Department of Computer Science
Series of Publications A
Report A-2013-10

Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond

Doris Entner

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Hall 5 (University Main Building, Fabianinkatu 33) on November 20, 2013, at twelve o'clock.

University of Helsinki
Finland

Supervisor
Patrik O. Hoyer, University of Helsinki, Finland

Pre-examiners
Joris Mooij, University of Amsterdam, The Netherlands
Ilya Shpitser, University of Southampton, United Kingdom

Opponent
Kun Zhang, Max Planck Institute for Intelligent Systems, Tübingen, Germany

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: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/
Telephone: +358 9 1911, telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-9406-4 (paperback)
ISBN 978-952-10-9407-1 (PDF)
Computing Reviews (1998) Classification: G.3, G.4, I.2.6
Helsinki 2013
Unigrafia

Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond

Doris Entner

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
entnerd@hotmail.com
http://www.cs.helsinki.fi/u/entner/

PhD Thesis, Series of Publications A, Report A-2013-10
Helsinki, November 2013, 79 + 113 pages
ISSN 1238-8645
ISBN 978-952-10-9406-4 (paperback)
ISBN 978-952-10-9407-1 (PDF)

Abstract

In many fields of science, researchers are keen to learn *causal* connections among quantities of interest. For instance, in medical studies doctors want to infer the effect of a new drug on the recovery from a particular disease, or economists may be interested in the effect of education on income.

The preferred approach to causal inference is to carry out controlled experiments. However, such experiments are not always possible due to ethical, financial or technical restrictions. An important problem is thus the development of methods to infer cause–effect relationships from *passive observational* data. While this is a rather old problem, in the late 1980s research on this issue gained significant momentum, and much attention has been devoted to this problem ever since. One rather recently introduced framework for causal discovery is given by linear non-Gaussian acyclic models (LiNGAM). In this thesis, we apply and extend this model in several directions, also considering extensions to non-parametric acyclic models.

We address the problem of causal structure learning from time series data, and apply a recently developed method using the LiNGAM approach to two economic time series data sets. As an extension of this algorithm, in order to allow for non-linear relationships and latent variables in time series models, we adapt the well-known Fast Causal Inference (FCI) algorithm to such models.

We are also concerned with non-temporal data, generalizing the LiNGAM model in several ways: We introduce an algorithm to learn the causal structure among multidimensional variables, and provide a method to find pairwise causal relationships in LiNGAM models with latent variables. Finally, we address the problem of inferring the causal effect of one given variable on another in the presence of latent variables. We first suggest an algorithm in the setting of LiNGAM models, and then introduce a procedure for models without parametric restrictions.

Overall, this work provides practitioners with a set of new tools for discovering causal information from passive observational data in a variety of settings.

Computing Reviews (1998) Categories and Subject Descriptors:
G.3 [Probability and Statistics]: Correlation and regression analysis, Multivariate statistics, Time series analysis
G.4 [Mathematical Software]: Algorithm design and analysis
I.2.6 [Artificial Intelligence]: Learning - Parameter learning

General Terms:
Algorithms, Theory

Additional Key Words and Phrases:
Machine Learning, Causality, Graphical Models, Passive Observational Data, Latent Variables, Non-Gaussianity

Acknowledgements

First and foremost, I thank my supervisor Patrik Hoyer, without whose advice, support and patience this thesis had not been possible. In particular, the right mixture of freedom and guidance in conducting research as well as an always-open office door for clarifying and inspiring discussions made my time as Ph.D. student successful and enjoyable.

I am grateful to the neuroinformatics research group for the good working atmosphere as well as scientific and non-scientific discussions over lunch, in our meetings and study groups. I also thank my co-authors Peter Spirtes, Alessio Moneta and Alex Coad for the fruitful collaboration, significantly adding to this thesis.

For valuable comments on this manuscript I am indebted in particular to Patrik Hoyer, who repeatedly and untiringly read the draft, as well as to Michael Gutmann and Antti Hyttinen, and the two pre-examiners Joris Mooij and Ilya Shpitser.

The Department of Computer Science and the Helsinki Institute for Information Technology (HIIT) provided a great working and studying environment. Both these institutions as well as the Academy of Finland and the Helsinki Graduate School in Computer Science and Engineering (HeCSE) financially supported my studies and trips to summer schools, conferences and research visits.

A big thanks goes to my colleagues and friends for their entertainment outside of work, in particular for the weekly badminton games and Friday-night drinks and dinners, but also the many other activities.

I am much obliged to my parents Helga and Helmut for their support throughout my life, making it possible to fulfill my dreams. I thank my whole family as well as my friends back home for always welcoming me on my visits to Austria making it easy to recharge my batteries and enjoy my holidays.

Finally, I thank Dennis for supporting and encouraging me throughout my Ph.D. studies, being my personal IT-support, distracting me from work and simply being there.

Contents

List of Symbols

$\mathbf{A}$	connection matrix in reduced form of linear SEMs
$\mathbf{A}_i$	connection matrices in VAR models, $i = 1, \ldots, q$
$\mathbf{B}$	connection matrix in linear SEMs
$\mathbf{B}_i$	connection matrices in SVAR models, $i = 0, \ldots, q$
$\mathrm{cov}(\boldsymbol{v}_1, \boldsymbol{v}_2)$	matrix of covariances of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$
e	scalar disturbance or error term
$\boldsymbol{e}$	multidimensional disturbance or error term
$\mathrm{E}(\boldsymbol{v})$, $\mu_{\boldsymbol{v}}$	expected value of $\boldsymbol{v}$, mean of $\boldsymbol{v}$
$\mathcal{E}$	set of edges in a graph
$\mathcal{G}$	graph
$\mathbf{I}$	identity matrix
K	causal order of variables
p	probability distribution
pa_i, pa_x	parent set of variable v_i, or x, respectively
q	order of a VAR or SVAR model, number of time-lags
r	scalar residual in a regression model
$\boldsymbol{r}$	multidimensional residual in a regression model
$\rho_{\boldsymbol{v}_1,\boldsymbol{v}_2}$	matrix of correlations of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$
$\rho_{\boldsymbol{v}_1,\boldsymbol{v}_2 \cdot \boldsymbol{v}_3}$	matrix of partial correlations of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ given $\boldsymbol{v}_3$
σ_v^2	variance of v
$\boldsymbol{\Sigma}_{\boldsymbol{v}}$	covariance matrix of $\boldsymbol{v}$
$\mathcal{U}$	set of latent variables
$\mathcal{V}$	set of vertices in a graph, set of variables
v, v_i	scalar random variable
$\boldsymbol{v}$, $\boldsymbol{v}_i$	multidimensional random variable
$\mathcal{W}$	set of observed variables
x	scalar random variable denoting the cause
$\boldsymbol{x}$	multidimensional random variable denoting the cause
y	scalar random variable denoting the effect
$\boldsymbol{y}$	multidimensional random variable denoting the effect
$\mathcal{Z}$	subset of the observed variables $\mathcal{W}$
$\perp\!\!\!\perp_p$	statistically independent in probability distribution p
$\not\perp\!\!\!\perp_p$	statistically dependent in probability distribution p
$\perp\!\!\!\perp_{\mathcal{G}}$	d-separated in graph $\mathcal{G}$
$\not\perp\!\!\!\perp_{\mathcal{G}}$	not d-separated in graph $\mathcal{G}$
$\prec$	relation in causal order: $v_i \prec v_j$ means that v_i is prior to v_j in the causal order

List of Abbreviations

CBN	Causal Bayesian Network
DAG	Directed Acyclic Graph
FCI	Fast Causal Inference
ICA	Independent Component Analysis
LiNGAM	Linear Non-Gaussian Acyclic Model
lvLiNGAM	latent variable LiNGAM
MAG	Maximal Ancestral Graph
OLS	Ordinary Least Squares
PAG	Partial Ancestral Graph
SEM	Structural Equation Model
SVAR	Structural Vector Autoregression
VAR	Vector Autoregression

Chapter 1

Introduction

In the field of machine learning and statistics, scientists are commonly interested in inferring regularities and features concerning the real world from data. To model the real world, one may often assume that everything follows rules (like physical laws), and that the data (i.e. observations) are generated according to these rules. Researchers are then interested in learning (parts of) this data generating process, or certain characteristics of it, from the available observations.

This thesis is concerned with the subfield of *causal discovery*, aiming at learning cause-effect relationships from data. In this chapter, we first discuss the concept of causality and demonstrate the general problems of inferring causal relationships from data by means of examples. We then pose two main research questions in the field of causality, parts of which are addressed in this thesis, give an overview of the organization of the rest of this document, and list the original publications on which this thesis is based.

1.1 Correlation, Causation, and Interventions

The specific topic we address in this thesis is how to learn *causal* relationships among variables of interest. One central observation is that a correlation or dependence between two variables typically results from any (combination) of several causal relationships, as stated in Reichenbach's (1956) *principle of the common cause*: A correlation or dependence between two variables x and y usually indicates that x causes y, or y causes x, or x and y are joint effects of a common cause. This is demonstrated in the following two examples.

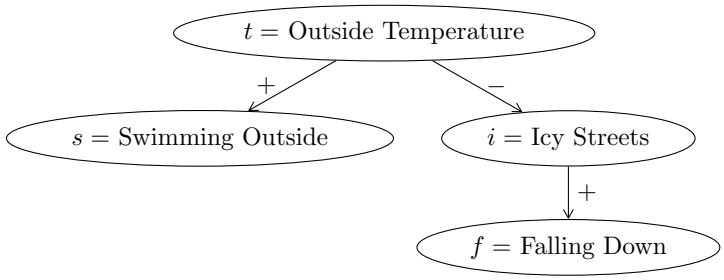


Figure 1.1: A graph depicting causal relationships. The variables are assumed to be binary, so that t can take the values 'low' and 'high', and all other variables can take the values 'yes' and 'no'. For details see Examples 1.1, 1.2, and 1.3.

Example 1.1 (Correlation due to a Cause-Effect Relationship)**.** *Consider the subgraph over the two variables i and f in Figure 1.1, showing the data generating process of $i =$'icy streets' and $f =$'falling down'. The arrow from i to f depicts a* direct causal effect*, i.e. i is the cause and f is the effect. The '+' indicates a positive causal effect, since the probability of falling is greater when the streets are icy. This implies a positive correlation between i and f.*

The joint probability distribution over i and f, $p(i, f)$, can be factorized in two ways, $p(i)\,p(f \mid i)$ and $p(f)\,p(i \mid f)$, both representing a statistical dependence *between i and f. Intuitively, only the former factorization corresponds to the data generating process represented by the graph $i \to f$: The value of the cause i is first sampled from $p(i)$, independently of f. Secondly, the value of the effect f is sampled from $p(f \mid i)$, which depends on the value of i.*

Example 1.2 (Correlation due to a Common Cause)**.** *In the subgraph over the variables t, s, and i in Figure 1.1, the data generating process of the variables $i =$'icy streets', $s =$'swimming outside', and $t =$'outside temperature' is depicted. While t has a positive causal effect on s (if the temperature is high, people are likely to swim outside), it has a negative effect on i (if the temperature is low, streets are likely to be icy). As the data generating process shows, there is no causal effect of i on s, nor of s on i.*

Nevertheless, there is a negative correlation between i and s: Observing people swimming outside suggests that the streets are not icy. Since this correlation is not due to a direct cause-effect relationship, but due to the common cause t, the correlation is called spurious.

These two examples illustrate that knowledge solely of correlations among the variables (or a probability distribution over them) is not enough to infer causal relationships. However, as the influential work of Spirtes et al. (1993) and Pearl (2000) showed, with appropriate assumptions on the data generating process, as discussed in detail in later parts of this thesis, this may well be possible.

The major difference between correlation and causation is that the former is *symmetric*, i.e. if variable x is correlated with variable y, then y is correlated with x. Causation, on the other hand, is (typically) *antisymmetric*: if x is a cause of y, then y is not a cause of x. Spirtes et al. (1993) stated that causation is usually, in addition to antisymmetric, also transitive (if x is a cause of y, and y is a cause of z, then x is an (indirect) cause of z, see Example 1.3), and irreflexive (a variable is not a cause of itself).

Example 1.3 (Transitivity of Causation, Direct and Indirect Causes)**.** *The arrows in the graph of Figure 1.1 represent* direct *causal relationships, such that t is a direct cause of i, and i a direct cause of f, with regard to the variable set $\{t, s, i, f\}$. By transitivity, t is an (indirect) cause of f.*

A key tool to discover cause-effect relationships are *interventions*: Intervening on the cause by *setting* it to a certain value (as opposed to merely observing this variable at that value) influences the value of the effect. However, intervening on the effect has no impact on the value of the cause. Thus, interventions break the symmetry of correlation, and add a direction to it.

Example 1.4 (Interventions)**.** *In Examples 1.1 and 1.2, by intervening on $i =$'icy streets', for instance by building a heating or cooling system beneath the streets, we are able to distinguish between causation and absence of causation. In Example 1.1, when turning the heating or cooling system on, the value of the variable $f =$'falling down' is affected: For instance, if we make sure (by intervention) that the streets are not icy, people are less likely to fall. This allows us to infer that i is a cause of f. In Example 1.2, on the other hand, the variable $s =$'swimming outside' will not be affected by the value of the icy street under the intervention, which implies that i is not a cause of s. Furthermore, in the former example, intervening on $f =$'falling down' (for example by building traps), the value of i would not change and hence, f is not a cause of i.*

More realistic applications of inferring causal relations through interventions are, for example, medical drug trials, where patients are randomly assigned to either taking the drug or a placebo, and the effect of the drug

is measured. Another example is testing whether the use of a fertilizer has a causal effect on the crop yield, for instance by intervening on the dose of a fertilizer. If such interventions are actively carried out and data are collected under such an intervention, one talks about *experimental data*. In this case, the desired causal effect can be directly inferred from the data.

However, such experiments cannot always be carried out. In Example 1.4, for instance, intervening on 'icy streets' would be very costly, intervening on 'falling down' unethical, and intervening on 'outside temperature' simply technically impossible. Other more realistic situations in which such interventions cannot be carried out are, for instance, in epidemiology, when evaluating the effect of a potentially dangerous substance (like lead in paint, PVC in pipes and flooring) on the health of people, or the effect of drinking alcohol or smoking during pregnancy on the development of the unborn.

In these cases, causal relationships have to be inferred from *passive observational* (i.e. *non-experimental*) data, which are merely observed without performing any interventions. One main concern when using passive observational data is bias in the causal effect due to *confounding*, that is due to variables that are related to both the cause and the effect. For instance, when inferring the effect of drinking alcohol on the unborn from passive observational data one has to take into account that women who drink alcohol during pregnancy may also be less aware of healthy nutrition. If not appropriately controlled for, the diet of a pregnant woman can introduce spurious correlation between the drinking of alcohol and the development of the unborn, since a poor diet may also have a negative effect on the unborn. Note that this kind of spurious correlation is removed when carrying out experiments: In this example, pregnant women would be randomly assigned to drink alcohol or not (which is of course ethically not justifiable), and hence both groups (the drinking and non-drinking one) would contain women from any background (healthy or unhealthy nutrition).

In this thesis we focus on learning causal relationships from passive observational data, following the seminal work by Spirtes et al. (1993) and Pearl (2000). One main reason for concentrating on such data is that many of the collected data sets are in fact non-experimental rather than experimental, since it is generally easier to collect passive observational data.

1.2 Research Questions

There are at least two core research questions in the field of causal discovery, both of which are partly addressed in this thesis.

Q1: How can one infer the effect of an intervention?

First, it is important to distinguish between predicting a (future) observation in a system that remains undisturbed, and predicting the effect of an intervention. The former is a purely statistical task, relying on common occurrences (i.e. correlations) of two variables. For instance, in Example 1.2, seeing people swimming outside helps in predicting whether the streets are icy *given that the data generating process is not altered.* In practice, prediction problems are often solved using classification or regression methods (see for example Hastie et al., 2009). In this thesis, however, we are concerned with the task of predicting the effect of an *intervention.* For instance, in the above example, we want to predict what would happen to the icy streets if we made sure that people are swimming outside. Although a (negative) correlation exists between these two variables, it is clear that the condition of the streets would not change under this intervention. Thus, for answering Q1 knowledge of correlations is not sufficient.

Question Q1 can be posed in several settings. First of all, what is known about the data generating process? In some cases, the graph of this process is given, for instance, by expert knowledge (i.e. we know which variables are involved in the process and how they are connected, but not the strength of the effects). In other situations, only certain parts of the graph are known, or the data generating process is completely unknown, and we only assume that the data are generated by such a graph.

Secondly, what kind of observations do we have? As already mentioned, the data set can be passive observational or experimental. A further aspect to take into account is whether all 'relevant' variables of the data generating process are observed, or if some variables are unobserved (i.e. no observations are available for these variables).

In the case of experimental data sets, if the intervention of which we want to predict the effect is carried out, it is possible to infer the effect directly from the data. However, if the required intervention was not performed, it is interesting to pose Q1 in the various settings above.

In this thesis, however, we will address research question Q1 in the case of passive observational data when only parts of the relevant variables are observed. Furthermore, the underlying graph is unknown, though some other background knowledge on the variables is available (such as a partial ordering of the variables).

Q2: How can one learn the structure of the underlying causal model?

In many cases, the underlying graph of the data generating process is not known, and the main interest lies in inferring the graph or certain

characteristics of it. This may allow answering Q1, but also gives a deeper insight into how certain dependencies are produced, and helps to understand the system in general.

As for Q1, we can distinguish between the type of data set at hand: Is it an experimental or passive observational data set? Are all relevant variables of the data generating process observed?

Furthermore, in some cases several data sets may be available. For instance, experiments may have been carried out under various interventions, each of which yields a separate data set. Alternatively, data sets (passive observational or experimental) may only share parts of the variables, resulting from different studies on related problems. The aim then is to combine the information of these data sets to learn a data generating process over the involved variables.

This thesis addresses research question Q2 in the setting of a single passive observational data set. In some of the presented work not all relevant variables of the data generating process need to be observed.

Which of the two research questions should be posed depends on the problem. In general, if the interest lies on inferring the effect of one specific intervention then the less general question Q1 is appropriate, since one should not solve a harder problem (Q2) than needed. However, if the main task is to better understand the causal connections among the involved variables and to learn features of the underlying causal system, Q2 is the appropriate question to pose.

1.3 Outline

In Chapters 2 to 5 we discuss the necessary background and existing work: Chapter 2 contains basic concepts and notations of graph theory and probability theory, which are required for the later chapters. The causal models considered in this thesis as well as related definitions and theorems are introduced in Chapter 3. Relevant existing methods towards answering research question Q1 are presented in Chapter 4, whereas the relevant existing work addressing research question Q2 is given in Chapter 5.

The contributions of this thesis to the research field are presented in Chapter 6. The results are based on the publications listed in the following section and reprinted at the end of the thesis. Finally, Chapter 7 concludes the thesis by summarizing the results and pointing out future research directions.

1.4 Publications and Authors' Contributions

The thesis is based on the following publications, referred to as Article I to Article VI. While the authors' contributions are listed here below each article, the content of the articles is discussed in Chapter 6.

I. Moneta, A., Entner, D., Hoyer, P. O., and Coad, A. (2013). Causal Inference by Independent Component Analysis: Theory and Applications. *Oxford Bulletin of Economics and Statistics*, Volume 75, Issue 5, pages 705-730.

   The present author implemented the algorithm and performed a large part of the calculations for the application sections, and assisted in writing the manuscript. Dr. Moneta drafted most of the article and performed parts of the data analysis. Dr. Hoyer and Dr. Coad helped with analyzing the results and with writing the manuscript.

II. Entner, D. and Hoyer, P. O. (2010). On Causal Discovery from Time Series Data using FCI. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, pages 121-128. HIIT Publications 2010-2.

   The idea was suggested by Dr. Hoyer, and the algorithm was jointly developed with the present author. The present author implemented the method, performed the data analysis, and wrote the section summarizing the results of these experiments, as well as assisted in writing the other parts of the article.

III. Entner, D. and Hoyer, P. O. (2012). Estimating a Causal Order among Groups of Variables in Linear Models. In *Artificial Neural Networks and Machine Learning - ICANN 2012*, LNCS 7553, pages 84-91, Springer Berlin Heidelberg.

   The idea arose from a discussion between M.Sc. Ali Bahramisharif and the present author. The present author suggested the general algorithm. Ideas for the trace method and the pairwise measure were discussed with Dr. Hoyer and Prof. Aapo Hyvärinen. The present author finalized the methods, performed all simulations, and wrote the paper. Dr. Hoyer commented on the draft at several stages and assisted in writing in the final stage.

IV. 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*, LNAI 6797, pages 181-195, Springer Berlin Heidelberg.

Dr. Hoyer proposed the basic idea, which, after further development jointly with the present author, led to the problem statement of the article. The present author developed the algorithm and proved the theorems and lemmas, assisted by Dr. Hoyer. The present author performed all the simulations and wrote the article. Dr. Hoyer provided valuable comments on the draft at several stages, and helped with editing in the later stages.

V. Entner, D., Hoyer, P. O., and Spirtes, P. (2012). Statistical Test for Consistent Estimation of Causal Effects in Linear Non-Gaussian Models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, Journal of Machine Learning Research Workshop and Conference Proceedings 22: 364-372.

The motivation of the underlying problem was given by Dr. Hoyer. The present author developed the algorithm, stated the theorems and lemmas, and proved them with the support of Dr. Hoyer. The present author performed all the simulations and drafted most of the article. Dr. Hoyer co-edited the manuscript, and Prof. Spirtes gave valuable comments at several stages.

VI. Entner, D., Hoyer, P. O., and Spirtes, P. (2013). Data-Driven Covariate Selection for Nonparametric Estimation of Causal Effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013).* Journal of Machine Learning Research Workshop and Conference Proceedings 31: 256-264.

The idea came up in a discussion between Dr. Hoyer and the present author, who then jointly developed the novel method, and proved its soundness and completeness. Prof. Spirtes suggested the comparison algorithm based on FCI. The present author implemented the methods, and performed all the simulations. Dr. Hoyer and the present author drafted the article, and obtained valuable comments from Prof. Spirtes at several stages.

Chapter 2

Background

We first introduce the necessary notation and terminology related to graphs used in the causal models of this thesis. Furthermore, we summarize some principles of probability theory and statistics, which are relevant to the theorems and methods stated in later chapters.

2.1 Graph Terminology

Here, we introduce terms and notation related to graphs, following Spirtes et al. (1993) and Pearl (2000).

A *directed graph* $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{v_1, \ldots, v_n\}$ being a set of vertices, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ a set of edges. A pair $(v_i, v_j) \in \mathcal{E}$ is also denoted as $v_i \to v_j$. We assume that there is at most one edge between any pair of vertices, and that there are no self loops, i.e. no edges from any vertex to itself.

A *path* π between v_1 and v_k is a sequence of edges $(d_1, \ldots d_{k-1})$, $d_j \in \mathcal{E}$, $j = 1, \ldots, k-1$, such that there exists a sequence of vertices $v_1, \ldots, v_k$ with edge d_j having endpoints v_j and v_{j+1}, i.e. $(v_j, v_{j+1}) \in \mathcal{E}$ or $(v_{j+1}, v_j) \in \mathcal{E}$. A *directed path* is a path π such that for all edges d_j, $j = 1, \ldots, k-1$, $(v_j, v_{j+1}) \in \mathcal{E}$, i.e. $v_1 \to \ldots \to v_j \to v_{j+1} \to \ldots \to v_k$. A *directed cycle* is a directed path starting and ending in the same vertex, i.e. $v_1 = v_k$. A directed graph not containing any directed cycles is called a *directed acyclic graph* (DAG).

If there is an edge $v_i \to v_j$, then v_i and v_j are called *adjacent*, v_i is the *parent* of v_j, and v_j the *child* of v_i. A node v_i is called a *root* or *source* if it has no parents, and a *sink* if it has no children. If there is a directed path from v_i to v_j, then v_i is called an *ancestor* of v_j, and v_j a *descendant* of v_i.

A *causal (topological) order* among the vertices $v_1, \ldots, v_n$ of a DAG $\mathcal{G}$

is a permutation $K = (K_1, \ldots, K_n)$ of the indices $1, \ldots, n$, such that for every $i > j$, v_{K_i} is not an ancestor of v_{K_j}, also denoted as $v_{K_j} \prec v_{K_i}$.

A triple (v_i, v_k, v_j) is called a *collider* if $(v_i, v_k) \in \mathcal{E}$ and $(v_j, v_k) \in \mathcal{E}$, i.e. $v_i \to v_k \leftarrow v_j$. A collider (v_i, v_k, v_j) is *unshielded* if there is no edge between v_i and v_j.

The *skeleton* of a DAG $\mathcal{G}$ is an undirected graph, i.e. its edges are of the form $v_i - v_j$, which is obtained by removing all arrowheads from the edges of $\mathcal{G}$. A *pattern* is obtained from a DAG by removing some of the arrowheads, meaning that it can contain two types of edges, directed ($v_i \to v_j$) and undirected ones ($v_k - v_l$), and cannot contain any directed cycles.

A *mixed graph* is a graph that can contain three kinds of edges: directed ($\to$), bidirected ($\leftrightarrow$), and undirected ($-$); between any pair of vertices, there can be more than one edge type. Directed paths and cycles, parents, children, ancestors and descendants are defined as in directed graphs. Additionally, if $v_i \leftrightarrow v_j$ in $\mathcal{G}$, then v_i is a *spouse* of v_j. If $v_i - v_j$ in $\mathcal{G}$, then v_i is a *neighbor* of v_j. An *almost directed cycle* occurs when there exist v_i and v_j, $i \neq j$, such that v_i is a spouse and an ancestor of v_j (Richardson and Spirtes, 2002; Zhang, 2008).

An *ancestral graph* is a mixed graph with no directed cycles, no almost directed cycles, and for any undirected edge $v_i - v_j$, v_i and v_j have no parents or spouses. This definition implies that ancestral graphs contain at most one edge between any pair of vertices. A *partial ancestral graph* (PAG) is obtained from an ancestral graph by changing some edge marks into circles '$\circ$', i.e. it may contain six kinds of edges: $-$, $\to$, $\leftrightarrow$, $\circ\!\!-$, $\circ\!\!-\!\!\circ$, and $\circ\!\!\to$ (Richardson and Spirtes, 2002; Zhang, 2008).

2.2 Probability Theory and Statistics

We give some basic definitions of probabilities, and introduce statistical concepts used in this thesis. For further details see for example Wasserman (2004), or the introductory chapters of Spirtes et al. (1993) and Pearl (2000).

2.2.1 Random Variables

Given a (multidimensional) random variable $\boldsymbol{v} = (v_1, \ldots, v_n)$, we denote the *joint probability distribution* as $p(\boldsymbol{v})$ or $p(v_1, \ldots, v_n)$. We will use lowercase p for probability distributions for both discrete and continuous random variables. In the former case p is a probability mass function, in the latter a probability density function.

Let $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ being two (possibly) multidimensional random variables. The *marginal probability distribution* of $\boldsymbol{v}_1$ is given by

$$p(\boldsymbol{v}_1) = \begin{cases} \int p(\boldsymbol{v}_1, \boldsymbol{v}_2)\, d\boldsymbol{v}_2 & \text{(for continuous variables)} \\ \sum_{\boldsymbol{v}_2} p(\boldsymbol{v}_1, \boldsymbol{v}_2) & \text{(for discrete variables).} \end{cases} \tag{2.1}$$

Given $p(\boldsymbol{v}_2) > 0$, the *conditional probability distribution* of $\boldsymbol{v}_1$ given $\boldsymbol{v}_2$ is defined as

$$p(\boldsymbol{v}_1 \,|\, \boldsymbol{v}_2) = \frac{p(\boldsymbol{v}_1, \boldsymbol{v}_2)}{p(\boldsymbol{v}_2)}. \tag{2.2}$$

The *chain rule* is a direct consequence of this definition, stating that the joint probability distribution can be factorized using conditional probability distributions as follows

$$p(v_1, \ldots, v_n) = \prod_{i=1}^{n} p(v_i \,|\, v_1, \ldots, v_{i-1}). \tag{2.3}$$

We use standard definitions of the expectation of $\boldsymbol{v}$ (denoted as $\mathrm{E}(\boldsymbol{v})$ or $\mu_{\boldsymbol{v}}$), the covariance matrix of $\boldsymbol{v}$ ($\boldsymbol{\Sigma}_{\boldsymbol{v}}$ or $\mathrm{cov}(\boldsymbol{v}, \boldsymbol{v})$, which reduces for scalar variables to the variance, σ_v^2), the matrix of (cross-)covariances of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ ($\mathrm{cov}(\boldsymbol{v}_1, \boldsymbol{v}_2)$), the matrix of correlations of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ ($\rho_{\boldsymbol{v}_1, \boldsymbol{v}_2}$), as well as the matrix of partial correlations of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ given $\boldsymbol{v}_3$ ($\rho_{\boldsymbol{v}_1, \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}$).

2.2.2 Statistical Independence

Two (multidimensional) random variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are said to be *statistically independent*, denoted as $\boldsymbol{v}_1 \perp\!\!\!\perp \boldsymbol{v}_2$ (Dawid, 1979), if and only if their joint probability distribution is equal to the product of their marginals, i.e.

$$\boldsymbol{v}_1 \perp\!\!\!\perp \boldsymbol{v}_2 \quad \Leftrightarrow \quad p(\boldsymbol{v}_1, \boldsymbol{v}_2) = p(\boldsymbol{v}_1)\, p(\boldsymbol{v}_2). \tag{2.4}$$

Conditional independence of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ given $\boldsymbol{v}_3$ is defined similarly:

$$\boldsymbol{v}_1 \perp\!\!\!\perp \boldsymbol{v}_2 \mid \boldsymbol{v}_3 \quad \Leftrightarrow \quad p(\boldsymbol{v}_1, \boldsymbol{v}_2 \,|\, \boldsymbol{v}_3) = p(\boldsymbol{v}_1 \,|\, \boldsymbol{v}_3)\, p(\boldsymbol{v}_2 \,|\, \boldsymbol{v}_3). \tag{2.5}$$

There exist a variety of statistical tests for independence, some of which are discussed below. The null hypothesis of such tests is that the variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are (conditionally) independent (given $\boldsymbol{v}_3$), i.e.

$$\mathcal{H}_0 : \boldsymbol{v}_1 \perp\!\!\!\perp \boldsymbol{v}_2 \qquad \text{or} \qquad \mathcal{H}_0 : \boldsymbol{v}_1 \perp\!\!\!\perp \boldsymbol{v}_2 \,|\, \boldsymbol{v}_3. \tag{2.6}$$

From the obtained p-value of such an independence test we can conclude, given a threshold α, whether the null hypothesis should be rejected.

There are two types of errors: The null hypothesis is true and is rejected (type 1 error), or the null hypothesis is wrong and is not rejected (type 2 error). The rate of type 1 errors can be directly controlled for by the threshold α. However, if this threshold is set too low in order to avoid type 1 errors, typically the number of type 2 errors becomes larger.

A central point in several of the methods discussed in this thesis is to, contrary to standard statistical principles, *accept* the null hypothesis if it is not rejected. This can be justified by using consistent tests, i.e. for growing sample size, and when appropriately decreasing the threshold α, both the type 1 and the type 2 error rates converge to zero, so that such methods are correct in the limit of large sample size. More precisely, these methods are *pointwise consistent*, meaning that for every $\varepsilon > 0$ and for every probability distribution p there exists a sample size $n_{\varepsilon,p}$ such that for every sample larger than $n_{\varepsilon,p}$ the probability of making a wrong inference is smaller than ε. However, they are *not* uniformly consistent, i.e. there exists no single sample size n_ε, which is independent of the probability distribution p, for which the above holds (Spirtes et al., 1993 (2nd edition, Ch.12.4); Robins et al., 2003).

For discrete variables *Pearson's* χ^2 *test* is often used to test independence between variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ given $\boldsymbol{v}_3$. In essence, it compares the number of observed counts for $p(\boldsymbol{v}_1, \boldsymbol{v}_2 \,|\, \boldsymbol{v}_3)$, and the number of expected counts under $\mathcal{H}_0$ (i.e. using that $p(\boldsymbol{v}_1, \boldsymbol{v}_2 \,|\, \boldsymbol{v}_3) = p(\boldsymbol{v}_1 \,|\, \boldsymbol{v}_3)\, p(\boldsymbol{v}_2 \,|\, \boldsymbol{v}_3)$) to develop a test statistic, which is χ^2-distributed under the null hypothesis.

For continuous variables, we distinguish between Gaussian (i.e. normal) and non-Gaussian (non-normal) ones. For normally distributed variables, independence is equivalent to *zero correlation.*[1] In this case, *Fisher's Z*, which follows a standard normal distribution under $\mathcal{H}_0$, can be used to test for zero (partial) correlation.

For non-Gaussian variables, we describe here only two ways of testing independence. A recently developed method, termed HSIC (Hilbert Schmidt Independence Criterion, Gretton et al., 2008), is a kernel-based test for marginal dependence. In the limit of large sample size this test will detect any form of statistical dependence. However, due to its computational complexity it can only be applied to relatively small sample sizes. Zhang et al. (2011) used a similar kernel-based approach to develop a test for conditional independence.

The second approach relies on the fact that two variables v_1 and v_2 are independent if and only if for all functions g and h it holds that

[1] Note that independence implies uncorrelatedness regardless of the form of the distribution, but the converse is only true for Gaussian distributions.

$\mathrm{E}(g(v_1)h(v_2)) = \mathrm{E}(g(v_1))\,\mathrm{E}(h(v_2))$, see for example Hyvärinen et al. (2001). Thus, we can test independence by testing for vanishing correlations between the transformed variables (for which there exist standard tests). The obvious drawback is that one can never test *all* functions g and h. However, a test based on a few carefully selected functions g and h, which detect various forms of dependence, is a computationally efficient alternative to the HSIC test.

Finally, the Darmois-Skitovitch Theorem (Darmois, 1953; Skitovitch, 1953) states an interesting property about dependence and independence of two sums of independent random variables.

Theorem 2.1 (Darmois-Skitovitch Theorem)**.** *Let $e_1, \ldots, e_n$ be independent random variables ($n \geq 2$), $v_1 = \beta_1 e_1 + \ldots + \beta_n e_n$ and $v_2 = \gamma_1 e_1 + \ldots + \gamma_n e_n$ with constants β_i, γ_i, $i = 1, \ldots, n$. If v_1 and v_2 are independent, then those e_j which influence both sums v_1 and v_2 (i.e. $\beta_j \gamma_j \neq 0$) are Gaussian.*

This theorem directly implies that if there exists a j such that $\beta_j \gamma_j \neq 0$ and e_j non-Gaussian, then the variables v_1 and v_2 are dependent.

2.2.3 Linear Regression

As we use linear regression models in several articles of this thesis, we briefly introduce the ordinary least squares (OLS) estimator and some of its properties. Let w and $\boldsymbol{v} = (v_1, \ldots, v_n)$ be random variables with zero mean. The linear regression model of w on $\boldsymbol{v}$ is given by

$$w = \sum_{i=1}^{n} b_i v_i + e \tag{2.7}$$

with b_i, $i = 1, \ldots, n$, constants and e a disturbance term. For the OLS estimator, the vector $\boldsymbol{c} = (c_1, \ldots, c_n)^T$ is chosen to minimize the sum squared error between w and its estimate $\hat{w} = \boldsymbol{c}^T \boldsymbol{v}$. The estimator has the closed form solution

$$\boldsymbol{c} = \mathrm{cov}(\boldsymbol{v}, \boldsymbol{v})^{-1} \mathrm{cov}(\boldsymbol{v}, w). \tag{2.8}$$

The resulting residuals $r = w - \hat{w}$ are by construction uncorrelated with the regressors $\boldsymbol{v}$, i.e. $\rho_{r,\boldsymbol{v}} = 0$.

If the covariance matrix of $\boldsymbol{v}$ is finite and non-singular, and e has zero mean and is uncorrelated with $\boldsymbol{v}$, the OLS estimator $\boldsymbol{c}$ is a consistent estimator of the regression coefficients $\boldsymbol{b} = (b_1, \ldots, b_n)^T$ (Verbeek, 2008).

Chapter 3

Causal Models

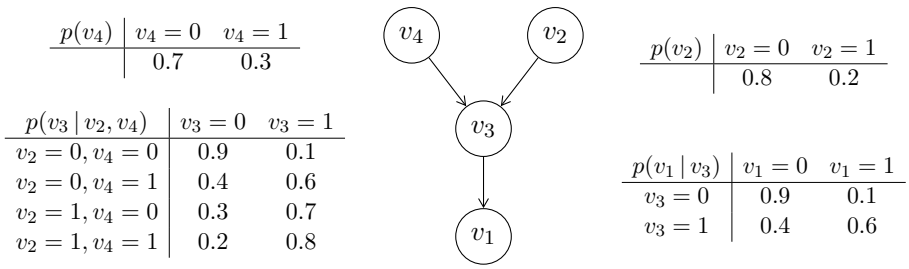
We formalize the notion of causality using models based on directed acyclic graphs in which edges represent causal relationships (Spirtes et al., 1993; Pearl, 2000), as demonstrated in the graph of Figure 1.1 (page 2). We first introduce models for non-temporal data, in particular causal Bayesian networks (CBNs) and structural equation models (SEMs), and some basic concepts and assumptions relating causality to DAGs. In the later part of this chapter, we generalize these models to time series data.

An alternative approach to causal modeling is the potential-outcome framework of Neyman (1923) and Rubin (1974). Since Pearl (2000) showed that this approach is equivalent to SEMs, we do not present the potential-outcome framework here. Details can be found, for instance, in the recent book of Berzuini et al. (2012).

3.1 Examples

We start with demonstrating CBNs and SEMs by examples; formal definitions are given in the next section. In a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the set $\mathcal{V}$ contains random variables $v_1, \ldots, v_n$, and there is an edge $v_i \to v_j$ in $\mathcal{E}$ if and only if v_i is a *direct* cause of v_j (with respect to the full set of variables $\mathcal{V}$).[1] These models can be seen as data generating processes, explaining how the real world works. In CBNs conditional probability distributions are directly linked to the variables, whereas in SEMs each variable is associated with a deterministic function and an unknown

[1] Originally, (non-causal) Bayesian networks were introduced to efficiently represent joint probability distributions, and to facilitate probabilistic reasoning (see for instance Pearl, 1988, or Koller and Friedman, 2009). In such models, the edges are not interpreted as causal relationships, but merely reflect statistical dependencies.



$p(v_4)$	$v_4 = 0$	$v_4 = 1$
	0.7	0.3

$p(v_3 \mid v_2, v_4)$	$v_3 = 0$	$v_3 = 1$
$v_2 = 0, v_4 = 0$	0.9	0.1
$v_2 = 0, v_4 = 1$	0.4	0.6
$v_2 = 1, v_4 = 0$	0.3	0.7
$v_2 = 1, v_4 = 1$	0.2	0.8

$p(v_2)$	$v_2 = 0$	$v_2 = 1$
	0.8	0.2

$p(v_1 \mid v_3)$	$v_1 = 0$	$v_1 = 1$
$v_3 = 0$	0.9	0.1
$v_3 = 1$	0.4	0.6

Figure 3.1: Example of a causal Bayesian network (CBN). Each variable in the underlying DAG is associated with a conditional probability table. The variables could for example be v_1 = 'breaking wrist', v_2 = 'drinking beer', v_3 = 'falling down', and v_4 = 'icy streets'.

error term. In this way, data can be (stochastically) generated along a *causal order* K among the variables. The acyclicity assumption implied by the DAG ensures that at least one such order always exists.

The power of CBNs and SEMs lies in their ability to predict the effects of interventions. As discussed in the introduction, an intervention occurs when a variable is forced to take on a specific value, meaning that a causal system is actively disturbed by setting a variable to some constant value.

Example 3.1 (Causal Bayesian Network)**.** *Figure 3.1 shows an example of a CBN over four binary variables. To each variable v_i a (conditional) probability table is attached, giving the probability distribution $p(v_i \mid pa_i)$, $i = 1, \dots, 4$, with pa_i the parents of v_i in the DAG.*

There are two causal orders compatible with this CBN, $K = (2, 4, 3, 1)$, denoted by $v_2 \prec v_4 \prec v_3 \prec v_1$, and $K = (4, 2, 3, 1)$, i.e. $v_4 \prec v_2 \prec v_3 \prec v_1$.

The data are generated along either of these two causal orders, for instance for $K = (2, 4, 3, 1)$ we

1. *draw v_2 using the probability table $p(v_2)$,*
2. *draw v_4 using the probability table $p(v_4)$,*
3. *draw v_3 using the conditional probability table $p(v_3 \mid v_2, v_4)$, and*
4. *draw v_1 using the conditional probability table $p(v_1 \mid v_3)$.*

The causal order ensures that the values of the conditioning variables have been assigned in a previous step of the data generating process. The joint probability distribution factorizes according to the underlying DAG as

$$p(v_1, v_2, v_3, v_4) = p(v_2)\, p(v_4)\, p(v_3 \mid v_2, v_4)\, p(v_1 \mid v_3).$$

If we intervene, for instance, on v_3 by setting its value to 1 (instead of observing v_3 taking the value 1) we replace the conditional probability table

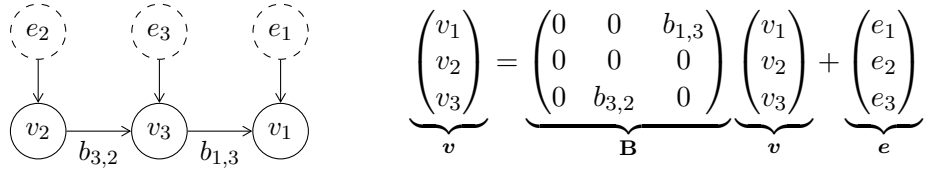


Figure 3.2: Example of a linear structural equation model (linear SEM). The variables of the underlying DAG are linked to linear equations, given here in matrix notation. The connection matrix $\mathbf{B}$ contains non-zero entries representing the edges of the DAG. The disturbances e_i, following distributions $p(e_i)$, $i = 1, 2, 3$, are unobserved and mutually independent.

$p(v_3 \,|\, v_2, v_4)$ *with* $p(v_3 = 1) = 1$. *This affects the data generating process in step 3, and the joint probability distribution under the intervention* $v_3 = 1$, *termed the* postinterventional probability distribution*, is given by*

$$p(v_1, v_2, v_4 \,|\, do(v_3 = 1)) = p(v_2)\, p(v_4)\, p(v_1 \,|\, v_3 = 1),$$

with $do(v_3 = 1)$ *indicating the intervention on* v_3. *In the underlying DAG this translates to deleting the edges from* v_4 *and* v_2 *to* v_3, *since under the intervention the former two variables are no longer causes of* v_3.

Example 3.2 (Structural Equation Model)**.** *Figure 3.2 shows an example of a* linear SEM*, in which each variable is associated with an equation defining its value as a linear combination of its parents and an unobserved disturbance term. These disturbances are assumed to be mutually independent.*

For this DAG, there is only one compatible causal order, namely $K = (2, 3, 1)$, *i.e.* $v_2 \prec v_3 \prec v_1$. *The data are generated along this causal order:*

1. *draw* e_2 *from its corresponding distribution and set* $v_2 = e_2$,
2. *draw* e_3 *from its corresponding distribution and set* $v_3 = b_{3,2}\, v_2 + e_3$,
3. *draw* e_1 *from its corresponding distribution and set* $v_1 = b_{1,3}\, v_3 + e_1$.

Similar to CBNs, the causal order ensures that the values of the variables occurring in the right hand side of the equations are determined in a previous step of the data generating process.

The probability distribution of each variable given its parents $p(v_i \,|\, pa_i)$ *is determined by the distributions of the disturbances. For example, if for* $i = 1, 2, 3$, $e_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ *(a Gaussian distribution with mean* μ_i *and variance* σ_i^2*), then* $p(v_i \,|\, pa_i)$ *also follows a Gaussian distribution:*

$$p(v_1 \mid v_3) \sim \mathcal{N}(\mu_1 + b_{1,3}\, v_3,\ \sigma_1^2),$$
$$p(v_2) \sim \mathcal{N}(\mu_2, \sigma_2^2),$$
$$p(v_3 \mid v_2) \sim \mathcal{N}(\mu_3 + b_{3,2}\, v_2,\ \sigma_3^2).$$

When intervening, for instance, on v_3 by setting its value to a constant c_3, the equation of $v_3 = b_{3,2}\, v_2 + e_3$ is replaced with $v_3 = c_3$. The implications on the joint probability under the intervention as well as on the underlying DAG of the SEM are as explained for CBNs.

3.2 Formal Definitions of CBNs and SEMs

Following Pearl (2000), we here give the formal definitions of CBNs and SEMs, based on the concept of interventions. As shown in Examples 3.1 and 3.2, changing one conditional probability distribution or structural equation by intervention does not affect the other distributions or equations. Formally, an *atomic intervention* arises when a variable v_i is *set* to some specific constant value c_i without affecting any other causal mechanism.

Definition 3.1 (Causal Bayesian Network)**.** *A* causal Bayesian network *consists of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a probability distribution over $\boldsymbol{v} = (v_1, \ldots, v_n)$ factorizing according to $\mathcal{G}$ as in*

$$p(v_1, \ldots, v_n) = \prod_{i=1}^{n} p(v_i \mid pa_i), \tag{3.1}$$

with pa_i the parents of v_i in $\mathcal{G}$, and postinterventional probability distributions *resulting from intervening on a set $\mathcal{V}_k \subset \mathcal{V}$ setting $\boldsymbol{v}_k = \boldsymbol{c}_k$ defined by the* truncated factorization formula

$$p(\mathcal{V} \setminus \mathcal{V}_k \mid do(\boldsymbol{v}_k = \boldsymbol{c}_k)) = \prod_{i:\, v_i \notin \mathcal{V}_k} p(v_i \mid pa_i). \tag{3.2}$$

The second way of defining causal models is via SEMs, which were first introduced in the fields of genetics (Wright, 1921), and econometrics (Haavelmo, 1943), and are further discussed for example by Bollen (1989). Over the years the causal language embodied by SEMs has been partly forgotten, and was revitalized by Pearl (2000).

Definition 3.2 (Structural Equation Model)**.** *A (recursive)* structural equation model *consists of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a set of probability distributions $p(e_i)$, $i = 1, \ldots, n$, and a set of equations*

$$v_i = f_i(pa_i, e_i),\ i = 1, \ldots, n, \tag{3.3}$$

where f_i is a function mapping the parents pa_i of v_i in $\mathcal{G}$ and an unobserved disturbance term e_i to v_i. The disturbance terms e_i are assumed to be mutually independent, i.e. $p(e_1, \ldots, e_n) = \prod_{i=1}^{n} p(e_i)$. Under an intervention $v_k = c_k$, the structural equation $v_k = f_k(pa_k, e_k)$ is replaced with $v_k = c_k$.

If all functions f_i in a SEM are linear, as in Example 3.2, we refer to the model as a *linear SEM*.[2] Typically, in these models the disturbances e_i, $i = 1, \ldots, n$, are assumed to have zero mean, i.e. $\mathrm{E}(e_i) = 0$.

CBNs are most often used with discrete random variables, as they give a compact way to represent conditional probability distributions, whereas SEMs are commonly used with continuous random variables. As Example 3.2 shows, SEMs imply a probability distribution over $\boldsymbol{v}$, which is uniquely determined by the distributions of the disturbance terms e_i, $i = 1, \ldots, n$, so that SEMs can be transformed to CBNs. Furthermore, for every CBN there exists at least one SEM that generates the same joint probability distribution over the involved variables as the CBN, as well as the same postinterventional distributions (Druzdzel and Simon, 1993; Pearl, 2000).

Thus, in some way SEMs and CBNs are just two alternative ways to represent the causal relationships among a set of variables, and both can model interventions equally naturally, as the examples and definitions show. Note however that SEMs are inherently more powerful than CBNs when it comes to *counterfactual* reasoning (Pearl, 2000, 2nd edition Ch. 1.4.4, Ch. 7). We do not however consider counterfactuals further in this thesis.

3.3 Causal Markov Condition

The data generating process of a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is characterized by local probability distributions $p(v_i \,|\, pa_i)$, so that the value of a variable v_i is determined by the values of its *direct* causes pa_i. Once these are known, the values of the indirect causes and other variables prior to v_i in the causal order are irrelevant. For instance, in Example 3.1, once we know a person fell down (v_3), the conditions of the street (v_4) or whether the person has drunk beer (v_2) contain no further information on the person breaking the wrist (v_1). This is stated formally in the *causal Markov condition* (Spirtes et al., 1993; Pearl, 2000).

Definition 3.3 (Causal Markov Condition)**.** *In the probability distribution generated by a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each variable $v_i \in \mathcal{V}$ is*

[2] While we use the terms 'linear SEM' and 'SEM' to distinguish between models with linear functions f_i and arbitrary functions f_i, the terms 'SEM' and 'non-parametric SEM (NPSEM)', respectively, are sometimes used instead.

independent of all its non-effects (non-descendants) given its direct causes pa_i (parents), for all $i = 1, \ldots, n$.[3]

While Definitions 3.1 and 3.2 imply the causal Markov condition, this condition together with the chain rule for probabilities of Equation (2.3) (page 11), yields that the joint probability distribution $p(v_1, \ldots, v_n)$ over the variables in $\mathcal{V}$ factorizes according to the DAG $\mathcal{G}$, as in Equation (3.1). Furthermore, Spirtes et al. (1993) assumed the causal Markov condition and proved the so called *manipulation theorem*, a generalization of the truncated factorization formula of Equation (3.2).

3.4 Causal Sufficiency and Selection Bias

So far, the discussion focused on the data generating process, not on the data itself. If only part of the variables of a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ are observed, the set $\mathcal{V}$ is divided into two disjoint sets, $\mathcal{W}$ containing the *observed* variables, and $\mathcal{U}$ containing the *latent* (i.e. *hidden*, *unobserved*) variables. The causal Markov condition is only assumed to hold for the set $\mathcal{V}$, i.e. when disregarding, or not observing, some variables, there can be additional dependencies, also termed *spurious correlations*, among the observed variables, see Example 1.2 (page 2) and Example 3.3 below.

The troublesome variables introducing such dependencies are so called *confounders*, which are variables not included in $\mathcal{W}$ but having a (direct or indirect) causal effect on two or more of the observed variables in $\mathcal{W}$, i.e. unobserved common causes of variables in $\mathcal{W}$.[4] Towards this end the following assumption is often made (Spirtes et al., 1993; Pearl, 2000).

Definition 3.4 (Causal Sufficiency)**.** *A set $\mathcal{W}$ of observed variables is* causally sufficient *if and only if every common cause of two or more variables in $\mathcal{W}$ is contained in $\mathcal{W}$. In this case, we also call the CBN or SEM over the DAG $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ causally sufficient.*

Example 3.3 (Confounder, Causal Sufficiency)**.** *In the generating DAG of Example 1.2, redrawn in Figure 3.3 (a) with v_1 = 'outside temperature', v_2 = 'swimming outside', and v_3 = 'icy streets', the causal Markov condition holds if all three variables are considered: people swimming (v_2) is independent of the streets being icy (v_3) given the outside temperature (v_1).*

[3] While the *causal* Markov condition is stated in terms of non-effects and direct causes, in non-causal Bayesian networks a similar, purely statistical condition, the *local* Markov condition, is stated in terms of non-descendants and parents in the underlying graph.

[4] A common cause of v_i and v_j is formally defined as a variable having a causal effect on v_i that is not via v_j, and a causal effect on v_j that is not via v_i (Spirtes et al., 1993).

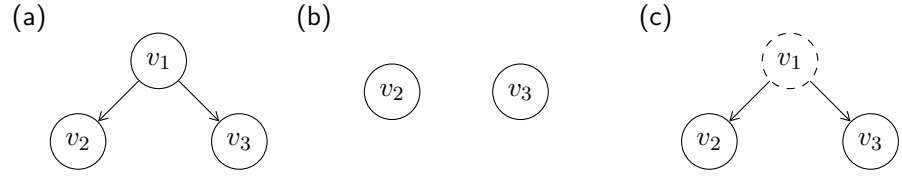


Figure 3.3: An example demonstrating causal sufficiency and the causal Markov condition. In (a) the set $\{v_1, v_2, v_3\}$ is causally sufficient and the causal Markov condition holds. When omitting v_1 from (a), the set $\{v_2, v_3\}$ shown in (b) is causally not sufficient and the causal Markov condition does not hold. In (c), the omitted variable v_1 is represented by a dashed circle.

On the contrary, setting $\mathcal{W} = \{v_2, v_3\}$ *and* $\mathcal{U} = \{v_1\}$*, and considering the graph over* $\mathcal{W}$ *only, as in Figure 3.3 (b), although there is no causal link between the two variables* v_2 *and* v_3*, they are negatively correlated. This* spurious *correlation is due to the* unobserved confounder v_1*.*

To represent unobserved confounders (and other unobserved variables) explicitly in a DAG $\mathcal{G}$ underlying a CBN or SEM, we will indicate observed variables by solid circles, and latent variables by dashed circles, as shown in Figure 3.3 (c) for Example 3.3.

Another way of introducing spurious correlation among two independent variables is *selection bias*. This rather is a property of the sampling method or design of a study than of the data generating model. Selection bias occurs when inclusion of a data point in the sample is affected by a variable which is causally related to some variable $v \in \mathcal{V}$. To put it differently, the value of a variable influences whether the data point is included in the data set or not. Selection bias can typically be avoided by appropriately collecting the data. For the rest of this thesis we assume that there is no selection bias.

3.5 Interventions and Causal Effects

In the introduced models, each variable of the associated DAG is linked to a (local) conditional probability distribution (in CBNs) or a structural equation (in SEMs), each representing an autonomous mechanism determining how the value of the corresponding variable is generated. Intervening on variable v_i only affects the corresponding conditional probability distribution or structural equation, as stated in the respective definitions. In the underlying DAG, this intervention simply means removing all edges with arrows into v_i (see Example 3.4 below). The postinterventional distribution

of y conditional on x, obtained from the truncated factorization formula of Equation (3.2), is also termed the causal effect of x on y (Pearl, 2000).[5]

Definition 3.5 (Causal Effect)**.** *Given a CBN or SEM, the* causal effect *of x on y, denoted as $p(y \mid do(x))$, is a function from x to the space of probability distributions on y, and is defined as the probability of y when intervening on x.*[6]

This definition of the causal effect marks the *total* effect of x on y, combining the direct effect (along the edge $x \to y$) as well as all indirect effects of x on y (along all directed paths from x to y other than $x \to y$). The definition of the *direct* effect requires that all paths between x and y other than the edge $x \to y$ are intervened on, which can in general be achieved by intervening on all variables other than y, or, if the DAG is known, by intervening on all parents of y, in addition to x (Pearl, 2000).

In linear SEMs, as in Example 3.2, the causal effect of x on y is typically not defined using the full postinterventional distribution. Rather, the (total) causal effect of x on y is defined as the rate of change in the expected value of y when intervening on x (Pearl, 2000), i.e.

$$\frac{\partial}{\partial x} \mathrm{E}(y \mid do(x)). \tag{3.4}$$

Causal effects in linear SEMs can also be read off the SEM directly, using the method of path coefficients (Wright, 1921, 1934), as demonstrated in the following example.

Example 3.4 (Interventions, Causal Effects)**.** *In the linear SEM of Figure 3.4 (a), intervening on the variable v_2 yields the model of Figure 3.4 (b), where in the DAG the intervened variable is marked with a double circle, and the updated linear equations are given below the DAG.*

The joint probability distribution over v_1, v_2, and v_3 in (a) and (b) are given by the factorizations of Equations (3.1) *and* (3.2)*, respectively:*

$$p(v_1, v_2, v_3) = p(v_1)p(v_2 \mid v_1)p(v_3 \mid v_1, v_2) \tag{3.5}$$

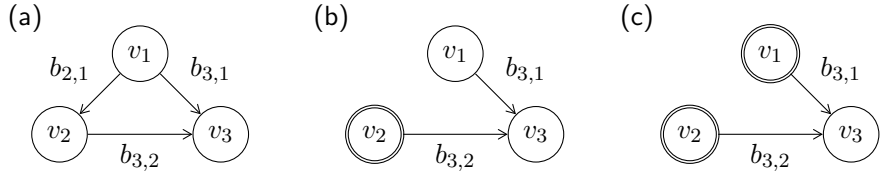
$$p(v_1, v_3 \mid do(v_2)) = p(v_1)p(v_3 \mid v_1, v_2). \tag{3.6}$$

Note that the postinterventional distribution is in general not *equal to the corresponding conditional distribution. Rewriting Equation* (3.6) *yields*

$$p(v_1, v_3 \mid do(v_2)) = \frac{p(v_1, v_2, v_3)}{p(v_2 \mid v_1)}, \tag{3.7}$$

[5] We will use the more convenient notation of x and y instead of v_i and v_j when talking about causes and effects.

[6] Note that for every possible assignment x_i of x, the causal effect gives a probability distribution over y, i.e. for each possible assignment y_j of y a value $p(y = y_j \mid do(x = x_i))$.



$v_1 = e_1$
$v_2 = b_{2,1}\, v_1 + e_2$
$v_3 = b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3$

$v_1 = e_1$
$v_2 = c_2$
$v_3 = b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3$

$v_1 = c_1$
$v_2 = c_2$
$v_3 = b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3$

Figure 3.4: An example demonstrating interventions. In (a), the DAG and linear equations of the data generating linear SEM are shown. The DAG and updated structural equations under intervention of v_2 are shown in (b), and under intervention of v_1 and v_2 in (c).

whereas the conditional probability distribution of v_1 and v_3 given v_2 is by Equation (2.2) (page 11) defined as

$$p(v_1, v_3 \,|\, v_2) = \frac{p(v_1, v_2, v_3)}{p(v_2)}. \tag{3.8}$$

Using Equation (3.4), the (total) causal effect of v_2 on v_3 is calculated as

$$\frac{\partial}{\partial v_2}\, E(v_3 \,|\, do(v_2)) = \frac{\partial}{\partial v_2}\, E(b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3) = \frac{\partial}{\partial v_2}\, b_{3,2}\, v_2 = b_{3,2}$$

since $E(b_{3,1}\, v_1) = b_{3,1}\, E(e_1) = 0$, $E(e_3) = 0$, and $E(b_{3,2}\, v_2) = b_{3,2}\, v_2$ due to the intervention of v_2. Note that in this case the direct and total causal effects are equal.

Direct *causal effects in linear SEMs are given by the corresponding coefficients in the structural equations. When intervening on v_1 and v_2, as shown in Figure 3.4 (c), we can calculate the direct causal effect of v_1 on v_3 as $\frac{\partial}{\partial v_1}\, E(b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3) = b_{3,1}$.*

In general, Wright (1921, 1934) stated that the total *causal effect of v_i on v_j in a linear SEM is the sum of the products of the coefficients along the various paths from v_i to v_j. For instance, the total causal effect of v_1 on v_3 is given by $b_{3,1} + b_{3,2}\, b_{2,1}$, i.e. by the direct effect ($b_{3,1}$) plus the indirect effect along the path via v_2 ($b_{3,2}\, b_{2,1}$). Note that while in linear SEMs indirect effects have this straightforward interpretation as products of coefficients along indirect paths, there is no such interpretation in general for SEMs and CBNs.*

3.6 DAGs and Independencies

The following theorems formally state how the underlying DAG of a CBN or SEM and the associated probability distribution are connected, as already illustrated in the examples using the causal Markov condition. Towards this end we define d-separation (Pearl, 1988), a concept which allows connecting independencies in a distribution p to a DAG $\mathcal{G}$.[7]

Definition 3.6 (d-Separation)**.** *Given a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a path π between v_i and v_j is said to be* blocked *by a set $\mathcal{Z} \subseteq \mathcal{V} \setminus \{v_i, v_j\}$ if and only if*

- *π contains a* chain $v_l \to v_k \to v_m$ *or a* fork $v_l \leftarrow v_k \to v_m$ *with $v_k \in \mathcal{Z}$, or*
- *π contains a* collider $v_l \to v_k \leftarrow v_m$ *with neither v_k nor any descendants of v_k in $\mathcal{Z}$.*

If a path π is not blocked, it is called active *or* open.

A set $\mathcal{Z}$ is said to d-separate *two vertices v_i and v_j, $i \neq j$, if and only if all paths between v_i and v_j are blocked by $\mathcal{Z}$, denoted as $v_i \perp\!\!\!\perp_{\mathcal{G}} v_j \,|\, \mathcal{Z}$. If $\mathcal{Z} = \emptyset$, we simply write $v_i \perp\!\!\!\perp_{\mathcal{G}} v_j$. If $\mathcal{Z}$ does not d-separate v_i and v_j, then v_i and v_j are called* d-connected *given $\mathcal{Z}$, denoted as $v_i \not\perp\!\!\!\perp_{\mathcal{G}} v_j \,|\, \mathcal{Z}$.*

A set $\mathcal{Z}$ is said to d-separate *two disjoint sets $\mathcal{V}_i$ and $\mathcal{V}_j$ if and only if all pairs $(v_i, v_j) \in \mathcal{V}_i \times \mathcal{V}_j$ are d-separated by $\mathcal{Z}$, denoted as $\mathcal{V}_i \perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z}$. If $\mathcal{Z}$ does not d-separate $\mathcal{V}_i$ and $\mathcal{V}_j$, then $\mathcal{V}_i$ and $\mathcal{V}_j$ are called* d-connected *given $\mathcal{Z}$, denoted as $\mathcal{V}_i \not\perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z}$.*

Example 3.5 (d-Separation)**.** *In the DAG of Figure 3.1 (Example 3.1) the following d-separation relationships (among others) hold*

$$v_4 \perp\!\!\!\perp_{\mathcal{G}} v_2 \qquad \textit{(by the second point of Definition 3.6)}$$
$$\{v_4, v_2\} \perp\!\!\!\perp_{\mathcal{G}} v_1 \,|\{v_3\} \qquad \textit{(by the first point of Definition 3.6).}$$

These are also reflected in the causal Markov condition. For example, the second d-separation relation means that 'icy streets' and 'drinking alcohol' are d-separated from 'breaking wrist', given 'falling down', as stated in Section 3.3 in terms of independencies. We also have, for instance, $v_4 \not\perp\!\!\!\perp_{\mathcal{G}} v_2 \,|\{v_3\}$, $v_4 \not\perp\!\!\!\perp_{\mathcal{G}} v_2 \,|\{v_1\}$, and $\{v_4, v_2\} \not\perp\!\!\!\perp_{\mathcal{G}} v_1$.

Denoting d-separation relationships with the symbol $\perp\!\!\!\perp_{\mathcal{G}}$ follows the notation of conditional (statistical) independence using $\perp\!\!\!\perp_p$ (Dawid, 1979), since these two concepts are closely related, as the following theorem shows (Verma and Pearl, 1988; Geiger and Pearl, 1988; Geiger et al., 1990).

[7] An equivalent formulation of d-separation, termed *moralization*, has been introduced by Lauritzen et al. (1990).

Theorem 3.1. *Given a SEM or CBN with underlying DAG $\mathcal{G}$, for any disjoint sets of random variables $\mathcal{V}_i, \mathcal{V}_j$, and $\mathcal{Z}$ hold*

(i) $\mathcal{V}_i \perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Rightarrow\; \mathcal{V}_i \perp\!\!\!\perp_p \mathcal{V}_j \,|\, \mathcal{Z}$ *in every probability distribution p which factorizes according to $\mathcal{G}$ (Global Markov Condition).*

(ii) $\mathcal{V}_i \not\perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Rightarrow\; \mathcal{V}_i \not\perp\!\!\!\perp_p \mathcal{V}_j \,|\, \mathcal{Z}$ *in at least one probability distribution p which factorizes according to $\mathcal{G}$.*

Statement (i) of Theorem 3.1 is referred to as the *global Markov condition* and, being a purely statistical property, is equivalent to the local Markov condition in DAGs (Lauritzen et al., 1990). For linear models, Spirtes et al. (1998) showed a result similar to Theorem 3.1 for partial correlations.

Theorem 3.2. *Given a linear SEM over a DAG $\mathcal{G}$, for any disjoint sets of random variables $\mathcal{V}_i, \mathcal{V}_j$, and $\mathcal{Z}$ hold*

(i) $\mathcal{V}_i \perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Rightarrow\; \rho_{\mathcal{V}_i,\mathcal{V}_j \cdot \mathcal{Z}} = 0$ *for every parameterization of the SEM over $\mathcal{G}$.*

(ii) $\mathcal{V}_i \not\perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Rightarrow\; \rho_{\mathcal{V}_i,\mathcal{V}_j \cdot \mathcal{Z}} \neq 0$ *for at least one parameterization of the SEM over $\mathcal{G}$.*

3.7 Faithfulness and Linear Faithfulness

For statement (ii) of Theorem 3.1, one can in fact show a stronger version saying that the implication holds for *almost all* probability distributions factorizing according to $\mathcal{G}$ (Spirtes et al., 1993). Those distributions which entail *additional* independencies to the ones entailed by the causal Markov condition (i.e. those distributions for which point (ii) of Theorem 3.1 does not hold) are said to be *unfaithful* to the DAG $\mathcal{G}$ (Pearl, 1988; Spirtes et al., 1993; Pearl, 2000). (In Pearl (2000) such distributions are termed 'unstable' with respect to the graph.) To get a one-to-one relationship between the d-separation relations of a DAG and the independencies of a probability distribution generated by a CBN or SEM over this DAG, the following assumption is needed.

Definition 3.7 (Faithfulness)**.** *Given a CBN or SEM over a DAG $\mathcal{G}$ with probability distribution p, p is said to be* faithful *to $\mathcal{G}$ if and only if every conditional independence in p is entailed by the causal Markov condition, i.e. is due to the structure of $\mathcal{G}$.*

The following example demonstrates faithfulness, and gives a basic intuition for the fact that almost all distributions are faithful to a DAG.

Example 3.6 (Faithfulness)**.** *Consider the graph in Figure 3.4 (a). Clearly, in the graph we have* $v_1 \not\perp\!\!\!\perp_{\mathcal{G}} v_3$*. However, if (and only if) the parameters* $b_{2,1}$*,* $b_{3,2}$*, and* $b_{3,1}$ *happen to be such that* $b_{3,1} + b_{3,2}\, b_{2,1} = 0$*, we obtain that* $v_1 \perp\!\!\!\perp_p v_3$ *in the distribution* p *due to the canceling paths between the two variables. Rewriting* v_3 *solely in terms of the disturbances* e_i*,* $i = 1, 2, 3$*, as*

$$\begin{aligned} v_3 &= b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3 = b_{3,1}\, e_1 + b_{3,2}\, (b_{2,1}\, v_1 + e_2) + e_3 \\ &= (b_{3,1} + b_{3,2}\, b_{2,1})\, e_1 + b_{3,2}\, e_2 + e_3 \end{aligned}$$

shows that if $b_{3,1} + b_{3,2}\, b_{2,1} = 0$*, the disturbance term* e_1 *has a zero effect on* v_3*, and hence* v_3 *is independent of* e_1*, and thus of* v_1*.*

However, the set where this constraint applies to the parameters is of (Lebesgue) measure zero among all possible parameter values, so for almost all parametrization $b_{2,1}$*,* $b_{3,2}$*, and* $b_{3,1}$ *the resulting distribution* p *will entail a dependence between* v_1 *and* v_3*.*

By adding the faithfulness assumption to Theorem 3.1, we obtain that d-separation relationships in a DAG $\mathcal{G}$ and independencies in a distribution p factorizing according to $\mathcal{G}$ are equivalent (Pearl, 2000), i.e. for all disjoint sets $\mathcal{V}_i$, $\mathcal{V}_j$, and $\mathcal{Z}$ holds

$$\mathcal{V}_i \perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Leftrightarrow\; \mathcal{V}_i \perp\!\!\!\perp_p \mathcal{V}_j \,|\, \mathcal{Z}. \tag{3.9}$$

This equivalence allows inferring d-separation relations (i.e. information about the underlying DAG) from testable dependencies and independencies in the observed data, and is used in some structure learning methods discussed in Chapter 5.

Similarly to the faithfulness assumption, we define linear faithfulness (Spirtes et al., 1993, 2nd edition p.47).

Definition 3.8 (Linear Faithfulness)**.** *Given a CBN or SEM over a DAG* $\mathcal{G}$ *with probability distribution* p*,* p *is said to be* linearly faithful *to* $\mathcal{G}$ *if and only if every zero partial correlation in* p *is entailed by the causal Markov condition.*

In linear SEMs with Gaussian disturbance terms, faithfulness and linear faithfulness are equivalent; for non-Gaussian disturbance terms, linear faithfulness is a stronger assumption. In non-linear models, neither implies the other (Robins, 1999). Adding the linear faithfulness assumption to Theorem 3.2, we obtain the equivalence of Equation (3.9) for linear SEMs, such that for all disjoint sets $\mathcal{V}_i$, $\mathcal{V}_j$, and $\mathcal{Z}$ holds

$$\mathcal{V}_i \perp\!\!\!\perp_{\mathcal{G}} \mathcal{V}_j \,|\, \mathcal{Z} \;\Leftrightarrow\; \rho_{\mathcal{V}_i, \mathcal{V}_j \cdot \mathcal{Z}} = 0. \tag{3.10}$$

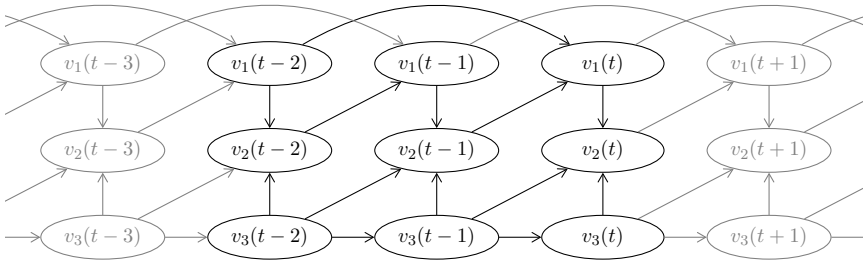


Figure 3.5: An example of a data generating structure of a time series.

3.8 Time Series Models

In some situations, variables evolve over time, forming a multivariate time series. We here only consider *discrete time* processes (as opposed to continuous time processes) where observations are obtained at regular time intervals, and each such observation is assumed to be generated by a combination of past and possibly present variables. Effects from variables in the past are called *lagged*, and from the present *instantaneous* effects. These instantaneous effects are assumed to follow a CBN or SEM.

Example 3.7 (Time Series)**.** *In Figure 3.5 we show the data generating process of a time series with 3 variables. The instantaneous effects are acyclic, and the model includes lagged effects one and two time steps back.*

If we assume linear relationships among the variables (similar to a linear SEM), the corresponding equations are defined as

$$\underbrace{\begin{pmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \end{pmatrix}}_{\boldsymbol{v}(t)} = \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ b^{(0)}_{2,1} & 0 & b^{(0)}_{2,3} \\ 0 & 0 & 0 \end{pmatrix}}_{\mathbf{B}_0} \underbrace{\begin{pmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \end{pmatrix}}_{\boldsymbol{v}(t)} + \underbrace{\begin{pmatrix} 0 & b^{(1)}_{1,2} & 0 \\ 0 & 0 & b^{(1)}_{2,3} \\ 0 & 0 & b^{(1)}_{3,3} \end{pmatrix}}_{\mathbf{B}_1} \underbrace{\begin{pmatrix} v_1(t-1) \\ v_2(t-1) \\ v_3(t-1) \end{pmatrix}}_{\boldsymbol{v}(t-1)}$$
$$+ \underbrace{\begin{pmatrix} b^{(2)}_{1,1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{\mathbf{B}_2} \underbrace{\begin{pmatrix} v_1(t-2) \\ v_2(t-2) \\ v_3(t-2) \end{pmatrix}}_{\boldsymbol{v}(t-2)} + \underbrace{\begin{pmatrix} e_1(t) \\ e_2(t) \\ e_3(t) \end{pmatrix}}_{\boldsymbol{e}(t)}, \tag{3.11}$$

with $e(t)$ unobserved disturbances. As we will shortly introduce, this is termed a 2^{nd}-order structural vector autoregressive model.

Chu and Glymour (2008) called the kind of structure of Figure 3.5 a *repetitive causal graph*. It is also closely related to unrolled dynamic

Bayesian networks (Koller and Friedman, 2009), which typically use only one time lag, and in addition contain a Bayesian network for the initial state. Furthermore, this structure is related to time series chain graphs (Dahlhaus and Eichler, 2003), which represent instantaneous dependencies as undirected edges. As in the examples of Section 3.1, the data generating process can be realized either by attaching a conditional probability distribution to each variable (as done in CBNs), or by using deterministic functional relationships and unobserved disturbance terms (as in SEMs). Most of the work on time series in this thesis uses the latter representation, in particular *Structural Vector Autoregressive Models* (SVAR), see for example Hamilton (1994), or Lütkepohl (2005).

Definition 3.9 (Structural Vector Autoregressive Model)**.** *A q^{th}-order structural vector autoregressive model (SVAR), with $q < \infty$, over a multivariate time series $\boldsymbol{v}(t) = (v_1(t), \ldots, v_n(t))$, $t = 1, \ldots, T$, consists of a graph representing the instantaneous and lagged effects, a set of probability distributions over the disturbances $\boldsymbol{e}$, and linear equations*

$$\boldsymbol{v}(t) = \mathbf{B}_0 \boldsymbol{v}(t) + \sum_{i=1}^{q} \mathbf{B}_i \boldsymbol{v}(t-i) + \boldsymbol{e}(t) \tag{3.12}$$

where $\mathbf{B}_0$ is the connection matrix containing the instantaneous effects, and can be permuted to strictly lower triangular form (by the acyclicity of the instantaneous effects), and $\mathbf{B}_i$, $i = 1, \ldots, q$, are connection matrices containing the lagged effects for lags $1, \ldots, q$. The vector $\boldsymbol{e}(t)$ contains the unobserved disturbance terms $e_i(t)$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, which are mutually independent both of each other and over time, and identically distributed over time.

Note that for $q = 0$ this model reduces to a linear SEM, as shown in Example 3.2. An example of a 2^{nd}-order SVAR is given in Example 3.7.

All assumptions made in the earlier parts of this chapter are assumed to apply also to the graphs involving time, and hence the theorems hold as well. Note however that one has to 'unroll' the graph sufficiently far back in time in order to capture dependencies happening in the past. Also note that for a first order SVAR the causal Markov condition implies that the future is independent of the past given the present.

Chapter 4

Causal Effect Identification

In this chapter, we present the relevant existing methods addressing research question Q1. The main focus lies on *identifying* a causal effect from passive observational data, i.e. estimating the postinterventional probability distribution under a *hypothetical* intervention. One main obstacle in this task is bias due to confounding variables which can introduce spurious correlations, as explained in the introduction by means of examples.

We start with formally defining causal effect identification in Section 4.1. In Section 4.2 we then discuss existing methods for identifying causal effects from a causally insufficient CBN or SEM, assuming that the underlying DAG is known. Finally, in Section 4.3 we present some solutions for estimating the effect of certain interventions when neither the DAG is known, nor the full set of variables is observed.

Two widely used approaches for estimating causal effects from passive observational data not further addressed are propensity score matching (Rosenbaum and Rubin, 1983), and instrumental variables, see for example Pearl (2000). Propensity score matching is one way of implementing the truncated factorization formula (Equation (3.2), page 18) in case of a single intervention, whereas the method of instrumental variables allows the identification of causal effects in certain linear SEMs. For recent illustrations of both approaches see Berzuini et al. (2012).

We use the following notation: We denote with x the cause, and with y the effect, and want to estimate the causal effect of x on y (with both x and y observed variables). Observed variables in general are denoted with w, sets of observed variables with $\mathcal{W}$, latent variables with u, and sets of latent variables with $\mathcal{U}$. If it is not specified whether a variable is observed or latent, we continue using v.

4.1 Formal Definition

As formally stated in Definition 3.5 (page 22), the causal effect of x on y is given by the postinterventional probability distribution $p(y \mid do(x))$, which is, for each value of x, the probability distribution over y when forcing x to that value. Thus, the preferred way of approaching the problem of estimating the effect of an intervention is to actually intervene and set the variable to a certain value, and observe the other variables under this intervention. The postinterventional distribution can then be estimated directly from the data set of this experiment. Examples were given in the introduction.

In many cases such experiments are however not possible, for ethical, financial, or practical reasons, and researchers have to estimate the effect of an intervention from *passive observational* (i.e. *non-experimental*) data, see the introduction for examples. The formal way of stating this problem is to determine whether a causal effect is *identifiable* from passive observational data (Pearl, 2000).

Definition 4.1 (Causal Effect Identifiability)**.** *The causal effect of x on y is* identifiable *from a CBN or SEM over $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if and only if $p(y \mid do(x))$ can be computed uniquely from any positive probability distribution over the observed variables $\mathcal{W} \subseteq \mathcal{V}$.*

4.2 Identifying Effects with DAG Known

In this section, the underlying DAG $\mathcal{G}$ over $\mathcal{V} = \mathcal{W} \cup \mathcal{U}$ of a CBN or SEM is assumed to be known, yet the parameters of the model are unknown. This situation may arise when enough knowledge about the domain in question is available, such that experts have a good understanding of which variables are causally connected. The goal is to determine whether we can identify a causal effect $p(y \mid do(x))$, for $x \in \mathcal{W}$ and $y \in \mathcal{W}$, and obtain an expression for it based on the probabilities of the observed variables.

Note that if all variables in $\mathcal{V}$ are observed, the causal effect of x on y is always identifiable by *adjusting* for the parents of the intervened variable x, and is given by

$$p(y \mid do(x)) = \sum_{pa_x} p(y \mid x, pa_x)\, p(pa_x), \tag{4.1}$$

where the sum is over all values of the parents of x. This implies in fact that it is enough that all parents of x are among the observed variables. Equa-

tion (4.1) follows from the truncated factorization formula (Equation (3.2), page 18). Pearl (2000) called this *Adjustment for Direct Causes*.

4.2.1 Back-Door Adjustment

When some arbitrary variables of the DAG are hidden, the following definition together with the back-door adjustment in the theorem below (Pearl, 1993a) state one criterion to possibly identify the causal effect.

Definition 4.2 (Back-Door Criterion, Admissible Set)**.** *We are given the underlying DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a CBN or SEM, and a set of observed variables $\mathcal{W} \subseteq \mathcal{V}$. Let $x \in \mathcal{W}$, $y \in \mathcal{W}$, and $\mathcal{Z} \subseteq \mathcal{W} \setminus \{x, y\}$. The set $\mathcal{Z}$ satisfies the* back-door criterion *with regard to the ordered pair (x, y) if and only if*

(i) no variable in $\mathcal{Z}$ is a descendant of x, and

(ii) $\mathcal{Z}$ blocks every back-door path *from x to y, i.e. every path between x and y that contains an arrow into x (i.e. '$x \leftarrow$').*

A set $\mathcal{Z}$ fulfilling the back-door criterion with respect to a pair (x, y) is called admissible*.*

Theorem 4.1 (Back-Door Adjustment)**.** *Let x, y and $\mathcal{Z}$ be as in Definition 4.2. If $\mathcal{Z}$ satisfies the back-door criterion with regard to (x, y) then the causal effect of x on y is identifiable and is given by*

$$p(y \,|\, do(x)) = \sum_{\mathcal{Z}} p(y \,|\, x, \mathcal{Z}) p(\mathcal{Z}). \tag{4.2}$$

The back-door criterion ensures that all paths that would introduce bias to the estimate of the causal effect of x on y are blocked. If all parents of x were observed, then $\mathcal{Z} = pa_x$ fulfills the back-door criterion, i.e. this criterion generalizes the adjustment for direct causes to an arbitrary adjustment set $\mathcal{Z}$.

Example 4.1 (Back-Door Criterion)**.** *In Figure 4.1 (a), the back-door criterion with regard to (x, y) holds with $\mathcal{Z} = \{w_1, w_2\}$, $\mathcal{Z} = \{w_2, w_3\}$, or $\mathcal{Z} = \{w_1, w_2, w_3\}$, but no other set $\mathcal{Z} \subseteq \mathcal{W}$. In Figure 4.1 (b), the back-door criterion is not fulfilled by $\mathcal{Z} = \{w\}$ nor by $\mathcal{Z} = \emptyset$.*

In small graphs it is relatively easy to check for a given set $\mathcal{Z}$ whether it fulfills the back-door criterion, and hence, by going through all possible sets $\mathcal{Z}$ to conclude whether an admissible set exists. For larger graphs, Tian et al. (1998) introduced an algorithm to efficiently search for an admissible set $\mathcal{Z}$ that is minimal, i.e. such that no subset of $\mathcal{Z}$ is admissible.

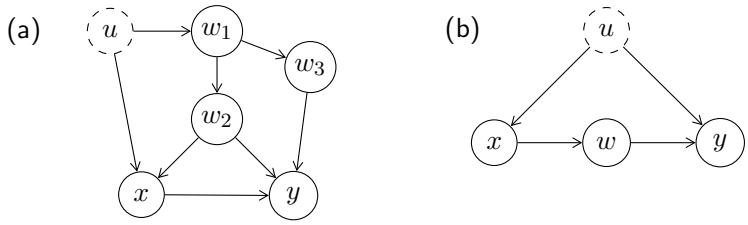


Figure 4.1: Adjustment Criterion. In (a) the back-door criterion holds for instance with $\mathcal{Z} = \{w_1, w_2\}$, whereas in (b) the back-door criterion is not fulfilled.

A special case that proves useful for our results is the back-door adjustment in *linear SEMs*, stated in the following theorem (Pearl, 1998).

Theorem 4.2 (Back-Door Adjustment in Linear SEMs)**.** *Assume that the underlying DAG $\mathcal{G}$ of a linear SEM is known, and let x, y and $\mathcal{Z}$ be as in Definition 4.2. If $\mathcal{Z}$ is an admissible set with regard to (x, y), then the* total *effect of x on y is identifiable, and given by c_x, the coefficient of x in the regression of y on x and $\mathcal{Z}$ using ordinary least squares:*

$$y = c_x x + \sum_{z \in \mathcal{Z}} c_z z + r_y. \tag{4.3}$$

Note that in general a regression coefficient does not need to be related in any way to the causal effect, but merely expresses some form of correlation between x and y given $\mathcal{Z}$. The back-door adjustment gives a criterion for when this expression coincides with the total effect (see Pearl, 2000, 2nd edition p.161), illustrated in the following example.

Example 4.2 (Adjustment in Linear SEMs)**.** *We use the linear SEM of Figure 3.4 (a) of Example 3.4 (page 22/23), and we want to estimate the causal effect of v_1 on v_3. Wright's method (1934) for path coefficients tells us that the total causal effect is given by $b_{3,1} + b_{3,2}\, b_{2,1}$. Since the empty set $\mathcal{Z} = \emptyset$ is admissible with regard to (v_1, v_3), Theorem 4.2 can be used to calculate the causal effect directly as the regression coefficient c_1 of the following model*

$$v_3 = c_1 v_1 + r_3,$$

using the OLS formula of Equation (2.8) *(page 13), and assuming that the variables have zero mean:*

$$c_1 = \frac{cov(v_1, v_3)}{\sigma^2_{v_1}} \quad \textit{and}$$

$$\begin{aligned} cov(v_1, v_3) &= E(v_1\, v_3) = E(v_1\, (b_{3,1}\, v_1 + b_{3,2}\, v_2 + e_3)) \\ &= E(e_1\, (b_{3,1}\, e_1 + b_{3,2}\, (b_{2,1}\, v_1 + e_2) + e_3)) \\ &= b_{3,1}\, E(e_1^2) + b_{3,2}\, b_{2,1} E(e_1^2) + \\ &\quad b_{3,2}\, E(e_1) E(e_2) + E(e_1) E(e_3) \\ &= (b_{3,1} + b_{3,2}\, b_{2,1})\, \sigma_{v_1}^2. \end{aligned}$$

4.2.2 Other Approaches

A similar approach to the back-door adjustment of Theorem 4.1 is the *Front-Door Adjustment* (Pearl, 1993b). This adjustment method allows the covariates in the adjustment set $\mathcal{Z}$ to lie between the cause x and the effect y, potentially yielding identifiability of the causal effect in cases where back-door adjustment fails (the simplest such example is the graph of Figure 4.1 (b)).

A more general tool to identify causal effects is provided by the three rules of the *do*-calculus introduced by Pearl (1994, 1995). Iteratively applying these rules may allow transforming a postinterventional distribution into an expression of observed (conditional) probability distributions only, and hence identifying the causal effect. Huang and Valtorta (2006), and Shpitser and Pearl (2006) showed that the *do*-calculus is *complete*, in the sense that a causal effect is identifiable if and only if it can be transformed into an expression over observed conditional probabilities using the three rules of the *do*-calculus.

One drawback of the *do*-calculus is that it does not directly provide a procedure specifying how to reach the transformation from the postinterventional distribution to observational probability distributions. Towards this end, general algorithms for the identification of (conditional) interventional distributions were introduced by Tian and Pearl (2002), Tian (2004), Shpitser and Pearl (2006), and Shpitser et al. (2011).[1]

4.3 Identifying Effects with DAG Unknown

In many situations the DAG corresponding to the data generating process is not known, so the methods of Section 4.2 cannot be applied directly. To still be able to infer causal effects from passive observational data, typically

[1] Conditional interventional distributions are defined as $p(y \mid \boldsymbol{z}, do(x)) = \frac{p(y, \boldsymbol{z}, \mid do(x))}{p(\boldsymbol{z} \mid do(x))}$ (Shpitser and Pearl, 2006), and are causal effects in a subpopulation determined by the values of $\boldsymbol{z}$, for instance, the effect of alcohol (x) on mortality (y) in the male population of a specified country ($\boldsymbol{z}$).

some other information is required. One common assumption in the methods discussed in this thesis is a known partial order among the observed variables.

Definition 4.3 (Partial ordering assumption)**.** *For a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \mathcal{U} \cup \mathcal{W} \cup \{x, y\}$, the variables in $\mathcal{U}$ being latent, and the variables in $\mathcal{W}$, as well as x and y, observed, the* partial ordering assumption *$\mathcal{W} \prec x \prec y$ holds if and only if there exists a valid causal order in which the variables in $\mathcal{W}$ precede x, which in turn precedes y.*

Note that this assumption implies that the *total* causal effect of x on y is equal to the *direct* causal effect (with regard to the set of observed variables $\mathcal{W} \cup \{x, y\}$).

This assumption is often reasonable, for example if a (partial) temporal ordering among the variables is known. Consider the field of medicine, where one wants to identify the effect of a treatment or exposure variable x (for example some specific form of surgery) on an outcome variable y (mortality), possibly adjusting for some observed covariates $\mathcal{Z} \subset \mathcal{W}$ (gender, age, and general health indicators at the time of the surgery).

Furthermore, under the partial ordering assumption, the back-door criterion is complete (Shpitser et al., 2010): If a set is not admissible in a DAG $\mathcal{G}$, then there exist models with underlying graph $\mathcal{G}$ in which adjusting for this set yields a biased (and inconsistent) estimator of the causal effect. Thus, finding admissible sets $\mathcal{Z}$ among the observed covariates $\mathcal{W}$ is an appropriate approach. This approach is taken in the methods discussed in the following two subsections.

4.3.1 Simple Approaches

We first discuss three simple methods for creating a possible adjustment set, given the partial ordering assumption. These strategies are (i) including all of the covariates in $\mathcal{W}$, (ii) including none of the covariates, and (iii) including all observed common causes of x and y (i.e. variables which are causes of x and also causes of y not via x). While for the first two strategies no background knowledge is needed, for the last approach it is necessary to identify causes of x and y using background information. In practice, especially the first strategy of adjusting for all covariates is commonly used. However, all these simple approaches may lead to an adjustment set that is *not* admissible, i.e. that does not block all back-door paths, as the following example shows (Spirtes et al., 1998; Greenland et al., 1999; Spirtes, 2000; VanderWeele and Shpitser, 2011).

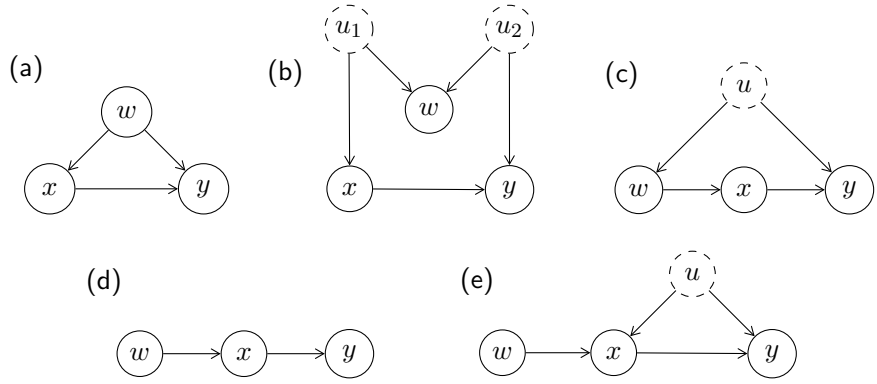


Figure 4.2: Figures (a) to (c) demonstrate simple adjustment criteria (for details see Example 4.3), with admissible sets with regard to (x, y) given by (a) $\mathcal{Z} = \{w\}$, (b) $\mathcal{Z} = \emptyset$, and (c) $\mathcal{Z} = \{w\}$. Figures (d) and (e) are used to illustrate the method of Section 4.3.2.

Example 4.3. *Consider the DAGs of the three data generating models in Figure 4.2 (a)-(c). Each of the simple methods for finding an adjustment set discussed above fails for one or more of these models: Adjusting for all observed covariates yields an admissible set for (a) and (c), however not for (b). In contrast, by adjusting for none of the variables we obtain an admissible set for (b), though not for (a) and (c). Finally, adjusting for all common causes, given the required background knowledge, yields in (a) the set $\mathcal{Z} = \{w\}$ and in (b) the set $\mathcal{Z} = \emptyset$, which are admissible in the respective models. However, in (c) the resulting adjustment set is the empty set, since w causes y only via x. This adjustment set is not admissible.*

VanderWeele and Shpitser (2011) introduced a criterion always yielding an admissible set, when such a set truly exists. In addition to the partial ordering assumption, they require that it is known which covariates are causes of x or y (as for the strategy of including all common causes). They show that the set of all those observed covariates which are causes of x, or of y, or of both yields an admissible set, i.e. blocks all back-door paths. For the graphs in Figure 4.2 (a)-(c) this criterion indeed yields the correct adjustment sets. However, in case there does not exist an admissible set, the criterion of VanderWeele and Shpitser (2011) fails to detect this.

4.3.2 Methods Based on Dependencies and Independencies

Statistical dependencies and independencies among the observed variables can be used to identify causal effects without knowing the underlying DAG.

We first want to point out one theoretical limitation of this approach. Even with the partial ordering assumption, some models entail the same set of independencies over the observed variables, even though they imply different admissible sets. For instance, the DAGs of Figure 4.2 (a) and (b) both do not entail any independencies among w, x, and y, and it is thus impossible to conclude from dependencies and independencies alone whether the variable w should be adjusted for to identify the causal effect. However, in some models it may well be possible to reach a conclusion.

Using the faithfulness and partial ordering assumptions, Spirtes and Cooper (1999), and Chen et al. (2007) both gave a simple criterion to identify a causal effect with an empty admissible set.[2] In essence, they require an *exogenous* variable w, i.e. a variable w which is a root in the underlying generating graph, and that x, y and w fulfill the following properties:

(i) $x \not\perp\!\!\!\perp y$,
(ii) $w \not\perp\!\!\!\perp y$, and
(iii) $w \perp\!\!\!\perp y \,|\, x$.

If these conditions hold, then x is a cause of y and there is no confounder (neither latent nor observed) between x and y. Hence, the effect of x on y is identifiable, and equal to the conditional probability $p(y \,|\, x)$ (which can be seen using the back-door criterion with the admissible set $\mathcal{Z} = \emptyset$).

This criterion is most easily understood with an example. The simplest DAG in which these conditions hold is given by the graph in Figure 4.2 (d). When adding a (latent) confounder between x and y, as in Figure 4.2 (e), the conditional independence of condition (iii) is destroyed. In general, if there is an active back-door path between x and y, conditions (i) to (iii) cannot all hold, and hence confounding can be detected.

[2] In fact, they only assume that x and y are the last two variables in the causal order, but they do not require any specific order among x and y.

Chapter 5

Structure Learning

In this chapter, the relevant existing work addressing research question Q2 is presented. We explain methods for learning the underlying DAG of a CBN or SEM from passive observational data. For this task clever search algorithms are needed, since the number of DAGs grows superexponentially in the number of nodes: for 3 nodes there are 25 DAGs (shown in Figure 5.1), for 4 nodes 543 DAGs, for 5 nodes 29281 DAGs, and for 6 nodes already 3781503 DAGs.

There is a vast literature on this topic, and we will only discuss the methods relevant for understanding the novel methods introduced in this thesis. These include the constraint based approach, and methods for estimating linear non-Gaussian acyclic models, as well as applications of these methods in SVAR models. One large bundle of alternative methods not considered here are score based methods (Heckerman et al., 1995; Meek, 1997; Chickering, 2002; Koivisto and Sood, 2004; Silander and Myllymäki, 2006). Furthermore, much attention has rather recently been devoted to the problem of learning non-linear models (Hoyer et al., 2009; Mooij et al., 2009; Zhang and Hyvärinen, 2010; Peters et al., 2011), as well as combining information from several (experimental or non-experimental) data sets (Tillman, 2009; Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Tsamardinos et al., 2012; Hyttinen et al., 2012). These methods are however out of the scope of this thesis.

5.1 Constraint Based Methods

Here, we discuss methods for structure learning based on statistical dependencies and independencies among the observed variables. We start with the PC algorithm (after its inventors' first names, Peter and Clark,

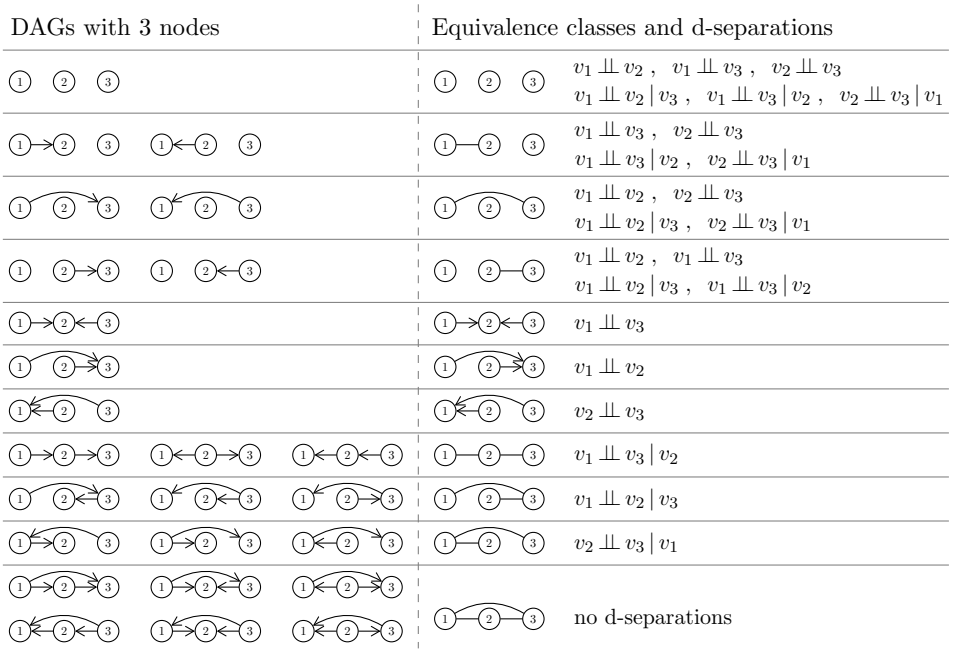


Figure 5.1: The left part of the figure shows all 25 DAGs over three variables v_1, v_2, and v_3, indicated as numbers 1, 2, and 3 in the graphs. These are grouped into the 11 equivalence classes in which these 25 DAGs can be divided, given by the rows of the figure. The pattern of each equivalence class and the d-separation relations entailed by the DAGs contained in this class are shown in the right part of the figure.

Spirtes and Glymour, 1991), a structure learning method for the causally sufficient case. We then move on to the FCI algorithm (Fast Causal Inference, Spirtes et al., 1993, 1999), which allows for latent variables and selection bias. These methods are suited for both discrete and continuous data (linear/non-linear, Gaussian/non-Gaussian), as all that is required is an appropriate independence test, some of which were reviewed in Section 2.2.2.

5.1.1 PC Algorithm

We assume that we are given a data set generated by a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which all variables $\mathcal{V}$ are observed (i.e. the set of observed variables is causally sufficient), and that the probability distribution p associated with the model is faithful to $\mathcal{G}$.

Since the PC algorithm is based on independencies, we first point out

an important property of DAGs, termed *Markov equivalence*, which limits the amount of information one can learn using this approach.

Definition 5.1 (Markov equivalence)**.** *Two DAGs are* Markov equivalent *if and only if they entail the same d-separation relations.*

To judge whether two DAGs are Markov equivalent, the conditions of the following theorem can be used (Verma and Pearl, 1990).

Theorem 5.1 (Markov equivalence)**.** *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same unshielded colliders.*

Based on this theorem, Markov equivalent DAGs can be represented by a pattern, termed the *Markov equivalence class*.

Example 5.1 (Markov equivalence)**.** *In Figure 5.1, all DAGs with three variables are shown. The DAGs are collected in rows according to their equivalence classes, which are shown next to the graphs, as are the d-separation relationships holding in each graph of the corresponding equivalence class. For example, all graphs with three edges form one equivalence class (bottom row in the figure), since they do not entail any d-separation relations and hence cannot be distinguished from each other. DAGs containing two edges can be divided into 6 equivalence classes: First of all, the condition of having the same skeleton narrows down the size of each equivalence class (for example having an edge between v_1 and v_2, and v_2 and v_3, but not between v_1 and v_3). Furthermore, the condition of having the same unshielded colliders yields that there is an equivalence class containing only one graph ($v_1 \to v_2 \leftarrow v_3$). However, it is not possible to distinguish between $v_1 \to v_2 \to v_3$, $v_1 \leftarrow v_2 \to v_3$, and $v_1 \leftarrow v_2 \leftarrow v_3$, yielding the equivalence class $v_1 - v_2 - v_3$.*

The aim of the PC algorithm is to, under the given assumptions, learn the pattern of the data generating graph solely from the dependencies and independencies in the data. The PC algorithm proceeds in two phases: the *adjacency phase* and the *orientation phase*, as schematically demonstrated in Figure 5.2, and further explained below. Two closely related but less efficient algorithms are the IC algorithm (Inductive Causation, Verma and Pearl, 1990), and the SGS algorithm (after its inventors' last names, Spirtes, Glymour and Scheines; Spirtes et al., 1990).

In the adjacency phase, the PC algorithm finds the *skeleton* of the pattern of the underlying DAG. This is based on the fact that there is an edge between two variables v_i and v_j in the DAG $\mathcal{G}$ if and only if v_i and v_j are d-connected given *every* possible set $\mathcal{Z} \subseteq \mathcal{V} \setminus \{v_i, v_j\}$. Since we assume that

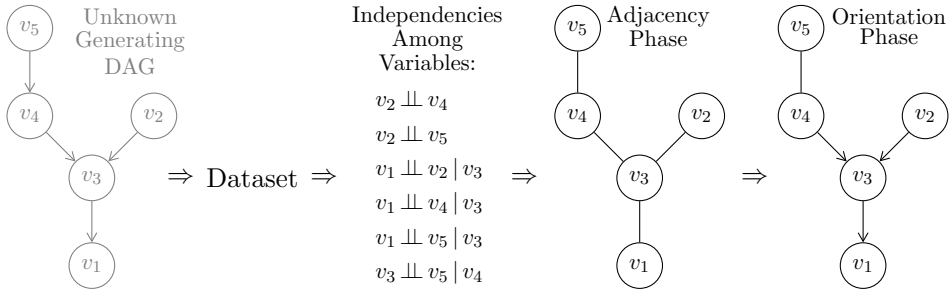


Figure 5.2: Demonstration of the PC algorithm. The data set is generated by an unknown causal model. The PC algorithm uses the (testable) independencies from the data to generate the skeleton of the underlying DAG in the adjacency phase, and then to orient as many edges as possible in the orientation phase to obtain the pattern representing the Markov equivalence class of the DAG.

the probability distribution p is faithful to $\mathcal{G}$, i.e. d-separation relations in $\mathcal{G}$ and (conditional) independencies in p are equivalent (see Equation (3.9), page 26), we thus can remove the edge between two variables v_i and v_j if we can find even a single set $\mathcal{Z}$ for which they are (conditionally) independent. Going through all these independencies is (roughly) done by starting with the empty conditioning set $\mathcal{Z}$, removing any possible edges, and then continuing with conditioning sets of cardinality one, two, and so on.

In the orientation phase, the skeleton of the adjacency phase together with Theorem 5.1 can be used to derive *orientation rules*, which allow orienting some of the edges to obtain the pattern of the Markov equivalence class of the underlying DAG. The first rule orients the unshielded colliders of the pattern: For any triple (v_i, v_k, v_j), with $v_i - v_k - v_j$ in the skeleton and v_i not adjacent to v_j, we can orient $v_i \to v_k \leftarrow v_j$ if and only if v_k is not in the conditioning set $\mathcal{Z}$ yielding $v_i \perp\!\!\!\perp v_j \mid \mathcal{Z}$ (by the definition of d-separation). In addition to these colliders, the orientation of some other edges may be determined uniquely for the equivalence class. For example, by the acyclicity assumption, if $v_i \to v_k \to v_j$ and $v_i - v_j$ are in the skeleton, we must orient $v_i \to v_j$. In total there are three such additional rules.

Meek (1995) showed that these orientation rules are sound and complete, i.e. the pattern contains all and only those orientations which are common to all the elements of the equivalence class. It is also straightforward to incorporate background knowledge into the PC algorithm, such as existence or non-existence of some edges, orientation of some edges, or a time order among the variables (Meek, 1995; Spirtes et al., 1993).

5.1.2 FCI Algorithm

We here assume again that the data are generated by a CBN or SEM over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. However, as opposed to the setting of the previous section, the set of observed variables may here only be a subset of $\mathcal{V}$, i.e. $\mathcal{V} = \mathcal{W} \cup \mathcal{U}$ with $\mathcal{W}$ observed, and $\mathcal{U}$ latent variables. This typically yields a set $\mathcal{W}$ of observed variables which is *not* causally sufficient. Note that we do not need to know the set $\mathcal{U}$ of latent variables, but only need to assume that such a set exists. Furthermore, we assume that the distribution associated with the model is faithful to $\mathcal{G}$. We can thus discuss the FCI algorithm (Fast Causal Inference, Spirtes et al., 1993, 1999), an extension of the PC algorithm to the causally *insufficient* case.[1]

As working explicitly with latent variables can be cumbersome (see for example Richardson and Spirtes (2002) for a list of reasons), the FCI algorithm is based on so called *maximal ancestral graphs* (MAGs, Richardson and Spirtes, 2002) over the observed variables $\mathcal{W}$.[2] MAGs are ancestral graphs, in which for any two non-adjacent nodes there exists a set of vertices that *m-separates* them, where m-separation is a generalization of the d-separation criterion for DAGs to ancestral graphs (Richardson and Spirtes, 2002). Furthermore, MAGs not only can represent latent variables (using bidirected arrows $\leftrightarrow$), but can also account for *selection bias* (using undirected arrows $-$). However, for simplicity, we will exclude selection bias from all following descriptions.

The interpretation of the edges of a MAG is as follows:

- A directed edge $w_1 \rightarrow w_2$ is interpreted as w_1 being a (direct or indirect) cause of w_2, and w_2 not being a cause of w_1, and
- a bidirected edge $w_1 \leftrightarrow w_2$ is interpreted as neither w_1 being a cause of w_2 nor w_2 being a cause of w_1.

First of all, note that every DAG is a special case of a MAG (simply not containing any bidirected edges). Furthermore, for every causal model over a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathcal{W} \cup \mathcal{U}$ as above, there exists a MAG over the observed variables $\mathcal{W}$ representing the same independencies (m-separation relations) among the observed variables $\mathcal{W}$ as in the DAG, and retaining ancestral relationships of the DAG. For example, the MAG over the observed variables w, x and y of the DAG in Figure 4.2 (e) (page 35)

[1] Verma and Pearl (1990) introduced a variant of the IC algorithm also accounting for latent variables (later termed IC* (Pearl, 2000)). Since the FCI algorithm is more efficient as well as further developed, we will only consider the FCI algorithm here.

[2] Earlier versions of the algorithm are based on so called inducing path graphs (Spirtes et al., 1993). The algorithm is essentially the same but the interpretation of the output differs.

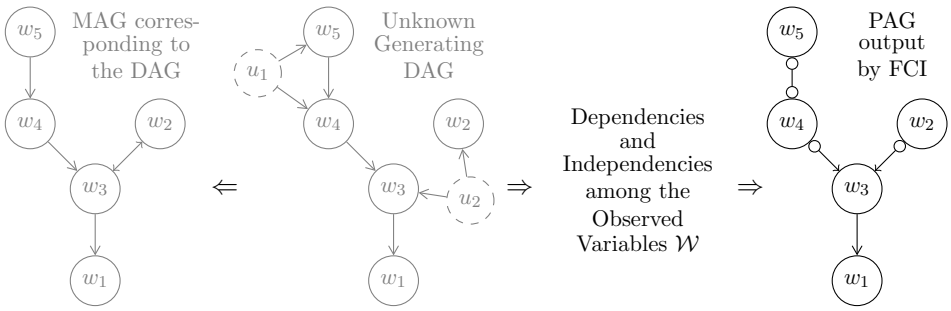


Figure 5.3: Demonstration of the FCI algorithm, MAGs and PAGs. For details see Example 5.2.

is given by the fully connected graph containing the three edges $w \to x$, $x \to y$, and $w \to y$, since (i) there are edges in the DAG from w to x, and from x to y, and (ii) w and y are always d-connected when disregarding u, and w is an ancestor of y. An algorithm to transform a DAG over observed variables $\mathcal{W}$ and hidden variables $\mathcal{U}$ into a MAG over the observed variables $\mathcal{W}$ is given by Richardson and Spirtes (2002). A further example of a MAG is given in Figure 5.3.

Similarly to the PC algorithm, the FCI algorithm uses dependencies and independencies among the observed variables to construct in two steps (adjacency and orientation phase) a so called *partial ancestral graph* (PAG, Richardson, 1996), which corresponds to the equivalence class of the underlying MAG. Uncertainty about the orientation of an edge is indicated with a circle mark '$\circ$' in the PAG. We illustrate FCI in the following example.

Example 5.2 (FCI Algorithm)**.** *In Figure 5.3, the graph in the middle depicts the (unknown) data generating process over observed variables $\mathcal{W} = \{w_1, \ldots, w_5\}$, and latent variables $\mathcal{U} = \{u_1, u_2\}$. This DAG can be transformed into a MAG over the observed variables, which is shown in the left graph of the figure.*

The FCI algorithm uses the testable independencies among the observed variables $\mathcal{W}$ to infer the PAG representing the equivalence class of the MAG (shown in right graph of the figure). Circle-marks represent unknown orientations; for instance, the circle at w_2 implies that from independencies alone it is not possible to infer whether w_2 is an ancestor of w_3.

While Theorem 5.1 gives a simple *necessary and sufficient* graphical criterion for two DAGs to be Markov equivalent, this condition is only necessary for two MAGs to be equivalent. There are different graphical criteria to define Markov equivalence for MAGs (Spirtes and Richardson,

1997; Zhao et al., 2005; Ali et al., 2009), however describing these is out of the scope of this thesis.

Furthermore, Zhang (2008) augmented the rules in the orientation phase (there are in total 11 rules, including the ones handling selection bias) such that FCI is sound and complete, i.e. all and only those edge marks which are common to all MAGs of the equivalence class represented by the PAG are oriented. As for the PC algorithm, it is straightforward to include background knowledge about a temporal ordering in FCI. However, it is not known whether FCI remains complete when including such knowledge.

5.2 Linear Non-Gaussian Acyclic Model Estimation

The methods presented in the previous section can only identify the underlying DAG/MAG of a causal model up to Markov equivalence. Here, we review a class of methods for *linear* SEMs with *non-Gaussian* error terms, also termed Linear Non-Gaussian Acyclic Models (LiNGAM, Shimizu et al., 2006). These methods exploit higher order statistics yielding full identifiability of the model when no latent variables are present. We start with reviewing two approaches to estimate such models. We then present a measure to infer the causal direction among only two variables, and how it can be used to estimate a LiNGAM model over several variables. Towards the end of this section we discuss extensions of these methods for models with latent variables, as well as with multidimensional variables.

5.2.1 ICA-LiNGAM

Given a LiNGAM model over a DAG $\mathcal{G}$, and assuming causal sufficiency for the observed variables, the ICA-LiNGAM algorithm (Shimizu et al., 2006) outputs, in the large sample limit, the underlying DAG $\mathcal{G}$. Note that the faithfulness assumption is *not* needed for this approach.

Estimating a LiNGAM model basically consists of two steps: obtaining a causal order among the variables, and estimating the connection strengths of the causal effects given the causal order. The second step is simple, and can be done using OLS by regressing the second variable in the causal order on the first, the third variable on the first and the second, and so on (or alternatively by using the Cholesky decomposition of the covariance matrix of the variables). Additionally, some of the causal effects (regression coefficients) might be set to zero if they are not statistically significant, which means that some of the edges in the DAG are cut out. The crucial

point in the algorithm thus is the first step of obtaining a causal order. Towards this end, we write a linear SEM using the matrix equation as shown in Example 3.2 (page 17)

$$\boldsymbol{v} = \mathbf{B}\boldsymbol{v} + \boldsymbol{e}. \tag{5.1}$$

In this section we only consider models where the distribution of the disturbances $\boldsymbol{e}$ are *non-Gaussian*, which is essential for full identification of the model.

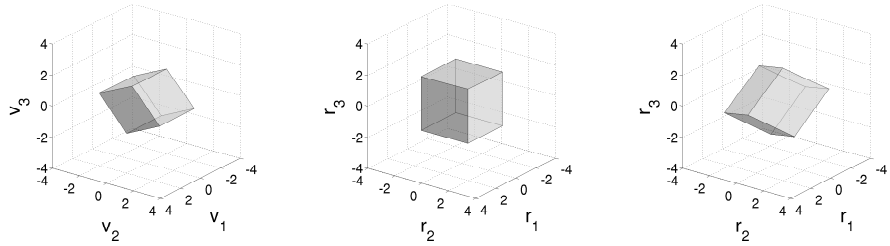
In ICA-LiNGAM, Equation (5.1) is rewritten in its *reduced form*, corresponding to an ICA (Independent Component Analysis, Hyvärinen et al., 2001) model

$$\boldsymbol{v} = \mathbf{A}\boldsymbol{e} \tag{5.2}$$

with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, so that the observed variables $\boldsymbol{v}$ can be seen as a linear mixture of the independent disturbances $\boldsymbol{e}$, termed *sources* in the ICA literature. In essence, ICA identifies a matrix $\mathbf{A}$ by whitening and rotating the data such that $\mathbf{A}^{-1}\boldsymbol{v}$ yields independent sources $\boldsymbol{e}$. In this step, non-Gaussianity is required, since Gaussian distributions are rotation symmetric (once they are whitened), and hence any rotation of white Gaussian data yields independent (uncorrelated) sources.

The matrix $\mathbf{A}$ inferred from ICA is unique up to arbitrary scaling, sign change and permutation of the columns. These ambiguities can be resolved using the acyclicity assumption of the underlying model (Shimizu et al., 2006), so that we obtain a unique matrix $\mathbf{A}$, which in turn yields a unique connection matrix $\mathbf{B}$. Note that the residuals $\boldsymbol{e}$ are only independent when estimating the matrix $\mathbf{B}$ in a correct causal order (which is found using the ICA matrix $\mathbf{A}$), as demonstrated in the following example.

Example 5.3 (ICA-LiNGAM)**.** *We use the linear SEM of Example 3.2 over the DAG $v_2 \to v_3 \to v_1$ (Figure 3.2, page 17), with the disturbances e_i, $i = 1, 2, 3$, following a uniform distribution with mean 0 and variance 1. The data are shown in Figure 5.4 (a). In (b) we show the estimated residuals $\boldsymbol{r}$ when estimating the matrix $\mathbf{B}$ using the correct causal order. These are independently and uniformly distributed, representing the distribution of the underlying disturbances $\boldsymbol{e}$. The independence of these residuals can be seen from the figure since the cube is aligned with the axes. Thus, for any given value of r_1, for instance, the values of r_2 and r_3 do not depend on this value. In contrast, in (c) we plot the residuals when estimating the matrix $\mathbf{B}$ along the order $v_1 \prec v_3 \prec v_2$, and can see that these are not independent, even though they are uncorrelated. The dependence of the residuals is indicated by the rotated cuboid, such that, for instance, for large values of r_1, the values of r_3 are restricted to be around 0.*



(a) Data generated by the model in Figure (3.2)

(b) Residuals when estimating $\mathbf{B}$ in the correct causal order $v_2 \prec v_3 \prec v_1$

(c) Residuals when estimating $\mathbf{B}$ in a wrong causal order $v_1 \prec v_3 \prec v_2$

Figure 5.4: Demonstration of ICA-LiNGAM. The schematic plots show the distributions of the data (a), and the residuals when estimating the model in two different orders, (b) and (c). The cuboid of each plot indicates the area in which the distribution is non-zero. Inside the cuboid the data are distributed uniformly.

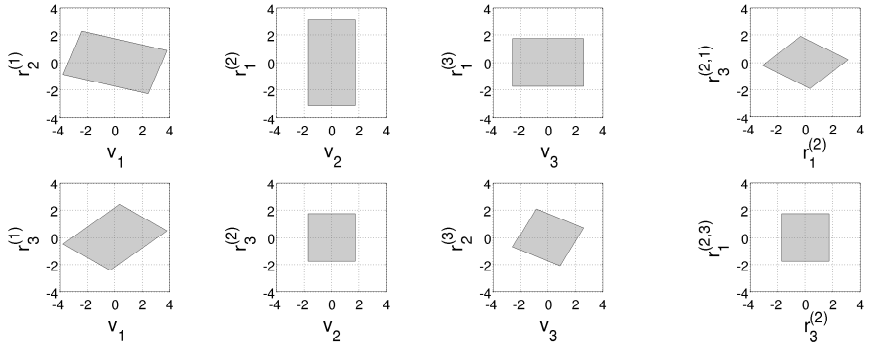
5.2.2 DirectLiNGAM

Under the same assumptions as for ICA-LiNGAM, DirectLiNGAM (Shimizu et al., 2011) tackles the problem of learning a causal order in an iterative manner. First, the algorithm searches for an exogenous variable (i.e. a root variable in the underlying model, which always exists due to the acyclicity assumption). Then, the effect of this exogenous variable is regressed out from all other variables, and the whole process is repeated on this smaller data set.[3] We restate two theorems by Shimizu et al. (2011), formalizing this method. The first one gives a criterion to find an exogenous variable.

Theorem 5.2. *Given that $\boldsymbol{v}$ follows a LiNGAM model, let $r_i^{(j)} = v_i - c_{ij}v_j$, $i \neq j$, be the residuals when regressing v_i on v_j using OLS. Then, v_j is exogenous if and only if $v_j \perp\!\!\!\perp r_i^{(j)}$ for all $i \neq j$.*

For this theorem it is essential that the residuals are non-Gaussian, since v_j and $r_i^{(j)}$ are always uncorrelated (by construction of the OLS estimator) and hence for Gaussian variables they are also independent. The second theorem of Shimizu et al. (2011) states that when regressing out the effect of an exogenous variable v_j from all other variables to obtain the residuals $r_i^{(j)}$, $i \neq j$, these residuals follow also a LiNGAM model with the same induced

[3]Note that for finite sample sizes the result of ICA-LiNGAM may depend on the initialization of the parameters in the ICA-algorithm, whereas DirectLiNGAM always yields the same solution.



(a) First Iteration of DirectLiNGAM (b) Second Iteration

Figure 5.5: Demonstration of DirectLiNGAM, plotting schematically the regressor versus the resulting residuals to find an exogenous variable, see Example 5.4 for details.

causal order. This allows searching for an exogenous variable among $r_i^{(j)}$ to find a second variable in the causal order. By iterating this process a full causal order can be found.

Theorem 5.3. *Given that $\boldsymbol{v}$ follows a LiNGAM model, and v_j is exogenous, the obtained residuals $r_i^{(j)} = v_i - c_{ij}v_j$, $i \neq j$, follow a LiNGAM model with the same induced causal order.*

Example 5.4 (Direct-LiNGAM)**.** *We demonstrate the algorithm using the same model as in Example 5.3. To find an exogenous variable, all pairwise OLS regressions are performed: For all $i \neq j$ we obtain $v_i = c_{ij}v_j + r_i^{(j)}$ and test whether $v_j \perp\!\!\!\perp r_i^{(j)}$. Plots of v_j versus $r_i^{(j)}$ are shown in Figure 5.5 (a). The plots show that independence with* both *residuals only holds for v_2 (since both rectangles are aligned with the axes). Thus, v_2 is exogenous and is chosen to be the first variable in the causal order.*

To find the second variable in the causal order, we repeat this step with the residuals of the regressions on v_2. We perform the pairwise regressions of $r_1^{(2)}$ and $r_3^{(2)}$ in both directions meaning that we estimate $r_3^{(2)} = \tilde{c}_{31}r_1^{(2)} + r_3^{(2,1)}$, and $r_1^{(2)} = \tilde{c}_{13}r_3^{(2)} + r_1^{(2,3)}$. The corresponding plots are shown in Figure 5.5 (b). These plots show that $r_3^{(2)}$ is exogenous among $r_1^{(2)}$ and $r_3^{(2)}$, yielding v_3 as the second variable in the causal order.

As now only v_1 is left, we found the causal order to be $v_2 \prec v_3 \prec v_1$.

5.2.3 Pairwise Measure of Causal Direction

Hyvärinen (2010) discussed the problem of inferring the causal order for a two variable LiNGAM model

$$x = e_x \tag{5.3}$$
$$y = bx + e_y \tag{5.4}$$

i.e. determining which of the variables x and y is the cause and which is the effect. The introduced measure is based on the likelihood ratio of the two possible models $x \to y$ and $y \to x$, and various ways of calculating this ratio based on differential entropy approximations, cumulants, and first-order approximations are suggested (Hyvärinen, 2010; Hyvärinen and Smith, 2013).

The pairwise measure can also be used to infer a causal order among n variables $v_1, \ldots, v_n$ following a LiNGAM model, using a DirectLiNGAM style approach (Hyvärinen, 2010). Since for an exogenous variable v_j, the causal model $v_j \to v_i$ holds for all $i \neq j$ (by marginalizing out all other variables), this measure can be used to find an exogenous variable. As in DirectLiNGAM, once an exogenous variable is found, its effect is regressed out from all other variables, and the procedure is repeated on this smaller set of variables.

5.2.4 Latent Variable LiNGAM

So far we have only discussed methods for LiNGAM models without latent variables. Hoyer et al. (2008) introduced the *latent variable LiNGAM* model (lvLiNGAM) and a method to estimate it. The generating equations can again be rewritten as an ICA like model, such that $\boldsymbol{w} = \mathbf{A}\boldsymbol{e}$, with $\boldsymbol{w}$ the observed variables, and $\boldsymbol{e}$ the disturbances of the observed variables as well as the hidden variables. Note that the matrix $\mathbf{A}$ has now more columns than rows (as opposed to the case without latent variables where $\mathbf{A}$ is a square matrix), termed an *overcomplete* ICA basis. However, with latent variables the model is in general not fully identifiable, and estimation of the overcomplete ICA basis is computationally very challenging (only up to 3 observed variables and one hidden variable were used in the simulations in Hoyer et al., 2008).

Recently, Tashiro et al. (2012) introduced a DirectLiNGAM style approach, searching for exogenous variables, as well as sink variables, to estimate a partial order among the observed variables in lvLiNGAM models.

5.2.5 GroupLiNGAM

Kawahara et al. (2010) generalized the LiNGAM model to multidimensional variables, also termed groups of variables, i.e. instead of having $\boldsymbol{v} = (v_1, \ldots, v_n)$ being a vector of scalar random variables, they assume $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ being a vector of multidimensional variables. The linear equations can again be written in matrix form as

$$\boldsymbol{v} = \mathbf{B}\boldsymbol{v} + \boldsymbol{e} \tag{5.5}$$

with $\mathbf{B}$ the connection matrix, which can be permuted to being lower block triangular, such that a causal order among the multidimensional variables $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ exists, and $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ multidimensional, non-Gaussian disturbance terms, such that $\boldsymbol{e}_i \perp\!\!\!\perp \boldsymbol{e}_j$, $i \neq j$. However, the components of each $\boldsymbol{e}_i$, that is, the disturbances within group i, can be correlated.

In their algorithm, termed GroupLiNGAM, Kawahara et al. (2010) aimed at *learning the partition* of the variables $\boldsymbol{v}$, and with it the causal order among these variables. This results in statistically and computationally challenging algorithms (exponential in the number of variables).

5.3 Trace Method

In this section, we describe the basic idea of the *Trace Method* of Janzing et al. (2010), a method designed to infer the causal order among *two multidimensional* variables $\boldsymbol{x}$ and $\boldsymbol{y}$, given linear relationships. The model is given as

$$\boldsymbol{y} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e} \tag{5.6}$$

where $\mathbf{B}$ is an arbitrary connection matrix, which is independently chosen of the covariance matrix $\mathbf{\Sigma}$ of the regressors $\boldsymbol{x}$, and the disturbances $\boldsymbol{e}$ are independent of $\boldsymbol{x}$. No assumptions are made on the distribution of $\boldsymbol{e}$.

The underlying idea is, intuitively, that in the correct causal direction, i.e. $\boldsymbol{x} \to \boldsymbol{y}$ in Equation (5.6), $p(\boldsymbol{x})$ and $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ originate from some 'independent mechanisms', whereas $p(\boldsymbol{y})$ and $p(\boldsymbol{x} \,|\, \boldsymbol{y})$ show some form of dependence. (Janzing and Schölkopf (2010) formally stated this in terms of algorithmic information theory.) The trace method gives an easily computable criterion to detect such dependencies for two multidimensional variables in linear models. Janzing et al. (2010) pointed out that the trace method does not yield any results for scalar variables, and the performance increases for higher dimensional variables.

5.4 SVAR Identification

SVAR models are introduced in Definition 3.9 (page 28) as

$$\boldsymbol{v}(t) = \mathbf{B}_0\boldsymbol{v}(t) + \sum_{i=1}^{q} \mathbf{B}_i\boldsymbol{v}(t-i) + \boldsymbol{e}(t), \tag{5.7}$$

and we here assume that all variables $\boldsymbol{v}(t)$ are observed (i.e. the system is causally sufficient). SVAR models can be used for policy analysis such as forecasting the effect of an intervention in the system, or analyzing causal influences of *shocks* to some variables, i.e. predicting the response (effect) of one variable to an impulse (change) in another variable. Before discussing these in more detail, we explain how SVAR models can be estimated.

Estimation of SVAR models can be done in three steps (see for example Hamilton (1994), Demiralp and Hoover (2003), Moneta (2003), Lütkepohl (2005), Moneta and Spirtes (2006), and Hyvärinen et al. (2010)):

(i) Estimation of the lagged effects using the reduced form *Vector Autoregressive model* (VAR)

$$\boldsymbol{v}(t) = \sum_{i=1}^{q} \mathbf{A}_i\boldsymbol{v}(t-i) + \boldsymbol{d}(t) \tag{5.8}$$

which is connected to the SVAR model of Equation (5.7) by $\mathbf{A}_i = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_i$, and $\boldsymbol{d}(t) = (\mathbf{I} - \mathbf{B}_0)^{-1}\boldsymbol{e}(t)$.

(ii) Estimation of the instantaneous effects matrix $\mathbf{B}_0$ using the error terms $\boldsymbol{d}(t)$ of Equation (5.8)

$$\boldsymbol{d}(t) = \mathbf{B}_0\boldsymbol{d}(t) + \boldsymbol{e}(t). \tag{5.9}$$

(iii) Correction of the estimates for the lagged effects using

$$\mathbf{B}_i = (\mathbf{I} - \mathbf{B}_0)\mathbf{A}_i. \tag{5.10}$$

We now discuss steps (i) and (ii) in more detail. In (i), since the VAR model does not contain instantaneous effects (i.e. the right hand side of Equation (5.8) does not include $\boldsymbol{v}(t)$), the matrices $\mathbf{A}_i$, $i = 1, \ldots, q$, can be relatively straightforwardly estimated from the data. One typical requirement is stability of the VAR process.

Definition 5.2 (Stability)**.** *The VAR model in Equation* (5.8) *is called* stable *if and only if* $\det(\mathbf{I}_n - \mathbf{A}_1 z - \ldots - \mathbf{A}_q z^q) \neq 0$ *for all* $|z| \leq 1$ *(with* det() *denoting the determinant of a matrix), i.e. all roots of this expression lie outside the unit circle.*

From the stability condition follows that the time series is *stationary*, i.e. the first and second moments are time invariant (Lütkepohl, 2005). In a stable VAR we can thus estimate the coefficient matrices of Equation (5.8) using OLS.

For non-stationary time series with roots of $det(\mathbf{I}_n - \mathbf{A}_1 z - \ldots - \mathbf{A}_q z^q)$ on the unit circle, taking differences of the variables may yield a stable VAR model. Such processes are termed *integrated*. However, in some other cases, for so called *co-integrated* processes, it is necessary to estimate the coefficients using a *Vector Error Correction Model* (VECM, Engle and Granger, 1987), from which the estimates of the matrices $\mathbf{A}_i$, $i = 1, \ldots, q$, of Equation (5.8) can be inferred.

Step (ii) of the estimation of an SVAR model is the crucial point to identify the model. Since the matrix $\mathbf{B}_0$ is assumed to be acyclic and the disturbances $\boldsymbol{e}(t)$ to be independent, Equation (5.9) is a linear SEM, and we can apply the techniques discussed in Sections 5.1 and 5.2, on the residuals $\boldsymbol{d}(t)$ of the VAR model, to obtain the instantaneous effects.[4]

In the case of *Gaussian* disturbances, the PC algorithm of Section 5.1 can be used to identify the DAG up to Markov equivalence, typically requiring background knowledge for full identification of the matrix $\mathbf{B}_0$. In contrast, for *non-Gaussian* disturbances, the LiNGAM estimation of Section 5.2 allows full identification of the matrix $\mathbf{B}_0$. Combining VAR estimation with the PC algorithm is for example demonstrated by Demiralp and Hoover (2003), Moneta (2003), and Moneta and Spirtes (2006), whereas VAR estimation combined with LiNGAM was introduced by Hyvärinen et al. (2010), and termed VAR-LiNGAM.

For forecasting a time series based on *observations* of the past, given that the system does not change (i.e. no interventions or shocks), VAR models are sufficient. However, for policy analysis it is important to identify the matrix $\mathbf{B}_0$: For predicting the effect of an intervention it is essential to obtain the correct matrices $\mathbf{B}_i = (\mathbf{I} - \mathbf{B}_0)\mathbf{A}_i$, $i = 1, \ldots, q$, and for predicting the effect of shocks in the disturbances it is important to obtain

[4] In econometrics it is common to use background knowledge to select a causal order among the variables and estimate the matrix $\mathbf{B}_0$ using the Cholesky decomposition of the covariance matrix of the estimated residuals $\boldsymbol{d}(t)$ along this causal order. However, since in many cases the theory is not sufficient to determine a causal order unambiguously, approaches such as the ones of Sections 5.1 and 5.2 may be required.

the independent disturbances $\boldsymbol{e}(t) = (\mathbf{I} - \mathbf{B}_0)\boldsymbol{d}(t)$, such that the shocked disturbance is not correlated with any other disturbances.

The effect of a unit shock in the disturbances $\boldsymbol{e}(t-\tau)$, $\tau \geq 0$ on the variables $\boldsymbol{v}(t)$ is given by the *impulse response* functions $\boldsymbol{\Psi}_\tau$ (Hamilton, 1994; Lütkepohl, 2005). Identification of these matrices $\boldsymbol{\Psi}_\tau$ requires the matrix $\mathbf{B}_0$, which can be seen from the Wold moving average representation of the VAR model of Equation (5.8) (Hamilton, 1994; Lütkepohl, 2005), as explained below:

$$\boldsymbol{v}(t) = \sum_{\tau=0}^{\infty} \boldsymbol{\Phi}_\tau \boldsymbol{d}(t-\tau) \tag{5.11}$$

$$= \sum_{\tau=0}^{\infty} \boldsymbol{\Phi}_\tau (\mathbf{I} - \mathbf{B}_0)(\mathbf{I} - \mathbf{B}_0)^{-1} \boldsymbol{d}(t-\tau) = \sum_{\tau=0}^{\infty} \boldsymbol{\Psi}_\tau \boldsymbol{e}(t-\tau). \tag{5.12}$$

The matrices $\boldsymbol{\Phi}_\tau$ of Equation (5.11) are the moving average coefficients of the VAR model, which can be transformed to the impulse response functions $\boldsymbol{\Psi}_\tau$ only if we know the matrix $\mathbf{B}_0$ yielding independent disturbances $\boldsymbol{e}(t)$, as can be seen from Equation (5.12).

5.5 Granger Causality

A commonly used concept for causal analysis in time series is *Granger causality* (Granger, 1969). The idea is that a cause x precedes its effect y in time, and hence, x should help in predicting y. More formally, assuming that we are given all the information in the universe up to and including time point t, we predict the value of $y(t+h)$, $h \geq 1$, once using all this information, and once using this information excluding $x(\tau)$ for all $\tau \leq t$. If for some $h \geq 1$ the prediction error (in the least square sense) of $y(t+h)$ is smaller for the former estimate than the latter, the time series x *Granger causes* y (Lütkepohl, 2005). In Example 3.7 (page 27), v_3 is a Granger cause of v_2, and v_2 is a Granger cause of v_1, whereas v_2 is not a Granger cause of v_3.

However, in general we do not observe all the information in the universe, i.e. the system may not be causally sufficient. In this case, Granger causality is only a necessary, but not sufficient condition for a time series x to be a 'real' cause of a time series y. Unobserved confounding time series can introduce spurious correlation between x and y, and hence, x may help in predicting y, even though it is not a cause of y.

Chapter 6

Contributions to the Research Field

After introducing the necessary background on the relevant existing work in Chapters 2 through 5, we are now ready to describe the contributions of the publications of this thesis to the research field. Throughout the Ph.D. studies, linear non-Gaussian models played a prominent role (Articles I, III, IV and V), but also more general models were used (Articles II and VI), extending the work on LiNGAM models to non-parametric settings. An overview of which research question the articles address, as well as the required model assumptions, is given in Figure 6.1.

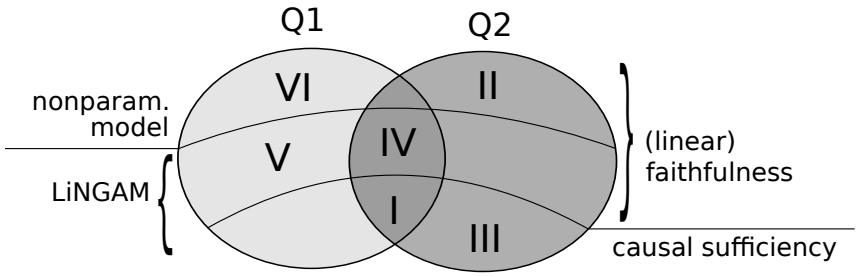


Figure 6.1: Overview of the articles. The two circles indicate the two research questions: Q1 (effect identification) and Q2 (structure learning). The articles in the intersection (partly) address both questions. The upper line separates the articles according to the used model (non-parametric model or LiNGAM model), the lower line according to additional assumptions (faithfulness or linear faithfulness but no causal sufficiency, and causal sufficiency but no faithfulness).

6.1 Structure Learning in Time Series Models

Articles I and II focus on causal inference from time series data. In Article I, we introduce the VAR-LiNGAM method for SVAR identification (Section 5.4) to the econometrics community, and apply it to a micro- and a macroeconomic data set. Article II generalizes the FCI algorithm (Section 5.1.2) to time series data, which allows learning the structure of a time series in the presence of latent variables. Both articles address the problem of structure learning (Q2) from time series data. In addition, the method used in Article I gives estimates of the causal effects (i.e. addresses Q1 as well).

6.1.1 SVAR Identification in Econometrics using LiNGAM

In Article I the SVAR model of Definition 3.9 (page 28) is used, assuming that the model is causally sufficient, and given that the probability distributions of the disturbances are *non-Gaussian* (which can be tested using standard statistical tests). The aim is to demonstrate the characteristics and the potential of the VAR-LiNGAM method to the econometrics community by applying it to two economic data sets.

The first time series data set consists of four variables: employment, sales, research and development (R&D) expenditure, and operating income. The data points are observed annually for manufacturing firms in the US for the years 1972 to 2004, with some years missing for some firms, yielding an unbalanced panel data set. (The data set is discussed in more detail by Coad and Rao (2010).) Under the standard panel assumption that all firms follow the same process, the firms are pooled together in our analysis.

As the process over these four variables is not stationary, we take log-differences yielding a stationary process over the growth rates of the variables, i.e.

$$v_i(t) = \log(\tilde{v}_i(t)) - \log(\tilde{v}_i(t-1)) \tag{6.1}$$

with $\tilde{v}_i(t)$ denoting any of employment, sales, R&D expenditure, and operating income at time point t, and $v_i(t)$ denoting the respective growth rates. Furthermore, statistical tests show that the variables are non-Gaussian, such that the VAR-LiNGAM method can be applied.

For the first step in the SVAR identification, we use a VAR model with $q = 2$ time lags, and estimate the coefficients of Equation (5.8) (page 49) using an estimator based on least absolute deviations, since Dasgupta and Mishra (2004) and Coad and Rao (2010) suggested that these estimators

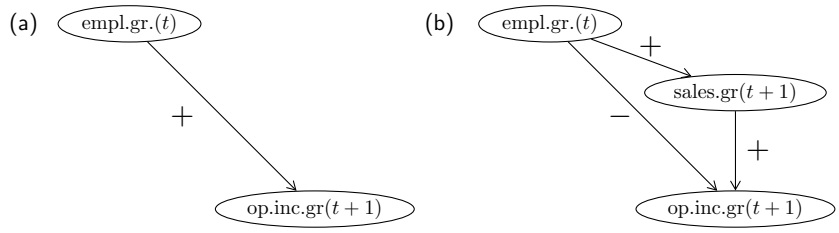


Figure 6.2: Graph over a subprocess of the firm growth data of (a) the VAR estimation, and (b) the SVAR estimation of the model, over the variables employment growth, operating income growth, and sales growth.

are more robust for non-Gaussian variables.[1] Estimating the instantaneous effects of Equation (5.9) is done with the ICA-LiNGAM approach (Section 5.2.1). Finally, the corrected lagged effects are calculated as in Equation (5.10).

Comparing the simple VAR coefficients $\mathbf{A}_1$ and $\mathbf{A}_2$, which disregard the instantaneous effects, to the SVAR coefficients $\mathbf{B}_1$ and $\mathbf{B}_2$, which are corrected for the instantaneous effects, reveals substantial differences in some of the estimates. Consider the subprocess depicted in Figure 6.2. In (a), the coefficient of the VAR estimation for employment growth at time t on operating income growth at time $t + 1$ is positive. However, in (b), the same coefficient turns into a negative effect in the corrected SVAR estimates. In addition, there is an indirect large positive effect from employment growth at time t on operating income growth at time $t + 1$ via sales growth at time $t + 1$ (and some other smaller indirect effects), so that the *total* effect is positive, but the *direct* effect is truly negative. Thus, using the SVAR estimates $\mathbf{B}_0$, $\mathbf{B}_1$, and $\mathbf{B}_2$ direct and indirect effects can be distinguished, giving a better understanding of the underlying system.

The second data set used in Article I consists of six variables: three macroeconomic variables (GDP, GDP deflator, and the Dow-Jones index of spot commodity prices), and three policy variables (borrowed bank reserves, non-borrowed bank reserves, and the federal funds rate), based on the data set of Bernanke and Mihov (1998).[2] The variables are observed

[1] Note that the OLS estimator is only equal to the maximum likelihood estimator in case of Gaussian variables.

[2] In the original data set of Bernanke and Mihov (1998), instead of the borrowed reserves, the total reserves were included. However, an important assumption of the VAR-LiNGAM method is independent shocks (disturbance terms). Since non-borrowed reserves are part of the total reserves, it is likely that a shock to one of them is correlated with a shock to the other. Hence, to make the independence assumption more plausible, we replace total bank reserves with borrowed bank reserves.

for the US, monthly in the period of January 1965 to December 1996. The main objective in this application is to analyze the effect of changes in the monetary policy of the Federal Reserve System, the central banking system of the US, on the macroeconomic variables. The goal is to find the best indicator among the policy variables for the monetary policy of the Fed, requiring the impulse response functions.

The variables form a co-integrated process, and hence, a VECM model is used to estimate the VAR coefficients in the first step of the VAR-LiNGAM procedure, using $q = 7$ time lags. The instantaneous effects matrix $\mathbf{B}_0$ required for the impulse response functions is obtained using ICA-LiNGAM.

As the results of this data set are less intuitive to understand than the ones of the first data set, we restrict ourselves here to state that the impulse response functions reflect economic theories on how the shocks to the policy variables affect the GDP and GDP deflator. Furthermore, analyzing sub-intervals of the time series reveals that the Fed has changed its policy instruments across the years, reflected in the fact that the policy variable corresponding most closely to the (by theory) expected policy shock changes.

In summary, Article I illustrates that the VAR-LiNGAM method may be a suitable approach for analyzing economic time series data. The obtained results from the two applications do not only make intuitive sense, but also reflect established economic theories.

6.1.2 FCI for Time Series Data

The VAR-LiNGAM method used in Article I is designed for linear models with non-Gaussian disturbances and no latent variables, i.e. for the causally sufficient case. In Article II, we present a method for time series data analysis, based on the FCI algorithm (Section 5.1.2), dropping all these assumptions, while however requiring faithfulness. Hence, in some sense Article II can be seen as a non-parametric extension of the VAR-LiNGAM method used in Article I.

The original FCI algorithm is designed for data generated by models such as CBNs and SEMs, not including a time dimension. To generalize this algorithm to time series models, such as the SVAR model, one possible approach is to 'unroll' the underlying graph for a certain 'window' length τ, and incorporate background knowledge given by the assumptions of a time series into the FCI algorithm. For instance, in the graph of Figure 3.5 (page 27), the model is shown in the window from time points $t-3$ to $t+1$. Typically, we consider windows between time points $t - \tau$ to t, $\tau > 0$.

In contrast to Article I, we allow latent variables, i.e. $\boldsymbol{v}(t) = (\boldsymbol{w}(t), \boldsymbol{u}(t))$

with $\boldsymbol{w}(t)$ observed, and $\boldsymbol{u}(t)$ unobserved variables. Note that the *full* set of variable-time point nodes fulfill the causal Markov condition in the completely unrolled graph. However, on a finite window of length τ additional dependencies can occur, even in the case of no latent variables $\boldsymbol{u}(t)$, due to the marginalization over the variables outside the window. For instance, in Figure 3.5, for any window of length τ, $v_2(t-\tau)$ and $v_3(t-\tau)$ are dependent since all conditioning sets $\mathcal{Z}$ for which $v_2(t-\tau) \perp\!\!\!\perp v_3(t-\tau) \,|\, \mathcal{Z}$ holds lie outside the window (for instance $\mathcal{Z} = \{v_3(t-\tau-1)\}$). Furthermore, when marginalizing out latent variables inside the window, dependencies may go back arbitrarily far in time: In the graph of Figure 3.5 with $\boldsymbol{w}(t) = (v_1(t), v_2(t))$ and $\boldsymbol{u}(t) = v_3(t)$, $v_2(t)$ and $v_2(t-k)$ are dependent *for all* k due to the latent variable v_3.

In this sense, FCI is precisely the right tool, since it can handle marginals of DAGs, represented by MAGs, and infer their equivalence class in form of PAGs. We thus apply a modified FCI algorithm (see below) to samples over the observed variables $\boldsymbol{w}(t-\tau), \ldots, \boldsymbol{w}(t)$, which are obtained using a 'sliding window' approach, i.e. for observations over a period of length T we obtain $T-\tau$ samples $(\boldsymbol{w}(1), \ldots, \boldsymbol{w}(\tau+1))$, ..., $(\boldsymbol{w}(T-\tau), \ldots, \boldsymbol{w}(T))$.

To ensure that the distribution over the variables $p(\boldsymbol{v}(t-\tau), \ldots, \boldsymbol{v}(t))$ is well-defined, we assume that the whole process has a strictly positive time invariant probability distribution, as for example is the case for stable VAR models. Furthermore, we assume that $p(\boldsymbol{v}(t-\tau), \ldots, \boldsymbol{v}(t))$ is faithful to the unrolled graph.

Even though FCI was originally designed for data without a time dimension, it is straightforward to incorporate temporal knowledge: As already mentioned by Spirtes et al. (1993), all edges can be oriented 'forward in time' (i.e. adding arrowheads pointing into the nodes in the future), since the cause must precede the effect. This can be implemented in the orientation phase of FCI (see below). In Article II, we additionally incorporate the knowledge of having a time invariant structure, i.e. a variable $v_i(t-t_1)$ is a cause of $v_j(t)$, $t_1 \geq 0$ if and only if the same applies for $v_i(t-t_1-k)$ and $v_j(t-k)$ for all k. This knowledge can be included in the adjacency phase and the orientation phase of FCI, resulting in the tsFCI (time series FCI) algorithm introduced in Article II, and summarized next.

In the adjacency phase, an edge between $w_i(t-t_1)$ and $w_j(t-t_2)$, $0 \leq t_2 \leq t_1 \leq \tau$, is removed if and only if we find a set $\mathcal{Z} \subseteq \{\boldsymbol{w}(t-\tau), \boldsymbol{w}(t-\tau+1), \ldots, \boldsymbol{w}(t-t_2)\} \setminus \{w_i(t-t_1), w_j(t-t_2)\}$ such that

$$w_i(t-t_1) \perp\!\!\!\perp w_j(t-t_2) \,|\, \mathcal{Z}. \tag{6.2}$$

Note that the set $\mathcal{Z}$ only needs to contain variables up to and including time point $t-t_2$ (or, if it is assumed that there are no instantaneous effects,

only up to and including time point $t - t_2 - 1$). If such a set $\mathcal{Z}$ is found, by the time invariant structure, we also know that there exists a set $\mathcal{Z}_k$ such that

$$w_i(t - t_1 - k) \perp\!\!\!\perp w_j(t - t_2 - k) \,|\, \mathcal{Z}_k, \tag{6.3}$$

with $\mathcal{Z}_k$ containing the 'same' variables as $\mathcal{Z}$ but k time steps earlier. Thus, for all k for which $\mathcal{Z}_k$ lies inside the window the edge between $w_i(t-t_1-k)$ and $w_j(t-t_2-k)$ can be removed (which follows from Lemma 1 of Article II).

Example 6.1 (Adjacency Phase in tsFCI)**.** *Taking the time series of Figure 3.5 on a window of length $\tau = 3$, we find that*

$$v_1(t) \perp\!\!\!\perp v_1(t-1) \,|\{v_2(t-1), v_1(t-2)\}$$

and hence we can remove the edge between $v_1(t)$ and $v_1(t-1)$. By the invariant time structure we thus know that in the unrolled graph the following also holds:

$$v_1(t-k) \perp\!\!\!\perp v_1(t-1-k) \,|\{v_2(t-1-k), v_1(t-2-k)\}.$$

However, only for $k = 1$ the conditioning set $\mathcal{Z}_1 = \{v_2(t-2), v_1(t-3)\}$ lies inside the window, and hence we can only additionally remove the edge between $v_1(t-1)$ and $v_1(t-2)$, but not between $v_1(t-2)$ and $v_1(t-3)$ (since in the considered window these two variables are not independent).

In the orientation phase, we start by orienting all edges forward in time, i.e. if there is an edge between $w_i(t - t_1)$ and $w_j(t - t_2)$ with $t_1 > t_2$, then we orient it as $w_i(t - t_1) \circ\!\!\rightarrow w_j(t - t_2)$. This simply follows from the fact that the cause precedes the effect, and in a PAG (the output of FCI) the arrowhead means that $w_j(t - t_2)$ is not a cause of $w_i(t - t_1)$. Furthermore, when applying the complete set of orientation rules of FCI (Zhang, 2008), whenever an endpoint of an edge between $w_i(t-t_1)$ and $w_j(t-t_2)$ is oriented, we orient the endpoint in the same way for the edge between $w_i(t - t_1 - k)$ and $w_j(t - t_2 - k)$, for all k (by Lemma 2 of Article II).

The advantage of using the tsFCI algorithm over applying the original FCI algorithm to a data set as explained above (including background knowledge about a temporal ordering) is twofold. In both the adjacency and the orientation phase tsFCI reduces the computation time using the time invariant structure by (i) restricting the number of independence tests carried out and (ii) reducing the number of orientation rules applied. Secondly, the output PAG of tsFCI always represents a time invariant structure, whereas for FCI this is not necessarily true due to finite sample effects.

Compared to SVAR identification (Section 5.4) as well as Granger causality (Section 5.5), the clear advantage of tsFCI is its ability to handle latent variables.

6.2 Structure Learning in Extended LiNGAM Models

Articles III through VI address both effect identification (Q1) and structure learning (Q2) in the case of static data. Most methods continue along the line of Article I, using linear SEMs with non-Gaussian error terms. Although designed for static data, using the 'sliding window' approach as explained for tsFCI in Article II, these methods may also be applied to time series data. In this section we discuss the work towards answering Q2, while the work addressing Q1 will be presented in the next section.

6.2.1 LiNGAM for Multidimensional Variables

In Article III, we introduce a set of methods to estimate the causal order among multidimensional variables $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, assuming that the data are generated by a similar model to the one in GroupLiNGAM (Section 5.2.5). The only difference in our model is that the matrix $\mathbf{B}$ is assumed to be *strictly* lower block triangular (as opposed to lower block triangular). However, any model following the definition of the GroupLiNGAM model can be transformed into an observationally equivalent model following the model definition of Article III. An important difference to the work by Kawahara et al. (2010) is that we assume to know a priori the partition of the variables in $\boldsymbol{v}$ into the vectors (also referred to as groups) $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.

The general algorithm follows the idea of DirectLiNGAM (Section 5.2.2), i.e. first finding an exogenous group, then regressing out the effect of the exogenous group on all other groups, and repeating the process on the resulting residuals. We introduce three alternatives to find an exogenous group.

In the first approach of finding an exogenous group, we generalize Theorem 5.2 (page 45) to multidimensional variables, stating that a group of variables $\boldsymbol{v}_j$ is exogenous if and only if it is independent of all the residuals $\boldsymbol{r}_i^{(j)}$ resulting from OLS regressions of $\boldsymbol{v}_i$ on $\boldsymbol{v}_j$ (Lemma 1 of Article III). As this result is based on independencies, it is essential that the disturbance terms $\boldsymbol{e}_i$ are non-Gaussian.

The second approach is based on the pairwise measure explained in Section 5.2.3, which also requires non-Gaussian disturbances. The naïve approach of utilizing this measure would be to test for each pair (v_i, v_j) of *scalar* variables, with v_i in group $\boldsymbol{v}_i$ and v_j in group $\boldsymbol{v}_j$, $i \neq j$, whether v_i is a cause of v_j. If the error terms within each group were independent then for any variable v_j of an exogenous group $\boldsymbol{v}_j$ the pairwise measure would infer

the correct causal direction between v_i and v_j. However, since correlated error terms within each group are allowed, it is necessary to transform the variables of the pair (v_i, v_j) appropriately to meet the model assumptions of the pairwise measure. After this transformation, we obtain a measure for each pair (v_i, v_j), indicating which variable is exogenous among the two. By combining these measures appropriately it is then possible to find an exogenous group.

Lastly, in the third approach we utilize the Trace Method (Section 5.3) to find an exogenous group. This method is designed to find a causal order among two multidimensional variables, and can be generalized to find an exogenous group among several groups by applying it to each pair $(\boldsymbol{v}_i, \boldsymbol{v}_j)$, $i \neq j$. For an exogenous group $\boldsymbol{v}_j$, the model $\boldsymbol{v}_j \to \boldsymbol{v}_i$ holds for all i (by marginalizing out intermediate groups). Appropriately combining these measures yields an exogeneity measure for each group. Note that this approach does not require non-Gaussian disturbances, so can also be applied to Gaussian data.

After finding an exogenous group, a direct generalization of Theorem 5.3 (page 46) shows that after regressing out this group from all other variables, the resulting (multidimensional) residuals follow again the same model (Lemma 2 of Article III).

Comparing the three approaches on simulated data shows that the method based on the pairwise measure performs very well, also for small sample sizes, whereas the generalization of DirectLiNGAM requires more data points to reach the same performance. The approach based on the trace method always makes more mistakes than the former two methods. The simulations also show that for large models and small data sets it may even be advantageous to use the pairwise measure in the naïve way explained above. All three variants outperform the simple approach of replacing each group with an aggregate, such as the mean over all variables, and then applying methods for scalar variables, like DirectLiNGAM.

The methods introduced in Article III are closely related to the work by Kawahara et al. (2010), both using a DirectLiNGAM-style approach for the overall algorithm. However, there are some clear distinctions and advances in the methods of Article III. The major difference of our approach to GroupLiNGAM is, as mentioned above, that in our method we assume to know how the vector $\boldsymbol{v}$ is partitioned into the variables $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, yielding efficient algorithms to infer the causal order among these variables. In GroupLiNGAM, on the other hand, Kawahara et al. (2010) aimed at learning this partition, which results in computationally and statistically more challenging algorithms. Furthermore, we give three strategies to find

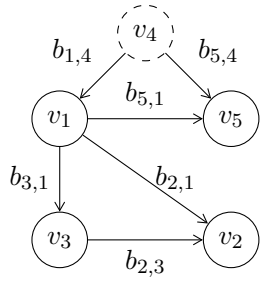


Figure 6.3: An lvLiNGAM model with $\mathcal{U} = \{v_4\}$ and $\mathcal{W} = \{v_1, v_2, v_3, v_5\}$.

an exogenous group, whereas Kawahara et al. (2010) only considered the linear non-Gaussian case.

6.2.2 Pairwise Causal Relationships in lvLiNGAM

In Article IV we assume that we are given data over the observed variables $\mathcal{W}$, generated by a faithful lvLiNGAM model (Section 5.2.4) over variables $\mathcal{V} = \mathcal{W} \cup \mathcal{U}$. As already mentioned, it is in general impossible to infer the whole lvLiNGAM model (since generally it is not identifiable), and estimation of models, up to equivalence, is computationally very challenging. Thus, the goal here is to develop more efficient techniques to learn (identifiable) *parts* of the underlying model.

The basic idea is that there might be subsets $\mathcal{Z} \subseteq \mathcal{W}$ for which a LiNGAM model (without latent variables) fits. More precisely, as shown in Lemma 1 of Article IV, the variables in $\mathcal{Z}$ follow a LiNGAM model if and only if every confounder of any two variables $z_i \in \mathcal{Z}$ and $z_j \in \mathcal{Z}$ lies in $\mathcal{Z}$ (i.e. the set $\mathcal{Z}$ is causally sufficient). In Article IV we call such sets *unconfounded*.

Example 6.2 (Confounded and Unconfounded Sets)**.** *For the graph in Figure 6.3, the sets* $\{v_1, v_5\}$, $\{v_2, v_5\}$, $\{v_1, v_2, v_3, v_5\}$, *for instance, are confounded by the latent variable* v_4, *and the set* $\{v_2, v_3\}$ *is confounded by the observed variable* v_1. *When estimating a LiNGAM model for any of these sets, the resulting residuals are* not *independent of each other (no matter which causal order is learned). For instance, for* v_1 *and* v_5 *we may obtain that* $v_1 = r_1$, *and* $v_5 = c_{5,1}\, v_1 + r_5$. *Expressing* r_1 *and* r_5 *in terms of the disturbance terms* e_1, e_4, e_5 *of variables* v_1, v_4, v_5 *yields*

$$r_1 = b_{1,4}\, e_4 + e_1$$
$$r_5 = v_5 - c_{5,1}\, v_1 = (b_{5,1}\, b_{1,4} + b_{5,4} - c_{5,1}\, b_{1,4})\, e_4 + (b_{5,1} - c_{5,1})\, e_1 + e_5.$$

By the non-Gaussianity assumption of the disturbances and the Darmois-Skitovitch Theorem (Section 2.2.2), for independence between r_1 and r_5 we require $b_{1,4} = 0$ (yielding $b_{5,1} = c_{5,1}$), or $b_{5,1} = c_{5,1}$ and $b_{5,1}\, b_{1,4} + b_{5,4} - c_{5,1}\, b_{1,4} = 0$, which is only possible when $b_{5,4} = 0$. A similar calculation shows that if the LiNGAM model was estimated in the reverse direction, the residuals would also be dependent.

Thus, by testing independence between the residuals, it is possible to detect confounding in the case of non-Gaussian *variables. Note that for* Gaussian *variables the residuals r_1 and r_5 are always independent since these two residuals are by construction of the OLS estimator (and also by definition of ICA) uncorrelated.*

In contrast, the sets $\{v_1, v_3\}$, $\{v_1, v_2\}$, and $\{v_1, v_2, v_3\}$ are unconfounded (and no other sets of two or more observed variables are), and a LiNGAM model fits.

However, going through all subsets of $\mathcal{W}$ to test whether they follow a LiNGAM model is in general not feasible, as their number grows exponentially in the number of variables in $\mathcal{W}$. Fortunately, as the previous example hints at, unconfounded sets with three or more variables have unconfounded subsets. Thus, an incremental search approach could be used to find all *maximally* unconfounded sets, i.e. when adding any variable to the set its unconfoundedness is destroyed. However, even the number of these sets grows exponentially in the worst-case graphs.

Thus, to obtain a computationally efficient algorithm, in Article IV we address the problem of determining all *pairwise* causal relations, for which there exists an unconfounded set $\mathcal{Z}$ containing this pair. Such pairs are called *unconfounded with respect to $\mathcal{Z}$*, and for each such pair the introduced algorithm returns which variable is the cause and which the effect, and the total effect of the former on the latter. The resulting method is termed the 'pairwise lvLiNGAM' algorithm.

Example 6.3 (Pairwise lvLiNGAM)**.** *Given a data set generated from the model of Figure 6.3 with observed variables $\mathcal{W} = \{v_1, v_2, v_3, v_5\}$, the proposed algorithm iteratively searches for all pairs which are part of an unconfounded set.*

First, it searches for pairs (v_i, v_j), $i \neq j$, $v_i \in \mathcal{W}$ and $v_j \in \mathcal{W}$, which are unconfounded by estimating a LiNGAM model over (v_i, v_j) and testing its fit. As explained in Example 6.2, the only pairs for which the LiNGAM model fits are (v_1, v_2) and (v_1, v_3), yielding the causal directions from v_1 to v_2, and from v_1 to v_3, and total causal effects $b_{2,1} + b_{2,3}\, b_{3,1}$, and $b_{3,1}$, respectively.

Next, for any pair (v_i, v_j) which was not found to be unconfounded yet, the algorithm tries to build up a set $\mathcal{Z} \cup \{v_i, v_j\}$ such that the pair is unconfounded with regard to this set. The set $\mathcal{Z}$ includes all those variables v_k for which (v_k, v_i) and (v_k, v_j) were found to be unconfounded in a previous iteration of the algorithm (i.e. any potential confounders of v_i and v_j). To test whether $\mathcal{Z} \cup \{v_i, v_j\}$ is unconfounded, a LiNGAM model is estimated and its fit is evaluated. In the example, for the pair (v_2, v_3) we obtain the set $\mathcal{Z} = \{v_1\}$, and a LiNGAM model fits to $\{v_1, v_2, v_3\}$, as mentioned in Example 6.2. Furthermore, this LiNGAM model reveals that v_3 is the cause of v_2 with total causal effect $b_{2,3}$.

The algorithm iterates this step until no more new pairs are found to be unconfounded with regard to some set.

As the main result of Article IV, we prove in Theorem 1 that, in the large sample limit, the pairwise lvLiNGAM algorithm is sound (i.e. whenever a pair is judged to be unconfounded with regard to a set, this is correct, and the estimate of the total causal effect converges to the true total effect) and complete (i.e. any pair which is part of an unconfounded subset is discovered). The completeness in particular means that *any* pairwise causal effect which is detectable by fitting a LiNGAM model to a subset of variables and testing its fit is detected.

Fitting a LiNGAM model as done in Example 6.3 to find unconfounded sets is computationally inefficient. In Lemma 3 of Article IV, a more efficient procedure is suggested to find such sets: For a set $\mathcal{Z}$ and a pair (v_i, v_j) as in Example 6.3, we estimate the following two regression models

$$v_j = c_{j,i} v_i + \boldsymbol{c}_j^T \boldsymbol{z} + r_j \tag{6.4}$$

$$v_i = c_{i,j} v_j + \boldsymbol{c}_i^T \boldsymbol{z} + r_i \tag{6.5}$$

and conclude as follows:

(i) If $r_j \perp\!\!\!\perp v_i$ and $r_i \not\perp\!\!\!\perp v_j$, then v_i is a cause of v_j with total effect $c_{j,i}$.

(ii) If $r_j \not\perp\!\!\!\perp v_i$ and $r_i \perp\!\!\!\perp v_j$, then v_j is a cause of v_i with total effect $c_{i,j}$.

(iii) If $r_j \perp\!\!\!\perp v_i$ and $r_i \perp\!\!\!\perp v_j$, and $c_{j,i} = c_{i,j} = 0$, then $v_i \perp\!\!\!\perp v_j \mid \mathcal{Z}$.

(iv) If none of the above holds, then (v_i, v_j) is not unconfounded with respect to $\mathcal{Z} \cup \{v_i, v_j\}$.

In simulations we evaluate the algorithm using two independence tests (HSIC, and non-linear correlations, see Section 2.2.2), and compare the results to ICA-LiNGAM (which does not account for latent variables). For

pairwise lvLiNGAM, the larger the sample size grows, the closer the estimated total effects are to the true effects. On small sample sizes, the method performs better with HSIC as the independence test, whereas on larger sample sizes using the test based on non-linear correlations seems beneficial. ICA-LiNGAM, on the other hand, keeps making a significant number of mistakes even for large sample sizes.

In some sense, the pairwise lvLiNGAM algorithm straddles both research questions, structure learning (Q2) and effect identification (Q1). If a LiNGAM model was actually estimated to find unconfounded pairs with regard to some set $\mathcal{Z} \subset \mathcal{W}$, this would reveal parts of the underlying structure if this model fit. However, in the computationally more efficient approach, only two simple regression models need to be estimated to test for unconfoundedness of a pair with regard to a given set, and only the total effect between the variables of the corresponding pair is returned.

The approach of estimating regressions and testing for independence among the residual and the regressor, taken in Article IV, partly motivated the work in Article V, discussed in the next section.

6.3 Effect Identification under the Partial Ordering Assumption

In Articles V and VI we address the problem of effect identification (Q1) when the generating DAG $\mathcal{G}$ of the causal model is *not* known and the set of observed variables $\mathcal{W} \cup \{x, y\}$ is *not* assumed to be causally sufficient. For both articles we require the partial ordering assumption $\mathcal{W} \prec x \prec y$ (Definition 4.3, page 34), and aim at inferring the direct causal effect of x on y (which is equal to the total causal effect under the partial ordering assumption). In Article V, we further assume that the data generating process follows an lvLiNGAM model (Section 5.2.4) and that the resulting probability distribution is linearly faithful to $\mathcal{G}$, whereas in Article VI we do not have any restrictions on the model (i.e. it can be any CBN or SEM) except that we need to assume faithfulness.

6.3.1 Consistency Test for Causal Effects in lvLiNGAM

Under the just mentioned assumptions, in Article V we introduce a statistical test for lvLiNGAM models to infer whether adjusting for a given set $\mathcal{Z} \subseteq \mathcal{W}$ yields a consistent estimator of the causal effect of x on y. The

test for the given set $\mathcal{Z}$ is based on two simple OLS regressions:

$$x = \boldsymbol{c}_x^T \boldsymbol{z} + r_x, \quad \text{and} \tag{6.6}$$

$$y = c_{y,x} x + \boldsymbol{c}_y^T \boldsymbol{z} + r_y. \tag{6.7}$$

If the residual r_x is Gaussian (which can be tested using standard tests), the test terminates without conclusion. For *non-Gaussian* r_x, we perform a statistical test of dependence between r_x and r_y, and conclude as follows (based on Theorem 1 of Article V):

(i) If independence is rejected, the estimated effect $c_{y,x}$ of x on y is inferred to be inconsistent.
(ii) If independence is not rejected, the effect is inferred to be consistent.

The main result of Article V is that, under the given assumptions, the just described test will, in the large sample limit, correctly identify sets $\mathcal{Z}$ which yield a consistent estimator of the causal effect (Theorem 1 of Article V). The reason why non-Gaussianity is required becomes clear when looking at the residuals r_x and r_y: By construction of the OLS estimator $cov(r_x, r_y) = 0$, and hence r_x and r_y are always independent for Gaussian variables.

Example 6.4 (Statistical Test for Consistency in lvLiNGAM models)**.** *We demonstrate the method using lvLiNGAM models over the DAGs in Figures 4.2 (a) and (b) (page 35). We first consider the model in (a), with linear equations*

$$\begin{pmatrix} w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ b_{x,w} & 0 & 0 \\ b_{y,w} & b_{y,x} & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \end{pmatrix} + \begin{pmatrix} e_w \\ e_x \\ e_y \end{pmatrix}. \tag{6.8}$$

Using $\mathcal{Z} = \emptyset$ in Equations (6.6) *and* (6.7)*, the resulting residuals r_x and r_y are dependent, which can be seen when expressing these residuals in terms of the disturbances $\boldsymbol{e} = (e_w, e_x, e_y)^T$:*

$$\begin{aligned} r_x &= x = (b_{x,w}\,,\ \mathbf{1}\,,\ 0)\,\boldsymbol{e} \\ r_y &= y - c_{y,x}x = (b_{y,x}b_{x,w} + b_{y,w}\,,\ b_{y,x}\,,\ 1)\,\boldsymbol{e} - c_{y,x}(b_{x,w}\,,\ 1\,,\ 0)\,\boldsymbol{e} \\ &= ((b_{y,x} - c_{y,x})b_{x,w} + b_{y,w}\,,\ \boldsymbol{b_{y,x} - c_{y,x}}\,,\ 1)\,\boldsymbol{e}. \end{aligned}$$

Using the formula for the OLS estimator (Equation (2.8)*, page 13) or the back-door criterion (Definition 4.2, page 31), one can show that $c_{y,x} \neq b_{y,x}$ in the large sample limit,*[3] *and thus the coefficients of e_x in both representations r_x and r_y, marked in bold, are non-zero. By the non-Gaussianity of*

[3] To be more precise, $c_{y,x}$ does not converge in probability to $b_{y,x}$.

e_x and the Darmois-Skitovitch Theorem (Section 2.2.2) it then follows that r_x and r_y are dependent*, allowing us to detect the* inconsistent *estimator.*

On the other hand, using $\mathcal{Z} = \{w\}$ in Equations (6.6) *and* (6.7) *yields that $r_x = e_x$ and $r_y = e_y$ (using a similar calculation as before, and that $c_{y,x} = b_{y,x}$ by the OLS estimator). These residuals are by assumption* independent*, and hence we can detect the* consistent *estimator.*

In (b) the situation is reversed: $\mathcal{Z} = \emptyset$ yields independent *residuals r_x and r_y, and a* consistent *estimator $c_{y,x}$ of the causal effect $b_{y,x}$, whereas for $\mathcal{Z} = \{w\}$ the residuals are* dependent *and the estimator is* inconsistent*. The calculations are a bit more cumbersome, and can be found in the Supplementary Material of Article V.*

When searching in small models for an admissible set $\mathcal{Z}$ among all possible subsets of the observed covariates $\mathcal{W}$, it is possible to perform a brute force search by applying the statistical test for consistency to all subsets. However, for larger models this is not feasible as the number of subsets grows exponentially in the number of covariates. Hence, we suggest in Article V simple forward and backward selection procedures, which are quadratic in the number of covariates. In essence, for the forward selection we first apply the statistical test to the empty set. Then, we apply the test to all subsets $\mathcal{Z}$ with one variable, and pick the 'best' subset, i.e. the one yielding the most independent residuals. We then augment this best subset by one variable, and find among those sets again the 'best' one, and repeat this process until there are no more variables to add. The backward elimination starts with the full set of variables as adjustment set, and removes in a similar fashion variables from this set.

In simulations on rather small models (up to ten covariates) these approaches have performed equally well as the brute force procedure. However, with the forward and backward selection there is no guarantee that an admissible set is found, in the large sample limit, even if one exists. We also compare the average error in the estimate given by our procedure (when the estimate was deemed consistent) and given by the simple adjustment criteria of including all or none of the covariates, as well as using ICA-LiNGAM. While for our procedure the error decreases substantially for growing sample size, the control methods keep making, on average, much larger errors. Performance differences due to the used independence test (HSIC or non-linear correlations) are negligible.

6.3.2 Non-parametric Approach

Article VI can be seen as a non-parametric version of Article V, and as an extension of the work by Spirtes and Cooper (1999), and Chen et al.

(2007) (Section 4.3.2) allowing for non-empty admissible sets. Under the stated assumptions, Article VI gives conditions solely based on dependencies and independencies among the observed variables to determine whether an effect is identifiable.

The aim is to reach one of the following decisions:

`'±'` The causal effect is non-zero and can be estimated using a found admissible set $\mathcal{Z}$ by back-door adjustment (Theorem 4.1, page 31).

`'0'` The causal effect is zero.

`'?'` We do not know, i.e. we cannot infer `'±'` or `'0'`.

Note that when inferring `'±'` or `'0'` the causal effect of x on y is identified. Towards this end, we introduce two simple rules:

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and

(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

then infer `'±'` and give $\mathcal{Z}$ as an admissible set.

R2: If there exists a set $\mathcal{Z} \subseteq \mathcal{W}$ such that

(i) $x \perp\!\!\!\perp y \mid \mathcal{Z}$,

or, if there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(ii) $w \not\perp\!\!\!\perp x \mid \mathcal{Z}$, and

(iii) $w \perp\!\!\!\perp y \mid \mathcal{Z}$,

then infer `'0'`.

If neither R1 nor R2 applies, infer `'?'`.

The idea of R1 is as follows: Condition (i) ensures that there exists at least one active path from w to y, given $\mathcal{Z}$. By condition (ii) these paths must pass through x, since including x in the conditioning set blocks all these paths. This implies that there is at least one active path π from w to x given $\mathcal{Z}$ (pointing into x by the partial ordering assumption). If there existed an active back-door path from x to y, condition (ii) could not hold, since concatenating this back-door path with the active path π at x would yield an active path from w to y given $\mathcal{Z} \cup \{x\}$, with a collider at x. An example for which R1 applies is given by the graph of Figure 4.1 (a) (page 32) with $w = w_1$ and $\mathcal{Z} = \{w_2, w_3\}$.

Rule R2 consists of two parts: First, if condition (i) holds, then, by the faithfulness assumption, there is no edge between x and y in the underlying DAG and hence, the causal effect is 0. Secondly, condition (ii) and (iii) together also ensure that there is no edge between x and y: By condition

(ii) there exists at least one active path π from w to x given $\mathcal{Z}$. If there existed an edge from x to y, appending this edge to the path π would yield an active path from w to y given $\mathcal{Z}$ such that condition (iii) could not hold. Note that the second part of R2 may allow us to detect a zero effect of x on y even in the case of latent confounding.

The main result of Article VI is that these two simple rules are, in the large sample limit, both *sound* (i.e. whenever we make a decision, it is correct) and *complete* (i.e. whenever we infer `'?'`, it is impossible to reliably infer `'±'` or `'0'` based on dependencies and independencies alone), as stated in Theorems 2 and 3 of Article VI.

The rules are also related to the FCI algorithm (Section 5.1.2). In fact, when incorporating the background knowledge of the partial ordering, the PAG output by FCI over the observed variables $\mathcal{W} \cup \{x, y\}$ may be utilized to reach the same decisions as with our rules. However, for FCI with background knowledge it is not known whether it is complete, whereas our rules are.

On finite sample data, combining the (possibly conflicting) results when applying the rules with different pairs $(w, \mathcal{Z})$ or sets $\mathcal{Z}$ is done using an ad-hoc procedure based on a Bayes classifier. In simulations, the novel rules and the approach based on the FCI algorithm clearly outperform the simple adjustment criteria presented in Section 4.3.1, as the former two procedures may output `'?'` if no admissible set exists, or if there is not enough evidence to make decisions `'±'` or `'0'`. The simple adjustment criteria, on the other hand, always output an estimate.

Neither of the two approaches of Articles V and VI subsumes the other. While the two simple rules of Article VI can be applied to any kind of model (not only to lvLiNGAM models), the statistical test of Article V may be able to reach conclusions for models, under the given assumptions, in which the rules of Article VI cannot yield a decision: For the two models of Example 6.4, the statistical test of Article V successfully identifies whether w should be adjusted for. However, the two models imply the same independencies over the observed variables (there are none), and hence, the two models cannot be distinguished based on the rules of Article VI.

Chapter 7

Conclusions

In this thesis we provided novel methods addressing two important problems in the field of causal discovery: causal effect identification (research question Q1) and structure learning (research question Q2). After describing the existing work related to Articles I to VI in Chapters 2 through 5, we discussed the contributions of these articles in Chapter 6. Common themes to the articles were the LiNGAM model and the handling of latent variables. All introduced methods were developed for passive observational data.

The main contribution of this thesis is twofold: In Articles I, III, IV and V we used the LiNGAM model and extended it in various directions (partly including latent variables). These articles provide powerful tools addressing Q1 and Q2 in situations where the data are linear and non-Gaussian. If these assumptions are not met, Articles II and VI present solutions to some of the addressed problems in the nonparametric setting.

In particular, in Article I we discussed the application of an SVAR identification method based on LiNGAM requiring the causal sufficiency assumption. To overcome this limitation, Article II introduced the tsFCI algorithm, which does not rely on the parametric assumptions of a LiNGAM model nor on causal sufficiency. Article III extended the LiNGAM model to multidimensional variables, and introduced a bundle of methods to learn a causal order among these variables. In Article IV, we then presented a complete algorithm to identify pairwise total causal effects in lvLiNGAM models. Finally, in Article V we aimed at identifying one specific direct causal effect in lvLiNGAM models, given the partial ordering assumption; article VI addressed the same problem in the non-parametric setting.

While some gaps in the literature were filled by the contributions of Articles I to VI, there are many open questions. All methods were based on models over *acyclic* causal structures, which in some applications may

not be suitable. It would thus be interesting to develop and apply methods which allow for cyclic connections among the variables.

For the problem of SVAR identification, for instance, instead of using the LiNGAM method one could use the LiNG algorithm (Linear non-Gaussian, Lacerda et al., 2008), a generalization of the LiNGAM model to cyclic models. Similarly, the approach using the PC algorithm for SVAR identification could be modified by replacing the PC algorithm by the CCD method (Cyclic Causal Discovery, Richardson, 1996; Richardson and Spirtes, 1999), which is a constraint based method similar to PC for cyclic models.

Additionally, in Articles V and VI cyclic connections among the observed covariates $\mathcal{W}$ could be allowed, even when keeping the partial ordering assumption, i.e. the assumption that the covariates $\mathcal{W}$ precede the treatment x, which precedes the outcome y. On the other hand, relaxing the partial ordering assumption in Articles V and VI would also be an interesting topic for future research. One such possible scenario is to allow (some of) the observed covariates to lie between x and y. This could yield additional possibilities to identify a (direct or indirect) causal effect by, for instance, utilizing front-door adjustment (Section 4.2.2).

In terms of algorithms, in Article IV it would be interesting to combine the introduced method with other approaches, such as the ones by Tashiro et al. (2012), or Kawahara et al. (2010), as these methods may learn different parts of the model, for example identify a sink, and the combination could help to better understand the underlying model. Furthermore, replacing the forward and backward search procedures of Article V by more sophisticated approaches could be investigated, as the current methods, although working well in simulations, are only heuristics.

Finally, as most of the presented work has focused on theoretical issues of model identifiability, much work remains in terms of investigating the performance of the presented methods in various real-world applications. For instance, it would be interesting to apply the method introduced in Article III to functional magnetic resonance imaging (fMRI) data to analyze functional connectivity in the brain, or the algorithms of Articles V and VI to estimate causal effects in various problems in epidemiology or economics. In such applications, it would be important to evaluate to what extent model violations may affect the result, as well as to further investigate statistical issues due to small sample sizes and potentially large numbers of variables.

References

Ali, R. A., Richardson, T. S., and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837.

Bernanke, B. S. and Mihov, I. (1998). Measuring monetary policy. *The Quarterly Journal of Economics*, 113(3):869–902.

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

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

Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8(R219).

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

Chu, T. and Glymour, C. (2008). Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9:967–991.

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

Coad, A. and Rao, R. (2010). Firm growth and R&D expenditure. *Economics of Innovation and New Technology*, 19(2):127–145.

Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly structured stochastic systems*, pages 115–137. University Press, Oxford.

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21:2–8.

Dasgupta, M. and Mishra, S. K. (2004). Least absolute deviation estimation of linear econometric models: A literature review. *MPRA Paper*, 1781.

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

Demiralp, S. and Hoover, K. D. (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65(s1):745–767.

Druzdzel, M. J. and Simon, H. A. (1993). Causality in bayesian belief networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 3–11. Morgan Kaufmann.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Geiger, D. and Pearl, J. (1988). On the logic of causal models. In *Proceedings of the Fourth Conference on Uncertainty in Artificial Intelligence (UAI-88)*, pages 136–147. AUAI Press.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534.

Granger, C. W. J. (1969). Investigating causal relationships by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

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 (NIPS*2007)*, pages 585–592.

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

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition.

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.

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 (NIPS*2008)*, pages 689–696.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 217–224. AUAI Press.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI-12)*, pages 387–396. AUAI Press.

Hyvärinen, A. (2010). Pairwise measures of causal direction in linear non-Gaussian acyclic models. In *Proceedings of 2nd Asian Conference on Machine Learning (ACML2010)*. Journal of Machine Learning Research Workshop and Conference Proceedings 13:1-16.

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

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731.

Janzing, D., Hoyer, P. O., and Schölkopf, B. (2010). Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 479–486. Omnipress.

Janzing, D. and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194.

Kawahara, Y., Bollen, K., Shimizu, S., and Washio, T. (2010). GroupLiNGAM: Linear non-Gaussian acyclic models for sets of variables. *arXiv*, 1006.5041.

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*. The MIT Press.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 366–374. AUAI Press.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed markov fields. *Networks*, 20(5):491–505.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410. Morgan Kaufmann.

Meek, C. (1997). *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University.

Moneta, A. (2003). Graphical models for structural vector autoregressions. *LEM Working Paper 07, Sant'Anna School of Advanced Studies, Pisa, Italy*.

Moneta, A. and Spirtes, P. (2006). Graphical models for the identification of causal structures in multivariate time series. In *Proceedings of the 2006 Joint Conference on Information Sciences (JCIS 2006)*. Atlantis Press.

Mooij, J., 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 (ICML-09)*, pages 745–752, Montreal. Omnipress.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51. English translation of excerpts (1990) by D. Dabrowska and T. Speed, in Statistical Science, 5:465-472.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (1993a). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (1993b). Mediating instrumental variables. Technical Report R-210, University of California, Los Angeles, CA.

Pearl, J. (1994). A probabilistic calculus of actions. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pages 454–462. Morgan Kaufmann.

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

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. (2nd edition 2009).

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2011). Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 589–598. AUAI Press.

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

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 454–461. Morgan Kaufmann.

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

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

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation, and Discovery*, pages 349–405. AAAI Press/The MIT Press.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

Rosenbaum, P. R. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

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.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 437–444. AUAI Press.

Shpitser, I., Richardson, T. S., and Robins, J. M. (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 661–670. AUAI Press.

Shpitser, I., VanderWeele, T., and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536. AUAI Press.

Silander, T. and Myllymäki, P. (2006). A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452. AUAI Press.

Skitovitch, W. P. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217–219.

Spirtes, P. (2000). The limits of causal discovery from observational data. *American Economic Association, Boston, MA*.

Spirtes, P. and Cooper, G. F. (1999). An experiment in causal discovery using a pneumonia database. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics (AISTATS 1999)*. Morgan Kaufmann.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Spirtes, P., Glymour, C., and Scheines, R. (1990). Causality from probability. In McKee, G., editor, *Evolving Knowledge in Natural and Artificial Intelligence*, pages 181–199. London: Pitman.

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

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In Glymour, C. and Cooper, G. F., editors, *Computation, Causation, and Discovery*, pages 211–252. AAAI Press/The MIT Press.

Spirtes, P. and Richardson, T. (1997). A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (AISTATS 1997)*.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research*, 27(2):182–225.

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*, pages 491–498. Springer Berlin Heidelberg.

Tian, J. (2004). Identifying conditional causal effects. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 561–568. AUAI Press.

Tian, J., Paz, A., and Pearl, J. (1998). Finding minimal d-separators. Technical Report R-254, Computer Science Department, University of California, Los Angeles, CA.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, pages 567–573. The AAAI Press.

Tillman, R. E. (2009). Structure learning with independent non-identically distributed data. In *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, pages 1041–1048, Montreal. Omnipress.

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 Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. Journal of Machine Learning Research Workshop and Conference Proceedings 15: 3-15.

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.

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

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons, 3rd edition.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Conference on Uncertainty in Artificial Intelligence (UAI-88)*, pages 352–359. AUAI Press.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 255–270. AUAI Press.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):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. (2010). Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment (NIPS 2008 Workshop)*. Journal of Machine Learning Research Workshop and Conference Proceedings 6:157-164.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 804–813. AUAI Press.

Zhao, H., Zheng, Z., and Liu, B. (2005). On the markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562.