Department of Computer Science
Series of Publications A
Report A-2013-4

# Discovering Causal Relations in the Presence of Latent Confounders

Antti Hyttinen

*To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XII (University Main Building, Unioninkatu 34) on May 8th, 2013, at twelve o'clock noon.*

University of Helsinki
Finland

**Supervisor**
Patrik Hoyer, University of Helsinki, Finland

**Pre-examiners**
Alexander Statnikov, New York University, USA
Jaakko Hollmén, Aalto University, Finland

**Opponent**
Ioannis Tsamardinos, University of Crete, ICS-FORTH, Greece

**Custos**
Jyrki Kivinen, University of Helsinki, Finland

**Contact information**

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.helsinki.fi
URL: http://www.cs.Helsinki.fi/
Telephone: +358 9 1911, telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-8778-3 (paperback)
ISBN 978-952-10-8779-0 (PDF)
Computing Reviews (1998) Classification: I.2.6, G.3
Helsinki 2013
Unigrafia

## Discovering Causal Relations in the Presence of Latent Confounders

Antti Hyttinen

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
antti.hyttinen@helsinki.fi
http://www.cs.helsinki.fi/u/ajhyttin/

PhD Thesis, Series of Publications A, Report A-2013-4
Helsinki, May 2013, 107+138 pages
ISSN 1238-8645
ISBN 978-952-10-8778-3 (paperback)
ISBN 978-952-10-8779-0 (PDF)

### Abstract

The causal relationships determining the behaviour of a system under study are inherently directional: by manipulating a cause we can control its effect, but an effect cannot be used to control its cause. Understanding the network of causal relationships is necessary, for example, if we want to predict the behaviour in settings where the system is subject to different manipulations. However, we are rarely able to directly observe the causal processes in action; we only see the statistical associations they induce in the collected data. This thesis considers the discovery of the fundamental causal relationships from data in several different learning settings and under various modeling assumptions. Although the research is mostly theoretical, possible application areas include biology, medicine, economics and the social sciences.

Latent confounders, unobserved common causes of two or more observed parts of a system, are especially troublesome when discovering causal relations. The statistical dependence relations induced by such latent confounders often cannot be distinguished from directed causal relationships. Possible presence of feedback, that induces a cyclic causal structure, provides another complicating factor. To achieve informative learning results in this challenging setting, some restricting assumptions need to be made. One option is to constrain the functional forms of the causal relationships to be smooth and simple. In particular, we explore how linearity of the causal relations can be effectively exploited. Another common assumption under study is causal faithfulness, with which we can deduce the lack of causal relations from the lack of statistical associations. Along with these assumptions, we use data from randomized experiments, in which the system under study is observed under different interventions and manipulations.

In particular, we present a full theoretical foundation of learning linear cyclic models with latent variables using second order statistics in several experimental data sets. This includes sufficient and necessary conditions on the different experimental settings needed for full model identification, a provably complete learning algorithm and characterization of the underdetermination when the data do not allow for full model identification. We also consider several ways of exploiting the faithfulness assumption for this model class. We are able to learn from overlapping data sets, in which different (but overlapping) subsets of variables are observed. In addition, we formulate a model class called Noisy-OR models with latent confounding. We prove sufficient and worst case necessary conditions for the identifiability of the full model and derive several learning algorithms. The thesis also suggests the optimal sets of experiments for the identification of the above models and others. For settings without latent confounders, we develop a Bayesian learning algorithm that is able to exploit non-Gaussianity in passively observed data.

**Computing Reviews (1998) Categories and Subject Descriptors:**

I.2.6 [Artificial Intelligence]: Learning – Knowledge acquisition, Parameter learning

G.3 [Probability and Statistics]: Multivariate statistics, Correlation and regression analysis, Experimental design

**General Terms:**

Algorithms, Theory

**Additional Key Words and Phrases:**

Machine learning, Graphical models, Causality, Causal discovery

# Acknowledgements

The research for this thesis was conducted at Department of Computer Science, University of Helsinki during 2008-2012. The other affiliations included Helsinki Institute for Information Technology (HIIT) and Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN). The research was funded by the Academy of Finland, University of Helsinki Research Funds and HIIT.

First and foremost, I would like to thank Dr. Patrik Hoyer for his excellent guidance and tireless supervision. I am also grateful to Dr. Frederick Eberhardt, a co-author in many of the research articles, for co-operation and additional advice. I thank Professor Aapo Hyvärinen for leading the Neuroinformatics research group.

I would like to thank Doris Entner and Dr. Michael Gutmann for entertaining and sometimes even scientific discussions. The comments on the manuscript by P. H., F. E., D. E. and the pre-examiners helped to improve this dissertation. In addition, I would like to acknowledge the Department of Computer Science and HIIT for being a great place for doctoral studies. In particular, I would like to remember studying, discussing and hanging out with Stefan S., Pekka P., Esther G., Miika P., Antti H. and Kun Z..

I would also like to thank my parents Satu and Markku Hyttinen for support throughout my education. Finally and also especially, I thank Laura Alhola for all the support and welcomed distraction.

# Original Publications

This thesis consists of an introductory part and six research articles printed in their original form. The articles are not used in other dissertations. The contributions of the present author are detailed in Section 7.5, p. 91.

**Article I** P. O. Hoyer and A. Hyttinen: **Bayesian Discovery of Linear Acyclic Causal Models**. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 240–248, Montreal, Canada, 2009. AUAI Press.

**Article II** A. Hyttinen, F. Eberhardt, P. O. Hoyer: **Causal discovery for linear cyclic models with latent variables**. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pages 153–160, Helsinki, Finland, 2010. HIIT Publications 2010-2.

**Article III** A. Hyttinen, F. Eberhardt, P. O. Hoyer: **Learning Linear Cyclic Causal Models with Latent Variables**. *Journal of Machine Learning Research*, 13(Nov):3387–3439, 2012.

**Article IV** A. Hyttinen, F. Eberhardt, P. O. Hoyer: **Causal Discovery of Linear Cyclic Models from Experimental Data Sets with Overlapping Variables**. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 387–396, Catalina Island, California, USA, 2012. AUAI Press.

**Article V** A. Hyttinen, F. Eberhardt, P. O. Hoyer: **Noisy-OR Models with Latent Confounding**. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 363–372, Barcelona, Spain, 2011. AUAI Press.

**Article VI** A. Hyttinen, F. Eberhardt, P. O. Hoyer: **Experiment Selection for Causal Discovery**. Accepted to a journal conditioned on minor revisions, 2013.

# Contents

# Chapter 1

# Introduction

To truly understand a phenomenon is to know the underlying causes behind the phenomenon. Thus, the notion of causation is built deep into human understanding and language. Causal terms like ‘causes’, ‘prevents’, ‘inhibits’ and ‘contributes’ are often used in everyday language. But what do we exactly mean when we say for example that ‘smoking causes cancer’? We certainly do not mean that smoking always results in cancer nor that smoking is always necessary for developing cancer, the relation is probabilistic (Suppes, 1970). We imply that there is some statistical connection between smoking and developing cancer: smokers tend to develop cancer more often than non-smokers. However, this connection may have any number of explanations: there may exist genes inducing both nicotine addiction and cancer. Thus, the statement ‘smoking causes cancer’ also means something more. It means that there is some sort of physical, biological or chemical process or mechanism, from the cause (smoking) to the effect (cancer) (Salmon, 1984). However, exactly understanding these complicated mechanisms in detail may be very difficult, if not entirely impossible. Perhaps a better answer can be obtained by considering the use of such causal knowledge: if you are not particularly keen on developing cancer, you should not smoke. If you all of a sudden start to smoke more, you will be more likely to develop cancer, if you start smoking less, you are less likely to develop cancer. Thus, if we manipulate or intervene on the cause, and determine its value independent of its respective natural causes, we get a changed outcome of the effect (Woodward, 2003). On the contrary, intervening on the effect leaves the cause unchanged. These simple clarifications using processes and manipulations allow for mathematical formalization and development of computational tools for scientific utilization of the concept of causality.

As in many scientific inquiries, in this thesis we examine various kinds of systems, with characteristics that can be measured and modeled with random variables. Often we can also control or manipulate some characteristic of such a system. From measurements in various experimental settings we collect data, and from data we try to infer the mechanisms working in the system and understand how the different parts of the system interact. The systems considered may come from a number of application areas such as biology (e.g. gene regulation in biological cells), medicine (e.g. effects of treatments on patients), economics (e.g. interaction between different factors such as supply and demand), or the social sciences (e.g. connections between health and wealth). These systems are often very *complex*, the measurements are *noisy* and the workings of the system are also inherently *uncertain*: they include a number of stochastic factors. In the first example, it is clear that smoking only sometimes causes cancer, and the different biological mechanisms from the cause to the effect are indeed very complex.

Machine learning is a research field interested in modeling such complex systems exhibiting uncertainty from a general point of view. Perhaps opposed to traditional statistics, it is characteristic for machine learning to exploit the vast computational power of modern computers and the efficient algorithms of computer science. An often useful description of machine learning is implied by the name: some autonomous and intelligent machine is learning from experience, in order to predict the future and infer the best possible next action. This is also why machine learning can be seen as subfield of artificial intelligence: one of the requirements for a truly intelligent being is the ability to learn from experience and adapt the gained knowledge to new scenarios. However in practice, this autonomous and intelligent machine is perhaps more of an unachievable goal to work towards. What we want is more understanding of the world, in a somewhat efficient and cost effective manner. Thus, machine learning can essentially be seen as a collection of problem settings, efficient algorithms and successful principles for handling complexity and uncertainty.

In many machine learning settings, it is enough to model and understand the dependence relations, such as correlations, between different observed variables in the probability distribution generated by the system. With the understanding of such dependencies, it is possible to predict the behavior of the system in its natural unmanipulated state. But, such statistical dependencies are most often merely manifestations of the more fundamental causal processes and relations working in the system (Pearl, 2000). In causal discovery our aim is to understand these underlying causal relationships.

What are the uses for this deeper, causal understanding? Cooper (1999) divides the uses into three categories. First, we get insight into the *structure* of the causal processes working in the system. For example, we might be interested in the structure of the regulatory network controlling expressions of different genes in a biological cell. The mathematical objects used to present this causal structure are directed graphs, where nodes depict different random variables modeling the measured aspects of the system, and the edges correspond to direct causal relationships (Wright, 1934). See Figure 1.1 for some examples. Once we understand the structure of a system, we might in some cases be able to build a better one, perhaps fix the system if it is broken, or at least help the system to work more efficiently.

Second, causal understanding is needed for us to *predict* the outcomes of any manipulations we might consider doing in the future. Often we want to predict the outcomes of new, previously unseen manipulations. For example, a doctor needs to know whether the new mixture of drugs he is about to assign to a patient is likely to improve the condition or not.

Third, we need causal understanding to *explain* why a system produced the observed behavior. For example, if we observe a statistical dependence between the number of siblings a child has and obesity at a later age, explanations for this dependence are to be given in causal terms.

In this thesis, the focus is on *learning* causal relationships from data, under various different assumptions and settings. As causal relationships form a fundamental structure of a system, learning them is also perhaps an order of magnitude more demanding than learning of mere statistical relationships: we have to look deeper. A simple statistical dependence between two observed quantities in a passive observational data set might be the result of any combination of the causal structures in Figure 1.1 (a-c): either variable might be the cause of the other, or the dependence could be the result of a common cause of the two (Reichenbach, 1956). Yet another reason for dependence is selection bias in Figure 1.1 (d), but selection bias can often be ruled out with background knowledge on the data collection process. This important principle is often formulated as the slogan 'correlation does not imply causation'. In situations like this, where the causal structure is not uniquely determined by the data at hand, we say that the causal structure is not *identifiable* (Fisher, 1966). Note that observed correlation in combination with a temporal ordering is not sufficient indication for a causal edge between variables: the correlation might still be produced by a confounder (a third variable that is a cause of both, see Figure 1.1 (c)).

*Randomized experiments* (randomized controlled trials) have been used

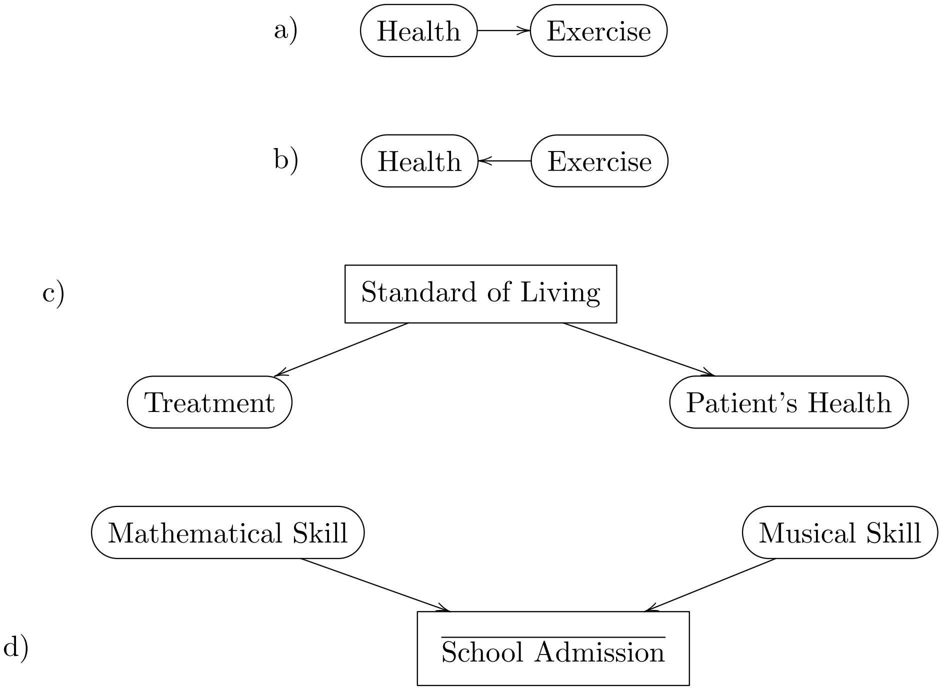


Figure 1.1: Some different causal structures possible for producing a dependence between two measured variables. Unmeasured variables are in rectangles. An arrow $x \to y$ represents a causal effect of $x$ on $y$. Correlation between health and exercise might be due to the need of good health to do exercise (a), or exercising actually benefiting the health (b). Latent confounding (c): standard of living may influence the access to both treatment and better nutrition. Thus, people receiving treatment might survive more often than those not receiving it, even if the treatment does not really affect the patient at all. Selection bias (d): if pupils are chosen into a school based on combination of mathematical and musical test scores, among the admitted pupils there might be a negative correlation between mathematical and musical skills. Overlining is used to mark conditioning: 'School Admission' gets value 'Yes' in the considered population.

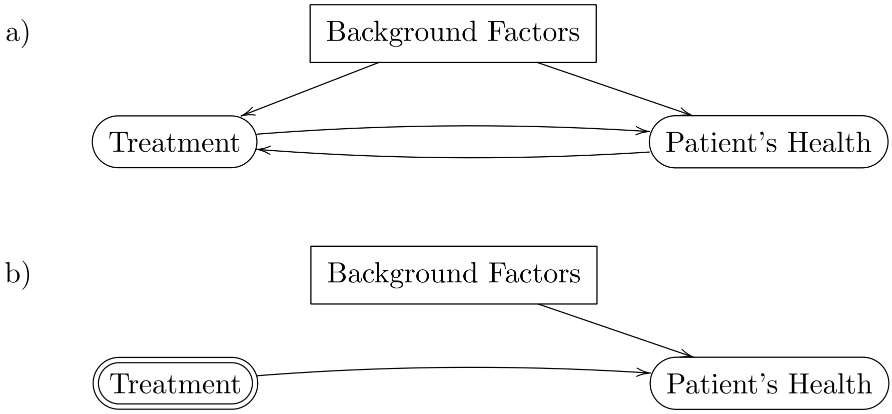


Figure 1.2: Causal discovery from experimental data (randomized controlled trials). (a) The causal effect of the treatment on patient's health is obscured by the latent confounding of the background factors. (b) If we assign the treatment to the patients randomly, for example according to a coin flip, any influence that background factors and patient's health have on the treatment are broken; this is denoted by removing all edges into the intervened variable. Then, any correlation between the treatment and health is due to direct causal influence of the treatment on the patient's health.

to discover causal relationships since the first half of the last century (Fisher, 1935; Rubin, 1974), and still provide the golden standard. Figure 1.2 shows a simple example of this setting. By randomly assigning the treatment to the patients, we can break the influence any other factors have on receiving the particular treatment, thereby rendering the causal effect the treatment has on health identifiable. However, often randomized controlled trials are not possible due to cost or ethical reasons. Performing a randomized controlled trial on every suspected causal relationship quickly becomes infeasible even for small systems. If we suspect a substance might be unhealthy, intentionally exposing some people to it is definitely unethical. Sometimes the interventions may not even be technically possible.

During recent decades, principles and assumptions for inferring causality also from *passive observation* have been formalized (Pearl and Verma, 1991; Pearl, 2000; Spirtes et al., 1993). An underlying idea in these basic results is to consider more variables than just two at the same time. The observed joint distribution for this larger set of variables can in some cases narrow down the possible causal structures significantly. Often used assumptions include acyclicity of the causal structure, causal sufficiency (no latent confounders like in Figure 1.1 (c) present), absence of selection bias and faithfulness. With such assumptions on reality there is always a trade-off: adding in more assumptions allows for more powerful learning methods and more causal relations discovered, but also limits the situations the methods are applicable for. Clearly, there is no single set of assumptions suitable for all learning settings. The field of causal discovery considers formalization, relaxation and assessment of various assumptions by which causal discovery is possible.

*Causal sufficiency*, the absence of latent confounders, is often one of the assumptions considered to be too restricting (Spirtes et al., 1993). But as we saw earlier, presence of *latent confounders*, unobserved common causes that affect two or more observed parts of a system, is a severe difficulty when discovering causal relations. The statistical dependence induced by such a latent confounder often cannot be distinguished from directed causal relationships. Latent confounding often makes causal relationships also harder to detect. Thus, one of the key research questions in this thesis is: How can we learn causal relationships in the presence of latent confounding variables?

One assumption particularly investigated in this thesis is the restriction of the parametric form of the causal relationships. In many articles we restrict the causal relationships to be *linear* (Geiger and Heckerman, 1994; Shimizu et al., 2006). Figure 1.3 shows an example of this. The ability

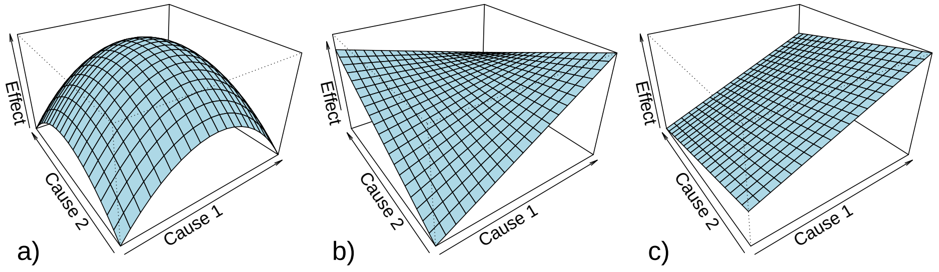


Figure 1.3: The surfaces show the values of the effect given forced values of its two causes. The functions producing the effect from the causes are non-linear in (a) and (b), and linear in (c). (a) Middle range values of both causes tend to produce large values for the effect. (b) Large values of both causes tend to produce large values for the effect individually, but if both causes are set to large values, the value of the effect is considerably smaller. (c) Larger values for the first cause increase (linearly) the value of the effect, while higher values for the second cause decrease (linearly) the value of the effect. Each cause affects the effect roughly in the same way regardless of the value of the other cause.

to estimate the parameters of this restricted form more reliably may well compensate for the possible bias arising from the slightly incorrect parametric assumptions (Koller and Friedman, 2009). By the use of different parametric restrictions, we are able to formulate interesting identifiability results and devise powerful learning algorithms also in settings where some of the other common assumptions are violated. By restricting the parametric form we can also often learn causal models with *cyclic* structures, which has been a recent interest in the causal discovery community (Richardson, 1996; Schmidt and Murphy, 2009). But, what exactly are the useful parametric restrictions for identifying the causal structure? How much do these restrictions diminish the applicability of the learning methods exploiting them?

After the formalization of many causal concepts and assumptions, the field of causal discovery has again turned its eye on experimental data from randomized controlled trials (Eberhardt, 2007; Cooper and Yoo, 1999). This is because sometimes the assumptions needed for learning causality from non-experimental data are too restricting or do not really hold in finite sample data, while at other times much of the causal relations may be left unidentified. In particular, a high number of latent confounders may cause the algorithms using only passive observational data to output uninformative results. In this thesis we concentrate on the somewhat unexplored possibility of exploiting experimental data to learn full causal models in conditions suffering from significant latent confounding. To accomplish this, we are combining the methodologies of randomized controlled experiments with assumptions exploited in causal discovery from passive observation. In many cases we can identify much of the causal structure from only a few different experimental settings. The important research questions are: How to exploit the given experimental data efficiently? Which characteristics should the experiments satisfy if our aim is to learn a full causal model? How can we select the experiments optimally?

Another assumption relevant to this thesis is the causal *faithfulness* assumption (Pearl, 1988). In plain terms it can be expressed as the slogan 'no correlation implies no causation'. If the causal relationship does not manifest itself through a statistical dependence in the data, it is considered non-existent by this assumption. Such a preference towards simpler models is commonly used in all fields of science in one way or another. But, how can we effectively and reliably exploit the causal faithfulness assumption?

This thesis is structured as follows. The first six chapters provide a general overview of the field of causal discovery, serving as necessary background material for understanding the original research presented in Chapter 7 and the reprinted articles at the end of the thesis. In Chapter 2 (Causal Modeling) we describe mathematical models of real world causal systems. In Chapter 3 (Causal Discovery) the basic approaches for causal discovery from passive observation are introduced. Rather than describing the learning algorithms in detail, we focus on the general principles and the underlying assumptions. Then, Chapter 4 (Experiments) considers extending these basic learning approaches to also exploit experimental data. In Chapter 5 (Latent Confounding) we consider what can be learned when latent confounders may be present. The interpretations and learning algorithms allowing for cyclic model structures are examined in Chapter 6 (Cycles).

The remainder of the thesis then focuses on the contributions of the original research articles. In Chapter 7 (Contributions to the Field) we briefly summarize the most important findings in the articles and relate the conducted research to the general field of causal discovery. The articles are printed in full at the end of this thesis. The contributions of the present author are summarized in Section 7.5. Chapter 8 (Conclusion) offers final remarks and suggests some further research possibilities.

# Chapter 2

# Causal Modeling

This chapter introduces causal models that formalize the concept of causality. Thus, we explain how causal relationships can be modeled mathematically. When building causal models one should always keep in mind the more informal interpretations[1] of causality laid out in the beginning of Section 1:

- A causal relation corresponds to a physical, chemical or biological process from a cause to an effect.
- When *ideally* intervening on the cause the effects and only the effects should generally change.

Causal models should describe the generated probability distribution under any possible (ideal) manipulations on the investigated system. Causal models accomplish this by exploiting the similarities of the probability distributions observed in the different experimental settings. Note that the causal models introduced here are in a sense static, they do not describe the dynamic behavior of the system as a time series as such. Nor do the models explain a single chain of events (token causality), but rather stochastic causal processes that relate different types of events.

Already in Chapter 1 we saw how complicated systems of causal structures can be described and understood using directed graphs. As it happens, most causal models use some sort of directed graph for describing the directed causal connections. As the graph notations are fairly simple, the definitions and concepts are introduced when needed. Another building block for causal models are probabilistic models, or simply probability distributions. They are introduced in Section 2.1 (Probabilistic Models).

[1] For a comparison of different philosophical interpretations of causality, see Woodward (2003).

Section 2.2 (Causal Bayesian Networks) then introduces perhaps the most intuitive causal model family, Bayesian networks. In Section 2.3 (Structural Equation Models) we model causality using an alternative framework called structural equation models (SEM); the original research of the thesis is most cleanly described using this framework. Instead of building these models by sequentially adding in assumptions and definitions, we will first introduce the model and only then examine what sort of assumptions we are making when using the model, and whether these are intuitive to our understanding and applicable to the real world. This is done because the different assumptions and definitions are inherently intertwined; they are best understood as a whole. Finally, Section 2.4 (How Do the Models Answer Our Causal Inquiries?) considers how the formal causal models can be used to answer our causal questions.

Note that alternative formalizations of causal concepts have been offered in the potential outcome framework[2] (Neyman, 1927; Rubin, 1974; Holland, 1986) and as Granger causality (Granger, 1969) in the field of time series analysis. See for example Berzuini et al. (2012) for an up-to-date discussion.

## 2.1 Probabilistic Models

Probabilistic models or probability distributions are designed to model (stochastic) uncertainty. A discrete probability distribution $P()$ defines a probability for all configurations of the values of a vector of discrete random variables $X_1, \ldots, X_n$. Such a function can be represented by a probability table, for example:

| $P(X_1, X_2)$ | $X_2 = 0$ | $X_2 = 1$ | $X_2 = 2$ |
|---|---|---|---|
| $X_1 = 0$ | 0.1 | 0.1 | 0.2 |
| $X_1 = 1$ | 0.2 | 0.3 | 0.1 |

For continuous random variables $x_1, \ldots, x_n$ the probabilities of different events can be defined using a probability density function $p()$. The probability of getting any single configuration is infinitesimal, but the probability of a set $\mathcal{S}$ of configurations can be calculated as

$$P(x_1, \ldots, x_n \in \mathcal{S}) \quad = \quad \int_{\mathcal{S}} p(x_1, \ldots, x_n) dx_1 \ldots dx_n. \tag{2.1}$$

[2] The potential outcome framework is mathematically equivalent to structural equation models (Pearl, 2000).

A common example of a probability distribution for continuous variables is the multivariate Gaussian distribution:

$$\begin{aligned} p(x_1, \ldots, x_n) &= \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (2.2) \\ &= \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \end{aligned}$$

where the random variables are in a vector $\mathbf{x} = [x_1, \ldots, x_n]^T$, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ a symmetric positive definite covariance matrix.

If we have access to a joint distribution of two disjoint sets of variables $\mathcal{X}$ and $\mathcal{Y}$, we can do two basic operations. If the variables are discrete, we can calculate the *marginal* distribution of a subset of variables $\mathcal{X}$ by summing over the different configurations of variables $\mathcal{Y}$:

$$P(\mathcal{X}) = \sum_{\mathcal{Y}} P(\mathcal{X}, \mathcal{Y}). \qquad (2.3)$$

For continuous random variables this sum is interpreted as an integral:

$$p(\mathcal{X}) = \int p(\mathcal{X}, \mathcal{Y}) dy_1, \ldots, dy_k,$$

where $\mathcal{Y} = \{y_1, \ldots, y_k\}$. These formulas are often called the sum-rule of probability.

We can also determine the *conditional* distribution of $\mathcal{X}$ given a certain configuration of $\mathcal{Y}$:

$$P(\mathcal{X}|\mathcal{Y}) = \frac{P(\mathcal{X}, \mathcal{Y})}{P(\mathcal{Y})}, \quad \text{when } P(\mathcal{Y}) > 0. \qquad (2.4)$$

When using the notion of conditional probability, the product-rule of probabilities is often useful:

$$P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X}|\mathcal{Y})P(\mathcal{Y}) = P(\mathcal{Y}|\mathcal{X})P(\mathcal{X}). \qquad (2.5)$$

The notion of conditional probability as well as the product rule apply also for probability density functions $p()$ of continuous variables, keeping in mind the (slightly) different interpretation of marginalization.

An essential part of probability theory is the concept of (marginal) independence[3] between two sets of random variables:

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \Leftrightarrow P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X})P(\mathcal{Y}). \qquad (2.6)$$

[3] The useful notation $\perp\!\!\!\perp$ is related to $\perp$ marking the orthogonality of vectors (Dawid, 1979).

Intuitively this means that when predicting the values of $\mathcal{X}$, knowing the values of variables $\mathcal{Y}$ does not help in the prediction task: $P(\mathcal{X}|\mathcal{Y}) = P(\mathcal{X})$. Equivalently, knowing the value of variables $\mathcal{X}$ does not help when predicting $\mathcal{Y}$: $P(\mathcal{Y}|\mathcal{X}) = P(\mathcal{Y})$. We define conditional independence similarly:

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \quad \Leftrightarrow \quad P(\mathcal{X},\mathcal{Y}|\mathcal{Z}) = P(\mathcal{X}|\mathcal{Z})P(\mathcal{Y}|\mathcal{Z}), \text{ when } P(\mathcal{Z}) > 0, \tag{2.7}$$

where the constraint $P(\mathcal{Z}) > 0$ denies conditioning on an event of probability zero. Intuitively, when predicting the values of $\mathcal{X}$, once we know values of variables $\mathcal{Z}$, knowing the values of variables $\mathcal{Y}$ does not help in the prediction task: $P(\mathcal{X}|\mathcal{Y},\mathcal{Z}) = P(\mathcal{X}|\mathcal{Z})$. This condition should apply for every configuration of the conditioning variables in $\mathcal{Z}$. If the condition applies only to some configuration of the conditioning variables, we talk about *context-specific independence.* Two sets of variables are independent in some context[4] $\mathcal{Z} = \mathbf{z}$ if and only if their distribution factorizes accordingly:

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} = \mathbf{z} \quad \Leftrightarrow \quad P(\mathcal{X},\mathcal{Y}|\mathcal{Z} = \mathbf{z}) = P(\mathcal{X}|\mathcal{Z} = \mathbf{z})P(\mathcal{Y}|\mathcal{Z} = \mathbf{z}), \tag{2.8}$$
$$\text{when } P(\mathcal{Z} = \mathbf{z}) > 0.$$

Probabilistic models handle uncertainty, but the complex structure of the statistical relationships may not be explicitly visible. Graphical models handle this complexity by combining some sort of a graph and a condition that connects the graph with some independence statements among the variables. One example of such a model class, non-causal Bayesian networks, is given in Section 2.2.4 (Bayesian Networks as Probabilistic Models).

## 2.2 Causal Bayesian Networks

The most familiar causal modeling framework is given by causal Bayesian networks (Pearl, 1988, 2000). Usually, this model is presented using discrete variables. For a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with the set of nodes as the observed variables $\mathcal{V} = \{X_1, \ldots, X_n\}$ and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, we will also use the notion of parent set

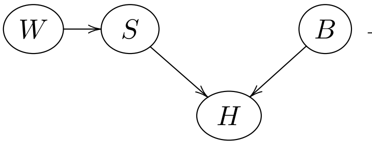
$$\mathrm{pa}(X_i) \quad = \quad \{X_j|(X_j, X_i) \in \mathcal{E}\}, \tag{2.9}$$

where the associated directed graph $\mathcal{G}$ is implicit. We present the model here as defining the distributions in all possible experimental settings (Pearl, 2000).

[4] The notation $\mathcal{Z} = \mathbf{z}$ represents the fact that all members of set $\mathcal{Z}$ obtain the given values.

| $P(W)$ | $W=0$ | $W=1$ |
|---|---|---|
| | 0.3 | 0.7 |

| $P(S\|W)$ | $S=0$ | $S=1$ |
|---|---|---|
| $W=0$ | 0.5 | 0.5 |
| $W=1$ | 0.9 | 0.1 |



| $P(B)$ | $B=0$ | $B=1$ |
|---|---|---|
| | 0.4 | 0.6 |

| $P(H\|S,B)$ | $H=0$ | $H=1$ |
|---|---|---|
| $S=0,B=0$ | 0.3 | 0.7 |
| $S=0,B=1$ | 0.6 | 0.4 |
| $S=1,B=0$ | 0.5 | 0.5 |
| $S=1,B=1$ | 0.8 | 0.2 |

Figure 2.1: Causal Bayesian network for curing a headache. Whether a person is at work or not ($W$) affects the possibility to sleep ($S$). Sleeping and taking painkillers ($B$) may prevent the person having a headache an hour later ($H$).

**Definition 1 (Causal Bayesian Network)** *A causal Bayesian network over variables $\mathcal{V} = \{X_1, \ldots, X_n\}$ consist of*

- *a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and*
- *conditional probability distributions $P(X_i|\mathrm{pa}(X_i))$ defined by (a vector of) parameters $\boldsymbol{\theta}$.*

*In the non-experimental setting it produces the distribution*

$$P(X_1, \cdots, X_n) \quad = \quad \prod_{X_i \in \mathcal{V}} P(X_i|\mathrm{pa}(X_i)) \tag{2.10}$$

*and when variables in set $\mathcal{J} \subseteq \mathcal{V}$ are intervened on, then the remaining passively observed variables $\mathcal{U} = \mathcal{V} \setminus \mathcal{J}$ have distribution*

$$P(\mathcal{U}) \quad = \quad \prod_{X_i \in \mathcal{U}} P(X_i|\mathrm{pa}(X_i)). \tag{2.11}$$

Figure 2.1 gives an example of a causal Bayesian network modeling the causal processes for curing a headache. The parameters $\boldsymbol{\theta}$ in this example model consist of the decimal quantities placed in the conditional probability tables.

First of all, the model formalizes the notion of causality. The model considers variable $X_i$ as a cause of some other variable $X_j$, if and only if there is a directed path[5] from $X_i$ to $X_j$ in the graph $\mathcal{G}$. The causes from

[5] A path between $X_i$ and $X_j$ is a sequence of nodes $X_i, \ldots, X_j$ with edges between all adjacent nodes in the sequence. A directed path from $X_i$ to $X_j$ is a path between $X_i$ and $X_j$ where all edges are directed towards $X_j$.

which there is a direct edge to the effect $Y$ are called *direct*, other causes are *indirect*. This interpretation of the edges postulates that causality is *transitive*: if $X_i$ is a cause of $X_j$ and $X_j$ is a cause of $X_k$, then $X_i$ is a cause of $X_k$. In the example for figure 2.1 $S$ and $B$ are direct causes of $H$, while $W$ is a direct cause of $S$ and an indirect cause of $H$.

Since the graph is acyclic, the model also assumes that in the modeled system no variable can be a cause of itself. Similarly if some variable $X_i$ is a cause of another variable $X_j$, then $X_j$ cannot be a cause of $X_i$. Causal relations in this type of a model are thus *asymmetric*.

**Assumption 1 (Acyclicity)** *The graph of causal relations does not include any directed cycles.*

For now, we take acyclicity as a working assumption. In Section 6 (Cycles) we will discuss this assumption in more detail. One useful implication of acyclicity is that the variables can be ordered in at least one *causal order*, where a cause always precedes its effects. In the example of Figure 2.1, the possible causal orders are $W, S, B, H$ or $W, B, S, H$ or $B, W, S, H$. Formally, a causal order of an acyclic graph can be defined as follows:

$$\mathrm{o} : \{1, \ldots, n\} \mapsto \{1, \ldots, n\} \text{ such that } \forall i, \forall j > i : X_{\mathrm{o}(j)} \notin \mathrm{pa}(X_{\mathrm{o}(i)}). \tag{2.12}$$

Then, $X_{\mathrm{o}(1)}, X_{\mathrm{o}(2)}, \ldots, X_{\mathrm{o}(n)}$ is a causal order of the variables $X_1, \ldots, X_n$.

Now, we can examine how an individual sample of variable values is obtained from the model. This sampling procedure should respect the idea of real causal processes, continuous in time and space, that produce the measured values of the variables. A sample can be generated by the following ancestral sampling recipe (Bishop, 2006):

1. Sample $X_{\mathrm{o}(1)}$ from $P(X_{\mathrm{o}(1)})$.

   $\vdots$

i. Sample $X_{\mathrm{o}(i)}$ from $P(X_{\mathrm{o}(i)} | \mathrm{pa}(X_{\mathrm{o}(i)}))$.

   $\vdots$

n. Sample $X_{\mathrm{o}(n)}$ from $P(X_{\mathrm{o}(n)} | \mathrm{pa}(X_{\mathrm{o}(n)}))$.

The variables are sampled here in the causal order: causes always get their values before the effects. Then, before sampling $X_i$, all of its direct causes $\mathrm{pa}(X_{\mathrm{o}(i)})$ have already obtained their values. The time in the sampling process seems to resemble the time in the modeled real processes.[6] The conditional probability distribution $P(X_{\mathrm{o}(i)}|\mathrm{pa}(X_{\mathrm{o}(i)}))$ models the stochastic causal processes that bring about the value of $X_i$ from values of its direct causes $\mathrm{pa}(X_{\mathrm{o}(i)})$. As the values of the direct causes considered in the model do not deterministically determine $X_i$, this is a probability distribution. We can easily verify that the probability of sampling any single configuration of variables $\mathcal{V}$ is indeed given by the right side of Equation 2.10. In the example of Figure 2.1, we would first sample $W$ and $B$ from their marginal distributions, $S$ from the row indicated by $W$ and finally $H$ from the row specified by the values of $S$ and $B$.

### 2.2.1 Causal Markov Condition

Equation 2.10 implies that the joint distribution can be factorized: the joint distribution is a product of the conditional distributions of each variable given its direct causes. This factorization property is equivalent[7] to the following assumption on the causal relations of the real world (Pearl, 2000; Spirtes et al., 1993).

**Assumption 2 (Local Causal Markov Condition)** *A variable is independent of its non-effects conditional on its direct causes.*

One of the consequences of this assumption is the following equation:

$$P(X_{\mathrm{o}(i)}|X_{\mathrm{o}(1)},\ldots,X_{\mathrm{o}(i-1)}) \quad = \quad P(X_{\mathrm{o}(i)}|\mathrm{pa}(X_{\mathrm{o}(i)})). \tag{2.13}$$

Here the conditioning set on the left consists of variables that are all non-effects of $X_{\mathrm{o}(i)}$ due to the definition of the causal order. As this set includes all direct causes of $X_i$, Equation 2.13 is a direct consequence of Assumption 2. Previously, we already saw this property in action in the sampling process: when sampling $X_{\mathrm{o}(i)}$ from $P(X_{\mathrm{o}(i)}|\mathrm{pa}(X_{\mathrm{o}(i)}))$ only the values of its direct causes $\mathrm{pa}(X_{\mathrm{o}(i)})$ were taken into account. The indirect causes did affect the value $X_i$, but only indirectly, through the sampled values of the direct causes. The factorization of Equation 2.10 is implied by Equation 2.13 and the product rule of probabilities (Equation 2.5):

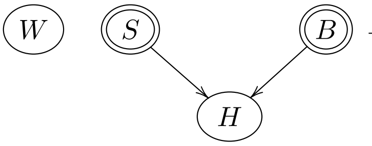
$$P(X_1,\ldots,X_n) \quad = \quad \prod_{i=1}^{n} P(X_{\mathrm{o}(i)}|X_{\mathrm{o}(1)},\ldots,X_{\mathrm{o}(i-1)}) = \prod_{i=1}^{n} P(X_i|\mathrm{pa}(X_i)).$$

[6] This assumes that there is a well defined true causal order and we are sampling in that particular order.

[7] The discussion hereafter shows that Assumption 2 implies the factorization. See for example Hausman and Woodward (1999) for the other direction.

| $P(W)$ | $W=0$ | $W=1$ |
|---|---|---|
| | 0.3 | 0.7 |

| $P(S)$ | $S=0$ | $S=1$ |
|---|---|---|
| | 0 | 1 |



| $P(B)$ | $B=0$ | $B=1$ |
|---|---|---|
| | 0 | 1 |

| $P(H\|S,B)$ | $H=0$ | $H=1$ |
|---|---|---|
| $S=0,B=0$ | 0.3 | 0.7 |
| $S=0,B=1$ | 0.6 | 0.4 |
| $S=1,B=0$ | 0.5 | 0.5 |
| $S=1,B=1$ | 0.8 | 0.2 |

Figure 2.2: Manipulated Causal Bayesian network of curing a headache, when a person decides to sleep ($S=1$) and take a painkiller ($B=1$). By marking trivial distributions $P(S)$ and $P(B)$ for the intervened variables, the manipulated situation can be cast as another causal Bayesian network. Compare to Figure 2.1.

Note that the (local) causal Markov condition (Assumption 2) is not merely a theoretical definition, it says something about causal relationships of the real world.[8] Indirect causes do not help in predicting the value of a variable once the values of the direct causes are known: all information is already in the values of the direct causes. However, knowing the value of an effect might still provide additional information for the prediction. The applicability of the causal Markov condition is still under some philosophical debate (Spirtes et al., 1993; Hausman and Woodward, 1999; Sober, 2001). In most causal systems, causal Markov condition is a useful and valid modeling assumption (Spirtes et al., 1993). But there also exist systems where the assumption is not valid or its application needs very careful inspection: in the presence of merely accidental correlations, when the considered population is a mixture of subpopulations, when the variables are measured inaccurately, and in quantum mechanics (Hausman and Woodward, 1999). In the model of Figure 2.1 the local causal Markov condition implies, for example, that $H \perp\!\!\!\perp W \mid S, B$, which seems plausible: the headache persisting does not depend on where you are, given the information on whether you slept or took a painkiller.

## 2.2.2 Manipulation

The causal model definition also describes what is common between the different distributions observed under different experimental settings in Equation 2.11: the conditional distributions of the passively observed variables $\mathcal{U}$ given their respective direct causes. This is formalized by the following assumption (Dawid, 2010; Woodward, 2003).

[8] Assumption 8 defines a global Markov condition, the discussion here applies to both.

**Assumption 3 (Invariance/Modularity)** *For any variable $X_i \in \mathcal{V}$, the causal processes producing its value, defined by $P(X_i|\mathrm{pa}(X_i))$, are unaltered no matter which variables of the system, other than $X_i$, are intervened on.*

Note in particular that for the conditional distribution $P(X_i|\mathrm{pa}(X_i))$ determining the value of $X_i$ it is indifferent whether a direct cause $X_j \in \mathrm{pa}(X_i)$ gets its value by its respective natural causes or simply set by the experimenter.

On the other hand, Assumption 3 can be seen to define what we mean by ideal, surgical interventions: the interventions do not disturb the conditional distributions of the non-intervened variables. They only break the influence the natural direct causes have on the intervened variables and determine the values of these intervened variables. This edge-breaking is shown in the manipulated model of Figure 2.2. Any edges into the intervened variables are simply cut out. This reflects our intuition: the dependence of a cause and effect persists when intervening on the cause (cause $S$ and effect $H$) but disappears when intervening on the effect (cause $W$ and effect $S$).

The causal modeling framework presented here does not always guarantee our second intuition about causality: intervening on a cause (as defined in the paragraph after Definition 1) might have absolutely no impact on the effect. In the example of Figure 2.1, if the table for $P(S|W)$ had identical rows, then intervening on $W$ would have had no effect on the values of its direct effect $S$, nor on its indirect effect $H$. The framework does guarantee that for intervening on a variable $X$ to change another variable $Y$, $X$ has to be a cause of $Y$ (Hausman and Woodward, 1999). In addition, we can say that *generally*, intervening on a cause does change its effects. The case[9] described above is in a sense pathological. The assumption of faithfulness discussed in Section 3.1.2, commonly assumed in causal discovery at least in some form, forbids such problematic situations.

### 2.2.3 Gaussian (Bayesian) Networks

Although a Bayesian network is often used with discrete variables, the same theory applies without major modifications also to continuous variables. The probability distributions $P()$ are just replaced by probability density functions $p()$. The framework describes the similarities of the probability density functions in different experimental setting. Instead of using

[9] More complicated examples of models where intervening on a cause does not influence the effect are given in Figure 3.2.

conditional probability tables that do not restrict the form of the causal relationship in the discrete case, for continuous variables we resort to using some particular parametric form for the conditional probability density functions. An often used parametric model is the following linear Gaussian form:

$$p(x_i|x_{\mathrm{pa}(x_i)}) \quad = \quad \mathcal{N}(x_i; \mu_i + \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j, \sigma_i^2), \tag{2.14}$$

where $\mathcal{N}$ denotes the density function of the normal distribution, and the parameters of the conditional probability distribution are the coefficients $b_{ij}$, mean $\mu_i$ and variance $\sigma_i^2$. When distinction to discrete variable Bayesian networks is needed, such networks are called Gaussian (Bayesian) networks (Geiger and Heckerman, 1994).

### 2.2.4 Bayesian Networks as Probabilistic Models

Bayesian networks are often used as probabilistic models, without the causal interpretation given in the previous section. Then from Definition 1, Equation 2.11 concerning interventions is simply dropped. Thus, the statistical interpretation of a non-causal Bayesian network defines only the distribution of the system in its passive observational state. This also means that the causal Markov condition (Assumption 2) turns into merely a definition that links the graph with the independence and factorization properties of the passively observed joint distribution. Any edges in the graph lose their causal meaning, they only indicate statistical dependence relations between the variables in the model. Note that such Bayesian networks, like any probabilistic models can be used to model the behavior of a system in a single experimental setting, but then a model for one experimental setting does not accurately describe the system behavior in other experimental settings.

Why then have non-causal Bayesian networks been such successful modeling tools? One answer could be that Bayesian networks are so understandable for our causally trained minds. In addition, the independence properties that can be exploited fit so well into the world of causal processes. Often when using Bayesian networks, there is a hint of the causal interpretation present: at least some edges are interpreted causally. In other fields (e.g. image processing) undirected graphical models, able to represent different kinds of independence properties, have been more useful.

## 2.3 Structural Equation Models

Another option to model the influence direct causes have on their effect is to use functional relationships and structural equations. Such *structural equation modeling* (SEM) originates from genetics (Wright, 1921) and econometrics (Haavelmo, 1943). We will consider here the interpretation of SEMs advocated by the causal discovery community and in particular Pearl (2000). Sometimes these models are also called *functional causal models.*

These models dig directly into to the idea of causal process determining the value of the effect. We will present these models using continuous variables. The workings of the processes producing a value for each variable are modeled by *deterministic* functions and structural equations:

$$\begin{array}{rcl} x_1 & := & f_1(\mathrm{pa}(x_1), e_1), \\ & \vdots & \\ x_i & := & f_i(\mathrm{pa}(x_i), e_i), \\ & \vdots & \\ x_n & := & f_n(\mathrm{pa}(x_n), e_n), \end{array} \tag{2.15}$$

where *disturbances* $e_1, \cdots, e_n$ are independent random variables responsible for making the system stochastic.

Note that we are using here the assignment sign ':=' and not the equality sign '='. This is because the equation is structural: $x_i$ is determined as a function of its direct causes and the stochastic disturbance. The mathematical use of the equality sign doesn't convey this asymmetry (Pearl, 2000). The assignment sign used here corresponds closely to the use of the assignment sign in programming languages of computer science, where a variable on the left-hand side gets its value from the formula on the right-hand sign.

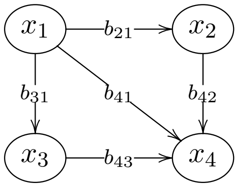
When discussing the properties of these causal models we will for simplicity focus again on acyclic models, i.e. *recursive* structural equation models. In addition, the functional relationships are constrained to be linear.

**Assumption 4 (Linearity)** *Each variable gets its value by a linear combination of its parents and an additive disturbance.*

This way the similarity to Bayesian networks is apparent.

**Definition 2 (Linear Acyclic Causal Model)** *A linear acyclic causal model consists of an acyclic graph $\mathcal{G}$ and structural equations of the type*

$$x_i \;:=\; \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j + e_i, \tag{2.16}$$



$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_{21} & 0 & 0 & 0 \\ b_{31} & 0 & 0 & 0 \\ b_{41} & b_{42} & b_{43} & 0 \end{bmatrix}$$

Figure 2.3: Linear Acyclic Causal Model. Edges in the graph $\mathcal{G}$ correspond to non-zero entries in matrix $\mathbf{B}$.

*where disturbances $e_i$ are distributed independently with some distributions $p_1(), \cdots, p_n()$. This system can be written in matrix notation as*

$$\mathbf{x} \quad := \quad \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{2.17}$$

*where $\mathbf{x} = [x_1, \ldots, x_n]^T$, $\mathbf{e} = [e_1, \ldots, e_n]^T$, and the zero entries in the coefficient matrix $\mathbf{B}$ correspond to missing edges in the graph. In the passive observational setting the model produces the distribution*

$$p(x_1, \cdots, x_n) \quad = \quad \prod_{i=1}^{n} p_i(x_i - \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j) \tag{2.18}$$

*and when variables $\mathcal{J}$ are intervened on, then the remaining passively observed variables $\mathcal{U} = \mathcal{V} \setminus \mathcal{J}$ have distribution*

$$p(\mathcal{U}) \quad = \quad \prod_{x_i \in \mathcal{U}} p_i(x_i - \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j). \tag{2.19}$$

Note that the definition leaves the exact form of the disturbance distributions $p_1, \ldots, p_n$ undefined. The parameters of the model are thus $\mathbf{B}$ and whatever parameters are used to determine the disturbance distributions. Figure 2.3 shows an example of the $\mathbf{B}$-matrix and the corresponding graph.

Much of the discussion for Bayesian networks applies directly to these recursive SEM models. In many ways the models are equivalent, different ways of representing the same ideas. However, as we will see later, SEMs allow for a neater representation of concepts such as latent confounding and cycles. The sampling process deserves further inspection. We will again sample the variables in their causal order (see Equation 2.12). One sample from the model can be generated by the following procedure:

1 (a) Sample the disturbance $e_{\mathrm{o}(1)}$ from $p_{\mathrm{o}(1)}()$.

1 (b) Determine $x_{\mathrm{o}(1)}$ from its structural equation.

$\vdots$

i (a) Sample the disturbance $e_{\mathrm{o}(i)}$ from $p_{\mathrm{o}(i)}()$.

i (b) Determine $x_{\mathrm{o}(i)}$ from its structural equation.

$\vdots$

n (a) Sample the disturbance $e_{\mathrm{o}(ni)}$ from $p_{\mathrm{o}(n)}()$.

n (b) Determine $x_{\mathrm{o}(n)}$ from its structural equation.

Note that again when determining the value of a variable, all the required elements needed to evaluate the linear function have already been sampled previously. The sampling process respects again the idea of a continuous process in space and time. Essentially the two steps at each stage of the sampling process perform the same action as the single step when sampling from the Bayesian network. The sampling produces the distribution in Equation 2.18, because Assumption 2 (Local Causal Markov Condition) holds here as well. In matrix notation, given a drawn set of disturbances **e**, **x** will get a value

$$\mathbf{x} \;=\; (\mathbf{I}-\mathbf{B})^{-1}\mathbf{e}, \tag{2.20}$$

where the matrix $(\mathbf{I}-\mathbf{B})$ is always invertible: since the model is acyclic, $\mathbf{B}$ is always permutable to a lower triangular matrix. Note that this is simply a solution of Equation 2.17.

Similarly like Bayesian networks, SEMs define also the distribution under any ideal interventions. The invariance assumption corresponds closely to Assumption 3.

**Assumption 5 (Invariance/Modularity)** *For any node $x_i \in \mathcal{V}$, the causal processes defining its value, defined by the structural equations and the distribution of the disturbance $e_1, \cdots, e_n$, remain unaltered no matter which variables of the system (other than $x_i$) are (ideally) intervened on.*

The consequence of this invariance is that any structural equation corresponding to an intervened variable $x_i \in \mathcal{J}$ is wiped out, and replaced by another structural equation that sets the value of the intervened variable to some constant $c_i$:

$$x_i \;:=\; c_i. \tag{2.21}$$

This corresponds to cutting all edges directed into the intervened variables in the associated graph $\mathcal{G}$. Thus, the experimental situation can be modeled by another SEM.

$$\begin{array}{lcll} W & := & E_W, & P(E_W) = (0.3, 0.7) \\ B & := & 1 & \\ S & := & 1 & \\ H & := & f_H(S, B, E_H), & P(E_H) = (0.3, 0.3, 0.2, 0.2) \end{array}$$

| $f_H(S,B,E_H)$ | $E_H = 0$ | $E_H = 1$ | $E_H = 2$ | $E_H = 3$ |
|---|---|---|---|---|
| $S = 0, B = 0$ | 0 | 1 | 1 | 1 |
| $S = 0, B = 1$ | 0 | 0 | 1 | 1 |
| $S = 1, B = 0$ | 0 | 1 | 0 | 1 |
| $S = 1, B = 1$ | 0 | 0 | 0 | 1 |

Figure 2.4: Structural equation model for curing a headache. The causal graph structure is the same as in Figure 2.2, and this model produces the exact same distribution as the Bayesian network in Figure 2.2.

When assuming some specific distribution for disturbances $e_i$, acyclic linear SEMs correspond to causal Bayesian networks, with conditional probability distributions defined by

$$p(x_i|\mathrm{pa}(x_i)) \quad = \quad p_i(x_i - \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j). \tag{2.22}$$

If we assume $p_i$ are univariate normal distributions, the SEM corresponds to a Gaussian (Bayesian) network (Section 2.2.3) with the same parameters. Conversely, one Bayesian network can be generally written as many different structural equation models, that nevertheless produce the exact same distribution in all ideal experimental settings. The structural equation model in Figure 2.4 corresponds to the Bayesian network in Figure 2.2 but defines additional structure. The headaches seem to come in four distinct unobserved types: those that do not need curing ($E_H = 0$), those that are cured by a painkiller ($E_H = 1$), those that are cured by sleeping ($E_H = 2$) and those that persist whatever you do ($E_H = 3$).

## 2.4 How Do the Models Answer Our Causal Inquiries?

One of the aims of causal modeling was to get insight on the complicated *structure* of the causal relationships working in a system. The causal models defined in the previous sections define this structure in an interpretable and explicit form using the associated graph. For example the Bayesian network in Figure 2.1 spells out the structure of the processes for curing a headache.

Now assume the causal process of curing a headache is modeled well enough by the model in Figure 2.1. You have a headache and you would like to know what the best action is, sleeping or taking a painkiller. Figure 2.2 gives the manipulated version of the model. Using this model we can *predict* the value of $H$ under different manipulations. The probabilities of headache persisting ($H = 1$) under manipulations of $S$ and $B$ (denoted here by '$||$' to distinguish from plain conditioning) are the following:

$$\begin{aligned} P(H = 1||S = 0, B = 0) &= 0.7, \\ P(H = 1||S = 0, B = 1) &= 0.4, \\ P(H = 1||S = 1, B = 0) &= 0.5, \\ P(H = 1||S = 1, B = 1) &= 0.2. \end{aligned}$$

Clearly, taking a painkiller and sleeping seems the best course of action in this case.

Causal models can in some cases *explain* the outcomes of certain events. Say again you got a headache and you took a painkiller and slept for one hour. This chain of events is given by the structural equation model in Figure 2.4. Say that after the course of action the headache was gone ($H = 0$). Was it the sleeping or the painkiller that cured your headache? Thus if you would have just slept or just taken a painkiller would the headache still be present? Pearl (2000) describes a procedure for answering this type of counterfactual queries given a structural equation model. We simply update the disturbance distribution of $E_H$ (in this case $E_W$ is irrelevant) by the observed evidence using the Bayes-formula:

$$\begin{aligned} P(E_H|S = 1, B = 1, H = 0) &= \frac{P(S = 1, B = 1, H = 0|E_H)P(E_H)}{\sum_{E'_H} P(S = 1, B = 1, H = 0|E'_H)P(E'_H)} \\ \Rightarrow P(E_H|S = 1, B = 1, H = 0) &= (0.375, 0.375, 0.2, 0). \end{aligned}$$

Clearly, the headache wasn't the incurable type $E_H = 3$. Then the probability of the headache can be evaluated under different actions using this updated disturbance probability distribution:

$$\begin{aligned} P'(H = 1||S = 0, B = 0) &= 0.375, \\ P'(H = 1||S = 0, B = 1) &= 0.375, \\ P'(H = 1||S = 1, B = 0) &= 0.2, \\ P'(H = 1||S = 1, B = 1) &= 0.0. \end{aligned}$$

Thus, with a probability of 37.5% your headache would have been cured without any actions, with a 37.5% probability your headache would have been cured by the painkiller alone, and with 20% probability sleeping was the effective measure taken.

# Chapter 3

# Causal Discovery

This chapter discusses the learning of causal models from data, collected from a system in its *natural passive observational state* under certain simplifying assumptions. Although many methods learn a full causal model of the form described in the previous section, we will mostly focus on the subproblem of *structure discovery*: finding the *directed* causal graph structure of the underlying model, ignoring the parameters defining the causal relations. This is because the structure of the causal relations is often what interests us the most. Finding the structure is also perhaps the most cumbersome part.

An inescapable fact is that we are very rarely able to observe all the important variables. In a sense, we are always observing only a subset of variables involved in the data generating process. Nevertheless, we would like to understand the causal structure among the variables we have observed; what happens when intervening on a quantity that is not measured is not in our immediate interest. Fortunately, the causal models under consideration are closed under marginalization of variables connected with many types of structures (see Figure 3.1 (a-c)). This means that if the data generating process over some 'original' set of variables can be described by a causal model in a given class, there is also a causal model in the same model class accurately describing the causal relations over a subset of the original variables.

However, in Figure 3.1 (d) the situation is more difficult: an unobserved variable $U$ confounds $X$ and $Y$, and is thus called a *latent confounder*. Because of the confounder $X$ and $Y$ are found dependent in the marginalized distribution $P(X, Y)$. The dependence should disappear when intervening on $X$ and when intervening on $Y$. This is not possible by the simple model structures used in Section 2, if only the observed variables $X$ and $Y$ are considered in the model. The assumption of causal sufficiency denies the

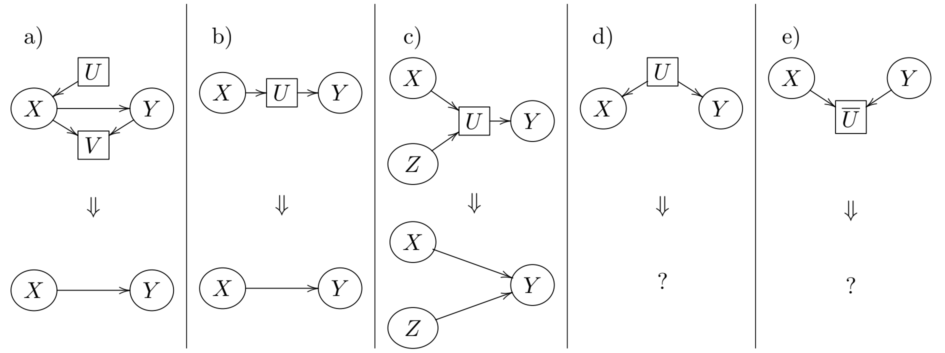


Figure 3.1: Marginalizing the structure. If the variables in the squares happen to be unobserved, in the first three cases (a), (b) and (c) this is entirely unproblematic: the structure learned is still a valid causal structure, just among the observed variables. The latent confounder in (d) and selection bias in (e) cannot be modeled by simple directed graphs.

existence of such problematic variables.

**Assumption 6 (Causal Sufficiency)** *There are no unobserved common causes (latent confounders) of two or more of the observed variables.*

As we will see, this assumption simplifies the learning methods a great deal. In Chapter 5 (Latent Confounding) the possibilities of learning causal models without such a restrictive assumption are considered.

Another type of variable troublesome when unobserved is $U$ in Figure 3.1 (e). Note that such a variable is only troublesome when conditioned on, as the variable $V$ connected with a similar structure in Figure 3.1 (a) is not problematic: it does not affect the distribution $P(X, Y)$ in any way. Thus, the following simplifying assumption is often made as well.

**Assumption 7 (No Selection Bias)** *No common effects of two or more of the variables*[1] *are conditioned on.*

For selection bias to occur, values of the variables influence whether the sample is included in the data set or not. In many cases this can be ruled out by the properties of the data collection process.

Thus, in this chapter we will cover the most basic algorithms for causal discovery from passive observational data, assuming causal sufficiency and

[1] The variables are here either observed variables or latent confounders (if their presence is allowed).

| | | | |
|---|---|---|---|
| parametric restrictions | | Score-Based with Linear Gaussian CPDs (Section 3.2) | `LiNGAM`, Additive Noise Models etc. (Section 3.3) |
| non-parametric | `PC` (Section 3.1) | | Score-Based with Discrete Variables (Section 3.2) |
| | independence relations | 2nd order statistics | full distribution |

Table 3.1: One classification of causal discovery methods when assuming causal sufficiency.

the absence of selection bias. One way of classifying the different methods is along the different parametric assumptions on the data generating model and the extent to which the joint distribution is exploited (Table 3.1). Some discovery methods make some assumptions about the parametric forms of the causal relations and the way noise affects the system, some try to manage without making any such restricting assumptions. On the whole these parametric restrictions aid in causal discovery, but sometimes diminish the applicability of the algorithms. Often data sets with discrete variables can be analyzed without restrictions, but with continuous variables some restrictions are commonly made.

The different methods exploit different aspects of the observed distribution. Some methods exploit the independence relations detected in the data, others use only 2nd order statistics (mean and covariance information), some exploit the whole distribution. Generally, more samples are needed for accurately describing the more intricate structure in the distribution, such as higher order statistics. This more detailed structure can in some cases be used to identify the causal structure uniquely, whereas methods using only independence relations may leave (part of) the structure underdetermined.

The chapter is divided into three parts: first we will discuss finding causal models from detected independence relations in Section 3.1 (Constraint-based Approach). Then, Section 3.2 (Score-based Approach) describes a more Bayesian approach based on calculating posterior probabilities for graph structures. Finally, we will present methods that make up a third category in Section 3.3 (Exploiting Higher Moments in Continuous Data) and use higher order statistics of continuous data together with assuming parametric restrictions on the causal relations.

## 3.1 Constraint-based Approach

As perhaps already hinted at in Chapter 2 (Causal Modeling), lack of causal relations in causal models produces conditional independence relations in the generated distributions. Thus, independence relations observed in the generated distribution may allow us to infer the absence of some causal relations. On the other hand, any dependence is an indication of structures able to produce such a dependence. By considering the independence and dependence relations between a larger group of variables, we might be capable of narrowing down the set of possible causal structures. This is the underlying idea of constraint-based causal discovery. In the following, the data generating process is assumed to be a Bayesian network or a recursive structural equation model, and the variables may be discrete or continuous.

### 3.1.1 From Graphs to Independence Relations

Which independence relations do the models of Section 2 then produce? Some independence relations are always observed in the generated distribution as a consequence of Assumption 2 (Local Causal Markov Condition). However, Assumption 2 implies also additional independence properties that are not explicitly given by its definition. For example, any causal model with the structure in Figure 2.1 will necessarily yield the independence[2] $W \perp\!\!\!\perp B \mid S, H$. This independence relation is not given by Assumption 2 explicitly since the conditioning set $\{H, S\}$ is not the parent set of either $W$ nor $B$.

Thus, let us aim for a sufficient condition on the causal graph structure of the true data generating model, such that a conditional independence statement $X \perp\!\!\!\perp Y \mid \mathcal{C}$ holds in the generated distribution. First, the condition should not hold between the observed variables in the causal structures of Figure 1.1 (p. 4). Second, the condition should imply the independence relations entailed by the local causal Markov condition both explicitly and implicitly. A concept called d-separation (d for directed) has been introduced for this purpose (Pearl, 1988).

[2] By a straight-forward calculation one can verify that $P(W|S, B, H) = P(W|S, H)$.

| | $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathcal{C}$ | $X \not\perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathcal{C}$ |
|---|---|---|
| $X \perp\!\!\!\perp Y \mid \mathcal{C}$ | OK | UNFAITHFUL |
| $X \not\perp\!\!\!\perp Y \mid \mathcal{C}$ | NOT POSSIBLE | OK |

Table 3.2: Correspondence between the graphical criterion of d-separation on graph $\mathcal{G}$ and conditional independence relations in the probability distribution generated by a model with causal structure $\mathcal{G}$. Models producing extra independence relations that are not consequences of the global causal Markov condition are labeled as unfaithful.

**Definition 3 (D-separation)** *An (undirected) path $p$ between nodes $X$ and $Y$ is said to be blocked (or d-separated) by a set of nodes $\mathcal{C} \subseteq \mathcal{V} \backslash \{X, Y\}$ if and only if $p$ contains*

- *a chain $U \to Z \to V$ such that $Z \in \mathcal{C}$, or*
- *a fork $U \leftarrow Z \to V$ such that $Z \in \mathcal{C}$, or*
- *a collider $U \to Z \leftarrow V$ such that neither $Z$ nor any effect of $Z$ are in $\mathcal{C}$.*

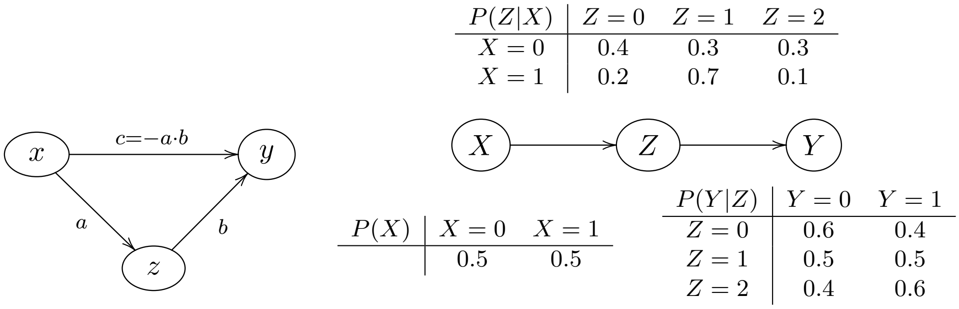
*Nodes $X$ and $Y$ are said to be d-separated by set $\mathcal{C}$, denoted by $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathcal{C}$, if and only if all paths between them are d-separated by the set $\mathcal{C}$.*

Given this graphical definition we can reformulate the local Markov condition (Assumption 2) in an alternative form.[3]

**Assumption 8 (Global Causal Markov Condition)** *If two variables $X$ and $Y$ are d-separated by a set $\mathcal{C}$ in the graph describing the true causal relations, then they are independent in the generated distribution given $\mathcal{C}$.*

Table 3.2 shows further implications of the satisfaction and dissatisfaction of the d-separation condition. If the d-separation condition does not apply for a pair with respect to a conditioning set, this does not yet guarantee that the corresponding dependence is found in the generated distribution. But, for every graph structure there exists a causal model (Bayesian network) that produces only the independencies given by the d-separation condition and no other (Pearl, 1988). The models that generate distributions with independence relations that are not implications of the global Markov condition are termed unfaithful and assumed unlikely models for generating

[3] Assuming the causal models of Section 2, one can prove that independence follows from d-separation. As an alternative to Assumption 2, the global causal Markov condition is an assumption on reality.



| $P(Z\|X)$ | $Z=0$ | $Z=1$ | $Z=2$ |
|---|---|---|---|
| $X=0$ | 0.4 | 0.3 | 0.3 |
| $X=1$ | 0.2 | 0.7 | 0.1 |

| $P(X)$ | $X=0$ | $X=1$ |
|---|---|---|
| | 0.5 | 0.5 |

| $P(Y\|Z)$ | $Y=0$ | $Y=1$ |
|---|---|---|
| $Z=0$ | 0.6 | 0.4 |
| $Z=1$ | 0.5 | 0.5 |
| $Z=2$ | 0.4 | 0.6 |

Figure 3.2: Two unfaithful models. On the left a linear model where the paths from $x$ to $y$ cancel out exactly, thus $x \perp\!\!\!\perp y$. On the right the influence of $X$ on $Y$ is completely randomized by variable $Z$, thus $X \perp\!\!\!\perp Y$.

the observed data (see Section 3.1.2 for justification of this). This characterization of independence properties can be trivially extended to the experimental distribution produced by a manipulated model, by examining the corresponding manipulated graph.

### 3.1.2 The Faithfulness Assumption

If two variables are found to be independent in some distribution, we are often tempted to draw the conclusion that there is no causal influence between the variables.[4] This sort of deduction can also be seen as a version of the Occam's razor principle: if several models fit the data equally well, choose the simplest one (see Rasmussen and Ghahramani (2001)). If there is no indication of a causal relation between some pair variables in the data, the simplest model fitting the data postulates that the causal connection is not there. The assumption generalizing such deductions, commonly made in all fields of science at least in some form, can be formalized here nicely using global causal Markov condition (Pearl, 1988, 2000; Spirtes et al., 1993).

**Assumption 9 (Faithfulness)** *A causal model is faithful if all (conditional) independence relations in the probability distribution produced by the model are consequences of the global causal Markov condition.*

Assuming faithfulness of the underlying data generating model ensures that if the variables are independent in the generated distribution, there is no

[4] The overlying idea of faithfulness might be easily remembered as the inadequate but useful slogan 'no correlation implies no causation'.

unblocked path in the true causal graph structure (Table 3.2). This is in converse to the causal Markov condition (Assumption 8) which ensures the existence of an unblocked path between any variables observed to be dependent in the distribution. For faithful models the concepts of d-separation and conditional independence are equivalent, this allows us to infer causal structure from detected independence relations.

Another interpretation of faithfulness states that the independence relations present in the generated distribution do not disappear with (small) perturbation of the model parameters (while keeping the causal structure fixed). Thus, the independence relations are consequences of the model structure. This interpretation leads to measure theoretic reasons to assume faithfulness: under many sampling procedures determining the parameters of causal model with a fixed structure, drawing an unfaithful model constitutes a set of measure 0 (Spirtes et al., 1993; Meek, 1995b). This means that if we draw the parameters of the model randomly, as opposed to carefully fixing them by hand, we will never create unfaithful models. How much such a measure theoretic claim says about causal systems in nature is debatable.

Figure 3.2 shows two 'pathological' examples of models that violate the faithfulness assumption. The parameters for these two models are handpicked to get a non-structural independence relation $X \perp\!\!\!\perp Y$ (or $x \perp\!\!\!\perp y$). If the parameters are perturbed even a little bit, this independence relation disappears. Both cases are examples of a more general setting where parameters are not independently drawn at random, instead, they have equality constraints (such as $c = -a \cdot b$). Also, if the variables are deterministically (instead of stochastically or probabilistically) related, faithfulness should not be assumed (Richardson and Spirtes, 1999).

Although the causal model can often be safely assumed to be faithful, finite sample data may appear to contain (conditional) independence relations not due to the structure of the model. This *effective unfaithfulness* leads to errors when detecting independence relations in sample data. Glymour et al. (1999) call the assumption denying such problematic models and data sets the 'sample causal faithfulness assumption'. Often large, dense models may be effectively unfaithful: almost exact canceling of the unblocked paths between variables often occurs for some of the $2^{n-2}$ possible conditioning sets. Methods not assuming faithfulness are better suited for such situations.

### 3.1.3 The Hunt for an Independence Oracle

The technical part of the puzzle concerns detecting the independence relations in the observed data. For deriving some of the theoretical results and explaining the algorithms, it is useful to assume we posses an *independence oracle*: from the data produced by a causal model we can infer without error the truth value of the statement $X \perp\!\!\!\perp Y \mid \mathcal{C}$. However, detecting independence relations from sample data is not a totally straight-forward task.

Currently, perhaps the best candidate for an independence oracle is traditional statistical testing. In such tests we usually have two competing hypotheses:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid \mathcal{C}, \qquad \mathcal{H}_1 : X \not\!\perp\!\!\!\perp Y \mid \mathcal{C},$$

one hypothesizing independence and one hypothesizing dependence. Then, a test statistic $t(\mathbf{X})$, a function of the generated data $\mathbf{X}$, is formulated measuring how incompatible the data is with the hypothesis $\mathcal{H}_0$. The distribution of the test statistic $t(\mathbf{X})$ is derived under $\mathcal{H}_0$, often exploiting the Central Limit Theorem. Now, if the test statistic $t(\mathbf{X})$ calculated from the data $\mathbf{X}$ lands on an unlikely region in the distribution of $p(t|\mathcal{H}_0)$, then the null-hypothesis $\mathcal{H}_0$ may be rejected as implausible. This unlikely region is judged by some assigned significance level.

One caveat here is that in causal discovery based on faithfulness, we need to recognize *independence* relations and an orthodox statistician might argue that the statistical test only gives evidence for significant *dependence* relations. When the test statistic happens to land on the likely region according to $\mathcal{H}_0$, we just don't have any evidence against $\mathcal{H}_0$. One can then resort to the Occam's razor principle: one should not assume a more complicated model if a simpler model explains the data. Here the structure corresponding to $\mathcal{H}_0$ is simpler than the structure corresponding to $\mathcal{H}_1$. Not obtaining any evidence of dependence is perhaps sufficient for accepting $\mathcal{H}_0$ in many cases. We can still consider what would happen if we would only be allowed to reject $\mathcal{H}_0$ and never accept it. We would not be able to detect any independence relations, and thus, we would never be able to reject full (acyclic) graphs as the data generating structure. The output from a learning algorithm would not be very informative.

With finite number of samples the independence tests are bound to produce at least some errors. A type I error occurs when the variables are independent but the test indicates that they are dependent. The probability of this type of errors is fixed by the significance level. A type II error occurs when the variables are in reality dependent but the statistical test indicates independence. This is a more troublesome error type for causal discovery, as mistaken independence judgments may result in deleting important edges from the learned graph structure. In addition, the rate of errors of type II cannot be directly controlled. Causal discovery algorithms may also be prone to various multiple testing complications. When the number of data samples increase towards infinity, the independence test works like an independence oracle: when the significance threshold is lowered systematically with increasing sample size, the probabilities of both types of error approach zero (Richardson and Spirtes, 1999).

Different independence tests have been designed for various different situations. When testing $X \perp\!\!\!\perp Y \mid \mathcal{C}$ with discrete variables, an often used independence test is the Pearson's $\chi^2$-test for goodness of fit. Let $N_{ijk}$ denote the number of samples when $X$ gets its $i$:th value, $Y$ gets its $j$:th value and the variables in the conditioning set $\mathcal{C}$ get their $k$:th configuration. Under $\mathcal{H}_0$ the conditional probability distribution $P(X, Y|\mathcal{C})$ should factorize as $P(X|\mathcal{C})P(Y|\mathcal{C})$. Thus, we can calculate the *expected* number of samples of each configuration of $X$, $Y$ and variables in $\mathcal{C}$ under $\mathcal{H}_0$ using the sample probability estimates:

$$E_{ijk} \quad := \quad \frac{\sum_{j'} N_{ij'k}}{\sum_{i',j'} N_{i'j'k}} \cdot \frac{\sum_{i'} N_{i'jk}}{\sum_{i',j'} N_{i'j'k}} \cdot \sum_{i',j'} N_{i'j'k}. \tag{3.1}$$

The test statistic measuring the deviation of the observed counts $N_{ijk}$ from the expected counts $E_{ijk}$ used is then the following:

$$t \quad = \quad \sum_{i,j,k} \frac{(N_{ijk} - E_{ijk})^2}{E_{ijk}}. \tag{3.2}$$

Under $\mathcal{H}_0$ and for sufficiently large sample sizes, the test statistic $t$ will be distributed as $\chi^2$ with $(m_X - 1)(m_Y - 1)m_\mathcal{C}$ degrees of freedom, where $m_X$ is the number of categories for $X$, $m_Y$ is the number of categories for $Y$ and $m_\mathcal{C}$ is the number of configurations for the conditioning set $\mathcal{C}$.

When testing $x \perp\!\!\!\perp y \mid \mathcal{C}$ for continuous variables, it is common to fit a linear regression model explaining $y$ with $x$ and the variables in the conditioning set $\mathcal{C}$. Then assuming Gaussian conditional probability distributions, testing conditional dependence $x \perp\!\!\!\perp y \mid \mathcal{C}$ corresponds to testing the significance of the regression coefficient of $x$ with the commonly used t-test. There exist also independence tests that aim to detect more complicated, non-linear dependencies between continuous variables (Gretton et al., 2008; Zhang et al., 2011).

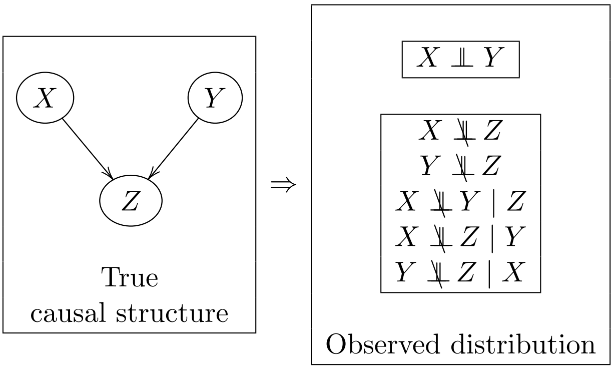

X
Y
Z
True
causal structure
⇒
$X \perp\!\!\!\perp Y$
$X \not\perp\!\!\!\perp Z$
$Y \not\perp\!\!\!\perp Z$
$X \not\perp\!\!\!\perp Y \mid Z$
$X \not\perp\!\!\!\perp Z \mid Y$
$Y \not\perp\!\!\!\perp Z \mid X$
Observed distribution
⇒
X
Y
Z
Learned
causal structure

Figure 3.3: Causal discovery from independence and dependence relations detected in passive observational data, assuming causal sufficiency. The d-separation condition implies only one independence, and by assuming faithfulness of the data generating model, this is the only independence present in the passively observed distribution. In this particular case there are no other (faithful) models with different structure that would produce exactly these independence and dependence relations.

### 3.1.4 From Independence Relations to Graphs

It should now be clear that different causal structures produce different independence and dependence properties in the joint, passively observed distribution. Figure 3.3 shows an example setting where the learning works out nicely, we are able to detect a causal structure just from the independence relations in passive observational data. First, we can read off the independence facts from the true causal structure on the left using the d-separation criterion (Definition 3). Only one independence, $X \perp\!\!\!\perp Y$ is present, and (assuming faithfulness) for all other pairs and conditioning sets we have dependence. This is shown in the middle. Then, consider not knowing the true structure on the left, but only the independence facts in the middle. One way of deducing the structure from independence fact goes as follows. Since $X$ and $Y$ were marginally independent, according to faithfulness there should not be edges between them. Since $X$ and $Z$ are marginally dependent, there should be a path between them. As this path cannot go through $Y$, we must have an edge between $X$ and $Z$, and symmetrically between $Y$ and $Z$. Finally, since $X$ and $Y$ were marginally independent, both edges must be oriented towards $Z$, otherwise the independence would be the unfaithful kind. Note that this simple structure actually allows orienting the causal edges. The structure is called an *unshielded collider* (or a v-structure) and it plays an important part in struc-

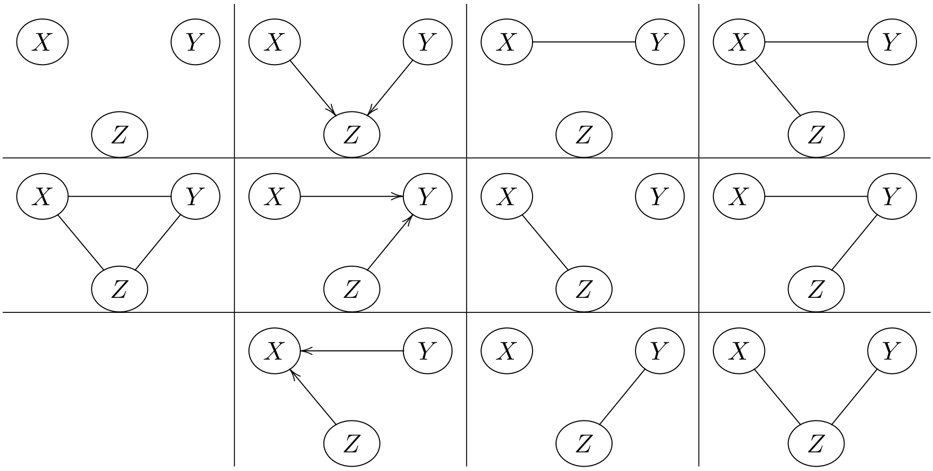


Figure 3.4: All 11 Markov equivalence classes of the 25 directed acyclic graphs (DAGs) over three variables, represented by partially oriented DAGs. Forming the actual, fully oriented DAGs in each equivalence class is easy: the unoriented edges of the partially oriented DAG may be oriented in either direction, as long as no new unshielded colliders or cycles are formed. Here, the unoriented edges should not be confused with undirected edges used for example in undirected graphical models.

ture discovery from independence relations. An edge $X \rightarrow Y$ would in a way 'shield' this collider, and there would no longer be any independence relations present in the data.

More commonly, we are only able to recognize a possible group of graph structures; such a group is called a *Markov equivalence class.* Graphs inside a Markov equivalence class produce the exact same sets of dependence and independence relations. Two graphs belong to the same equivalence class if and only if they have 1) the same skeleton (unoriented structure) and 2) the same set of unshielded colliders (Pearl, 2000). Figure 3.4 shows all such equivalence classes for graphs with three nodes. The equivalence classes are represented by partially oriented DAGs: any unoriented edges can be oriented in either direction as long as no directed cycles or additional unshielded colliders are formed.

The characterization of equivalence classes suggest a two phase procedure (Pearl, 2000) for searching for causal structures compatible with the data. First, we try to find the *skeleton*: the causal structure without the edge orientations. In fact, we can determine the existence of any edge given an independence oracle: conditioning for example on all true parents of both variables (variables themselves excluded), the variables are dependent if and only if there is an edge between them. Thus, if the edge is not present and we test all possible conditioning sets, we will eventually find an independence. After finding the skeleton, we continue to orient edges by adding v-structures where the detected independence and dependence relations indicate. For this second phase there exists four rules that are necessary and sufficient for orienting all edges that have the same orientation in a equivalence class (Meek, 1995a). Such an algorithm is thus *complete*: all possible information in the independence relations is utilized.

In practice, with finite sample data, the situation is a little more complicated. The order in which the independence queries are performed can greatly impact the learned structure. Spirtes et al. (1993) define an optimal strategy in their `PC`-algorithm (the authors' first names are Peter and Clark). First we can run all marginal independence tests, then conditional tests with singleton conditioning sets and so on. The subsequent tests with larger conditioning sets are run only when needed. The tests with larger conditioning sets tend to be less reliable. In addition, if we can place a limit on the (undirected) degree of nodes in the true graph, the `PC`-algorithm will have a polynomial time complexity (Spirtes et al., 1993).

The good thing about constraint-based causal discovery is that it can be made completely non-parametric, thus we do not have to assume any parametric restrictions on the causal relationship. On the other hand the algorithms are based on hard judgments of dependence and independence. If the judgments go wrong, the recovered structure may be far from the truth. Often any constraint-based algorithms should be run several times with different significance levels for the independence tests, in order get some idea on which alternative structures are possible.

## 3.2 Score-based Approach

In Bayesian score-based[5] causal discovery we assume a Bayesian network $(\mathcal{G}, \boldsymbol{\theta})$ as the data generating process and try to assign posterior probabilities $P(\mathcal{G}|\mathbf{X})$ (or, more generally, scores) to different graph structures $\mathcal{G}$ given some data $\mathbf{X}$ obtained in the passive observational setting. The data $\mathbf{X}$ is assumed here to be in $N \times n$ matrix form with each of the $N$ samples in rows while the different columns represent the $n$ observed variables.

[5] A score-based approach may also be non-Bayesian, then instead of posterior probability, the score is considered to be some other function of the model and the data.

### 3.2.1 Derivation of the Posterior Probability

The calculation of the posterior probability $P(\mathcal{G}|\mathbf{X})$ is fairly complicated (Cooper and Herskovits, 1992). The first step is to invert the conditional probability using Bayes-formula:

$$P(\mathcal{G}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathcal{G})P(\mathcal{G})}{P(\mathbf{X})}, \text{ where} \tag{3.3}$$

$$P(\mathbf{X}) = \sum_{\mathcal{G}} P(\mathbf{X}|\mathcal{G})P(\mathcal{G}). \tag{3.4}$$

Here $P(\mathcal{G})$ is a prior probability of the graph structure $\mathcal{G}$. Often this is taken to be uniform $P(\mathcal{G}) \sim 1$ for all considered graph structures, since no graph structure is favored before observing the data. The likelihood of the data $P(\mathbf{X})$ can be calculated by summing the numerators of Equation 3.3 over all considered graph structures (Equation 3.4).

The term $P(\mathbf{X}|\mathcal{G})$ is the marginal likelihood of the data: the probability of drawing data $\mathbf{X}$ from a model with graph structure $\mathcal{G}$. In order to calculate this, we have interpret the parameters $\boldsymbol{\theta}$ of the model as random variables. Then, we integrate over the parameters $\boldsymbol{\theta}$, and use the product rule of probabilities (Equation 2.5, p. 11):

$$P(\mathbf{X}|\mathcal{G}) = \int P(\mathbf{X}, \boldsymbol{\theta}|\mathcal{G})d\boldsymbol{\theta} = \int P(\mathbf{X}|\boldsymbol{\theta}, \mathcal{G})P(\boldsymbol{\theta}|\mathcal{G})d\boldsymbol{\theta}. \tag{3.5}$$

Here $P(\mathbf{X}|\boldsymbol{\theta}, \mathcal{G})$ is simply the likelihood of the data given the fully defined model $(\mathcal{G}, \boldsymbol{\theta})$, while the term $P(\boldsymbol{\theta}|\mathcal{G})$ is the prior distribution on the model parameters $\boldsymbol{\theta}$. The notation $d\boldsymbol{\theta}$ represents $d\theta_1 d\theta_2 \ldots$, i.e. the integral is taken over all individual parameters.

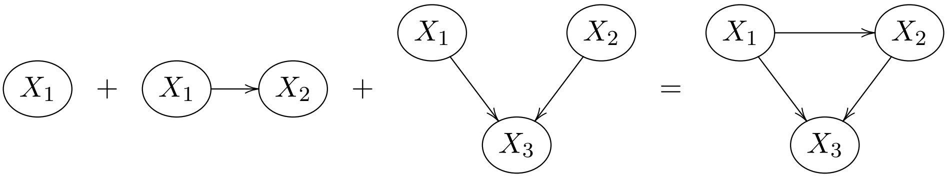
Since the samples, appearing in the rows of the data matrix $\mathbf{X}$, are independently and identically distributed we have that

$$P(\mathbf{X}|\boldsymbol{\theta}, \mathcal{G}) = \prod_{j=1}^{N} P(\mathbf{X}_{j\cdot}|\boldsymbol{\theta}, \mathcal{G}). \tag{3.6}$$

The probability of a single sample $\mathbf{X}_{j\cdot}$ given a uniquely defined model $(\mathcal{G}, \boldsymbol{\theta})$ can be calculated[6] from Equation 2.10 (p. 13):

$$P(\mathbf{X}_{j\cdot}|\boldsymbol{\theta}, \mathcal{G}) = \prod_{i=1}^{n} P(\mathbf{X}_{ji}|\mathbf{X}_{j,\mathrm{pa}(X_i)}, \boldsymbol{\theta}_i). \tag{3.7}$$

[6] With slight abuse of notation $\mathbf{X}_{j,\mathrm{pa}(X_i)}$ denotes the values the parents of $X_i$ receive in the $j$:th sample.



$$\text{score}(X_1, \emptyset) + \text{score}(X_2, \{X_1\}) + \text{score}(X_3, \{X_1, X_2\}) \quad = \quad \log P(\mathbf{X}|\mathcal{G})$$

Figure 3.5: Combining local scores to calculate the posterior probability of a full graph structure.

Here, the parameters $\boldsymbol{\theta}$ are divided into $n$ groups $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, such that $\boldsymbol{\theta}_i$ defines the local probability distribution $P(X_i|\text{pa}(X_i))$. In this expression the graph $\mathcal{G}$ defines the parent sets of each variable $\text{pa}(X_i)$. Thus, overall we get:

$$P(\mathbf{X}|\boldsymbol{\theta}, \mathcal{G}) \quad = \quad \prod_{j=1}^{N} \prod_{i=1}^{n} P(X_i = \mathbf{X}_{ji}|\mathbf{X}_{j,\text{pa}(X_i)}, \boldsymbol{\theta}_i). \qquad (3.8)$$

In order to further factorize and simplify Equation 3.5, it is useful if the prior factorizes similarly as $P(\mathbf{X}|\boldsymbol{\theta}, \mathcal{G})$. Heckerman et al. (1995) call this property *parameter independence*. The different parameter groups, each defining one conditional probability distribution, are independent given the graph structure:

$$\boldsymbol{\theta}_i \perp\!\!\!\perp \boldsymbol{\theta}_j \mid \mathcal{G}.$$

In addition, each parameter group $\boldsymbol{\theta}_i$ depends only on the local structure: the parents of node $X_i$, denoted by $\text{pa}(X_i)$. The other structure in $\mathcal{G}$ is irrelevant to the parameters $\boldsymbol{\theta}_i$. With these assumptions, the prior $P(\boldsymbol{\theta}|\mathcal{G})$ factorizes as follows:

$$P(\boldsymbol{\theta}|\mathcal{G}) \quad = \quad \prod_{i=1}^{n} P(\boldsymbol{\theta}_i|\mathcal{G}) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_i|\text{pa}(X_i)). \qquad (3.9)$$

Using Equations 3.8 and 3.9 the likelihood of the data given a structure in Equation 3.5 can be written as the integral:

$$P(\mathbf{X}|\mathcal{G}) \quad = \quad \int \prod_{i=1}^{n} \prod_{j=1}^{N} P(\mathbf{X}_{ji}|\mathbf{X}_{j,\text{pa}(X_i)}, \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i|\text{pa}(X_i)) d\boldsymbol{\theta}_1 \ldots d\boldsymbol{\theta}_n.$$

First consider integrating over parameters $\boldsymbol{\theta}_1$. In the product over index $i$, only the first term depends on the parameters $\boldsymbol{\theta}_1$. The rest can be taken out of the integral over $\boldsymbol{\theta}_1$ as constants. Similar simplification applies to the integrals over all parameter groups. Thus, given the obtained neat factorizations, the total integral can be calculated as $n$ separate integrals:

$$P(\mathbf{X}|\mathcal{G}) = \prod_{i=1}^{n} \int \prod_{j=1}^{N} P(\mathbf{X}_{ji}|\mathbf{X}_{j,\mathrm{pa}(X_i)}, \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i|\mathrm{pa}(X_i)) d\boldsymbol{\theta}_i. \quad (3.10)$$

The log-value of the integral for a given $i$ and local graph substructure $\mathcal{G}_i$ is interpreted as a *local score*. The logarithm of the likelihood can be calculated as the sum of the local scores (Figure 3.5):

$$\mathrm{score}(X_i, \mathrm{pa}(X_i)) = \log \int \prod_{j=1}^{N} P(\mathbf{X}_{ji}|\mathbf{X}_{j,\mathrm{pa}(X_i)}, \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i|\mathrm{pa}(X_i)) d\boldsymbol{\theta}_i, \quad (3.11)$$

$$\log P(\mathbf{X}|\mathcal{G}) = \sum_{i=1}^{n} \mathrm{score}(X_i, \mathrm{pa}(X_i)). \quad (3.12)$$

For each variable $X_i$, the local score in Equation 3.11 depends only on the local structure, defined by $\mathrm{pa}(X_i)$, of the associated graph. Thus, the calculated value of a single local score can be reused for several different full graph structures that share the same local structure. Heckerman et al. (1995) call this property *parameter modularity.* This important property makes calculation of the posterior probabilities of several structures somewhat tractable. Once we have obtained the likelihoods $P(\mathbf{X}|\mathcal{G})$ for all graph structures, the posterior probabilities can be calculated from Equations 3.3 and 3.4.

### 3.2.2 Local Scores

The local scores can be calculated in closed form for discrete variables when 1) local conditional probability distributions are multinomial, and 2) the parameters $\boldsymbol{\theta}_i$ have Dirichlet (conjugate prior for multinomial) distributions (Cooper and Herskovits, 1992). If $N_{ijk}$ marks the number of times the variable $X_i$ gets its $k$:th value, while its parents get their $j$:th configuration, the local score is:

$$\mathrm{score}(X_i, \mathrm{pa}(X_i)) = \log \left( \prod_j \frac{\Gamma(\sum_k \alpha_{ijk})}{\Gamma(\sum_k \alpha_{ijk} + N_{ijk})} \prod_k \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \right), \quad (3.13)$$

where $\Gamma$ denotes the gamma function, and $\alpha_{ijk}$ are hyper-parameters for the prior Dirichlet distributions. One common option here is to use the so-called

BDeu-prior (Bayesian Dirichlet with likelihood Equivalence, Uniform):

$$\alpha_{ijk} \quad = \quad \frac{\alpha}{m_{X_i} \cdot m_{\mathrm{pa}(X_i)}}, \tag{3.14}$$

where $m_{X_i}$ denotes the number of categories for $X_i$ and $m_{\mathrm{pa}(X_i)}$ denotes the number of configurations of parents, and $\alpha$ is some given hyper-parameter (for example $\alpha = 1$). Note that this formulation is particularly nice: we are not restricting the parametric form of the conditional probability distributions and still the posterior probabilities can be calculated without approximation errors. If the discrete variables are not categorical, but for example ordinal, other priors may exhibit better properties.

For continuous variables with linear Gaussian conditional probability distributions, the local score can also be calculated in closed form using a particular application of the Normal-Wishart prior (Geiger and Heckerman, 1994). In other cases we can resort to approximate the integral numerically.

Note that both local scores presented here satisfy a property of *likelihood-equivalence*: graphs in the same Markov equivalence class (Section 3.1.4) receive the same overall score. This seems natural with multinomial and linear-Gaussian conditional probability distributions: if two graph structures can produce the exact same distributions, then no data should help in discriminating them (Heckerman et al., 1995).

### 3.2.3 Search over Possible Structures

In the true Bayesian fashion, any inference problem should be solved by marginalizing over all possible network structures. We might for example calculate the confidence of $X$ being a cause of $Y$ given the data $\mathbf{X}$. We can utilize the fact that '$X$ causes $Y$' is independent of the data given the causal graph structure $\mathcal{G}$:

$$\begin{aligned} P(\text{`}X \text{ causes } Y\text{'}|\mathbf{X}) &= \sum_{\mathcal{G}} P(\text{`}X \text{ causes } Y\text{'}, \mathcal{G}|\mathbf{X}) \\ &= \sum_{\mathcal{G}} P(\text{`}X \text{ causes } Y\text{'}|\mathcal{G}) P(\mathcal{G}|\mathbf{X}), \end{aligned}$$

where $P(\text{`}X \text{ causes } Y\text{'}|\mathcal{G}) = 1$ whenever $X$ is an ancestor of $Y$ in graph $\mathcal{G}$, otherwise 0. The posterior probabilities $P(\mathcal{G}|\mathbf{X})$ for all graph structures have been obtained as described in the previous sections. Unfortunately, this approach is often intractable due to the high number of graphs. One option is then to use, for example, a few of the highest scoring structures for the inference.

Instead of model averaging we may have to resort to model selection. Often the most informative structure is the one that maximizes the overall score. This structure is called the MAP structure (Maximum a Posteriori). Thus, given pre-calculated local scores, we try to find the structure $\mathcal{G}$ solving

$$\max_{\mathcal{G}} P(\mathcal{G}|\mathbf{X}). \tag{3.15}$$

Note that we do not need to calculate the divisor $P(\mathbf{X})$ in Equation 3.3 for finding this maximum, as it is only a normalizing factor and a constant for all graphs. Even finding this MAP structure is unfortunately NP-hard (Chickering, 1996). One complicating issue here is the acyclicity assumption, because it is a *global* assumption. The highest scoring structure cannot be found efficiently considering only the local neighborhoods separately. Current state-of-the-art exact structure discovery algorithms find the MAP structure using dynamic programming and other optimization techniques (Koivisto and Sood, 2004; Silander and Myllymäki, 2006; Jaakkola et al., 2010).

Fortunately, there exists methods that are faster, and still able to offer some guarantees on finding the maximum scoring structure. One such algorithm is `GES` (Generalized Equivalence Search, Chickering (2002); Meek (1997)). `GES` starts from the equivalence class of an empty graph, first greedily adding edges one by one, such that the improvement on the score is maximized in each step. After no addition improves the score, a succession of edge deletions is performed in a similar, greedy fashion. Using the Bayesian score introduced, such a greedy algorithm produces the equivalence class of the true graph structure as an output provided that the faithfulness assumption holds, in the infinite sample limit (Chickering, 2002; Koller and Friedman, 2009). Note that this result does not mean that we would be guaranteed to find the maximum scoring graph with a limited number of samples. However, `GES` has shown to achieve good results even when the assumptions are slightly violated (Koller and Friedman, 2009). For even larger graphs calculating the best possible edge addition or deletion in `GES` may still be computationally too heavy. Markov Chain Monte Carlo -based (MCMC) resampling methods can be used in such situations (Koller and Friedman, 2009). Another algorithm often considered is `Max-Min Hill-Climbing`, which is a combination of the score-based and constraint-based approaches (Tsamardinos et al., 2006).

In situations where score-based learning can be applied, the results are usually better than for the constraint-based methods of Section 3.1. One reason for this is the fact that the approach does not make hard decisions about existing independence relations. In addition, the method also outputs many alternative probable structures. However, with the common local scores, the methods give the same score for all graphs in a single Markov equivalence class. Thus, the power of the algorithms when exploiting only passive observational data is limited to finding Markov equivalence classes.

### 3.2.4 Model Complexity

The Bayesian paradigm holds a preference towards simpler models. If two different model structures are able to produce the same dependencies in the data, the simpler one will usually get a higher posterior probability. Thus, there is no need to penalize complex structures in the prior for the graphs.

However, if the true model structure turns out to not be the simplest one, the approaches usually fail to produce the equivalence class of the correct causal structure. Thus, the approaches enforce some sort of simplicity assumption very similar to the faithfulness assumption. In fact, unfaithful models can be problematic. If the true causal model is as Figure 3.2 (a), score-based structure search would usually return an unshielded collider $x \rightarrow z \leftarrow y$. This is because such a structure can explain the dependencies in the joint distribution just as well as the true structure, and it is one edge simpler. For learning the model of Figure 3.2 (b), unfaithfulness turns out not to be a problem. Although faithfulness essentially has to be assumed when using score-based methods, its violations may not be as critical as in the constraint-based framework.

## 3.3 Exploiting Higher Order Statistics

This section describes methods exploiting the full joint distribution of continuous variables. Often methods are first crafted with restrictions on the functional forms of the causal relations, such as linearity. Further research has shown that the parametric assumptions can be somewhat relaxed. The class of methods presented here are more powerful than the previous methods: they can often identify the structure uniquely from passive observational data and are thus not restricted to finding only Markov equivalence classes. In addition, the methods can also find unfaithful or effectively unfaithful models accurately. However, they all require some restrictions on the parametric form of the causal relations, and produce unreliable results when the restrictions are not respected by the data generating system.

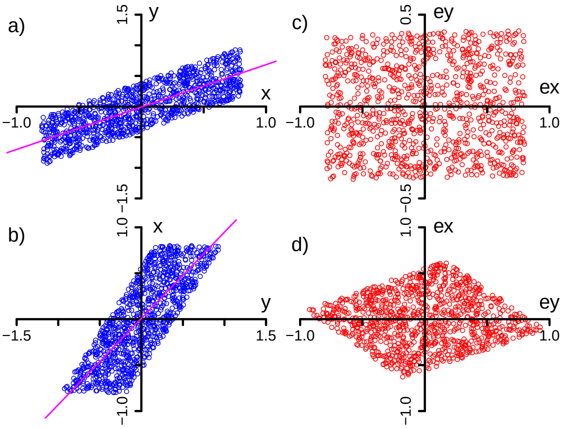


Figure 3.6: Principle of the `LiNGAM`-algorithm. The data points of $x$ and $y$ are shown in (a), and same points with axis reversed are in (b). The magenta lines show the estimated linear causal relation for both causal directions. Plot (c) shows the estimated disturbances for the model $x \rightarrow y$, and (d) for the model $y \rightarrow x$. Because the estimated disturbances in (c) are independent while in (d) they are not, we can deduce that $x$ causes $y$.

### 3.3.1 Linear Non-Gaussian Acyclic Models

Consider that we have been given the passive observational data shown in Figure 3.6 (a) (and with the direction of the axis switched in (b)). We would like to know whether $x$ causes $y$ or $y$ causes $x$. Since the relationship seems to be linear, we compare two models:

$$\begin{aligned} x &:= e_x \\ y &:= b_{yx} \cdot x + e_y \end{aligned} \qquad\qquad \begin{aligned} x &:= b_{xy} \cdot y + e_x \\ y &:= e_y \end{aligned}$$

The first model postulates that $x$ causes $y$ while according to the second $y$ causes $x$.

One can fit the first model to the data as the following simple procedure. First, one can regress the dependent variable $y$ on $x$. The least squares solution will give an estimate on $b_{yx}$, this is represented by the magenta line in Figure 3.6 (a). Then, estimates for the disturbances of both observed variables can be obtained by setting $e_x := x$ and $e_y := y - b_{yx} \cdot x$. Fitting the second model can be done similarly.

Now, according to the model definition and the causal sufficiency assumption the disturbance terms of the two variables should be independent. The disturbances of both models are plotted in Figure 3.6 (c-d). The disturbances for the first model are independent in (c): the value of $e_x$ does not help in predicting the value of $e_y$. On the other hand, the disturbances for the second model in (d) are dependent: for example if $e_y \approx 1$, then $e_x \approx 0$. Thus, we can deduce that the first model is correct, variable $x$ causes variable $y$.

Note, that due to the properties of linear regression, the disturbances will necessarily be uncorrelated for both models. For variables in a jointly Gaussian distribution uncorrelated variables are also independent, the disturbances must have non-Gaussian distributions for this type of inference to work.

**Assumption 10 (Non-Gaussianity)** *The distributions of the independent disturbances $e_1 \dots e_n$ are non-Gaussian.*

The uniform distribution used in the example is one such a non-Gaussian distribution.

The inference conducted in the previous example is possible using the `LiNGAM`-algorithm (Linear Non-Gaussian Acyclic Model) (Shimizu et al., 2006). The model class considered is a linear structural equation model (SEM)

$$\mathbf{x} \;:=\; \mathbf{Bx} + \mathbf{e}, \tag{3.16}$$

where each $e_i = \mathbf{e}[i]$ is distributed independently with a Non-Gaussian distribution $p_i()$. The basic `LiNGAM`-algorithm uses numerical methods of Independent Component Analysis (ICA) (see Hyvärinen et al. (2001)) to find the matrix $\mathbf{B}$ (in Equation 3.16) such that the residuals $e_1, \dots, e_n$ are as independent as possible (Shimizu et al., 2006). A very basic result is that we can find the true $\mathbf{B}$-matrix just from passive observational data for any number of variables, in the infinite sample limit. By pruning the obtained $\mathbf{B}$-matrix we will also find the true causal structure.

Recently, Shimizu et al. (2011) have published an alternative method called `DirectLiNGAM` exploiting these same principles. This method can extract the structure more reliably from a fewer number of data points. They fit regression models, similarly as in the previous example, and compare (general) independence of the residuals for the different models.

### 3.3.2 Non-linear Causal Relations

The linearity assumption of the `LiNGAM`-method has been relaxed in recent literature. Hoyer et al. (2009) consider structural equation models where the equations are of the form:

$$x_i := f_i(\mathrm{pa}(x_i)) + e_i. \tag{3.17}$$

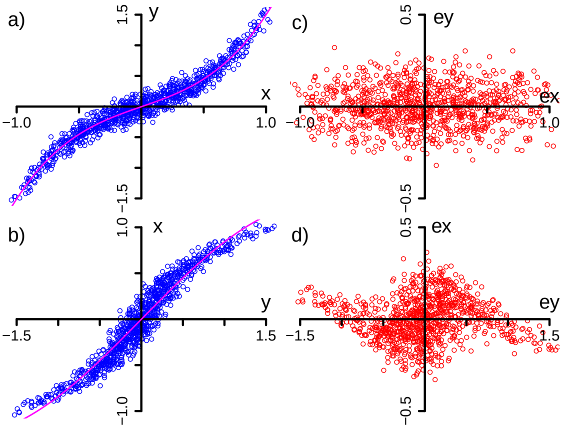


Figure 3.7: Discovery of additive noise models. The data points of $x$ and $y$ are shown in (a), and the same points with axis reversed appear in (b). The magenta lines show the estimated function causal relation for both directions. Plot (c) shows the estimated disturbances for model $x \to y$, and (d) for the model $y \to x$. Because the estimated disturbances in (c) are independent while in (d) they are not, we can again deduce that $x$ causes $y$.

Here the functions $f_i$ are possibly non-linear. Notice that the stochastic disturbance term $e_i$ is *additive*, hence the model is called an additive noise model.

**Assumption 11 (Additive Noise)** *The structural equations are deterministic apart from an additive stochastic term.*

Figure 3.7 (a-b) shows data generated by such an additive noise model with two variables. Similarly as in Section 3.3.1 we can try to fit two models, one postulating $x \to y$ and the other $y \to x$. The non-linear causal relation (line in magenta) is discovered by non-linear regression, fitting a 5th degree polynomial in both cases. The estimated disturbances are independent for $x \to y$ in Figure 3.7 (c), but dependent for $y \to x$ in Figure 3.7 (d). The additive noise model fits only for one direction of the causal relation. If the noise was indeed additive, $x \to y$ is the correct structure.

Hoyer et al. (2009) show that in many cases we can discover the causal direction between two variables in this way just from passive observational data. One exception occurs when $f_i$ is linear and the disturbances $e_i$ have Gaussian distributions: in this case identifiability is still possible only up to the Markov equivalence class (Section 3.1.4). But if either requirement is dropped (like the linearity in the previous example and Gaussianity in Section 3.3.1), the model can in fact be identified. Thus, non-linearity of the causal relations may actually aid in the identification of causal models in some situations (Hoyer et al., 2009).

Zhang and Hyvärinen (2009) generalize the model to include an additional nonlinear function $g_i$ in the structural equation:

$$x_i := g_i(f_i(\mathrm{pa}(x_i)) + e_i), \tag{3.18}$$

and show that the causal direction can be identified with these models as well. Peters et al. (2011) generalize the identifiability results for both model families to any number of variables. The actual learning of these models requires the use of various non-linear optimization techniques and general independence measures (Mooij et al., 2009). Note that these results all assume causal sufficiency, and due to use of the whole continuous distribution require a relatively high number of samples. The methods are also fairly sensitive to the assumptions: if the underlying system cannot be described by the structural equations allowing only additive noise, the results may be unreliable.

# Chapter 4

# Experiments

Experiments are a very powerful way of finding causal relations, as causal relations can be considered to be determined by interventions in experimental settings. Discovering causal relations from passive observational data may be considered a more risky way of obtaining causal knowledge, as causal relations are only inferred from statistical associations under often unverifiable assumptions. On the contrary, in experiments causality may be considered more directly observed. Experiments may also provide confirmation of the causal relationships inferred from passive observational data.

Experiments were also the first formalized way of obtaining scientific causal knowledge (Fisher, 1935). However, in these traditional medical or in some cases agricultural experiments (see Figure 1.2, p. 5) the considered variables are often divided into treatment variables and outcome variables. The goal is to find the causal effects the treatment variables have on the outcomes, if there are any. In causal discovery this distinction is not made. Instead, the goal is to discover the whole causal structure without a priori division of variables to different groups. This is a more realistic setting for some application areas.

The benefits of using experiments for causal discovery are twofold. First, a surgical (i.e. ideal) experiment cuts the effects any confounders have on an intervened variable. Thus, the association between an intervened variable and an observed variable can be interpreted to be causal, directed from the intervened variable to the observed variable (Figure 1.2). In addition, experiments allow us to orient edges that may have been left unoriented by constraint or score-based methods exploiting only passive observational data. Although exploiting higher order statistics may allow us to orient edges, in many situations experiments provide a more reliable way of obtaining the causal orientation. We will focus on this latter benefit in this section, since it has been covered more extensively in the causal discovery literature. The use of experiments to break latent confounding is considered briefly in Section 5.5 (Experiments in the Presence of Latent Confounding) and more extensively in the original research of this thesis in Chapter 7 (Contributions to the Field).

We will formalize an *experiment* to be an experimental setting where a specific set of variables $\mathcal{J}$ are intervened on, while the remaining variables $\mathcal{U} = \mathcal{V} \setminus \mathcal{J}$ are (passively) observed. The intervened variables $\mathcal{J}$ are considered to be randomized (independently). The number of samples extracted in the experiment and how the values of the intervened variables are determined, are not considered explicitly. This is because the experimental settings are the most vital thing for causal discovery: they generally define which causal relations can and which cannot be identified. We will refer to the problem of choosing the experimental settings, in such a way that as many as possible causal relationships can be identified, as *experiment selection.* After the experiments have been selected, we may want to lower the uncertainty of our parameter estimates by optimally choosing the values of the intervened variables. In statistics such questions are more extensively covered under the term *design of experiments.*

In Section 4.1 (Combining Several Experimental Data Sets) we will show how several experimental data set can be combined to reveal more of the causal structure than each data set alone. In section 4.2 (Experiment Selection) the choice of most informative experiments is considered in different settings, and some general results on the number of needed experiments are given. Section 4.3 (Different Types of Experiments) suggests alternative kinds of experiments and explains how they can be modeled with the causal modeling frameworks.

## 4.1 Combining Several Experimental Data Sets

First one may consider the benefits of experimental data for the constraint-based approach. As interventions in the experiments break some of the edges, more independencies can be found than in the passive observational case. Since association between an intervened variable and an observed variables can be attributed to the causal influence from the intervened variable to the observed variable, more edges can also be oriented. However, the independence tests run on data from different experiments are likely to give contradictory results due to the finite number of samples available in any realistic situation. Even with one data set the different conditional independence tests may produce contradictory answers, but with several experimental data sets this complication is amplified. One option might be to develop some sort of voting scheme for example on the conflicting structural features implied by the different data sets (Eberhardt, 2008b).

For the Bayesian score-based learning algorithms the incorporation of surgical experiments is more straight-forward (Cooper and Yoo, 1999). Experimental data induces differences only in the step of calculation of the local scores. Equation 3.13 (p. 39) gives the correct local score when $N_{ijk}$ is interpreted as the times variable $X_i$ gets its $k$:th value *by passive observation* while its parents get their $j$:th configuration. That is, any samples where $X_i$ gets its value by intervention are disregarded in the calculation of the local scores associated with node $X_i$. This is intuitive: the clamped value of $X_i$ does not give any information on the natural processes that bring about the value of $X_i$. Note that the samples disregarded in the calculation of local scores associated with $X_i$, are not disregarded in the calculation of local scores associated with other variables. Similar modification of local scores to account for experimental data also applies for any scores for continuous variables.

Most of the score-based algorithms introduced in Section 3.2.3 work straight-forwardly with the local scores considering also experimental data as explained. Although GES has been used with experimental data (Hauser and Bühlmann, 2012a), it is not yet clear whether this generalization has the same performance guarantees as GES. The score-based way of learning from experimental data sets has one clear advantage: all data are in a sense pooled together, no hard decision are made, so no hard conflicts can arise.[1]

## 4.2 Experiment Selection

Which experiments are sufficient and necessary in order to fully learn a causal model under the assumption of causal sufficiency? This question has been considered in three different settings, each explained in the following subsections. Note that although this characterization is mostly given considering only the constraint-based framework, it is directly applicable to the score-based approaches as well.

### 4.2.1 Predetermined Set of Experiments

The first interesting characterization of experiments needed for full structure identification can be stated under the common assumptions of acyclic-

[1] On the other hand, it may be good for a method to report and notice conflicts: they may indicate for example a faulty data collection process.

ity (Assumption 1), sufficiency (Assumption 6) and faithfulness (Assumption 9). Then, the graph structure can be uniquely identified if for each unordered variable pair $(X, Y)$ at least two of the following options are satisfied (Eberhardt, 2007):

- There is an experiment where $X$ is intervened and $Y$ is passively observed.
- There is an experiment where $Y$ is intervened and $X$ is passively observed.
- There is an experiment where $X$ and $Y$ are both passively observed.

This characterization of experiments is also worst case necessary: for each set of experiments not satisfying the condition there exist model structures that are indistinguishable from the given set of experiments.

Note in particular that the condition ensures that we can orient each edge between the variables. For any edge $X - Y$ the experiments will have at least one experiment where one of the variables is intervened on and the other one is passively observed. Say the intervened variable happens to be $X$. Now if $X$ and $Y$ are dependent in the distribution obtained when intervening on $X$, the edge can be oriented as $X \rightarrow Y$, otherwise the edge must be oriented as $Y \rightarrow X$.

If we aim for full identifiability using only experiments where a single variable is intervened on in every experiment, then $n - 1$ experiments are needed (Eberhardt et al., 2006). It can also be shown that experiments where one variable is randomized and other variables are forced to certain constant values do not aid in this inference. On the other hand, if we can intervene on and randomize any number of variables in each experiment we need only $\lceil \log_2(n) \rceil$ experiments plus possibly an additional passive observational data set (Eberhardt et al., 2005). One derivation of this latter bound and instructions on how to select experiments such that this condition is satisfied are considered in detail in Article VI.

### 4.2.2 Orienting Edges in a Markov Equivalence Class

Another relevant setting for selecting a set of experiments occurs when we have previously obtained the Markov equivalence class of the true graph from passive observational data (Figure 4.1). The partially oriented DAG of the equivalence class defines the graph structure uniquely up to the orientation of some edges (Section 3.1). The task is then to find a minimal set of experiments that provides information sufficient for orienting all unoriented edges.

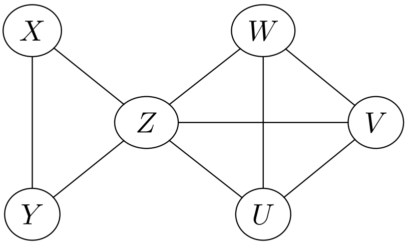


Figure 4.1: Orienting edges by experiments from a given Markov equivalence class. There are total of 72 DAGs in this equivalence class. The partially oriented DAG has two maximal cliques, $\{X, Y, Z\}$ and $\{Z, W, U, V\}$, and the edges may be oriented by two experiments intervening on $\mathcal{J}_1 = \{Z, W\}$ and $\mathcal{J}_2 = \{X, W, V\}$ respectively.

The number of experiments needed to orient all edges is related to the size $C_{max}$ of the largest maximal unoriented clique (Eberhardt, 2008a). An unoriented clique is a set of nodes such that any two nodes in the set are connected by an unoriented edge. The clique is maximal if it is not part of any larger clique. In the partially oriented DAG of Figure 4.1 the maximal unoriented cliques are $\{X, Y, Z\}$ and $\{Z, W, U, V\}$, and thus $C_{max} = 4$. Now, finding the orientations inside each clique using experiments is very similar to learning a causal model over the variables in the clique as in Section 4.2.1. The slight difference here comes from the fact that the different cliques may contain some common nodes, such as $Z$ in Figure 4.1. Apart from this, the edges in the different cliques can be oriented in parallel. Then, at most $\lceil \log_2 C_{max} \rceil$ experiments are needed (Eberhardt, 2008a; Hauser and Bühlmann, 2012b). In the example of Figure 4.1, two experiments ($\lceil \log_2 4 \rceil = 2$) intervening on $\mathcal{J}_1 = \{Z, W\}$ and $\mathcal{J}_2 = \{X, W, V\}$ are sufficient for orienting the unoriented edges.

### 4.2.3 Active Learning

Another type of setting for learning causally sufficient models is termed active learning (Tong and Koller, 2001; Murphy, 2001; He and Geng, 2008). Active learning applies to a situation where an agent is trying to learn as much as it can from the world by conducting various experiments. After each experiment the agent is allowed to contemplate which next experiment would be most informative and cost-effective. The next experiment is often chosen to minimize some measure of uncertainty on the possible causal structure. Due to the high number of DAGs all probable in the early stages of the learning setting, active learning procedures resort to approximate sampling techniques when selecting the next experiment. Note that the results of Section 4.2.1 and 4.2.2 give a characterization of the experiments that an active learning procedure must perform in order to fully identify a model structure.

## 4.3 Different Types of Experiments

Although surgical, ideal experiments are very useful in formalizing the notion of causality, the randomized experiments performed may not live up to this standard. In many situations it might not be possible to entirely break the influence the natural causes have on an intervened variable. For example in medical experiments there is sometimes non-compliance: the patients may not always take the medicine assigned to them. Many different types of non-ideal experiments have been considered in the literature. Again, the application field dictates which type of experiments are possible, useful and cost effective.

One particularly useful type of experiment has been termed (in different contexts) a *soft intervention* (Eberhardt and Scheines, 2007), a quasi-experiment (He and Geng, 2008), or a parametric intervention (Eberhardt, 2007). In such an experiment the conditional probability distributions (or similarly the structural equations) $P(X|\mathrm{pa}(X))$ of the intervened variables are replaced by different conditional probability distributions $Q(X|\mathrm{pa}(X))$. Often at least some aspects of $Q(X|\mathrm{pa}(X))$ are known to the experimenter.

One example of a soft intervention important for this thesis[2] can be given for linear causal models (Eberhardt et al., 2010). A soft intervention corresponds to adding a vector $\mathbf{c}$ to the system of structural equations in Equation 2.17 (p. 20):

$$\mathbf{x} \;:=\; \mathbf{B}\mathbf{x} + \mathbf{c} + \mathbf{e}, \tag{4.1}$$

where the elements of $\mathbf{c}$ corresponding to non-intervened variables are always zero. Note that unlike for surgical experiments, the matrix $\mathbf{B}$ remains unaltered: the natural causes of the intervened variables still have some say on determining the values of the intervened variables. For example when analyzing biological systems, this kind of intervention would correspond to artificially increasing or decreasing the concentration of some substances already present in the system, without having complete control over the variable in question.

[2] Eberhardt et al. (2010) show that the experimental effects heavily used in Articles II-IV can be estimated also from experiments with this type of interventions.

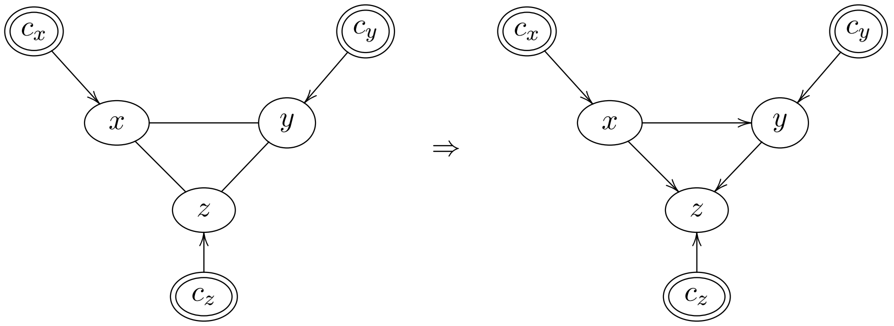


Figure 4.2: Causal discovery from a soft intervention. One soft intervention is sufficient to determining the full structure between the observed variables $x$, $y$ and $z$. When running PC on the observed variables $\mathbf{x}$ *and* the intervention variables $\mathbf{c}$, all edges between the observed variables get oriented.

Sometimes soft interventions may be *more* informative than surgical interventions. In the example of Figure 4.2 all variables of a 3 variable causal model are simultaneously subject to soft interventions. This one experiment allows us to identify the full structure among the observed variables: each edge becomes part of an unshielded collider and is thus oriented correctly (Eberhardt and Scheines, 2007). It really depends on the application area whether this experiment where all variables are subject to soft intervention is cheaper than a set of surgical experiments allowing for full identification of the structure.

In addition to soft interventions, there exist several other types of experiments in the literature (Eaton and Murphy, 2007). For example, uncertain interventions are only sometimes able to influence their target variables. Fat hand -interventions may also influence other variables than the intended target variables.

# Chapter 5

# Latent Confounding

In some sense, assuming causal sufficiency in Section 3 is ducking the real problems in inferring causal relationships. In real situations, we often cannot assume that there are no latent confounders present. There exists so many potential latent confounders that it is unlikely that not a single one is a confounder for the set of observed variables (Robins and Wasserman, 1999). In many situations the influence of latent confounding is also strong enough for the methods assuming causal sufficiency to be unreliable. It has been argued that in empirical studies with relatively high sample sizes one often observes highly significant dependencies between variables that are firmly believed not to be causally related (Robins and Wasserman, 1999).

Not only are the confounders unobserved, we also do not know how many confounders affect the system under study. In addition these unobserved variables may have a fairly complicated structure among themselves.[1] Since the variables are unobserved, we usually have no hope in obtaining this structure. We abstract away from the causal structure between the latent variables by using *mixed graphs.* In addition to directed edges, mixed graphs include *bidirected edges* between pairs of observed variables whenever there exists at least one latent confounder affecting both of the variables (Figure 5.1 (a-c)). In this notation any latent confounder causing three or more variables will be represented by bidirected edges between all pairs of the affected variables (Figure 5.1 (c)). Many of the algorithms presented in this section also allow for selection bias, which can be represented by *undirected*[2] *edges* between observed variables (Figure 5.1 (d)).

Allowing for latent confounding presents several difficulties and com-

[1] In some cases the scientist might actually be more interested in the causal structure between the latent variables (Silva et al., 2006).

[2] This notation is not to be confused with unoriented edges that indicate a causal effect from one variable to another, with the orientation unknown, as used in Section 3.1.4.

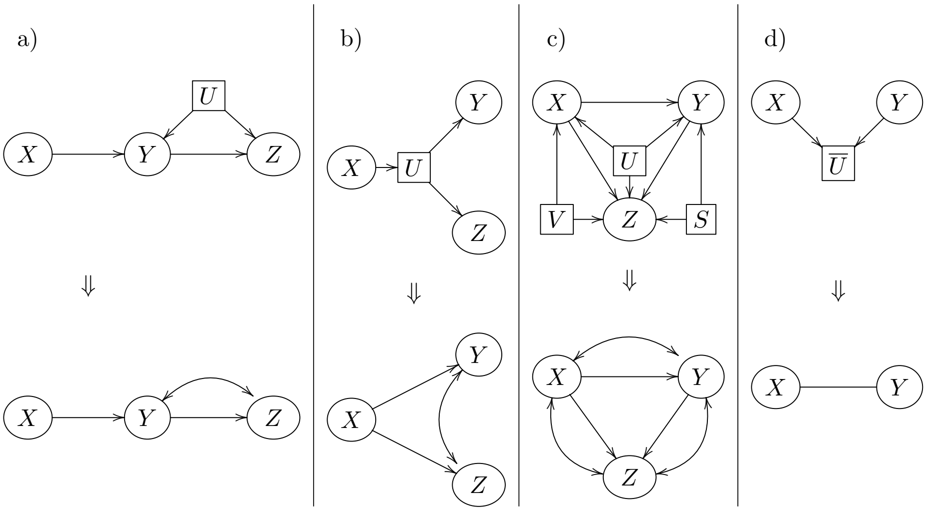


Figure 5.1: Representing latent confounding by bidirected edges (a-c) and selection bias by undirected edges (d).

plications for the approaches given in Section 3 (Causal Discovery). Especially under a lot of latent confounding, the models learned from passive observational data leave a large part of the structure unidentified. Thus without causal sufficiency, it becomes increasingly important to incorporate background knowledge and to combine information from several (possibly experimental) data sets.

In Section 5.1 (FCI-algorithm) we consider the basic constraint-based algorithm robust against latent confounding and selection bias. Then in Section 5.2 (Data Sets with Overlapping Variables) the basic idea behind constraint-based causal discovery is applied to a setting where we have several data sets that do not all share the same set of variables. Sections 5.3 (Approximating the Score-based Approach) and 5.4 (LiNGAM and Latent Variables) briefly explain how the other discovery approaches can be used to discover causal structures under latent confounding. Finally Section 5.5 (Experiments in the Presence of Latent Confounding) considers exploiting experimental data.

## 5.1 FCI-algorithm

The general idea of constraint-based causal discovery of Section 3.1 can also be exploited under latent confounding. The detected independence

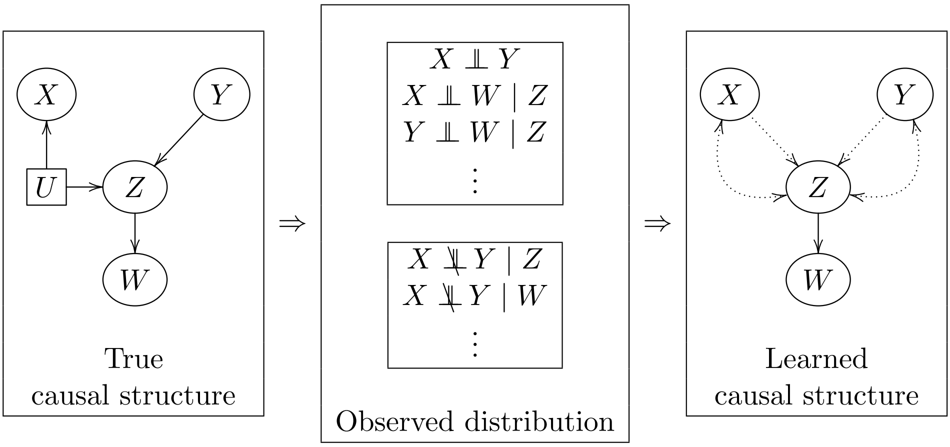

X
Y
U
Z
W
True
causal structure
⇒
$X \perp\!\!\!\perp Y$
$X \perp\!\!\!\perp W \mid Z$
$Y \perp\!\!\!\perp W \mid Z$
⋮
$X \not\!\perp\!\!\!\perp Y \mid Z$
$X \not\!\perp\!\!\!\perp Y \mid W$
⋮
Observed distribution
⇒
X
Y
Z
W
Learned
causal structure

Figure 5.2: Causal discovery from independence and dependence relations in passive observational data, not assuming causal sufficiency. Again d-separation implies some independencies on the observed variables, and by assuming faithfulness of the true data generating model other independencies are not present in the passively observed distribution. In the learning result on the right panel several relations remain undetermined (dotted edges) but existence of some relations can be inferred (missing edges and the solid edge).

and dependence relations in the data can in some cases narrow down the possible causal structures.

The connection between the graph notation (Figure 5.1) and the set of independence relations is the familiar d-separation condition (Definition 3, p. 28) with the following conventions. The bidirected edges $X \leftrightarrow Y$ can be interpreted like there would be an implicit latent variable in between: $X \leftarrow L \rightarrow Y$ (Richardson and Spirtes, 1999). Similarly, any undirected edge $X - Y$ can be interpreted as a structure $X \rightarrow \overline{S} \leftarrow Y$ and implicitly including the variable $S$ in each conditioning set. Given the conventions on d-separation, the faithfulness assumption (Assumption 9, p. 30) extends to this setting without modifications.

Figure 5.2 presents a situation where the possible causal structures producing the detected independence and dependence relations can be narrowed down to have several informative features in common. The true causal structure is represented on the left, note that it includes a latent confounder $U$. Some of the independence and dependence relations between the observed variables are shown in the middle. The right panel shows the learning result. Even in this favorable case a significant part of the causal structure remains unidentified, these features are marked by the dashed edges. For example, it is not clear whether $X$ causes $Z$ or if there is a latent confounder between the variables. One of the relations must be present. Some important features can be recovered uniquely from the independence relations. Variable $Z$ is discovered to be a direct cause of $W$. There are no confounders or causal connections between $X$ and $Y$. Furthermore, neither $X$ nor $Y$ is a direct cause of $W$, their possible causal influence on $W$ must be mediated by $Z$.

The `FCI` (Fast Causal Inference) algorithm automates the inference of the causal structure from the detected independence relations in passive observational data (Spirtes et al., 1993). There exists variants of the procedure that use Assumption 7 (No Selection Bias) and ones that do not. As indicated in the example, even in the best case scenario `FCI` can infer the graph structure only up to an equivalence class. A formal graph language for dealing with the equivalence classes has been developed (Richardson and Spirtes, 2002). Partial ancestral graphs (PAG) describe the common features of mixed graphs in the equivalence classes (such as ancestral relationships) and have many favorable properties in the task of inferring the causal structure from passive observational data. The current version of the `FCI` algorithm has been shown to be *complete*: it can discover all aspects of the causal structure that are uniquely determined by the independence and dependence relations (Zhang, 2008). However, the several different types of edges used in PAGs require a quite intricate causal interpretation. Furthermore, the good properties of PAGs do not carry over to the learning settings vital to this thesis: experimental data, possibly cyclic causal structure, data sets with overlapping variables (Section 5.2), or available background knowledge (Zhang, 2008; Borboudakis et al., 2011; Tsamardinos et al., 2012). Developing `FCI` to also be complete with background knowledge and experimental data is worth investigating (Zhang, 2008).

Overall, `FCI` can give important information on the causal structure but more often may leave the structure severely underdetermined. This is especially the case with a highly confounded set of variables. For example the models in Figure 5.1 do not produce any conditional independencies among the observed variables: any conditioning just unblocks new paths through the confounders. For example in Figure 5.1 (a) conditioning on $Y$ unblocks the path $X \rightarrow \overline{Y} \leftrightarrow Z$ and thus $X \not\perp\!\!\!\perp Z \mid Y$. The considerably different structures in (a-c) thus cannot be distinguished based on independence and dependence relations in passive observational data.

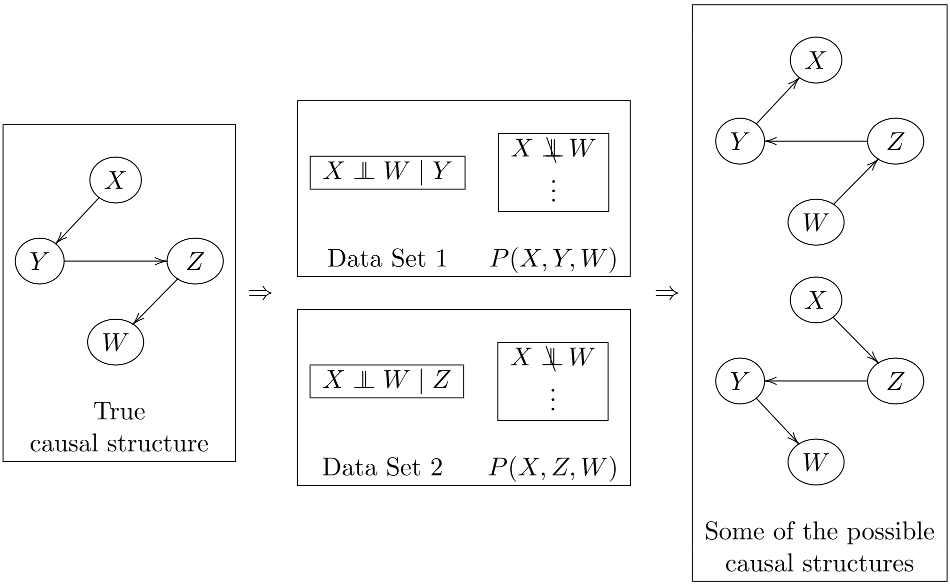

X
Y
Z
W
True
causal structure
⇒
$X \perp\!\!\!\perp W \mid Y$
$X \not\perp\!\!\!\perp W$
⋮
Data Set 1
$P(X, Y, W)$
$X \perp\!\!\!\perp W \mid Z$
$X \not\perp\!\!\!\perp W$
⋮
Data Set 2
$P(X, Z, W)$
⇒
Some of the possible
causal structures

Figure 5.3: Several data sets with overlapping variables. Assume we have observed data from the data generating model in the left panel in two settings: in the first set we have $X, Y, W$ and in the second $X, Z, W$. The independence relations detected in the marginal distributions are given in the two panels in the middle. The right panel shows the learning result. This example was originally considered by Tsamardinos et al. (2012).

## 5.2 Data Sets with Overlapping Variables

Sometimes we may have several data sets (generated by the same system in a passive observational state) that do not share the same set of variables, but nevertheless there is significant overlap (Danks, 2002). The goal is to learn the causal structure over the joint set of observed variables. It is clear that in this setting no single data set is usually causally sufficient. Some work has assumed *joint causal sufficiency* (Danks, 2002, 2005): the joint set of observed variables are assumed to be causally sufficient. Then there cannot be any confounders that would not be observed in any of the data sets. In a more challenging setting, no such joint sufficiency is assumed (Tillman et al., 2009). This way the setting is even more challenging than when observing a single joint, causally insufficient data set. We will focus on this latter setting here.

Figure 5.3 shows an example of such a learning setting. The single model (on the left) is observed in two circumstances, such that in each data set some of the variables remain unobserved. The important observations are shown in the two panels in the middle. Unfortunately, the different graph structures explaining the independence and dependence facts may have very little in common. The two data sets indicate that $X$ and $W$ are only indirectly linked through $Y$ and $Z$. Any unblocked path between $X$ and $W$ would explain this, two examples are shown on the right panel while the true causal structure provides a third. One of the key observations here is that we can learn the presence of causal relations between variables that are not observed together in any data set. In the example of Figure 5.3 we learn for example that there is at least some sort of connection between variables $Y$ and $Z$, otherwise the detected dependencies would simply not be possible.

Some of the methods designed for this particular learning setting exploit the theory surrounding the `FCI`-algorithm. The `ION`-algorithm (Integration of Overlapping Networks) integrates the results of the different structures learned by `FCI` for the individual data sets (Tillman et al., 2009). The ION algorithm is similarly *complete* with respect to the independence relations in the data sets (assuming acyclicity, faithfulness) as `PC` and `FCI`. A further development called `IOD` (integration of overlapping data sets) gives more accurate results from a finite number of samples (Tillman and Spirtes, 2011).

Triantafillou et al. (2010) also present an algorithm (`cSAT+`) for integrating the detected independence constraints into a causal structure. Their algorithm exploits the power of a general computer science technique called SAT-solving (satisfiability of logical statements). One benefit of this general framework is that the methods can be more easily adapted also to other type of learning settings.

## 5.3 Approximating the Score-based Approach

Ideally, one would like to learn causal models under latent confounding using a score-based approach as in Section 3.2 (Score-based Approach). Unfortunately, the posterior probability of different graph structures cannot be calculated in closed form. One option then may be to select the graph structure with the maximum (passive observational) likelihood $P(\mathbf{X}|\boldsymbol{\theta}_{ML}, \mathcal{G})$, where $\boldsymbol{\theta}_{ML}$ are the parameters that maximize the likelihood function given a fixed $\mathcal{G}$. A better approximation is to use the BIC-score (Bayesian Information Criterion, Gelman et al. (2004)), which adds an additional term to the likelihood function penalizing unnecessarily complex structures. Calculating these objectives for all graph structures is infeasible, and some sort of step-wise search over structures, such as GES (see Section 3.2.3, p. 40), must be performed. Given this basic idea, there are two different ways to proceed: we can consider the latent variables in the causal graph either explicitly or only implicitly.

When we explicitly mark some given number of latent variables in the causal graph structure $\mathcal{G}$, the calculation of the likelihood would require marginalizing (integrating) over the latent variables as their values are not observed. Thus, this is essentially a problem of missing data. The unknowns are the structure $\mathcal{G}$, the parameters $\boldsymbol{\theta}$ and the missing values for the latent variables. One option is then to use the structural EM -algorithm (Friedman, 1998; Koller and Friedman, 2009), which attempts to maximize the objective function by sequentially alternating three steps, each maximizing the objective with respect to one type of unknowns while keeping the other two fixed.

Alternatively, we may consider $\mathcal{G}$ to be a mixed graph, and again only implicitly account for the latent variables. In some cases one can find parameterization directly on the mixed graph structure, then marginalizing over the latent variables is not needed. Evans and Richardson (2010) consider binary variables in this way, while Drton and Richardson (2004) give an estimation procedure for linear causal models with Gaussian disturbances. Silva and Ghahramani (2009) give a more Bayesian approach based on sampling and variational approximations under some parametric restrictions (linear Gaussian and Probit).

Unfortunately, latent confounding variables introduce several severe complications: the objective may be multimodal and not generally decomposable, the complexity of the causal structure may not be well defined and the search space may not be tractable (Richardson and Spirtes, 2002). Learning models with latent variables often requires hand-tuned engineering (Koller and Friedman, 2009). It is unclear what sort of guarantees the above algorithms have on finding (the equivalence class of) the true structure.

Recently, Claassen and Heskes (2012) consider an approach without parametric restrictions that in a way combines the score-based and constraint-based approaches. They use score-based methods to approximate the posterior probabilities of certain structural independence statements. Then, a search (resembling that of FCI) can be performed, respecting the different confidence measures on the different independence statements.

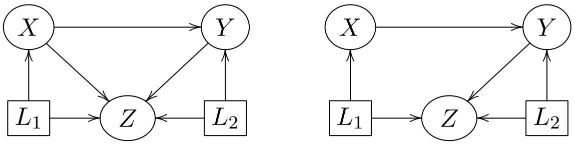


Figure 5.4: Two causal structures that are indistinguishable from experiments intervening only a single variable at a time. However, the models will have different behavior when intervening on $X$ and $Y$ simultaneously.

## 5.4 LiNGAM and Latent Variables

Attempts have been made to extend the basic idea of exploiting non-Gaussianity of the data and linearity of causal relations to the setting where some of the variables remain unobserved. The `lvLiNGAM`-method (latent variable `LiNGAM`) exploits an overcomplete basis ICA algorithm to find the **B**-matrix (Equation 3.16) of the linear Non-Gaussian Acyclic model (Hoyer et al., 2008b). It has been shown that the overcomplete basis of an ICA model is identifiable (Eriksson and Koivunen, 2004). However, turning the overcomplete ICA result into the **B**-matrix cannot always be done uniquely, and the method may only return an equivalence class of possible **B** matrices. Although the method seems theoretically promising, the estimation of the overcomplete ICA basis is only possible for very small models and even for small models the estimation may produce inaccurate results in practice. The development of more reliable methods not using the overcomplete ICA algorithms explicitly is ongoing (Entner and Hoyer, 2011; Tashiro et al., 2012).

## 5.5 Experiments in the Presence of Latent Confounding

In medical experiments, interventions are the primary way of obtaining information on specific causal relationships between possibly confounded variables. However in causal discovery, using experiments for learning full causal models under latent confounding has only been a fairly recent interest, perhaps because of the focus on learning causal relationships from passive observation.

Figure 5.4 illustrates one example on the use of experiments under latent confounding (Article IV). The figure shows two distinct causal structures of Bayesian networks. When passively observing these models, no independence relations can be detected. In addition, the models share the same independence relations also in all experiments intervening on a single variable at a time. In fact, there exist Bayesian networks with the two structures that produce the exact same distribution in the passive observational setting and in all single intervention experiments (Article IV). To discover whether $X$ has a direct causal influence on $Z$, an experiment intervening on $X$ and $Y$ simultaneously is needed. The example seems to suggest that no interesting characterizations on the type of experiments needed to identify the full causal structure can be developed, if we do not want to make any parametric restrictions on the complexity of the individual causal relationships. Essentially, in order to determine the direct causal links to a single variable we always have to intervene on all $n-1$ other variables.

Another difficulty is the incorporation of experimental data into the learning procedures. Experiments may constrain the possible causal structures able to produce the independence relations in arbitrary ways, so perhaps no nice characterization of equivalence classes (such as PAGs) can be developed either.

Borboudakis et al. (2011) and Borboudakis and Tsamardinos (2012) have considered experimental data in their general framework of constraint-based causal discovery using SAT-solvers, originally developed to learn causal models from overlapping data sets (see Section 5.2). They are able to exploit the ancestral causal information given by experimental data. In their `MCI`-algorithm, Claassen and Heskes (2010) exploit independence relations found in multiple experimental settings, that can be seen to correspond to soft interventions. One of the main themes of the original research articles in this thesis is the use of experimental data under latent confounding, together with certain parametric restrictions on the causal relationships (Articles II-V).

# Chapter 6

# Cycles

In Chapter 2 we made the 'traditional' assumption of the causal structure being acyclic. The assumption proved itself useful in many instances in deriving the theory, starting from the sampling processes in the causal order, the interpretation of the models and the description of powerful learning algorithms, and ending with unique identifiability of many types of models in the causally sufficient case (Section 3.3). But when can we really assume a priori that the causal structure is acyclic?

Perhaps the acyclicity of the causal structure seems intuitive. When considering events in time, a cause $X$ must precede its effect $Y$, and if $Y$ happens later than $X$, $Y$ cannot cause $X$. But this intuition applies only to token causation, where we are considering only a single chain of events in a single system (Richardson, 1996). The presented algorithms in this thesis aim to detect type-level causation: when one type of event causes some other type of event in some population. Then, although in different individual systems one type of event caused another type of event, in another system this may have happened in the opposite direction. For example in Figure 6.1 (a) disease A may influence the immune system of the patient making him vulnerable to some disease B. In another case the disease B may cause the patient to acquire disease A in a similar way.

The previous consideration is related to the argument that the causal structure is acyclic over time. True systems are often dynamic like in Figure 6.2 (left). $X$ at time $t$ can influence $Y$ at time $t+1$ and the value of $Y$ at $t$ can influence the value of $X$ at time $t+1$. The structure seems acyclic. But the practical setting is often not quite as simple. First, we may observe only a static approximation of this dynamic process (Figure 6.2 right), perhaps in some sort of equilibrium (Richardson, 1996). Then, although the dynamic process may be acyclic, the static model describing our observations is inherently cyclic. But sometimes even the dynamic process cannot

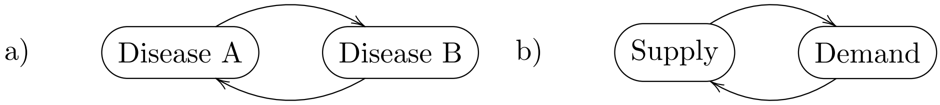


Figure 6.1: Cyclic models of causality are natural when time-series behaviour is not explicitly considered. (a) Disease $A$ may influence a patients immune system making the patient vulnerable for disease B. For another patient this may happen in the other direction. (b) Supply of a product affects its demand, increased demand would also yield a higher supply.

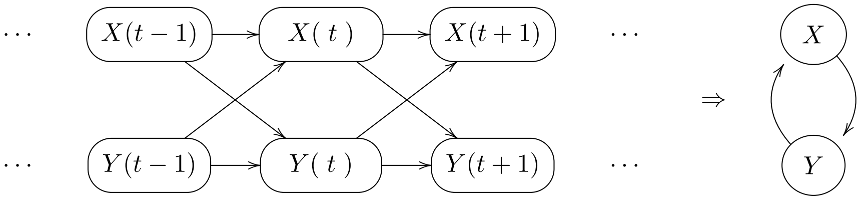


Figure 6.2: Cyclic models arising from acyclic dynamic processes. Although the dynamic process itself is acyclic, the causal structure of the observed static model approximating this dynamic process is cyclic.

be accurately modeled with an acyclic model. The time steps between the observations are often coarse such that $X$ of time step $t$ may have time to influence $Y$ at time $t$. If also $Y$ has time to influence $X$ within the same time step, we will have a cycle in the model. One can also argue that the measured quantities may not in fact be instantaneous but aggregate values over time. For example any quantity measuring supply or demand of some stock or a product (Figure 6.1 (b)) is usually an aggregate over several weeks or months. Thus the accurate causal model for the behavior would once again be cyclic.

In the applications of causal discovery there is definitely a demand for cyclic models. For example, linear non-recursive SEMs are commonly used in econometrics to represent feedback processes that have reached equilibrium (Spirtes, 1995). In biology, the causal structure of cellular signaling networks are known to be inherently cyclic (Sachs et al., 2005). Thus, several important causal relationships may be missed when using acyclic models.

Overall, cycles in the causal structure bring about similar complications as latent confounding. Assuming acyclicity simplifies the possible graphs significantly: if we have discovered a causal relation $X \to Y$ we can immediately rule out the causal relation $Y \to X$. In cyclic models both edges may be present. If they are not, often experimental data is needed to infer the absence of such cycles. Also the factorization property of the joint distribution exploited in the development of fast Bayesian score-based learning methods (Section 3.2) is lost when considering cyclic structures. Due to these extra difficulties, there is definitely also use for the methods relying on acyclicity in some learning settings.

In Section 6.1 (Cyclic Causal Models), we will first present a way of defining linear cyclic models. This interpretation is also used in Articles II-IV in the thesis. We will also describe and comment on interpretations of some other possibly non-linear cyclic causal models. Section 6.2 (Independence Properties of Cyclic Models) explains the current understanding of independence relations produced by cyclic causal structures. Section 6.3 (Discovery Algorithms) describes several learning algorithms.

## 6.1 Cyclic Causal Models

A particularly nice interpretation of cyclic models can be given when the causal relationships are linear. Consider a linear structural equation model (also given in Equation 2.17 on p. 20)

$$\mathbf{x} := \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{6.1}$$

where this time there is no restrictions on matrix $\mathbf{B}$, i.e. it does not need to be permutable to lower triangularity like for the acyclic interpretation. In the structural equation system, the variables $\mathbf{x}$ get their values as a linear combination of other variables plus a disturbance term $\mathbf{e}$. If the model is cyclic there is no causal order in which to sample the elements of $\mathbf{x}$ in such a way that always the parents of a variable would be sampled before the variable itself. Then, one way to interpret Equation 6.1 is that the values of $\mathbf{x}$ at time $t$ depend causally on the values of $\mathbf{x}$ at time $t-1$:

$$\mathbf{x}(t) := \mathbf{B}\mathbf{x}(t-1) + \mathbf{e}. \tag{6.2}$$

Notice that here the disturbance term $\mathbf{e}$ is interpreted to stay constant throughout the iteration. Since we often do not have access to the actual time series of this process, one option is to assume that we obtain samples once the system in Equation 6.2 has reached its equilibrium. Then the following sampling process may be used.

1. Sample the disturbances $\mathbf{e} = [e_1, \ldots, e_n]^T$ from their respective distributions $p_1(), \ldots, p_n()$.

2. Iterate Equation 6.2 until convergence, starting from some $\mathbf{x}(0)$, for example $\mathbf{x}(0) = \mathbf{0}$.

Iterating Equation 6.2 $t$ times gives the following value for $\mathbf{x}(t)$:

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{B}\mathbf{x}(t-1)+\mathbf{e}\\ &= \mathbf{B}(\mathbf{B}\mathbf{x}(t-2)+\mathbf{e})+\mathbf{e}\\ &\vdots\\ &= \mathbf{B}^t\mathbf{x}(0)+\sum_{i=0}^{t-1}\mathbf{B}^i\mathbf{e}.\end{aligned}$$

It is clear that for some coefficient matrices $\mathbf{B}$ this expression will fail to converge with increasing time $t$. As we hope to obtain samples in the equilibrium the following assumption is made disallowing such divergent systems (Fisher (1970), Article III).

**Assumption 12 (Asymptotic Stability)** *A linear cyclic model with coefficient matrix* $\mathbf{B}$ *is* asymptotically stable *if and only if for every possible experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$*, the eigenvalues* $\lambda_i$ *of the (manipulated) matrix* $\mathbf{B}$ *satisfy* $\forall i : |\lambda_i| < 1$.

Under this assumption we have that $\mathbf{B}^t\mathbf{x}(0) \to \mathbf{0}$ and the equilibrium point is completely independent of $\mathbf{x}(0)$. In addition, the other term converges as a geometric matrix series $\sum_{i=0}^{t-1}\mathbf{B}^i \to (\mathbf{I}-\mathbf{B})^{-1}$. Thus, the equilibrium point reached is simply

$$\mathbf{x} \;:=\; (\mathbf{I}-\mathbf{B})^{-1}\mathbf{e}. \tag{6.3}$$

Notice that this value for $\mathbf{x}$ is also a solution of Equation 6.1. As one would expect, when $\mathbf{B}$ happens to be acyclic this interpretation coincides with the simpler acyclic interpretation (Equation 2.20, p. 21). The important point to notice here is that the mapping from a configuration of disturbances $\mathbf{e}$ to equilibrium point $\mathbf{x}$ is a bijection under the assumption of asymptotic stability. The produced distribution can be obtained by change of variables for probability distributions (Gelman et al., 2004):

$$p(\mathbf{x}) \;=\; \det(\mathbf{I}-\mathbf{B})\cdot p_{\mathbf{e}}((\mathbf{I}-\mathbf{B})\mathbf{x}), \tag{6.4}$$

where $p_{\mathbf{e}}(\mathbf{e}) = \prod_{i=1}^{n} p_i(e_i)$.

An intervention can be interpreted as follows. An intervened variable $x_i$ is held at a constant value $c_i$ throughout the iteration. Then similarly to acyclic models the structural equation determining the value of $x_i$ is replaced with equation

$$x_i \;:=\; c_i. \tag{6.5}$$

The asymptotically stable manipulated system is then iterated until equilibrium.

A similar interpretation can also be given for non-linear cyclic models with additive noise (Mooij et al., 2011). There the structural equation system is described by

$$\mathbf{x} \;:=\; \mathbf{f}(\mathbf{x}) + \mathbf{e}, \tag{6.6}$$

where $\mathbf{f}$ is a nonlinear function $R^n \to R^n$. If we assume[1] that for any configuration of the disturbances $\mathbf{e}$ there is a single fixed point $\mathbf{x}$, then there is again a bijection from $\mathbf{e}$ to $\mathbf{x}$ and the distribution produced is given by

$$p(\mathbf{x}) \;:=\; \det(\mathbf{I} - \nabla\mathbf{f}(\mathbf{x})) \cdot p_{\mathbf{e}}(\mathbf{x} - \mathbf{f}(\mathbf{x})), \tag{6.7}$$

where $\nabla\mathbf{f}(\mathbf{x})$ denotes the Jacobian of $\mathbf{f}$ at $\mathbf{x}$ and again $p_{\mathbf{e}}(\mathbf{e}) = \prod_{i=1}^{n} p_i(e_i)$.

For discrete cyclic models, the product of conditional probability distributions $P(X_i|\text{pa}(X_i))$ does not generally sum up to one if the graph defining the parent relations is allowed to be cyclic. Schmidt and Murphy (2009) fix this by adding a new normalization constant in front of the factorization:

$$P(X_1, \ldots, X_n) \;=\; \frac{1}{Z}\prod_{i=1}^{n} \phi_i(X_i|\text{pa}(X_i)), \text{ where} \tag{6.8}$$

$$Z \;=\; \sum_{X_1,\ldots,X_n} \prod_{i=1}^{n} \phi_i(X_i|\text{pa}(X_i)). \tag{6.9}$$

Here the terms $\phi_i(X_i|\text{pa}(X_i))$ are called *interventional potentials.* If the causal structure happens to be acyclic the interventional potentials reduce to conditional probability distributions $P(X_i|\text{pa}(X_i))$ and $Z = 1$. Unfortunately, Schmidt and Murphy (2009) do not give a causal sampling process for their model formulation.[2]

Another interpretation of Equation 6.1 that would allow for cycles could be

$$\mathbf{x}(t) \;:=\; \mathbf{B}\mathbf{x}(t-1) + \mathbf{e}(t), \tag{6.10}$$

where the disturbances are no longer constant over the iteration. Clearly, $\mathbf{x}$ does not converge to any particular value. One option is to assume the system is sampled when it has reached some sort of *equilibrium distribution.*

[1] Mooij et al. (2011) give a sufficient condition for models with 2 variables.

[2] When using interventional potentials $\phi_i(x_i|\text{pa}(x_i)) = \mathcal{N}(x_i; \sum_{x_j \in \text{pa}(x_i)} b_{ij}x_j, \sigma_i^2)$ for continuous variables, the distribution $\frac{1}{Z}\prod_{i=1}^{n} \phi_i(x_i|\text{pa}(x_i))$ is equal to the distribution in Equation 6.4 under Gaussian disturbances $p_i(e_i) = \mathcal{N}(e_i; 0, \sigma_i^2)$. At least then, the sampling process for linear cyclic models gives a causal sampling process for the model of Schmidt and Murphy (2009) as well.

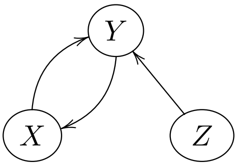


Figure 6.3: Failure of the local causal Markov condition. There exists linear cyclic models with this structure that produce distributions where $X \not\perp\!\!\!\perp Z \mid Y$ because of the unblocked path $X \rightarrow \overline{Y} \leftarrow Z$.

## 6.2 Independence Properties of Cyclic Models

The local causal Markov condition (Assumption 2, p. 15) does not describe the factorization and independence properties of cyclic models. Instead, Spirtes (1995) and Koster (1996) have shown that global causal Markov condition (Assumption 8, p. 28) characterizes the necessary independence relations for *linear* cyclic models observed in equilibrium. If $X$ and $Y$ are d-separated given $\mathcal{C}$ in the cyclic causal structure, then $X$ will be independent of $Y$ given $\mathcal{C}$ in joint distribution. If $X$ is not d-separated from $Y$ given $\mathcal{C}$ then $X$ and $Y$ will be dependent given $\mathcal{C}$ at least in one model with the given structure. Figure 6.3 shows an example of a case where the causal Markov condition fails: variable $X$ is not always independent of its non-effect $Z$ given its direct cause $Y$.

In nonlinear cyclic models the independence relations are a more complicated issue. The intuition behind d-separation and global causal Markov condition would seem applicable also for non-linear cyclic models in some circumstances but there exist some pathological counterexamples. Two d-separated variables in nonlinear cyclic causal models may still be dependent in the joint distribution (Spirtes, 1995). Details of the independence properties of discrete cyclic models have not yet been resolved either (Pearl and Dechter, 1996; Neal, 2000).

## 6.3 Discovery Algorithms

Some methods for discovering cyclic causal models have been devised assuming causally sufficiency. Since the global causal Markov assumption and d-separation condition apply for linear cyclic models in the same way as for acyclic models, the CCD (Cyclic Causal Discovery) algorithm discovers the cyclic structures from the observed independencies in passively observed data in a similar fashion to that presented in Section 3.1 (Constraint-based Approach) (Richardson and Spirtes, 1999). The algorithm is not complete:

the output PAG may not represent all features that are common to all the graphs with the given d-separation properties. However, a weaker form of completeness property has been shown. The equivalence classes are considerably larger than for acyclic models.

Schmidt and Murphy (2009) exploit experimental data heavily in their method for learning discrete cyclic models. Basically they maximize the likelihood given by Equation 6.8 in the different experimental settings. There are no completeness or identifiability results, but the approach shows good empirical performance (Schmidt and Murphy (2009), Article III). The numerical optimization method becomes computationally heavy with larger models.

The basic idea of exploiting non-Gaussianity has been extended for discovering cyclic models as well: in this case the **B**-matrix is not always uniquely identifiable from passive observation, but the equivalence class is considerably smaller than for constraint-based methods (Lacerda et al., 2008). Mooij et al. (2011) show that in the case of additive Gaussian noise also cyclic models with 2 variables can be discovered consistently in many cases. The estimation of this model is also done by a numerical optimization procedure. There is hope that these methods can be extended to general additive noise models, similarly to when the structure is assumed acyclic.

# Chapter 7

# Contributions to the Field

This chapter summarizes the original research articles of this thesis. As the articles are reprinted at the end of this thesis in their original form, the summaries are brief and focus on the motivation for the research, and how the articles connect to the existing work described in earlier chapters. This chapter also emphasizes the most important and interesting findings.

The main contributions of the original research are characterized by the contributed learning methods and their respective learning settings. One classification of the methods is given in Table 7.1; the articles where a particular method appears are given in parentheses. Section 7.1 (Discovery of Linear Acyclic Models) describes the contributions of Article I on developing a score-based method able to exploit linearity and higher order statistics in passively observed data. Section 7.2 (Linear Cyclic Models with Latent Variables) summarizes the research conducted for a series of articles II-IV about learning linear models under latent confounding. This includes methods that assume faithfulness and ones that do not. Section 7.3 (Noisy-OR Models with Latent Confounding) describes a model class with binary variables allowing for latent confounders and for which we were able to prove similar identifiability results as for the linear models. Section 7.4 (Experiment Selection) explains how to select the experiments such that the requirements for full causal model identification can be guaranteed. Notice that Article VI is not in Table 7.1 as it does not describe any specific learning algorithms. Finally, Section 7.5 describes the contributions of the present author in the preparation of the original articles.

| | Faithfulness Not Assumed | Faithfulness Assumed | |
|---|---|---|---|
| Causal Sufficiency and Passive Observational Data | | BayesLiNGAM[1] (I) | Linearity |
| Latent Confounding and Experimental Data | Direct-LLC (II,III)<br>Overlapping-LLC (IV) | Faithful-LLC (II)<br>Bilinear Approach (IV)<br>Linear Inference (IV) | Linearity |
| Latent Confounding and Experimental Data | Noisy-OR EM (V)<br>Efficient Conditioning (V) | | Noisy-OR (generalized) |

Table 7.1: Classification of different algorithms and learning settings considered in the research articles. The characteristics are given in the gray cells on the same row (on the left and on the right) and column as the algorithm. The distinction of the different developments of the LLC method does not appear in the original articles in exactly this from.

## 7.1 Discovery of Linear Acyclic Models

Article I (Bayesian Discovery of Linear Acyclic Causal Models) considers learning the causal structure of linear acyclic models (Definition 2, p. 19) under causal sufficiency. The graph structure of linear acyclic models can be learned in several ways in this setting. Constraint-based methods such as PC (Section 3.1.4) can consistently return the Markov equivalence class of the data generating graph structure. Score-based methods using linear Gaussian local scores (see Sections 3.2 and 2.2.3) can assign posterior probabilities to different graph structures. Graphs in each Markov equivalence class get the same score. This is because linear acyclic models with Gaussian disturbances are inherently identifiable only up to the Markov equivalence class. In any case, such a method suggests several alternative structures to the highest scoring one, and thus the method is able to express its uncertainty in its output result. On the other hand, LiNGAM (Section 3.3.1) can find the structure *uniquely*, when the disturbances are non-Gaussian. Unfortunately, it is not able to express its uncertainty on

[1] With Gaussian data BayesLiNGAM behaves like other score-based methods (Section 3.2.4, p. 42) and thus the assumption of faithfulness is essentially made. If all disturbances are non-Gaussian the faithfulness assumption is not needed.

the result, and it is not able to function consistently if the disturbances happen to be Gaussian.

Thus, our aim in Article I is to develop a method enjoying the identification power of the `LiNGAM`-method for non-Gaussian data, but in such a way that it would still be able describe its uncertainty on the output similarly as score-based approaches. In addition, we want a method that works consistently also for Gaussian data, similarly as the method of Hoyer et al. (2008a). To achieve these goals, we formulate new local scores for the Bayesian score-based approach for causal discovery. The resulting learning algorithm is appropriately called `BayesLiNGAM`.

The first step in devising a local score is to assign a probabilistic model for each conditional probability distribution $p(x_i|\mathrm{pa}(x_i))$. When assuming linearity of the causal relations, this amounts to assigning a probabilistic model for the disturbances $e_1, \ldots, e_n$ appearing in the corresponding linear structural equations. We model the distributions in two different ways. The goal is to have a general model able to represent any distribution, in particular the Gaussian distribution as well as some basic non-Gaussian distributions. The first option is to use a Gaussian-Laplacian (GL) model

$$p_i(e_i) \quad \sim \quad \exp(-\alpha_i|e_i| - \beta_i e_i^2) \tag{7.1}$$

that can perfectly model the Gaussian distribution when $\alpha = 0$, and the (non-Gaussian) Laplacian distribution when $\beta = 0$. As an alternative, we also consider the mixture of Gaussians (MoG):

$$p_i(e_i) \quad = \quad \sum_{k=1}^{K} \pi_{ik} \cdot \mathcal{N}(e_i; \mu_{ik}, \sigma_{ik}^2), \tag{7.2}$$

where $\pi_{ik} \geq 0$ and $\sum_{k=1}^{K} \pi_{ik} = 1$. We generally use the mixture of two Gaussians ($K = 2$) for computational efficiency, but still allowing for modeling of non-Gaussian disturbances.

For both alternative disturbance models the local score can be calculated by integrating over the model parameters (see Equation 3.11, p. 39):

$$\mathrm{score}(x_i, \mathrm{pa}(x_i)) \quad = \quad \log \int p_i(x_i - \sum_{x_j \in \mathrm{pa}(x_i)} b_{ij} x_j) d\boldsymbol{\theta}_i. \tag{7.3}$$

The integral is taken over the coefficients $b_{ij}$ and the parameters defining the disturbances. For GL the parameters are $\alpha_i$ and $\beta_i$ and for MoG $\mu_{i1}, \ldots, \mu_{iK}$, $\sigma_{i1}^2, \ldots, \sigma_{iK}^2$ and $\pi_{i1}, \ldots, \pi_{iK}$. We use some fairly non-informative prior distributions for all model parameters. The integral cannot be

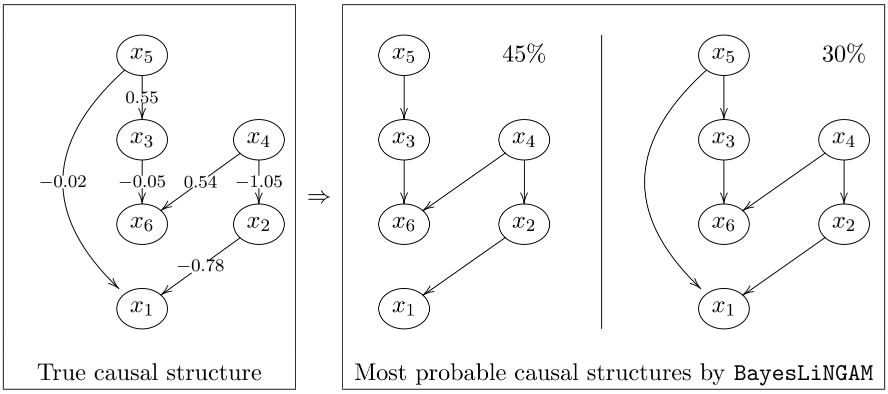


Figure 7.1: One Learning Result of `BayesLiNGAM`. Disturbances for $x_2, x_3, x_6$ are Gaussian and for $x_1, x_4, x_5$ non-Gaussian.

calculated in closed form and we resort to numerical approximations. We use the Laplace approximation (Gelman et al., 2004) and MCMC-sampling, both are common practices in machine learning.

With the general tools of the Bayesian score-based approach presented in Section 3.2 (p. 36) one is able to calculate the posterior probabilities for all directed acyclic graphs (for a fairly small number of variables). Notice that these scores do not satisfy the principle of likelihood equivalence, so different graphs in a single Markov equivalence class can now receive different scores, and identification of the structure may go beyond the equivalence class if the data is non-Gaussian. One learning result using the MoG-score is presented in Figure 7.1 (see Article I for technical details). The `BayesLiNGAM`-algorithm is able to learn most of the structure correctly. It is also able to express its uncertainty on the edge $x_5 \to x_1$ that has a fairly small coefficient. The method is able to utilize non-Gaussianity but still allows for Gaussian disturbances. The two local scores performed fairly similarly in all tests.

One future option would be to develop this method to allow for non-linear causal relations. The identifiability results for additive noise models suggest that such a procedure might be powerful. Note that in non-linear additive noise models non-Gaussianity is not actually needed for unique identifiability. Returning alternative structures for characterizing the uncertainty would be beneficial also for the non-linear discovery methods.

## 7.2 Linear Cyclic Models with Latent Variables

In the series of articles II-IV we consider learning linear causal models from experimental data, under some different settings and assumptions. The origin for the line of research was a previously published paper by two of the co-authors (Eberhardt et al., 2010). Article II (Causal Discovery for Linear Cyclic Models with Latent Variables) expands their results. Article III (Learning Linear Cyclic Causal Models with Latent Variables) combines and expands parts of the results by Eberhardt et al. (2010) and in Article II into a journal article. Article IV (Causal Discovery of Linear Cyclic Models from Several Data Sets with Overlapping Variables) generalizes the learning methods to exploit new kinds of experimental data sets.

The aim of the research is to learn causal models under a severe difficulty: latent confounding. To make the situation even more challenging we also allow the model structures to be cyclic. In order to allow for a clear interpretation of cyclic causal models, we constrain the causal relationships to be *linear*. Apart from this restricting assumption, the learning setting is very general. This challenging learning setting requires the use of experimental data. Combining several experimental data sets under latent confounding has not previously been extensively covered in the literature.

### 7.2.1 Model

We use a particularly clean and useful formulation of linear cyclic models with latent variables (Bollen, 1989). The formulation exploits the following observation concerning unobserved variables. Consider the following SEM model:

$$\begin{aligned} x_1 &:= b_{12}x_2 + b_{13}x_3 + e_1, \\ x_2 &= b_{21}x_1 + b_{23}x_3 + e_2, \\ x_3 &:= e_3. \end{aligned}$$

If $x_3$ is not observed, the SEM describing the causal relations between the observed variables $x_1$ and $x_2$ can be obtained by inputing the third equation into the first two, and aggregating the stochastic terms into new disturbances $e'_1$ and $e'_2$:

$$\begin{aligned} x_1 &:= b_{12}x_2 + e'_1, \quad \text{where } e'_1 := b_{13}e_3 + e_1, \\ x_2 &:= b_{21}x_1 + e'_2, \quad \text{where } e'_2 := b_{23}e_3 + e_2. \end{aligned}$$

Now the new disturbances are no longer independent, but correlated due to the latent confounder $x_3$. Thus, latent confounding can be represented by allowing the disturbances to be correlated.

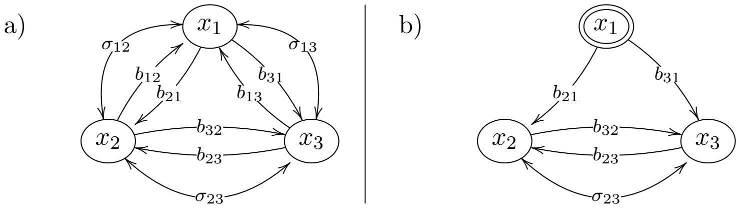


Figure 7.2: (a) An example of a linear cyclic model with latent variables. Any nonzero elements of $\mathbf{B}$ are represented by directed edges and nonzero elements of the symmetric covariance matrix $\boldsymbol{\Sigma}_\mathbf{e}$ are represented by bidirected edges. (b) The model structure when $x_1$ is intervened on.

**Definition 4 (Linear Cyclic (Causal) Model with Latent Variables)** *A linear Cyclic (Causal) Model with Latent Variables* $(\mathbf{B}, \boldsymbol{\Sigma}_\mathbf{e})$ *is a structural equation model*

$$\mathbf{x} := \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{7.4}$$

*where the disturbances* $\mathbf{e}$ *are distributed with mean* $\mathbf{0}$ *and covariance* $\boldsymbol{\Sigma}_\mathbf{e}$.

Figure 7.2 shows an example of such a model. Notice that in the model definition only the covariance matrix $\boldsymbol{\Sigma}_\mathbf{e}$ of the disturbances is given. This is because we only consider the second order statistics in our theory: the mean (assumed zero) and the covariance matrix. Thus, the distribution of the disturbances can be arbitrary. Any results apply for example for a multivariate Gaussian disturbance distribution $p_\mathbf{e} = \mathcal{N}(\mathbf{e}; \mathbf{0}, \boldsymbol{\Sigma}_\mathbf{e})$, then, the second order statistics uniquely characterize the distributions and there are no higher order statistics to exploit.

The interpretation of this linear cyclic model follows the description in Section 6 (Cyclic Models). The difference in the sampling process is that now the disturbances are *dependent* and they have to be sampled from their joint distribution. Under suitable stability conditions[2], the covariance of the passive observational distribution is given by

$$\mathbf{C}_\mathbf{x} \quad = \quad (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Sigma}_\mathbf{e}(\mathbf{I} - \mathbf{B})^{-T}. \tag{7.5}$$

This follows directly from Equation 6.3 (p. 68).

One complicating factor here are the possible self-loops, i.e. diagonal elements of $\mathbf{B}$. Such self-loops are inherently unidentifiable from data

[2] Article II assumes asymptotic stability (Assumption 12, p. 68). Article III formulates a slight generalization called *weak stability*. The learning setting of Article IV requires another stability related assumption: *no unit cycles.*

observed in equilibrium (Lacerda et al., 2008). This is why we consider *canonical models* where any self-loops are removed. The self-loops do not affect the equilibrium points, they only affect the convergence rate. Thus, any prediction results we make in the equilibrium are still correct even for models with self-loops.

### 7.2.2 Experimental Effects

Eberhardt et al. (2010) introduced the concept and notion of experimental effect, particularly useful when learning linear models:

**Definition 5 (Experimental Effect)** *An experimental effect $t(x_i \rightsquigarrow x_j || \mathcal{J})$ denotes the observed covariance between an intervened variable $x_i \in \mathcal{J}$ and an observed variable $x_j \in \mathcal{U}$ in the infinite sample limit, when the intervened variables $\mathcal{J}$ are randomized independently with unit variance.*

The experimental effects corresponding to pairs $(x_i, x_j) \in \mathcal{J} \times \mathcal{U}$ (where $\mathcal{J}$ denotes the intervened and $\mathcal{U}$ denotes the passively observed variables in the experiment in question) can be usually estimated from the conducted experiments, even if the randomization was not done independently with unit variance. Experimental effects can even be estimated from soft interventions (Eberhardt et al., 2010).

From the analysis of linear models (Wright, 1921), it follows that the experimental effect $t(x_i \rightsquigarrow x_j || \mathcal{J})$ equals the sum-product of coefficients on all unblocked directed paths from the intervened variable $x_i$ to the observed variable $x_j$. In cyclic models there may be an infinite number of such paths, as we can go through a cycle any number of times. The stability assumption ensures that the sum-product always converges. Consider the experimental effects in an experiment intervening on $x_1$ (Figure 7.2 (b)). The experimental effects are related to the model parameters as follows:

$$\begin{aligned} t(x_1 \rightsquigarrow x_2 || x_1) &= (b_{21} + b_{23}b_{31})(1 + b_{23}b_{32} + (b_{23}b_{32})^2 + \ldots) \\ &= \frac{b_{21} + b_{23}b_{31}}{1 - b_{23}b_{32}}, \qquad (7.6) \\ t(x_1 \rightsquigarrow x_3 || x_1) &= \frac{b_{31} + b_{32}b_{21}}{1 - b_{23}b_{32}}. \qquad (7.7) \end{aligned}$$

Here we have used the familiar geometric sum formula as $|b_{23}b_{32}| < 1$ by the asymptotic stability assumption (Assumption 12, p. 68). The experimental effects are *non-linear* functions of the model parameters. Note that experimental effects are always independent of the covariances between disturbances (such as $\sigma_{23}$ in the previous example). This property makes them useful when learning linear cyclic models in the presence of latent confounding.

| | |
|---|---|
| $\mathcal{J}_1 = \{x_1\}$, $\mathcal{U}_1 = \{x_2, x_3\}$ | $t(x_1 \rightsquigarrow x_2 \| x_1) = b_{21} + t(x_1 \rightsquigarrow x_3 \| x_1) b_{23}$<br>$t(x_1 \rightsquigarrow x_3 \| x_1) = b_{31} + t(x_1 \rightsquigarrow x_2 \| x_1) b_{32}$ |
| $\mathcal{J}_2 = \{x_2\}$, $\mathcal{U}_2 = \{x_1, x_3\}$ | $t(x_2 \rightsquigarrow x_1 \| x_2) = b_{12} + t(x_2 \rightsquigarrow x_3 \| x_2) b_{13}$<br>$t(x_2 \rightsquigarrow x_3 \| x_2) = b_{32} + t(x_2 \rightsquigarrow x_1 \| x_2) b_{31}$ |
| $\mathcal{J}_3 = \{x_3\}$, $\mathcal{U}_3 = \{x_1, x_2\}$ | $t(x_3 \rightsquigarrow x_1 \| x_3) = b_{13} + t(x_3 \rightsquigarrow x_2 \| x_3) b_{12}$<br>$t(x_3 \rightsquigarrow x_2 \| x_3) = b_{23} + t(x_3 \rightsquigarrow x_1 \| x_3) b_{21}$ |

Table 7.2: Linear equation system for solving coefficients $\mathbf{B}$ for a three variable model. All coefficients $b_{ji}$ are unknown, experimental effects denoted by $t(x_i \rightsquigarrow x_j \,||\, x_i)$ can be estimated from the experimental data sets.

### 7.2.3 The Basic Learning Algorithm

Although generally one estimated experimental effect puts a non-linear constraint on the model coefficients $b_{ji}$ (Equations 7.6 and 7.7), many observed experimental effects can be used to construct also simpler, linear constraints:

$$t(x_1 \rightsquigarrow x_3 || x_1) = b_{31} + t(x_1 \rightsquigarrow x_2 || x_1) b_{32}. \tag{7.8}$$

The formula counts the sum-product of all directed paths $x_1 \rightsquigarrow x_3$ by considering the paths in two separate sets. First, the contribution of paths that do not go through $x_2$ is simply the edge coefficient $b_{31}$. Second, the contributions of paths that do go through $x_2$ can be shown to sum up to $t(x_1 \rightsquigarrow x_2 || x_1) b_{32}$. When this constraint is used to learn the model the experimental effects $t(x_1 \rightsquigarrow x_3 || \{x_1, x_2\})$ and $t(x_1 \rightsquigarrow x_4 || \{x_1, x_2\})$ are considered to be estimated from the data, while the coefficients $b_{31}$ and $b_{32}$ are unknown model parameters. Then, a model can be identified when enough such linear constraints have been gathered in the conducted experiments.

This idea[3] is exploited in the `Direct-LLC` (Linear Latents Cyclic) learning algorithm. We explain the idea behind this algorithm here with a simple example. Consider learning a three variable model from the combination of passive observational data and three experiments where always a different variable is intervened on. Note that we fix the structure of the learned model to be completely connected when learning the parameters $(\mathbf{B}, \mathbf{\Sigma_e})$. The correct structure can be later estimated by pruning any edges corresponding to close to zero parameter values.

In the first step the coefficient matrix $\mathbf{B}$ is estimated by forming a linear equation system on the coefficients. The equations of this system are shown

[3] Eberhardt et al. (2010) formulated their original version of the `LLC`-algorithm by forming linear constraints on the total effects $t_{ji} = t(x_i \rightsquigarrow x_j || x_i)$ and solving for coefficients $b_{ji}$ only later. This resembles closely to `Overlapping-LLC` of Section 7.2.5.

in Table 7.2. Each experiment induces two equations. The total number of equations $3 \times 2 = 6$ equals the number of unknown elements in $\mathbf{B}$, since we are assuming that the diagonal is zero. Under suitable stability conditions, the system has a unique solution, and hence the $\mathbf{B}$-matrix can be identified.

In the second step we estimate the covariance matrix $\mathbf{\Sigma_e}$. Since the matrix $\mathbf{B}$ is now known we can solve $\mathbf{\Sigma_e}$ from Equation 7.5:

$$\mathbf{\Sigma_e} = (\mathbf{I} - \mathbf{B})\mathbf{C_x}(\mathbf{I} - \mathbf{B})^T, \tag{7.9}$$

where $\mathbf{C_x}$ is the passively observed covariance matrix. The `Direct-LLC`-algorithm uses only linear equations so it is very efficient compared to many other causal discovery methods.

### 7.2.4 Identifiability & Completeness

What other sets of experiments would allow the `Direct-LLC`-algorithm to identify the model parameters? It turns out that we can formulate intuitive requirements on the set of experiments, that are not only sufficient but also necessary for the model identification. The identifiability properties of linear cyclic models with latent variables are characterized by the following conditions.

**Definition 6 (Ordered Pair Condition)** *An experiment satisfies the ordered pair condition for an ordered pair of variables $(x_i, x_j) \in \mathcal{V} \times \mathcal{V}$ if and only if $x_i$ is intervened on and $x_j$ is passively observed.*

**Definition 7 (Covariance Condition)** *An experiment satisfies the covariance condition for an unordered pair of variables $\{x_i, x_j\} \subseteq \mathcal{V}$ if and only if both variables are passively observed.*

Already Eberhardt et al. (2010) realized that experiments satisfying the ordered pair condition for all pairs are sufficient for identifying $\mathbf{B}$. In articles II and III we show that a linear cyclic model with latent variables is identified if and only if the set of experiments satisfies the ordered pair condition for all ordered pairs of variables and the covariance condition for all unordered pairs of variables. In particular, the proofs show that the formulated equation system for coefficients $\mathbf{B}$ has only a single solution, provided that the model satisfies the stability conditions.

Furthermore we show that the `Direct-LLC`-method is *complete*: it identifies as many coefficients as it is possible from the covariance information

| | |
|---|---|
| $\mathcal{J}_1 = \{x_1, x_2\}$, $\mathcal{U}_1 = \{x_3\}$ | $t_{31} = t(x_1 \rightsquigarrow x_3 \| \| x_1, x_2) + t_{21} t(x_2 \rightsquigarrow x_3 \| \| x_1, x_2)$<br>$t_{32} = t(x_2 \rightsquigarrow x_3 \| \| x_1, x_2) + t_{12} t(x_1 \rightsquigarrow x_3 \| \| x_1, x_2)$ |
| $\mathcal{J}_2 = \{x_1\}$, $\mathcal{U}_2 = \{x_2\}$ | $t_{21} = t(x_1 \rightsquigarrow x_2 \| \| x_1)$ |
| $\mathcal{J}_2 = \{x_2\}$, $\mathcal{U}_2 = \{x_1\}$ | $t_{12} = t(x_2 \rightsquigarrow x_1 \| \| x_2)$ |
| $\mathcal{J}_3 = \{x_3\}$, $\mathcal{U}_3 = \{x_1, x_2\}$ | $t_{13} = t(x_3 \rightsquigarrow x_1 \| \| x_3)$<br>$t_{23} = t(x_3 \rightsquigarrow x_2 \| \| x_3)$ |

Table 7.3: Linear equation system for total effects $\mathbf{T}$ for a three variable model from data sets with partial overlap. All total effects $t_{ji}$ are unknown, experimental effects denoted by $t(x_i \rightsquigarrow x_j || \mathcal{J})$ can be estimated from the experimental data sets.

in the given set of experiments.[4] We also consider the underdetermination left in the model coefficients when the identifiability conditions are not met. For example, if the ordered pair condition is not satisfied for a pair $(x_i, x_j)$, then coefficients on all edges into $x_j$ are generally unidentified. This characterizes the equivalence class of models capable of producing the observed data in the given set of experiments.

Another interesting observation is that given any two experiments intervening on $\mathcal{J}_1$ and $\mathcal{J}_2$ respectively, we can deduce the second order statistics in the *union experiment*, where we intervene on $\mathcal{J}_1 \cup \mathcal{J}_2$ (Lemma 9 in Article III). With an additional passive observational data set we can also calculate the statistics in the *intersection experiment*, where variables $\mathcal{J}_1 \cap \mathcal{J}_2$ are intervened on. This means that in order to predict in a novel experimental setting we do not always have to learn the full model. Such predictions without learning the full model may be more accurate, as noisy data irrelevant to the prediction task are disregarded.

### 7.2.5 Overlapping Experimental Data Sets

In Article IV we considered learning linear cyclic models with latent variables from several data sets that do not all share the same variables. This resembles the situation when learning from overlapping data sets in Section 5.2 (p. 59), but we consider the data sets to be experimental. Notice that in this setting we are not assuming joint causal sufficiency: there can be latent confounding variables that are not observed in any of the data sets.

[4] Article III shows completeness by noting that the coefficients are underdetermined if the ordered pair condition is not satisfied for all pairs. Article IV shows a slightly stronger result that is only implicit in Article III: if `Direct-LLC` leaves a coefficient underdetermined, it is inherently underdetermined.

The left column in Table 7.3 shows a set of partially overlapping experimental data sets. For example, in the second experiment $x_1$ is intervened on and $x_2$ is observed, but $x_3$ is unobserved. Unfortunately, here linear equations on the coefficients $\mathbf{B}$ are not generally possible. For example, the experimental effect $t(x_1 \rightsquigarrow x_3 || x_1)$ needed for Equation 7.8 is unobserved. However, we can form equations on the total effects $t_{ji}$ (which correspond to experimental effects $t(x_i \rightsquigarrow x_j || x_i)$). Table 7.3 shows such a system. Assuming stability, the system can be solved for all total effects $t_{ji}$. These solved total effects can be substituted into the linear system in Table 7.2 and solve for coefficients $\mathbf{B}$. The general algorithm performing this inference[5] is `Overlapping-LLC`.

In Article IV we show that satisfying the ordered pair condition for all pairs of variables is sufficient and worst case necessary for full identification of the coefficient matrix $\mathbf{B}$. Furthermore, the `Overlapping-LLC`-algorithm is shown to be *complete*.

### 7.2.6 Exploiting the Faithfulness Assumption

The rather demanding identifiability conditions suggest that we can rarely learn the full linear cyclic model with latent variables. Especially a set of partially overlapping data sets is not likely to satisfy the ordered pair condition for all pairs of variables. Given that the conditions for identifiability are sufficient and necessary we have to add assumptions to achieve more powerful learning results from the data at hand. One assumption commonly made in causal discovery is causal faithfulness (Assumption 9, p. 30). The general idea of exploiting faithfulness here is to perform independence tests in a similar fashion as done by `PC` and `FCI` -algorithms (Section 3.1.4 and 5.1) and deduce constraint equations that can be added to the linear equation systems.

Consider passively observing $x_i \perp\!\!\!\perp x_j$. Faithfulness ensures that there cannot be any unblocked paths between the variables $x_i$ and $x_j$. A special case of such paths would be edges $x_i \rightarrow x_j$ and $x_j \rightarrow x_i$. Article II considered adding constraints of the following form into the linear equation system:

$$b_{ji} = 0, \tag{7.10}$$

$$b_{ij} = 0. \tag{7.11}$$

[5] This algorithm happens to resemble the original formulation of the `LLC`-algorithm for fully observed experiments presented by Eberhardt et al. (2010). By then, the formulation of linear equations directly on the model coefficients was not yet understood.

The presented algorithm `Faithful-LLC` also uses some additional constraints that we were able to formulate as linear equations on the coefficients $\mathbf{B}$.

The previous constraints do not exhaust the information of a detected independence relation $x_i \perp\!\!\!\perp x_j$. If we place constraints on general experimental effects the independence implies the following constraints:

$$\begin{aligned}
t(x_i \rightsquigarrow x_j || x_i) &= 0, && (7.12)\\
t(x_j \rightsquigarrow x_i || x_j) &= 0, && (7.13)\\
t(x_i \rightsquigarrow x_k || x_i) t(x_k \rightsquigarrow x_j || x_k) &= 0, \quad (\forall x_k \in \mathcal{V} \setminus \{x_i, x_j\}) && (7.14)\\
t(x_k \rightsquigarrow x_i || x_k) t(x_k \rightsquigarrow x_j || x_k) &= 0. \quad (\forall x_k \in \mathcal{V} \setminus \{x_i, x_j\}) && (7.15)
\end{aligned}$$

Equations 7.12, 7.13 and 7.14 are implied by the absence of any directed paths between the variables. Equation 7.15 follows from the fact that $x_i$ and $x_j$ cannot be confounded by $x_k$. Furthermore, the intervention sets in each experimental effect in each equation can be extended to any supersets. For example we have that

$$t(x_i \to x_j || \mathcal{J}) = 0 \quad \Rightarrow \quad t(x_i \to x_j || \mathcal{K}) = 0, \tag{7.16}$$

where $\mathcal{K} \supset \mathcal{J}$ s.t. $x_j \notin \mathcal{K}$. This is because the paths contributing to the experimental effect $t(x_i \to x_j || \mathcal{K})$ are a subset of paths contributing to the experimental effect $t(x_i \to x_j || \mathcal{J})$. According to the equation on the left there are no paths contributing to $t(x_i \to x_j || \mathcal{J})$, so there cannot be any paths contributing to $t(x_i \to x_j || \mathcal{K})$. Enforcing all of the implied equations would allow us to exploit the different faithfulness constraints quite extensively. For example the constraints of `Faithful-LLC` are included: Equation 7.10 is a consequence of Equations 7.12 and 7.16.

However, the previous constraints (such as the ones in Equations 7.12-7.15) cannot be directly inputted to the systems of equations constraining the coefficients $\mathbf{B}$, except in a few special cases. Article IV represents two methods for exploiting the constraints in the overlapping experimental data sets setting: `Bilinear Approach` and `Linear Inference`. The better performing algorithm, `Linear Inference`, is based on heavy use of the vast number of equations relating different experimental effects, only some which are presented in this section.

The power of this algorithm is highlighted in Figure 7.3. Notice that all causal arcs between $x_1$ and $x_3$ are discovered to be absent, although variables $x_1$ and $x_3$ are not observed together in either of the observed data sets.

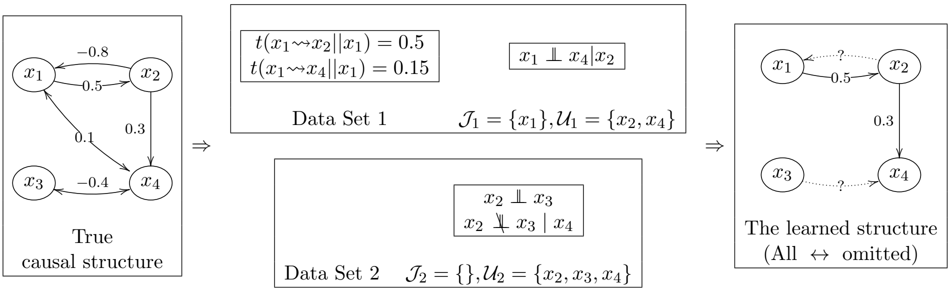

−0.8
$x_1$
$x_2$
0.5
0.3
0.1
$x_3$
−0.4
$x_4$
True causal structure
⇒
$t(x_1 \rightsquigarrow x_2 || x_1) = 0.5$
$t(x_1 \rightsquigarrow x_4 || x_1) = 0.15$
$x_1 \perp\!\!\!\perp x_4 | x_2$
Data Set 1
$\mathcal{J}_1 = \{x_1\}, \mathcal{U}_1 = \{x_2, x_4\}$
$x_2 \perp\!\!\!\perp x_3$
$x_2 \not\perp\!\!\!\perp x_3 \mid x_4$
Data Set 2
$\mathcal{J}_2 = \{\}, \mathcal{U}_2 = \{x_2, x_3, x_4\}$
⇒
?
0.5
0.3
?
The learned structure
(All ↔ omitted)

Figure 7.3: Learning from several experimental data sets with overlapping variables. On the left the true data generating model is displayed. In the middle boxes the most important observations in the two data sets are highlighted. The right panel shows the causal structure over the union of observed variables learned by the `Linear Inference` -algorithm (omitting any double-headed edges). Dotted edges indicate undetermined features.

In Article II we also suggest an experiment selection procedure similar to considerations of active learning (Section 4.2.3), where the experiments are conducted in a sequence and the results of the previously run experiments can influence the next chosen experiment. We consider the independence induced faithfulness constraints to satisfy the ordered pair condition for some appropriate pairs. For example, finding $x_1 \perp\!\!\!\perp x_2$ in passive observational data implies that $b_{12} = b_{21} = 0$ (under faithfulness) and thus the ordered pair condition can be considered to be satisfied for pairs $(x_1, x_2)$ and $(x_2, x_1)$. Then, the experiment that is guaranteed to satisfy the ordered pair condition for most new pairs is always conducted. Such a greedy procedure does not guarantee the identifiability with the least set of experiments in the worst case, when no independence relations are found in the experiments. But more often, the additional constraints due to faithfulness save the few extra experiments this greedy procedure would perform in the worst case.

### 7.2.7 Discussion

The main contributions of this line of research were theoretical identifiability and completeness results and the derived learning algorithms in the very challenging and general learning setting. Improving the algorithms to more optimally handle the uncertainty due to finite sample data through the use of a more Bayesian or maximum likelihood based approach would perhaps be possible. In all articles we apply the methods to the problem of structure learning from finite number of samples: we identify the sig-

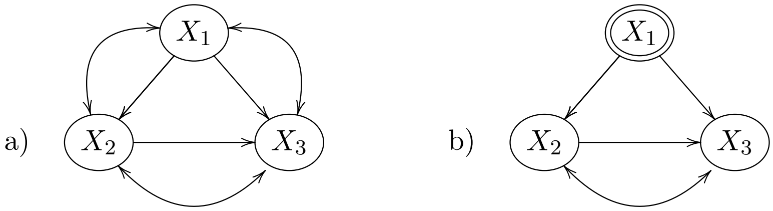


Figure 7.4: (a) A complete Noisy-OR model with latent confounding. (b) The same model when variable $X_1$ is intervened on.

nificantly non-zero coefficients corresponding to present edges for example by using resampling based approaches. Penalized maximum likelihood or Bayesian versions of the algorithms may give more robust structure estimates. However, all of this may limit the size of the models we can learn. The algorithms exploiting faithfulness constraints could also be developed more towards completeness. Maybe, we could exploit these faithfulness constraints also without assuming the linearity of the causal relations.

## 7.3 Noisy-OR Models with Latent Confounding

In Article V we also consider causal discovery in the presence of latent confounding variables using experiments and parametric restrictions. The article presents a model class called noisy-OR models with latent confounding. Interestingly, the identifiability properties of the models in this class are also characterized by the ordered pair condition (Definition 6). The parametric restriction considered constrains the structural equations to be noisy-OR expressions[6] (Peng and Reggia, 1986; Pearl, 1988).

**Definition 8 (Noisy-OR Model with Latent Confounding)** *A Noisy-OR Model with Latent Confounding is a structural equation model over binary variables $X_1, \ldots, X_n$ where each structural equation obeys the form*

$$X_i \quad := \quad \Big( \bigvee_{X_j \in \mathrm{pa}(X_i)} (X_j \wedge B_{ij}) \Big) \vee E_i. \tag{7.17}$$

*The* links *$B_{ji}$ are distributed independently $P(B_{ji} = 1) = b_{ji} > 0$ and the* leaks *$E_1, \ldots, E_n$ have a free distribution $P(E_1, \ldots, E_n)$. The corresponding graph $\mathcal{G}$ defining the parent relations is acyclic.*

[6] Cozman (2004) shows two sets of reasonable axioms on causal relations that lead to noisy-OR conditional probability distributions.

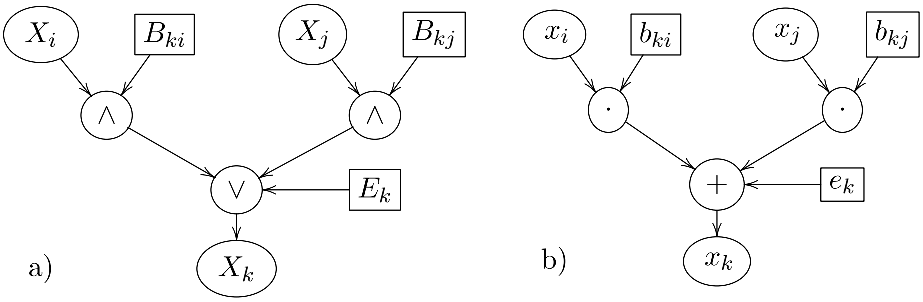


Figure 7.5: Connection between linear SEM equations and noisy-OR SEM equations.

Figure 7.4 shows the graph structure of a noisy-OR model with latent confounding. Similarly as for linear cyclic models with latent variables we allow the leaks $E_1, \dots, E_n$ to be jointly dependent to account for latent confounding. In Figure 7.4, all pairs of observed variables have double headed edges between them to represent this fact. The probability distributions in the passive observational and experimental settings are somewhat complicated and not presented here. The sampling process is anyway simple: first we sample the configuration of the leaks $E_1, \dots, E_n$ from their joint distribution, then the links $B_{ij}$ from their independent distributions and finally we calculate the values of the observed variables $X_1, \dots, X_n$ from the structural equations in the causal order.

The structural equation in Equation 7.17 is further illustrated in Figure 7.5 (a). The parent $X_i$ has the tendency of turning the child $X_k$ ON but this effect can be suppressed if the corresponding link $B_{ki}$ is OFF. The value of the child is an OR-expression of the parents, but the leak $E_k$ can turn the child ON, even if all parents are OFF. Note that the structural equations are quite similar to the linear structural equations (Figure 7.5 (a) and (b) ): the leaks of the noisy-OR expression seem to correspond to the disturbances of the linear equation, ORs correspond to additions and finally ANDs correspond to multiplications. In both models the influence of each parent is aggregated by a deterministic function to produce the value of the child (see Heckerman and Breese (1994) for a discussion on this type of models).

Unfortunately, a similar type of linear mediation of causal effects as for linear models does not occur in these noisy-OR models. However, another property, a type of context specific independence (Section 2.1, p. 10), can be used to identify the causal model. The property implies that if the parents $X_i$ and $X_j$ of some variable $X_k$ are independent given some set $\mathcal{C}$, then further conditioning on their child $X_k$ being zero does not destroy this independence. The situation can be illustrated using Figure 7.5 (a). Assume that $X_i \perp\!\!\!\perp X_j$. Now if $X_k = 0$, the we can deduce that $E_k = 0$, $X_j \wedge B_{kj} = 0$ and $X_i \wedge B_{ki} = 0$. Under these restrictions knowing the value of $X_j$ still does not aid in prediction of $X_i$, hence $X_i$ and $X_j$ are still independent. If we had conditioned on $X_k = 1$, then $X_i = 0$ would increase the probability of $X_j = 1$ as something must turn $X_k$ ON.

The general problem when discovering causal models under latent confounding is that conditioning on variables generally induces dependencies through the latent confounders. The context specific independence property of noisy-OR expressions has the consequence that conditioning on nodes being zero does not induce any dependencies through the confounders. In Figure 7.4 (b), conditioning on $X_2 = 0$ does not unblock the path $X_1 \rightarrow \overline{X_2} \leftrightarrow X_3$. Then, an arc $X_1 \rightarrow X_3$ exists if and only if $X_1 \not\perp\!\!\!\perp X_3 \mid X_2 = 0$ in the distribution where $X_1$ is intervened on. Note that this deduction still requires that $X_1$ is intervened on to break the confounding edge $X_1 \leftrightarrow X_3$.

The basic identifiability result states that the noisy-OR model with latent confounding can be identified if the set of experiments satisfies the ordered pair condition for all pairs and we have an additional passive observational data set. The passive observational data is needed to uncover the distribution of the leaks $P(E_1, \ldots, E_n)$. This condition is also worst case necessary: for each set of experiments not satisfying the ordered pair condition for all pairs there exists models that cannot be distinguished from this set of experiments. In addition Article V shows that these result apply also for slight generalization of the basic noisy-OR models where a cause being OFF may induce its effect to be ON.

Noisy-OR models can be learned in several ways from data sets satisfying the identifiability conditions. The basic idea of the identifiability proof was to condition on various variables being zero. Note however, that such conditioning reduces the available sample size for estimating the model parameters, as we are essentially throwing away samples not in line with the conditioning. The `Efficient Conditioning` -algorithm (EC) conditions on as few variables as possible when estimating the model parameters. A more accurate and robust option is to use the EM-algorithm (Expectation Maximization, Dempster et al. (1977)) for maximizing the likelihood (this is referred to as `Noisy-OR-EM`), but this optimization procedure is slow and can be run only for models with up to 8 variables.

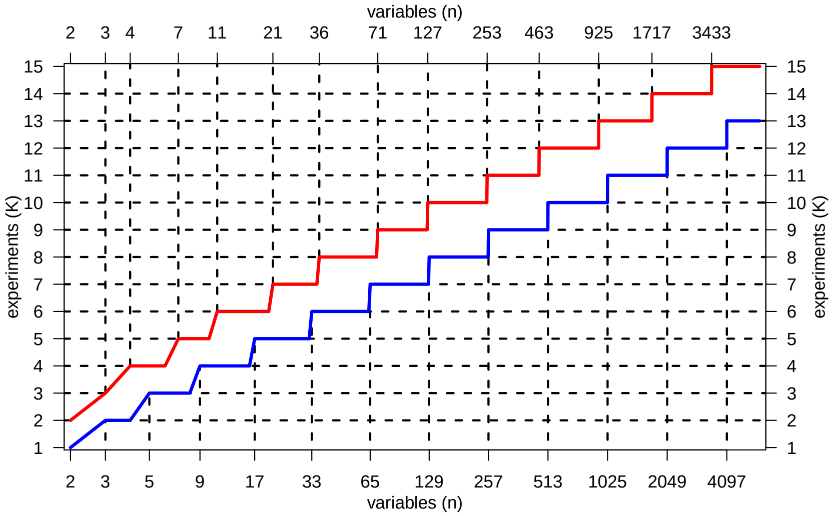


Figure 7.6: Sufficient and necessary number of experiments needed to satisfy the unordered pair condition (blue) and the ordered pair condition (red). The ticks on the x-axis appear only when an additional experiment is needed.

One interesting way of extending this model class would be to allow for cycles in the structure. It would also be interesting to know how good of an approximation the noisy-OR type of expression provides in various situations. Using some parametric form on $P(E_1, \ldots, E_n)$ would allow for faster convergence and more accurate estimation of the model parameters. Another interesting question which remains is finding the general property of the causal models that can be identified from data satisfying the ordered pair condition for all pairs of observed variables.

## 7.4 Experiment Selection

Article VI considers selecting experiments such that the identifiability of various causal models is guaranteed. In particular we show how to select experiments for the models considered in this thesis. The article applies different results previously established by the combinatorics community. The focus is on finding optimal sets of experiments, such that the total number of experiments is minimized. We also consider situations where the number of intervened variables per experiment is constrained.

Our previous articles presented conditions for the experiments sufficient and (worst case) necessary for full model identifiability. They are most easily described in relation to each other when we assume that passive observational data is always available (this trivially satisfies the covariance condition for all variables). Linear cyclic models with latent variables can then be identified if and only if the ordered pair condition is satisfied for all pairs (Section 7.2). Satisfying the ordered pair condition for all pairs is then also sufficient and worst case necessary for fully identifying Noisy-OR models with latent confounding (Section 7.3). In addition, assuming causal sufficiency and faithfulness, acyclic causal models can be identified if and only if the set of experiments satisfies the following unordered pair condition for all unordered pairs (Section 4.2.1, p. 49).

**Definition 9 (Unordered Pair Condition)** *An experiment satisfies the unordered pair condition for an unordered pair of variables $\{x_i, x_j\} \subseteq \mathcal{V}$ if and only if either $x_i$ is intervened on and $x_j$ is passively observed, or $x_j$ is intervened on and $x_i$ is passively observed.*

The unordered and ordered pair conditions seem intuitive enough to characterize the identifiability of causal models in other types of settings and under different assumptions as well. For example, three experiments intervening on

$$\mathcal{J}_1 = \{x_2, x_4, x_6\}, \mathcal{J}_2 = \{x_3, x_4, x_7\}, \mathcal{J}_3 = \{x_5, x_6, x_7\}$$

satisfy the unordered pair condition for all unordered pairs among seven variables, while four experiments intervening on

$$\mathcal{J}_1 = \{x_1, x_2, x_3\}, \mathcal{J}_2 = \{x_1, x_4, x_5\}, \mathcal{J}_3 = \{x_2, x_4, x_6\}, \mathcal{J}_4 = \{x_3, x_5, x_6\}$$

satisfy the ordered pair condition for all ordered pairs among six variables.

Figure 7.6 shows the number of experiments necessary and sufficient for satisfying either pair condition for all pairs. These numbers along with the procedures for finding the actual sets of experiments can be found in the combinatorics literature under the term (completely) separating systems. Note that the x-axis is in logarithmic scale, so the number of experiments needed for full identification grows only logarithmically with increasing number of variables. Another observation is that only a few more experiments are needed to satisfy the ordered pair condition compared to the unordered pair condition.

Although relatively few experiments are needed to satisfy the pair conditions, the experiments for these optimal designs may be unrealistic: the experiments require us to intervene on quite many variables at the same time. For example the optimal design that uses 13 experiments for satisfaction of the ordered pair condition for all pairs of 1024 variables, needs to intervene on average on more than 380 variables per experiment (Article VI, Figure 9 middle). To find more realistic experiments, Article VI also describes procedures that output sets of experiments such that the (unordered or ordered) pair condition is satisfied for all pairs, but the number of intervened variables per experiment is bounded or minimized. For example there exists a set of 50 experiments with at most 41 intervened variables per experiments that satisfies the ordered pair condition for all ordered pairs of 1024 variables (Article VI, Figure 9 top).

Sometimes when learning causal models we may have background knowledge that narrows down the possible causal structures and relations. Such background knowledge often allows satisfaction of the pair conditions for some specific pairs of variables. Then we would like to select the experiments such that the pair conditions are satisfied for *some* variable pairs. Such problem settings have been considered in graph theory. This more complicated setting does not allow for simple construction of optimal experiment selection procedures. Basically, finding the set of experiments with minimum size is NP-hard. Algorithms for finding such sets of experiments may be constructed using, for example, different graph coloring algorithms. If our aim is also to limit the number of intervened variables, the situation is even more complicated. No non-exhaustive solutions for this general problem seem to exist at present.

## 7.5 Contributions of the Present Author

For Article I, the present author derived and applied the general theory to the specific setting based on an original idea by Dr. Hoyer during 2008-2009. The present author also implemented and tested the method. Furthermore, the present author took part in the writing and editing process with Dr. Hoyer.

The theory for the Article II, building on then recently published work of the two other writers of the paper (Eberhardt et al., 2010), was derived in co-operation with a large influence from the present author during 2010. Especially the characterization of the underdetermination, the proof of the sufficient and necessary identifiability conditions, and the proof of completeness originated from the research of the present author. In addition, the present author implemented and tested the learning methods, and participated in the writing process with the other authors. Although some of the results in this paper are presented in more detail in Article III, the article also contains results that are not in Article III, for example on faithfulness and experiment selection.

Article III gives a more detailed journal version of part of the research presented in preliminary form in Article II and by Eberhardt et al. (2010). The present author implemented and tested the learning method, and performed the realistic data analysis. Theoretical details appearing in the theorems and their proofs were also mostly worked out by the present author. The three authors co-wrote the paper.

Article IV applied the previously derived theory appearing in Article III to a new learning setting. The present author was responsible for proving many of the theoretical results. This time all three authors contributed equally to the derivation, implementation and testing of the learning methods. The best performing algorithm (`Linear Inference`) for processing the different faithfulness constraints was conceived, derived and implemented by the present author. The three authors co-wrote the paper in 2012.

Article V was conceived during 2011. From an example of identifiability for a three variable model by the other authors, the present author derived the general identifiability result, and noticed the connection to the context specific independence property of noisy-OR expressions. Especially the possibility for a free distribution for the leaks (instead of a less general form) was found by the present author. The present author derived, implemented and tested the discovery methods. Again the article was co-written by all authors.

The research for Article VI started in autumn 2011 by considerations of the present author on how to optimally select experiments such that the identification of the models used in our previous research would be guaranteed. The previously used experiment selection methods were suboptimal. Although an optimal solution for the simplest scenario was found independently, we also found a rather large range of existing combinatorics research considering equivalent problems. As the connection of this research to problem of causal discovery was not totally straight-forward and largely unknown to the causal discovery community, an article was prepared in spring 2012. A large part of the literature review and the application of the combinatorics results to the experiment selection problem was conducted by the present author. Again, all three authors contributed equally to the writing process. The implementation of the selection procedures was the present author's work.

# Chapter 8

# Conclusion

This thesis presented methods for learning causal relationships from data. Rather than learning specific single causal relationships the aim was on learning the full causal structure among the observed variables and characterizing the underdetermination when the full structure cannot be learned. The learning settings included causally sufficient situations but mainly focused on discovering causal models in the presence of latent confounding. Some of the learning algorithms also allowed for cyclic causal structures. The results suggest that if our aim is to learn the full causal structure in these challenging settings, several experimental data sets and parametric restrictions on the individual causal relations are needed. Care should be taken in selecting the experiments in order to achieve informative learning results from limited number of experiments and interventions.

The introductory part of the thesis included a quite broad description of the field of causal discovery. This was provided in order to show the principles and motivation for the research as well the connections of the original research to the existing literature. The original research papers can be seen to generalize and provide connections between different approaches previously considered, as well as develop new ways of learning causal models.

Perhaps the main contribution of the thesis is the general theory for learning linear cyclic models with latent variables from experimental data sets without assuming faithfulness. We formulated necessary and sufficient conditions on the set of experiments such that the causal model could be fully identified. The ordered pair condition was shown to be the important condition characterizing the identifiability properties and possible underdetermination of the learned model. We also updated, modified and applied the `LLC` learning method to different learning settings, and proved its completeness when faithfulness is not assumed. These different versions of the basic idea exploit linear equations and are thus fast and scalable. Unlike most other algorithms, the presented methods are able to exploit and combine experimental data sets while still allowing for latent confounding. We also considered learning linear cyclic models with latent variables from overlapping experimental data sets. This kind of data has not been commonly exploited previously.

We introduced a new model class called noisy-OR models with latent confounding. We showed that satisfying the ordered pair condition for all pairs is sufficient and worst case necessary also for the identification of the models in this class. Thus, restricting the parametric form of the causal relations to be linear or to follow a noisy-OR parameterization proved to be useful in discovering the presence and absence of causal relations.

In addition to showing how to exploit experimental data efficiently in many settings, we also provided guidelines on selecting the experiments to be conducted. These sets of experiments minimize the number of experiments and the number of variables needed to be intervened on in the experiments. The experiment selection procedures were given in the causally sufficient case and when allowing for latent confounding. For the completeness of the thesis, we thus provided the optimal sets of experiments for the models which we provided learning algorithms for.

The causal faithfulness assumption was a common theme in the research. We showed how to efficiently use the faithfulness assumption when learning linear cyclic models with latent variables. The use of the faithfulness assumption helps to achieve more informative learning results when the underlying structure is sufficiently sparse. On the other hand, some of our methods did not use this assumption, and are thus more applicable for learning denser causal structures, from data sets with limited number of samples. With the use of experiments and parametric restrictions on the causal relationships, we are able to learn models also without the faithfulness assumption.

We also developed the `LiNGAM` method appearing in the causal discovery literature to a more Bayesian direction. This allowed the new method to be robust against situations where the non-Gaussianity assumption fails. Bayesian inference allowed us to output several alternative structures in addition to the most probable one. This research shows a fruitful combination of two somewhat separate causal discovery principles: assigning posterior probabilities for graph structures while still exploiting the information in the higher order statistics of the possibly non-Gaussian data.

The research conducted was theoretic and it had a somewhat theoretical motivation. Various difficulties may arise when applying the methods in application fields because of the simplifying assumptions on the parametric form. However, as we showed in this thesis, without some parametric restrictions not a whole lot of causal relations can be learned, if we still want to allow for latent confounding and cycles. In addition, when learning from a limited amount of data, the causal models with simpler relationships may provide better predictions than more complicated models in which all parameters cannot be estimated accurately. On the other hand, research is often conducted by starting from simpler parametric models and only later aiming at models with fewer parametric constraints. Machine learning has experienced several cases where linear methods have been generalized to allow for non-linear characteristics: principles behind `LiNGAM` led to additive noise models, linear classification methods can now be used with non-linear kernels, and the ideas in the linear models of Wright (1921) led to Bayesian networks. So the hope is that at least some of the ideas and methods in the thesis can be exploited also without the parametric restrictions.

In particular, developing the ideas for exploiting faithfulness induced constraints in the `Linear Inference` -method to a more non-parametric direction deserves more consideration. Such a method would allow for learning causal models in an even more general setting, allowing for the incorporation of background knowledge and experimental data. While we were able to utilize much of the independence information in the `Linear Inference` -procedure, developing the procedure to completely and provably utilize all independence and dependence information is an interesting research direction.

The benefits of the Bayesian approach for learning causal models are clear: no hard decision about the existence or nonexistence of causal relations need to be made. Thus, one future research opportunity would be to develop the different existing methods presented in this thesis and in the original research to a more Bayesian direction. The major challenges include for example the efficiency of the procedures, especially when allowing for cyclic structures and latent confounding.

The identifiability properties of linear cyclic models with latent variables and noisy-OR models with latent confounding were based on the satisfaction of the ordered pair condition. These model classes have their similarities: in both models the direct causes influence the effect through their respective independent stochastic mechanisms; the influences are then aggregated by a deterministic function to produce a value for the effect. But surprisingly, the theoretical properties leading to the essentially same identifiability results were quite different. Finding a general property of models allowing for identifiability when the ordered pair condition is satisfied for all pairs is thus a very interesting question.

# References

Berzuini, C., Dawid, P., and Bernardinelli, L., editors (2012). *Causality, Statistical Perspectives and Applications.* John Wiley & Sons.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons.

Borboudakis, G., Triantafilou, S., Lagani, V., and Tsamardinos, I. (2011). A constraint-based approach to incorporate prior knowledge in causal models. In *Proceedings of the 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 321–326. ESANN.

Borboudakis, G. and Tsamardinos, I. (2012). Incorporating causal prior knowledge as path-constraints in Bayesian networks and maximal ancestral graphs. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1799–1806. Omnipress.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In Fisher, D. and Lentz, H., editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Claassen, T. and Heskes, T. (2010). Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems 23*, pages 415–423.

Claassen, T. and Heskes, T. (2012). A Bayesian approach to constraint based causal discovery. In *Proceedings of the Twenty-Eight Conference on Uncertainty in Artificial Intelligence*, pages 207–216. AUAI Press.

Cooper, G. and Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 116–125. Morgan Kaufmann.

Cooper, G. F. (1999). Overview of the representation and discovery of causal relationships using Bayesian networks. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation & Discovery*, pages 3–62. AAAI / MIT Press.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Cozman, F. G. (2004). Axiomatizing noisy-OR. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 981–982. IOS Press.

Danks, D. (2002). Learning the causal structure of overlapping variable sets. In *Discovery Science*, Lecture Notes in Computer Science 2534, pages 178–191. Springer.

Danks, D. (2005). Scientific coherence and the fusion of experimental results. *British Journal for the Philosophy of Science*, 56:791–807.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

Dawid, A. P. (2010). Beware of the DAG! In *Proceedings of the NIPS 2008 Workshop on Causality*, Journal of Machine Learning Research Workshop and Conference Proceedings 6, pages 59–86.

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Drton, M. and Richardson, T. S. (2004). Iterative conditional fitting for Gaussian ancestral graph models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 130–137. AUAI Press.

Eaton, D. and Murphy, K. (2007). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Workshop and Conference Proceedings 2, pages 107–114.

Eberhardt, F. (2007). *Causation and Intervention.* PhD thesis, Carnegie Mellon University.

Eberhardt, F. (2008a). Almost optimal intervention sets for causal discovery. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 161–168. AUAI Press.

Eberhardt, F. (2008b). Sufficient condition for pooling data from different distributions. *Synthese*, 163:433–442.

Eberhardt, F., Glymour, C., and Scheines, R. (2005). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 178–184. AUAI Press.

Eberhardt, F., Glymour, C., and Scheines, R. (2006). N-1 experiments suffice to determine the causal relations among N variables. In Holmes, D. and Jain, L., editors, *Innovations in Machine Learning, Theory and Applications Series: Studies in Fuzziness and Soft Computing*, pages 97–112. Springer-Verlag.

Eberhardt, F., Hoyer, P. O., and Scheines, R. (2010). Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Workshop and Conference Proceedings 9, pages 185–192.

Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74:981–995.

Entner, D. and Hoyer, P. O. (2011). Discovering unconfounded causal relationships using linear non-Gaussian models. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2010 Workshops*, Lecture Notes in Computer Science 6797, pages 181–195. Springer Berlin Heidelberg.

Eriksson, J. and Koivunen, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604.

Evans, R. J. and Richardson, T. S. (2010). Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 177–184. AUAI Press.

Fisher, F. M. (1966). *The Identification Problem in Economics*. McGraw-Hill.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica*, 38(1):73–92.

Fisher, R. A. (1935). *The design of experiments*. Hafner.

Friedman, N. (1998). The Bayesian structural EM algorithm. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 129–138. Morgan Kaufmann.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall.

Glymour, C., Spirtes, P., and Richardson, T. (1999). Response to rejoinder. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation & Discovery*, pages 343–345. AAAI / MIT Press.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12.

Hauser, A. and Bühlmann, P. (2012a). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.

Hauser, A. and Bühlmann, P. (2012b). Two optimal strategies for active learning of causal models from interventions. In *Proceedings of the The 6th European Workshop on Probabilistic Graphical Models*, pages 123–130.

Hausman, D. M. and Woodward, J. (1999). Independence, invariance and the causal Markov condition. *British Journal of Philosophy of Science*, 50:521–583.

He, Y. and Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547.

Heckerman, D. and Breese, J. S. (1994). A new look at causal independence. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 286–292. Morgan Kaufmann.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistics Association*, 81:945–960.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008a). Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 282–289. AUAI Press.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008b). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley and Sons.

Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010). Learning bayesian network structure using lp relaxations. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Workshop and Conference Proceedings 9, pages 358–365.

Koivisto, M. and Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5):2148–2177.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings* *of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 366–374. AUAI Press.

Meek, C. (1995a). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann.

Meek, C. (1995b). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–418. Morgan Kaufmann.

Meek, C. (1997). *Graphical Models: Selecting Causal and Statistical Models.* PhD thesis, Carnegie Mellon University.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24*, pages 639–647.

Mooij, J. M., Janzing, D., Peters, J., and Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th International Conference on Machine Learning*, pages 745–752. ACM.

Murphy, K. P. (2001). Active learning of causal Bayes net structure. Technical report, U.C. Berkeley.

Neal, R. (2000). On deducing conditional independence from d-separation in causal graphs with feedback. *Journal of Artificial Intelligence Research*, 12:87–91.

Neyman, J. (1927). On the application of probability theory to agricultural experiments: essay on principles. *Rocniki Nauk Rolniiczych*, 10:1–51.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Pearl, J. and Dechter, R. (1996). Identifying independencies in causal graphs with feedback. In *Proceedings of the 12th Conference in Uncertainty in Artificial Intelligence*, pages 420–426. Morgan Kaufmann.

Pearl, J. and Verma, T. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, pages 441–452. Morgan Kaufmann.

Peng, Y. and Reggia, J. (1986). Plausibility of diagnostic hypotheses: The nature of simplicity. In *In Proceedings of the 5th National Conference on AI*, pages 140–145. AAAI.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2011). Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 589–598. AUAI Press.

Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's razor. In *Advances in Neural Information Processing Systems 13*, pages 294 – 300. MIT Press.

Reichenbach, H. (1956). *Direction of Time.* University of California Press.

Richardson, T. and Spirtes, P. (1999). Automated discovery of linear feedback models. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation & Discovery*, pages 253–302. AAAI / MIT Press.

Richardson, T. and Spirtes, P. (2002). Ancestral graph markov models. *Annals of Statistics*, 30(4):962–1030.

Richardson, T. S. (1996). *Feedback Models: Interpretation and Discovery.* PhD thesis, Carnegie Mellon University.

Robins, J. M. and Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation & Discovery*, pages 305–321. AAAI / MIT Press.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Schmidt, M. and Murphy, K. (2009). Modeling discrete interventional data using directed cyclic graphical models. In *Proceedings of the 25th Conference Conference on Uncertainty in Artificial Intelligence*, pages 487–495. AUAI Press.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248.

Silander, T. and Myllymäki, P. (2006). A simple approach to finding the globally optimal Bayesian network structure. In *Proceedings of the 22th Conference on Uncertainty in Artificial Intelligence*, pages 445–452. AUAI Press.

Silva, R. and Ghahramani, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246.

Sober, E. (2001). Venetian sea levels, british bread prices, and the principle of the common cause. *British Journal of Philosophy of Science*, 52:331–346.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 491–498. Morgan Kaufmann.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag. (2nd edition MIT Press 2000).

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. (2012). Estimation of causal orders in a linear non-gaussian acyclic model: A method robust against latent confounders. In *Artificial Neural Networks and Machine Learning – ICANN 2012*, Lecture Notes in Computer Science 7552, pages 491–498. Springer.

Tillman, R. E., Danks, D., and Glymour, C. (2009). Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems 21*, pages 1665–1672.

Tillman, R. E. and Spirtes, P. (2011). Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Workshop and Conference Proceedings 15, pages 3–15.

Tong, S. and Koller, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 863–869. Morgan Kaufmann Publishers.

Triantafillou, S., Tsamardinos, I., and Tollis, I. G. (2010). Learning causal structure from overlapping variable sets. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Workshop and Conference Proceedings 9, pages 860–867.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

Tsamardinos, I., Triantafillou, S., and Lagani, V. (2012). Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research*, 13:1097–1157.

Woodward, J. (2003). *Making Things Happen, A Theory of Causal Explanation.* Oxford University Press.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press.