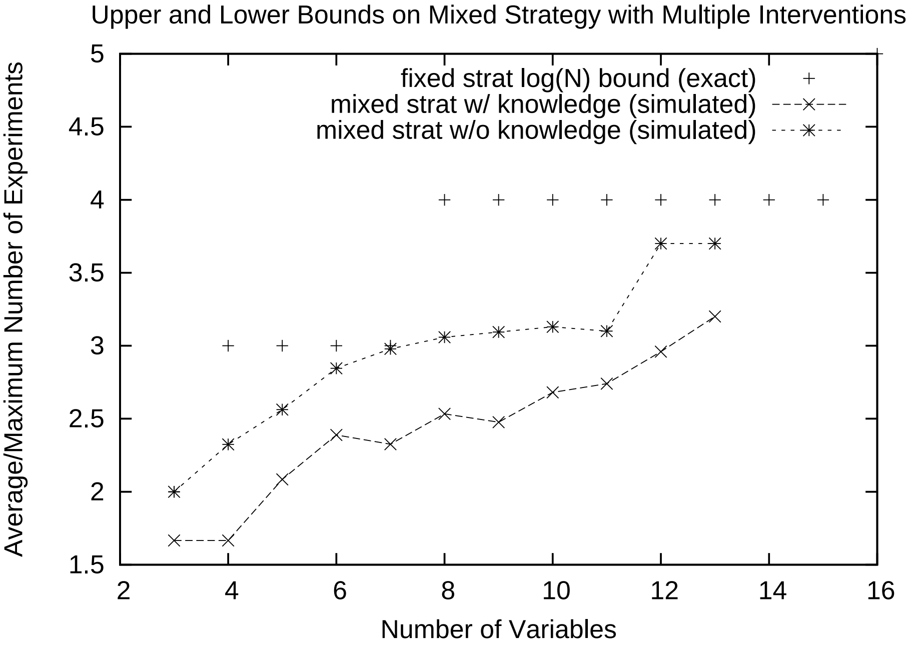
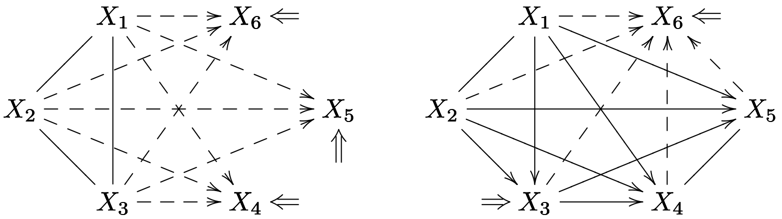


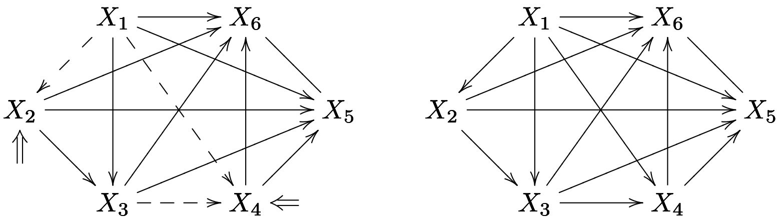
Figure 4.2: *Bounds for the worst case expected number of experiments for different N. The lower line is based on a uniform distribution over complete graphs, where knowledge of the distribution is used in the structure search. The upper line is based on the same distribution, but where the knowledge is not used. The dots indicate the* $\log_2(N) + 1$ *bound of fixed Strategy 3.3.5 with multiple simultaneous interventions.*

tion set sizes for the first experiment shows that mixed strategies may provide enormous savings in the number of variables that are subject to intervention. The upper and lower bound do not differ much in the number of variables subject to intervention in the first experiment, but there is a significant difference to the *minimum* number of variables necessary for the fixed strategy. This stark difference results from the fact that in the case of the mixed strategy, the search algorithm benefits if the root and the sink of the graph are *not* subject to intervention, and instead some variables evenly distributed across the interior of the tier-ordering of the graph are manipulated (as in the case of the middle variable of the three variable example earlier). Under such circumstances additional acyclicity constraints aid the search and allow recovery of more structure. In contrast, the fixed strategy always considers the worst case selection of variables for intervention, which generally includes variables at the extremes of the tier-ordering.

A simple example over six variables will help illustrate the point: Let the true graph be a complete graph over six variables, with a tier-ordering $X_1 \succ \ldots \succ X_6$, i.e. X_1 is the root and X_6 is the sink. If it is *known* that the graph is complete, a fixed strategy using multiple simultaneous interventions per experiment will still take three experiments, since it must ensure against unlucky choices of intervention sets, such as $\mathbf{S}_1 = \{X_4, X_5, X_6\}, \mathbf{S}_2 = \{X_3, X_6\}$ and $\mathbf{S}_3 = \{X_1, X_4\}$. But a mixed strategy, will select with probability $p = 1/\binom{N}{2}$ the intervetion set $\mathbf{S}_1 = \{X_2, X_4\}$, which will resolve the graph in two experiments. The figure below shows the knowledge graph after the first experiment for the fixed strategy, with $\mathbf{S}_1 = \{X_4, X_5, X_6\}$, on the left. Semi-directed edges are dashed, absence of edges indicates no-knowledge edges. Of course, since it is known that the graph is complete, no-knowledge edges really represent undirected edges here, and semi-directed edges can be resolved to directed edges. But we use this representation, so that it is clear what information has been established in the first experiment. The knowledge graph after the second experiment, combining the new information with the previous experiment is shown on the right.



While the semi-directed edges can be resolved with the help of background knowledge (it is a complete graph), the undirected edges between X_1, X_2 and X_4, X_5 make the third experiment necessary. In contrast, for the mixed strategy, the knowledge graph after an intervention on $\mathbf{S}_1 = \{X_2, X_4\}$ in the first experiment is shown on the left below and on the right after all semi-directed edges are resolved with the background knowledge.



In this case some remaining edges can be oriented due to acyclicity constraints: Since the graph is known to be complete, there must be a $X_2 - X_4$-edge, and since there is a path $X_2 \rightarrow X_3 \rightarrow X_4$, that missing edge must be oriented away from the cycle, i.e. $X_2 \rightarrow X_4$. That leaves only the X_5X_6-edge to be resolved in the next experiment.[2] With just two experiments and subjecting three variables to interventions, the mixed strategy is able to recover the graph. The point is that mixed search strategies may turn out to be very useful across a variety of cost measures.

4.3 Discussion: Mixed Strategies

We presented results on mixed strategies with single interventions and gave an account of how these strategies could be understood in game theoretic terms. Figure 4.3 differences between the different search strategies.[3] On the horizontal axis we have all possible distributions over graphs with N variables for some specific value of N.[4] The vertical axis shows the pay-off to Nature. The pay-off is infinite near the intersection of the axes and zero at the top. The only reason for the inverted axis is to ensure that up is better for the Scientist in the diagram, i.e. zero pay-off to Nature (zero experiments) is the best the Scientist can achieve.

The figure shows how fixed Strategy 3.3.2 provides a maxi-min guarantee of $N - 1$ experiments, no matter what the true graph is, and since it is fixed, it is independent of the outcome of particular experiments. But it never touches the best response surface. For any distribution over graphs it is strictly dominated by an adaptive or mixed strategy. However, as discussed in the section on adaptive strategies, any particular adaptive strategy touches the fixed strategy bound in the worst case, i.e. only weakly dominates the $N - 1$-fixed strategy. Mixed strategies lie between the best response surface and the lower bounds of the pure adaptive strategies, since mixed strategies contain the adaptive component and only randomize over indifferent intervention sets.[5] $2/3N - 1/3$ is

[2] If the mixed strategy had had an intervention set of size three in the first experiment, an intervention on $\{X_2, X_4, X_6\}$ would have resolved the graph in one experiment. But on average across complete graphs, an initial intervention set size of three is suboptimal.

[3] I am very grateful to Teddy Seidenfeld for repeatedly in different contexts pointing me to the importance of understanding game strategies in terms of this diagram. The figure is his, just adapted for my purposes and results.

[4] Since the diagram is only schematic, it is irrelevant whether one can actually arrange all distributions in one dimension.

[5] Mixed strategies touch the best response surface more often than is indicated in the

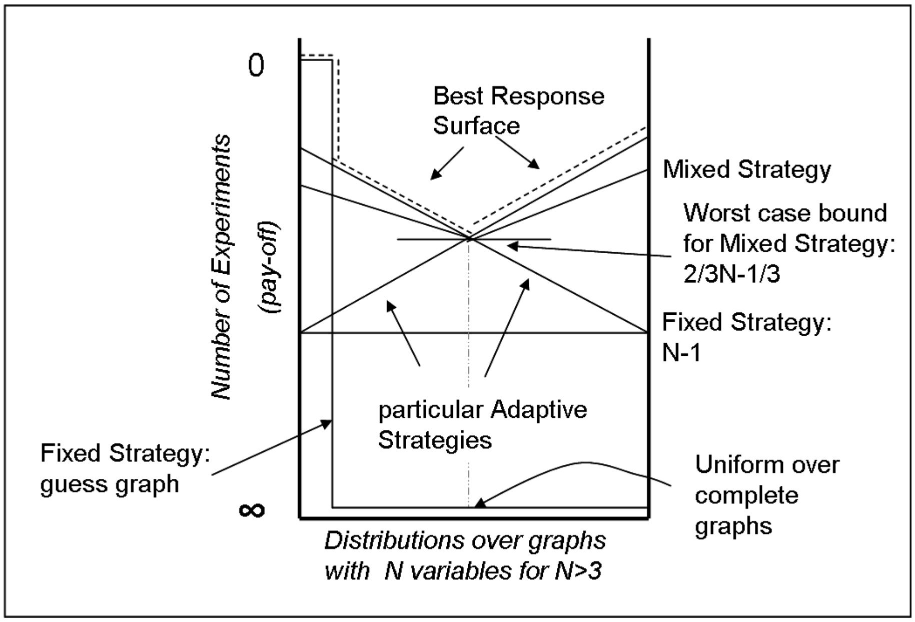


Figure 4.3: *Mixed, fixed and adaptive strategies using single structural interventions.*

the maxi-min for mixed strategies. It is the worst case expected number of experiments for any distribution over graphs. It is the best any mixed strategy can do in expectation against an adversarial opponent in this game; it specifies the security for the game. And since it is a zero-sum game, it is also a Nash equilibrium against a strategy by Nature that uses a uniform distribution over complete graphs. Mixed Strategy 4.2.2 attains this bound. Strategy 4.2.2 is a Bayes solution, since it touches the best response surface for a uniform distribution over complete graphs, which is the worst case prior for Nature. But Strategy 4.2.2 is not an equalizer, since for some graphs it can perform better and for some worse than $2/3N - 1/3$ (obviously, since experiments are counted discretely). The best response surface is not covered entirely by adaptive or mixed strategies. For particular (extreme) distributions over graphs there are fixed strategies that provide the best response, e.g. annoucing (guessing) graph G without any experiment when Nature has a degenerate distribution over the graph G only. But clearly, this strategy is an abismal candidate if the distri-

diagram here, especially for non-trivial distributions over graphs. But they do not touch the best response surface for all distributions (unless one considers pure strategies to be a subset of mixed strategies).

bution over graphs is not degenerate, and disastrous if the distribution has no support on the guessed graph, i.e. the strategy is not Nash. For $N = 3$, the worst case mixed strategy bound drops to the $N - 1 = 2$ fixed strategy bound.

The diagram is similar for multiple simultaneous interventions, only that the vertical separation of the different strategies (number of experiments) is much smaller. Since we do not have full results for mixed strategies involving multiple simultaneous interventions, we could only provide preliminary analyses giving upper and lower bounds for $N \leq 13$.

There is another way to frame the analysis of mixed strategies that removes the difference between single and multiple interventions. So far we considered the pay-off structure purely in terms of the number of experiments. However, if the pay-off structure were determined by the number of *different* variables that are subject to an intervention[6] throughout a sequence of experiments, then there is – at least for the worst case distribution – no difference between strategies that only intervene on a single variable per experiment and those that intervene on multiple simultaneously. In both cases, the same number of different variables have to be subject to an intervention at some point in the sequence. Consequently, all the bounds we gave so far for mixed strategies using single interventions can also be read as describing the relationship between the number of variables in the graph and the cost of causal structure search if the cost is measured in terms of the number of different variables subject to intervention.[7]

Before turning to search algorithms, there is a different way of framing the case for fixed strategies in game theoretic terms. Following Wald's [48] approach we framed the search for causal structure as a game between Nature and the Scientist. Nature gets the first move to determine the graph after which the Scientist has free reign. Nature is left to reveal whatever the experiments of the Scientist demand given the initial choice. But there is a different way of thinking that does not take Nature to be the "constant gardener", stuck within the constraints set by the initial choice. Nature might be just as devious as the Scientist and change her mind about the true underlying structure in light

[6] Note the emphasis on *different* variables. If the same variable is subject to an intervention in two different experiments, that would still count as one unit of cost. In most cases this is presumably an inadequate measure of cost. If a variable is subject to intervention in two different experiments, then the cost is generally not the same as when it is subject to intervention in just one experiment.

[7] I am grateful to Teddy Seidenfeld for alerting me to this connection.

of the experiment the Scientist proposes. On this view, the game alternates at each move between Nature and Scientist. But to keep the game interesting earlier choices must be respected. That is, Nature cannot pick a graph that is inconsistent with the information that was revealed in previous experiments. Nature is constrained to the equivalence class of graphs that each imply all the independence constraints revealed in the previous experiments.

In such a game, the worst case analysis reverts to the fixed strategy bounds. Since Nature can choose to rotate the true graph at each step in such a way that the Scientist's proposed experiment results in the least desirable intervention set, Nature can force the Scientist to perform $N-1$ or $\log_2(N)+1$ experiments even *in expectation*, depending on whether multiple simultaneous interventions are permitted. Analyses of adaptive or mixed strategies become redundant, since – if Nature is paying attention to the game – there is no hope of hitting a lucky intervention set.

Lastly, there is a third way of framing the game between the Scientist and Nature that we do not discuss here. Rather than Nature not being able to adapt at all after the first choice (our first analysis) or being able to adapt in light of the experiment proposed by the Scientist (the last analysis mentioned), Nature and the Scientist may adapt after each experiment, but without knowing what the next proposed experiment is – an intermediary view. So rather than being an alternating sequential game, we would have a sequence of games in which both players have to announce their move simultaneously (and, of course, Nature must remain consistent with the sequence of experiments so far).

Chapter 5

Search Algorithms

The bounds and strategies of the previous two chapters specify sequences of experiments that are sufficient and for some worst case necessary to recover the causal structure among a set of N variables. That is, these sequences of experiments guarantee under the specified assumptions that sufficient constraints can be recovered to uniquely identify the true causal graph. These constraints come in the form of independence constraints or in terms of constraints on differences in correlations generated by the different experiments in a sequence. In this chapter we supply the algorithms that peform the inference to the causal structure from the constraints generated over the sequence of experiments. We separate the discussion of algorithms into algorithms based on independence constraints and those based on differences in correlations, and among those based on independence constraints we distinguish posthoc and online algorithms. Posthoc algorithms are used for fixed strategies with pre-determined sequences of experiments. The algorithms are run after all experiments have been performed. Online algorithms are run after each experiment for in adaptive and mixed search strategies.

5.1 Posthoc Algorithms based on Independence Constraints

On a high level, structure search algorithms using independence relations have three components:

1. A stage that determines the information relevant to the structure search

from the experimental set-up. This is computed easily from the intervention set.

2. A structure search that incorporates knowledge of the experimental set-up and searches for the manipulated causal structure in one experiment.

3. A combining algorithm, that combines manipulated structures from different experiments to form one structure.

All of the algorithms we present here concentrate on the last point, the combination. For the first two parts we rely on existing structure search algorithms that can be supplemented with knowledge about the experimental set-up. All of the following algorithms are designed on the basis of the PC-algorithm [43]. But in principle any structure search algorithm with the appropriate guarantees could be substituted in the second stage. For the simulation studies in the next chapter we also tested the more cautious cPC-algorithm [33] and the score-based GES-algorithm [5]. All the bounds for search based on independence tests (Theorems 3.3.1, 3.3.4, 3.3.6, 3.3.8 and 3.3.9) are agnostic with regard to the specific structure search algorithm employed on the data of each experiment, as long as the structure search recovers the equivalence class of graphs whose manipulated structures given the experimental intervention(s) imply all and only the independence relations true in the manipulated distribution. For the PC-algorithm, and the cPC-algorithm, whose structure search is based explicitly on independence constraints, and for the GES-algorithm such guarantees are known for the large sample limit. Although originally designed for structure search in passive observational data, these algorithms can be supplemented with additional knowledge that allows search for structure in data that derives from a distribution subject to interventions.

In the case of sequences of experiments information about the true underlying structure has to be combined from different experiments. Some information about the causal structure can be obtained in different ways and consequently, decisions have to be made about how the overall search algorithm determines such information. For example, information about the orientation of edges can be obtained in several ways. The PC-algorithm performs an adjacency search in its first stage and determines orientations, where possible, in a second step on the basis of unshielded colliders and the further edge-orientations these imply (see section on the PC-algorithm). But orientation information can also be obtained from interventions: Structural interventions destroy all incoming causal arrows and leave only the outgoing ones. Parametric interventions create collider tests. In addition, edge-orientation can also be obtained from the combination of different experiments. For example, if there is a (causally sufficient) set of two variables X and Y and the first experiment is a structural intervention on X and the second a passive observation of both variables, then if the first experiment returns the variables to be independent, and the second shows an adjacency between the two, then one can infer from the combination, that the true graph must be $Y \rightarrow X$. In general, these different ways in which orientation is determined will have an implication on how many experiments are required to recover a particular graph. Under the idealizing assumption that an experiment returns the independencies true in the (manipulated) population, different approaches to resolving orientation do not conflict. But once statistical variability is considered, these different approaches might have quite different reliability and lead to a variety of different conflicts.

For the worst case bounds of fixed strategies we cannot count on the discovery of unshielded colliders (unless they are created by parametric interventions), since the worst case graphs are complete and therefore shield all their colliders. There is no way of ensuring that a structural intervention is placed in such a way that shields are guaranteed to be broken. Consequently, fixed strategies must ensure that they create sufficient constraints on the distributions that the combining algorithms can discover the causal structure no matter what it is. We consider the strategies in the order of the complexity of the combining algorithm.

5.1.1 Multiple Parametric Interventions: One Experiment

For the one experiment necessary and sufficient for causal discovery when multiple simultaneous parametric interventions are possible (Theorem 3.3.9), the selection of the intervention set is already given, namely all but one variable, and since there is only one experiment, no combining stage is needed. Consequently, any passive observational search algorithms only needs to be supplemented with the knowledge of which $N-1$ variables in $\mathbf{V}$ are subject to a parametric intervention. An intervention variable is added to these variables that is uncaused and that has a single direct edge into the intervened variable. That is, the set of variables $\mathbf{V}$ is augmented by the set on intervention variables $\mathbf{Pol}$ and all edges between intervention variables, and from any variable in $\mathbf{V}$ to any variable in $\mathbf{Pol}$ are excluded, as are all edges from intervention variables to variables in $\mathbf{V}$

other than one edge from each intervention variable to its intervened variable, which is forced. On the basis of this augmented set of variables and this background knowledge, any standard passive observational search algorithm can be applied.

If the intervention variables have no marginal distribution and only a joint distribution over **V** conditional on states of the intervention variables is given, then tests for differences in conditional probability are required, independence tests are insufficient.

5.1.2 Single Parametric Interventions: $N-1$ Experiments

The structure search using single parametric interventions (Theorem 3.3.8) is – like the case for multiple parametric interventions – based on creating collider tests. The intervention set is a single variable in each experiment. Which one, or in which order does not matter, as long as it is a different one each time. The structure search for a single experiment proceeds as in the previous case by augmenting the set of variables by – this time – a single intervention variable with a forced direct edge into the variable it intervenes upon. Edges into the intervention variable are excluded, as are all other edges from the intervention variable to other variables in **V**. The results of the collider tests must be combinined from different experiments.

Algorithm 5.1.1: Combining - Single Parametric Interventions
Assuming all experiments have been performed:

1. Initialize the algorithm with the $N-1$ knowledge graphs $K_1, \dots, K_{N-1}$, each K_i resulting from experiment $\mathcal{E}_i$, in which variable X_i was subject to a parametric intervention.

2. For each structure K_i, let K_i determine all edges (adjacency and direction), if any, *into* X_i and the edge, if it exists, between X_i and X_N, where X_N is the variable that was never subject to a parametric intervention in any of the experiments in the sequence.

3. If the addition of any edge creates a directed cycle, we have a conflict, and an error is returned.

The combining algorithm for single paramentric interventions ignores in the knowledge graphs K_i from each experiment all edges other than those found by a collider test at the intervened variable (and the edge into X_N). Of course, information about other edges might be useful (e.g. to resolve conflicts), once statistical variability is taken into account, but under the ideal assumptions of an independence oracle it is redundant and in principle the structure search could be restricted to search for only these edges in each experiment.

5.1.3 Structural Interventions: Linear or Logarithmic Number of Experiments

For fixed strategies using single or multiple structural interventions, the intervention set is again already pre-determined by the strategies (Strategies 3.3.2 and 3.3.5) or the OPTINTER algorithm (Strategy 3.3.28). The structure search algorithms are supplemented with slightly different information than in the parametric intervention case. There is no need to augment the set of variables. Only knowledge about the manipulated structure needs to be added, that is: For each intervened variable $X \in \mathbf{S}$, all edges *into* X from any other variable in $\mathbf{V}$ are excluded. The structure search algorithms for each experiment need not recover any orientation information. Adjacency information is sufficient as edge orientations can be inferred from the combinations of experiments or the experimental set-up. If structure search in each experiment is limited to adjacency search, the PC- and cPC-algorithms can be terminated before entering the orientation stage. (Again, it may be useful to nevertheless determine orientation information in multiple ways to help resolve conflicts.)

If one takes the experimental set-up into account, the structure search in each experiment returns a knowledge graph with

1. No-knowlegde edges between all pairs of variables X, Y, where $X, Y \in \mathbf{S}$. We say that X and Y were subject to a *structural zero-information test* in this experiment.

2. Adjacency edges between all pairs of variables X, Y, where $X, Y \in \mathbf{U}$. We say that X and Y were subject to a *structural adjacency test* in this experiment.

3. Semi-directed edges between all pairs of variables X, Y, where (i) $X \in \mathbf{S}$ and $Y \in \mathbf{U}$ and (ii) $X \perp\!\!\!\perp Y | \mathbf{C}$ for some conditioning set $\mathbf{C}$. We say that X and Y were subject to a *structural X-orientation test* in this experiment.

4. Direct cause edge between all pairs of variables X, Y, where (i) $X \in \mathbf{S}$ and $Y \in \mathbf{U}$ and (ii) $X \not\!\perp\!\!\!\perp Y | \mathbf{C}$ for all conditioning sets $\mathbf{C}$. Again, we say that X

and Y were subject to a *structural X-orientation test* in this experiment, but this time with a different outcome than the previous one (dependence instead of independence).

After all experiments are completed, the combining algorithm (for structural interventions) infers the orientation of an edge from the combination of different conditions that a pair of variables is subject to in the sequence of experiments.

Algorithm 5.1.2: Combining - Structural Interventions
Assuming all experiments have been performed:

1. For each pair of variables $X, Y \in \mathbf{V}$, select all the experiments in which it is not the case that both $X, Y \in \mathbf{S}$.

2. Output a directed edge $X \rightarrow Y$ in the final output graph if all of the following hold:

    (a) In all experiments in which only X (of the pair) is subject to a structural intervention, X and Y are found to be adjacent.

    (b) In all experiments in which only Y is subject to a structural intervention, X and Y appear non-adjacent.

    (c) In all experiments in which X and Y are passively observed, they appear adjacent.

    Analagously, if X and Y are switched. If the addition of any directed edge creats a cycle, then we have a conflict and report an error.

3. The pair of variables are determined to be non-adjacent in the final output graph if all of the following hold.

    (a) X and Y appear non-adjacent in all the experiments in which only X (of the two) is subject to a structural intervention.

    (b) X and Y appear non-adjacent in all experiments in which only Y (of the two) is subject to a structural intervention.

    (c) X and Y appear non-adjacent in all experiments in which they are both passively observed.

4. In all other cases, there is a conflict between the results of the individual experiments and an error is reported.

5.1.4 Multiple Structural Interventions, Causally Insufficient: N Experiments

For multiple simultaneous interventions on a causally insufficient set of variables (Theorem 3.3.6), the fixed strategy algorithm is trivial – none of the standard structure search algorithms are needed, since all but one variable are subject to a structural intervention in each experiment. One simply has to determine in each experiment whether the non-intervened variable covaries with any of the intervened variables. If it does, one adds an edge from the intervened variable to the non-intervened one. It is a simple procedure of multiple linear regression in each experiment.

5.2 Online Algorithms based on Independence Constraints

For adaptive and mixed strategies that consider the outcome of each experiment, intervention sets can no longer be pre-determined, but are contingent on the information gained from previous experiments.

A combining algorithm for adaptive and mixed strategies starts with a knowledge graph with no-knowledge edges between every pair of variables and then updates the knowledge graph after each experiment, making use of all the orientation information that can be gained.

Algorithm 5.2.1: Adaptive Combining - Structural Interventions
Initialize a knowledge graph K with no-knowledge edges between each pair of variables in $\mathbf{V}$. For each experiment perform a structure search using the information about the experimental set-up (forced and prohibited edges). Perform the following update of the knowledge graph after each experiment:

1. For each new experiment determine the post-manipulation graph M (with all orientation information that can be determined). Add no-knowledge edges between all pairs of variables X, Y, where both X and Y are in $\mathbf{S}$. Add semi-directed edges from X to Y if $X \in \mathbf{U}$ and $Y \in \mathbf{S}$.

2. For each pair of variables X and Y in $\mathbf{V}$ consider their connection in the knowledge graph K and the post-manipulation graph M. If X and Y are connected by a directed edge or are known to be non-adjacent in K, then

they are *known*. If the edge between X and Y is *unknown*, then there are 17 cases:

(a) $X -?- Y$ in K and $X -?- Y$ in M: Do nothing.

(b) $X -?- Y$ in K and $X \longrightarrow Y$ in M: Substitute $X \longrightarrow Y$ in K. If it creates a cycle of directed edges, there is a conflict and an error is reported. Similarly, if X and Y are switched.

(c) $X -?- Y$ in K and $X \text{ —— } Y$ in M: Substitute $X \text{ —— } Y$ in K.

(d) $X -?- Y$ in K and $X \dashrightarrow Y$ in M: Substitute $X \dashrightarrow Y$ in K. Similarly if X and Y are switched.

(e) $X -?- Y$ in K and $X \qquad Y$ in M: Substitute $X \qquad Y$ in K.

(f) $X \text{ —— } Y$ in K and $X -?- Y$ in M: Do nothing.

(g) $X \text{ —— } Y$ in K and $X \text{ —— } Y$ in M: Do nothing.

(h) $X \text{ —— } Y$ in K and $X \longrightarrow Y$ in M: Substitute $X \longrightarrow Y$ in K. If it creates a cycle of directed edges, there is a conflict and an error is reported. Similarly, if X and Y are switched.

(i) $X \text{ —— } Y$ in K and $X \dashrightarrow Y$ in M: Substitute $X \longrightarrow Y$ in K. If it creates a cycle of directed edges, there is a conflict and an error is reported. Similarly, if X and Y are switched.

(j) $X \text{ —— } Y$ in K and $X \qquad Y$ in M: There is a conflict and an error is reported.

(k) $X \dashrightarrow Y$ in K and $X -?- Y$ in M: Do nothing. Similarly, if X and Y are switched.

(l) $X \dashrightarrow Y$ in K and $X \dashrightarrow Y$ in M: Do nothing. Similarly, if X and Y are switched.

(m) $X \dashrightarrow Y$ in K and $X \dashleftarrow Y$ in M: Substitute $X \qquad Y$ in K. Similarly, if X and Y are switched.

(n) $X \dashrightarrow Y$ in K and $X \longrightarrow Y$ in M: Substitute $X \longrightarrow Y$ in K. If it creates a cycle of directed edges, there is a conflict and an error is reported. Similarly, if X and Y are switched.

(o) $X \dashrightarrow Y$ in K and $X \longleftarrow Y$ in M: There is a conflict and an error is reported. Similarly, if X and Y are switched.

(p) $X \dashrightarrow Y$ in K and $X \text{ —— } Y$ in M: Substitute $X \longrightarrow Y$ in K. If it creates a cycle of directed edges, there is a conflict and an error is reported. Similarly, if X and Y are switched.

(q) $X \dashrightarrow Y$ in K and $X \quad\quad Y$ in M: Substitute $X \quad\quad Y$ in K. Similarly, if X and Y are switched.

3. Infer all implied orientations of edges using the Meek Rules [23] and one additional rule: For any directed path from $X \rightarrow \ldots \rightarrow Y$ in knowledge graph K, if $X \dashleftarrow Y$ in K, then resolve the XY-edge as non-adjacent in K.[1]

4. If all edges in K are *known*, end the sequence of experiments, otherwise perform the next experiment.

The algorithm is greedy in the sense that it does not reconsider edges it has previously resolved. If we have an independence oracle, this is not a problem, since the oracle supplies the independence relations true for the manipulated distribution and so no errors, and hence no conflicts occur. We discuss in the next chapter what can be done when conflicts do occur.

In the case of mixed strategies, the analysis in combining the information from the sequence of experiments is the same as in the adaptive case. The random component of the mixed strategy is restricted to the selection of the intervention set *prior* to the experiment. There is no random component in the analysis.

5.3 Posthoc Algorithms based on Correlation-Tests: Structural Interventions

Causal structure search on the basis of differences in correlation becomes particularly relevant when causal sufficiency of the set of variables can no longer be guaranteed. Structural interventions make the intervened variable independent of its causes, latent or measured, while parametric interventions do not. Consequently, the search algorithms based on differences in correlations differ with regard to the type of intervention. We also separate the algorithm that searches for causal structure among the observed variables from the algorithm that then searches for latent variables given the structure over the observed variables.

We use a particular partial order over variables and paths such that connections between variables closely related in the tier ordering are considered before those that span several levels of the tier ordering:

[1] It is not known whether these rules are complete in the sense that they resolve all implied edges that sequences of experiments could give rise to in knowledge graphs. Probably not.

Definition 5.3.1: Partial Order over Variables: $O(\mathbf{V}, \succ)$
Given a set of N experiments on a set of N (causally insufficient) variables $\mathbf{V}$ such that each experiment $\mathcal{E}_i$ is a structural intervention on a different single variable X_i, define a partial oder ($\succ$) over the set of variables $\mathbf{V}$, such that $X_i \succ X_j$ if and only if $X_i \not\perp\!\!\!\perp X_j$ in $\mathcal{E}_i$.

Definition 5.3.2: Partial Order over Paths: $O(\mathcal{P}, \prec)$
Given a set $\mathcal{P}$ of directed paths, define a partial order over paths in $\mathcal{P}$ such that for any two paths $p_1 = X_1 \rightarrow \ldots \rightarrow X_r$ and $p_2 = Y_1 \rightarrow \ldots \rightarrow Y_s$, with $p_1, p_2 \in \mathcal{P}$ and $X_1, \ldots, X_r, Y_1, \ldots Y_s \in \mathbf{V}$, $p_1 \prec p_2$ if and only if there exists a path $p_3 \in \mathcal{P}$, such that $p_1 \subset p_3$ and $p_3 = Y_1 \rightarrow \ldots \rightarrow Y_s$ (i.e. p_1 is contained in p_3 and the endpoints of p_3 are the same as those of p_2).[2]

Definition 5.3.3: Manipulated Knowledge Graph
Given a knowledge graph $K = (\mathbf{V}, \mathbf{E})$ and an experiment $\mathcal{E}$ in which a set $\mathbf{S} \subseteq \mathbf{V}$ of variables is subject to a structural intervention, the manipulated knowledge graph of K is the graph where all edges incident on any intervened variable ($X \in \mathbf{S}$), all no-knowledge edges, all undirected edges and all semi-directed edges are removed. We will refer to the manipulated knowledge graph of K as $MK(K|\mathcal{E})$. It is a DAG.

We can now specify the algorithm:[3]

Algorithm 5.3.4: Single Structural Interventions and Correlation-Tests: Observed Structure
The algorithm assumes that N experiments, each involving a structural intervention on a single variable (Strategy 3.3.16) have been performed and that the model is a linear structural equation model. All correlations are assumed to be appropriately normalized so as to be interchangable with edge coefficients.

1. Initialize a knowledge graph K over the variables in $\mathbf{V}$, where each pair of variables is connected by a no-knowledge edge.

2. Given the N experiments, sort the variables in $\mathbf{V}$ according to the partial order $O(\mathbf{V}, \succ)$ over variables.

[2] p_1 and p_2 are not ordered when both their endpoints coincide, i.e. $X_1 = Y_1$ and $X_r = Y_s$. But they are ordered if p_3 starts or ends with p_1, i.e. when p_1 and p_2 share one endpoint and there is a path p_3 containing p_1 and connecting the endpoints of p_2.

[3] Following the defense of this thesis, Patrik Hoyer and the author developed a simpler implementation of this algorithm using matrix algebra. Incidentally, the revised version of the algorithm can also be used for the discovery of cyclic causal structures. Anyone interested should search for the appropriate publication before implementing this version of the algorithm.

3. For each pair of variables X, Y such that $X \succ Y$ and for which there is no other variable $Z \in \mathbf{V}$ such that $X \succ Z \succ Y$, substitute an edge $X \rightarrow Y$ in K and determine the correlation ρ_{XY} from the experiment where X is subject to an intervention. Let $e_{XY} = \rho_{XY}$ and associate e_{XY} with the directed edge $X \rightarrow Y$.

4. Sort all directed paths in K of length greater than two into a partial order over paths $O(\mathcal{P}, \prec)$.

5. For each path p with endpoints X and Y in $O(\mathcal{P}, \prec)$, starting from the smallest ones in the order, compute the total correlation ρ_{XY} between X and Y from the experiment where X was subject to a structural intervention.

6. Let $\mathcal{P}^*_{XY}$ be the set of all (unconditionally) *active* paths between X and Y in $MK(K|\mathcal{E}_X)$, where $\mathcal{E}_X$ is the experiment in which X was subject to a structural intervention.

7. Test whether the total correlation ρ_{XY} between the endpoints X and Y in $\mathcal{E}_X$ can be accounted for in terms of the correlation due to the paths in $\mathcal{P}^*_{XY}$ alone, i.e. if $\rho_{XY} = \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$, where e_i is an edge coefficient of an edge on one such path. If so, X and Y are determined to be non-adjacent in K. If not, substitute an edge $X \rightarrow Y$ in K and associate the difference in correlation $e_{XY} = \rho_{XY} - \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$ with that edge.

8. Paths created by the addition of the new edge are NOT included in $O(\mathcal{P}, \prec)$, i.e. the partial ordering over paths is not recomputed.

After all paths in $O^{\mathcal{P}}_{\succ}$ are considered and the appropriate edges added, the causal structure (and the edge coefficients) among the observed variables are established. The knowledge graph over the observed variables is a DAG. We can now turn to the latent variables.

Algorithm 5.3.5: Single (Structural) Interventions and Correlation-Tests: Latent Variables

Given the knowledge graph K determined by the previous algorithm with edge coefficients over the observed variables, let $T(\mathbf{V}, \succ_T) = T_1 \succ_T \ldots \succ_T T_w$ be a tier ordering over the variables in K with w tiers.

For i from 1 to w (in that order),

1. For all pairs of variables $X, Y \in T_i$, choose an experiment $\mathcal{E}_{XY}$ in which X and Y are passively observed.

2. Compare the known correlation τ^p_{XY} due to the paths in $\mathcal{P}^*_{XY}$ with the total passive correlation τ_{XY} measured in $\mathcal{E}_{XY}$. If $\tau^p_{XY} = \tau_{XY}$, then there is no latent common cause of X and Y. If not, then add a latent common cause $X \leftarrow L \rightarrow Y$ to the knowledge graph K and associate the difference between the correlations with the path $X \leftarrow L \rightarrow Y$.

3. For j in $i+1$ to w, while $i < j$,

   (a) For all pairs of variables X, Y such that $X \in T_i$ and $Y \in T_j$ choose an experiment $\mathcal{E}_{XY}$ in which X and Y are passively observed.

   (b) Let $\mathcal{P}^*_{XY}$ be the set of all (unconditionally) active paths between X and Y in $MK(K|\mathcal{E}_{XY})$.

   (c) Compare the correlation τ^k_{XY} due to the paths in $\mathcal{P}^*_{XY}$ with the total passive correlation τ_{XY} measured in $\mathcal{E}_{XY}$. If $\tau^p_{XY} = \tau_{XY}$, then there is no latent common cause of X and Y. If not, then add a latent common cause $X \leftarrow L \rightarrow Y$ to the knowledge graph K and associate the difference between the correlations with the path $X \leftarrow L \rightarrow Y$.

Return K with all the latent common causes and all its edge coefficients (or correlations due to the latent common cause).

Care must be taken when determining the active paths in a particular experiment. The comparisons are easier when one can test against an experiment in which all variables are passively observed, but such an experiment is not necessary. In some cases, reducing the number of active paths by interventions will make the test for comparison simpler and more reliable.

One of the major advantages of these two algorithms is that correlations can be computed on the basis of the entire data set of an experiment and so the algorithm does not run the risk of any conditional independence tests that – if the conditioning set is large – are impossible, because there is insufficient data. The real work is done in accounting for and testing differences between correlations due to different sets of pathways. These can be intricate tests, since the difference in correlation that a long pathway adds, might be very small. This is less of a problem in the first stage of the algorithm – the search for structure among the observed variables – since here the long paths are built up piece by piece and the test is whether an additional short direct path exists, which would presumably make a significant difference to the total correlation. So the problem of weak correlations due to long pathways mainly bites in the search for the presence of latent variables.

The algorithm adds a latent common cause only for pairs of variables. Consequently, if there is (in fact) one latent common cause of three variables, the algorithm will render this as three pair-wise common causes. However, we conjecture, that in certain circumstances we can recover such structure and structure *among* latent variables by applying the Build-Pure-Clusters-algorithm [42]. We have not implemented such a move, but there are two approaches one could take:

1. Initialize the BPC-algorithm with the known structure and edge coefficients over the observed variables.

2. Or, remove the correlation due to the observed structure from the overall correlations between variables and run the BPC algorithm on the residual correlations.

In the case of multiple simultaneous interventions the analysis of the data is largely similar. However, there are two advantages: First, one can establish the tier ordering in fewer experiments — $O(\log_2(N))$ experiments instead of N experiments — and second, more direct connections between variables can be read off immediately since any indirect path in one experiment can be at most as long as the number of variables not simultaneously subject to an intervention. We only show the steps that change in comparison to the single intervention algorithm:

Algorithm 5.3.6: Multiple Structural Interventions and Correlation Test: Observed Variables:

Step 2: Sort the variables in $\mathbf{V}$ into a partial ordering, where $X \succ Y$ if $X \not\perp\!\!\!\perp Y$ in some experiment in which X is subject to a structural intervention and Y is passively observed.

Step 5: For each path p with endpoints X and Y in $O(\mathcal{P}, \prec)$, starting from the smallest ones in the order, compute the total correlation ρ_{XY} between X and Y in an experiment $\mathcal{E}^*$ where $X \in \mathbf{S}$ and $Y \in \mathbf{U}$. Choose $\mathcal{E}^*$ such that the largest number of known active paths between X and Y are broken.

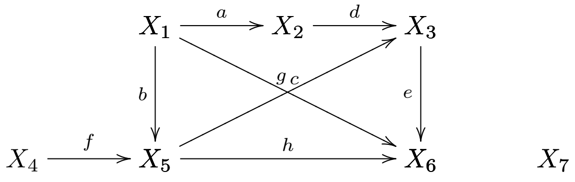
Step 6: Let $\mathcal{P}^*_{XY}$ be the set of all (unconditionally) *active* paths between X and Y in $MK(K|\mathcal{E}^*)$.

Step 7: Test whether the total correlation ρ_{XY} between the endpoints X and Y in $\mathcal{E}^*$ can be accounted for in terms of the correlation due to the paths in $\mathcal{P}^*_{XY}$ alone, i.e. if $\rho_{XY} = \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$, where e_i is an edge coefficient of an edge on one such path. If so, X and Y are determined to be non-adjacent in K. If not, substitute an edge $X \rightarrow Y$ in K and associate the difference in correlation $e_{XY} = \rho_{XY} - \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$ with that edge.

There is no difference in the search for latent variables. One must only ensure that there is an experiment from which the appropriate passive observational correlations can be determined (which is not guaranteed automatically, unlike the single intervention case).

Example

Let the true graph among seven observed variables $X_1, \ldots, X_7$ be the graph below, with the edge coefficients specified as small letters.



X_7 is disconnected from the causal structure among the observed variables. But, assume that for each pair off variables there is a latent common cause resulting in an additional correlation between the pair of variables specified by the following table:

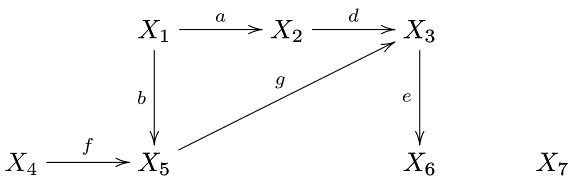
Pair	Latent	Correlation due to Latent	Pair	Latent	Correlation due to Latent
(X_1, X_2)	L_1	δ_1	(X_3, X_4)	L_{12}	δ_{12}
(X_1, X_3)	L_2	δ_2	(X_3, X_5)	L_{13}	δ_{13}
(X_1, X_4)	L_3	δ_3	(X_3, X_6)	L_{14}	δ_{14}
(X_1, X_5)	L_4	δ_4	(X_3, X_7)	L_{15}	δ_{15}
(X_1, X_6)	L_5	δ_5	(X_4, X_5)	L_{16}	δ_{16}
(X_1, X_7)	L_6	δ_6	(X_4, X_6)	L_{17}	δ_{17}
(X_2, X_3)	L_7	δ_7	(X_4, X_7)	L_{18}	δ_{18}
(X_2, X_4)	L_8	δ_8	(X_5, X_6)	L_{19}	δ_{19}
(X_2, X_5)	L_9	δ_9	(X_5, X_7)	L_{20}	δ_{20}
(X_2, X_6)	L_{10}	δ_{10}	(X_6, X_7)	L_{21}	δ_{21}
(X_2, X_7)	L_{11}	δ_{11}			

With passive observational data alone, the FCI-algorithm [43] would even in the large sample limit only recover a complete graph of non-directed edges.

We will use the algorithm for the single intervention case (Algorithms 5.3.4 and 5.3.5). Suppose we have N experiments $\mathcal{E}_1, \ldots, \mathcal{E}_N$ such that each $\mathcal{E}_i$ is a structural intervention on X_i. In Step 2 of the algorithm we find that

$$\begin{aligned}
X_1 \not\perp\!\!\!\perp \{X_2, X_3, X_5, X_6\} \quad &\text{and} \quad X_1 \perp\!\!\!\perp \{X_4, X_7\} \text{ in } \mathcal{E}_1 \\
X_2 \not\perp\!\!\!\perp \{X_3, X_6\} \quad &\text{and} \quad X_2 \perp\!\!\!\perp \{X_1, X_4, X_5, X_7\} \text{ in } \mathcal{E}_2 \\
X_3 \not\perp\!\!\!\perp X_6 \quad &\text{and} \quad X_3 \perp\!\!\!\perp \{X_1, X_2, X_4, X_5, X_7\} \text{ in } \mathcal{E}_3 \\
X_4 \not\perp\!\!\!\perp \{X_3, X_5, X_6\} \quad &\text{and} \quad X_4 \perp\!\!\!\perp \{X_1, X_2, X_7\} \text{ in } \mathcal{E}_4 \\
X_5 \not\perp\!\!\!\perp \{X_3, X_6\} \quad &\text{and} \quad X_5 \perp\!\!\!\perp \{X_1, X_2, X_4, X_7\} \text{ in } \mathcal{E}_5 \\
X_6 \not\perp\!\!\!\perp \emptyset \quad &\text{and} \quad X_6 \perp\!\!\!\perp \{X_1, X_2, X_3, X_4, X_5, X_7\} \text{ in } \mathcal{E}_6 \\
X_7 \not\perp\!\!\!\perp \emptyset \quad &\text{and} \quad X_7 \perp\!\!\!\perp \{X_1, X_2, X_3, X_4, X_5, X_6\} \text{ in } \mathcal{E}_7
\end{aligned}$$

From this information, we can generate a partial order: $X_1 \succ \{X_2, X_5\} \succ X_3 \succ X_6$ and $X_4 \succ X_5 \succ X_3 \succ X_6$. This can be used to build a partial order graph (POG), shown below, and according to the POG, we can obtain the edge coefficients for an edge $X \rightarrow Y$ in the POG by determining the correlation between X and Y in the experiment in which X was subject to a structural intervention.



In the POG there are 8 directed paths of length greater than one:

$$\begin{gathered}
X_5 \rightarrow X_3 \rightarrow X_6 \\
X_2 \rightarrow X_3 \rightarrow X_6 \\
X_1 \rightarrow X_2 \rightarrow X_3 \\
X_1 \rightarrow X_5 \rightarrow X_3 \\
X_4 \rightarrow X_5 \rightarrow X_3 \\
X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_6 \\
X_1 \rightarrow X_5 \rightarrow X_3 \rightarrow X_6 \\
X_4 \rightarrow X_5 \rightarrow X_3 \rightarrow X_6
\end{gathered}$$

If they are sorted according to the partial ordering of paths described earlier, one order is given in the above list. We now consider the paths in that order to identify additional edges. We will use ρ_{XY} to refer to the unconditional correlation between X and Y in the experiment where X was subject to a structural intervention. We assume that the correlations are appropriately normalized so that $\rho_{XY} = \sum_p \prod_{i \in p} e_i$, where p are the active paths between X and Y in the experiment where X is subject to a structural intervention and e_i is an edge coefficient of an edge on an active path p:

1. Since $\rho_{X_5 X_6} \neq ge$, a directed edge $X_5 \rightarrow X_6$ is added with an edge coefficient $h = \rho_{X_5 X_6} - ge$. Note that h corresponds to the edge in the true graph, since we know all paths other than the direct one that are active between X_5 and X_6 in the experiment where X_5 is subject to a structural intervention. The added edge will be taken into account in the following tests.

2. Since $\rho_{X_2 X_6} = de$, no direct edge is added between X_2 and X_6.

3. Since $\rho_{X_1 X_3} = ad + bg$, no direct edge is added between X_1 and X_3.

4. X_1 and X_3 have already been considered for a direct connection, hence this path can be skipped.

5. Since $\rho_{X_4X_3} = fg$, no direct edge is added between X_4 and X_3.

6. Since $\rho_{X_1X_6} \neq ade + bge + bh$, a direct edge $X_1 \rightarrow X_6$ is added with an edge coefficient $c = \rho_{X_1X_6} - (ade + bge + bh)$.

7. The X_1X_6-connection has already been considered. Skip.

8. Since $\rho_{X_4X_6} = fge + fh$, no direct edge is added between X_4 and X_6.

At this point, the causal structure among the observed variables has been established. We now proceeed to search for latent variables. We use τ_{XY} to refer to the passive observational correlation between X and Y. In general, the algorithm only requires that X and Y are passively observed when they are considered for latent confounding. Other variables may be subject to an intervention, since the relevant active paths can be computed given the intervention sets, so that the total correlation can be adjusted accordingly. But we will assume for simplicity of reference that τ_{XY} refers to an experiment in which *all* variables are passively observed.

The tier ordering of the variables given the causal structure among the observed variables is

$$\{X_1, X_4, X_7\} \succ \{X_2, X_5\} \succ \{X_3\} \succ \{X_6\}$$

First, all pairs within the top tier are considered, then all pairs with one variable in the first and one variable in the second tier, then pairs of one and three, and only after all pairs of one and four (the lowest tier) have been considered, do pairs within tier two follow.

1. Since $\tau_{X_1X_4} \neq 0$, a latent common cause L_3 is added and a correlation $\delta_3 = \tau_{X_1X_4}$ due to L_3 is associated with the latent variable. The correlation due to this latent path is taken into account in the further tests.

2. Since $\tau_{X_1X_7} \neq 0$, a latent common cause L_6 is added and a correlation $\delta_6 = \tau_{X_1X_7}$ due to L_6 is associated with the latent variable.

3. Since $\tau_{X_4X_7} \neq 0$, a latent common cause L_{18} is added and a correlation $\delta_{18} = \tau_{X_4X_7}$ due to L_{18} is associated with the latent variable.

4. Since $\tau_{X_1X_2} \neq a$, a latent common cause L_1 is added and a correlation $\delta_1 = \tau_{X_1X_2} - a$ due to L_1 is associated with the latent variable.

5. Since $\tau_{X_1X_5} \neq b+f\delta_3$, a latent common cause L_4 is added and a correlation $\delta_4 = \tau_{X_1X_5} - (b + f\delta_3)$ due to L_4 is associated with the latent variable.

6. Since $\tau_{X_4X_2} \neq a\delta_3$, a latent common cause L_8 is added and a correlation $\delta_8 = \tau_{X_2X_4} - a\delta_3$ due to L_8 is associated with the latent variable.

7. Since $\tau_{X_4X_5} \neq f + b\delta_3$, a latent common cause L_{16} is added and a correlation $\delta_{16} = \tau_{X_4X_5} - (f + b\delta_3)$ due to L_{16} is associated with the latent variable.

8. Since $\tau_{X_7X_2} \neq a\delta_6$, a latent common cause L_{11} is added and a correlation $\delta_{11} = \tau_{X_2X_7} - a\delta_6$ due to L_{11} is associated with the latent variable.

9. Since $\tau_{X_7X_5} \neq b\delta_6 + f\delta_{18}$, a latent common cause L_{20} is added and a correlation $\delta_{20} = \tau_{X_7X_5} - (b\delta_6 + f\delta_{18})$ due to L_{20} is associated with the latent variable.

10. Since $\tau_{X_1X_3} \neq ad + bg + d\delta_1 + g\delta_4 + fg\delta_3$, a latent common cause L_2 is added and a correlation $\delta_2 = \tau_{X_1X_3} - (ad + bg + d\delta_1 + g\delta_4 + fg\delta_3)$ due to L_2 is associated with the latent variable.

11. Since $\tau_{X_4X_3} \neq fg + g\delta_{16} + d\delta_8 + (bg + ad)\delta_3$, a latent common cause L_{12} is added and a correlation $\delta_{12} = \tau_{X_4X_3} - (fg + g\delta_{16} + d\delta_8 + (bg + ad)\delta_3)$ due to L_{12} is associated with the latent variable.

12. Since $\tau_{X_7X_3} \neq d\delta_{11} + (ad + bg)\delta_6 + g\delta_{20} + fg\delta_{18}$, a latent common cause L_{15} is added and a correlation

$$\delta_{15} = \tau_{X_7X_3} - (d\delta_{11} + (ad + bg)\delta_6 + g\delta_{20} + fg\delta_{18})$$

due to L_{15} is associated with the latent variable.

13. Since $\tau_{X_1X_6} \neq ade + bge + bh + c + de\delta_1 + e\delta_2 + (ge + h)\delta_4 + (ge + h)f\delta_3$, a latent common cause L_5 is added and a correlation

$$\delta_5 = \tau_{X_1X_6} - (ade + bge + bh + c + de\delta_1 + e\delta_2 + (ge + h)\delta_4 + (ge + h)f\delta_3)$$

is associated with the latent variable.

14. Since $\tau_{X_4X_6} \neq fge + fh + (ge+h)\delta_{16} + de\delta_8 + e\delta_{12} + (bge+ade+bh+c)\delta_3$, a latent common cause L_{17} is added and a correlation

    $$\delta_{17} = \tau_{X_4X_6} - (fge + fh + (ge+h)\delta_{16} + de\delta_8 + e\delta_{12} + (bge+ade+bh+c)\delta_3)$$

    due to L_{17} is associated with the latent variable.

15. Since $\tau_{X_7X_6} \neq e\delta_{15} + de\delta_{11} + (ade+c+bge+bh)\delta_6 + (ge+h))\delta_{20} + (ge+h)f\delta_{18}$, a latent common cause L_{21} is added and a correlation

    $$\delta_{21} = \tau_{X_7X_6} - (e\delta_{15} + de\delta_{11} + (ade+c+bge+bh)\delta_6 + (ge+h))\delta_{20} + (ge+h)f\delta_{18})$$

    due to L_{21} is associated with the latent variable.

16. Since $\tau_{X_2X_5} \neq ab + a\delta_4 + b\delta_1 + af\delta_3 + f\delta_8$, a latent common cause L_9 is added and a correlation $\delta_9 = \tau_{X_2X_5} - (ab + a\delta_4 + b\delta_1 + af\delta_3 + f\delta_8)$ due to L_9 is associated with the latent variable.

17. Since $\tau_{X_2X_3} \neq d + abg + bg\delta_1 + a\delta_2 + afg\delta_3 + ag\delta_4$, a latent common cause L_7 is added and a correlation

    $$\delta_7 = \tau_{X_2X_3} - (d + abg + bg\delta_1 + a\delta_2 + afg\delta_3 + ag\delta_4)$$

    due to L_7 is associated with the latent variable.

18. Since $\tau_{X_5X_3} \neq g + abd + db\delta_1 + ad\delta_4 + adf\delta_3 + d\delta_9 + +fd\delta_8 + b\delta_2 + f\delta_{12}$, a latent common cause L_{13} is added and a correlation

    $$\delta_{13} = \tau_{X_5X_3} - (g + abd + db\delta_1 + ad\delta_4 + adf\delta_3 + d\delta_9 + +fd\delta_8 + b\delta_2 + f\delta_{12})$$

    due to L_{13} is associated with the latent variable.

19. Since $\tau_{X_2X_6} \neq de + ac + ab(h+ge) + e\delta_7 + (c+bh+bge)\delta_1 + ae\delta_2 + (ge+h)a\delta_4 + a\delta_5 + (ge+h)af\delta_3 + (ge+h)\delta_9 + (ge+h)f\delta_8$, a latent common cause L_{10} is added and a correlation

    $$\begin{aligned}\delta_{10} &= \tau_{X_2X_6} - (de + ac + ab(h+ge) + e\delta_7 + (c+bh+bge)\delta_1 + ae\delta_2 + \\ &\quad (ge+h)a\delta_4 + a\delta_5 + (ge+h)af\delta_3 + (ge+h)\delta_9 + (ge+h)f\delta_8)\end{aligned}$$

    due to L_{10} is associated with the latent variable.

20. Since

$$\begin{aligned}\tau_{X_5X_6} \neq\ & h + ge + bc + bade + dbe\delta_1 + be\delta_2 + (ade + c)f\delta_3 + (ade + c)\delta_4 \\ & + b\delta_5 + def\delta_8 + de\delta_9 + fe\delta_{12} + e\delta_{13} + f\delta_{17}\end{aligned}$$

a latent common cause L_{19} is added and a correlation

$$\begin{aligned}\delta_{19} =\ & \tau_{X_5X_6} - [h + ge + bc + bade + dbe\delta_1 + be\delta_2 + (ade + c)f\delta_3 + (ade + c)\delta_4 \\ & + b\delta_5 + def\delta_8 + de\delta_9 + fe\delta_{12} + e\delta_{13} + f\delta_{17}]\end{aligned}$$

due to L_{19} is associated with the latent variable.

21. Since

$$\begin{aligned}\tau_{X_3X_6} \neq\ & e + adc + gh + adbh + (bh + c)d\delta_1 + (bh + c)\delta_2 + (had + gc)f\delta_3 + \\ & (adh + gc)\delta_4 + (ad + bg)\delta_5 + dfh\delta_8 + dh\delta_9 + d\delta_{10} + fh\delta_{12} \\ & + h\delta_{12} + fg\delta_{17} + g\delta_{19}\end{aligned}$$

a latent common cause L_{14} is added and a correlation

$$\begin{aligned}\delta_{14} =\ & \tau_{X_3X_6} - [e + adc + gh + adbh + (bh + c)d\delta_1 + (bh + c)\delta_2 + (had + gc)f\delta_3 + \\ & (adh + gc)\delta_4 + (ad + bg)\delta_5 + dfh\delta_8 + dh\delta_9 + d\delta_{10} + fh\delta_{12} \\ & + h\delta_{12} + fg\delta_{17} + g\delta_{19}]\end{aligned}$$

due to L_{14} is associated with the latent variable.

At this point all latent variables have been discovered. The algorithm outputs a latent variable for every pair of variables even if in fact there may be latent variables that cause more than two of the observed variables. Discovering this structure among the latent variables has to be determined in a further step, using something like the BPC-Algorithm [42].

5.4 Posthoc Algorithms based on Correlation-Tests: Parametric Interventions

If the sequence of experiments involves parametric interventions, then the algorithms of the previous section are inadequate since the correlation due to latent variables is never broken to reveal the structure among the observed variables. However, as in the causally sufficient case, we can adapt the algorithm here to take the intervention variables into consideration. Under the assumption that the intervention variables are not counfounded with any other variable in $\mathbf{V}$ and that no variable in $\mathbf{V}$ is a cause of any of the intervention variables, we can re-analyze on the basis of differences in correlation the single experiment where each variable is subject to a parametric intervention. The procedure is largely the same as the algorithm for structural interventions, only that this time one does not test the correlation between the *intervened* variable and some other variable, but between the *intervention* variable and some other variable. Again we separate the algorithm into two stages: the stage for structure search among observable variables and the stage for search for latent variables, but since the latent variable search matches the algorithm for structural interventions, we only describe the search for observable structure:

Algorithm 5.4.1: Parametric Interventions and Correlation-Tests: Observable Structure

The algorithm assumes that the set of variables $\mathbf{V}$ is causally insufficient, that one experiment $\mathcal{E}$ involving a parametric intervention on each variable in $\mathbf{V}$ has been performed and that the model is linear.

1. Initialize a knowledge graph K over the variables in $\mathbf{V}$ and their intervention variables in **Pol**, where each pair of variables in $\mathbf{V}$ is connected by a no-knowledge edge and each intervention variable in **Pol** has a directed edge into the variable it intervenes on and is otherwise considered non-adjacent to any other variable in K.

2. Sort the variables in $\mathbf{V}$ into a partial ordering, where $X \succ Y$ iff $I_X \not\perp\!\!\!\perp Y$ in the experiment. (Note that the independence check is against the intervention variable I_X and not against the intervened variable X.)

3. For each X in $\mathbf{V}$, determine the correlation $\rho_{I_X,X}$ between X and I_X and associate $e_{I_X,X} = \rho_{I_X,X}$ with the edge $I_X \rightarrow X$.

4. For each pair of variables X, Y such that $X \succ Y$ and for which there is no other variable $Z \in \mathbf{V}$ such that $X \succ Z \succ Y$, substitute an edge $X \rightarrow Y$ in K and determine the edge-coefficient by dividing the total correlation between I_X and Y by the edge coefficient $e_{I_X,X}$, i.e. $e_{XY} = \rho_{I_X,Y}/e_{I_X,X}$. Associate e_{XY} with the directed edge $X \rightarrow Y$.

5. Sort all directed paths in K of length greater than two into a partial order over paths $O(\mathcal{P}, \prec)$ (see Algorithm 5.3.4).

6. For each path p with endpoints X and Y in $O(\mathcal{P}, \prec)$, starting from the smallest ones in the order, compute the total correlation ρ_{XY} between X and Y by $\rho_{X,Y} = \rho_{I_X,Y}/e_{I_X,X}$.

7. Let $\mathcal{P}^*_{XY}$ be the set of all (unconditionally) *active* paths between X and Y in the knowledge graph K[4]

8. Test whether the total correlation ρ_{XY} between the endpoints X and Y in $\mathcal{E}$ can be accounted for in terms of the correlation due to the paths in $\mathcal{P}^*_{XY}$ alone, i.e. if $\rho_{XY} = \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$, where e_i is an edge coefficient of an edge on one such path. If so, X and Y are determined to be non-adjacent in K. If not, substitute an edge $X \rightarrow Y$ in K and associate the difference in correlation $e_{XY} = \rho_{XY} - \sum_{p \in \mathcal{P}^*_{XY}} \prod_{e_i \in p} e_i$ with that edge.

9. Paths created by the addition of the new edge are NOT included in $O(\mathcal{P}, \prec)$, i.e. the partial ordering over paths is not recomputed.

After all paths in $O^{\mathcal{P}}_{\succ}$ are considered and the appropriate edges added, the causal structure (and the edge coefficients) among the observed variables are established. The knowledge graph over the observed variables is a DAG. The search for latent variables proceeds in exactly the same way as for structural interventions.

The search algorithm depends on a joint distribution over the variables of interest, $\mathbf{V}$, and the set of intervention variables $\mathbf{Pol}$. It is an open question whether tests of differences in correlations could be replaced with tests of differences in the conditional probability distribution $P(\mathbf{V}|\mathbf{Pol})$. It is also important to note that the parametric intervention may not influence the edge coefficients of the causal relations between variables under consideration. This is guaranteed in a linear model, in which the parametric intervention simply adds a linear factor, but it need not hold for general parametric interventions.

[4] No need to consider a manipulated knowledge graph here since the interventions are only parametric. The active paths are determined relative to the directed edges in the knowledge graph.

Chapter 6

Statistical Variability

[1] Throughout the entire thesis so far we have assumed that we are able to resort to an oracle that returns true responses to our tests for constraints. This separation allowed us to consider the combinatorics of the experiments for discovery independently of statistical variability. Before we turn to simulation results in the next chapter we discuss in this chapter some of the additional issues that arise when the search strategies must take statistical variability into account. We already indicated in describing the algorithms where issues resulting from statistical variability enter.

6.1 Conflicts

In search for the causal structure among a set of variables based on independence tests the absence of a particular causal arrow between X and Y is determined by the existence of *some* conditioning set that makes the two variables independent. However, the presence of an edge implies that the two variables in question remain dependent for *all* possible conditioning sets. In large dense networks the search may require a very large number of conditioning sets to determine adjacency. Consequently, the likelihood of all independence tests returning the correct result decreases as the number of tests increases. This is exaggerated by the fact that the available number of data points for a particular independence test gets smaller as the conditioning sets increase. In some cases we are able to detect errors. In sequences of experiments, results from many

[1] This section contains verbatum text and results from [9].

different experiments have to be combined. Some experiments can be used without additional cost in the number of experiments to repeat statistical tests while others provide constraints on the causal structure that are only consistent with particular results from other experiments.

We have a *conflict* if the results from different constraint tests are inconsistent with any causal structure that – appropriately manipulated given the interventions of the specific experiment – is assumed to generate the data in the different experiments. For example, one of the simplest conflicts occurs if we have two variables X and Y and collect two separate data-sets in which both variables are passively observed, but in one data-set they appear dependent (suggesting a causal connection), while they are determined to be independent in the other. No single causal structure among the two variables is consistent with these results. In this case possible resolutions seem obvious (e.g. pooling the data), but the problem is more general:

Conflicts can arise for a variety of reasons in an inference process. As in the example above, we can test for the same independence constraint when the distribution over the relevant variables is the same and we nevertheless may obtain conflicting results because of sample variability. Such a conflict may arise when one experiment is repeated or when two different experiments have the same marginal distribution with regard to the variables being tested. We refer to this as a *within constraint within distribution conflict*. A second type of conflict may arise when the constraint being tested is the same in two cases, but the (marginal) distribution over the relevant variables is different as a result of different intervention sets in different experiments. A conflict occurs when the test results are not consistent with one generating structure appropriately manipulated for the two experiments.[2] We refer to this type of conflict as a *within constraint across distribution conflict*. It can only occur across two different experiments. Lastly, conflicts may arise as a result of tests on different constraints that cannot be combined consistently. This may occur within one or across several experiments. We refer to this type of conflict as a *within-* or *across experiment across constraint conflict*. In particular it may be the case that the constraints within one experiment can be combined consistently, i.e. that the constraints from each experiment are individually consistent with a (equivalence class of) generating graph(s), but that the sets of generating graphs

[2] In principle a conflict might occur under these circumstances even if the sampled data is exactly the same in both cases, since the different distributions might draw inconsistent conclusions with regard to the same constraint on the same data.

have an empty intersection, which implies an inconsistency across experiments.

We will not address conflicts here that arise within one experiment, since with regard this thesis these conflicts are handled entirely by the structure search algorithm within on experiment.[3] Different structure search algorithms might handle conflicts of this type differently. Most efficient constraint based algorithms do not consider all possible independence tests, but infer structure on the basis of the least possible number of independence tests. However, the cPC-algorithm distinguishes itself from the PC-algorithm in being particularly sensitive to conflicts arising from tests for colliders.

We will restrict the discussion here to combinatorial conflicts that occur when equivalence classes of graphs from different experiments are combined. If the combining algorithm is greedy, then – as in the case of structure search for a single experiment – conflicts may not occur as often since resolved pairs of variables are not reconsidered. But if accuracy of the output graph is the aim and additional information is available, being greedy might not be the best approach.

Let X and Y refer to some pair of variables in $\mathbf{V}$ and let $\mathcal{E}_i$ refer to some experiment in the sequence of experiments, different indices indicate different experiments. If only adjacencies are determined in each experiment (and directionality is inferred by combining experiments), as is the case in the combining algorithm for structural interventions (Algorithm 5.1.2), then a conflict occurs in the following situations:

1. X and Y are passively observed in $\mathcal{E}_k$ and $\mathcal{E}_l$, but $\mathcal{E}_k$ indicates they are non-adjacent, whereas $\mathcal{E}_l$ indicates they are adjacent.

2. X is subject to a structural intervention in $\mathcal{E}_k$ and found to be adjacent to Y (in fact one would conclude that it is a direct cause of Y), but in $\mathcal{E}_l$ both variables are passively observed and found to be non-adjacent.

3. X is subject to a structural intervention in $\mathcal{E}_k$ and found to be a direct cause of Y, and Y is subject to a structural intervention in $\mathcal{E}_l$ and found to be a direct cause of X (a conflict, since we assume acyclicity).

4. X and Y are passively observed in $\mathcal{E}_k$ and found to be adjacent, but X is subject to a structural intervention in $\mathcal{E}_l$ and not found to be adjacent to

[3] Given a sequence of experiments it may be of interest at a later stage to reconsider whether this approach can be improved, i.e. whether one can do better by considering constraints generated by a sequence of experiments all together, rather than for each experiment individually.

Y, and Y is subject to a structural intervention in $\mathcal{E}_m$ and not found to be adjacent to X.

There are many cases where orientations of edges can be determined within a single experiment as well:

1. If an adjacency is found between an (structurally) intervened variable and any other variable in the system, then the orientation of the edge is out of the intervened variable.

2. If three passively observed variables X, Y, Z form an unshielded collider in the post-manipulation graph then the collider can be discovered due to the particular independence relations it implies.

3. There are a few cases where the discovery of some edge orientations imply others (e.g. orienting away from an unshielded collider, orientation to avoid a violation of acyclicity etc.).

If these techniques to determine direction are included in the structure search algorithm for a particular experiment, as is the case for the adaptive search algorithm (Algorithm 5.2.1), a variety of further conflicts can arise pertaining to orientation information:

1. $X, Y \in \mathbf{U_i}$ and $X, Y \in \mathbf{U_j}$, $\mathcal{E}_i$ indicates that $X \rightarrow Y$ whereas $\mathcal{E}_j$ indicates $X \leftarrow Y$.

2. $\mathcal{E}_i$ indicates that $X \rightarrow Y$, but $\mathcal{E}_j$ indicates that the two variables are only adjacent *although* $\mathcal{E}_j$ should have discovered the orientation *if*, in fact, it is the case that $X \rightarrow Y$. This type of situation occurs if the directionality of the edge is determined by an unshielded collider and this unshielded collider would be present in both experiments $\mathcal{E}_i$ and $\mathcal{E}_j$.

3. If the directed edges from individual experiments combine to form a cycle which is not present in any of the individual results, then we have a further conflict (similar to the first bullet in this list, although several variables may be involved).

Conflicts may also arise from a failure of the basic assumptions, such as acyclicity, Markov or faithfulness, or a failure of causal sufficiency when it is assumed. In this case different solution approaches then the ones discussed here may be required. We will not address this point here.

6.2 Conflict Resolution

Under what circumstances is it possible to resolve the above conflicts? Of course one could re-run one of the experiments, possibly with a larger sample size, pool the data from the original and the repeated experiment and perform the crucial independence tests again, now with a larger sample. This is a simple but possibly expensive solution. One may also consider introducing some type of reliability hierarchy of tests. It is not clear how such a hierarchy might work. A simple approach may be that tests with smaller conditioning sets are more reliable. But this only provides a partial order over the tests and each of the conflicts can occur with conditioning sets of the same size in the conflicting tests. Something more sophisticated is needed.

Simple Pooling:

One cannot simply pool the data relevant to a particular independence test from two experiments, because different experiments in a sequence have different joint distributions over the variables resulting from the different interventions. If different variables are subject to interventions, this implies different manipulated graphs over the variables, representing the different joint distributions. Pooling data from different distributions may lead to spurious changes in correlations. The change can go either way. Depending on how data is pooled from experiments with different interventions, a spurious dependence may arise between variables that are in fact causally separated, or a spurious independence may arise between variables that are causally connected. Straight-forward pooling of data is therefore not a solution to conflicts that arise from statistical variation.

Voting:

One may attempt to divise some voting procedure to resolve conflicts without re-doing one of the experiments. For example, given an independence test on X and Y, select from a sequence of experiments those that do not simultaneously intervene on both X and Y, i.e. those experiments that are informative about X and Y. Among these experiments let a simple vote decide whether $X \rightarrow Y$, $X \leftarrow Y$ or X and Y are non-adjacent. Unfortunately it is not that simple: Although there are three possibilities for the true structure between a pair of variables, the tests are only binary. An independence test on a pair of variables, where one is subject to an intervention, can decide whether there is an edge from the intervened variable to the other variable, but cannot distinguish between an edge incident on the intervened variable and no edge at all. Similarly, an independence test where both variables are passively observed can tell whether there is an edge at all, but is unable to distinguish directions. It seems therefore, that the vote of a particular test should be evenly split between the options it cannot distinguish. That is, if after an intervention on X we find X and Y to be independent, then both non-adjacency and $X \leftarrow Y$ should receive half a vote each.

This approach does not yet take into account that votes from different experiments are votes from different joint distributions, which may make the discovery of a particular (in)dependence harder or easier. In order to reflect the significance of the result from any particular experiment, the decision could be a function of the p-values of the independence tests they represent, e.g. a threshold of the average p-value. But now we run into trouble with the asymmetry of the search procedure: In order to discover a *non-adjacency* we have to find *one* conditioning set that makes the two variables independent. The PC-algorithm iterates through the independence tests in order of complexity (size of the conditioning set), so that there always is a well-defined independence test that determines non-adjacency. The p-value from this test could be used to determine the weight of the vote from this particular experiment. However, *adjacency* is established if there is *no* conditioning set that makes the variables independent, i.e. *all* the independence tests fail. Consequently there is no unique priviledged p-value to contribute to the decision. Further, there is no guarantee that there is a corresponding independence test in each experiment so that one could reduce the conflict to a set of independence tests. And even if there were, then one would be aggregating p-values from different distributions. It is not clear what the justification for such a procedure would be.

Quite apart from the above matters, issues of judgment aggregation arise. Since the combination of independence relations imply other (in particular, higher order) independence relations, the outcome of a voting procedure depends on how votes are aggregated and it is not clear at all, how an aggregation procedure here would have to be designed to be in some sense “truth tracking”, i.e. that we could have any hope that using some voting method will get us closer to the true graph.[4]

[4] In [2] Bradley et al. suggest a method for aggregating causal graphs that conflict. However, their method is not truth tracking, nor is it sensitive to the incoherencies of the different causal graphs that are aggregated. It is not clear at all, what the aggregated graph represents

The bottom line is that voting may well work as a useful heuristic to resolve conflicts, but the worry is that the ad hoc decisions made in order to have a well-defined voting procedure destroy the consistency guarantees of the overall search algorithm.

6.2.1 A sufficient Condition for Conflict Resolution

The key difficulty in resolving conflicts is to figure out how and when the pooling of data from different joint distributions affects the independence tests relevant to a conflict. Failure to identify these cases can lead to spurious correlations or independencies when data is pooled. However, if one can ensure that the distribution *relevant* to the conditional independence test in question is the same in the conflicting experiments, then the data can be pooled to obtain an independence test with a larger sample size. A larger sample size can be used to increase the power, to lower the significance level of the test while maintaining the same power, or some improvement of both.

For example, suppose two variables X and Y, whose independence is in question, are graphically disconnected, i.e. causally separate, from the other variables $W_1, \dots, W_n$ in the causal structure. If there are two experiments, one which is an intervention on W_i and another with an intervention on W_j, then clearly the changes in the interventions will have no effect on the marginal distribution over X and Y and the data from the experiment can be pooled for the independence tests on X and Y. Causal separation is a very strong condition to ensure the validity of pooling, but we show that it can be weakened.

The basic idea is that we track the interventions that differ between the experiments whose data we want to pool for a particular independence test, and ensure that these "changing interventions" are screened off in each experiment from the variables whose independence we are testing. If that is the case, we can pool the data and perform an independence test with larger sample size.

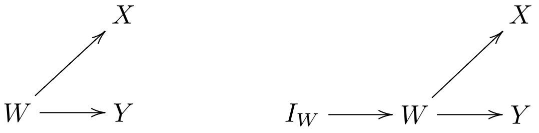
Consider two experiments $\mathcal{E}_i$ and $\mathcal{E}_j$, whose data we intend to pool for the conditional independence test T of $X, Y | \mathbf{C}$. Let $\mathbf{Pol_i}^*$ contain all the interven-

and why it should be considered informative about the true causal structure. All it offers is an aggregation of opinions about causal structure that respects certain desirable features of judgment aggregation. But these features (e.g. Pareto) are only preserved with regard to certain aspects of the causal structure (e.g. faithfulness is neglected). The solution they provide is unsatisfactory for causal models, which presumably are supposed to represent something true about the world. If judgments of causal relations are incosistent and one can perform tests to resolve the inconsistency before aggregation then surely they should not just be aggregated. Furthermore, it might be beneficial to consider the aggregation procedure at the constraint level as opposed to the level of causal structure.

tion variables in $\mathcal{E}_i$ that differ with respect to those in $\mathcal{E}_j$ and let $\mathbf{Pol_j}^*$ contain all the intervention variables in $\mathcal{E}_j$ that differ with respect to those in $\mathcal{E}_i$.[5] We then have the following theorem:

Theorem 6.2.1: Pooling under d-separation
If the set of variables $\{X, Y\}$ is d-separated from the set of changing intervention variables $\mathbf{Pol_i}^*$ given the conditioning set $\mathbf{C}$ in the manipulated graph of experiment $\mathcal{E}_i$ and if the set of variables $\{X, Y\}$ is d-separated from the set of changing intervention variables $\mathbf{Pol_j}^*$ given the conditioning set $\mathbf{C}$ in the manipulated graph of experiment $\mathcal{E}_j$, then the distributions relevant for independence test $T_{X,Y|C}$ are invariant across experiments $\mathcal{E}_i$ and $\mathcal{E}_j$ and the data relevant to the test can be pooled.

This theorem specifies a sufficient condition that allows for pooling of data. A simple example will illustrate the main claim. Suppose the left graph below specifies the true causal structure.



Consider two experiments, one a passive observation and one an intervention on W (right graph above). The theorem says that we cannot straightforwardly pool data for the (unconditional) test $T_{X,Y}$ (whether X is independent of Y), because I_W is not d-separated from $\{X, Y\}$ in the second experiment. But we can pool for the test $T_{X,Y|W}$ (whether X and Y are independent conditional on W), since the conditioning set $\{W\}$ d-separates the changing intervention I_W from $\{X, Y\}$. Consequently, if we had a conflict regarding the X, Y-connection, we could now at least perform a test $T_{X,Y|W}$ with larger sample size.

The theorem does not specify a *necessary* condition, since the intervention distributions in different experiments can be tweaked in such ways as to preserve the invariance properties of the distributions relevant to the independence test even if the d-separation condition is not satisfied. Trivially, this can be done if the intervention distribution of a variable is essentially the same as the passive observational distribution for that variable.

[5] See detail and proof in appendix.

6.2.2 Discussion of Conflict Resolution

While the above theorem specifies a sufficient condition for pooling which might resolve some conflicts in sequences of experiments, it requires substantial knowledge about the causal structure. Whether or not one can pool for a particular independence test depends on whether the variables subject to the test are d-separated from the changing interventions. But if one is trying to discover the causal structure, one will rarely know whether the relevant d-separation is satisfied. The theorem appears to be of little help in our circumstances.

The situation is not quite so bad: Given a (sub-)sequence of experiments, the theorem specifies for any possible causal structure invariance conditions for certain marginal and conditional distributions. These can be used. One can check the likelihood of the observed invariances over the sequence of experiments for any possible graph G over $\mathbf{V}$. Invariances that are implied but not observed or observed but not implied then reduce the set of possible graphs and resolve previous conflicts. This may turn out to be a computational nightmare, but at least progress can be made.

Such a likelihood based approach extracts from the Bayesian approach the key feature that resolves conflicts. For a Bayesian, conflicts of the type described above do not arise explicitly. A strictly Bayesian approach would place a prior over all possible structures and all possible parameterizations of those structures. Evidence from the experiments is integrated by an updating procedure involving the likelihood. Conflicts are thereby taken care of implicitly in the update. The computation is expensive and without obvious short-cuts. Priors are not going to be simple after the first experiment and it is not going to be sufficient to just keep track of the most likely graph(s). Nevertheless, it is another way of achieving a solution to the problem.

In very sparse graphs, or if only very few conflicts occur in a sequence of experiments, one may be able to determine directly whether the d-separation condition is satisfied in particular cases. However, even here, the theorem requires the search algorithm to store information about which independence test determined a non-adjacency in each experiment, so that the problematic test can be identified for possible conflict resolution later (or one has to find it again).

The theorem applies generally as it is not specific to particular families of distributions. It is therefore more generally relevant to techniques in meta-analysis. In particular, if the d-separation relation is known to hold, – say, the causal structure is known – then parameter estimates can be obtained by combining data from different experiments whose manipulated joint distributions are known to differ.

While the theorem provides a start of a solution to conflicts in some cases, the feasibility of its application is not entirely clear. In general, we consider the efficient resolution of conflicts one of the main open problems for search with sequences of experiments.

Chapter 7

Testing

Given the entire account so far we could test many different variations of set-up, assumptions and strategies in simulations. We can vary the background assumptions, the model space assumptions (discrete, continuous), causal sufficiency, the type of search strategy, the type of interventions, the number of interventions per experiment, the structure search algorithms that are used in each experiment, the parameters of the combining algorithms (greediness, resolution of orientation, etc.), the way conflicts are resolved, sample size and a whole slew of conditions pertaining to the graphs, e.g. number of variables, sparsity, parameterization etc. There is no way we would have been able to consider a set of simulations that could be deemed representative of the space these dimensions span. However, we have set up a simulation suite building on the TETRAD program[1] that can be used to test a large variety of combinations of the above assumptions and algorithms. We have only explored a relatively small part of the space and we show results here of a representative subset of that small part.

In [8] we simulated the fixed search strategies with single and multiple simultaneous interventions (Strategies 3.3.2 and 3.3.5). We performed $N-1$ or $\log_2(N)+1$ experiments, respectively, on sets with different numbers of (causally sufficient) variables, with true graphs of different sparsity, under the assumption that the causal model was a discrete binary model. Orientation information was only determined by the combining algorithm given the different experimental set-ups, i.e. the structure search algorithms in each experiment only returned

[1] http://www.phil.cmu.edu/projects/tetrad/

adjacency information. If conflicts occurred, no effort was made to resolve the conflict and an error was output for the conflicted pair of variables. Accuracy of the search procedure was measured in terms of the percentage of pairs of variables whose direct connection or non-connection was determined exactly correctly. The direct connection between a pair of variables X and Y is discovered exactly correctly if (i) $X \rightarrow Y$ in G_{true} if and only if $X \rightarrow Y$ in G_{output}; (ii) $X \leftarrow Y$ in G_{true} if and only if $X \leftarrow Y$ in G_{output}; and (iii) $X \quad Y$ in G_{true} if and only if $X \quad Y$ in G_{output}. We used the PC-algorithm to determined adjacency in each experiment, and we varied across sequences of experiments the sample *per experiment* from 100 to 10,000 samples. If a fixed sequence of experiments did not complete the discovery within the allocated number of experiments, unresolved edges were determined to be errors.

Apart from the more intuitive and obvious connections between sample size and accuracy, the interesting result of that simulation study was that the fixed strategy using multiple simultaneous interventions per experiment was much more accurate than its single intervention counterpart. That is, in fewer ($\log(N) + 1$ as opposed to $N - 1$) experiments, and consequently with much fewer samples (since samples were allocated at a fixed rate per experiment), these strategies provided more accurate results. This difference, however, disappeared when the analysis was restricted to outputs that did not contain any conflicted pairs of variables. We took those studies to provide some support for the idea that single intervention search strategies create many conflicts between experiments that jeopardize the output accuracy.[2] The conclusion is, of course, *not* that single intervention search strategies are necessarily less accurate even with larger sample sizes. It may well be the case that once a good way has been found to resolve conflicted pairs of variables, the accuracy of single intervention strategies will be boosted to the level of multiple intervention strategies. The need for a good conflict resolution technique was a general conclusion of that study. Among the searches for which no conflicts occurred, accuracy of output was around 90% (exactly correctly resolved pairs of variables) at only 1,000 samples per experiment.

[2] The intuition is that since each experiment considers the causal relations of all pairs of variables, unreliable conclusions are drawn (about, for example, pairs of variables not closely connected to the intervened variable). A single intervention experiment on a set of variables makes, so to speak, claims about causal connections it does not really know much about.

7.1 Simulations

In the simulations presented here we focus exclusively on mixed strategies with structural interventions. This is the most plausible approach one would use in an actual search for causal structure if there are no restrictions on which variables can be subject to intervention. Furthermore, mixed strategies provide the most interesting case since they imply the tightest theoretical bounds and they generate – in contrast to fixed strategies – a unique output: A mixed strategy stops its search when it has discovered a directed acyclic graph among the set of variables under consideration. For a fixed strategy, the set of experiments is fixed a priori and so the output might still contain pairs of variables whose connection is not yet resolved. (Adaptive strategies share this early stopping aspect with mixed strategies, but have none of the other features.)

We do not consider parametric interventions since the search for structure using parametric interventions is very similar to structure search in passive observational data once the set of variables has been augumented by the intervention variables and their edges into the intervened variables. The main issue with parametric interventions is the reliability of collider tests, for which the cPC-algorithm provides some additional guarantees.

Throughout the simulations, all the following assumptions are made:

1. Causal Markov
2. Causal Faithfulness
3. Acyclicity of true causal structure
4. Causal Sufficiency of the set of variables[3]
5. Interventions are structural, uncaused and not confounding

In addition, a whole list of pragmatic decisions had to be made to get the simulations up and running:

We used different algorithms for structure search in each experiment. The structure search algorithm in each experiment uses a knowledge graph for that experiment. The knowledge graph includes knowledge about the experimental set-up: No-knowledge edges are placed between variables that have both been subject to an intervention, directed edges are placed where an adjacency is found

[3] Simulations on causally insufficient sets of variables, or where other background assumptions fail, would be most interesting, but some of us would like to graduate first.

between an intervened and a non-intervened variable, and semi-directed edges into the intervened variable are placed if an intervened and a non-intervened variable are found to be non-adjacent. All other edges are determined by the structure search algorithm. Some of the algorithms return bi-directed edges (representing independence constraints indicative of latent variables). Since we are only considering causally sufficient sets of variables, these bi-directed edges must result from a statistical test being subject to error. These edges are treated as conflicts within one experiment and are removed and replaced by no-knowledge edges.

No matter what information the structure search algorithm returns in the knowledge graph, the algorithm combining knowledge graphs from different experiments (Algorithm 5.2.1) is greedy, i.e. it only consider edges that have not been resolved by previous experiments. However, for any unresolved edge, the combining algorithm will use all *orientation* information that a structure search algorithm discovers in a single experiment. We thus avoid, by not reconsidering a resolved edge, conflicts for edges that are already resolved, but that would, if reconsidered in light of the current experiment, be considered conflicted.[4] On the other hand, conflicts may arise that would have been avoided if we only had taken *adjacency* information from each experiment into account.

Conflicts between experiments are resolved by...not resolving them: The conflicted edge in the knowledge graph is replaced with a no-knowledge edge, thereby voiding the conflict and trying again. Subsequently, only findings from experiments performed after occurrence of the conflict are considered for the resolution of this edge. If the conflict is due to a cycle that would be created if an edge were added, then *all* edges in the *smallest* cycle that would be created are replaced with no-knowledge edges and are reconsidered only in terms of the experiments that follow.[5] Needless to say that this form of conflict resolution is not really a form of resolution but rather a form of restarting the search with regard to a subset of the edges. All the information from previous experiments regarding the conflicted edges is wasted. It was a decision that was made for

[4] The only exception to this is when a resolved edge forms part of a potential cycle. Then all edges of the cycle are removed.

[5] For example, if the knowledge graph contains a path $X \rightarrow Y \rightarrow Z$ and also a path $X \rightarrow W \rightarrow Y$, and the latest experiment suggests adding an edge $Z \rightarrow X$, then that would create two cycles, $\{X, Y, Z\}, \{X, W, Y, Z\}$, which would be recognized as a conflict. The suggested resolution here is that all edges in the $\{X, Y, Z\}$-cycle are removed and replaced with no-knowledge edges in the knowledge graph, while the XWY-path is left intact. There is no more principled reason for this resolution other than that it is a minimal change that resolves the conflict without picking out one individual edge.

computational efficiency, since no computationally feasible solution for a more principled conflict resolution was available.

As the discussion of mixed strategies using multiple simultaneous interventions indicated, there is no obvious choice for the size of the intervention set of the first experiment. One does not want to include too many variables, since then information about causal structure among the intervened variables would be lost. But one also does not want to intervene on too few (at least, if one is – as we are here – concerned with minimizing the number of experiments), in case the true graph is very dense. In Section 4.2.2 we listed the optimal sizes for the first intervention set when (a) the true graph is complete and it is known that the true graph is complete, and (b) the true graph is complete but it is not known whether the true graph is complete. Both cases were subject to the assumption that one had an independence oracle. In the following simulations of the mixed strategy with multiple simultaneous interventions graphs are not complete, nor do we have an independence oracle. We do not know what the optimal size for the first intervention set is under these circumstances, but we use the intervention set sizes we established for the (b) case. We do not yet have a good sense of the impact of this choice.

Directed acyclic graphs over N variables are sampled randomly by using the Tetrad-implementation of an MCMC-algorithm described in [27]. The parametrization is also determined by the default Tetrad implementations, which in the case of linear models bounds the edge-coefficients away from zero. The PC- and the cPC-algorithm use a χ^2-test in the discrete case and Fisher's Z-test in the linear case with a significance level of $\alpha = 0.05$. The GES-algorithm is initialized with a flat prior over structures. We performed simulations for each combination of items from the following categories:

Causal Model: discrete (binary) or linear normal

Structure Search Algorithm in each Experiment: PC-Algorithm, conservative PC-Algorithm or GES-algorithm

Interventions: Single or Multiple simultaneous interventions

Sample Size per Experiment: 100, 1,000 or 10,000 samples in each experiment in a sequence

Number of Variables: The number of variables is varied between 4 and 10.

Graph Density: Graph density was varied by changing the maximum degree of any node in the graph.

Number of Nodes	Maximum Node Degrees Considered
4	$\{3\}$
5	$\{3, 4\}$
6	$\{3, 5\}$
7	$\{3, 6\}$
8	$\{3, 5, 7\}$
9	$\{3, 5, 8\}$
10	$\{3, 6, 9\}$

We allocate a fixed sample size per experiment, independent of the number of experiments in a sequence. In general, one could use many other ways of allocating samples. Our choice implies that if the sequence of experiments is longer, then the result graph is based on a larger overall sample.[6]

Graph density can also be determined by several other measures. We use maximum node-degree, but one could also restrict the degree to in- or out-degree, or one could limit the total number of edges in the graph. Maximum node-degree seemed to work well in our case and is reasonably closely related to the total number of edges in the graph (see discussion on graph density in [8]). In summary, then, we have the following simulations:

Overview:

```
For a discrete binary or a linear normal causal model,
using one of the three search algorithms (call it SA)
using a single or multiple intervention search strategy
(call it Strat), do:
For the number N of nodes, from N = 4 to N=10{
  For each maximum node degree d for N{
    For each sample size in (100, 1000, 10000){
      For 100 iterations{
        Sample a directed acyclic graph G uniformly from DAGs with N
        variables and max degree d and parameterize it with a random
        parameterization that is Markov and faithful to the causal model.
        Initialize strategy Strat for the graph.
```

[6] One could also fix the overall sample, but would then have to decide how it is divided up among experiments in a sequence that varies in length. Or one could sample variables selectively.

```
10.        Perform experiment E determined by OPTINTER for
           mixed strategies.
11.        Use SA supplemented with knowledge about the experimental
           set-up to determine the knowledge graph for E.
12.        Use Adaptive Combining Algorithm for structural interventions
           to determine knowledge graph K for the sequence of experiments
           performed so far.
13.        If there is a conflict, replace the conflicted (or cycle-)
           edges with no-knowledge edges and note which experiments
           may not be considered for the resolution of the conflicted edges.
14.        If K is not a resolved acyclic causal structure, return to
           10., otherwise output K.
15.        Measure the accuracy of K compared to G.
    } } } }
```

We measure the accuracy of the output graph G_{output} in two different ways: First, we count the number of times out of the 100 iterations that the search algorithm discovered the true graph G_{true} *exactly* correctly. That is, there are no false positive and no false negative edges and no incorrect edge-endpoints in the output; the output is – in qualitative terms – perfect. Second, we measure the percentage of pairs of variables whose direct connection (or non-connection) is discovered exactly correctly. We introduced this second measure because we found that for larger graphs it was very rare that the graph could be recovered exactly correctly with *any* of the search algorithms and strategies we considered, despite the fact that a decent portion of the graph had been recovered correctly. This was a pragmatic decision to render a simple intuitive qualitative measures of accuracy. We discuss some alternatives in the section on future research.

7.2 Results

We do not present the full output, but representative cases:

1. Figure 7.1 shows simulation results, when the true graph is a linear normal causal model over 4 variables with maximum node-degree equal to 3 (i.e. all DAGs over 4 variables). The graph shows the average number of experiments that were required to recover a DAG, for each of the three search algorithms combined with a mixed strategy using either single or

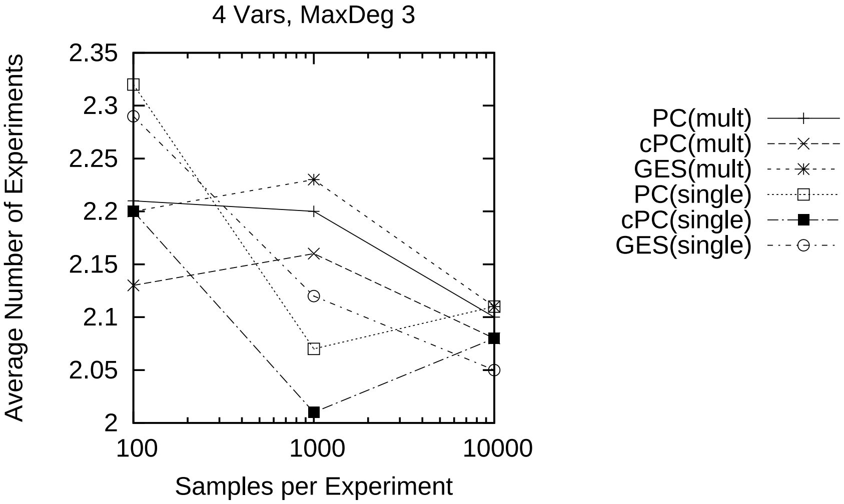
multiple interventions per experiment. The second graph shows the number (out of 100) of graphs that were determined exactly correctly for each of the different search strategies.

2. Figure 7.2 shows the same results as the previous figure ($N = 4, maxDegree = 3$), just for discrete binary models.

3. Figure 7.3 shows the same type of results for linear normal models over 6 variables with maximum node-degree of 5 (i.e. all DAGs over 6 variables).

4. Figure 7.4 shows the same results as the previous figure ($N = 6, maxDegree = 5$), just for discrete binary models.

5. Figure 7.5 , Figure 7.6 and Figure 7.7 show results for linear normal models over 10 variables. However, here we vary the graph density across the figures. For Figure 7.5 , we have ($N = 10, maxDegree = 3$), i.e. sparse graphs only, in Figure 7.6 , we have ($N = 10, maxDegree = 6$), i.e. sparse and medium dense graphs only, and in Figure 7.7 , we have ($N = 10, maxDegree = 9$), i.e. all graphs over 10 variables. Again we show the average number of experiments for the six types of searches (three structure search algorithms crossed with two strategy types: single and multiple interventions per experiment). Instead of a count of exactly correct graphs, we show the average (over 100 iterations) of the percentage of pairs of nodes (out of $\binom{10}{2} = 45$) that were resolved exactly correctly.



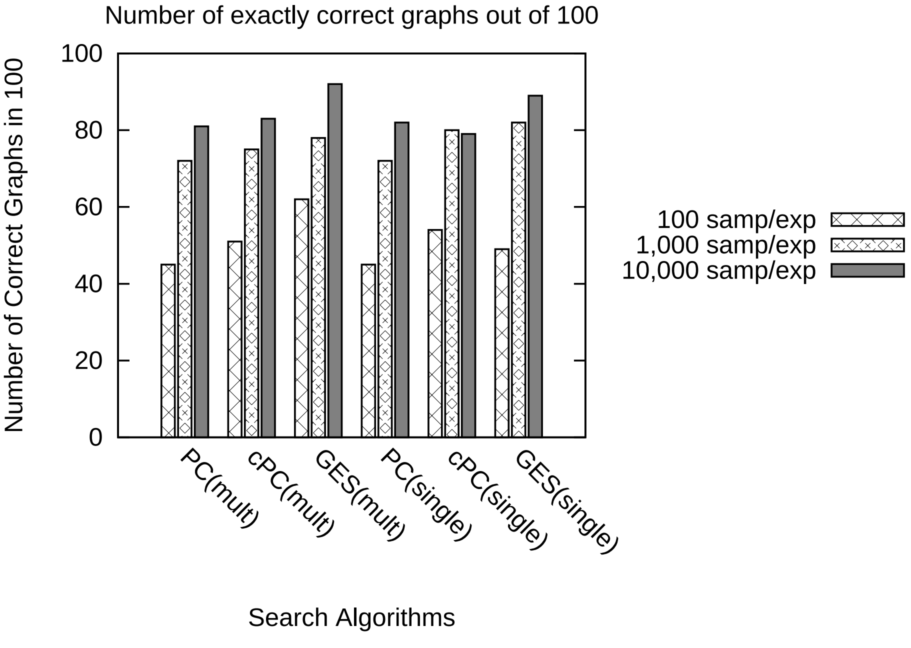
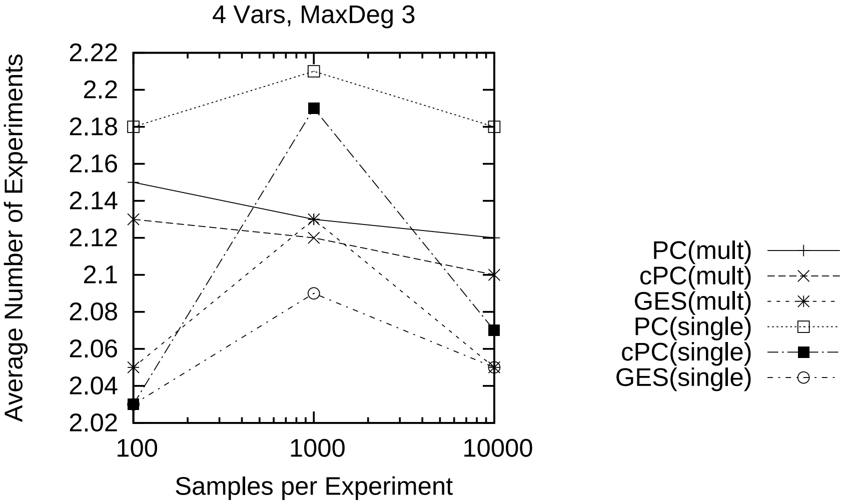


Figure 7.1: ***Linear Normal Models (Var 4, MaxDeg 3):*** *Graphs over 4 variables with maximum node-degree 3. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Number (out of 100) of graphs that were discovered exactly correctly, for different seach strategies and sample sizes (bottom).*



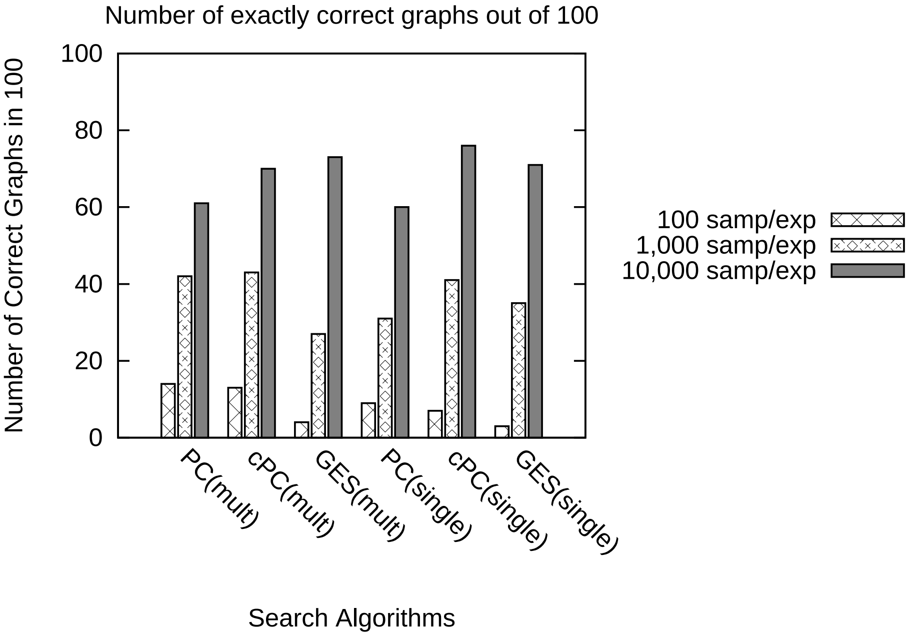
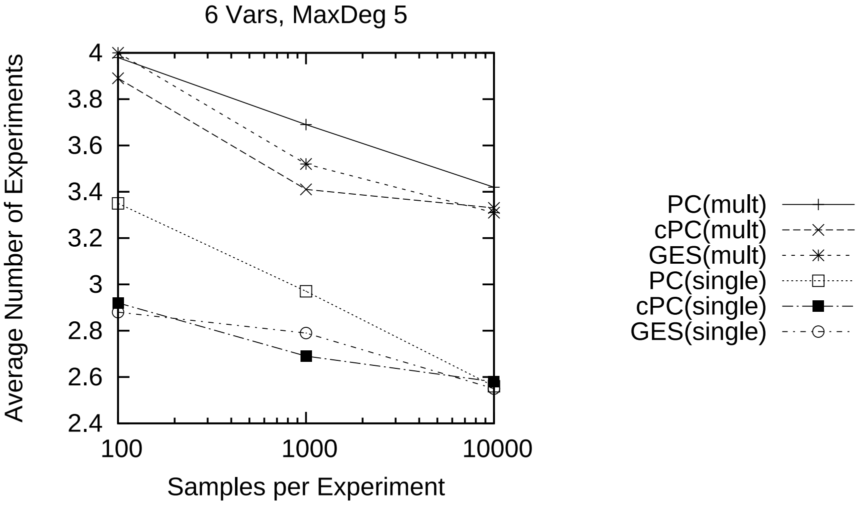


Figure 7.2: ***Discrete Binary Models (Var 4, MaxDeg 3):*** *Graphs over 4 variables with maximum node-degree 3. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Number (out of 100) of graphs that were discovered exactly correctly, for different seach strategies and sample sizes (bottom).*



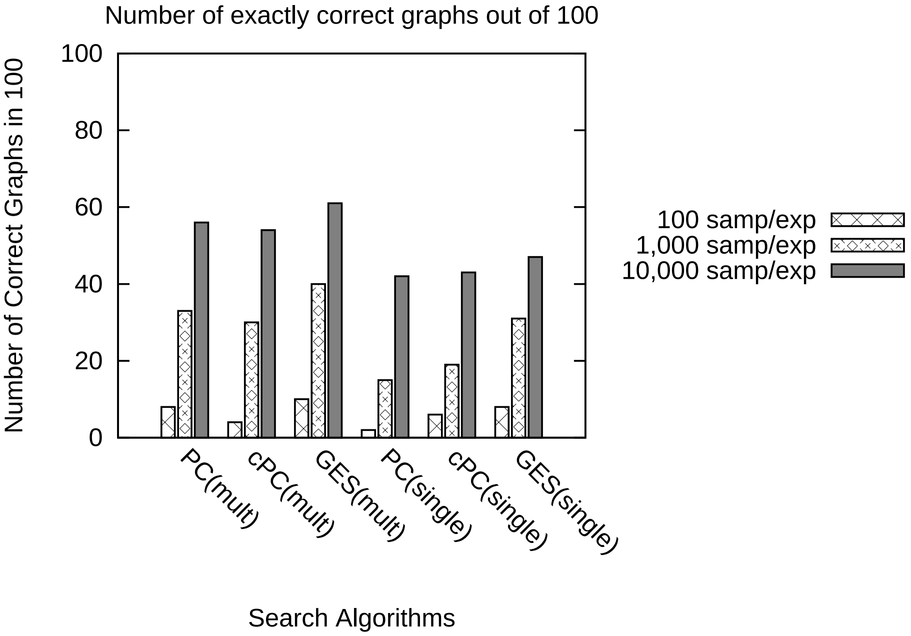
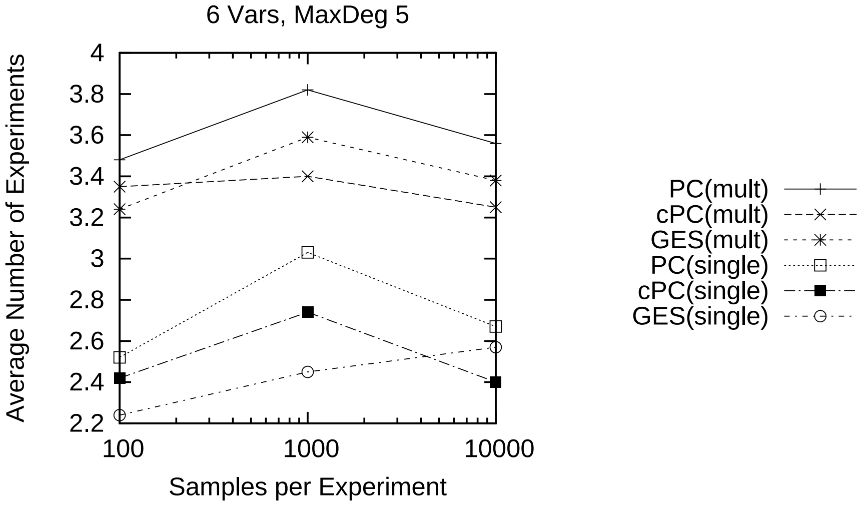


Figure 7.3: ***Linear Normal Models (Var 6, MaxDeg 5):*** *Graphs over 6 variables with maximum node-degree 5. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Number (out of 100) of graphs that were discovered exactly correctly, for different seach strategies and sample sizes (bottom).*



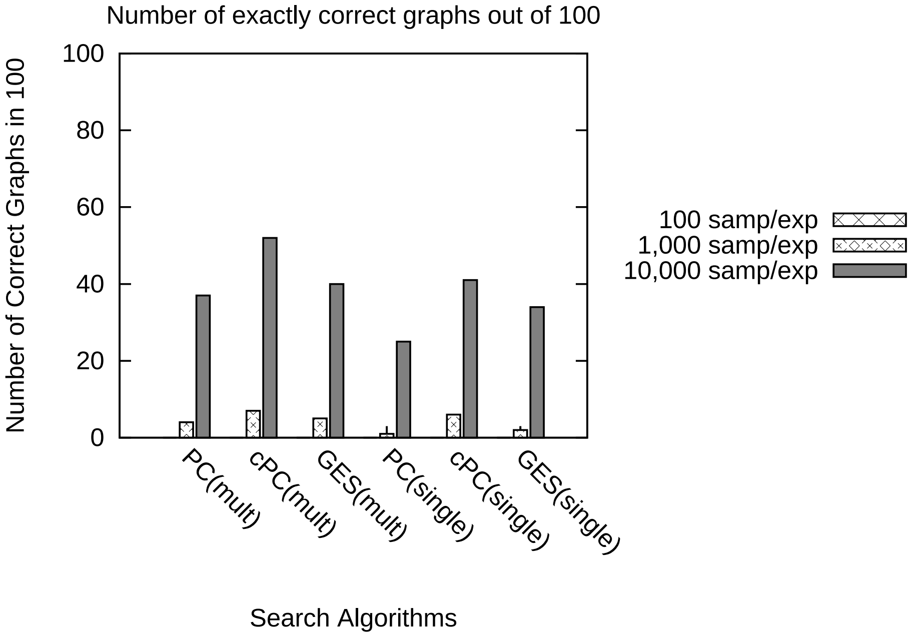


Figure 7.4: ***Discrete Binary Models (Var 6, MaxDeg 5):*** *Graphs over 6 variables with maximum node-degree 5. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Number (out of 100) of graphs that were discovered exactly correctly, for different seach strategies and sample sizes (bottom).*

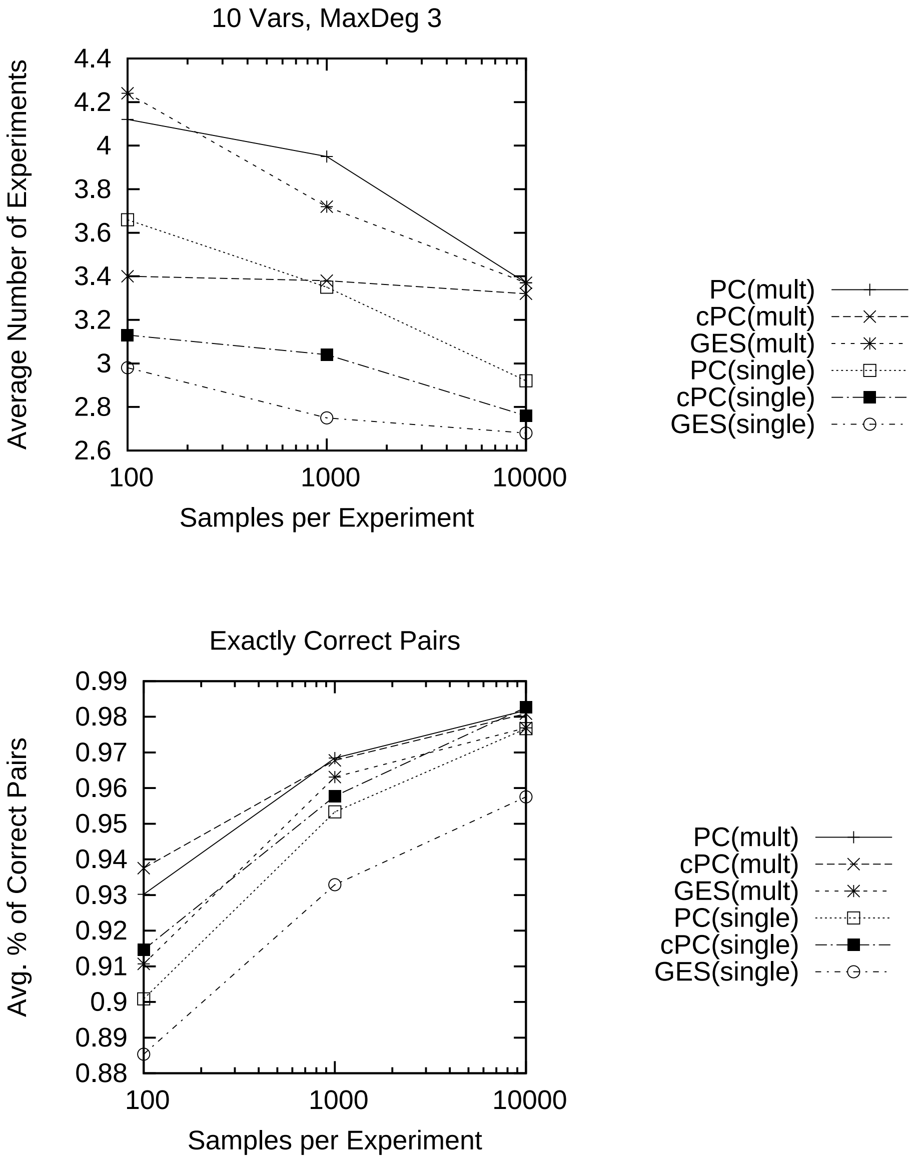


Figure 7.5: ***Linear Normal Models (Var 10, MaxDeg 3):*** *Graphs over 10 variables with maximum node-degree 3. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Percentage of pairs of nodes (of 45 possible pairs), whose direct (or non-)connection was discovered exactly correctly, for different seach strategies and sample sizes (bottom).*

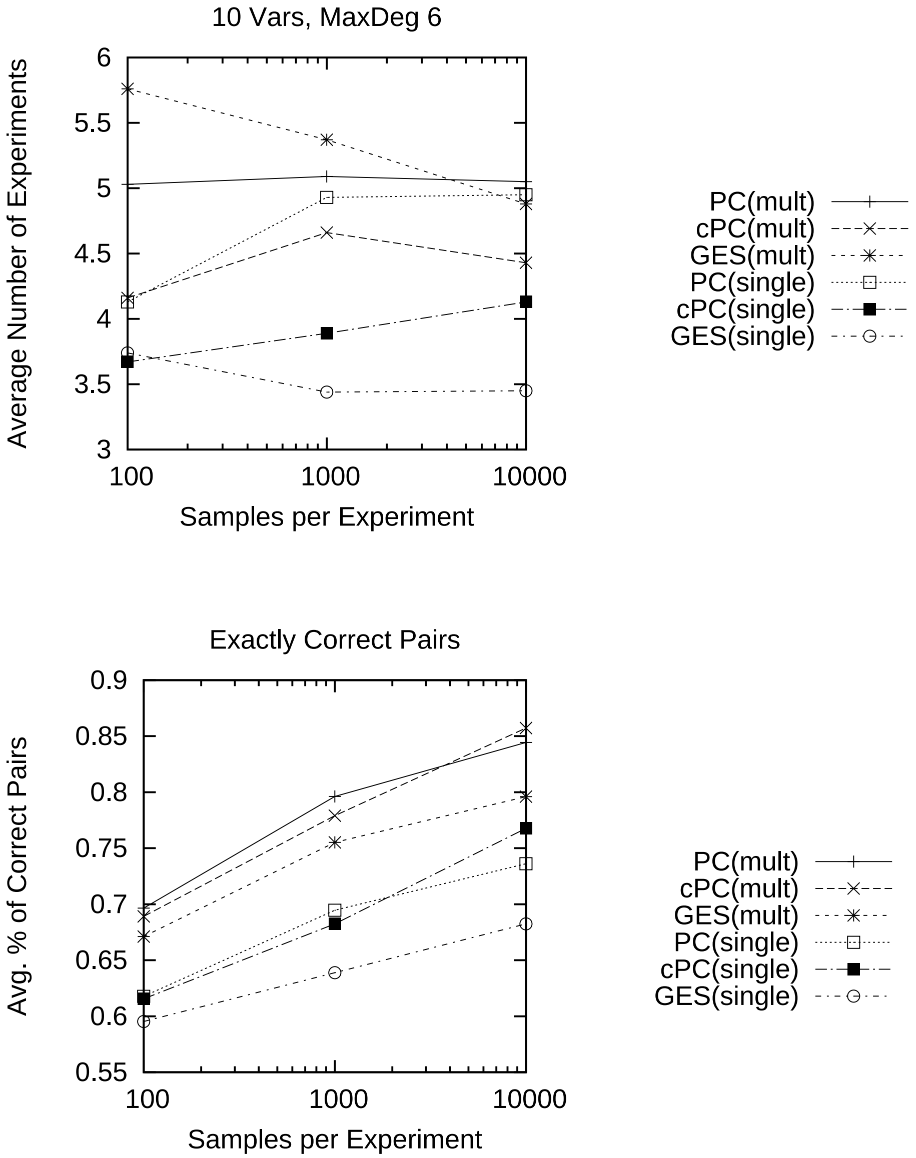
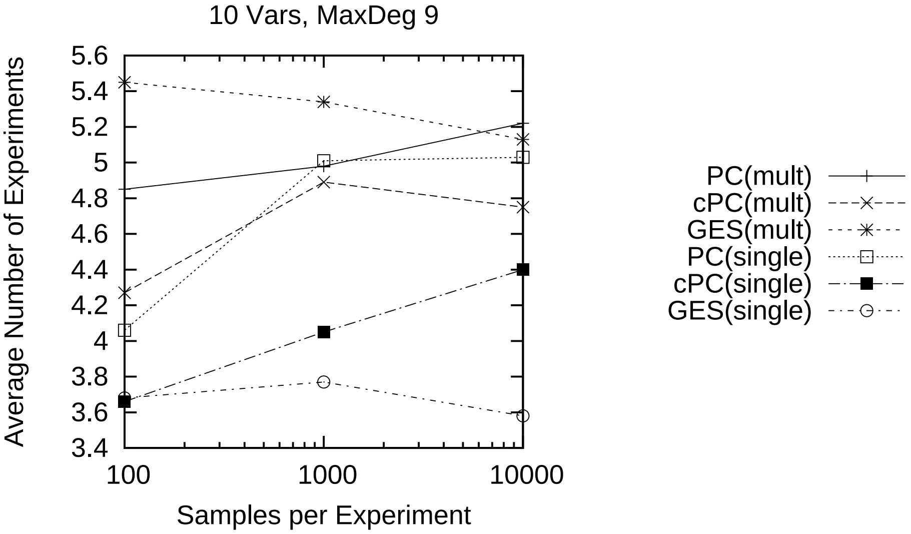


Figure 7.6: ***Linear Normal Models (Var 10, MaxDeg 6):*** *Graphs over 10 variables with maximum node-degree 6. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Percentage of pairs of nodes (of 45 possible pairs), whose direct (or non-)connection was discovered exactly correctly, for different seach strategies and sample sizes (bottom).*



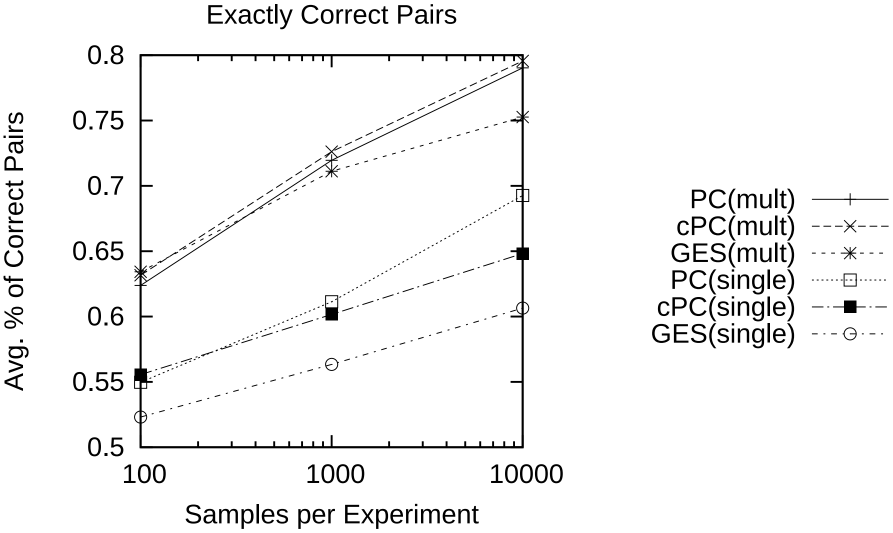


Figure 7.7: ***Linear Normal Models (Var 10, MaxDeg 9):*** *Graphs over 10 variables with maximum node-degree 9. Average number of experiments (over 100 iterations) to discover a causal graph for single and multiple interventions per experiment with the PC-, cPC- or GES-algorithm for structure search (top). Percentage of pairs of nodes (of 45 possible pairs), whose direct (or non-)connection was discovered exactly correctly, for different seach strategies and sample sizes (bottom).*

The plots do not include error bars to preserve clarity. Assuming that 100 iterations are sufficient to make the distribution of the averages more or less normal, we found that for the linear normal models a 95% confidence interval of the average number of experiments is a band of between ±0.2 and ±0.3 units around the average values shown in the plots. The width of the band was largely independent of the structure search algorithm or search strategy used. But, as sample size per experiment went up, the confidence interval got slightly smaller (unsurprisingly). As the graph density increased (greater maximum node-degree), the confidence interval also increased slightly. A 95% confidence interval for the average number of exactly correct edges (shown here only for $N = 10$) covers a band of between ±1% and ±2% around the mean values in the plots. The confidence interval is smaller for sparse graphs.

The results shown in the graphs are representative of the other values we considered for the simulations. Many of the results are unsurprising: The accuracy, in the two measures we considered, goes up as the sample size per experiment increases. The accuracy decreases as graph density and number of variables increases. At least for many of the linear normal models, the number of experiments descreases as the sample size per experiment increases. The accuracy with linear normal models is much higher than with discrete binary models.

The more interesting results are in the contrast of single and multiple intervention search strategies. While search strategies that employ single interventions per experiment return a causal structure in about the same or less number of experiments (depending on the structure search algorithm) than their multiple intervention counterparts, the causal structure they return has many more errors: Up to 10% more pairs of variables are resolved exactly correctly (or up to 10% more graphs are resolved exactly correctly) with the multiple intervention strategies. This can be seen most vividly for larger numbers of variables, dense graphs and linear normal models (Figures 7.3, 7.6 and 7.7). 7.2). This provides an interesting extension of the results we found in the simulation study on fixed strategies [8], discussed above. There we found the multiple simultaneous intervention strategies to be much more accurate, even though their overall sample size (in the sequence of experiments) was much smaller than the single intervention case (due to only $\log_2(N) + 1$ experiments instead of $N - 1$). Here, with mixed strategies, we find the opposite: Single intervention mixed strategies find a causal structure in the same number of or *fewer* experiments than their multiple intervention counterparts, but the causal structure has more errors. Since a mixed strategy can return after any number of experiments and need not wait for some fixed length sequence of experiments to complete, the single intervention mixed strategy returns too early (presumably also due to the greedy search) with an incorrect graph. A mixed strategy with multiple simultaneous interventions performs slightly longer sequences of experiments, but then appears to have fewer errors in the output. While it would be far too strong to draw a general conclusion, we do think this is evidence for a point indicated already in the discussion of the simulation study in [8]: Search strategies using multiple simultaneous interventions seem to strike a better balance between output accuracy and search investment (in the sense of sample size or number of experiments).[7] This suggested result is not the same as a perhaps more well known, or to some people, more obvious conclusion that in any one experiment, if we are just interested in the effect of a set of treatment variables on a set of outcome variables, then accuracy of the result can be increased by intervening on other variables as well (e.g. by clamping them to a particular value). The proposal we are making is that with regard to learning about the *entire structure* among a set of variables, sequences of experiments with *multiple* interventions, are better in output accuracy than those with single interventions.

In all the previous chapters our results concerned some kind of worst case analysis: absolute worst case or average worst case. However, we did not consider statistical variability. Here in the simulations we consider a much broader class of graphs than the worst case. In fact, when we bound above the maximum node-degree by some value smaller than $N-1$, where N is the number of variables in the graph, we are explicitly bounding ourselves away from the worst case. Instead of the worst case graphs, statistical errors now cause problems. It turns out that in the grand scheme of things, the number of experiments the mixed search strategies perform in the simulations are in the same ball-park of the worst case bound for a given number of variables and a given graph density, no matter what the structure search algorithm is in each experiment and no matter whether the true model is discrete or continuous. Furthermore, $\log(N)$, where N is the number of variables, seems to be a decent rough guide of how many experiments are required. The point here is again not a formal one, but a suggestion that despite their worst case nature, the bounds, especially the

[7] Please note the caution in the statement. No-one said that there was a proof for this statement in this thesis. No-one said that the graphs show this result without any doubt. Nor did anyone say that there cannot be other reasons why this difference came about. It is, to use someone else's infamous term, a little corroboration.

tighter ones using multiple simultaneous interventions, give an idea of what to expect from a sequence of experiments.

There are many other points these simulations raise, but which we do not have a good explanation for or where we lack confidence in the results of the simulation. Much more careful analysis needs to be done to tease out the differences resulting from the structure search algorithms. Much more analysis could be done on the nature and occurrence of conflicts and how best orientation information should be determined. We have not done more analysis here, since we believe that there are many more theoretical issues to sort out before a proper practical analysis is worth the time and effort. At this point there are still too many ad hoc choices in the implementation of the algorithms that, if they had been made otherwise, might yield different results. As a reminder, a short list of the most important concerns: the size of the first experiment, the different options to resolve orientations of edges, resolution of conflicted edges and cycles, greedy search or reconsideration of already resolved edges in later experiments and the allocation of samples to the sequence of experiments.

Consequently, it is not clear how representative the simulations are of how the theoretical results spell out in practise. Many of the results from earlier chapters are possibility or impossibility results and as such, are not practical benchmarks. The implemented algorithms are not tuned to optimize the computation involved and several decisions were taken to make the test of a reasonable space of graphs feasible. As a result of these decisions, the algorithms are at times wasteful with information that is or may be available in the data. Nevertheless, we think the simulations show some broad differences between approaches and are indicative of the additional problems that arise when sample variability is taken into account. So in summary, we recommend the results of this section with a grain of salt and a PG rating: parental guidance suggested.

Chapter 8

Conclusion

This thesis project set out to analyze causal discovery using interventions. The aim was to look beyond a single experiment to sequences of experiments and provide guidelines of how sequences of experiments should be performed. How should intervention sets be chosen if the aim is to discover the entire causal network among a potentially large set of variables? Under what circumstances can such discovery be successful in one experiment and what should be done if a single experiment is insufficient?

The guidelines we provided in this thesis are worst case bounds on the number of experiments sufficient (and sometimes necessary) to discover the causal structure. We provide bounds under a variety of different model space assumptions and for several search strategies and types of interventions. For each bound we supply at least one search strategy that guarantees (given the appropriate oracle) discovery of the causal structure within that bound, and we provide algorithms that compute the intervention sets and combine results from different experiments to yield the search result. The bounds apply to the worst case or worst case expectation, but the adaptation to more benevolent situations is obvious and the adaptive and mixed strategies are sensitive to such cases. In some cases, as with the algorithms based on the differences in correlation, the algorithms push the limits of what can be learned into the space of unmeasured variables and connections are indicated to algorithms for structure search among latent variables. We framed the entire discussion of search procedures in a language that has a straightforward game-theoretic interpretation and consequently, many of the discovery problems can be understood as searches for solutions of a particular game. Such a game-theoretic framework also supplies a much more general approach to cost of discovery, which here has almost exclusively been considered in terms of the number of experiments. Thus, the guidelines we provide are not heuristics of what should be done in particular circumstances, but guidelines to what is possible given a set of circumstances and assumptions.

The bounds on the number of experiments derive from the combinatorics of experiments. Since the bounds are based on oracles that report the appropriate independence or correlation information in each experiment, the bounds are insensitive to structure search algorithms that have the same asymptotic properties. That is, one can plug any of a variety of constraint or score based algorithms into the structure search stage for any one experiment. As long as the algorithms have the same asymptotic properties, such as discovery of the Markov equivalence class, the same bound applies. Of course, the convergence rate may not be the same. Certain structure search algorithms might be better or worse on particular structures or for particular sample sizes and so any actual sequence of experiments may be shorter or longer or return more or less accurate results for different algorithms. For algorithms with different asymptotic properties, different bounds apply. For example, ICA-based procedures can, under a non-normality assumption recover the exact causal structure without performing a single intervention, while the PC-algorithm is limited to the Markov equivalence class of causal structures. By contrasting algorithms based on independence tests with those based on differences in correlation, we showed how the discovery bounds change under otherwise the same assumptions.

In the simulations and discussion of conflicts in sequences of experiments we dipped into issues relating to statistical variability. This can only be seen as the snowflake on the tip of an iceberg. Quite apart from convergence rates of different search procedures, the entire discovery problem is ultimately much more general than we have presented it here, just in terms of the right sequence of experiments. The optimization ranges over the cost of different experiments, the allocation of samples in experiments and which variables to measure in which experiments. So perhaps this thesis raises more questions than it answers. But then, after all, it was a thesis submitted in a philosophy department.

We leave the reader with a short list of some of the issues and ideas we consider important or interesting, but are still underdeveloped and left for future research, which – as is usual in these cases – may never be done, or turn out quite differently.

8.1 Future Research

The closer the thesis has come to completion the more unfinished it appeared to become. Several ideas that were originally going to be considered more thoroughly have not even been touched properly yet, because there was plenty of work to be done before. In the meantime many new ideas and problems have arisen and so we summarize just a few. All of the following sections are to be taken as indicators of future plans only. We have not had the time to consider any of them more carefully and they are – more so than the previous chapters – prone to errors and misguided thoughts.

8.1.1 Measure of Accuracy for Causal Structure

In our simulations we measured the accuracy of the output of the search procedure in terms of the number of exactly correctly discovered graphs in 100 trials or the average percentage of exactly correctly resolved pairs of variables in each graph in 100 trials. These are straightforward qualitative measures of the correctness of the output. They are coarse functions of more detailed qualitative accuracy measures that report false positive and false negative edges and/or edge endpoints. On these qualitative measures, each edge is treated equally. However, unless the score is perfect, the score may not reflect what one cares about in discovering the causal structure. A highly-accurate output may not correspond to a highly-useful tool for policy making.

Generally one cares about causal structure in cases of prediction under intervention and for the estimation of counterfactual states. We have focused on discovery of the entire causal structure among a set of variables, not just on particular connections. Knowledge of the entire causal structures might be of interest as a compact representation of knowledge, in some sense like a statistic, when the particular prediction or estimation task is not known in advance: Ship the causal graph to the policy maker when it is not known what policy decisions the policy maker will be considering. Ship the causal graph to the Mars Rover if you do not know what it might bump into. In these cases the aim is to provide with the causal graph the best guide for predictions under interventions, when the prediction query is not known. Consequently, the relevant measure of accuracy of the causal graph should account for the result of possible interventions.

Qualitative Measures of Accuracy

Given that the discussion within this thesis has concentrated for the most part on the discovery of the causal structure, the measure of accuracy should reflect the accuracy of this qualitative aspect of the causal model under interventions, i.e. the accuracy of the post-manipulation graphs for different interventions. One way to think qualitatively about the accuracy of a prediction of the effect of an intervention is whether the post manipulation graph correctly represents all and only the paths connecting intervened variables to their – direct or indirect – effects.[1] Intuitively, what we care about in a prediction of the effect of an intervention is that we know how the intervention affects other variables. Pathways constitute a qualitative component of this knowledge.

Obviously, the primitive measures of accuracy we used in the simulations are not independent of this proposed measure of accuracy of the post-manipulation graphs, but they are not fully indicative of it either. Missed edges that are part of many pathways in the true graph weigh more heavily in the error suggested here than missed edges that only feature in a few paths between variables. Each missed edge is treated equally in the measure we used in the main part of the thesis. Given these considerations, the qualitative measure of output accuracy should ideally measure the number of false positive and false negative paths between intervened and non-intervened variables given by the 2^N post-manipulation graphs resulting from all possible interventions on N variables (interventions on all possible subsets of N variables) in the output graph.

Consideration of 2^N post-manipulation graphs obviously makes this effort untenable very quickly. Even restricting ourselves to N single variable interventions and determining for each of the non-intervened variables the pathways to the intervened variable in the appropriate post-manipulation graph is a significant task if one intends to consider decent sized simulations.[2]

[1] Of course, there are other ways to consider the qualitative effect of an intervention. One could consider the accuracy of the separation into effects and non-effects of the intervened variables, one could restrict consideration to direct cause-effect relations only or to the independencies created by the intervention (if it is structural), or one of many other ways. If the aim is to ensure that the discovered causal graph corresponds to the true causal graph then the measure of accuracy should not incentivize search algorithms that output graphs with excellent scores, but that do not represent the true causal relations. It is an open question whether the suggested qualitative measure entails such incentives.

[2] There are ways to record pathways dynamically during the search procedure that can then be easily checked for a particular intervention set. This would prevent re-computing connections after the entire graph has been determined.

Quantitative Measures of Accuracy

[3] More generally, the policy maker is not only concerned with the qualitative information of the kind described above. It is not sufficient to know that a certain intervention will affect some other variable via particular causal connections. How strong the effect will be matters just as much: Quantitative information is relevant. Missing a weak causal link is less of a worry than missing a strong one. The notion of weak and strong causal links is not defined easily for an individual causal connection in a discrete model. But for linear models with edge coefficients, accuracy can be measured by associating two values with each pair of variables, one value for the edge coefficient in each direction. That is, a directed edge $X \to Y$ is associated with a value for the XY-direction that corresponds to the edge coefficient and a zero for the YX-direction, since there is no edge. For non-adjacent pairs of variables, both values are zero. The measure of accuracy then compares the squared difference of each value with the corresponding value in the true model. However, it is not clear how to aggregate these measures, since the sums of squares are not independent.

This concern applies more generally: For any one parameterized output, there is a long list of standard accuracy scores one could apply (AIC, BIC, Kullbach-Leibler distance, χ^2-distance etc.), each with their own pros and cons. But the more general difficulty for the circumstances considered in this thesis is to ensure that (i) the accuracy measure is sensitive to the particular features relevant to a causal model, i.e. the accuracy of the manipulated model, (ii) it is computationally feasible, and (iii) the accuracy measure is a metric. It is not clear how one could aggregate accuracy measures from the different manipulated distributions into one value that would provide an overall score. The problem is illustrated by the following example:

Cooper and Yoo [6] consider accuracy measures for models estimated on the basis of observational and experimental data. Their framework adds an additional variable E specifying the experimental state of the model. The variable functions as a switch that changes the state of the causal model from a passive observational state to a particular manipulated state. They define conditional probability distributions $P(\mathbf{V}|E)$ over the set of variables $\mathbf{V}$ conditional on the state of the variable E, where E can either take the value "observational" or "manipulate X" for some variable $X \in \mathbf{V}$. That is, E specifies whether the joint

[3] I am very grateful to Peter Spirtes for discussions and ideas I got from him on this topic. Only the errors in this section are purely mine.

distribution over the variables of interest is passive observational or whether it is manipulated by an intervention on one particular variable. Cooper and Yoo then define an observational error rate as the expected prediction error of one variable given the state of another (averaged over all pairs) based on the observational distribution. Similarly, for an error rate based on a manipulated distribution. The observational and manipulated error measures are not combined, since they are based on different conditional distributions.

However, if we were able to turn the conditional distribution $P(\mathbf{V}|E)$ into a joint distribution $P(\mathbf{V}|E)P(E)$, then accuracy measures from different experiments could be aggregated, the overall error rate would be defined on the basis of the joint distribution $P(\mathbf{V}, E)$. The main problem with such an account is to justify a particular distribution $P(E)$ over experiments. It is not clear under what circumstances one could consider such a marginal distribution or what its support would be.

For the circumstances we have been considering in the thesis, accuracy measures based on the parameterized models bring their own separate problems: So far we have only given an account of structure search, we have not made any suggestions about how to recover the parameterization.

8.1.2 Information Theoretic Measures and VC-Dimension

In information theory entropy is used as a measure of uncertainty about the value of a particular quantity [40]. It is defined as:

$$H(X) = -\sum_i p(X = x_i) \log p(X = x_i)$$

Reduction in entropy corresponds to a reduction in uncertainty. The terminology can be directly applied to the circumstances considered here. Initially, before any experiment is performed, we do not know which graph G is the true graph over N variables. If we let G be a random variable over graphs, then the entropy is initially maximal at $H(G) = \log_2(D_N)$, where D_N is the number of DAGs over N variables. Given an experiment, the uncertainty is reduced since the experiment determines (assuming an oracle) the Markov equivalence class of graphs for the manipulated distribution, which for most experiments, contains a subset of the original graphs. The *information gain* due to the experiment $\mathcal{E}$ is defined as

$$IG(G|\mathcal{E}) = H(G) - H(G|\mathcal{E})$$

where the conditional entropy $H(G|\mathcal{E}) = \sum_i p(\mathcal{E} = MME_i)H(G|\mathcal{E} = MME_i)$ and MME_i is a particular manipulated Markov equivalence class. So the information gain of one experiment is to be understood as the reduction in entropy from the initial set of possible graphs to the average entropy in each Markov equivalence class of the manipulated distribution generated by the experiment. This notion can be extended easily to sequences of experiments. A sequence of experiments uniquely discovers a graph when the information gain of the sequence of experiments is equal to the initial entropy, i.e. all uncertainty is resolved. A sequence of experiments that has an information gain equal to the initial entropy need not be made up of experiments that maximize information gain at each step in the sequence. This is another way of saying that a search using sequences of experiments need not be greedy.

For example, there are 25 DAGs over three variables, so the initial entropy is $H_0 = \log_2(25) = 4.644$. The following table shows the number of equivalence classes created by an experiment involving either a passive observation, an intervention on a single variable, an intervention on two variables or an intervention on three variables simultaneously, and its corresponding information gain. All interventions are assumed to be structural.

Experiment	# of Markov Equiv. Classes	Information Gain
Passive Observation	11	3.70
Single Intervention	10	3.64
Double Intervention	4	2.19
Triple Intervention	1	0

The passive observation has the highest information gain, but we know from earlier results that any combination of two different single intervention experiments is sufficient to recover the true causal graph over 3 variables, so a passive observation (with its maximum information gain) is unnecessary for best worst case discovery.

Computing the information gain of a sequence of experiments amounts to determining the Markov equivalence classes of causal graphs that for each experiment in the sequence imply the same conditional independence relations. In contrast to common Bayesian uses of information gain in graph search, which generally apply information gain to the probability distribution over graphs, the proposed use of information gain here is based on the qualitative features relevant to constraint based search. The computations involved to determine the equivalence classes of graphs are non-trivial and so we are not sure how fruitful this approach may be, but it seemed worth making the connection and adaptation to this area of learning theory.

In 1971 Vapnik and Chervonenkis defined the VC-dimension [45] as a measure of the capacity of a classification algorithm. The VC-dimension is defined as the cardinality of the largest set of points the classification algorithm can *shatter*, i.e. classify correctly, no matter how the points are arranged in space or how they are assigned to classes. For example, a classifier that amounts to a straight line in two dimensional space has a VC-dimension of three, since it can shatter three points that are not colinear, but cannot shatter all (non-colinear) arrangements of four points.

It seems intuitive that a similar measure of capacity should apply to search strategies using sequences of experiments. Essentially, the sequences of experiments shatter the space of graphs over N variables. No matter which graph is the true graph, the search strategy must be able to separate it out from the others. The classification problem is slightly different than the original case the VC-dimension was designed for, since in our case there is only one point, the *true* graph, that must be separated from the others, as opposed to separating two *sets* of points. Nevertheless, it seems that one could adapt the VC-dimension accordingly.

8.1.3 Cost of Discovery

Within this thesis the cost of discovery was only considered in terms of the number of experiments. We only briefly indicated that other measures of cost are not minimized by the sequences of experiments proposed here. In particular, we mentioned that if we aim to minimize the total number of different variables subject to an intervention over the sequence of experiments, then the distinction between the multiple simultaneous and single intervention strategies disappears. And if the aim is to minimize the number of variables subject to interventions over the sequence of experiments, then the single intervention strategies are superior to those with multiple simultaneous interventions. But of course, cost of discovery can take many more forms. Not every variable is equally easy to subject to an intervention. In some cases it might be impossible or unethical to perform a randomized trial on a certain variable. Cost can also be a function of sample size.

The game theoretic interpretation that we gave of the discovery problem lends itself quite naturally to a more detailed and differentiated consideration of the cost of discovery. While such an effort was originally going to form part of this thesis, too many other issues arose for the simple case of cost in terms of number of experiments already, so that a more detailed analysis of costs of experiments has been relegated to future research. We can here only give a brief introduction and indicate some of the difficulties.

Meganck et al. [26] assume that we are given a Markov Equivalence class of graphs and that the aim is to determine the next best intervention (on a single variable X_i) given some utility function. This follows closely the set-up by Tong and Koller [44] and Murphy [29], only that Meganck et al. consider the approach in a constraint based framework. They present various utility functions that essentially determine the best next intervention by how many edges the experiment resolves (at best, at least or expected). Consequently, their problem has the following form: Find the variable Y such that

$$Y = \arg\max_{X_i \in \mathbf{V}} U(X_i \in \mathbf{S})$$

where X_i is some variable and $U(X_i \in \mathbf{S})$ is the utility of intervening on X_i in the next experiment. The utility is given as

$$U(X_i \in \mathbf{S}) = \frac{gain(X_i \in \mathbf{S})}{cost(X_i \in \mathbf{S}) + cost(M(Ne(X_i)))}$$

where the $gain()$ is a function specifying the best/worst/expected number of edges (depending on the function) inferred by intervening on X_i, $cost(X \in \mathbf{S})$ is the cost of performing the intervention and $cost(M(Ne(X)))$ is the cost of measuring the neighbors of X_i.

Meganck et al. only consider a single structural intervention per experiment, but their utility functions could easily be extended to include multiple structural interventions per experiment. More importantly, however, their optimization only considers the next experiment. It is a greedy approach to discovery and it is not clear why a repeated optimization of utility for the next experiment would be optimal for the overall sequence of experiments – in particular if the framework were extended to multiple simultaneous experiments. But in any case, it presents the basic framework to address questions of utility in the search procedure.

The advantage of placing the issue of a more general cost structure in a game-theoretic framework, as we have suggested here, is that it straightfowardly enables the representation of different distributions over possible graphs and that many computational techniques for solving the optimization are already available. We think this is going to be one of the most interesting further extensions of the work presented in this thesis.

Appendix A

Proofs

The proofs are sorted according to the order they appear in the chapters of the main text. The theorems are repeated here with the original reference number. In many cases the proofs of theorems are preceded by some lemmas that are needed in the proof of the theorem.

A.1 Fixed Strategies

A.1.1 Structural Interventions

Lemma A.1.1:
If G is a causal graph over a set of variables $\mathbf{V}$, and G' the manipulated graph resulting from a structural intervention on a set of variables $\mathbf{S} \subset \mathbf{V}$, then for all pairs of variables $X, Y \notin \mathbf{S}$, X and Y are d-separated by some set $\mathbf{C} \subseteq \mathbf{V} \backslash \{X, Y\}$ in G if and only if X and Y are d-separated by some $\mathbf{C}' \subseteq \mathbf{V} \setminus \{X, Y\}$ in G'.

Proof. G' is identical to G except that all edges into variables in $\mathbf{S}$ in G do not occur in G'.
LTR: First assume X and Y are d-separated by some $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ in G. Then no undirected path between X and Y in G d-connects those variables relative to $\mathbf{C}$. Suppose for reductio that X and Y are not d-separated by $\mathbf{C}$ in G'. Then some path between X and Y in G' must be active, i.e., there is a d-connection. The paths between X and Y in G' are a subset of those in G. Thus some path between X and Y that was inactive in G must now be active in G'. Thus all nodes on such a path that were inactive in G must now be active in G'. But a node that is inactive on a path relative to $\mathbf{C}$ in G cannot become

active on the same path relative to $\mathbf{C}$ in G' when G' has fewer edges than G. Any inactive node must remain inactive and consequently no d-connection can arise. Consequently, X and Y are d-separated by $\mathbf{C}' = \mathbf{C}$ in G'.
RTL: Assume that X and Y are not d-separated by any $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ in G, that is, they are d-connected by every $\mathbf{C}$ in G. Then X and Y are adjacent in G, and an intervention on the variables in $\mathbf{S}$ does not remove this adjacency (since it does not contain X or Y), thus they are still adjacent in G' and thus d-connected by every $\mathbf{C}$ in G'. □

Lemma A.1.2:
If G is a causal graph over a set of variables $\mathbf{V}$, and G' the manipulated graph resulting from a structural intervention on a set of variables $\mathbf{S} \subset \mathbf{V}$, then for all pairs of variables X, Y with $X \in \mathbf{S}$ and $Y \notin \mathbf{S}$, X and Y are d-separated by some set $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ in G' if and only if X and Y are non-adjacent or $Y \to X$ in G.

Proof. LTR: Suppose that $X \to Y$ in G, then X and Y remain adjacent in G' and consequently d-connected for all $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$.
RTL: If $Y \to X$ in G then there cannot be any directed path from X to Y (acyclicity) and all incoming arrows on X, including this direct one are removed by the structural intervention on X, so there is no causal connection, so X and Y are d-separated for the empty set in G'. If X and Y are non-adjacent in G, then any d-connection in G' must be due to indirect paths. It can be proved that any set of indirect paths between two non-adjacent variables can be blocked with some conditioning set $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ to d-separate the variables (see foundation for PC-algorithm in [43]). □

Definition A.1.3: Structural Tests
We say an experiment on a set of causally sufficient set of variables is a *structural X-orientation test* for variables X, Y, if X but not Y is subject to a structural intervention. It is a *structural adjacency test* for X, Y if neither is subject to an intervention. The experiment is a *structural zero information test* for X, Y if both are subject to a structural intervention simultaneously. Two structural orientation tests for X, Y, are opposing if one is an X-orientation test and the other is a Y-orientation test.

Lemma A.1.4:
For discovery of the causal structure among any pair of variables X and Y either (i) a structural orientation test and a structural adjacency test or (ii)

two opposing structural orientation tests are sufficient and in the worst case necessary.

Proof. Sufficient (i): If a structural adjacency test determines X and Y to be d-separated for some conditioning set, then by Lemma A.1.1 they are determined to be non-adjacent, and we are done. If the adjacency test determines X and Y are d-connected for all conditioning sets, then they are, by Lemma A.1.1, adjacent and we require further test. If the structural X-orientation test determines X and Y are d-separated for some conditioning set, then by Lemma A.1.2, X and Y are either non-adjacent or $X \leftarrow Y$. Combined with the information from the adjacency test, this implies that $X \leftarrow Y$. If a structural X-orientation test determines X and Y are d-connected for all conditioning sets, then by Lemma A.1.2, $X \rightarrow Y$. This covers the three possible structures over two variables. Similarly for an orientation test with an intervention on Y.
Sufficient (ii): If a structural X-orientation test determines X and Y are d-connected for all conditioning sets, then $X \rightarrow Y$ and we are done, as in the previous case. Similarly for an intervention on Y. If a structural X-orientation test determines d-separation for some conditioning set then X and Y are non-adjacent or $Y \rightarrow X$, and a further test is needed. If a structural Y-orientation test also determines d-separation, then it implies the analogous disjunction with X and Y switched, and the combination of the results implies that X and Y are non-adjacent.
Necessary (i): Suppose the true graph is $X \rightarrow Y$, then an adjacency test alone is insufficient to determine the causal structure.
Necessary (ii): Suppose X and Y are non-adjacent in the true graph, then either structural orientation test is insufficient to determine the causal structure alone. □

Lemma A.1.5:
Let $G = (\mathbf{V}; \mathbf{E})$ be a graph on N vertices and let $\mathcal{E}$ be an experiment on G consisting of a simultaneous intervention on $K \leq N$ variables. Let $\mathbf{S} \subset \mathbf{V}$ be the set of variables subject to a structural intervention, i.e. $|\mathbf{S}| = K$, and $\mathbf{U} = \mathbf{V} \setminus \mathbf{S}$. Then

1. $\mathcal{E}$ is a orientation test for $K(N-K)$ pairs of variables, namely all pairs X, Y where $X \in \mathbf{S}$ and $Y \in \mathbf{U}$.

2. $\mathcal{E}$ is an adjacency test for $\binom{N-K}{2}$ pairs of variables, namely all pairs $X, Y \in \mathbf{U}$

3. $\mathcal{E}$ is a zero-information test for $\binom{K}{2}$ pairs of variables, namely all pairs $X, Y \in \mathbf{S}$.

Note that the number of pairs for which $\mathcal{E}$ is an orientation test is maximized at $K = \frac{N}{2}$.

Theorem: (fixed strategy) Single Structural Interventions, Causally Sufficient (3.3.1)
$N-1$ experiments are sufficient and in the worst case necessary to determine the causal graph among $N > 2$ variables[1] when only a single structural intervention is allowed in each experiment.

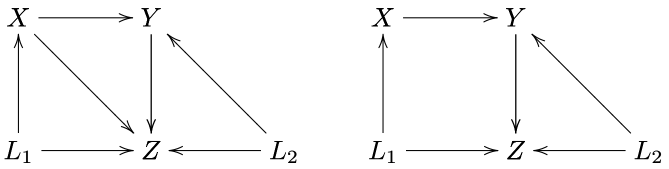
Proof. Sufficient: Consider $N-1$ experiments, each intervening structurally on a different variable. In each experiment $\mathcal{E}_i$ where X_i is subject to a structural intervention all $N-1$ pairs of variables (X_i, Y) for some $Y \in \mathbf{V}\backslash\{X_i\}$ are subject to a structural orientation test. By the end of the sequence of experiments every pair of variables has been subject to a structural intervention test. Except for all $N-1$ pairs of variables (Y, X_N), where X_N is the variable that is never subject to an intervention in the sequence and $Y \in \mathbf{V}\backslash\{X_N\}$, all other pairs have also been subject to an opposing orientation tests. However, each of the (Y, X_N) pairs was subject to an adjacency test by the time the second experiment completed (as were all other pairs). Hence, by Lemma A.1.4 the causal structure among each pair of variables can be uniquely determined, and hence the whole graph is discovered.

Necessary: Suppose only $N-2$ experiments were performed, one each on X_1 to X_{N-2}. Suppose that in the true underlying causal graph X_{N-1} and X_N happen to both be (direct) causes of each X_i, where $1 \leq i \leq N-2$, and that X_{N-1} and X_N are adjacent. Without loss of generality, assume that $X_N \rightarrow X_{N-1}$. In this case all of the interventions on $X_1, \ldots, X_{N-2}$ will indicate that there is an edge between X_N and X_{N-1}, but none are able to supply orientation information. Hence, an $(N-1)$th experiment is required. □

Theorem: (fixed strategy) Single Structural Intervention, Causally Insufficient (3.3.3)
Given a causally insufficient set of variables, no sequence of experiments is sufficient to determine the worst case causal graph among N variables when only a single structural intervention is permitted in each experiment.

[1] For $N = 2$, two experiments are sufficient and in the worst case necessary.

Proof. There is a very simple counter-example. Consider the following two graphs over three variables with two latent variables L_1 and L_2:



If X is subject to a structural intervention, the manipulated distribution over the observed variables contains no independence constraints in either graph. For a structural intervention on Y only $X \perp\!\!\!\perp Y$, but again for both graphs. Similarly for a structural intervention on Z, we only find that $Z \perp\!\!\!\perp \{X, Y\}$, but again for both graphs. No independence constraints obtain in the passive observational distribution. Consequently, the two graphs cannot be distinguished. The graphs can be embedded as subgraphs in cases where there $N > 3$ As long as X, Y, Z are either all common causes or all common effects of all other variables in the larger graph, the problem still occurs. □

Lemma A.1.6:
$\lceil \log_2(N) \rceil$ experiments are sufficient to subject all pairs of variables in a causal graph among N variables to a structural orientation test.

Proof. For $N = 2$ and for $N = 3$, one and two experiments, respectively, are sufficient to subject all pairs of variables to a structural orientation test, satisfying the Lemma. In each case the experiment consists of an intervention on a single variable. Now suppose that the theorem holds for all $N \leq r$. Then for $N = r + 1$ let the first experiment $\mathcal{E}_1$ consist of an intervention on $K = \lfloor \frac{N}{2} \rfloor$ variables. It follows from Lemma A.1.5, that $\mathcal{E}_1$ is a structural orientation test for the $\lfloor \frac{N}{2} \rfloor * \lceil \frac{N}{2} \rceil$ pairs of variables with one variable in $\mathbf{S}$ and the other in $\mathbf{U}$. Now, $|\mathbf{S}| = \lfloor \frac{N}{2} \rfloor$ and $|\mathbf{U}| = \lceil \frac{N}{2} \rceil$. Since the structural orientation tests for pairs of variables within $\mathbf{S}$ and $\mathbf{U}$ are independent of each other, we can perform them simultaneously. Hence, we need only worry about how many experiments it requires to resolve the larger one, $\mathbf{U}$. By the induction hypothesis we know that for $N' = \lceil \frac{N}{2} \rceil$, $\lceil \log_2(N') \rceil$ experiments are sufficient to subject all pairs in a causal graph among N' variables to a structural orientation test. Then adding the one experiment we started off with, we have

$$\lceil \log_2(N') \rceil + 1 = \lceil \log_2(\lceil \frac{N}{2} \rceil) \rceil + 1 = \lceil \log_2(2 \lceil \frac{N}{2} \rceil) \rceil = \lceil \log_2(N) \rceil$$

and so all pairs of variables in G are subject to a structural orientation test. □

Lemma A.1.7:
$\lfloor \log_2(N) \rfloor + 1$ experiments are sufficient to determine the causal graph among N variables.

Proof. Construct the sequence of experiments as follows:

# of variables	# of experiments	variable	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	...
1	1	X_1	0	0	0	0	0	...
2	2	X_2	1	0	0	0	0	...
3	2	X_3	0	1	0	0	0	...
4	3	X_4	1	1	0	0	0	...
5	3	X_5	0	0	1	0	0	...
6	3	X_6	1	0	1	0	0	...
7	3	X_7	0	1	1	0	0	...
8	4	X_8	1	1	1	0	0	...
9	4	X_9	0	0	0	1	0	...
10	4	X_{10}	1	0	0	1	0	...
11	4	X_{11}	0	1	0	1	0	...
12	4	X_{12}	1	1	0	1	0	...
13	4	X_{13}	0	0	1	1	0	...
14	4	X_{14}	1	0	1	1	0	...
15	4	X_{15}	0	1	1	1	0	...
16	5	X_{16}	1	1	1	1	0	...
17	5	...	...	...	...	...	...	...
...								

The first column specifies the number of variables N, the second column specifies how many experiments are appropriate for that number of variables and the third column lists the names of the N variables. The fourth column alternates 1s and 0s, and for each subsequent column, 1s and 0s alternate with half the frequency of the previous column. The table can then be read as follows: For any N and corresponding number of experiments k, columns $\mathcal{E}_1$ to $\mathcal{E}_k$, read from to row 1 to N specify the intervention sets of the k experiments. A 1 means that the variable is included in the intervention set, a 0 means that is is not.

For example, for $N = 7$, three experiments are necessary. So the entries up to row 7 and column $\mathcal{E}_3$ of the table are relevant. The three columns specify the three intervention sets: $\mathbf{S}_1 = \{X_2, X_4, X_6\}$, $\mathbf{S}_2 = \{X_3, X_4, X_7\}$ and $\mathbf{S}_3 = \{X_5, X_6, X_7\}$.

From Lemma A.1.4 we know that two opposing structural orientation tests or one orientation and one adjacency test are sufficient to determine the causal structure between two variables. Consequently, to prove that the sequence of experiments constructed as above is sufficient to discover the causal graph we show that for each pair of variables Lemma A.1.4 is satisfied. That is, for any number of variables N consider a pair of variables X_i, X_j with $i, j \leq N$. Each variable is associated with a sequence of 1s and 0s of length k (correponding to the appropriate number of experiments) given in the row of the variable in the table. Lemma A.1.4 can be satisfied if in these two sequences of length k there is one index such that both sequences are 0 at that index and one index where one is zero and the other 1 (i.e. an adjacency test and an orientation test) or there is one index where one sequence is 0 and the other 1, and another index where the first is 1 and the other 0 (i.e. two opposing orientation tests).

If N is a power of 2, this is trivially satisfied, since the kth column is always filled with 0s (hence an adjacency test), and for any two variables X_i, X_j with $i, j \leq N$ the sequences of length k always must differ for at least one index, since the rows are binary expansions of different integers by construction (hence an orientation test) .

When N is not a power of 2, suppose Lemma A.1.4 were not satisfied for all pairs of variables, i.e. it is not the case that for all pairs of variables, their two sequences (a) have an index where they are both 0 and (b) another index where one is 1 and the other 0; or (c) they have an index where one is 1 and the other 0, and another index where the former is 0 and the latter is 1. Passing the negation through, this implies that there is a pair of variables such that not (a) and not (c), or not (b) and not (c), which simplifies to (i) not (a) and not (c), or (ii) not (b), i.e. (i) there is no index for which both sequences are 0 and there are no two indices in which the two sequences differ in opposite ways, or (ii) there is no index for which the two sequences differ.

Note that (ii) leads to a contradiction because sequences from two different variables necessarily differ by construction (as in the case when N is a power of 2).

Consider (i) and suppose without loss of generality that the sequence for X_i is of the form $11001001\ldots$. If there are no indices in which both sequences are 0, then the sequence for X_j must have 1s in all the places that the sequence of X_i has 0s, i.e. $xx11x11x\ldots$, where x just means that the value is undetermined.

Now, since the sequences can also not have two indices in which they differ in opposite ways, none of the x's can be 0's. Hence they must all be 1's. But in that case the sequence of X_j consists entirely of 1s, in which case $j = N$ and N is a power of 2, which is a contradiction. Hence (ii) is impossible. It follows that all pairs of variables are subject to a pair of tests that satisfy Lemma A.1.4, and therefore the causal graph can be discovered.

□

Note, that the construction of the intervention sets given in the table is not unique, and that, depending on the value of N, even different sizes of intervention sets are possible without increasing the number of experiments, but this is not true in general. In the case of $N = 7$ we indicated for Strategy 3.3.5 that the bound of three experiments cannot be achieved when we intervene on 4 variables in one experiment. Whether or not it is necessary to intervene on the floor depends on how far N is from the closest power of 2. See the discussion of flexibility in the main text following Strategy 3.3.5.

Lemma A.1.8:
$\lceil \log_2(N) \rceil$ experiments are in the worst case necessary to subject all pairs in a causal graph among N variables to a structural orientation test.

Proof. It can easily be shown that for $N = 2, 3, 4$ the numbers of experiments necessary to subject all pairs to a structural orientation test are 1, 2 and 2 respectively, satisfying the above bound.

Suppose the theorem holds for all $N \leq r$. Then let $N = r + 1$. Consider all possibilities for the first experiment $\mathcal{E}_1$. It can consist of an intervention on K variables, where $0 \leq K < N = r + 1$. This implies that $\mathcal{E}_1$ subjects $K(N - K)$ pairs of variables to a orientation test. If the underlying true graph is complete and the choice of intervention set the least fortunate, $\mathcal{E}_1$ results in a complete undirected graph among the $(N - K)$ variables that were not subject to an intervention and constitutes a zero information test for all $\binom{K}{2}$ pairs of variables in the intervened set $\mathbf{S}$. Note that $|\mathbf{S}| = K < N = r + 1$ and $|\mathbf{U}| = N - K < N = r + 1$. Hence, we know by the inductive hypothesis that $\lceil \log_2(\max(K, N - K)) \rceil$ experiments are necessary to test the remaining subgraphs among variables in $\mathbf{U}$ and among variables in $\mathbf{S}$. Counting $\mathcal{E}_1$, it follows that the total number of experiments necessary to subject all pairs in a causal graph among N variables to an orientation test is given by:

$$x_{total} \quad = \quad 1 + \lceil \log_2(\max(K, N - K)) \rceil$$

$$
\begin{aligned}
&= 1 + \lceil \log_2(N/2) \rceil \\
&= 1 + \lceil \log_2(N) \rceil - 1 \\
&= \lceil \log_2(N) \rceil
\end{aligned}
$$

since $\max(K, N - K)$ is minimized at $K = N/2$. If N is odd, we intervene on $K = \frac{N-1}{2}$, since this maximizes the number of orientation tests and also subjects $\binom{(N-K)}{2}$ pairs to adjacency tests, but the above result still holds, since if N is odd, $\lceil \log_2(N) \rceil = \lceil \log_2(N+1) \rceil$. □

Lemma A.1.9:
$\lfloor \log_2(N) \rfloor + 1$ experiments are in the worst case necessary to determine the causal graph among N variables.

Proof. If N is not a power of two, then $\lfloor \log_2(N) \rfloor + 1 = \lceil \log_2(N) \rceil$ and we know from Lemma A.1.8 that $\lceil \log_2(N) \rceil$ experiments are necessary to subject each pair of variables to an orientation test, which is in the worst case necessary to discover the causal graph (see proof of Theorem 3.3.1). When N is a power of 2, consider the rows of 1s and 0s in the construction of the intervention sets in the table of Lemma A.1.8 . Since each pair of variables must have a sequence such that they differ for at least one index (one orientation test), the sequence must be at least of length $\log_2(N)$. But if the sequence for each variable is only of length $\log_2(N)$, then there one variable, say W, that is subject to an intervention in every experiment. Unless one further experiment is performed, any pair of variables W, Y for some $Y \in \mathbf{V} \setminus \{W\}$ is only subject to a W-orientation test. A further experiment is necessary to determine the causal graph. □

Theorem: (fixed strategy) Multiple Structural Interventions, Causally Sufficient (3.3.4)
$\lfloor \log_2(N) \rfloor + 1$ experiments are sufficient and in the worst case necessary to determine the causal graph among N variables when multiple simultaneous and independent structural interventions are allowed in each experiment.

Proof. By combining Lemma A.1.7 and A.1.9. □

Lemma A.1.10:
Two variables X and Y in a possibly causally insufficient set of variables $\mathbf{V}$ are unconditionally dependent in a structural X-orientation test in an experiment $\mathcal{E}$ if and only if all paths connecting the variables are directed from X to Y.

(Note that this implies that the d-connecting paths therefore do not involve latent variables, nor other variables in $\mathbf{S}$.)

Proof. If X and Y are connected by a path directed from X to Y in the manipulated distribution of experiment $\mathcal{E}$, then by the faithfulness condition, they will be dependent. If X and Y are unconditionally dependent in the manipulated distribution, then they are d-connected (by the Markov assumption), but the d-connection cannot be due to a path from Y to X, nor due to a (latent) common cause of Y and X, since all incoming arrows on X are broken by the structural intervention. Consequently, the d-connection must be due to paths (among the observed variables, excluding other variables in $\mathbf{S}$, since their incoming arrows would also be broken) from X to Y. □

Theorem: (fixed strategy) Multiple Structural Interventions, Causally Insufficient (3.3.6)
Given a causally insufficient set of variables, N experiments are sufficient and in the worst case necessary to discover the causal structure among the N observed variables if multiple variables can be subject to a structural intervention simultaneously and independently in each experiment.

Proof. In N experiments where each experiment consists of an intervention on all but one variable, while a different one is left out each time, unconditional independence tests imply that each pair of variables is subject to two opposing orientation tests. Since all other variables are subject to intervention, unconditional dependence implies by Lemma A.1.10 a direct edge. All ordered pairs of variables are tested. □

A.1.2 Parametric Interventions

Theorem: (fixed strategy) Single Parametric Intervention, Causally Sufficient (3.3.8)
$N-1$ experiments are sufficient and in the worst case necessary to determine the causal graph among N variables when only a single parametric intervention is allowed in each experiment.

Proof. Any experiment involving only parametric interventions is sufficient to establish adjacency information. Let each of the $N-1$ experiments $\mathcal{E}_i$, with $0 < i < N$ consist of a parametric intervention on X_i. We need to show how these experiments are sufficient to determine all orientation information. In each case, the intervention variable I_i forms an unshielded collider with any cause of X_i. Hence, for any variable Y, where Y and I_i are unconditionally independent, but dependent conditional on $\mathbf{C} \cup \{Y\}$ for all conditioning sets $\mathbf{C}$, we know that Y is a cause of X_i. Further, in each experiment we check whether X_i and X_N (which is not subject to an intervention) are dependent for all conditioning sets. If so, and if I_i, X_i and X_N do not form an unshielded collider, then X_i is a cause of X_N. Since we perform a parametric intervention on $N-1$ variables, all causes of these $N-1$ variables can be discovered, and since we check for each variable whether it is a cause of X_N, all of X_N's causes are determined as well. Hence, $N-1$ experiments are sufficient to discover the causal structure. $N-1$ experiments are in the worst case necessary, since $N-2$ parametric interventions would imply that two variables, say X_1 and X_2 are not subject to an intervention. If all other variables are common causes of X_1 and X_2 and X_1 and X_2 are adjacent then it is impossible to determine the orientation of the edge between them. □

This proof depends on a faithful distribution over the variables $\mathbf{V}$ *and* the intervention variables. But this can be weakened (see [31]). If only a conditional distribution over the observed variables given the intervention variables is available then one can check for each triple I_X, X, Y, where X and Y are known to be adjacent, whether I_X and Y covary for all conditioning sets $\mathbf{C} \setminus \{X\}$. If they do, orient the edge $X \rightarrow Y$, otherwise orient $X \leftarrow Y$. This latter strategy does not depend on distributions over the intervention variables since in principle these correlation checks can be based on changes in the conditional probability $P(Y|I_X, \mathbf{C})$.

Theorem: (fixed strategy) Multiple Parametric Interventions, Causally Sufficient (3.3.9)
One experiment is necessary and sufficient to determine the causal graph among N variables when multiple simultaneous parametric interventions are allowed in each experiment.

Proof. Since the parametric interventions described in the proof of the previous theorem do not interfere with each other, they can be performed all at the same time. That is, in a single experiment $N-1$ variables are subject to a parametric intervention and the causal structure is discovered all in one go. One experiment is obviously necessary and a passive observation is insufficient under the given assumptions. $N-1$ simultaneous parametric interventions are necessary. □

The theorem also follows directly from a theorem on rigid indistinguishability ([43], Theorem 4.6, Chapter 4). The same considerations as in the previous proof with regard to faithfulness and distributions over the intervention variables apply. Furthermore, it should be noted that – as the name suggests – parametric interventions interfere with the parameterization. The proofs here apply only to discovery of the causal structure, not to discovery of the parameterization.

Theorem: (fixed strategy) Parametric Interventions, Causally Insufficient (3.3.11)
No sequence of experiments is sufficient to determine the worst case causal graph among N causally insufficient variables if only parametric interventions (single or multiple) are allowed in the experiments.

Proof. A counter-example is shown in Figure 3.2. No independence constraints among the observed and intervention variables distinguish the first two graphs in the figure. Since the number of independence constraints is enormous, we do not include them here. They were automatically checked by the Causality Lab Program.[2] □

Theorem: Parametric Interventions and Inducing Paths (3.3.13)
Let G be a graph over a set of variables $\mathbf{V}$ and let $\mathbf{O}$ be a subset of $\mathbf{V}$ containing the observed variables. Let G_{man} be the graph G where each variable $X \in \mathbf{O}$ is extended with an intervention variable $I_X \rightarrow X$. The subgraph $G_{\mathbf{O}}$ of G over the observed variables can be uniquely determined by parametric interventions on each variable in $\mathbf{O}$ if and only if for each pair of variables $X, Y \in \mathbf{O}$ that are non-adjacent in G, there is no inducing path between I_X and Y and no inducing path between I_Y and X relative to $\mathbf{V} \cup \{I_X | X \in \mathbf{O}\}$ in G_{man}.

Proof. LTR: If there is some pair of variables $X, Y \in \mathbf{O}$ that are non-adjacent in G, there is an inducing path between I_X and Y and there is an inducing path between I_Y and X relative to $\mathbf{V} \cup \{I_X | X \in \mathbf{O}\}$ in G_{man} then by Theorem 6.1 in [43] I_X and Y are not d-separated by any subset of $(\mathbf{V} \cup \{I_W | W \in \mathbf{O}\}) \setminus \{I_X, Y\}$. Similarly, for I_Y and X. If neither I_Y and X nor I_X and Y can be d-separated, then the existence of an edge between X and Y cannot be determined from independence constraints, since they are the same, if the edge exists or not. So the subgraph $G_{\mathbf{O}}$ over the observed variables cannot be determined uniquely.

RTL: Suppose the graph over the N variables cannot be uniquely determined. Then there are two graphs G_1, G_2 that – despite any number of parametric

[2] http://www.phil.cmu.edu/projects/causality-lab/

interventions – imply the same conditional independence contraints. G_1 and G_2 can differ in several different ways: (i) in one graph two nodes are adjacent, which are non-adjacent in the other, or (ii) in one graph an edge between two variables is oriented in the opposite direction of the other. Suppose that (i) is the case, and without loss of generality, assume that variables X and Y are adjancent with $X \to Y$ in G_1 and non-adjacent in G_2. If the two graphs cannot be distinguished, then they must imply the same independence constraints, i.e. in this case I_X and Y must be d-connected for all conditioning sets for both graphs. By Theorem 6.1 in [43] it follows that if I_X and Y cannot be d-separated in G_2, and there must be an inducing path between them. Similarly for I_Y and X. If (ii) is the case, and $X \to Y$ in G_1 and $Y \to X$ in G_2, then I_Y and X must be d-connected in G_1 and I_X and Y in G_2. Again, from Theorem 6.1 in [43] it follows that there must be the appropriate inducing paths. □

A.1.3 Correlation Tests

Definition A.1.11: Partial Order Graph (POG)
Given a set of experiments on a set of (causally insufficient) variables $\mathbf{V}$ such that for each pair of variables X_i, X_j in $\mathbf{V}$ there is an experiment $\mathcal{E}_i$ with $X_i \in \mathbf{S_i}$ and $X_j \in \mathbf{U_i}$ and an experiment $\mathcal{E}_j$ with $X_i \in \mathbf{U}_j$ and $X_j \in \mathbf{S}_j$, let $\mathbf{O}_{\succ}$ be a partial order over the set of variables $\mathbf{V}$, such that $X_i \succ X_j$ if and only if $X_i \not\perp\!\!\!\perp X_j$ in $\mathcal{E}_i$. A partial order graph (POG) over variables $\mathbf{V}$ has an edge $X_i \to X_j$ if and only if $X_i \succ X_j$ and there does not exist $X_k \in \mathbf{V}$ with $X_i \succ X_k \succ X_j$.

Lemma A.1.12:
A POG is a subgraph (not necessarily strict) of the true graph over the set of observed variables $\mathbf{V}$.

Proof. If there is an edge between X_i and X_j in the POG, then $X_i \not\perp\!\!\!\perp X_j$ in $\mathcal{E}_i$, which implies by Lemma A.1.10 that they are d-connected via the observed variables. But there cannot be any other variable X_k on the path between the two in the true graph, since then we would have found that $X_i \not\perp\!\!\!\perp X_k$ in $\mathcal{E}_i$ and $X_k \not\perp\!\!\!\perp X_j$ in $\mathcal{E}_k$, which would have resulted in $X_i \succ X_k \succ X_j$, which was explicitly excluded in the construction of the POG. Hence, every edge in the POG must be an edge in the true graph. □

Lemma A.1.13:
The POG connects each variable by an oriented path (possibly indirectly) to all and only its descendents in the true graph.

Proof. Since the POG is a subgraph (Lemma A.1.12) of the true graph it cannot make a variable a descendent, if it is not a descendent in the true graph. Hence the "only" part. For the "all"-part, suppose for contradiction that there is a variable X_m that is a descendent of X_i in the true graph but is not connected to X_i in the POG. If X_m is a descendent, then there is a longest directed path $X_i \to X_{i+1} \to \ldots \to X_m$ with all variables on the path in $\mathbf{V}$ in the true graph. By the faithfulness condition, for all $0 \leq p < m$, $X_{i+p} \not\perp\!\!\!\perp X_{i+p+1}$ in $\mathcal{E}_{i+p}$ where $X_{i+p} \in \mathbf{S}$ and $X_{i+p+1} \in \mathbf{U}$. By Lemma A.1.10 and the definition of a POG it follows that $X_{i+p} \succ X_{i+p+1}$ for all $0 \leq p < m$. Hence, in the POG, each variable must be connected to all its descendents in the true graph. □

Definition A.1.14: Partial Order over Paths: $O(\mathcal{P}, \prec)$

Given a set $\mathcal{P}$ of directed paths, define a partial order over paths in $\mathcal{P}$ such that for any two paths $p_1 = X_1 \to \ldots \to X_r$ and $p_2 = Y_1 \to \ldots \to Y_s$, with $p_1, p_2 \in \mathcal{P}$ and $X_1, \ldots, X_r, Y_1, \ldots Y_s \in \mathbf{V}$, $p_1 \prec p_2$ if and only if there exists a path $p_3 \in \mathcal{P}$, such that $p_1 \subset p_3$ and $p_3 = Y_1 \to \ldots \to Y_s$ (i.e. p_1 is contained in p_3 and the endpoints of p_3 are the same as those of p_2).[3]

Lemma A.1.15:

For all paths in the POG, with some start vertex X_i and some end vertex X_j in $\mathbf{V}$, any direct edge between X_i and X_j that exists in the true graph, and its correlation can be determined.

Proof. Induction over the partial order over paths $O(\mathcal{P}, \prec)$: Clearly, this is trivially true for all the paths lowest in the partial order, since those are those variables that are connected by a directed edge. Suppose it is true for all paths in the partial order lower than some ordering rank k. Consider a path p with the ordering rank $k+1$ from some X_i to some X_j in the POG. The correlation between X_i and X_j in $\mathcal{E}_i$ where $X_i \in \mathbf{S}$ and $X_j \in \mathbf{U}$, equals by Lemma A.1.10 the sum of all the directed paths between X_i and X_j among observed variables in the true graph. By construction of the ordering, all directed connections between variables on p are considered lower ordering rank than p and by the induction hypothesis we know that all these direct connections and their correlations are known. Consequently, all directed paths between X_i and X_j except for the direct connection are known. Suppose all these paths add up to a correlation of ρ_{known} between X_i and X_j. If there is a direct edge $X_i \to X_j$ in the true graph,

[3] p_1 and p_2 are not ordered when both their endpoints coincide, i.e. $X_1 = Y_1$ and $X_r = Y_s$. But they are ordered if p_3 starts or ends with p_1, i.e. when p_1 and p_2 share one endpoint and there is a path p_3 containing p_1 and connecting the endpoints of p_2.

then the correlation between X_i and X_j in $\mathcal{E}_i$ where $X_i \in \mathbf{S}$ and $X_j \in \mathbf{U}$ does not equal the correlation due to the known paths, i.e. $\rho_{known} \neq \rho_{total}$, since the correlation due to the direct edge has not been accounted for in ρ_{known}. Hence the direct edge, and the correlation associated with it (the difference of the two values), is discovered. By induction all direct edges between any two variables X_i and X_j connected by a path in the POG, and their correlations are known. □

Theorem: (fixed strategy) Single Structural Intervention, Correlation-Test, Causally Insufficient (3.3.14)
Given a set of N causally insufficient variables, N experiments are sufficient and in the worst case necessary to determine the causal graph among the observed variables when only a single structural intervention is allowed in each experiment and the model is linear.

Proof. If each experiment $\mathcal{E}_i$ for $1 \leq i \leq N$ is an experiment with a single structural intervention on X_i, a POG can be constructed. Algorithm 5.3.4 can be used to determine the causal graph.

Suppose there is an edge $X_i \rightarrow X_j$ in the output of Algorithm 5.3.4 that is not in the true graph. Then either it was in the POG initially. But that is impossible, since the POG is a subgraph of the true graph (Lemma A.1.12). Or it was added in the phase in which correlations were computed. In this case there must be one or more indirect paths between X_i and X_j. Without loss of generality, assume the indirect path is $X_i \rightarrow \ldots \rightarrow X_k \rightarrow \ldots \rightarrow X_j$. Then by Lemma A.1.15 all direct connections between variables on the path, except for the endpoints are known at the time the direct connection is added. Consequently, the correlation between X_i and X_j in $\mathcal{E}_i$ due to all indirect paths is known: ρ_{known}. If there is no edge $X_i \rightarrow X_j$, then the $\rho_{known} = \rho_{total}$ and no edge is added. At no other point in the algorithm is an edge between these two variables added. So a false positive edge is impossible.

Suppose there is no edge $X_i \rightarrow X_j$ in the output of Algorithm 5.3.4 although there is an edge between the variables in the true graph. If there is no indirect path between X_i and X_j, then $X_i \rightarrow X_j$ is added in the POG, since the POG connects each variables to its descendents (Lemma A.1.13 and X_j is not a descendent of any variable that is a descendent of X_i). If there is an indirect path, then as in the previous case, by Lemma A.1.15 all indirect connections between X_i and X_j are known when the direct connection is considered. But since $\rho_{known} \neq \rho_{total}$, the direct edge $X_i \rightarrow X_j$ is discovered. At no other point in the algorithm is this pair of variables considered for edge-addition. So a false negative edge is impossible.

Hence, N experiments are sufficient. N experiments are necessary, since if only $N-1$ experiments are performed, then one variable X_N is not subject to a structural intervention. If there is a latent common cause of each variable and X_N and no variable is a cause of X_N, then it is impossible to determine whether X_N is a cause of the other observed variables or non-adjacent. □

Theorem: Search for Latent Common Causes: Single Structural Interventions (3.3.15)
Given a set of N causally insufficient variables and assuming the model is linear, N experiments, with a single structural intervention only per experiment, are sufficient and in the worst case necessary to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

Proof. Necessary: If one variable is not subject to intervention, then one cannot separate active from passive correlations, which implies that the correlation due to the latent variable could not be distinguished from the correlation due the paths among the observable variables. Since the structure among the observables could not be fully recovered, the latent variables cannot be identified.

Sufficient: Theorem 3.3.14 guarantees that one can discover the structure and correlations due to the paths among observed variables. We show that all latent common causes can be discovered by induction over the tier ordering of the observed variables in the graph. Consider all the roots of the POG. For any pair of root variables X, Y, any passively observed correlation must be due to a latent common cause, since neither variable is a cause of the other and no other observed variable is a cause of the root variables of the graph. Similarly, for any nodes disconnected in the graph among observed variables. Suppose all the latent common causes between roots and disconnected variables have been discovered. Consider any pair of variables R_i, X where R_i is a root and X is in the second tier of the tier ordering (since the roots are in the first tier). Any correlation between such a pair of variables is due to the direct edge $R_i \rightarrow X$ (known), any confounder of R_i and another root R that is connected by a directed edge to X (known), or a confounder of the pair itself. The latter can now be determined from the passive observational correlation by subtracting out the other components. Now, suppose that for any variable X in tier $r < k$ for some k that is connected to any root R by a path, all latent common causes of R and X are known. Now consider a variable Z in tier k connected to a root

R_j by a path of length $k-1$. Any correlation between R_j and Z is due to (i) directed paths from R_j to Z among the observed variables (known), (ii) paths involving a latent common cause of a pair of variables R_j and Y, where there is a directed path from Y to Z (known by induction hypothesis), and (iii) a latent common cause of R_j and Z. Since (i) and (ii) are known, (iii) can be determined and the residual correlation can be assigned to that latent common cause. Hence all latent common causes of the root and any variable are discovered. Once the latent common causes of the root and any other variable are determined, the induction is repeated now starting with tier two. Consider a pair of variables T_1, T_2 in tier two. Any passive observational correlation between those variables is due to (i) confounders of the roots that are connected to T_1 and T_2 by directed paths among the observables (known), (ii) confounders of T_1 and a root that is connected to T_2 by a directed path (known from the first induction), (iii) the same as (ii) with the indices switched (known), or (iv) a confounder of the T_1 and T_2, which can now be determined. The induction proceeds as before, now holding one variable fixed in tier two while the other is moved down the tier ordering. Similarly then for all tiers until all pairs of variables have been checked in this top down manner. □

We do not give an argument here that all pairs of variables are passively observed in the sequence of experiments, but it should be obvious for single intervention experiments that this condition is satisfied after the second experiment. In finding the appropriate experiment to determine the passive observational correlation between a pair of variables X, Y, one must take into account which paths are broken by any other variable that may be subject to an intervention in that experiment. We do not yet have any empirical data or further formal results on which circumstances are desirable, but presumably it often helps to choose experiments where the variance of the estimator of the residual correlation is small.

Theorem: (fixed strategy) Multiple Structural Interventions, Correlation-Test, Causally Insufficient (3.3.17)
$2\lceil \log_2(N) \rceil$ experiments are sufficient to determine the causal graph among N causally insufficient variables when multiple simultaneous structural interventions can be performed in each experiment and the model is linear.

Proof. By using $\log_2(N)$ intervention sets of different sets of $N/2$ variables (using the construction of a Cantor set outlined for various N variables in Strategy

3.3.5 or the table in the proof of Lemma A.1.7) and $\log_2(N)$ intervention sets that are the complements of those intervention sets, the total set of experiments satisfies the conditions for construction of a POG. The remainder of the proof then matches the sufficiency part of the proof of Theorem 3.3.14. □

Theorem: Search for Latent Common Causes: Multiple Structural Interventions (3.3.18)
Given a set of N causally insufficient variables and assuming the model is linear, $2\lceil \log_2(N) \rceil + 1$ experiments, with multiple simltaneous interventions per experiment, are sufficient to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

Proof. The proof is exactly analogous to the proof for Theorem 3.3.15. The extra experiment, a passive observation, is just to trivially ensure that for each pair of variables there is an experiment in which the pair is passively observed. □

Theorem: (fixed strategy) Single Parametric Interventions, Correlation-Test, Causally Insufficient (3.3.20)
N experiments are sufficient and in the worst case necessary to determine the causal graph among N causally insufficient variables when only a single parametric intervention can be performed in each experiment and the model is linear.

Proof. The proof is analogous to the proof of Theorem 3.3.14 except that construction of the partial order is now given by $X \succ Y$ if and only if $I_X \not\perp\!\!\!\perp Y$ in the experiment in which X is subject to a parametric intervention. For details of the computation of the correlations, see the corresponding Algorithm 5.4.1. □

Theorem: (fixed strategy) Multiple Parametric Interventions, Correlation-Test, Causally Insufficient (3.3.21)
One experiment is sufficient and in the worst case necessary to determine the causal graph among N causally insufficient variables when multiple simultaneous parametric interventions can be performed in each experiment and the model is linear.

Proof. Since parametric interventions do not interfere with each other, they can be combined in one experiment. Otherwise the proof follows the previous theorem. □

Theorem: Search for Latent Common Causes: Multiple Parametric Interventions (3.3.22)
Given a set of N causally insufficient variables and assuming the model is linear, one experiment, with multiple simultaneous interventions per experiment, is sufficient and in the worst case necessary to determine for each pair of observed variables, whether the pair is confounded by a latent common cause.

Proof. The proof is exactly analogous to the proof for Theorem 3.3.15. □

A.1.4 Restrictions

Theorem: Limited Structural Intervention Set, Causally Sufficient (3.3.29)
Given N causally sufficient variables, if the number of simultaneous structural interventions is limited by $k_{max} < \frac{N}{2}$ in any one experiment, then

$$(\frac{N}{k_{max}} - 1) + \frac{N}{2k_{max}} \log_2(k_{max})$$

experiments are sufficient to discover the causal graph.

Proof. Suppose without loss of generality that k_{max} divides N by some integer p where p is an even number.[4] Divide the N variables into p disjoint subsets of k_{max} variables. Let the first $p-1$ experiments each be a k_{max}-interventions on one of the p sets of variables. These $p-1$ experiments will result in structural orientation tests for all pairs of variables that go between the p sets, and in structural adjacency tests for all pairs of variables in the graph. So after $p-1$ experiments every pair of variables has been subject to one adjacency test and only the pairs of variables within each of the p sets have not yet been subjected to a orientation test. From Lemma A.1.6 it follows that $\log_2(k_{max})$ experiments are sufficient to subject all pairs in a causal graph among k_{max} variables to a structural orientation test. Since the maximum size of the intervention set used in Lemma A.1.6 is (in our case here) $\frac{k_{max}}{2}$ and since we are restricted in this case by k_{max} as the maximum size of the intervention set, it follows that we can perform interventions on two of the p sets concurrently in one experiment. Consequently, $\frac{p}{2} = \frac{N}{2k_{max}}$ sequences of $\log_2(k_{max})$ experiments each are sufficient to subject all the pairs of variables in the p sets to an orientation test. Once all

[4] If not, then we do not have integer values for the bound, however the results still hold for the ceiling of the resulting value.

these experiments have been performed, every pair of variables in the N-graph has been subject to an adjacency and an orientation test, which is sufficient to determine the causal graph. □

The bound is not necessary, since, for example, in the case of 24 variables and $k_{max} = 4$, the following 10 (instead of $(6-1)+3*2 = 11$ intervention sets are sufficient:

$$\{X_1, X_2, X_3, X_4\}, \{X_5, X_6, X_7, X_8\}, \{X_9, X_{10}, X_{11}, X_{12}\},$$
$$\{X_{13}, X_{14}, X_{15}, X_{16}\}, \{X_{17}, X_{18}, X_{19}, X_{20}\}, \{X_1, X_5, X_9 X_{13}\},$$
$$\{X_{17}, X_{21}, X_2 X_6\}, \{X_{10}, X_{14}, X_{18} X_{22}\}, \{X_3, X_7, X_{11} X_{15}\}, \{X_{19}, X_{23}\}$$

Not only are these fewer intervention sets, but the last one does not even contain 4 variables. For larger N and other values of k the discrepancy can be made even larger.

Theorem: Limited Parametric Intervention Set, Causally Sufficient (3.3.30)
Given N causally sufficient variables, if the number of simultaneous parametric interventions is limited by $k_{\max} < N-1$ in any one experiment, $\lceil \frac{N-1}{k_{\max}} \rceil$ experiments are sufficient and in the worst case necessary to discover the causal graph.

Proof. Since parametric interventions can be combined independently, one only needs to ensure that all but one variable are subject to a parametric intervention and then the proof for necessity and sufficiency follows the proofs of Theorems 3.3.8 and 3.3.9. □

A.2 Adaptive Strategies

Theorem: Adaptive vs. Fixed Strategies (3.4.1)
Under the same assumptions, no adaptive strategy can improve on the worst case bounds of theorems 3.3.1, 3.3.4, 3.3.6, 3.3.8 and 3.3.9 if the true graph is a worst case graph.

Proof. The result should be obvious given that none of the proofs of the theorems relied on the search strategy being *fixed*. But the proofs can also be given explicitly by describing the search for causal structure as a game between experimenter and nature, and by specifying an explicit strategy for nature such that an adaptive strategy by the scientist is no better than a fixed strategy: The experimenter specifies an experiment and nature returns the independence relations true of the graph, possibly modified by the experimental intervention. At each point in the game, however, nature may return the independence relations implied by the equivalence class of graphs that make discovery most difficult (in terms of the number of experiments) but where all graphs in the equivalence class are consistent with the independence relations supplied to the experimenter in the previous experiments. The claim of the theorem amounts then to the claim that there always exists a strategy for nature that ensures that the experimenter requires the same number of experiments as in the fixed strategy case to reduce the equivalence class of graphs over the N variables to one, i.e. to identify the underlying causal structure uniquely.

Nature's strategy is as follows: Let $X_1, \ldots, X_N$ be the variables the experimenter can intervene upon. When the experimenter intervenes on the set of variables $\mathbf{S}$ in the first experiment, nature maintains the equivalence class of graphs that satisfy the following conditions: The class contains all the graphs that have complete subgraphs among the non-intervened variables, and for all $X_k \notin \mathbf{S}$, X_k is a direct cause of every variable in $\mathbf{S}$. In other words, whichever set of variables the experimenter intervenes upon, they are the common effects of all the variables that have not been intervened upon.

Now consider the adaptive strategy of the experimenter trying to identify the graph. After the first experiment, she has no information about the directions of the edges among the non-intervened variables and consequently no information to adapt the next intervention set.

For Theorems 3.3.1 and 3.3.8 that involve single interventions only, the game just repeats from the sink of the graph upwards. After $N-2$ experiments, the scientist has no information on the orientation of the edge between X_{N-1} and X_N. Hence an $(N-1)$th experiment is required.

For Theorem 3.3.6 the scenario is similar, if the first variable to be left out of the intervention set is the root X_N of the graph, and the game proceeds down the graph-hierarchy, then one outgoing edge from X_N would be revealed at a time. But after $N-1$ experiments, the X_N to X_1 connection remains unresolved, so an Nth experiment is required.

For Theorem 3.3.9 an adaptive strategy trivially cannot do better than the fixed strategy, since even the fixed strategy only needs one experiment.

For Theorem 3.3.4 the proof follows from Lemm A.1.9, the *necessary* part of Theorem 3.3.4. Nature's strategy is a generalization of the strategy used in Theorem 3.3.1: Start with an equivalence class over all graphs (a complete no-knowledge graph). Maintain for any pair of variables in $\mathbf{U}$ an undirected edge, for any pair of variables in $\mathbf{S}$ no-knowledge edges, and for any pair of variables with $X \in \mathbf{S}$ and $Y \in \mathbf{U}$ a semi-directed edge $X \prec - - - Y$. This use of semi-directed edges implies that the first experiment with an intervention set of size K splits the discovery problem among the N variables into two subproblems of size K and $N - K$, respectively. No inferences can be made from the orientations in one subproblem to the orientations in the other if all the semi-directed edges between the two sub-problems are resolved in the same way (all non-adjacent, or all directed). That is, there is no adaptive advantage in the subproblems. The remainder then follows the inductive proof of Lemma A.1.8 and the *necessary* part of Theorem 3.3.4. □

An example with $N = 7$ variables will illustrate the strategy: If there are 7 variables, then the experimenter must intervene on three variables to guarantee to stay within the bound of three experiments, say she chooses $\{X_1, X_2, X_3\}$. Nature can return a complete adjacency graph over the four non-intervened variables with semi-directed edges into the each of the intervened ones. There is no useful information in such a return that would improve an *adaptive* strategy. So in the second experiment, the best the scientist can do is choose two from the non-intervened set and one from the intervened set, say $\{X_1, X_4, X_5\}$. Again, nature can retain the adjacency between X_5 and X_6 and return an adjacency between X_2 and X_3 (the other edges are not relevant). The scientist still has three edges whose orientation need to be resolved: (X_4, X_5) from the first experiment, and (X_2, X_3) and (X_6, X_7) from the second experiment. A third experiment with an intervention set of three variables cannot be avoided. There was no information that could be used to adapt the sequence.

A.3 Mixed Strategies

Lemma A.3.1:
For $N \geq 4$ the worst case expected number of experiments necessary and sufficient to uniquely determine the causal graph is greater than 2 if only single interventions are permitted per experiment.

Proof. Even if it is known that the true graph is a complete graph over at least four variables then the best case sequence of interventions does not resolve the orientations in less than two experiments. The worst case requires three experiments, so the average case must be greater than two. □

Lemma A.3.2:
The uniform distribution over complete graphs of N variables maximizes the expected number of experiments necessary and sufficient to discover the true graph uniquely when only single intervention experiments are permitted.

Proof. Suppose the uniform over complete graphs is not a worst case distribution. Then there is a distribution that is non-uniform over complete graphs or a distribution that has support over graphs that are not complete that implies a higher expected number of experiments. Suppose there is a such a distribution that has support on graphs that are not complete. But then, since by Lemma A.3.1 at least two experiments each intervening on a single but different variable are going to be performed, any missing edge of an incomplete graph will be discovered in those first two experiments. Discovery of the same graph with edges added to make it complete would take at least the same number of experiments, possibly more, since the additional orientations would have to be resolved. Since the graph is complete, the additional edges cannot make discovery easier, since no unshielded colliders are created. Hence, all incomplete graphs could be completed without increasing the average number of experiments. Hence, there is a worst case distribution with support over complete graphs only. Now suppose there is a worse distribution over complete graphs that is non-uniform. There are two types of non-uniformity: First, the distribution may assign zero probability to some complete graphs, but maintain a subset of graphs that is symmetric in the sense that each node is still equally likely to occur at any position in the graph (e.g. graph rotations). In this case the uniform distribution has a greater expected number of experiments since there are simply more graphs to distinguish. Second, if the non-uniformity arises in a way that makes a particular variable more likely to occur in the interior of the graph (rather than the root or sink), then one can intervene with greater probability on that variable. Whenever the intervened variable actually is in the interior of the graph, additional acyclicity constraints are created that imply a reduction in the number of experiments.[5] Since this happens more often in the non-uniform distribution, the expectation cannot be higher than in the uniform distribution over all complete graphs. □

[5] See the example Section 4.2.2 with an intervention on the middle variable of a complete graph over three variables: the graph is resolved in one experiment, as opposed to two.

Theorem: (mixed strategy) Single Structural Interventions, Causally Sufficient (3.3.4)
Given a set of $N > 3$ causally sufficient variables, the worst case expected number of experiments necessary and sufficient to discover the causal structure is $\frac{2}{3}N - \frac{1}{3}$ experiments if only one variable can be subject to a structural intervention per experiment.

Proof. By Lemma A.3.2, the uniform distribution over complete graphs is a worst case distribution. If all complete graphs over N variables are equally likely, then there is no priviledged node. Suppose without loss of generality that the true complete graph over the variables $X_1, \ldots, X_N$ is such that for all $i < j$, $X_i \rightarrow X_j$. Under these circumstances an intervention on X_i is

1. uninformative with respect to edge-orientation about all pairs of variables X_j, X_k with $j, k < i$.
2. uninformative with respect to edge-orientation about all pairs of variables X_j, X_k with $j, k > i$.
3. informative for the remaining edges: It resolves
   (a) edges between variables X_j, X_k with $j > i > k$
   (b) outgoing edges from X_i and,
   (c) since it is known that the graph is complete (distribution is only over complete graphs), all semi-directed edges can be resolved into directed edges, and so all edges incident on X_i are resolved.

In other words, an intervention on X_i splits the discovery problem into two subproblems, one with $N - i$ variables and the other with $i - 1$ variables. The intervention on X_i is uninformative with regard to these subproblems.

Given the uniform distribution over complete graphs, the problem is entirely symmetric in the sense that each node is equally likely to be at any of the possible positions in a complete graph. With a uniform distribution selecting among the unintervened variables, each variable is equally likely to be subject to an intervention in the first experiment. Consequently, we can give the expected number of experiments for this worst case distribution in terms of the numbers required for the subproblems the intervention creates:

$$E(\#\mathcal{E}(N)) = \frac{1}{N}\sum_{i=1}^{N}(E(\#\mathcal{E}(i-1)) + E(\#\mathcal{E}(N-i)) + 1)$$

where $E(\#\mathcal{E}(N))$ is the expected number of experiments required to discover the true graph if the graph is sampled from a Uniform over complete graphs of N variables, i.e. the expected number of experiments for N variables is one plus the average of the sum of the number of experiments that it takes to resolve the two subproblems of size $N-i$ and $i-1$, respectively. This can be simplified to

$$E(\#\mathcal{E}(N)) = 1 + \frac{2}{N}\sum_{i=1}^{N} E(\#\mathcal{E}(i-1))$$

with initial values that can be determined by hand:

Number of Variables	$E(\#\mathcal{E})$ for complete Graphs
0	0
1	0
2	1
3	5/3

We claim that the sum is equal to:

$$E(\#\mathcal{E}(N)) = \frac{2}{3}N - \frac{1}{3} \qquad \text{for } N \geq 2.$$

It is certainly true for $N = 2$. Suppose it is true for all integers up to some $N-1$. Then

$$\begin{aligned} E(\#\mathcal{E}(N)) &= 1 + \frac{2}{N}\sum_{i=1}^{N} E(\#\mathcal{E}(i-1)) \\ &= 1 + \frac{2}{N}\sum_{i=1}^{N}(\frac{2}{3}(i-1) - \frac{1}{3}) \\ &= 1 - \frac{2N}{3N} + \frac{4}{3N}\sum_{i=1}^{N}(i-1) \\ &= 1 - \frac{2}{3} + \frac{4N(N-1)}{6N} \\ &= -\frac{1}{3} + \frac{2}{3}N \end{aligned}$$

□

A.4 Conflicts

Definition A.4.1: Sets of Changing Interventions
For two experiments $\mathcal{E}_1 = \{\mathbf{S}_1, \mathbf{U}_1, \mathbf{Pol}_1\}$ and $\mathcal{E}_2 = \{\mathbf{S}_2, \mathbf{U}_2, \mathbf{Pol}_2\}$. The two sets of changing interventions between the two experiments are $\mathbf{Pol_1}^* = \mathbf{Pol_1} \setminus \mathbf{Pol_2}$ and $\mathbf{Pol_2}^* = \mathbf{Pol_2} \setminus \mathbf{Pol_1}$.

$\mathbf{Pol_1}^*$ contains all policy variables that occur in $\mathcal{E}_1$, but not in $\mathcal{E}_2$, and those that occur in both experiments, but are different in each case, e.g. even though variable Z may be subject to an intervention in both experiments, the type of intervention may be different each time, i.e. $Pol_1(Z) \neq Pol_2(Z)$. Similarly for $\mathbf{Pol_2}^*$. $\mathbf{Pol_1}^*$ and $\mathbf{Pol_2}^*$ contain all the intervention variables that *change* between the two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$.

Theorem: Pooling under d-separation (6.2.1)
If the set of variables $\{X, Y\}$ is d-separated from the set of changing intervention variables $\mathbf{Pol_i}^*$ given the conditioning set $\mathbf{C}$ in the manipulated graph of experiment $\mathcal{E}_i$ and if the set of variables $\{X, Y\}$ is d-separated from the set of changing intervention variables $\mathbf{Pol_j}^*$ given the conditioning set $\mathbf{C}$ in the manipulated graph of experiment $\mathcal{E}_j$, then the distributions relevant for independence test $T_{X,Y|C}$ are invariant across experiments $\mathcal{E}_i$ and $\mathcal{E}_j$ and the data relevant to the test can be pooled.

Proof. Let $P_i(\mathbf{V})$ be the distribution over variables $\mathbf{V}$ in experiment $\mathcal{E}_i$, similarly for $P_j(\mathbf{V})$ and $\mathcal{E}_j$. If $\{X, Y\}$ is d-separated from the changing intervention variables $\mathbf{Pol_i}^*$ in $\mathcal{E}_i$ given $\mathbf{C}$, and if $\{X, Y\}$ is d-separated from the changing intervention variables $\mathbf{Pol_j}^*$ in $\mathcal{E}_j$ given $\mathbf{C}$, then we have (following Theorem 7.1 in [43]):

$$P_i(X, Y|\mathbf{C}, \mathbf{Pol_i}) = P_i(X, Y|\mathbf{C}) = P_j(X, Y|\mathbf{C}) = P_j(X, Y|\mathbf{C}, \mathbf{Pol_j})$$

The joint (conditional) distribution over $X, Y|\mathbf{C}$ is invariant to the changing interventions in $\mathcal{E}_i$ and $\mathcal{E}_j$. It follows as a trivial consequence that the marginals $P(X|\mathbf{C})$ and $P(Y|\mathbf{C})$ are also invariant. Consequently, the independence test $T_{X,Y|\mathbf{C}}$ is invariant to the distributions of both experiments and the data relevant to this test can be pooled. Note, that the invariance is based on the distribution conditional on the policy variables, so there is no commitment to a marginal distribution over the intervention variables. □

Index

Bibliography

[1] F. Bacon. *Novum Organum.* Parry & MacMillan, 1620, 1854.

[2] R. Bradley, F. Dietrich, and C. List. Aggregating causal judgments. (unpublished at time of submission), 2006.

[3] J. Campbell. An interventionist approach to causation in psychology. In A. Gopnik and L. Schulz, editors, *Causal Learning: Psychology, Philosophy and Computation.* Oxford University Press, 2006.

[4] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

[5] D. M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

[6] G. Cooper and C. Yoo. Causal discovery from a mixture of data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 116–125, San Francisco, CA, 1999. Morgan Kaufmann.

[7] D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of 11th Conference on Artificial Intelligence and Statistics (AISTATS-07)*, 2007.

[8] F. Eberhardt. Error rates for strategies using sequences of experiments to discover the causal structure. *North Eastern Student Colloquium on Artificial Intelligence (NESCAI)*, 2006.

[9] F. Eberhardt. Sufficient condition for pooling data from different distributions. *Symposium on Philosophy, History, and Methodology of ERROR*, 2006. Forthcoming in Synthese, special issue, 163 (3): 433-442, 2008 Springer.

[10] F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21st Conference on Uncertainty and Artificial Intelligence(UAI-05)*, pages 178–184. AUAI Press, Corvallis, Oregon, 2005.

[11] F. Eberhardt, C. Glymour, and R. Scheines. N-1 experiments suffice to determine the causal relations among n variables. In D. E. Holmes and L. C. Jain, editors, *Innovations in Machine Learning*, volume 194 of *Theory and Applications Series: Studies in Fuzziness and Soft Computing*. Springer-Verlag, 2006.

[12] F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, pages 981–995, 2007.

[13] R. Fisher. *The design of experiments*. Hafner, 1935.

[14] J. Folkman. Graphs with monochromatic complete subgraphs in every edge coloring. *SIAM Journal on Applied Mathematics*, 18, No. 1:19–24, 1970.

[15] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42,6:1115–1145, 1995.

[16] W. S. Gossett. Comparison between balanced and random arrangements in field plots. *Biometrica*, 29:363–379, 1937.

[17] C. Hitchcock. On the importance of causal taxonomy. In A. Gopnik and L. Schulz, editors, *Causal Learning: Psychology, Philosophy and Computation*. Oxford University Press, 2005.

[18] P. O. Hoyer, S. Shimizu, and A. J. Kerminen. Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. In *Proceedings of 3rd European Workshop on Probabilistic Graphical Models (PGM-06)*, pages 155–162, 2006.

[19] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In R. Dechter and T.S. Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty and Artificial Intelligence (UAI-06)*, pages 437–444. AUAI Press, Corvallis, Oregon, 2006.

[20] J. B. Kadane. *Bayesian Methods and Ethics in a Clinical Trial Design.* Wiley-Interscience, 1996.

[21] J. B. Kadane and T. Seidenfeld. Randomization in a Bayesian perspective. In J. B. Kadane, , M. J. Schervish, and T. Seidenfeld, editors, *Rethinking the Foundations of Statistics.* Cambridge University Press, 1999.

[22] K. B. Korb, L. R. Hope, A. E. Nicholson, and K. Axnick. Varieties of causal intervention. In C. Zhang, H. W. Guesgen, and W. K. Yeap, editors, *Proceedings of the 8th Pacific Rim International Conferences on Artificial Intelligence*. Springer, 2004.

[23] C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–418. Morgan Kaufmann, August 1995.

[24] C. Meek. *Graphical Models, selecting Causal and Statistical Models*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 1996.

[25] C. Meek and C. Glymour. Conditioning and intervening. *British Journal of Philosophy of Science*, 45:1001–1021, 1994.

[26] S. Meganck, B. Manderick, and P. Leray. A decision theoretic approach to learning Bayesian networks. Technical report, Vrije Universiteit Brussels, 2005.

[27] G. Melançon, I. Dutour, and G. Bousque-Melou. Random generation of dags for graph drawing. Technical Report INS-R0005 February (2000), Dutch Research Center for Mathematical and Computer Science (CWI), 2000.

[28] J. S. Mill. *Philosophy of scientific method.* Hafner, 1843, 1950.

[29] K. P. Murphy. Active learning of causal Bayes net structure. Technical report, Department of Computer Science, U.C. Berkeley, 2001.

[30] R. Neapolitan. *Learning Bayesian Networks.* Pearson Prentice Hall, 2004.

[31] E. Nyberg and K. Korb. Informative interventions. In F. Russo and J. Williamson, editors, *Causality and Probability in the Sciences.* College Publications, London, 2006.

[32] J. Pearl. *Causality.* Oxford University Press, 2000.

[33] J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, 2006. AUAI Press.

[34] H. Reichenbach. *The Direction of Time.* University of California Press, 1956.

[35] T. Richardson. *Feedback Models: Interpretation and Discovery.* PhD thesis, Department of Philosophy, Carnegie Mellon University, 1996.

[36] J M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

[37] D. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[38] D. Rubin. Assignment to treatment group on the basis of covariate. *Journal of Educational Statistics*, 2:1–26, 1977.

[39] D. Rubin. Bayesian inference for causal effects: The role of randomizations. *Annals of Statistics*, 6:34–58, 1978.

[40] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[41] S. Shimizu, A. Hyvärinen, Y. Kano, and P. O. Hoyer. Discovery of non-gaussian linear causal models using ICA. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence (UAI-05)*, pages 525–533. AUAI Press, 2005.

[42] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning measurement models for unobserved variables. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 543–55, San Francisco, CA, 2003. Morgan Kaufmann.

[43] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* MIT Press, 2 edition, 2000.

[44] S. Tong and D. Koller. Active learning for structure in Bayesian networks. In B. Nebel, editor, *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 863–869, San Francisco, CA, 2001. Morgan Kaufmann.

[45] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[46] I. Verdinelli and J. B. Kadane. Bayesian designs for maximizing information and outcome. *Journal of the American Statistical Association*, 87:510–515, 1992.

[47] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 255 – 270, New York, NY, 1991. Elsevier Science.

[48] A. Wald. *Statistical Decision Functions.* Wiley, New York, 1950.

[49] J. Woodward. *Making Things Happen.* Oxford University Press, 2003.

[50] F. Yates. A fresh look at the basic principles of the design and analysis of experiments. In F. Yates, editor, *Experimental Design.* Griffin, London, 1970.