Diss. ETH No. 20756

Restricted Structural Equation Models for Causal Inference

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
JONAS MARTIN PETERS

Dipl.-Math. University of Heidelberg
born May 28, 1984
citizen of Germany

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Bernhard Schölkopf, co-examiner
PD Dr. Dominik Janzing, co-examiner

2012

To my family

Acknowledgements

I thank my mathematics teachers Irmgard Tieben-Kluge and Axel Schönfeld and my mathematics professors Prof. Dr. Matthias Kreck, Prof. Dr. Michael Leinert, Prof. Dr. Rainer Dahlhaus as well as Prof. Dr. Markus Reiß who introduced me to the beautiful and powerful language of mathematics and probability theory.

I am especially grateful to my supervisors PD Dr. Dominik Janzing, Prof. Dr. Bernhard Schölkopf and Prof. Dr. Peter Bühlmann. I was lucky to be supervised not only by excellent researchers and teachers but also by people who demonstrated how to achieve scientific progress in honest and constructive discussions. Cooperation was always much more important than competition.

Many thanks to the two groups I was a member of during my PhD. Both consist of smart and friendly people who always managed to create an open atmosphere. Special thanks to the MPI for Intelligent Systems in Tübingen for tea and cake talks, retreats and for the causality meetings; and to the Seminar for Statistics at the ETH Zurich for Backgammon, music, table tennis and for all the fondue.

I thank Dr. Joris Mooij, Dr. Kun Zhang, Dr. Patrik Hoyer, Eleni Sgouritsa and Jakob Zscheischler for fruitful collaborations. And thanks to Dr. Leon Bottou for an exciting time at Microsoft Research.

I made many friends in the Studentenphilharmonie, in Bebenhausen, at the Arbeitskreis Klima and at the Deutsche SchülerAkademie. Without them life would be much less colorful.

And special thanks to my family, the most wonderful people in the world.

Contents

Abstract **11**

Zusammenfassung **13**

1. Introduction **17**
1.1. Problem of Causal Inference 17
1.2. Determining Causal Effects from the Graph 20
1.3. Relating Graph and Reality: The True Causal Graph . . 21
1.4. Applying Causal Concepts to Machine Learning 25
1.5. Publications 27

2. Relating Graph and Distribution **31**
2.1. Graph Notations 31
2.2. Graphical Models (GMs) 36
2.3. Structural Equation Models (SEMs) 36
2.4. Relation between GMs and SEMs 37
2.5. Identifiability of the Graph Given the Distribution . . . 38
2.5.1. Identifiability in GMs 39
2.5.2. Identifiability in SEMs 39
2.6. When Faithfulness does not hold 40
2.7. Relating True Causal Graph and Distribution 43
2.7.1. Discussion of Assumptions 43
2.7.2. Independence of Cause and Mechanism 46

3. Existing Algorithms **51**
3.1. Independence-Based Methods 51
3.2. Score-Based Methods 52
3.3. Linear Non-Gaussian Additive Models 54

4. Continuous Bivariate Additive Noise Models **57**
4.1. Introduction 57

4.2. Model Definition 57
4.3. Identifiability 58

5. Discrete Bivariate Additive Noise Models 63
5.1. Introduction 63
5.2. Model Definition 64
5.2.1. Integer Models 65
5.2.2. Cyclic Models 65
5.2.3. Relations 66
5.3. Identifiability 66
5.3.1. Integer Models 67
5.3.2. Cyclic Models 70
5.3.3. Special Case: X and Y binary 73
5.3.4. Mixed Models 76

6. From Bivariate to Multivariate Models 79
6.1. Introduction 79
6.2. Model Definition 80
6.3. Identifiability 84

7. Multivariate Gaussian Models with Same Error Variances 87
7.1. Introduction 87
7.2. Model Definition 87
7.3. Identifiability 88

8. Extension: Causal Inference on Time Series 91
8.1. Introduction 91
8.2. Existing Methods 92
8.2.1. Granger Causality 92
8.2.2. ANLTSM 94
8.2.3. TS-LiNGAM 94
8.2.4. Limitations of Existing Methods 94
8.3. SEMs for Time Series: TiMINo 95
8.4. Identifiability 96

9. Extension: Confounder Detection 99
9.1. Introduction 99
9.2. Model Definition 100

9.3. Identifiability 102

10. Algorithms **103**
10.1. Continuous Bivariate Models 103
10.2. Discrete Bivariate Models 104
10.3. Multivariate Models 107
10.3.1. Finding a Suitable Structure 107
10.3.2. Finding all Suitable Structures 110
10.4. Time Series 110
10.5. Confounders 114

11. Experiments **123**
11.1. Continuous Bivariate Models 123
11.2. Discrete Bivariate Models 129
11.3. Multivariate Models 144
11.4. Time Series 149
11.5. Confounder 156
11.6. Semi-Supervised Learning 161

12. Conclusions and Future Work **169**
12.1. Conclusions 169
12.2. Future Work 171

A. Proofs **175**
A.1. Proofs of Chapter 1 175
A.1.1. Proof of Proposition 1.4 175
A.2. Proofs of Chapter 2 175
A.2.1. Proof of Proposition 2.6 175
A.3. Proofs of Chapter 4 176
A.3.1. Proof of Theorem 4.1 176
A.3.2. Proof of Corollary 4.2 177
A.4. Proofs of Chapter 5 178
A.4.1. Proof of Theorem 5.3 178
A.4.2. Proof of Theorem 5.5 181
A.4.3. Proof of Theorem 5.9 184
A.5. Proofs of Chapter 6 188
A.5.1. Some Lemmata 188
A.5.2. Some Propositions 190

A.5.3. Proof of Theorem 6.6 191
A.6. Proofs of Chapter 7 194
A.6.1. Some Lemmata 194
A.6.2. Proof of Theorem 7.1. 196
A.7. Proofs of Chapter 8 200
A.7.1. A Lemma 200
A.7.2. Proof of Theorem 8.2 200

Bibliography **203**

Abstract

Causal inference tries to solve the following problem: given i.i.d. data from a joint distribution, one tries to infer the underlying causal DAG (directed acyclic graph), in which each node represents one of the observed variables.

For approaching this problem, we have to make assumptions that connect the causal graph with the joint distribution. Independence-based methods like the PC algorithm assume the causal Markov condition and faithfulness. These two conditions relate conditional independences and the graph structure; this allows to infer properties of the graph by testing for conditional independences in the joint distribution. Independence-based methods encounter the following difficulties: (1) One can discover causal structures only up to Markov equivalence classes, in particular one cannot distinguish between $X \to Y$ and $Y \to X$. (2) In practice, conditional independence testing is difficult. Especially when the conditioning set is large, their power is often relatively low. (3) When the data come fron a non-faithful distribution, the results may be wrong, but the user does not realize it. Also, when the set of variables is causally insufficient, i.e. some important variables have not been observed, those methods may draw wrong conclusions.

In structural equation models (SEMs) each variable X_j is a function of a set of nodes $\mathbf{PA}_j$ and some noise variable N_j:

$$X_j = f_j(\mathbf{PA}_j, N_j)\,, \qquad j = 1, \ldots, p$$

where the N_j are jointly independent. The corresponding graph is obtained by drawing directed arrows from each variable in $\mathbf{PA}_j$ to X_j (the $\mathbf{PA}_j$ become parents of X_j). In this form, SEMs are too general to be used for structure learning. Given a distribution, we can find for each DAG with respect to which the distribution is Markov to a

corresponding SEM. This changes, however, if we consider *restricted* SEMs, in which some combinations of function and the distribution of noise and parents are excluded. In Gaussian SEMs with linear functions and additive noise, for example, the graph can be identified from the joint distribution again up to Markov equivalence classes (assuming faithfulness). This, however, constitutes a somewhat exceptional case. If the functions are linear and the noise is non-Gaussian, the DAG becomes fully identifiable. In this thesis we present alternative directions of deviating from the linear Gaussian case: (i) apart from few exceptions, identifiability also holds for non-linear functions and arbitrarily distributed additive noise. And (ii), if we require all noise variables to have the same variances, again, the DAG can be recovered from the joint distribution. We also present restricted SEMs for discrete variables with similar identifiability results. Moreover, we apply restricted SEMs to time series data. We further investigate whether it is possible to distinguish between the cases "X is causing Y", "Y is causing X" and "both variables are caused by a third unobserved variable" (throughout this work we call a common cause a *confounder*).

From our point of view, SEM-based causal inference and the restriction of the function class leads to the following advantages: (1) We can identify causal relationships even within an equivalence class. (2) Fitting a model with additive noise is easier than general conditional independence testing. (3) We do not require faithfulness. (4) If the model assumptions are violated (e.g. the data do not follow an additive noise model or there are hidden common causes), the method is able to output "I do not know" instead of giving wrong answers.

For all of the proposed identifiability results we present practical methods and apply them to simulated and real data sets.

Zusammenfassung

Kausale Inferenz beschäftigt sich mit dem folgenden Problem: Seien unabhängig und identisch verteilte Daten einer gemeinsamen Verteilung gegeben. Das Ziel der kausalen Inferenz liegt darin, den zugrundeliegende kausalen Graphen zu schätzen, dessen Knoten die Zufallsvariablen repräsentieren. Wir nehmen dabei an, dass der Graph gerichtete Kanten, aber keine Zykel enthält (directed acyclic graph).

Um dieses Problem anzugehen, müssen wir Annahmen treffen, die den kausalen Graphen mit der gemeinsamen Verteilung in Verbindung bringen. Sogenannte unabhängigkeitsbasierte Methoden wie der PC Algorithmus nehmen an, dass die Verteilung bzgl. des Graphen Markov und treu ist. Diese beiden Bedingungen verbinden die (bedingten) Unabhängigkeiten in der Verteilung mit der Graphstruktur und ermöglichen mit Hilfe von Unabhängigkeitstests Teile des Graphen zu identifizieren. Unserer Meinung nach treten hierbei jedoch folgende Schwierigkeiten auf: (1) Man kann die Graphstruktur nur bis auf Markoväquivalenzklassen bestimmen. Insbesondere sind wir so nicht in der Lage zwischen $X \to Y$ und $Y \to X$ zu unterscheiden. (2) In der Praxis ist es schwierig, bedingte Unabhängigkeitstests durchzuführen. Vor allem, wenn die Menge an Variablen, auf die man bedingt, größer wird, besitzen die Tests oft nur eine relativ kleine Macht. (3) Wenn die Treuebedingung verletzt ist, liefern diese Methoden falsche Ergebnisse, ohne dass man dies erkennen kann. Gleiches gilt bei kausal unvollständigen Strukturen.

In sogenannten Structural Equation Models (SEMs) wird jede Variable X_j als Funktion einer Menge von Knoten $\mathbf{PA}_j$ und einer Rauschvariable N_j geschrieben:

$$X_j = f_j(\mathbf{PA}_j, N_j)\,, \qquad j = 1, \ldots, p$$

wobei alle Variablen N_j gemeinsam unabhängig sind. Den entsprech-

enden Graphen erhält man, indem man gerichtete Kanten von jeder Variablen auf der rechten Seite (den Eltern $\mathbf{PA}_j$) zu der entsprechenden Variable X_j auf der linken Seite zeichnet. In dieser Form sind SEMs jedoch zu allgemein, als dass man sie zum Strukturlernen verwenden könnte. Gegeben einer Verteilung kann man zu jedem Graphen, zu dem die Verteilung Markov ist, ein SEM mit genau dieser Struktur finden. Dies ändert sich, falls wir *eingeschränkte* oder *restricted* SEMs betrachten. In diesen schränkt man die Klasse der möglichen Kombinationen von Rauschen und Funktion ein. Betrachtet man beispielsweise nur SEMs mit linearen Funktionen und normalverteilten Rauschvariablen, kann man den Graphen bis auf Markoväquivalenzklasse aus der gemeinsamen Verteilung bestimmen (unter der zusätzlichen Annahme von Treue). Dieser Fall stellt allerdings eine Art Ausnahme da. Im Falle von linearen Funktionen und nicht-normalverteiltem Rauschen wird der Graph identifizierbar. In dieser Dissertation präsentieren wir alternative Modelle, die ebenfalls zur Identifizierbarkeit führen. (i) Mit Ausnahme weniger Beispiele erhalten wir Identifzierbarkeit ebenfalls bei nicht-linearen Funktionen mit beliebig verteiltem additiven Rauschen. Und (ii), im linearen normalverteilten Fall beweisen wir Identifzierbarkeit des Graphens aus der Verteilung unter der Annahme, dass alle Rauschvariablen die gleiche Varianz besitzen. Wir führen eingeschränkte SEMs für diskrete Variablen ein und erhalten analoge Ergebnisse. Ebenfalls zeigen wir, dass unsere Methoden auch auf Zeitreihen anwendbar sind. Ferner untersuchen wir, in wie weit man bei zwei beobachteten Variablen die Fälle "X verursacht Y", "Y verursacht X" und "beide Variablen werden durch eine dritte Variable verursacht" unterscheiden können (Detektion einer unbeobachteten gemeinsamen Ursache).

Unserer Meinung nach bringt die SEM-basierte kausale Inferenz und die Einschränkung der Funktionenklasse folgende Vorteile mit sich: (1) Wir können kausale Relationen auch innerhalb einer Markoväquivalenzklasse bestimmen. (2) Ein Model mit additivem Rauschen zu fitten (für jede Variable müssen wir eine multivariate Regression durchführen) ist ein einfacheres Problem als das Testen von bedingten Unabhängigkeiten. (3) Die Methoden basieren nicht auf der Treuebedingung. (4) Falls die (z.T. starken) Modellannahmen verletzt sind (beispielsweise sind die Daten nicht durch ein Modell mit additivem

Rauschen erzeugt oder es gibt unbeobachtete gemeinsame Ursachen), ist die Methode in der Lage, unentschlossen zu bleiben anstatt eine falsche Antwort zu geben.

Für alle vorgestellten Identifizierbarkeitsergebnisse stellen wir praktische Methoden vor, die wir auf simulierte und reale Datensätze anwenden.

Chapter 1.

Introduction

1.1. Problem of Causal Inference

Consider the random variables

$$\begin{array}{ll} X: & \text{birth rate} \\ Y: & \#\,\text{storks}\,, \end{array} \tag{1.1}$$

for which the samples are drawn in different locations. For many countries the correlation between X and Y is significantly different from zero (for German data see [Matthews, 2000]). Although children are sometimes told that babies are delivered by storks, we do not expect that this is the causal explanation for this correlation. The correlation between X and Y can probably rather be explained by the influence of a third variable Z indicating whether the data is sampled from a rural area:

$$Z: \quad \text{rural area (yes/no)}\,. \tag{1.2}$$

Most people would agree that this (or a related) variable is necessary for explaining the causal relationship between X and Y and some would accept Figure 1.1 as a causal explanation. To reject the initial hypothesis $Y \to X$, one could perform an experiment by randomly distributing storks among different areas and then analyze the difference of birth rates between areas with different stork populations. Clearly, such an experiment would show that there is no influence from the stork population on the birth rate. Randomizing the birth rate by introducing corresponding laws, however, will not show an effect on the stork population either.

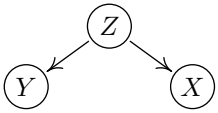


Figure 1.1.: A possible explanation for the correlation between X and Y including Z, see Equations (1.1) and (1.2).

Often, randomized studies (a special case of interventional experiments) cannot be performed in practice. They may be too expensive, unethical or even impossible to perform. This thesis investigates approaches to infer the causal graph from observational data only. In this work we concentrate mainly on acyclic models and introduce the notion of DAGs (directed acyclic graphs) in Section 2.2. This manifests a strong assumption, but simplifies the problem; only in Chapter 8 we investigate the case of time series, and can allow for causal loops. Given the notion of the true causal graph (that we specify in Section 1.3), we can now formulate the main problem of this thesis.

Problem 1.1 Given i.i.d. samples from $X_1, \ldots, X_p$ try to infer the underlying true causal DAG (see Figure 1.2).

For attacking Problem 1.1, we clearly need to relate the graph structure to properties of the joint distribution. In this work, we establish such connections between graphs and distributions and investigate under what assumptions we can identify the graph from the distribution. Only in Section 2.7 we discuss whether the assumptions made are sensible assumptions about the true *causal* graph. This way, we strictly separate mathematical statements from their causal interpretation.
In graphical models the Markov condition and faithfulness establish the connection between graph and distribution. It has been shown that under these assumptions the causal graph can be identified up to Markov equivalence (some arrows remain undirected) using methods like the PC algorithm, see Sections 2.1 and 3.1.
In this work, we mainly focus on structural equation models (SEMs) that are also referred to as functional models. In these models, each variable can be written as a function of its parents and some noise variable. The noise variables are then assumed to be jointly independent.

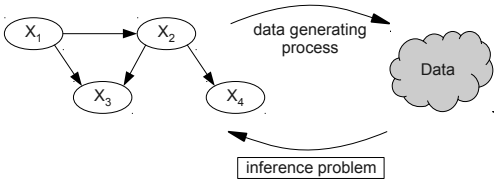


Figure 1.2.: The main problem addressed in this thesis.

Knowing that a distribution was generated by such an SEM is not enough to recover the graph. However, if we consider *restricted* SEMs, in which the functions have to belong to a certain function class, one can show for some of these classes that the graphical structure is identifiable from the joint distribution (under some conditions). For additive noise models we provide identifiability results for two random variables both in the continuous and in the discrete case (Chapters 4 and 5). Chapter 6 generalizes these results to the multivariate case. Although the linear Gaussian setting is not identifiable in general, we show in Chapter 7 that restricting the error variables to have the same variance leads to identifiability, again. Chapter 9 aims at identifying whether there is a causal relationship between two variables or whether their dependence can be explained by a hidden common cause. In Chapter 10 we present practical algorithms, which we test on simulated and real data sets in Chapter 11. Throughout this thesis we concentrate on the case of i.i.d. data. Only in Chapter 8 we drop this assumption and investigate the causal inference between time series.

In some situations, solving Problem 1.1 might be interested in itself since it provides further insight about the system under examination. In the example above, knowing that X is not caused by Y would tell us to look for another explanation of children being born. Section 1.2 shows how we can use the information of the causal graph to compute causal effects from some variables on others and thus predict the result of randomized experiments. As an alternative application and a possibly fruitful research path, we briefly describe in Section 1.4, how ideas from causal inference may have an impact on more traditional

inference problems considered in Machine Learning.

Finally, we stress our believe that inferring the causal graph is a very ambitious task. Especially since assumptions like causal sufficiency (Section 1.3) will be often violated in practice, we think results should be interpreted carefully. Our hope is that causal inference on observational data may help to detect the strong causal effects or provide hints how to design the next interventional experiment [as in Stekhoven et al., 2012, Maathuis et al., 2010].

1.2. Determining Causal Effects from the Graph

Given a directed acyclic graph (DAG)[1] $\mathcal{G}$, Pearl [2009] introduces the *do*-notation as a mathematical description of interventional experiments. More precisely, $do(X_j = \tilde{p}(x_j))$ stands for setting the variable X_j randomly according to the distribution $\tilde{p}(x_j)$, irrespective of its parents, while leaving all other variables unchanged. Formally:

Definition 1.2 Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a collection of variables with joint distribution $\mathcal{L}(\mathbf{X})$ that we assume to be absolutely continuous with respect to the Lebesgue measure or the counting measure (i.e. there exists a probability density function or a probability mass function). Given a DAG $\mathcal{G}$ over $\mathbf{X}$, we define the *interventional distribution* $do(X_j = \tilde{p}(x_j))$ of $X_1, \ldots, X_p$ by

$$p\big(x_1, \ldots, x_p \mid do(X_j = \tilde{p}(x_j))\big) := \prod_{i \neq j}^{p} p(x_i | x_{\mathbf{PA}_i}) \cdot \tilde{p}(x_j)\,,$$

where $\tilde{p}(x_j)$ is either a probability density function or a probability mass function. Similarly, we can intervene at different nodes at the same time by defining the interventional distribu-

[1] In Sections 1.2 and 1.3 we already use terminology that we introduce only in Chapter 2. Readers not familiar with the language of graphs and structural equation models may want to skip to that chapter first.

tion $do(X_j = \tilde{p}(x_j))$ for $j \in J$ as

$$p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{p}(x_j)), j \in J\big) := \prod_{i \notin J} p(x_i | x_{\mathbf{PA}_i}) \cdot \prod_{j \in J} \tilde{p}(x_j)\,.$$

Here, $x_{\mathbf{PA}_i}$ denotes the tuple of all x_j for X_j being a parent of X_i in $\mathcal{G}$. Pearl [2009] introduces Definition 1.2 with the special case of $\tilde{p}(x_j) = \delta_{x_j, \tilde{x}_j}$, where $\delta_{x_j, \tilde{x}_j} = 1$ if $x_j = \tilde{x}_j$ and $\delta_{x_j, \tilde{x}_j} = 0$ otherwise. Although we have not seen our Definition in literature, we regard it as a natural extension, which may have been introduced before. Note that in general

$$p(x_1, \ldots, x_n \,|\, do(X_j = \tilde{x}_j)) \neq p(x_1, \ldots, x_n \,|\, X_j = \tilde{x}_j)\,.$$

1.3. Relating Graph and Reality: The True Causal Graph

In this section we clarify what we mean by the true causal graph $\mathcal{G}_c$. In short, we use this term if one can read off the results of randomized studies from $\mathcal{G}_c$ and the joint distribution. This means that the graph and the joint distribution lead to causal effects that one observes in practice.

Definition 1.3 Given variables $X_1, \ldots, X_p$, we call the graph $\mathcal{G}_c$ the true causal graph if

- $\mathcal{G}_c$ is a directed acyclic graph.
- the distribution is Markov with respect to $\mathcal{G}_c$ (see Definition 2.1)
- the distribution satisfies causal minimality with respect to $\mathcal{G}_c$ (see Definition 2.1)
- for all J and $\tilde{p}(x_j)$ with $j \in J$ the distribution obtained by randomizing variables X_j with $\tilde{p}(x_j)$ coincides with the distribution $p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{p}(x_j)), j \in J\big)$, computed as in Definition 1.2.

Whenever we attach causal meaning to a graph we say "true causal graph" and use the notation $\mathcal{G}_c$. Most of the results in this thesis can be stated without relating them to causality. In the sense of Definition 1.3, the graph $X \to Y$ is certainly not causal for the example shown in Figure 1.1. In some situations, the precise design of a randomized experiment is not obvious. While most people would agree on how to randomize over medical treatment procedures, there is probably less agreement how to randomize over the tolerance of a person (does this include other changes of his personality, too?). Only sometimes, this problem can be resolved by including more variables and taking a less coarse-grained point of view. We do not go into further detail since we believe that this would require philosophical deliberations which lie beyond the scope of this work. Instead, we may explicitly add the requirement that "most people agree on what a randomized experiment should look like in this context" to Definition 1.3. From this point of view, Figure 1.1 may not be regarded as the true causal DAG for X, Y and Z since it is not apparent how to randomize over Z (does making a city rural include changing salaries and religious attitudes?). This way we gain a solid foundation of what we mean by causal relationships, but lose the ability of making causal statements in some situations.
If there exists a causal DAG, it is the only one (see Appendix A.1.1 for a proof):

Proposition 1.4 *Let $\mathcal{G}_c$ be the true causal DAG of $X_1, \dots, X_p$. Then $\mathcal{G}_c$ is unique.*

From the following example[2] we can draw two conclusions: (1) Although causal minimality sounds like a very natural assumption, we may miss some causal relationships. And (2), the true causal graph need not be unique if we do not require causal minimality in Definition 1.3.

Example 1.5 Suppose that X, Y and N are binary variables with $X, N \overset{\text{iid}}{\sim} Ber(0.5)$ and

$$Y = X + N\,, \tag{1.3}$$

[2]This example emerged from a personal discussion with Dominik Janzing, but it may have been looked at before.

where addition is computed in $\mathbf{Z}/2\mathbf{Z}$ (see Section 5.2.2). In the resulting distribution we find $X \perp\!\!\!\perp Y$ and we thus consider the empty graph as the true causal graph $\mathcal{G}_c$ (the distribution does not satisfy causal minimality with respect to the graph $\mathcal{G} : X \to Y$). Some people may argue that this is the wrong causal graph because it gives wrong answers to some counterfactual questions. Suppose we have observed the sample $X = 1$ and $Y = 1$. From the empty graph $\mathcal{G}_c$ we would conclude: "Here, Y would have been 1 if X had been 0.". The SEM (1.3) with graph $\mathcal{G}$, however, leads to the counterfactual statement: "Here, Y would have been 0 if X had been 0.". We are willing to accept that we cannot make the correct counterfactual statement here. After all, there are no differences in the interventional distributions (any *do*-statements will be the same) and thus, there is no way to distinguish between $\mathcal{G}_c$ and $\mathcal{G}$ in real life, even if one has access to interventional experiments.

We now investigate the behavior of the true causal graph under marginalization.

Example 1.6 (i) If $X \to Y \to Z$ is the true causal graph for variables X, Y and Z and $X \not\perp\!\!\!\perp Z$, then $X \to Z$ is the true causal graph for X and Z.

(ii) If the graph in Figure 1.1 is the true causal graph for X, Y and Z, there is no true causal graph for the variables X and Y (the *do*-statements do not coincide).

(iii) Assume that the graph $X \to Y \to Z$ with additional $X \to Z$ is the true causal graph for X, Y and Z and assume further that $\mathcal{L}(X, Y, Z)$ is faithful with respect to this graph. Then, the true causal graph for the variables X and Z is $X \to Z$.

(iv) If the situation is the same as in (iii) with the difference that $X \perp\!\!\!\perp Z$ (i.e. $\mathcal{L}(X, Y, Z)$ is not faithful with respect to the true causal graph), the true causal graph for X and Z is the empty graph.

Cases (iii) and (iv) show that

Remark 1.7 There are no solely graphical criteria for marginalization of the true causal graph.

Some people may want to define the true causal graph rather using a structural equation model (see Section 2.3), which is more restrictive in the following sense.

Proposition 1.8 *For given variables $X_1, \ldots, X_p$ let us assume that there is a structural equation model with graph $\mathcal{G}_c$, such that*

- *$\mathcal{G}_c$ is a directed acyclic graph.*
- *The distribution satisfies causal minimality with respect to $\mathcal{G}_c$ (see Definition 2.1)*
- *The structural equations describe the true data generating process.*
- *Intervening at one (or more nodes) does not change the structural equations. This means that an interventional experiment $do(X_j = \tilde{p}(x_j))$ is described by replacing the structural equation for X_j with an equation $X_j = \tilde{N}_j$, where $\tilde{N}_j$ has a distribution according to $\tilde{p}(x_j)$ and all noise variables are jointly independent. (See the notion of causal stability in [Pearl, 2009, Chapter 1.3.2]).*

Then $\mathcal{G}_c$ is the true causal graph defined in Definition 1.3.

The proof follows from the fact that the Markov assumption is satisfied in a structural equation model [Pearl, 2009, Theorem 1.4.1]. A further discussion of the differences between the two approaches described in Definition 1.3 and Proposition 1.8 lies beyond the scope of this work.

The following assumption requires that all "relevant" variables have been observed.

Definition 1.9 $X_1, \ldots, X_p$ are causally sufficient if there is a true causal DAG $\mathcal{G}_c$.

Probably, the variables X, Y and Z described in Figure 1.1 can be made causally sufficient, by including variables like education, income and religion, too. Richardson and Spirtes [2002] introduces a representation of graphs with hidden variables that is closed under marginalization. The algorithm FCI [Spirtes et al., 2000] exploits the (conditional) independences in the data to reconstruct the graph. On the side of structural equation models less work covers the case of hidden variables. Hoyer

et al. [2008] deals with linear equations and non-Gaussian noise. In Chapter 9 we make a first attempt to allow for hidden variables in the special case of two observed variables. Except for Chapter 9, however, in this work, we assume causal sufficiency.

1.4. Applying Causal Concepts to Machine Learning

We now give two examples how causal information can help for solving problems in the field of machine learning.

Causal and Anticausal Learning

This paragraph is based on [Schölkopf et al., 2012]. We consider two random variables X and Y and assume that they are causally sufficient (see Definition 1.9). Knowing, which variable is the cause (we denote this variable by C) and which is the effect (E) has direct implications on how to approach prediction problems. In this thesis, we only discuss the example of semi-supervised learning, although more applications can be found in [Schölkopf et al., 2012]. In semi-supervised learning, we are given i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{L}(X, Y)$ and additional samples from the input $X_{n+1}, \ldots, X_m$; the goal is to learn the conditional distribution $\mathcal{L}(Y|X)$. Section 2.7.2 describes how causal inference techniques sometimes assume that cause ($\mathcal{L}(C)$) and mechanism ($\mathcal{L}(E|C)$) are independent. The independence of noise variables in structural equation models constitutes only one important example. Those causal inference methods often exploit that in the generic cases there is a dependence in the opposite direction, that is between $\mathcal{L}(C|E)$ and $\mathcal{L}(E)$. If $X = C$ and $Y = E$ it follows that semi-supervised learning cannot work: any new knowledge about $\mathcal{L}(X)$ does not tell us anything about $\mathcal{L}(Y|X)$. Only if $X = E$ and $Y = C$ (or there is a common cause), the idea of semi-supervised learning is compatible with the independence assumption. In Section 11.6 we provide empirical evidence for this observation.

Finding Optimal Strategies Using Interventions

In this section we briefly explain how knowing the causal graph may help us to find optimal strategies. This idea is based on [Bottou et al., 2012]. Consider a set of random variables $X_1, \ldots, X_p$, and some objective function $Y = \ell(X_1, \ldots, X_p)$ of interest (it is possible to choose $Y := X_i$). Consider further the Markov factorization [Lauritzen, 1996] according to the (most often unknown) true causal DAG $\mathcal{G}_c$:

$$p(x_1, \ldots, x_p) = \prod_{i=1}^{p} p(x_i | x_{\mathbf{PA}_i}) \,. \tag{1.4}$$

Suppose further that we can control some part of the system. For example, $p(x_3|x_{\mathbf{PA}_3})$. In the example of online advertisement placement X_3 could be the number of ads shown depending on some user information and $Y = X_p$ the revenue for one advertisement placement. In particular, being able to control $p(x_3|x_{\mathbf{PA}_3})$ means that we know the parents $\mathbf{PA}_3$ of X_3. We aim at answering questions of the sort: How much money would we make on average if we use $p^*(x_3|x_{\mathbf{PA}_3})$ instead of $p(x_3|x_{\mathbf{PA}_3})$? Or, in mathematical terms: what is $\mathbf{E}_{p^*}\ell(X_1, \ldots, X_p)$, where ℓ is the objective function and

$$p^*(x_1, \ldots, x_p) = \prod_{i \neq 3} p(x_i | x_{\mathbf{PA}_i}) \cdot p^*(x_3 | x_{\mathbf{PA}_3}) \,?$$

This can be seen as a generalization of an interventional distribution. Analogously to Definition 1.2 we can define

$$\begin{aligned} &p\big(x_1, \ldots, x_p \mid do(X_j | X_{\mathbf{PA}_j} = \tilde{p}(x_j | x_{\mathbf{PA}_j}))\big) \\ &\qquad := \prod_{i \neq j}^{p} p(x_i | x_{\mathbf{PA}_i}) \cdot \tilde{p}(x_j | x_{\mathbf{PA}_j}) \,. \end{aligned}$$

Further, what would be the optimal choice of $p^*(x_3|x_{\mathbf{PA}_3})$? If the part we can control is parameterized by θ, i.e. we have $p^*(x_3|x_{\mathbf{PA}_3}) = p^*_\theta(x_3|x_{\mathbf{PA}_3})$: what is $\operatorname{argmax}_\theta \mathbf{E}_{p^*_\theta} Y$? For answering this question it may be helpful to look at $\frac{\partial}{\partial\theta} \mathbf{E}_{p^*_\theta} Y$. In [Bottou et al., 2012] we provide answers to these questions that are based on data, namely on i.i.d. samples from the joint distribution, with some variables being

unobserved. Precisely, we require observations from X_3, $\mathbf{PA}_3$ and $Y = \ell(X_1, \ldots, X_p)$. The key idea is to note that

$$\begin{aligned}\mathbf{E}_{p^*} Y &= \mathbf{E}_{p^*} \ell(X_1, \ldots, X_p) \\ &= \int \ell(x_1, \ldots, x_p) \frac{p^*(x_3 \,|\, x_{\mathbf{PA}_3})}{p(x_3 \,|\, x_{\mathbf{PA}_3})} \, p(x_1, \ldots, x_p) \, dx_1 \cdots dx_p\end{aligned}$$

and estimate this quantity by using the observed samples

$$\frac{1}{N} \sum_{i=1}^{N} \ell(X_{1,i}, \ldots, X_{p,i}) \, \frac{p^*(X_{3,i} \,|\, X_{\mathbf{PA}_3,i})}{p(X_{3,i} \,|\, X_{\mathbf{PA}_3,i})} \, .$$

In [Bottou et al., 2012] we further compute derivatives, discuss the trade-off between bias and variance of this estimator and construct approximate confidence intervals. In its Section 5.1 ("Better Reweighting Variables") we show how knowing other parts of the causal structure may help to obtain better estimators for the confidence intervals.
This problem is closely related to reinforcement learning [e.g. Sutton and Barto, 1998], multi-armed bandits [e.g. Robbins, 1952] and contextual bandits [e.g. Li et al., 2010]. The approach described in [Bottou et al., 2012] establishes the link to the causal point of view and further stresses how *causal* information can sometimes be beneficial to obtain better estimators (not discussed here).

1.5. Publications

This thesis is a cumulative dissertation. It is built upon and provides results from the publications shown in Table 1.1. We removed abstracts and changed notation in order to increase the homogeneity and readability of this work. Many parts of the publications are copied, only some parts are slightly modified. However, we did not change any content or result. Chapter 1 (apart from Section 1.4) and Chapter 2 (if not stated otherwise) are not built on any existing publication.
The publications shown in Table 1.2 have been written during the PhD studies. They are related to this thesis, but although their content is described, not all of their results are presented in detail.

publication	used in
P. Hoyer, D. Janzing, J. Mooij, J. Peters and B. Schölkopf: *Nonlinear causal discovery with additive noise models*, NIPS 2008 [Hoyer et al., 2009]	Chapter 4 Section 10.1 Section 11.1
J. Peters, D. Janzing and B. Schölkopf: *Causal inference on discrete data using additive noise models*, IEEE TPAMI 2011 [Peters et al., 2011a]	Chapter 5 Section 10.2 Section 11.2
J. Peters, J. M. Mooij, D. Janzing and B. Schölkopf: *Identifiability of Causal Graphs using Functional Models*, UAI 2011 [Peters et al., 2011b]	Chapter 6 Section 10.3.2 Section 11.3
J. Peters and P. Bühlmann: *Identifiability of Gaussian Structural Equation Models with Same Error Variances*, ArXiv e-print [Peters and Bühlmann, 2012]	Chapter 7
J. Peters, D. Janzing and B. Schölkopf: *Causal Inference on Time Series using Structural Equation Models*, ArXiv e-print [Peters et al., 2012]	Chapter 8 Section 10.4 Section 11.4
D. Janzing, J. Peters, J. M. Mooij and B. Schölkopf: *Identifying confounders using additive noise models*, UAI 2009 [Janzing et al., 2009]	Chapter 9 Section 10.5 Section 11.5
J. Mooij, D. Janzing, J. Peters and B. Schölkopf: *Regression by dependence minimization and its application to causal inference*, ICML 2009 [Mooij et al., 2009]	Section 10.3.1

Table 1.1.: Main publications this thesis is based on.

publication	mentioned in
B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang and J. Mooij: *On causal and anticausal learning*, ICML 2012 [Schölkopf et al., 2012]	Section 1.4
L. Bottou and J. Peters and J. Quiñonero-Candela and D. X. Charles and D. M. Chickering and E. Portugualy and D. Ray and P. Simard and E. Snelson: *Counterfactual Reasoning and Learning Systems*, ArXiv e-print [Bottou et al., 2012]	Section 1.4
K. Zhang, J. Peters, D. Janzing and B. Schölkopf: *Kernel-based Conditional Independence Test and Application in Causal Discovery*, UAI 2011 [Zhang et al., 2011]	Section 3.1
J. Peters, D. Janzing and B. Schölkopf: *Identifying Cause and Effect on Discrete Data using Additive Noise Models*, AISTATS 2010 [Peters et al., 2010]	Chapter 5 Section 10.2 Section 11.2

Table 1.2.: Further publications the main author have been involved in. Their results are mentioned, but not shown in detail.

Chapter 2.

Relating Graph and Distribution

The main theoretical question of this thesis is under what assumptions on the data generating process can we infer the causal graph from the joint distribution? Clearly, we have to establish a relation between graphs and the corresponding joint distribution. The approach taken by conditional independence-based methods relies on the assumptions that the distribution is Markov and faithful with respect to the graph. It is apparent that under these assumptions the graph can be identified from the distribution up to Markov equivalence (some arrows remain undirected), see Proposition 2.8. Constructive methods like the PC algorithm have been invented [Spirtes et al., 2000]. Our main focus, however, lies on structural equation models (SEMs) that we also call functional models, the functions of which are required to belong to a specified function class. For different function classes we prove that if the data generating process belongs to such a restricted SEM, one can identify the complete underlying graph.

2.1. Graph Notations

We start with some basic notation for graphs. Consider a finite family of random variables $\mathbf{X} = (X_1, \ldots, X_p)$ with index set $\mathbf{V} := \{1, \ldots, p\}$ (we use capital letters for random variables and bold letters for sets or vectors). We denote their joint distribution by $\mathcal{L}(\mathbf{X})$ and sometimes write $\mathbb{P}^{X_1}(x)$ for $\mathbb{P}(X_1 = x)$. We write $p_{X_1}(x)$ or simply $p(x)$ for the Radon-Nikodym derivative from $\mathcal{L}(X_1)$ either with respect to the Lebesgue or the counting measure and (sometimes implicitly) assume its existence. A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of nodes $\mathbf{V}$ and edges $\mathcal{E} \subseteq \mathbf{V}^2$. In a slight abuse of notation we sometimes identify the nodes

(or vertices) $j \in \mathbf{V}$ with the variables X_j. The following definitions can be found in [Spirtes et al., 2000, Koller and Friedman, 2009, Peters et al., 2011b], for example.

Definition 2.1 Let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be a graph with $\mathbf{V} := \{1, \ldots, p\}$ and corresponding random variables $\mathbf{X} = (X_1, \ldots, X_p)$.

- $\mathcal{G}_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a *subgraph* of $\mathcal{G}$ if $\mathbf{V}_1 = \mathbf{V}$ and $\mathcal{E}_1 \subset \mathcal{E}$. If additionally, $\mathcal{E}_1 \neq \mathcal{E}$, we call $\mathcal{G}_1$ a *proper subgraph* of $\mathcal{G}$.
- X_i is called a *parent* of X_j if $(i, j) \in \mathcal{E}$ and a *child* if $(j, i) \in \mathcal{E}$. The set of parents of X_j is denoted $\mathbf{PA}_j^{\mathcal{G}}$, the set of its children by $\mathbf{CH}_j^{\mathcal{G}}$. Two nodes i and j are *adjecent* if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$.
- We call $\mathcal{G}$ *fully connected* if all pairs of nodes are adjacent.
- Three nodes are called an *immorality* or a *v-structure* if one node is a child of the two others, which themselves are not adjacent.
- The *skeleton* of $\mathcal{G}$ is the set of all edges without taking the direction into account, that is all (i, j), such that $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$.
- A *path* in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $X_{i_1}, \ldots, X_{i_n}$, such that $(i_k, i_{k+1}) \in \mathcal{E}$ or $(i_{k+1}, i_k) \in \mathcal{E}$ for all $k = 1, \ldots, n-1$. If for all k the former holds we speak of a *directed path* between X_{i_1} and X_{i_n} and call X_{i_n} a *descendant* of X_{i_1}. We denote all descendants of X_i by $\mathbf{DE}_i^{\mathcal{G}}$ and all non-descendants of X_i by $\mathbf{ND}_i^{\mathcal{G}}$. If $(i_{k-1}, i_k) \in \mathcal{E}$ and $(i_{k+1}, i_k) \in \mathcal{E}$, X_{i_k} is called a *collider* on this path.
- $\mathcal{G}$ is called a *directed acyclic graph (DAG)* if there is no pair (X_j, X_k), such that there are directed paths from X_j to X_k and from X_k to X_j.
- A path between X_{i_1} and X_{i_n} is *blocked by a set* $\mathbf{S}$ (with neither X_{i_1} nor X_{i_n} in this set) whenever there is a node X_{i_k}, such that one of the following two possibilities hold:

1. $X_{i_k} \in \mathbf{S}$ and

$$X_{i_{k-1}} \to X_{i_k} \to X_{i_{k+1}} \text{ or}$$
$$X_{i_{k-1}} \leftarrow X_{i_k} \leftarrow X_{i_{k+1}} \text{ or}$$
$$X_{i_{k-1}} \leftarrow X_{i_k} \to X_{i_{k+1}}$$

2. $X_{i_{k-1}} \to X_{i_k} \leftarrow X_{i_{k+1}}$ and neither X_{i_k} nor any of its descendants is in $\mathbf{S}$.

We say that two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are *d-separated* by a third (also disjoint) subset $\mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$.

- The joint distribution $\mathcal{L}(\mathbf{X})$ is said to be *Markov with respect to the DAG* $\mathcal{G}$ if

$$\mathbf{A}, \mathbf{B} \; d\text{-sep. by } \mathbf{C} \;\Rightarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$.

- $\mathcal{L}(\mathbf{X})$ is said to be *faithful to the DAG* $\mathcal{G}$ if

$$\mathbf{A}, \mathbf{B} \; d\text{-sep. by } \mathbf{C} \;\Leftarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$. Throughout this thesis, the symbol $\perp\!\!\!\perp$ denotes (conditional) independence.

- A distribution satisfies *causal minimality* with respect to $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$.

- We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov with respect to $\mathcal{G}$:

$$\mathcal{M}(\mathcal{G}) := \{\mathcal{L}(\mathbf{X}) \,:\, \mathcal{L}(\mathbf{X}) \text{ is Markov wrt } \mathcal{G}\}\,.$$

- Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations, that means the Markov condition entails the same set of (conditional) independence conditions.

Verma and Pearl [1991] showed that

Lemma 2.2 *Two graphs are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

Figure 2.1 shows an example of two Markov equivalent graphs.

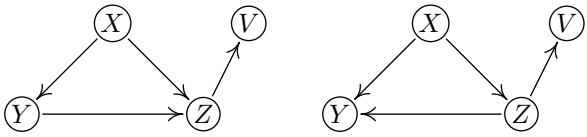


Figure 2.1.: Two Markov-equivalent DAGs.

The definition of faithfulness is not very intuitive at first glance. Note that

Remark 2.3 If $\mathcal{L}(\mathbf{X})$ is faithful with respect to $\mathcal{G}$, then causal minimality is satisfied.

In Example 1.5, the distribution is neither faithful nor is causal minimality satisfied. We now give another example of a distribution that is not faithful with respect to some DAG $\mathcal{G}_1$. This is achieved by making two paths cancel and thus rendering variables independent that should not be (according to the graph structure).

Example 2.4 Consider the following two graphs

$\mathcal{G}_1$ $\mathcal{G}_2$

Corresponding to the left graph we generate a joint distribution by the following equations.

$$X = N_X$$
$$Y = aX + N_Y$$
$$Z = bY + cX + N_Z$$

with $N_X \sim N(0, \sigma_X^2)$, $N_Y \sim N(0, \sigma_Y^2)$ and $N_Z \sim N(0, \sigma_Z^2)$. This is an example of a linear Gaussian structural equation model with graph $\mathcal{G}_1$ that we formally define below in Section 2.3. Now, if

$$a \cdot b + c = 0$$

the distribution is not faithful with respect to $\mathcal{G}_1$ (more precisely not triangle-faithful, see Section 2.7.1) since we obtain $X \perp\!\!\!\perp Z$.

Correspondingly, we generate a distribution related to graph $\mathcal{G}_2$:

$$\begin{aligned} X &= \tilde{N}_X \\ Y &= \tilde{a}X + \tilde{b}Z + \tilde{N}_Y \\ Z &= \tilde{N}_Z \end{aligned}$$

with $\tilde{N}_\cdot \sim N(0, \tau_\cdot^2)$. If we choose

$$\begin{aligned} \tau_X^2 &= \sigma_X^2, \\ \tilde{a} &= a, \\ \tau_Z^2 &= b^2\sigma_Y^2 + \sigma_Z^2, \\ \tilde{b} &= \frac{b\sigma_Y^2}{b^2\sigma_Y^2 + \sigma_Z^2} \qquad \text{and} \\ \tau_Y^2 &= \sigma_Y^2 - \frac{b^2\sigma_Y^4}{b^2\sigma_Y^2 + \sigma_Z^2} > 0 \end{aligned}$$

both models lead to the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & a\sigma_X^2 & 0 \\ a\sigma_X^2 & a^2\sigma_X^2 + \sigma_Y^2 & b\sigma_Y^2 \\ 0 & b\sigma_Y^2 & b^2\sigma_Y^2 + \sigma_Z^2 \end{pmatrix}$$

and thus to the same distribution. It can be checked that the distribution is faithful with respect to $\mathcal{G}_2$.

As stated in Remark 2.3, causal minimality is a strictly weaker assumption than faithfulness. The distribution from Example 2.4 is faithful with respect to $\mathcal{G}_2$, but not with respect to $\mathcal{G}_1$. Nevertheless, for both models, causal minimality is satisfied. The distribution is not Markov to any proper subgraph of $\mathcal{G}_1$ or $\mathcal{G}_2$ since removing any of the arrows would correspond to a new (conditional) independence that does not hold in the distribution. Note that $\mathcal{G}_2$ is not a proper subgraph of $\mathcal{G}_1$.

2.2. Graphical Models (GMs)

For random variables $\mathbf{X} = (X_1, \ldots, X_p)$, we define a graphical model as a tuple $(\mathcal{G}, \mathcal{L}(\mathbf{X}))$ with a joint probability distribution $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ that is Markov with respect to a directed acyclic graph $\mathcal{G}$.

2.3. Structural Equation Models (SEMs)

A structural equation model (SEM) (that we will also call a functional model) is defined as a tuple $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$, where $\mathcal{S} = (S_1, \ldots, S_p)$ is a collection of p equations

$$S_j : \quad X_j = f_j(X_{\mathbf{PA}_j}, N_j), \qquad j = 1, \ldots, p \tag{2.1}$$

and $\mathcal{L}(\mathbf{N}) = \mathcal{L}(N_1, \ldots, N_p)$ is the joint distribution of the noise variables, which we require to be jointly independent (thus, $\mathcal{L}(\mathbf{N})$ is a product distribution). Note that we consider SEMs only for real-valued random variables $X_1, \ldots, X_p$. The graph of a structural equation model $\mathbf{graph}(\mathcal{S}, \mathcal{L}(\mathbf{N}))$ is obtained simply by drawing direct edges from each parent to its direct cause, i.e. from each variable X_k occurring on the right-hand side of equation (2.1) to X_j. This graph is required to be acyclic. According to the notation defined in Section 2.1, $X_{\mathbf{PA}_j}$ are the parents of X_j. Each SEM generates a law $\mathcal{L}(\mathbf{X}) = \mathbf{law}_X(\mathcal{S}, \mathcal{L}(\mathbf{N}))$. Pearl [2009] shows in Theorem 1.4.1 that $\mathbf{law}_X(\mathcal{S}, \mathcal{L}(\mathbf{N}))$ is Markov with respect to $\mathbf{graph}(\mathcal{S}, \mathcal{L}(\mathbf{N}))$.

Structural equation models have been used for a long time in fields like agriculture or social sciences [e.g. Wright, 1921, Bollen, 1989]. Model selection, for example, was done by fitting different structures that were considered as reasonable given the prior knowledge about the system. These candidate structures were then compared using goodness of fit tests (see Section 3.2). The question of identifiability, however, has not been addressed until recently.

2.4. Relation between GMs and SEMs

We thus have the following mapping

$$\phi: \begin{array}{rcl} \mathcal{SM} := \{\text{SEMs}\} & \to & \{\text{graph models}\} =: \mathcal{GM} \\ \big(\mathcal{S}, \mathcal{L}(\mathbf{N})\big) & \mapsto & \big(\mathbf{graph}(\mathcal{S}, \mathcal{L}(\mathbf{N})), \mathbf{law}_X(\mathcal{S}, \mathcal{L}(\mathbf{N}))\big) \end{array}$$

The mapping ϕ is not injective. In fact, the structural equation model contains strictly more information than the corresponding graph and law. This information sometimes helps to answer counterfactual questions, as shown in the following example.

Example 2.5 Let $N_1, N_2 \sim Ber(0.5)$ and $N_3 \sim U(\{0,1,2\})$, such that the three variables are jointly independent. We define two different SEMs, first consider $\mathcal{S}_A$:

$$\mathcal{S}_A = \begin{cases} X_1 = N_1 \\ X_2 = N_2 \\ X_3 = \quad (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} \\ \qquad + \qquad\qquad N_3 \qquad\qquad \cdot 1_{X_1 = X_2} \end{cases}$$

If X_1 and X_2 have different values, depending on N_3 we either choose $X_3 = X_1$ or $X_3 = X_2$. Otherwise $X_3 = N_3$. Now, $\mathcal{S}_B$ differs from $\mathcal{S}_A$ only in the latter case:

$$\mathcal{S}_B = \begin{cases} X_1 = N_1 \\ X_2 = N_2 \\ X_3 = \quad (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} \\ \qquad + \qquad\quad (2 - N_3) \qquad\quad \cdot 1_{X_1 = X_2} \end{cases}$$

It can be checked that $\phi(\mathcal{S}_A, \mathcal{L}(\mathbf{N})) = \phi(\mathcal{S}_B, \mathcal{L}(\mathbf{N}))$, but the two models differ in some counterfactual questions: Suppose, we have seen a sample

$$(X_1, X_2, X_3) = (1, 0, 0)$$

and we are interested in the counterfactual question, what X_3 would have been if X_1 had been 0. From both $\mathcal{S}_A$ and $\mathcal{S}_B$ it follows that $N_3 = 0$, and thus the two SEMs "predict" different values for X_3 under a counterfactual change of X_1.

Note that in contrast to Example 1.5, the distribution is faithful with respect to the graph from the SEM (here: $X_1 \to X_3 \leftarrow X_2$). We are not aware of any reference that presents this example, although similar cases may have been looked at before.
On the other hand, the mapping ϕ is surjective if we allow for arbitrary functions f_i in the SEM (see Appendix A.2.1 for a proof)[1]:

Proposition 2.6 *Let $(\mathcal{G}, \mathcal{L}(\mathbf{X}))$ be a graphical model. Then there exists an SEM $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$ with $\phi(\mathcal{S}, \mathcal{L}(\mathbf{N})) = (\mathcal{G}, \mathcal{L}(\mathbf{X}))$.*

2.5. Identifiability of the Graph Given the Distribution

We have seen similarities between graphical models and structural equation models. Regarding identifiability, they are quite different. Let us first formulate the exact problem statement:

Problem 2.7 [infinite sample case] Suppose we are given a joint distribution $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ from a graphical model (or from an SEM) with (unknown) graph $\mathcal{G}_0$. Can we recover the graph $\mathcal{G}_0$?

By first considering graphical models we easily see that the answer to this problem is negative: The joint distribution $\mathcal{L}(\mathbf{X})$ is certainly Markov with respect to a lot of different graphs, e.g. to all fully connected acyclic graphs. Thus, there are many possible graphical models $(\mathcal{G}, \mathcal{L}(\mathbf{X}))$ for the same $\mathcal{L}(\mathbf{X})$. Using ϕ^{-1}, we see that the same is true for SEMs.
What can be done to overcome this indeterminacy? The hope is that by using additional assumptions one obtains sets of restricted models $\mathcal{SM}_{\text{restr}} \subset \mathcal{SM}$ and $\mathcal{GM}_{\text{restr}} \subset \mathcal{GM}$, in which we can identify the graph from the joint distribution. In our opinion, it is precisely here, where the difference between graphical and functional models becomes apparent. We think that it is easier to find "natural" restrictions for SEMs than for GMs. Below we will discuss restricted graphical models,

[1] A similar but in our opinion weaker statement than Proposition 2.6 can be found in [Druzdzel and van Leijen, 2001, Janzing and Schölkopf, 2010].

where we additionally assume faithfulness, that lead to an identifiability of the Markov equivalence class of the true DAG and restricted SEMs, that even lead to an identifiability of the unique DAG.

2.5.1. Identifiability in GMs

Some causal inference methods (see Sections 3.1 and 3.2) assume that $\mathcal{L}(\mathbf{X})$ is faithful with respect to the true graph $\mathcal{G}_0$, which further relates the joint distribution with the graph structure. Faithfulness means that each conditional independence found in $\mathcal{L}(\mathbf{X})$ is implied by the Markov condition (see Definition 2.1). If faithfulness holds, it is apparent that one can obtain the Markov equivalence graph of the true graph $\mathcal{G}_0$. The joint distribution $\mathcal{L}(\mathbf{X})$ satisfies a set of conditional independences that is exactly encoded in each graph $\mathcal{G}$ that is Markov equivalent to $\mathcal{G}_0$. If $\mathcal{G}$ is not Markov equivalent to $\mathcal{G}_0$, then there is at least one conditional independence in $\mathcal{G}$ that is not in $\mathcal{G}_0$ or vice versa. But then $\mathcal{L}(\mathbf{X})$ cannot be faithful with respect to $\mathcal{G}$. Given $\mathcal{L}(\mathbf{X})$ we can obtain the Markov equivalence class of $\mathcal{G}_0$ by finding all conditional independences in $\mathcal{L}(\mathbf{X})$.

Proposition 2.8 *If $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to the graph $\mathcal{G}_0$, the Markov equivalence class of $\mathcal{G}_0$ is identifiable from the joint distribution $\mathcal{L}(\mathbf{X})$.*

The Markov equivalence class may still be large [cf. Andersson et al., 1997] and the DAG $\mathcal{G}_0$ itself is not identifiable. The method briefly described in Section 3.1 shows how to reconstruct the graph from the set of conditional independence statements. It further tries to avoid checking all possible conditional independences in $\mathcal{L}(\mathbf{X})$.

2.5.2. Identifiability in SEMs

SEMs show us, what else is achievable under a different type of assumptions. First, consider only SEMs with linear functions and normally distributed noise variables. It can be shown that for each graph in the same Markov equivalence there is an SEM that leads to exactly the same $\mathcal{L}(\mathbf{X})$. (Heckerman and Geiger [1995, Assumption 4] observe that this result is obtained by combining [Shachter and Kenley, 1989, Theorem 1] and [Chickering, 1995, Theorem 2].) Recently,

however, it has been shown that this case is exceptional in the following sense: If we consider linear functions and non-Gaussian noise, one can identify the single correct DAG [Shimizu et al., 2006]. Clearly, assuming linear relationships is not suitable for many applications. In this thesis we extend the results in several ways and obtain the linear non-Gaussian model as a special case. We will show that if one restricts the functions to be additive in the noise component and excludes the linear Gaussian case, as well as a few other function-noise-input combinations (see Sections 4 and 6), $\mathcal{G}_0$ is identifiable from $\mathcal{L}(\mathbf{X})$. In Section 5, we show a similar result for discrete variables. Section 7 investigates a third direction of moving away from the linear Gaussian case. When all functions are linear, and the normally distributed noise variables share a common variance σ^2, we again obtain identifiability. For our results faithfulness is not required, but instead, we use causal minimality (see Definition 2.1 and Remark 7.2). Thus, many of this thesis' main results contribute to the question of identifiability in structural equation models.

2.6. When Faithfulness does not hold

Let us start with a distribution $\mathcal{L}(\mathbf{X})$ and consider the sets

$$\begin{aligned}\mathbf{G}(\mathcal{L}(\mathbf{X})) &:= \{\mathcal{G} \text{ DAG on } \mathbf{X} \,|\, \mathcal{L}(\mathbf{X}) \text{ is Markov wrt } \mathcal{G}\}\\ \mathbf{G}_F(\mathcal{L}(\mathbf{X})) &:= \{\mathcal{G} \text{ DAG on } \mathbf{X} \,|\, \mathcal{L}(\mathbf{X}) \text{ is Markov and faithful wrt } \mathcal{G}\}\,.\end{aligned}$$

$\mathbf{G}(\mathcal{L}(\mathbf{X}))$ contains several elements, for example, all fully connected graphs. If $\mathcal{L}(\mathbf{X})$ is faithful with respect to some DAG, then $\mathbf{G}_F(\mathcal{L}(\mathbf{X}))$ contains exactly all members of one Markov equivalence class. In Example 2.4 we found $\mathbf{G}(\mathcal{L}(\mathbf{X})) = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_1^1, \mathcal{G}_1^2, \mathcal{G}_1^3, \mathcal{G}_1^4, \mathcal{G}_1^5\}$, where $\mathcal{G}_1^i$ together with $\mathcal{G}_1$ are all fully connected graphs with three variables. Without further assumptions, no causal inference method is able to infer whether $\mathcal{G}_1$ or $\mathcal{G}_2$ is correct (and both could be). Example 2.4 suggests to choose the subset $\mathbf{G}_F(\mathcal{L}(\mathbf{X}))$ of all graphs with respect to which $\mathcal{L}(\mathbf{X})$ is faithful. Note, however, that this procedure does not work in general: For some distributions $\mathcal{L}(\mathbf{X})$, there is no DAG $\mathcal{G}$, such that $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to $\mathcal{G}$ and $\mathbf{G}_F(\mathcal{L}(\mathbf{X})) = \emptyset$.

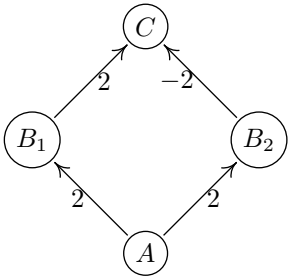


Figure 2.2.: Graph that is used to generate the distribution from Example 2.9.

Example 2.9 We generate a distribution for random variables A, B_1, B_2 and C from a linear Gaussian SEM according to the graph and coefficients shown in Figure 2.2. We find that

$$\begin{aligned} A &\perp\!\!\!\perp C \\ A &\perp\!\!\!\perp C \,|\, \{B_1, B_2\} \\ B_1 &\perp\!\!\!\perp B_2 \,|\, A \end{aligned}$$

A graph with respect to which $\mathcal{L}(A, B_1, B_2, C)$ is faithful and Markov has to satisfy

- no directed path between A and C
- no edge between B_1 and B_2
- all other edges should be included (thus the skeleton is the same as above)
- no v-structures at A, B_1 or B_2

Clearly, this is impossible.

Instead of taking the graphs, with respect to which $\mathcal{L}(\mathbf{X})$ is faithful, van de Geer and Bühlmann [2012] exploit a different route: they choose the subset of all graphs in $\mathbf{G}$ with the smallest number of edges. In a way, this follows the idea of Occam's razor by preferring the models that explain the data with the fewest number of parameters. As far

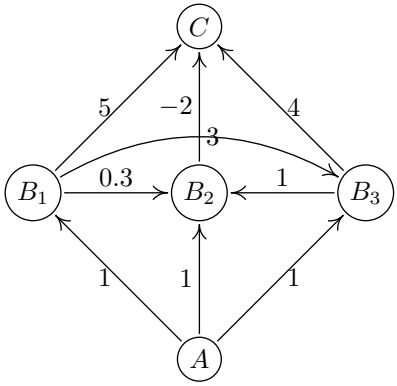


Figure 2.3.: Graph $\mathcal{G}_1$ used to generate the joint distribution of Example 2.10.

as we know, however, it is unsolved, how this set can be characterized. The following example shows that the graphs belonging to this set do not have to be Markov equivalent.

Example 2.10 In this example, we consider five variables, namely $\mathbf{X} = (A, B_1, B_2, B_3, C)$. Again we generate their distribution with a linear Gaussian SEM with structure and coefficients shown in Figure 2.3 and unit variances for the noise variables, i.e. $\mathbf{var}_{\mathcal{G}_1}(N_X) = 1$ for all $X \in \mathbf{X}$. In $\mathcal{L}(\mathbf{X})$ we find the independence constraints

$$A \perp\!\!\!\perp C \,|\, \{B_1, B_2, B_3\} \tag{2.2}$$

$$A \perp\!\!\!\perp C \,|\, \{B_1\} \tag{2.3}$$

It turns out that the obtained distribution can also be generated by an SEM with structure shown in Figure 2.4. The coefficients and noise variances for the SEM with graph $\mathcal{G}_2$ can be computed analytically from the coefficients in $\mathcal{G}_1$ using the covariance matrix of the distribution. In Figure 2.4 we show rounded values for the coefficients. For the variances use $\mathbf{var}_{\mathcal{G}_2}(N_A) = 1$, $\mathbf{var}_{\mathcal{G}_2}(N_{B_1}) = 1$, $\mathbf{var}_{\mathcal{G}_2}(N_{B_2}) = 0.2$, $\mathbf{var}_{\mathcal{G}_2}(N_{B_3}) = 0.5556$ and $\mathbf{var}_{\mathcal{G}_2}(N_C) = 9$.

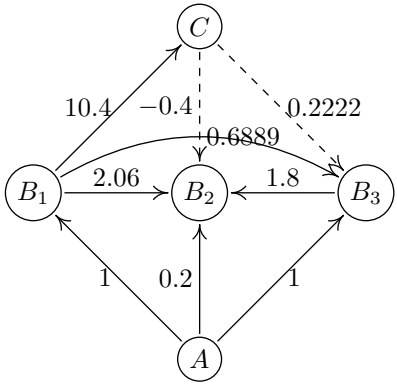


Figure 2.4.: The distribution from Example 2.10 can also be generated by an SEM from this graph $\mathcal{G}_2$ (the dashed arrows are different from $\mathcal{G}_1$). Both graphs have the minimal number of edges, but are not Markov equivalent.

The distribution is not faithful to any of the graphs. The first independence constraint (2.2) is encoded in $\mathcal{G}_1$ (Figure 2.3), the second one (2.3) in $\mathcal{G}_2$ (Figure 2.4). We cannot leave out any of the edges since this would introduce new independences that are not in $\mathcal{L}(\mathbf{X})$. Thus, $\mathcal{G}_1$ and $\mathcal{G}_2$ have a minimal number of edges, but they are not Markov equivalent.

2.7. Relating True Causal Graph and Distribution

2.7.1. Discussion of Assumptions

Because of the results we described in Section 2.5, we can now better understand the assumptions on the true causal graph that are made by popular causal inference methods (additional to Definition 1.3 or Proposition 1.8).

Assumption 2.11 *[Additional Assumption for Independence- and*

Score-based Methods] Let $\mathcal{G}_c$ be the true causal DAG of $X_1, \ldots, X_p$. Then additionally assume that $\mathcal{L}(\mathbf{X})$ is faithful *with respect to $\mathcal{G}_c$.*

Assumption 2.12 *[Additional Assumption for SEM-based Methods] Let $\mathcal{G}_c$ be the true causal DAG of $X_1, \ldots, X_p$. Then additionally assume that there is a* restricted SEM *with graph $\mathcal{G}_c$ and law $\mathcal{L}(\mathbf{X})$ and assume causal minimality.*

We are deliberately vague with the term "restricted SEM". Chapter 6 discusses identifiable functional model classes (IFMOC), which is one possibility to make this precise. Alternatively, Chapter 7 discusses Gaussian models with shared error variances. Both directions lead to identifiability: the graph can be inferred from the distribution.
We now go into a bit more detail about the comparison of the different conditions. Figure 2.5 represents their relationship graphically.

Markov condition Assume an SEM for $\mathbf{X} = (X_1, \ldots, X_p)$. Then Pearl [2009] shows in Theorem 1.4.1 that the joint distribution is Markov with respect to the corresponding graph. Recall also that for any $\mathcal{L}(\mathbf{X})$ that is Markov with respect to $\mathcal{G}$ we find an SEM with graph $\mathcal{G}$ (Proposition 2.6). Therefore, this part of the assumptions is common to conditional independence-based approaches and SEM-based approaches (see Figure 2.5).

Faithfulness Unfortunately, faithfulness in its full generality cannot be tested from the data [Zhang and Spirtes, 2008]. Further, a violation of faithfulness can lead to arbitrarily wrong DAGs. Zhang and Spirtes [2007] analyze the testability of faithfulness. They consider two special cases of faithfulness: *adjacency-faithfulness* (two adjacent variables are dependent conditional on any set of other variables) and *orientation-faithfulness* (a structure $X \to Y \leftarrow Z$ renders X and Z dependent given any set that contains Y and a structure $X - Y - Z$ with arrows other than above renders X and Z dependent given any set of variables that *does not* contain Y). They prove, for example, that under the assumption of Markov condition and adjacency-faithfulness, any violation of orientation-faithfulness is detectable. However, some violations of adjacency-faithfulness (e.g. $X \to Y \to Z$ and $X \to Z$ with X and Z independent) cannot be detected because they are faithful to

an alternative structure ($X \to Y \leftarrow Z$). To this kind of faithfulness they refer to as *triangular-faithfulness*.
One common argument in favor of faithfulness is the measure-zero argument. Suppose a Gaussian SEM with a given DAG and put any prior on the coefficients that is absolutely continuous with respect to the Lebesgue measure. Then with probability zero one obtains a distribution that is non-faithful with respect to that given DAG [Spirtes et al., 2000].
Robins et al. [2003] show that assuming the Markov condition and faithfulness is not enough to prove uniform consistency of the PC algorithm. Let us say $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to $\mathcal{G}$. Roughly speaking, the problem is that one can construct distributions that are Markov and (almost un-)faithful with respect to $\mathcal{G}'$, but $\mathcal{G}'$ is not Markov equivalent to $\mathcal{G}$ and still arbitrarily close to $\mathcal{L}(\mathbf{X})$. These other structures, however, may lead to completely different causal conclusions. For Gaussian SEMs, Zhang and Spirtes [2003] were able to prove uniform consistency assuming not only faithfulness, but *strong faithfulness*, which states that the smallest non-vanishing (partial) correlation needs to be bounded away from zero. Similar results were obtained in [Kalisch and Bühlmann, 2007] for high-dimensional data. Zhang and Spirtes [2003] further discuss that the measure-zero argument in its simple form does not hold for strong faithfulness anymore. For some families of graphs and linear Gaussian SEMs, Uhler et al. [2012] analyze how often distributions occur that are not strong faithful with respect to the corresponding graph (see also Experiment 1 in Section 11.3). To the best of our knowledge it is unknown what condition is required to prove uniform consistency of SEM-based methods. We believe that a condition is necessary that prevents the coefficients in a linear SEM, for example, to be close to zero. This is what we mean by the "β-min" condition in Figure 2.5.

Causal Minimality From our point of view, causal minimality is almost as natural as the Markov condition and in accordance with the intuitive understanding of a causal influence between two variables.

Restricted SEM assumption The assumptions made by the SEM-based approach can be violated in different ways. (1) The true data

generating process belongs to the considered class of SEMs (e.g., linear interactions and additive noise), but is not identifiable (e.g., the interactions are linear and all variables are Gaussian distributed). In this case the joint distribution allows several SEMs that lead to different graphs. Thus, our method described in Section 10.3 would output: "More than one graph possible, no answer proposed.". However, if we are willing to assume faithfulness, we can recover the Markov equivalence class by choosing the DAGs with the minimal number of edges and thus obtain asymptotically the same results as the PC algorithm[2]. (2a) The joint distribution does not belong to the considered class of SEMs. Here, the method would not be able to fit the data to any structure. Therefore the method would output: "Bad model fit. Try a different model class." (2b) The joint distribution does not allow for an instance of the considered SEM with respect to the true causal graph, but it does allow for an SEM with a different graph than the true causal graph (e.g., $X \to Y$ is the ground truth and the joint distribution does not allow an additive noise model from X to Y, but only from Y to X, see Chapter 4). This is the only situation, in which our method fails and gives a wrong answer. In Section 2.7.2 we argue, however, why we do not expect this case to happen in many situations.
Using restricted SEMs it is possible to define an asymmetry between time directions. In [Peters et al., 2009] we have shown that an ARMA time series $(X_t)_{t\in\mathbb{Z}}$ with a non-vanishing AR part satisfies an ARMA model in the opposite direction if and only if the noise follows a Gaussian distribution. Since for all recorded time series, the time ordering is known, this "application" constitutes an alternative way of examining the assumptions of restricted SEMs.

2.7.2. Independence of Cause and Mechanism

We now consider the special case of two variables X and Y. Let us assume that one is the cause of the other. If one is given samples only from the joint distribution $\mathcal{L}(X,Y)$, the question arises how one can break the symmetry between X and Y in

$$p_X(x) \cdot p_{Y|X}(y|x) = p_{X,Y}(x,y) = p_Y(y) \cdot p_{X|Y}(x|y)$$

[2]Proposition A.10 in the appendix proves this statement.

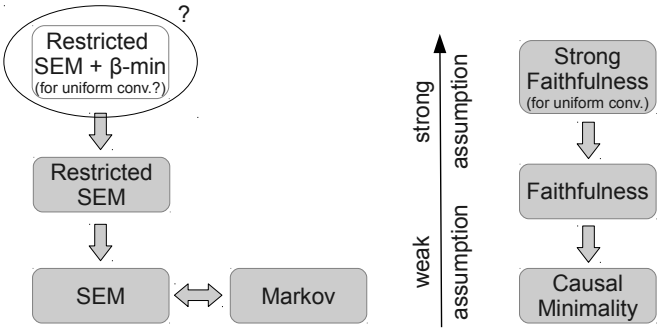


Figure 2.5.: Comparison of the different assumptions. Conditions imply all other conditions that are drawn below. Conditions to the left cannot easily be compared with conditions on the right of the axis.

Lemeire and Dirkx [2006], Janzing and Schölkopf [2010], Schölkopf et al. [2012, and references therein] suggest to use the following assumption that we will discuss a bit further below.

Assumption 2.13 *The mechanism from cause to effect, which we represent in a probabilistic setting with the conditional distribution* $\mathcal{L}(\textit{effect}|\,\textit{cause})$ *is "independent" of the cause, represented by the marginal distribution* $\mathcal{L}(\textit{cause})$.

It is obvious that the notion of *independence* of the marginal and the conditional has to be defined (see below). Assumption 2.13 is a crucial point relating causal properties of the true data generating process with statistics. It can be exploited for causal inference:

Causal Inference Principle 2.14 *Consider two random variables* X *and* Y. *When* $\mathcal{L}(X)$ *is "independent" of* $\mathcal{L}(Y\,|\,X)$ *but not vice versa, we infer* $X \to Y$.

Several formalizations of independence have been proposed, each concentrating on slightly different aspects. Let X denote the cause and Y the effect.

- Daniusis et al. [2010] investigates deterministic models $Y = f(X)$ for invertible functions f. They formulate a dependence between f and $\mathcal{L}(X)$ using *information geometry*.

- Janzing et al. [2010] and Zscheischler et al. [2011] consider a linear high-dimensional model $\mathbf{Y} = A \cdot \mathbf{X} + \mathbf{N}$, with $\mathbf{X}$ and $\mathbf{Y}$ being random vectors. They define independence between the matrix A and the covariance matrix $\Sigma_{\mathbf{X}}$ based on *free probability theory* which results in a condition that relates the traces of A, A^T and $\Sigma_{\mathbf{X}}$.

- Janzing and Schölkopf [2010] consider a slightly different setting where the observed data do not consist of i.i.d. samples but rather of single observations (strings). They then define *algorithmic mutual information* between the x and y using Kolmogorov complexity. For the probabilistic setting they propose to represent the marginal $\mathcal{L}(X)$ by a source S, a string that describes how to generate samples of X. The conditional $\mathcal{L}(X \mid Y)$ is a machine M that describes how to generate samples from Y for a given X. They define independence between M and S to be equivalent to zero algorithmic mutual information between the corresponding strings.

- The independent noise assumption in SEMs can also be seen as an instance of this principle. In the case of additive noise we have

$$Y = f(X) + N \qquad N \perp\!\!\!\perp X$$

  Here, $p_{Y|X}(y, x) = p_N(y - f(x))$ and the independence of the mechanism is thus represented by the noise variable. We will show that in the generic case this independence does not hold in the opposite direction.

- Certainly, there are many alternatives which may be investigated in future work.

Importantly, in the special case of two variables Assumption 2.13 and its interpretation described in Janzing and Schölkopf [2010] provide the

following justification for using restricted SEMs (apart from an argument based on Occam's razor). Consider two variables and assume that X is the cause and Y the effect. The inference principle 2.14 applied to additive noise models would fail if there is no additive noise model from X to Y, but there is one from Y to X. Janzing and Steudel [2010] use the concept of Kolmogorov complexity to show that this violates Assumption 2.13, provided that the complexity of $p(x)$ is not too small. Namely, it can only happen if the cause distribution $p(\text{cause}) = p(x)$ and the mechanism $p(\text{effect}|\text{cause}) = p(y|x)$ are matched in a precise way. Janzing and Steudel [2010] only consider the bivariate case, but we expect a similar statement to hold in general.

Note that Assumption 2.13 cannot be proved, even if we fix the notion of "independence". We can only check whether Assumption 2.13 applies to real world examples, for which most people would agree on the causal structure. In fact, we expect that it is actually false in some biological processes related to evolution: Over many years a biological system might adopt to the availability of certain resources. Nevertheless, we believe that it holds for many other processes and thus regard a further investigation as useful.
An argument *in favor* of the independence of cause and mechanism is the following idea related to Pearl's causal stability [Pearl, 2009]: On small time scales, we expect that even if the cause $P(\text{cause})$ is changed, the mechanism $P(\text{effect} \mid \text{cause})$ remains the same. That is why many scientists are concerned with the predicted changes in temperature and precipitation: the growth of plants, for example, still depends on its environment in the same way as before. This mechanism cannot be changed fast enough which may thus result in decreased crop yields [Parry et al., 2007].

Chapter 3.

Existing Algorithms

In this chapter we briefly introduce independence-based and score-based methods for causal inference, as well as a method based on SEMs with linear equations and non-Gaussian noise.

3.1. Independence-Based Methods

As always in this work we assume that the joint distribution is Markov with respect to the true graph (d-separation in the graph implies conditional independence in the distribution). If two variables, for example, are always dependent, no matter what other variables one conditions on, these two variables must be adjacent. Thus, properties of the joint distribution can help to infer parts of the graph structure. Conditional independence-based methods like the PC algorithm or FCI [Spirtes et al., 2000] additionally assume faithfulness (that means *all* conditional independences in the joint distribution are entailed by the Markov condition, cf. Definition 2.1). Then, one can use further reasonings like: If two variables are independent there is no collider-free path between them. Obviously, many more rules like this can be exploited.
Since both assumptions (Markov condition and faithfulness) put restrictions only on the conditional independences in the joint distribution, it is clear that these methods are not able to distinguish between two graphs that entail exactly the same set of (conditional) independences, i.e. between Markov equivalent graphs (see Figure 2.1 and Section 2.5.1). Since many Markov equivalence classes contain more than one graph, conditional independence-based methods usually leave some arrows undirected and cannot uniquely identify the true graph.

The first step of the PC algorithm determines the variables that are adjacent. One therefore has to test whether two variables are dependent given *any* other subset of variables. In the worst case (when there is no sparsity) this may result in conditional independence tests with conditioning sets of up to $p-2$ variables (where p is the number of variables in the graph). Although there is recent work on kernel-based conditional independence tests [Fukumizu et al., 2008, Zhang et al., 2011], such tests are difficult to perform in practice if one does not restrict the variables to follow a Gaussian distribution, for example [e.g. Bergsma, 2004].
From our perspective independence-based approaches potentially suffer from the following drawbacks: (1) We can identify the true DAG only up to Markov equivalence classes. (2) Conditional independence testing, especially with a large conditioning set, is difficult in practice. (3) The faithfulness condition in its general form cannot be tested given the data. (4) If faithfulness is violated we do not have any guarantees that the inferred graph(s) will be close to the original.

3.2. Score-Based Methods

Although the roots for score-based methods for causal inference may date back even further, we mainly refer to [Geiger and Heckerman, 1994, Heckerman, 1997, Chickering, 2002] and references therein. Given the data $\mathcal{D}$ from a vector $\mathbf{X}$ of variables, i.e. n i.i.d. samples, the idea is to assign a score $S(\mathcal{D}, \mathcal{G})$ to each graph $\mathcal{G}$ and search over the space of DAGs for the best scoring graph.

$$\hat{\mathcal{G}} := \underset{\mathcal{G} \text{ DAG over } \mathbf{X}}{\operatorname{argmax}} S(\mathcal{D}, \mathcal{G}) \tag{3.1}$$

There are several possibilities to define such a scoring function. Often a parametric model is assumed (e.g. linear Gaussian equations or multinomial distributions), which introduces a set of parameters $\theta \in \mathbf{\Theta}$.

From a **Bayesian** point of view, we may define priors $p_{pr}(\mathcal{G})$ and $p_{pr}(\theta)$ over DAGs and parameters and consider the log posterior as a score function (note that $p(\mathcal{D})$ is constant over all DAGs):

$$S(\mathcal{D}, \mathcal{G}) := \log p_{pr}(\mathcal{G}) + \log p(\mathcal{D}|\mathcal{G}) \,,$$

where $p(\mathcal{D}|\mathcal{G})$ is the marginal likelihood

$$p(\mathcal{D}|\mathcal{G}) = \int_{\theta \in \mathbf{\Theta}} p(\mathcal{D}|\mathcal{G}, \theta) \cdot p_{pr}(\theta)\,.$$

Here, $\hat{\mathcal{G}}$ is the mode of the posterior distribution, which is usually called maximum a posteriori (or MAP) estimator. Instead of a MAP estimator, one may be interested in the full posterior distribution over DAGs. In principle, even finer information as output is possible. One can average over all graphs to get a posterior of the hypothesis about the existence of a specific edge, for example.
In the case of parametric models, we call two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ *distribution equivalent* if for each parameter θ_1 there is a corresponding parameter θ_2, such that the distribution obtained from $\mathcal{G}_1$ in combination with θ_1 is the same as the distribution obtained from graph $\mathcal{G}_2$ with θ_2. It is known (see Section 2.5.2) that in the linear Gaussian case, for example, two graphs are distribution-equivalent if and only if they are Markov equivalent. One may therefore argue that $p(\mathcal{D}|\mathcal{G}_1)$ and $p(\mathcal{D}|\mathcal{G}_2)$ should be the same for Markov equivalent graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. Heckerman and Geiger [1995] discusses how to choose the prior over parameters accordingly.

From a more **frequentist** point of view, for each graph we may consider the maximum likelihood estimator $\hat{\theta}$. We may then define a different score function by the Bayesian Information Criterion (BIC)

$$S(\mathcal{D}, \mathcal{G}) = \log p(\mathcal{D}|\hat{\theta}, \mathcal{G}) - \frac{d}{2} \log n\,,$$

where n is the sample size. Chickering [2002] discusses, how these two approaches can be related using work by Haughton [1988].

Since the search space of all DAGs is growing super-exponentially in the number of variables [e.g. Chickering, 2002], greedy search algorithms is applied to solve Equation (3.1): at each step there is a candidate graph and a set of neighboring graphs. For all these neighbors one computes the score and considers the best-scoring graph as the new candidate. If none of the neighbors obtains a better score, the search procedure terminates (not knowing whether one obtained only a local optimum).

Clearly, one therefore has to define a neighborhood relation. Starting from a graph $\mathcal{G}$, we may define all graphs as neighbors from $\mathcal{G}$ that can be obtained by removing, adding or reversing one edge. In the linear Gaussian case, for example, one cannot distinguish between Markov equivalent graphs. It turns out that in those cases it is beneficial to change the search space to Markov equivalence classes instead of DAGs. The greedy equivalence search (GES) [Chickering, 2002] starts with the empty graph and consists of two-phases. In the first phase, edges are added until a local maximum is reached; in the second phase, edges are removed until a local maximum is reached, which is then given as an output of the algorithm.

3.3. Linear Non-Gaussian Additive Models

Although the work introduced by Shimizu et al. [2006], Kano and Shimizu [2003] covers the general case, the idea is maybe best understood in the case of two variables:

Example 3.1

$$Y = \phi X + N, \quad N \perp\!\!\!\perp X\,,$$

where X and N are normally distributed. It is easy to check that

$$X = \tilde{\phi} Y + \tilde{N}, \quad \tilde{N} \perp\!\!\!\perp Y\,.$$

with $\tilde{\phi} = \frac{\phi \mathbf{var}(X)}{\phi^2 \mathbf{var}(X) + \sigma^2} \neq \frac{1}{\phi}$ and $\tilde{N} = X - \tilde{\phi} Y$ [e.g. Peters, 2008].

If we consider non-Gaussian noise, however, the structural equation model gets identifiable.

Proposition 3.2 *Let X and Y be two random variables, for which*

$$Y = \phi X + N, \quad N \perp\!\!\!\perp X,\ \phi \neq 0$$

holds. Then we can reverse the process, i.e. there exists $\psi \in \mathbb{R}$ and a noise $\tilde{N}$, such that

$$X = \psi Y + \tilde{N}, \quad \tilde{N} \perp\!\!\!\perp Y\,,$$

if and only if X and N are Gaussian distributed.

The proof [e.g. Peters, 2008, Theorem 2.10] can be based on the Darmois-Skitovich theorem [Skitovic, 1954, 1962, Darmois, 1953]. This result has been discovered before, even applied to more than two variables. Shimizu et al. [2006] prove it using Independent Component Analysis (ICA) [Comon, 1994, Theorem 11], which itself is proved using the Darmois-Skitovic theorem.

Theorem 3.3 *[Shimizu et al. [2006]] Assume an SEM with graph* $\mathcal{G}_0$

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j \,, \qquad j = 1, \ldots, p \tag{3.2}$$

where all N_j *are jointly independent and non-Gaussian distributed. Additionally, for each* $j \in \{1, \ldots, p\}$ *we require* $\beta_{jk} \neq 0$ *for all* $k \in \mathbf{PA}_j^{\mathcal{G}_0}$*. Then, the graph* $\mathcal{G}_0$ *is identifiable from the joint distribution.*

The authors call this model a linear non-Gaussian acyclic model (LiNGAM) and provide a practical method based on ICA that can be applied to a finite amount of data. Later, an improved version of this method has been proposed in [Shimizu et al., 2011].

Chapter 4.

Continuous Bivariate Additive Noise Models

4.1. Introduction

In this chapter we show that nonlinearities can play a role quite similar to that of non-Gaussianity (Proposition 3.2): When causal relationships are nonlinear it typically helps break the symmetry between two observed variables and allows the identification of cause and effect. As Friedman and Nachman [2000] have pointed out, non-invertible functional relationships between the observed variables can provide clues to the generating causal model. However, we show that the phenomenon is much more general; for nonlinear models with additive noise *almost any* nonlinearities (invertible or not) will typically yield identifiable models.

In the next section, we start by defining the family of models under study, and then, in Section 4.3 we give theoretical results on the identifiability of these models from non-interventional data. We describe a practical method for inferring the generating model from a sample of data vectors later in Section 10.1, and show its utility in simulations and on real data in Section 11.1.

4.2. Model Definition

We assume that the observed data have been generated from a SEM with additive noise. Since we only consider the case of two variables X and Y in this chapter, the SEMs with corresponding graph $X \to Y$

are of the form

$$Y = f(X) + N \tag{4.1}$$

with $N \perp\!\!\!\perp X$, whereas the SEMs with corresponding graph $Y \to X$ can be written as

$$X = g(Y) + \tilde{N}\,, \tag{4.2}$$

with $\tilde{N} \perp\!\!\!\perp Y$. Here, we allow f and g to be possibly different arbitrary functions and the noise variables are assumed to be absolutely continuous with respect to the Lebesgue measure.
In this chapter we do not need to assume causal minimality (Definition 2.1) and do not have to explicitly exclude SEMs that result in the empty graph. Assume $Y = f(X) + N$, with $f(x) = x^2$ for $x \in A$, which is a small set satisfying $p_X(x) = 0$ for all $x \in A$ and $f(x) \equiv c$ otherwise. Then $X \perp\!\!\!\perp Y$ and one cannot distinguish this SEM from an SEM that results in the empty graph. This example violates causal minimality. In the case of two variables, we find a dependence between X and Y, whenever causal minimality is satisfied (which is not true for more than two variables). Theorem 4.1 as stated is correct and thus has to contain all non-faithful distributions.
Our data consist of a number of pairs (X_i, Y_i) sampled independently from the joint distribution $\mathcal{L}(X, Y)$. In the following section we discuss theoretical identifiability and tackle the practical case of a finite-size data sample in Section 10.1.

4.3. Identifiability

We already know that we cannot identify the graph from the joint distribution for all considered SEMs since they contain the non-identifiable linear Gaussian case. In the following we develop a theoretical result that gives necessary conditions for such a non-identifiable situation. We will see that these conditions are quite strong. Figure 4.1 illustrates the basic identifiability principle for the two-variable model. Denoting the two variables X and Y, we are considering the generative model $Y = f(X) + N$ where X and N are both Gaussian and statistically independent. In panel **(a)** we plot the joint density $p(x, y)$ of the observed variables, for the linear case of $f(x) = x$. As a trivial consequence of the model, the conditional density $p(y \,|\, x)$ has identical shape for all

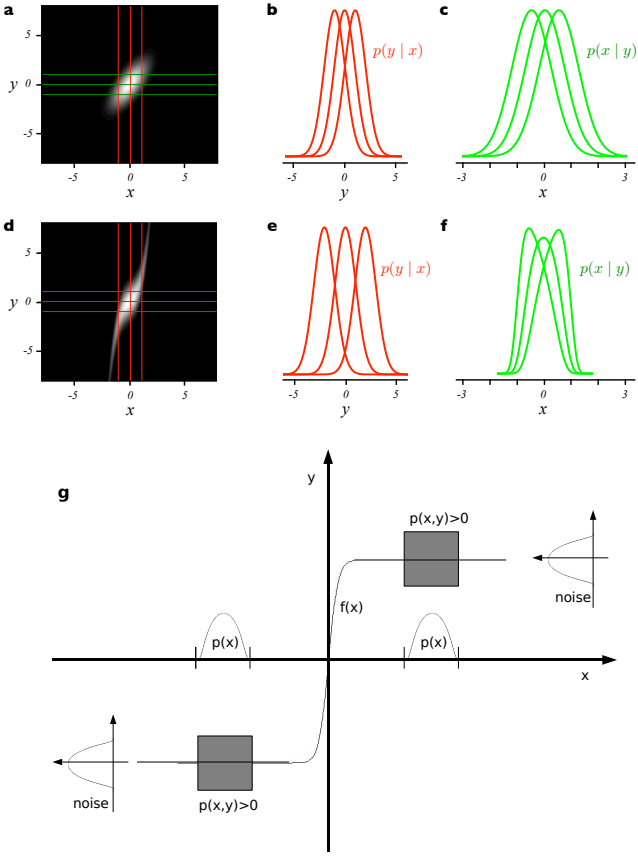


Figure 4.1.: Identification of causal direction based on constancy of conditionals. See main text for a detailed explanation of (a)–(f). (g) shows a non-identifiable example of a joint density $p(x, y)$; its support is given by the two gray squares. The input distribution p_X, the noise distribution p_N and f can in fact be chosen such that the joint density is symmetrical with respect to the two variables, i.e. $p(x, y) = p(y, x)$, making it obvious that there will also be a valid backward model.

values of x and is simply shifted by the function $f(x)$; this is illustrated in panel **(b)**. In general, there is no reason to believe that this relationship would also hold for the conditionals $p(x \mid y)$ for different values of y but, as is well known, for the linear–Gaussian model this is actually the case, as illustrated in panel **(c)**. Panels **(d-f)** show the corresponding joint and conditional densities for the corresponding model with a nonlinear function $f(x) = x + x^3$. Notice how the conditionals $p(x \mid y)$ look different for different values of y, indicating that a reverse causal model of the form $X = g(Y) + \tilde{N}$ (with Y and $\tilde{N}$ statistically independent) would not be able to fit the joint density. As we will show in this section, this will in fact *typically* be the case, however, not always.
To see the latter, we first show that there exist models other than the linear–Gaussian and the independent case which admit both a forward $X \to Y$ and a backward $X \leftarrow Y$ model. Panel **(g)** of Figure 4.1 presents a nonlinear functional model with $Y = f(X) + N$ where f is nonlinear, with additive non-Gaussian noise and non-Gaussian input distributions that nevertheless admits a backward model. The supports of the densities $p_X(x)$ and $p_N(n)$ are compact regions, and the function f is constant on each connected component of the support of p_X. The functions and probability densitities can be chosen to be (arbitrarily many times) differentiable.
Note that the example of panel **(g)** in Figure 4.1 is somewhat artificial: p has compact support, and x, y are independent inside the connected components of the support. Roughly speaking, the nonlinearity of f does not matter since it occurs where p is zero — an artifical situation which is avoided by the requirement that from now on, we will assume that all probability densities are strictly positive. Moreover, we assume that all functions (including densities) are three times differentiable. In this case, the following theorem shows that for generic choices of f, $p_X(x)$, and $p_N(n)$, there exists no backward model.

Theorem 4.1 *Let the joint probability density of X and Y be given by*

$$p(x, y) = p_N(y - f(x))p_X(x)\,, \tag{4.3}$$

where p_N, p_X are probability densities on $\mathbb{R}$. If there is a backward model of the same form, i.e.,

$$p(x, y) = p_{\tilde{N}}(x - g(y))p_Y(y)\,, \tag{4.4}$$

then, with $\nu := \log p_N$ *and* $\xi := \log p_X$*, the triple* (f, p_X, p_N) *must satisfy the following differential equation for all* x, y *with* $\nu''(y - f(x))f'(x) \neq 0$*:*

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'}, \tag{4.5}$$

where we have skipped the arguments $y - f(x)$*,* x*, and* x *for* ν*,* ξ*, and* f *and their derivatives, respectively. Moreover, if for a fixed pair* (f, ν) *there exists* $y \in \mathbb{R}$ *such that* $\nu''(y - f(x))f'(x) \neq 0$ *for all but a countable set of points* $x \in \mathbb{R}$*, the set of all* p_X *for which* p *has a backward model is contained in a* 3*-dimensional affine space.*

Loosely speaking, the statement that the differential equation for ξ has a 3-dimensional space of solutions (while a priori, the space of all possible log-marginals ξ is infinite dimensional) amounts to saying that in the generic case, our forward model cannot be inverted.
A simple corollary is that if both the marginal density $p_X(x)$ and the noise density $p_N(y - f(x))$ are Gaussian then the existence of a backward model implies linearity of f:

Corollary 4.2 *Assume that* $\nu''' = \xi''' = 0$ *everywhere. If a backward model exists, then* f *is linear.*

The proofs of Theorem 4.1 and Corollary 4.2 are provided in the appendix (Section A.3).
Finally, we note that even when f is linear and p_N and p_X are non-Gaussian, although a *linear* backward model has previously been ruled out Shimizu et al. [2006], there exist special cases where there is a *nonlinear* backward model with independent additive noise. One such case is when $f(x) = -x$ and p_X and p_N are Gumbel distributions: $p_X(x) = \exp(-x - \exp(-x))$ and $p_N(n) = \exp(-n - \exp(-n))$. Then taking $p_Y(y) = \exp(-y - 2\log(1 + \exp(-y)))$, $p_{\tilde{n}}(\tilde{n}) = \exp(-2\tilde{n} - \exp(-\tilde{n}))$ and $g(y) = \log(1 + \exp(-y))$ one obtains $p(x, y) = p_N(y - f(x))p_X(x) = p_{\tilde{n}}(x - g(y))p_Y(y)$.

Chapter 5.

Discrete Bivariate Additive Noise Models

5.1. Introduction

Independence-based methods are not able to distinguish between $X \to Y$ and $Y \to X$. For two continuous variables we have seen that $\mathcal{L}(X, Y)$ can only admit an additive noise model from X to Y

$$Y = f(X) + N\,, \tag{5.1}$$

and from Y to X only if the triple f and the densities of X and noise N satisfy a very specific differential equation. We say the model is identifiable in the "generic case". (In the remainder of this section we will use "genericness" in the meaning of "there are almost no exceptions").
For discrete variables, Sun et al. [2008] propose a method to measure the complexity of causal models via a Hilbert space norm of the logarithm of conditional densities and prefer models that induce smaller norms. Sun et al. [2006] fit joint distributions of cause and effect with conditional densities whose logarithm is a second order polynomial (up to the log-partition function) and show that this often makes causal directions identifiable, for example, when some or all variables are discrete. For discrete variables, several Bayesian approaches [Heckerman, 1997] are also applicable, but the construction of good priors are challenging and as we have mentioned in Section 3.2 often the latter are designed such that Markov equivalent DAGs still remain indistinguishable.
Here, we extend the model in Equation (5.1) to the discrete case in two different ways: (A) If X and Y take values in $\mathbb{Z}$ (the support may

be finite, though) ANMs can be defined analogously to the continuous case. (B) If X and Y take only finitely many values we can also define ANMs by interpreting the $+$ sign as an addition in the finite ring $\mathbb{Z}/m\mathbb{Z}$. We propose to apply this method to variables where the cyclic structure is appropriate (e.g., the direction of the wind after discretization, day of the year, season). Remark 5.2 in Section 5.2.2 describes how the second model can also be applied to structureless sets; this may be helpful whenever the random variables are categorical and when these categories do not inherit any kind of ordering (e.g. different treatments of organisms or phenotypes). In the following section we refer to (A) by *integer models* and to (B) by *cyclic models*.
We adopt the causal inference method from above: If there is an ANM from X to Y, but not vice versa, we propose that X is causing Y (more details in Section 5.2). Such a procedure is sensible if there are only few instances, in which there are ANMs in both directions. If, for example, all ANMs from X to Y also allow for an ANM from Y to X, we could not draw any causal conclusions at all. In Section 5.3 we show that these *reversible* cases are very rare and thereby answer this theoretical question.
For a practical causal inference method we have to test whether the data admit an ANM. We thus have to perform a discrete regression. But since in the discrete case regularization of the regression function is not necessary (given that there is a sufficient amount of data), in principle we could check all possible functions and test whether they result in independent residuals. This is highly intractable, of course, and we propose an efficient procedure that proved to work well in practice (Section 10.2).
In Section 5.2 we extend the concept of ANMs to discrete random variables and show the corresponding identifiability results in Section 5.3. In Section 10.2 we introduce an efficient algorithm for causal inference on finite data, for which we show experimental results in Section 11.2. Section A.4 contains the proofs.

5.2. Model Definition

Now we precisely define additive noise models in the case of discrete random variables. For simplicity we denote $p(x) = \mathbb{P}(X = x)$,

$q(y) = \mathbb{P}(Y = y)$, $n(l) = \mathbb{P}(N = l)$ and $\tilde{n}(k) = \mathbb{P}(\tilde{N} = k)$ and $\operatorname{supp} X$ is defined as $\operatorname{supp} X := \{k \,|\, p(k) > 0\}$.

5.2.1. Integer Models

Assume that X and Y are two random variables taking values in $\mathbb{Z}$ (their distributions may have finite support). We say that there is an additive noise model (ANM) from X to Y if there is a function $f : \mathbb{Z} \to \mathbb{Z}$ and a noise variable N such that the joint distribution $\mathcal{L}(X, Y)$ allows to write

$$Y = f(X) + N \text{ and } N \perp\!\!\!\perp X\,.$$

Furthermore we require $n(0) \geq n(j)$ for all $j \neq 0$. This does not restrict the model class, but is due to a freedom we have in choosing f and N: If $Y = f(X)+N$, $N \perp\!\!\!\perp X$, then we can always construct a new function f_j, such that $Y = f_j(X) + N_j$, $N_j \perp\!\!\!\perp X$ by choosing $f_j(i) = f(i) + j$ and $n_j(i) = n(i + j)$.
Such an ANM is called *reversible* if there is also an ANM from Y to X, i.e. if it satisfies ANMs in both directions.

5.2.2. Cyclic Models

We can extend ANMs to random variables that inherit a cyclic structure and therefore take values in a periodic domain. Random variables are usually defined as measurable maps from a probability space into the real numbers. Thus, we first make the following definition

Definition 5.1 Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbf{Z}/m\mathbf{Z}$ is called an m-cyclic random variable if $X^{-1}(k) \in \mathcal{H}$ $\forall k \in \mathbf{Z}/m\mathbf{Z}$. All other concepts of probability theory (like distributions and expectations) can be constructed analogously to the well-known case, in which X takes values in $\{0, \ldots, m-1\}$.

Let X and Y be m- and $\tilde{m}$-cyclic random variables, respectively. We say that Y satisfies an ANM from X to Y if there is a function $f : \mathbf{Z}/m\mathbf{Z} \to \mathbf{Z}/\tilde{m}\mathbf{Z}$ and an $\tilde{m}$-cyclic noise N such that

$$Y = f(X) + N \text{ and } N \perp\!\!\!\perp X\,.$$

Again we require $n(0) \geq n(j)$ for all $j \neq 0$ and call this model *reversible* if there is a function $g : \mathbf{Z}/\tilde{m}\mathbf{Z} \to \mathbf{Z}/m\mathbf{Z}$ and an m-cyclic noise $\tilde{N}$ such that $X = g(Y) + \tilde{N}$ and $\tilde{N} \perp\!\!\!\perp Y$.

Remark 5.2 Cyclic models are not restricted to random variables that take integers as values: Assume that X and Y take values in $\mathcal{A} := \{a_1, \ldots, a_m\}$ and $\mathcal{B} := \{b_1, \ldots, b_{\tilde{m}}\}$, which are structureless sets. Considering functions $f : \mathcal{A} \to \mathcal{B}$ and models with $\mathbb{P}(Y = b_j \,|\, X = a_i) = p$ if $b_j = f(a_i)$ and $(1-p)/(\tilde{m}-1)$ otherwise, is a special case of an ANM: Impose any cyclic structure on the data and use the additive noise $\mathbb{P}(N = 0) = p, \mathbb{P}(N = l) = (1-p)/(\tilde{m}-1)$ for $l \neq 0$.

5.2.3. Relations

The following two remarks are essential in order to understand the relationship between integer and cyclic models: (1) The difference between these two models manifests in the target domain. If we consider an ANM from X to Y it is important whether we put integer or cyclic constraints on Y (and thus on N). It does not make a difference, however, whether we consider the regressor X to be cyclic (with a cycle larger than $\#\text{supp}\, X$) or not. The independence constraint remains the same. (2) In the finite case ANMs with cyclic constraints are more general than integer models: Assume there is an ANM $Y = f(X) + N$, where all variables are taken to be non-cyclic and Y takes values between k and l, say. Then we still have an ANM $Y = f(X) + N$ if we regard Y to be $l - k + 1$-cyclic because $N \mod (l - k + 1)$ remains independent of X. It is possible, however, that $N \not\perp\!\!\!\perp X$, but $N \mod (l - k + 1) \perp\!\!\!\perp X$ (as shown in Example 5.7).

5.3. Identifiability

Whether or not there is an ANM between X and Y only depends on the form of the joint distribution $\mathcal{L}(X, Y)$. Let A be the set of all possible joint distributions and F its subset that allows an additive noise model from X to Y in the "forward direction", whereas B allows an ANM in the backward direction from Y to X (see Figure 5.1). Some trivial examples like $p(0) = 1, n(0) = 1$ and $f(0) = 0$ immediately show that

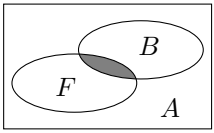


Figure 5.1.: How large is $F \cap B$?

there are joint distributions allowing ANMs in both directions, meaning $F \cap B \neq \emptyset$. But how large is this intersection? The proposed method would not be useful if we find out that F and B are almost the same sets. Then in most cases ANMs can be fit either in both directions or in none. Both, for ANMs with integer constraints and with cyclic constraints we identify the intersection $F \cap B$ and show that it is indeed a very small set. Imagine, we observe data from a natural process that allows an ANM in the causal direction. If we are "unlucky" and the data generating process happens to be in $F \cap B$, our method does not give wrong results, but answers "I do not know the answer".

5.3.1. Integer Models

Y or X has finite support

First we assume that either the support of X or the support of Y is finite. This already covers most applications. Figure 5.2 (the dots indicate a probability greater than 0) shows an example of a joint distribution that allows an ANM from X to Y, but not from Y to X. This can be seen easily at the "corners" $X = 1$ and $X = 7$: Whatever we choose for $g(0)$ and $g(4)$, the distribution of $\tilde{N} \,|\, Y = 0$ is supported only by one point, whereas $\tilde{N} \,|\, Y = 4$ is supported by 3 points. Thus $\tilde{N}$ cannot be independent of Y. Figure 5.3 shows a (rather non-generic) example that allows an ANM in both directions if we choose $p(a_i) = \frac{1}{36}, p(b_i) = \frac{2}{36}$ for $i = 1, \ldots, 4$ and $p(a_i) = \frac{2}{36}, p(b_i) = \frac{4}{36}$ for $i = 5, \ldots, 8$. We prove the following

Theorem 5.3 *Assume either X or Y has finite support. An ANM $X \to Y$ is reversible $\iff$ there exists a disjoint decomposition $\bigcup_{i=0}^{l} C_i = \operatorname{supp} X$, such that a) - c) are satisfied:*

a) The C_is are shifted versions of each other

$$\forall i \, \exists d_i \geq 0 \, : \, C_i = C_0 + d_i$$

and f is piecewise constant: $f\,|_{C_i} \equiv c_i \; \forall i$.

b) The probability distributions on the C_is are shifted and scaled versions of each other with the same shift constant as above: For $x \in C_i$, $\mathbb{P}(X = x)$ satisfies

$$\mathbb{P}(X = x) = \mathbb{P}(X = x - d_i) \cdot \frac{\mathbb{P}(X \in C_i)}{\mathbb{P}(X \in C_0)}.$$

c) The sets $c_i + \operatorname{supp} N := \{c_i + h \, : \, n(h) > 0\}$ are disjoint.

(Note that such a decomposition satisfying the same criteria also exists for $\operatorname{supp} Y$ by symmetry.) In the example of Figure 5.3 all a_i belong to C_0, all b_j to C_1 and $d_1 = 1$. As for the other theorems of this section the proof is provided in Section A.4. Its main point is based on the asymmetric effects of the "corners" of the joint distribution. In order to allow for an infinite support of X (or Y) we will thus generalize this concept of "corners".
Theorem 5.3 provides a full characterization of cases that allow for an ANM in both directions. Each of the conditions is very restrictive by itself, all conditions together describe a very small class of models: in almost all cases the direction of the model is identifiable. We have the following corollary:

Corollary 5.4 *Consider a discrete ANM from X, that takes values $x_1, \ldots, x_m$ ($m > 1$), to Y with a non-constant function f (otherwise X and Y are independent). Let the noise N take values from $N_{\min}$ to $N_{\max}$ and put any prior measure on the parameters $n(k)$ for $k = N_{\min}, \ldots, N_{\max}$ and $p(x_k), k = 1, \ldots, m$ that is absolutely continuous to the Lebesgue measure. If further $\min_{i,j \in \{1,\ldots,m\} \, : \, i \neq j} f(x_i) - f(x_j) \leq N_{\max} - N_{\min}$ we have the following statement: Only a parameter set of measure 0 admits an ANM from Y to X.*

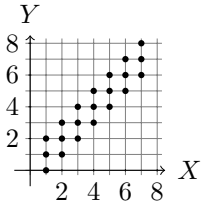


Figure 5.2.: This joint distribution satisfies an ANM only from X to Y.

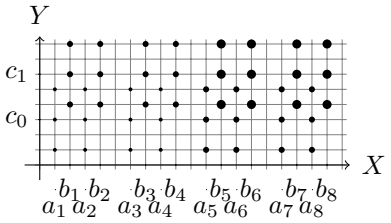


Figure 5.3.: Only carefully chosen parameters allow ANMs in both directions. (Radii correspond to probability values.)

X and Y have infinite support

Theorem 5.5 *Consider an ANM $X \to Y$ where both X and Y have infinite support. We distinguish between two cases*

a) **N has compact support:** $\exists m, l \in \mathbb{Z}$, *s.t.* $\operatorname{supp} N = [m, l]$. *Assume there is an ANM from X to Y and f does not have infinitely many infinite sets, on which it is constant. Then we have the following equivalence: The model is reversible if and only if there exists a disjoint decomposition $\bigcup_{i=0}^{\infty} C_i = \operatorname{supp} X$ that satisfies the same conditions as in Theorem 5.3.*

b) **N has entire $\mathbb{Z}$ as support:** $\mathbb{P}(N = k) > 0 \,\forall\, k \in \mathbb{Z}$. *Suppose X and Y are dependent and there is a reversible ANM $X \to Y$. Fix any $m \in \mathbb{Z}$. If f, $\mathbb{P}^N$ and $p(k)$ for all $k \geq m$ are known, then all other values $p(k)$ for $k < m$ are determined. That means even a small fraction of the parameters determine the remaining parameters.*

Note that the first case is again a complete characterization of all instances of a joint distribution, an ANM in both directions is conform with. The second case does not yield a complete characterization, but shows how restricted the choice of a distribution $\mathbb{P}^X$ is (given f and $\mathbb{P}^N$) that yields a reversible ANM.

5.3.2. Cyclic Models

Assume $Y = f(X) + N$ with $N \perp\!\!\!\perp X$. We will show that in the generic case the model is still not reversible, meaning there is no g and $\tilde{N}$, such that $X = g(Y) + \tilde{N}$ with $\tilde{N} \perp\!\!\!\perp Y$. However, as mentioned in Section 5.2.3, in finite domains this model class is larger than the class of integer models. We will see that correspondingly also the number of reversible cases increases.
Note that the model $Y = f(X) + N$ is reversible if and only if there is a function g, such that

$$p(x) \cdot n\big(y - f(x)\big) = q(y) \cdot \tilde{n}\big(x - g(y)\big) \;\forall x, y\,, \tag{5.2}$$

where $q(y) = \sum_{\tilde{x}} p(\tilde{x}) n(y - f(\tilde{x}))$ and $\tilde{n}(a) = p(g(\tilde{y}) + a) \cdot n(\tilde{y} - f(g(\tilde{y}) + a))/q(\tilde{y}) \quad \forall \tilde{y} : q(\tilde{y}) \neq 0$.

Non-Identifiable Cases

First, we give three (characteristic) examples of ANMs that are not identifiable. This restricts the class of situations in which identifiability can be expected. Figure 5.4 shows instances of Examples 1 and 2.

Example 5.6 Independent X and Y always admit an ANM from X to Y and from Y to X. We therefore have:

(i) If $Y = f(X) + N$ and $f(k) = \text{const}$ for all $k : p(k) \neq 0$, then the model is reversible.

(ii) If $Y = f(X) + N$ for a uniformly distributed noise N, then the model is reversible.

Proof. In both cases it X and Y are independent. Thus, $X = g(Y)+X$ with $g \equiv 0$ is a backward model. □

Example 5.7 If $Y = f(X) + N$ for a bijective and affine f and uniformly distributed X, then the model is reversible.

Proof. Since X is uniform and $f(x) = ax+b$ is bijective, Y is uniform, too. For $g(y) = f^{-1}(y)$ and $\tilde{n}(k) = n(b - f(k)) = n(y - f(g(y) + k))$ Equation (5.2) is satisfied. □

Example 5.8 We give two more examples of non-identifiable cases that show why an if-and-only-if characterization as in Theorem 5.3 is hard to obtain:

(i) Figure 5.5 (left) shows an example, where the sets on which f is constant neither satisfy condition c) nor are they shifted versions of each other.

(ii) The same holds for Figure 5.5 (right), this time even satisfying the additional constraint that $\mathbb{P}(N = 0) > \mathbb{P}(N = k)\,\forall k \neq 0$. Here, X is not uniformly distributed, either.

Identifiability Results

The counter examples from above already show that cyclic models are in some aspect more difficult than integer models and we thus do not

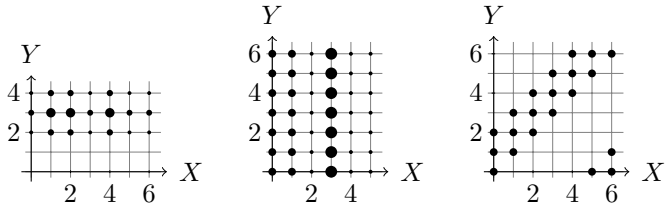


Figure 5.4.: These joint distributions allow ANMs in both directions. They are instances of Examples 1(*i*), 1(*ii*) and 2 (from left to right).

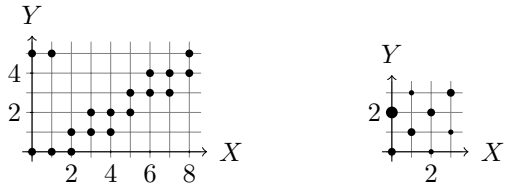


Figure 5.5.: These joint distributions allow ANMs in both directions. They are instances of Examples 3 (i) (left) and (ii) (right).

provide a full characterization of all reversible cases as we have done in the integer case. Nevertheless, we provide necessary conditions for reversibility, which is sufficient for our purpose.
Usually the distribution $n(l)$ (similar for $p(k)$) is determined by $\tilde{m}-1$ free parameters. As long as the sum remains smaller than 1, there are no (equality) constraints for the values of $n(0), \ldots, n(\tilde{m}-2)$. Only $n(\tilde{m}-1)$ is determined by $\sum_{l=0}^{\tilde{m}-1} n(l) = 1$. We show that in the case of a reversible ANM the number of free parameters of the marginal $n(l)$ is heavily reduced. The exact number of constraints depends on the possible backward functions g, but can be bounded from below by 2. Furthermore the proof shows that a "dependence" between values of p and n is introduced. Both of these constraints are considered to lead to non-generic models. That means for any *generic* choice of p and n we can only have an ANM in one direction.
Note further that $(\#\operatorname{supp} X \cdot \#\operatorname{supp} N)$ is the number of points (x, y) that have probability greater than 0. It must be possible to distribute these points equally to all points from $\#\operatorname{supp} Y$ in order to allow a backward ANM. Thus we have the necessary condition $\#\operatorname{supp} Y \mid (\#\operatorname{supp} X \cdot \#\operatorname{supp} N)$. (Here, $a \mid b$ denotes "a divides b", which we write if $\exists z \in \mathbb{Z} : b = z \cdot a$, and should not be confused with conditioning on a random variable.)

Theorem 5.9 *Assume $Y = f(X) + N$, $N \perp\!\!\!\perp X$ with non-uniform X (m-cyclic), Y ($\tilde{m}$-cyclic) and N ($\tilde{m}$-cyclic) and non-constant f.*

(i) There can only be an ANM from Y to X if

$$\#\operatorname{supp} Y \mid (\#\operatorname{supp} X \cdot \#\operatorname{supp} N)\,.$$

(ii) Assume that $\#\operatorname{supp} X = m, \#\operatorname{supp} N = \tilde{m}$. If there is an ANM from Y to X, at least one additional equality constraint is introduced to the choice of either p or n.

Again, the proof can be found in Section A.4.

5.3.3. Special Case: X and Y binary

We now investigate a special case, where X and Y are constrained to take binary values with probabilities $a := \mathbb{P}(X = 0, Y = 0), b := \mathbb{P}(X =$

$1, Y = 0), c := \mathbb{P}(X = 0, Y = 1)$ and $d := \mathbb{P}(X = 1, Y = 1)$. For this case we can compute a full characterization of reversible and irreversible ANMs. Therefore we assume the variables to be non-degenerate (i.e. $0 < \mathbb{P}(X = 0) = a + c < 1$ and $0 < \mathbb{P}(Y = 0) = a + b < 1$) and we use the following Lemma:

Lemma 5.10 *Let N and X be non-degenerate binary variables. Then*
$N \perp\!\!\!\perp X \Leftrightarrow \mathbb{P}(N = 1 \,|\, X = 0) = \mathbb{P}(N = 1 \,|\, X = 1)$.

The integer model is not very informative. The only two possibilities to form an ANM with integer constraints is to choose deterministic noise or a constant function f. Clearly, both cases lead to reversible ANMs. More interestingly, the results for the cyclic case are non-trivial:

1. f is constant.
   Here, X and Y are independent and the ANM is thus reversible (see Example 5.6(i)). Lemma 5.10 implies that $X \perp\!\!\!\perp N$ if and only if $\frac{c}{a+c} = \frac{d}{b+d}$. And this holds if and only if

   $$ad = bc$$

   (Here, neither of the parameters can be zero.)

2. f is non-constant.
   Without loss of generality let f be the identity function (we can always add an additive shift). This time we have $X \perp\!\!\!\perp N$ if and only if $\frac{c}{a+c} = \frac{b}{b+d}$, which is equivalent to

   $$ab = cd$$

   still assuming $a + c \neq 0 \neq b + d$.

Using symmetry it follows that there is an ANM from Y to X if and only if we have either $ac = bd$ or $ad = bc$.
We thus summarize (recall that only b and c or a and d can be zero at the same time):

- $ab = cd$ or $ad = bc$ leads to an ANM from X to Y.
- $ac = bd$ or $ad = bc$ leads to an ANM from Y to X.

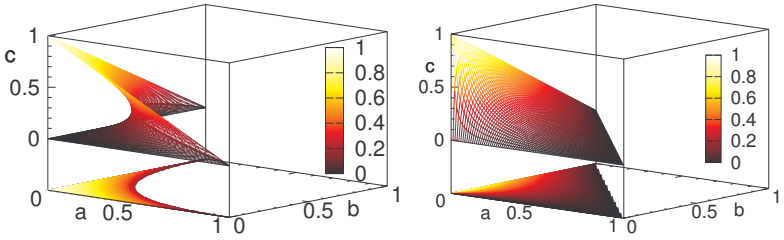


Figure 5.6.: For $X \not\perp\!\!\!\perp Y$ (both binary) these plots visualize the constraints of the joint distribution $\mathcal{L}(X, Y)$ in order to allow for an ANM: either from X to Y ($ab = cd$, left) or from Y to X ($ac = bd$, right). Note that the both surface are rotated versions of each other: the c-axis on the left corresponds to the b-axis on the right.

- $a = d$ and $b = c$ (this implies uniform X and Y) or $a = d = 0$ or $b = c = 0$ or $ad = bc$ leads to a reversible ANM.

This also fits with the theoretical result of Proposition A.1 in Section A.4: for bijective f and g (which is the only case that does not lead to independent X and Y) only uniformly distributed X and Y lead to reversible ANMs. Using $d = 1 - a - b - c$ one can plot these conditions as surfaces (see Figures 5.6 and 5.7). The models are represented by manifolds in a three-dimensional space.

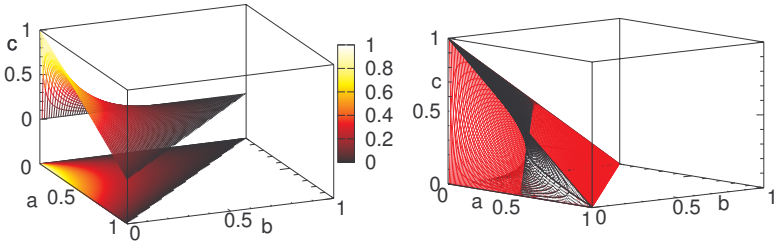


Figure 5.7.: These pictures characterize the joint distributions $\mathcal{L}(X, Y)$ that allow an ANM in both directions. This is fulfilled if both variables are independent ($ad = bc$, left) or (right) if $\mathcal{L}(X, Y)$ lies on the intersection of the $\text{ANM}_{X\to Y}$-surface (black) and the $\text{ANM}_{Y\to X}$-surface (red) from Figure 5.6: $b = c = 0$ corresponds to the a-axis, $a = d = 0$ and thus $c = 1 - b$ to the straight line between $(0, 0, 1)$ and $(0, 1, 0)$ and $a = d, b = c$ (ergo $c = 0.5 - a$) is represented by the intersection line between $(0.5, 0, 0)$ and $(0, 0.5, 0.5)$.

5.3.4. Mixed Models

With the results developed in the last two sections we can cover even models with mixed constraints if both variables have finite support. For the precise conditions of "usually" see Theorem 5.9 in Section 5.3.2.

$$
\begin{aligned}
& Y = f(X) + N,\ N \perp\!\!\!\perp X;\ X \text{ cyclic}, Y, N \text{ non-cyclic} \\
\overset{5.2.3}{\Rightarrow}\ & Y = f(X) + N,\ N \perp\!\!\!\perp X;\ X \text{ cyclic}, Y, N\ \tilde{m}\text{-cyclic} \\
\overset{\text{Thm } 5.9}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& \qquad X, \tilde{N} \text{ cyclic}, Y\ \tilde{m}\text{-cyclic} \\
\overset{5.2.3}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& \qquad X, \tilde{N} \text{ cyclic}, Y \text{ non-cyclic}
\end{aligned}
$$

And, conversely:

$$
\begin{aligned}
& Y = f(X) + N,\ N \perp\!\!\!\perp X;\ Y, N \text{ cyclic}, X \text{ non-cyclic} \\
\overset{5.2.3}{\Rightarrow}\ & Y = f(X) + N,\ N \perp\!\!\!\perp X;\ Y, N \text{ cyclic}, X\ m\text{-cyclic} \\
\overset{\text{Thm } 5.9}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& \qquad Y \text{ cyclic}, X, \tilde{N}\ m\text{-cyclic} \\
\overset{5.2.3}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& \qquad Y \text{ cyclic}, X, \tilde{N} \text{ non-cyclic}
\end{aligned}
$$

Acknowledgement

The authors want to thank Fabian Gieringer for collaborating on Section 5.3.3 and Joris Mooij, Kun Zhang and Stefan Harmeling for helpful comments.

Chapter 6.

From Bivariate to Multivariate Models

6.1. Introduction

Consider the case of two dependent random variables. Conditional independence-based methods (Section 3.1) cannot recover the graph since there is no (conditional) independence statement; $X \to Y$ and $Y \to X$ are Markov equivalent. Section 3.3 and Chapters 4 and 5 suggest the following procedure to tackle this problem: Whenever the joint distribution $\mathcal{L}(X, Y)$ allows an *additive noise model (ANM)* in one direction, i.e., there is a function f and a noise variable N, such that

$$Y = f(X) + N, \quad N \perp\!\!\!\perp X,$$

but not in the other, one infers the former direction to be the causal one (here: $X \to Y$). We have seen that under mild conditions (essentially some combinations of f, $\mathcal{L}(X)$ and $\mathcal{L}(N)$ have to be excluded) the model is identifiable. This means that whenever there is an ANM from X to Y the joint distribution does not allow for an ANM from Y to X. In this chapter we call these cases "bivariate identifiable". Another example of a bivariate identifiable model class are post-nonlinear models [Zhang and Hyvärinen, 2009].
Based on bivariate identifiability we define *Identifiable Functional Model Classes (IFMOCs)*, which we use to model distributions of more than two random variables. As a main result of this chapter we prove that whenever a data generating process belongs to an IFMOC, one can recover the corresponding graph from the joint distribution. To the best

of our knowledge this is the first identifiability statement of this kind that allows for nonlinear interactions.
Analogously to the two-variable case described above, practical methods for causal inference using ANMs have been suggested for the multivariate case [Hoyer et al., 2009, Zhang and Hyvärinen, 2009, Mooij et al., 2009, Tillman et al., 2010]: whenever a functional model (or SEM) with a certain graph models the data the method infer this graph as the causal graph. Our results fill a theoretical gap that has remained open so far: except for the linear case [Shimizu et al., 2006] the corresponding identifiability problem had not been solved yet.
What happens if the IFMOC assumption is not satisfied? We argue in Section 2.7.2 why we do not expect the true data generating process to belong to an IFMOC *only* for a different ordering of the variables. If one accepts this belief one can test whether the IFMOC assumption is valid: if it is not, one can either fit none or multiple models to the data. In order to exploit this deliberation, we provide an algorithm that outputs *all* structures that fit the data.
Section 6.2 defines the models we are looking at. Section 6.3 provides the identifiability results, Section 2.7.1 discusses the assumptions made by the model and the difference to independence-based and score-based methods. Section 10.3 provides an algorithm that identifies the causal graph given a data set and Section 11.3 contains experiments on artificial data. All proofs of this section are provided in tha appendix (Section 6).

6.2. Model Definition

First, we have to fix some notation. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a finite family of random variables and $\mathcal{G}$ be a graph over $\mathbf{X}$. In the remainder of this chapter (and the corresponding proofs in the appendix) we use the following notation: $p_X(x)$ denotes the pdf (or pmf) of a random variable X, $p_{\mathbf{S}}(x_{\mathbf{S}})$ denotes the joint pdf (or pmf) for a set of indices $\mathbf{S} \subset \{1, \ldots, p\}$. evaluated at the point $x_{\mathbf{S}}$. We will assume that $\mathcal{L}(\mathbf{X})$ is absolutely continuous with respect to either the Lebesgue measure or the counting measure (i.e., either we have a pdf or a pmf). Then $Y|_{X=x}$ is a RV that corresponds to the conditional density $p_{Y \mid X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$. We further identify the node i with the variable

X_i and the parents $\mathbf{PA}_i^{\mathcal{G}}$ with the variables $X_{\mathbf{PA}_i^{\mathcal{G}}}$.
First, we define *functional models* [e.g. Chapter 1.4 Pearl, 2009] that are also known as *Structural Equation Models*:

Definition 6.1 [$\mathcal{F}$-FMOC]

- p equations

$$X_i = f_i(X_{\mathbf{PA}_i}, N_i)\,, \qquad 1 \leq i \leq p$$

  with sets of variables $\mathbf{PA}_i \subseteq \{1, \ldots, p\} \setminus \{i\}$ and noise distributions $\mathcal{L}(N_i)$ are called a *functional model* if the $N_1, \ldots, N_p$ are jointly independent and the graph that is obtained by drawing arrows from all elements of $\mathbf{PA}_i$ to i (for each $i \in \{1, \ldots, p\}$) is acyclic.
- Given a set of functions

$$\mathcal{F} \subset \{f \mid f : \mathbb{R}^m \to \mathbb{R} \text{ for any } 2 \leq m \leq p\}$$

  we call a set of functional models a *functional model class with function class $\mathcal{F}$ ($\mathcal{F}$-FMOC)* if each of the functional models satisfies $f_i \in \mathcal{F}$ for all $i \in \{1, \ldots, p\}$ and induces $\mathcal{L}(\mathbf{X})$ that is absolutely continuous with respect to the Lebesgue measure or the counting measure.

Note that each functional model induces a unique joint distribution $\mathcal{L}(\mathbf{X})$.

We first concentrate on the bivariate case and consider the multivariate case in Definition 6.4. We have seen causal discovery methods that distinguish between cause and effect by means of the following observation. For some classes of bivariate functional models it has been shown that the structure of the model is in the "generic case" identifiable from the joint distribution: Consider, for example, only linear and additive functions $f(x, n) = a \cdot x + n$ and non-Gaussian noise. Then Shimizu et al. [2006] show that if $Y = f(X, N_Y)$ holds with $N_Y \perp\!\!\!\perp X$, one cannot find any function g such that $X = g(Y, N_X)$ with $N_X \perp\!\!\!\perp Y$. Thus, we will call the set of all triples $(f, \mathcal{L}(X), \mathcal{L}(N))$ of linear functions and non-Gaussian distributions *bivariate identifiable*. Hoyer et al. [2009], Peters et al. [2010] show a similar result for nonlinear additive

functions $f(x,n) = g(x)+n$, and Zhang and Hyvärinen [2009] for post-nonlinear models $f(x,n) = h\big(g(x)+n\big)$ with invertible h. Writing $\mathcal{F}_{|_2} := \{f \in \mathcal{F} \mid f : \mathbb{R}^2 \to \mathbb{R}\}$ and considering one-dimensional variables $(X, Y, N_X, N_Y \in \mathbb{R})$ we can now generalize these ideas and define:

Definition 6.2 [Bivariate Identifiable Set] Let $\mathcal{F}$ be a set of functions as above. We call a set $\mathcal{B} \subseteq \mathcal{F}_{|_2} \times \mathcal{P}_{\mathbb{R}} \times \mathcal{P}_{\mathbb{R}}$ containing combinations of functions $f \in \mathcal{F}_{|_2}$ and distributions $\mathcal{L}(X)$, $\mathcal{L}(N_Y)$ of input X and noise N_Y *bivariate identifiable in* $\mathcal{F}$ if

$$\begin{aligned} & (f, \mathcal{L}(X), \mathcal{L}(N_Y)) \in \mathcal{B} \text{ and } && Y = f(X, N_Y), N_Y \perp\!\!\!\perp X \\ \Rightarrow \quad & \nexists g \in \mathcal{F}_{|_2} : && X = g(Y, N_X), N_X \perp\!\!\!\perp Y \end{aligned}$$

holds. Additionally we require

$$f(X, N_Y) \not\perp\!\!\!\perp X \tag{6.1}$$

for all $(f, \mathcal{L}(X), \mathcal{L}(N_Y)) \in \mathcal{B}$ with $N_Y \perp\!\!\!\perp X$.

The first part of the definition requires that we cannot simultaneously fit both directions (one may think of $\mathcal{F}$ being the class of linear ANMs and $\mathcal{B}$ being all of those models, where input and noise are *not* jointly Gaussian). The left hand side of (6.1) corresponds to the effect, the right hand side to the cause. In the bivariate case one can imagine that we do not want them to be independent. We discuss this assumption below.
Note further that the function class needs to be restricted for the definition to be non-trivial, because for any joint distribution of (X, Y) we can find a function f and a noise $N_Y \perp\!\!\!\perp X$, such that $Y = f(X, N_Y)$ (Proposition 2.6). Proving that a set is bivariate identifiable is not trivial. The following lemma presents identifiability results that have been reported in literature and previous sections. In order to improve readability, we describe the classes and mention only the most important counter-examples. We denote all other exceptions by the sets $\tilde{B}_i$, which mostly contain constant functions and other, "non-generic" cases.

Lemma 6.3 *The following sets have been shown to be bivariate identifiable (again we have $X, N \in \mathbb{R}$):*

(i) linear ANMs: $\mathcal{F}_1 = \{f(x,n) = ax + n\}$

$$\mathcal{B}_1 = \{(f, \mathcal{L}(X), \mathcal{L}(N)) \,|\, (X,N) \textit{ not both Gaussian}\} \setminus \tilde{B}_1$$

(ii) discrete ANMs ($\tilde{m} \in \mathbb{N}$)*:* $\mathcal{F}_2 = \{f(x,n) \equiv \phi(x) + n \bmod \tilde{m}\}$

$$\mathcal{B}_2 = \{(f, \mathcal{L}(X), \mathcal{L}(N)) \,|\, (\phi, X) \textit{ not affine and uniform}\} \setminus \tilde{B}_2$$

(iii) nonlinear ANMs: $\mathcal{F}_3 = \{f(x,n) = \phi(x) + n\}$

$$\mathcal{B}_3 = \{(f, \mathcal{L}(X), \mathcal{L}(N)) \,| \\ (\phi, X, N) \textit{ not lin., Gaussian, Gaussian}\} \setminus \tilde{B}_3$$

(iv) post-nonlin.: $\mathcal{F}_4 = \{f(x,n) = \psi(\phi(x) + n), \psi \textit{ inv.}\}$

$$\mathcal{B}_4 = \{(f, \mathcal{L}(X), \mathcal{L}(N)) \,| \\ (\psi, \phi, X, N) \textit{ not lin., lin., Gaussian, Gaussian}\} \setminus \tilde{B}_4$$

Proof. Shimizu et al. [2006], Peters et al. [2010], Hoyer et al. [2009] and Zhang and Hyvärinen [2009] provide proofs and the precise definitions of the sets $\tilde{B}_i$ for (i)-(iv), respectively. □

We now generalize the concept of bivariate identifiable to more than two variables:

Definition 6.4 [$(\mathcal{B},\mathcal{F})$-IFMOC] Let $\mathcal{B}$ be bivariate identifiable in $\mathcal{F}$. We call an $\mathcal{F}$-FMOC a $(\mathcal{B},\mathcal{F})$-*Identifiable Functional Model Class*, for short $(\mathcal{B},\mathcal{F})$-IFMOC, if for all its functional models

$$X_i = f_i(\mathbf{PA}_i, N_i)\,, \qquad 1 \leq i \leq p$$

for each $i \in \{1,\dots,p\}$ and $j \in \mathbf{PA}_i$ and for all $x_{\mathbf{PA}_i \setminus \{j\}}$, we have

$$f_i(x_{\mathbf{PA}_i \setminus \{j\}}, \underbrace{\cdot}_{X_j}, \underbrace{\cdot}_{N_i}) \in \mathcal{F}_{|_2}\,. \tag{6.2}$$

Additionally, for all sets $\mathbf{S} \subseteq \{1,\dots,p\}$ with $\mathbf{PA}_i \setminus \{j\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_i \setminus \{i,j\}$, there exists an $x_{\mathbf{S}}$ with $p_{\mathbf{S}}(x_{\mathbf{S}}) > 0$ and

$$\Big(f_i(x_{\mathbf{PA}_i \setminus \{j\}}, \underbrace{\cdot}_{X_j}, \underbrace{\cdot}_{N_i}), \mathcal{L}(X_j \,|\, X_{\mathbf{S}} = x_{\mathbf{S}}), \mathcal{L}(N_i)\Big) \in \mathcal{B}\,. \tag{6.3}$$

Thus, an $(\mathcal{B}, \mathcal{F})$-IFMOC consists of many functional models, which are defined in Defintion 6.1.

Example 6.5
- In the bivariate case ($p = 2$), in Definition 6.4 we have $\mathbf{S} = \emptyset$ and thus equation (6.2) is always satisfied. (6.3) then reads that the triple $(f_2, \mathcal{L}(X_1), \mathcal{L}(N_2))$ is in the bivariate identifiable set $\mathcal{B}$ (if $X_1 \to X_2$ and thus $\mathbf{PA}_2 = \{1\}$).
- For more than two variables one can exploit Lemma 6.3. For ANMs equation (6.2) holds: the functions remain additive in the noise if some arguments are fixed. If one further uses linear ANMs $\mathcal{F} = \mathcal{F}_1$, for example, and restricts $\mathcal{B}$ to contain only non-Gaussian noise, also (6.3) holds and we recover LiNGAM [Shimizu et al., 2006]. Using the other $\mathcal{F} \neq \mathcal{F}_1$ from Lemma 6.3 we obtain analogous results for the nonlinear case.

6.3. Identifiability

Now we are able to state our main theoretical result:

Theorem 6.6 *Assume that $\mathcal{L}(\mathbf{X})$ is induced by a functional model from a $(\mathcal{B}, \mathcal{F})$-IFMOC with graph $\mathcal{G}_0$. Then it cannot be induced by a functional model from the same $(\mathcal{B}, \mathcal{F})$-IFMOC that corresponds to a different graph $\mathcal{G} \neq \mathcal{G}_0$.*

The proof can be found in Section A.5.3.

There is a connection between equation (6.1) and faithfulness. In the context of an IFMOC, (6.1) basically reads as Lemma 6.7.

Lemma 6.7 *Consider an instance of an IFMOC with DAG $\mathcal{G}_0$, a variable X_i and one of its parents X_j. For all sets $\mathbf{S}$ with $\mathbf{PA}_i^{\mathcal{G}} \setminus \{j\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_i^{\mathcal{G}}$ we have*

$$X_i \not\perp\!\!\!\perp X_j \,|\, X_{\mathbf{S}} \tag{6.4}$$

If the property (6.4) is violated, faithfulness is violated, too (in this sense, faithfulness is stronger). In Proposition 6.8 we show that

Lemma 6.7 implies *causal minimality*, a weak form of faithfulness (Definition 2.1). Causal minimality states that a joint distribution is not Markov with respect to a strict subgraph of the graph $\mathcal{G}_0$. Further, if $g(x, n) = n$ lies in $\mathcal{F}_{|_2}$, (6.1) is satisfied: If $Y = f(X, N_Y) \perp\!\!\!\perp X$ were true, $X = g(Y, N_X)$ with $N_X = X$ would be a valid backward model.

Proposition 6.8 *Property* (6.4) *in Lemma 6.7 implies causal minimality. If the joint distribution has a strictly positive density with respect to some product measure, causal minimality implies property* (6.4) *in Lemma 6.7.*

In the context of causal inference Theorem 6.6 reads as:

Assumption 6.9 *[causal IFMOC Assumption] Assume that the data generating mechanism belongs to an* $(\mathcal{B}, \mathcal{F})$*-IFMOC with the true causal graph* $\mathcal{G}_c$ *(i.e.,* $X_{\mathbf{PA}_i^{\mathcal{G}_c}}$ *are the direct causes of* X_i*).*

Corollary 6.10 *Under Assumption 6.9 we can identify the true causal DAG* $\mathcal{G}_c$ *from the joint distribution* $\mathcal{L}(\mathbf{X})$*.*

We do not claim that each natural process satisfies Assumption 6.9, only that *if it does*, we can *then* recover the true causal relationships from the joint distribution. Summarizing, this approach provides the following advantages: (1) We can identify the true causal graph even within the Markov equivalence class. (2) One can use IFMOCs to identify non-faithful causal models (even those "undetectable" versions of unfaithfulness mentioned in Section 2.7.1), for which conditional independence-based methods usually fail. (3) In our opinion the IFMOC assumption can be at least partially tested given the data (see Section 2.7.1).
Note that our result already includes discrete models, but only works for non-deterministic data.

Chapter 7.

Multivariate Gaussian Models with Same Error Variances

7.1. Introduction

In this chapter we again consider structural equation models (SEMs), see Section 2.3. Corresponding to each structural equation model, there is a directed acyclic graph (DAG). We have seen that in Gaussian SEMs with linear functions, the graph can be identified from the joint distribution only up to Markov equivalence classes (assuming faithfulness). Sections 3.3 and Chapters 4 and 6 have shown, however, that this constitutes an exceptional case. In the case of linear functions and non-Gaussian noise, the directed acyclic graph becomes identifiable. Apart from few exceptions the same is true for non-linear functions and arbitrarily distributed additive noise. In this chapter, we prove identifiability for a third modification: if we require all noise variables to have the same variances, again, the directed acyclic graph can be recovered from the joint Gaussian distribution.
Our result may come as a surprise that for a class of Gaussian structural equation models the underlying DAG is identifiable. The assumption of same error variances seems natural for a range of applications (with variables from a similar domain) and is commonly used in time series models, which often assume the innovations to be i.i.d..

7.2. Model Definition

We first recall some notation that we are going to use in this chapter. The index set $\mathbf{V} = \{1, \ldots, p\}$ corresponds to a set of vertices in a graph.

Associated with $j \in \mathbf{V}$ are random variables $X_j \in \mathbf{X}$. Given a DAG $\mathcal{G}$, we denote the parents of a node j by $\mathbf{PA}_j^{\mathcal{G}}$, the children by $\mathbf{CH}_j^{\mathcal{G}}$, the descendants by $\mathbf{DE}_j^{\mathcal{G}}$ and the non-descendants are denoted by $\mathbf{ND}_j^{\mathcal{G}}$. We now formally specify our model. Let $\mathbf{X} = \{X_1, \ldots, X_p\}$ be a finite set of variables. We consider an SEM (with DAG $\mathcal{G}_0$) of the form

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j \,, \tag{7.1}$$

where all N_j are independent and identically distributed according to $N(0, \sigma^2)$ with $\sigma^2 > 0$. Additionally, for each $j \in \{1, \ldots, p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$.

7.3. Identifiability

Theorem 7.1 *Let $\mathcal{L}(\mathbf{X})$ be generated from model (7.1). Then all coefficients can be reconstructed from $\mathcal{L}(\mathbf{X})$. In particular, $\mathcal{G}_0$ is identifiable.*

Remark 7.2 [Faithfulness and Causal Minimality] Theorem 7.1 assumes causal minimality, a weak form of faithfulness. From our point of view, causal minimality is as natural as the Markov condition and is in accordance with the intuitive understanding of a causal influence between variables. In its original form, Zhang and Spirtes [2008] define causal minimality as follows: Let $\mathcal{L}(\mathbf{X})$ be Markov to $\mathcal{G}_0$. Then it is not Markov with respect to any proper subgraph of $\mathcal{G}_0$. In the linear Gaussian case, causal minimality is implied by non-vanishing coefficients $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. This follows from Lemma A.14 (below) and Proposition 2 in [Peters et al., 2011b].

Section 3.1 mentions that methods based on conditional independence tests usually assume faithfulness. Zhang and Spirtes [2008] show that given the Markov condition and causal minimality some violations of faithfulness are detectable. They call the non-detectable part triangle-faithfulness, which is still stronger than causal minimality.

Remark 7.3 [Error Covariance with Unknown Scaling] Theorem 7.1 can be generalized to the case, where the error covariance matrix has the form

$$\Sigma_{\mathbf{N}} = \sigma^2 \times \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

with pre-specified $\sigma_1^2, \ldots, \sigma_p^2$ and unknown scaling σ^2.

Remark 7.4 [Causal Interpretation] Again, our result has implications for causal inference. If $\mathcal{G}_0$ is interpreted as the true causal graph of the data generating process for $X_1, \ldots, X_p$, the problem considered here is to infer the causal structure from the joint distribution. In this causal setting, our result reads as follows. If the observational data is generated by a Gaussian SEM that has the same error variances and whose graph is the true causal graph, then the causal graph is identifiable from the joint distribution. Despite this potentially important application, we have presented the statement and its proof without causal terminology.

Chapter 8.

Extension: Causal Inference on Time Series

8.1. Introduction

In the previous chapters, we have assumed to receive i.i.d. data. In this chapter, we extend the principles to time series. We consider finitely many time series X_t^i, $i \in \{1, \ldots, p\}$, with a maximal time order of π, that is we assume no influence from X_{t-k}^i on X_t^j for $k > \pi$. We further assume stationarity: the influence from X_{t-k}^i on X_t^j is required to be the same for all t. The question whether X^i is causing X^j now reads as whether there is a causal influence from some X_{t-k}^i on X_t^j, for $0 \leq k < \pi$. All models assume homoscedastic noise. In time series literature the maximal time order is often denoted by p. In order to keep the notation as consistent as possible, we chose π and denote the dimension of the time series by p.
Note that iid methods (see previous chapters) cannot be applied directly on time series data because a common history might introduce complicated dependencies between contemporaneous data X_t and Y_t. Motivated by the iid case, Chu and Glymour [2008] and Hyvärinen et al. [2008] propose approaches for the time series setting that include linear instantaneous effects. We describe these methods together with Granger causality in Section 8.2. All of them encounter similar problems: none of them are general enough to include nonlinear instantaneous effects or hidden common causes. Furthermore, when the model assumptions are violated the methods give incorrect results and one draws false causal conclusions without noticing. We propose to use time series models with independent noise (*TiMINo*) that include

nonlinear and instantaneous effects. The model is based on Functional Models (also known as Structural Equation Models) and assumes X_t to be a function of all direct causes and some noise variable, the collection of which is supposed to be jointly independent. This constitutes a relatively straight-forward extension on iid methods, but we regard the benefits in the setting of time series as substantial: In Section 8.3 we prove that for TiMINo models the full causal structure can be recovered from the distribution. Section 10.4 introduces an algorithm (*TiMINo causality*) that recovers the model structure from a finite sample. It covers a broader class of models than existing methods and can be run with any provided algorithm for fitting time series. If the data do not satisfy the assumptions, TiMINo causality remains mostly (see Section 10.4) undecided instead of drawing wrong causal conclusions. The methods are applied to simulated and real data sets in Section 11.4.

8.2. Existing Methods

For each i between 1 and p, let $\left(X_t^i\right)_{t\in\mathbb{N}_0}$ be a time series. $\mathbf{X}_t$ denotes the vector of time series values at time t. We call the infinite graph that contains each variable X_t^i as a node the *full time graph*. The *summary time graph* contains all p components of the time series as vertices and an arrow between X^i and X^j, $i \neq j$, if there is an arrow from X_{t-k}^i to X_t^j in the full time graph for some k. In this chapter we address the following

Problem: *Given a sample* $(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ *of a multivariate time series, recover the true causal summary time graph.*

8.2.1. Granger Causality

Granger causality [Granger, 1969] does not require complicated statistics, it is easy to implement, and it is based on the following idea: X^i does not Granger cause X^j if including the past of X^i does not help in predicting X_t^j given the past of all all other time series X^k, $k \neq i$. In principle, "all other" means all other information in the world. In practice, one is limited to X^k, $1 \leq k \leq p$. In order to translate the phrase "does not help" into the mathematical language we need to assume a multivariate time series model. If the data follow the assumed

model, e.g. the VAR model below, Granger causality is sometimes interpreted as testing whether $X^i_{t-h}, h > 0$ is independent of X^j_t given $X^k_{t-h},\ k \in \{1, \ldots, p\} \setminus \{i\}, h > 0$ [see Florens and Mouchart, 1982, Eichler, 2011, Chu and Glymour, 2008, and Section 8.2.2].

Linear Granger Causality

Linear Granger causality considers a VAR model:

$$\mathbf{X}_t = \sum_{\tau=1}^{\pi} \mathbf{A}(\tau)\mathbf{X}_{t-\tau} + \mathbf{N}_t\,, \tag{8.1}$$

where $\mathbf{X}_t$ and $\mathbf{N}_t$ are vectors and $\mathbf{A}(\tau)$ are matrices. For checking whether X^i G-causes X^j one fits a full VAR model M_{full} to $\mathbf{X}_t$ and a VAR model M_{restr} to $\mathbf{X}_t$ with the constraints $A_{ji}(\tau) = 0$ for all $1 \leq \tau \leq \pi$ that predicts X^j_t without using X^i. Then one tests whether the reduction of the residual sum of squares (RSS) of X^j_t is significant by using the following test statistic:

$$T := \frac{(RSS_{\text{restr}} - RSS_{\text{full}})/(p_{\text{full}} - p_{\text{restr}})}{RSS_{\text{full}}/(N - p_{\text{full}})}\,, \tag{8.2}$$

where p_{full} and p_{restr} are the number of parameters in the respective models. For the significance test we use $T \sim F_{p_{\text{full}}-p_{\text{restr}}, N-p_{\text{full}}}$.

Nonlinear Granger Causality

Granger causality has been extended to nonlinear relationships, [e.g. Chen et al., 2004, Ancona et al., 2004]. In this chapter we focus on an extension for the bivariate case proposed by Bell et al. [1996]. It is based on generalized additive models (gams) [Hastie and Tibshirani, 1990]:

$$X^i_t = \sum_{\tau=1}^{\pi} \sum_{j=1}^{n} f_{i,j,\tau}(X^j_{t-\tau}) + N^i_t\,, \tag{8.3}$$

where $\mathbf{N}_t$ is a p dimensional noise vector. In order to test whether X^2 G-causes X^1, for example with order 1, two models are fit: $X^1_t = g_1(X^1_{t-1}) + N_t$ and $X^1_t = g_2(X^1_{t-1}) + g_3(X^2_{t-1}) + M_t$. Bell et al. [1996] utilize the same F statistic as above; this time p_{full} and p_{restr} are the

estimated degrees of freedom of the corresponding models. They refer to simulation studies by Hastie and Tibshirani [1990].

8.2.2. ANLTSM

Following Bell et al. [1996], Chu and Glymour [2008] introduce additive nonlinear time series models (ANLTSM for short) for performing relaxed conditional independence tests: If including one variable, e.g. X^1_{t-1}, into a model for X^2_t that already includes X^2_{t-2}, X^2_{t-1}, and X^1_{t-2} does not improve the predictability of X^2_t, then X^1_{t-1} is said to be independent of X^2_t given $X^2_{t-2}, X^2_{t-1}, X^1_{t-2}$ (if the maximal time lag is 2). Chu and Glymour [2008] propose a method based on constraint-based methods like FCI [Spirtes et al., 2000] in order to infer the causal structure exploiting those conditional independence statements. The instantaneous effects are assumed to be linear and the confounders linear and instantaneous. Unfortunately, we did not find code for this method.

8.2.3. TS-LiNGAM

LiNGAM [Shimizu et al., 2006] infers causal graphs for linear, non-Gaussian data. It has been extended to time series by Hyvärinen et al. [2008] (for short: TS-LiNGAM). It allows for instantaneous effects, all relationships are assumed to be linear. Hidden confounders and nonlinearities may lead to wrong results.

8.2.4. Limitations of Existing Methods

From our point of view, the approaches described above suffer from the following methodological problems: (1) *Instantaneous effects:* The formulation of Granger causality has the intrinsic problem that it cannot deal with instantaneous effects. E.g., when X_t is causing Y_t, including any of the two time series helps for predicting the other. Thus Granger causality infers $X \to Y$ and $Y \to X$. ANLTSM and TS-LiNGAM only allow linear instantaneous effects. Theorem 8.2 shows that the causal summary time graph may still be identifiable when the instantaneous effects are linear and the variables are jointly Gaussian. TS-LiNGAM does not work in these situations. (2) *Confounders:*

Granger causality might fail when there is a confounder between X_t and Y_{t+1}, for example: The path between X_t and Y_{t+1} cannot be blocked by conditioning on any of the observed variables; Granger causality infers $X \to Y$. ANLTSM does not allow for nonlinear confounders or confounders with time structure and TS-LiNGAM may fail, too (Exp. 1 in Section 11.4). (3) *Bad model assumptions:* The methods share a similar problem: Performing general conditional independence tests is desirable, but not feasible, partially because the conditioning sets are too large [e.g. Bergsma, 2004]. Thus, the test is performed under a simple model, for example a linear one. If the model assumption is violated, one may draw wrong conclusions without noticing (e.g. Exp. 3 in Section 11.4). For TiMINo, that we define below, Lemma A.15 shows that after fitting and checking the model by testing for independent residuals, the difficult conditional independences have been checked implicitly.
Thus, a *model check* is a simple but effective improvement. Although Granger causality for two time series can easily be augmented with a cross-correlation test, we do not see a straight-forward extension to the multivariate Granger causality. Furthermore, testing for cross-correlation does not always suffice (see Section 10.4).

8.3. SEMs for Time Series: TiMINo

We define TiMINo, a model class including the models described above and prove its identifiability.

Definition 8.1 Consider a time series $\mathbf{X}_t = (X_t^i)_{1 \le i \le p}$, such that the finite dimensional distributions are absolutely continuous with respect to a product measure (i.e. there is a pdf or a pmf). We say the time series satisfies a *TiMINo* if there is a $\pi > 0$ and if $\forall 1 \le i \le p$ there are sets $\mathbf{PA}_0^i \subseteq \{1, \dots, p\} \setminus \{i\}$ and $\mathbf{PA}_k^i \subseteq \{1, \dots, p\}$ for all $1 \le k \le \pi$, s.t. $\forall t \ge \pi$

$$X_t^i = f_i\big((\mathbf{PA}_\pi^i)_{t-\pi}, \dots, (\mathbf{PA}_1^i)_{t-1}, (\mathbf{PA}_0^i)_t, N_t^i\big)\,, \tag{8.4}$$

with N_t^i and $\mathbf{X}_0, \dots, \mathbf{X}_{\pi-1}$ (jointly) independent and for each i, N_t^i identically distributed in t. Here, $(\mathbf{PA}_k^i)_{t-k}$ is a short hand notation for the set of $\#\mathbf{PA}_k^i$ variables $X_{t-k}^{\mathbf{PA}_k^i}$. The corresponding

full time graph is obtained by drawing arrows from any node that appears in the right-hand side of (8.4) to X_t^i. We require the full time graph to be acyclic.

8.4. Identifiability

Below we assume that equations (8.4) follow an identifiable functional model class (IFMOC), see Chapter 6 for a precise definition. Basically, it means that (I) *causal minimality* holds, a weak form of faithfulness that assumes a statistical dependence between cause and effect given all other parents [Spirtes et al., 2000]. And (II), all f_i come from a function class (e.g. additive noise) that is small enough to make the bivariate case identifiable (Chapter 4) if we exclude certain function-input-noise combinations like linear-Gaussian-Gaussian. The proof of the following theoretical result can be found in the appendix.

Theorem 8.2 *Suppose that* $\mathbf{X}_t$ *can be represented as a TiMINo with* $\mathbf{PA}(X_t^i) = \bigcup_{k=0}^{\pi} (\mathbf{PA}_k^i)_{t-k}$ *being the parents of* X_t^i *and assume further that one of the following holds:*

(i) Equations (8.4) *come from an IFMOC*[1]*.*

(ii) Each component of the time series exhibits a time structure (i.e. $\mathbf{PA}(X_t^i)$ *contains at least one* X_{t-k}^i*), the joint distribution is faithful with respect to the full time graph, and the summary time graph is acyclic.*

Then the full time graph can be recovered from the collection of finite-dimensional distributions

$$\mathcal{L}(\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \ldots, \mathbf{X}_{t_k}),$$

with $(t_1, \ldots, t_k) \in \mathbb{N}^k$*. In particular, the true causal summary time graph is identifiable.*

Remark 8.3

- Neither of the two conditions (i) and (ii) implies the other.

[1] To be precise, we only need the IFMOC condition in the instantaneous effects, see the proof for details.

- Regarding (i): Many choices of a function class are possible [Peters et al., 2011b]. In practice, however, one still needs to fit those functions f_i from the data, which means for additive noise that estimating $\mathbf{E}[X_t^i|\mathbf{X}_{t-p}, \ldots, \mathbf{X}_{t-1}]$ should be feasible. Different results show that strict stationarity and/or α mixing, or geometric ergodicity are required [e.g. Chu and Glymour, 2008]. For some sufficient conditions see [Tong, 1990, Chapter 4]. In this chapter we consider VAR fitting: $f_i(p_1, \ldots, p_r, n) = a_{i,1} \cdot p_1 + \ldots + a_{i,r} \cdot p_r + n$, gam regression: $f_i(p_1, \ldots, p_r, n) = f_{i,1}(p_1) + \ldots + f_{i,r}(p_r) + n$ [e.g. Bell et al., 1996], and GP regression: $f_i(p_1, \ldots, p_r, n) = f_i(p_1, \ldots, p_r) + n$. Note that linear functions lead to the model of Hyvärinen et al. [2008] as a special case.
- Regarding (ii): This condition nicely shows how the time structure makes the causal inference problem harder in some respect (the i.i.d. assumption is dropped), but easier in another respect: in the i.i.d. case, for example, the true graph is not identifiable if all components are jointly Gaussian and the relationships are linear; with time structure it is. (TS-LiNGAM would fail, though.)

Chapter 9.

Extension: Confounder Detection

9.1. Introduction

Until now we have used the assumption that all relevant variables have been observed. An interesting, and possibly even more important, question is how to proceed if not all the relevant variables have been observed. In that case, dependencies between observed variables may also be explained by *confounders* — for instance, if a dependence between the incidence of storks and the birthrate is traced back to a common cause influencing both variables. In general, the difficulty not only lies in the fact that the values of the latent variables have not been observed, but also that the causal structure is unknown. In other words, it is in general not clear whether and how many confounders are needed to explain the data and which observed variables are directly caused by which confounder. Under the assumption of linear relationships between variables, confounders may be identified by means of Independent Component Analysis, as shown recently by Hoyer et al. [2008], if the distributions are non-Gaussian. Other results for the linear case are presented in Silva et al. [2006]. In this chapter, we will not assume linear relationships, but try to tackle the more general, nonlinear case. In the case of two variables without confounder, we have argued that the causal inference task (surprisingly) becomes easier in case of nonlinear functional relationships. We have described a method to infer whether $X \to Y$ or $Y \to X$ from the joint distribution $\mathcal{L}(X, Y)$ of two real-valued random variables X and Y if the joint distribution has been generated from an IFMOC. These includes models where Y is a function

f of X up to an additive noise term, i.e.,

$$Y = f(X) + N\,, \tag{9.1}$$

where N is an unobserved noise variable that is statistically independent of X. We have shown in Section 4.3 that generic choices of functions f, distributions of X and distributions of N induce joint distributions on X, Y that do not admit such an additive noise model in the inverse direction, i.e., from Y to X. We believe that the situation with a confounder between the two variables is similar in that respect: nonlinear functional relationships enlarge the class of models for which the causal structure is identifiable.

9.2. Model Definition

We now state explicitly which assumptions we make in the rest of this chapter. First of all, we focus on the case of only two observed and one latent continuous random variables, all with values in $\mathbb{R}$. We assume that there is no feedback, or in other words, the true causal structure is described by a DAG (directed acyclic graph). Also, we assume that selection effects are absent, that is, the data samples are drawn i.i.d. from the probability distribution described by the model.

Definition 9.1 Let X, Y and T be random variables taking values in $\mathbb{R}$. We define a model for *Confounders with Additive Noise (CAN)* by

$$\begin{array}{ll} X = u(T) + N_X & \text{with } N_X, N_Y, T \\ Y = v(T) + N_Y & \text{jointly independent.} \end{array} \tag{9.2}$$

where $u, v : \mathbb{R} \to \mathbb{R}$ are continuously differentiable functions and N_X, N_Y are real-valued random variables.

The random variables N_X and N_Y describe additive "noise", of which one may think of as the net effect of all other causes which are not shared by X and Y. This model can be represented graphically by the DAG shown in Figure 9.1.

Definition 9.2 We call two CAN models equivalent if they induce the same distributions of N_X, N_Y and the same joint distribution of $(u(T), v(T))$.

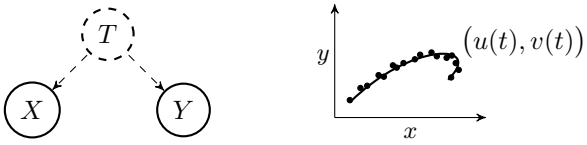


Figure 9.1.: Directed acyclic graph and a scatter plot corresponding to a CAN model for two observed variables X and Y that are influenced by an unobserved confounder variable T.

This definition removes the ambiguity arising from unobservable reparameterizations of T. We further adopt the convention $\mathbf{E}(N_X) = \mathbf{E}(N_Y) = 0$.

The method we propose here enables one to distinguish between (i) $X \to Y$, (ii) $Y \to X$, and (iii) $X \leftarrow T \to Y$ for the class of models defined in (9.2), and (iv) to detect that no CAN model fits the data (which includes, for example, generic instances of the case that X causes Y and in addition T confounds both X and Y). If $N_X = 0$ a.s. ("almost surely") and u is invertible, the model reduces to the model in (9.1) by setting $f := v \circ u^{-1}$. Given that we have observed a joint density on $\mathbb{R}^2$ of two variables X, Y that admits a unique CAN model, the method we propose identifies this CAN model and therefore enables us to do causal inference by employing the following decision rule: we infer $X \to Y$ whenever N_X is zero a.s. and u invertible, infer $Y \to X$ whenever N_Y is zero a.s. and v invertible, and infer $X \leftarrow T \to Y$ if neither of the alternatives hold, which corresponds in spirit with Reichenbach's principle of common cause [Reichenbach, 1956]. Note that the case of $N_X = N_Y = 0$ a.s. and u and v invertible implies a deterministic relation between X and Y, which we exclude here.

In practical applications, however, we propose to prefer the causal hypothesis $X \to Y$ already if the variance of N_X is small compared to the variance of N_Y (after we have normalized both X and Y to variance 1). To justify this, consider the case that X causes Y and the joint distribution admits a model (9.1), but by a slight measurement error, we observe $\tilde{X}$ instead of X, differing by a small additive noise term. Then $\mathcal{L}(\tilde{X}, Y)$ admits a proper CAN model because X is the

latent common cause, but we infer $\tilde{X} \to Y$ because, from a coarse-grained point of view, we should not distinguish between the quantity X and the measurement result $\tilde{X}$ if both variables almost coincide.
Finding the precise conditions under which the identification of CAN models is unique up to equivalence, is a non-trivial problem: If u and v are linear and N_X, N_Y, T are Gaussian, one obtains a whole family of models inducing the same bivariate Gaussian joint distribution. Other examples where the model is not uniquely identifiable are given in Hoyer et al. [2009]: any joint density which admits additive noise models from X to Y and also from Y to X is a special case of a non-identifiable CAN model.
In the next section, we provide theoretical motivation for our belief that in the generic case, CAN models are uniquely identifiable. A practical algorithm for the task is proposed in Section 10.5. It builds on a combination of nonlinear dimensionality reduction and kernel dependence measures. Section 11.5 provides empirical results on synthetic and real world data.

9.3. Identifiability

In [Janzing et al., 2009], we have proved a partial identifiability result for the special case that both u, v are invertible, where we have considered the following limit: first, let the variances of the noise terms N_X and N_Y be small compared to the curvature of the graph $(u(t), v(t))$; second, we assume that the curvature is non-vanishing nevertheless (ruling out the linear case), and third, that the density on the graph $(u(t), v(t))$ changes slowly compared to the variance of the noise.
These results do not provide a full answer to the question of identifiability. They indicate, however, that a stronger statement may hold under suitable technical conditions. They may also provide hints on how to achieve such a result. We do not provide any further details in this thesis, but refer to the original reference [Janzing et al., 2009].

Chapter 10.

Algorithms

We now present practical algorithms. Code of all presented methods can be found on the author's homepage.

10.1. Continuous Bivariate Models

Section 4.3 established for the case of an additive noise model for two continuous variables that given knowledge of the exact densities, the true model is (in the generic case) identifiable. We now consider practical estimation methods which infer the generating graph from sample data.

Again, we begin by considering the case of two observed scalar variables X and Y. Our basic method is straightforward: First, test whether X and Y are statistically independent. If they are not, we continue as described in the following manner: We test whether a model $Y = f(X) + N$ is consistent with the data, simply by doing a nonlinear regression of Y on X (to get an estimate $\hat{f}$ of f), calculating the corresponding residuals $\hat{N} = Y - \hat{f}(X)$, and testing whether $\hat{N}$ is independent of X. If so, we accept the model $Y = f(X) + N$; if not, we reject it. We then similarly test whether the reverse model $X = g(Y) + N$ fits the data.

The above procedure will result in one of several possible scenarios. First, if X and Y are deemed mutually independent we infer that there is no causal relationship between the two, and no further analysis is performed. On the other hand, if they are dependent but both directional models are accepted we conclude that either model may be correct but we cannot infer it from the data. A more positive result is when we are able to reject one of the directions and (tentatively)

accept the other. Finally, it may be the case that neither direction is consistent with the data, in which case we conclude that the generating mechanism is more complex and cannot be described using this model. The selection of the nonlinear regressor and of the particular independence tests are not constrained. Any prior information on the types of functional relationships or the distributions of the noise should optimally be utilized here. In our implementation, we perform the regression using Gaussian Processes Rasmussen and Williams [2006] and the independence tests using kernel methods Gretton et al. [2005]. Note that one must take care to avoid overfitting, as overfitting may lead one to falsely accept models which should be rejected.

10.2. Discrete Bivariate Models

Based on our theoretical findings in Section 5.3 we propose the following method for causal inference (see Section 10.1 for the continuous case):

1. Given: i.i.d. data of the joint distribution $\mathcal{L}(X, Y)$.

2. Regression of $Y = f(X) + N$ leads to residuals $\hat{N}$,
   regression of $X = g(Y) + \tilde{N}$ leads to residuals $\hat{\tilde{N}}$.

3. If $\hat{N} \perp\!\!\!\perp X$ and $\hat{\tilde{N}} \not\perp\!\!\!\perp Y$, infer *"X is causing Y"*,
   if $\hat{N} \not\perp\!\!\!\perp X$ and $\hat{\tilde{N}} \perp\!\!\!\perp Y$, infer *"Y is causing X"*,
   if $\hat{N} \not\perp\!\!\!\perp X$ and $\hat{\tilde{N}} \not\perp\!\!\!\perp Y$, infer *"I don't know (bad model)"*,
   if $\hat{N} \perp\!\!\!\perp X$ and $\hat{\tilde{N}} \perp\!\!\!\perp Y$, infer *"I don't know (both directions possible)"*.

(The identifiability results show that the last case will almost never occur.) This procedure requires discrete methods for regression and independence testing and we now discuss our choices. Code is available on the first author's homepage.

Regression Method

Given a finite number of iid samples of the joint distribution $\mathcal{L}(X, Y)$ we denote the sample distribution by $\mathcal{L}_n(X, Y)$. In continuous regression

we usually minimize a sum consisting of a loss function (like an ℓ_2-error) and a regularization term that prevents us from overfitting.
Regularization of the regression function is not necessary in the discrete case for large sampling. Since we may observe many different values of Y for one specific X value there is no risk in overfitting. This introduces further difficulties compared to continuous regression since in principle we now should try all possible functions from X to Y and compare the corresponding values of the loss function.
Minimizing a *loss function* like an ℓ_p error is not fully appropriate for our purpose, either: after regression we evaluate the proposed function by checking the independence of the residuals. Thus we should choose the function that makes the residuals as independent as possible [see also Mooij et al., 2009]. Therefore we consider a dependence measure (DM) between residuals and regressor as loss function, which we denote by $\mathrm{DM}(\hat{N}, X)$.
Two problems remain:
(1) Assume the different X values $x_1 < \ldots < x_n$ occur in the sample distribution $\mathcal{L}_n(X, Y)$. Then one only has to evaluate the regression function on these values. More problematic is the range of the function. Since we can only deal with finite numbers, we have to restrict the range to a finite set. No matter how large we choose this set, it is always possible that the resulting function class does not contain the true function. But since we used the freedom of choosing an additive constant to require $n(0) > n(k)$ and $\tilde{n}(0) > \tilde{n}(k)$ for all $k \neq 0$, we will always find a sample (X_i, Y_i) with $Y_i = f(X_i)$ if the sample size is large enough. Thus it would be reasonable to consider all Y values that occur together with $X = x$ as a potential value for $f(x)$. To even further reduce the impact of this problem we regard *all* values between $\min Y$ and $\max Y$ as possible values for f. And if there are too few samples with $X = x_j$ and the true value $f(x_j)$ is not included in $\{\min Y, \min Y + 1, \ldots, \max Y\}$ we may not find the true function f, but the few "wrong" residuals do not have an impact on the independence. In practice the following second deliberation is more relevant than the first one:
(2) Even if all values of the true function f are one of the $m := \#\{\min Y, \min Y + 1, \ldots, \max Y\}$ considered values, the problem of checking all possible functions is not tractable: If $n = 20$ and $m = 16$ there are $16^{20} = 2^{80}$ possible functions. We thus propose the following

heuristic but efficient procedure:

Start with an initial function $f^{(0)}$ that maps every value x to the y which occurred (together with this x) most often under all y. Iteratively we then update each function value separately. Keeping all other function values $f(\tilde{x})$ with $\tilde{x} \neq x$ fixed we choose $f(x)$ to be the value that results in the "most independent" residuals. This is done for all x and repeated up to J times as shown in Algorithm 1. Recall that we required $n(0) \geq n(k)$ for all k.

Algorithm 1 Discrete Regression with Dependence Minimization

Input: $\mathcal{L}_n(X, Y)$
Output: f
$f^{(0)}(x_i) := \operatorname{argmax}_y \mathbb{P}_n(X = x_i, Y = y)$
repeat
  $j = j + 1$
  for i in a random ordering **do**
    $f^{(j)}(x_i) := \operatorname{argmin}_y \mathrm{DM}\big(X, Y - f^{(j-1)}_{x_i \mapsto y}(X)\big)$
  end for
until residuals $Y - f^{(j)}(X) =: \hat{N} \perp\!\!\!\perp X$ **or** $f^{(j)}$ does not change anymore **or** $j = J$.

In the algorithm, $f^{(j-1)}_{x_i \mapsto y}(X)$ means that we use the current version of $f^{(j-1)}$ but change the function value $f(x_i)$ to be y. If the argmax in the initialization step is not unique we take the largest possible y. We can even accelerate the iteration step if we do not consider all possible values $\{\min Y, \ldots, \max Y\}$, but only the five that give the highest values of $\mathbb{P}_n(X = x_i, Y = y)$ instead.

Note that the regression method performs coordinate descent in a discrete space and $\mathrm{DM}\big(X, Y - f^{(j)}(X)\big)$ is monotonically decreasing (and bounded from below). Since $f^{(j)}$ is changed only if the dependence measure can be strictly decreased and furthermore the search space is finite, the algorithm converges towards a local optimum. Although it is not obvious why $f^{(j)}$ should converge towards the *global* minimum, the experimental results will show that the method works very reliably in practice.

Independence Test and Dependence Measure

Assume we are given joint iid samples (W_i, Z_i) of the discrete variables W and Z and we want to test whether W and Z are independent. In our implementation we use Pearson's χ^2 test (e.g. Agresti [2002]), which is most commonly used. It computes the difference between observed frequencies and expected frequencies in the contingency table. The test statistic is known to converge towards a χ^2 distribution, which is taken as an approximation even in the finite sample case. In the case of very few samples Cochran [1954] suggests to use this approximation only if more than 80% of the expected counts are larger than 5 ("Cochran's condition"). Otherwise, Fisher's exact test (e.g. Agresti [2002]) could be used. In the experimental section we denote the significance level of the test by α.
For a dependence measure DM we use the p-value (times -1) of the independence test. If the p-value is smaller than 10^{-16}, however, it is regarded as 0 and we take the test statistic instead.

10.3. Multivariate Models

The procedure for two variables described in Section 10.1 could be generalized to an arbitrary number p of observed variables, in the following straightforward way. For each DAG $\mathcal{G}_i$ over the observed variables, test whether it is consistent with the data by constructing a nonlinear regression of each variable on its parents, and subsequently testing whether the resulting residuals are mutually independent. If any independence test is rejected, $\mathcal{G}_i$ is rejected. On the other hand, if none of the independence tests are rejected, $\mathcal{G}_i$ is consistent with the data. This procedure is obviously feasible only for very small networks (roughly $p \leq 6$ or so) and also suffers from the problem of multiple hypothesis testing. Furthermore, the above simple algorithm returns *all* DAGs consistent with the data, including all those for which consistent subgraphs exist. We therefore now present an improved algorithm.

10.3.1. Finding a Suitable Structure

In this section, we propose a more efficient algorithm to find a causal model fitting the data than the algorithm that simply tests all possible

DAGs. We present the algorithm in Algorithm 2. It invokes two subroutines: `FittedNoiseValues`(X, Y), which fits Y as a function of X and returns the residuals (if X is empty, it should just return Y itself as the residuals); and `TestIndependence`(X, Y), which tests independence of X and Y, returning the p-value corresponding to the null hypothesis of independence.

Algorithm 2 Find a DAG consistent with the data

input: data matrix X of size $N \times p$, critical value α
$S \leftarrow \{1, \dots, p\}$
for $j = p$ **downto** 1 **do**
  for all $i \in S$ **do**
    $\hat{N}_i \leftarrow$ `FittedNoiseValues`$(X_{S \setminus \{i\}}, X_i)$
    $p_i \leftarrow$ `TestIndependence`$(X_{S \setminus \{i\}}, \hat{N}_i)$
  end for
  $i^* \leftarrow \operatorname{argmax} p_i$
  if $p_{i^*} < \alpha$ **then**
    return no consistent DAGs
  end if
  $\sigma_j \leftarrow i^*$
  $S \leftarrow S \setminus \{i^*\}$
end for
for $j = 1$ **to** p **do**
  $i \leftarrow \sigma_j$
  $\mathbf{PA}_i \leftarrow \{\sigma_1, \dots, \sigma_{j-1}\}$
  for $k = 1$ **to** $j - 1$ **do**
    $\hat{N}_i \leftarrow$ `FittedNoiseValues`$(X_{\mathbf{PA}_i \setminus \{\sigma_k\}}, X_i)$
    if `TestIndependence`$(X_{\mathbf{PA}_i}, \hat{N}_i) \geq \alpha$ **then**
      $\mathbf{PA}_i \leftarrow \mathbf{PA}_i \setminus \{\sigma_k\}$
    end if
  end for
end for
output: parent sets $(\mathbf{PA}_i)_{1 \leq i \leq p}$

In the first sweep, a possible causal ordering $\sigma \in S_p$ of the variables is inferred (where S_p denotes the symmetric group consisting of all permutations of $\{1, \dots, p\}$). In the second sweep, unnecessary arrows

are removed. The result is a minimal DAG consistent with the data. The time complexity of the algorithm is $\mathcal{O}(p^2)$ if we count regression and independence tests as atomic operations. This should be compared with the super-exponential number of DAGs with p variables which have to be tested for the enumeration algorithm proposed before.
We show that Algorithm 2 is asymptotically consistent under the following assumptions:

(1) Whenever $\{X_1, \ldots, X_l\}$ contains all the parents of Y and none of its descendants, the residuals $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ are independent of every set that contains no descendants of Y.

(2) Whenever $\{X_1, \ldots, X_l\}$ contains a child of Y, independence of $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ and $\{X_1, \ldots, X_l\}$ is rejected.

(3) Whenever there is a parent X of Y with $X \notin \{X_1, \ldots, X_l\}$ then $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ is not independent of X.

Assumption (1) is satisfied if the joint distribution is generated by an *additive* noise model, because the noise of a variable is only relevant for the variable itself and its descendants. Assumption (2) is satisfied in the generic case. Theorem 6.6 states that generic additive noise models generate distributions that do not admit additive noise models with a different structure. Assumption (3) follows from faithfulness because independence of the residual would imply $X \perp\!\!\!\perp Y \mid X_1, \ldots, X_l$, but conditional independence can only hold true for non-adjacent X, Y.
To obtain a causal ordering, we search for a variable X_i for which the regression on the remaining $p-1$ variables (i.e., on $X_{S\setminus\{i\}}$) yields a residual that is independent of $X_{S\setminus\{i\}}$. Every childless node will be accepted by assumption (1), which shows that our search cannot fail. Conversely, X_i is childless by assumption (2), and is thus the last variable with respect to an appropriate ordering of nodes. Since X_i is therefore causally irrelevant for the remaining variables we can repeat the same procedure with $p-1$ variables and so on, until we have identified the first node. Induction over p shows that we have indeed found an allowed causal ordering. The corresponding complete DAG $\mathcal{G}'$ differs from the true graph $\mathcal{G}$ only by unnecessary links.
To remove irrelevant parents, we use the following iterative method. For every X_i, let $\mathbf{PA}_i$ be the set of parents with respect to the current

preliminary graph. For every $Y \in \mathbf{PA}_i$, compute the regression of X_i on $\mathbf{PA}_i \setminus \{Y\}$ and check whether the residual is still independent of $\mathbf{PA}_i$. If Y is a true parent, independence will be rejected by assumption (3). Otherwise it will be accepted by assumption (1). Hence we keep exactly the links that are also present in $\mathcal{G}$.
To complete the consistency proof, the conjecture (assumption (2)) has to be proven.

10.3.2. Finding all Suitable Structures

We now slightly modify the algorithm proposed above, such that it can output all suitable graph structures. Again, given a data set the main idea of the method is as follows: for each graph structure it fits the corresponding functional model from the $\mathcal{F}$-FMOC and outputs all graphs, for which the residuals are independent. If the algorithm has either no or multiple outputs, Theorem 6.6 proves that Assumption 6.9 must be violated. Algorithm 3 shows how to avoid checking all possible DAGs: it finds the sink node, disregards it and continues with the smaller graph. The algorithm is based on Algorithm 2 but outputs *all* graphs that are consistent with the data by using depth-first search: whenever there is more than one way to proceed in building the DAG, instead of choosing the one that leads to the highest p-value of the independence test (see Algorithm 2, line 8) we keep track of all possibilities. Note that $\sigma_1, \ldots, \sigma_p$ give the causal order; they also depend on *currentcase* (omitted to improve readability). To increase robustness, we test for joint independence of the residuals at the end (not shown). The algorithm runs with any independence test and any regression method, our choices are described below.

10.4. Time Series

The algorithm for TiMINo causality is based on the theoretical finding in Theorem 8.2. It takes the time series data as input and outputs either a DAG that estimates the summary time graph or remains undecided. In principle, it tries to fit a TiMINo model to the data and outputs the corresponding graph. If no model with independent residuals is found, it outputs “I do not know”. For a time series with many components,

Algorithm 3 Finding all possible DAGs

input data matrix X of size $N \times p$, sign. value α
$totalcases \leftarrow 1, currentcase \leftarrow 1$
$S(1) \leftarrow \{1, \ldots, p\}, jj(1) \leftarrow p, \sigma_1 \leftarrow 0$
while $currentcase \leq totalcases$ **do**
  for $j = jj(currentcase)$ **downto** 1 **do**
    for all $i \in S$ **do**
      $\hat{N}_i \leftarrow \texttt{FittedNoiseValues}(X_{S\setminus\{i\}}, X_i)$
      $p_i \leftarrow \texttt{TestIndependence}(X_{S\setminus\{i\}}, \hat{N}_i)$
    end for
    $i^* \leftarrow \operatorname{argmax} p_i$
    if $p_i < \alpha$ for all i **then**
      break
    else if $p_i \geq \alpha$ for several i **then**
      increase $totalcases$ accordingly
      store jj, σ, S and those i (except i^*)
    end if
    $\sigma_j(currentcase) \leftarrow i^*$
    $S(currentcase) \leftarrow S(currentcase) \setminus \{i^*\}$
  end for
  $currentcase \leftarrow currentcase + 1$
end while
for $currentcase = 1$ **to** $totalcases$ **do**
  for $j = 1$ **to** p **do**
    $i \leftarrow \sigma_j$
    $\mathbf{PA}_i \leftarrow \{\sigma_1, \ldots, \sigma_{j-1}\}$
    for $k = 1$ **to** $j - 1$ **do**
      $\hat{N}_i \leftarrow \texttt{FittedNoiseValues}(X_{\mathbf{PA}_i\setminus\{\sigma_k\}}, X_i)$
      if $\texttt{TestIndependence}(X_{\mathbf{PA}_i}, \hat{N}_i) \geq \alpha$ **then**
        $\mathbf{PA}_i \leftarrow \mathbf{PA}_i \setminus \{\sigma_k\}$
      end if
    end for
  end for
end for
output all different DAGs
If #DAGs = 0 or ≥ 2, **output** "I do not know."

this gets intractable. In Section 11.4, we concentrate on time series without feedback loops, where we can exploit a more efficient method:

Full causal discovery

For additive noise models (ANMs) without time structure we have proposed a procedure that recovers the structure without enumerating all possible DAGs in Section 10.3.1. This procedure can be modified to be of use for time series (Algorithm 4). As reported in Section 10.3.1, the time complexity is $\mathcal{O}(p^2)$, where p is the number of time series, regarding fitting models and independence testing as atomic operations. To get the full time complexity, $\mathcal{O}(p^2)$ has to be multiplied by the sum of the complexity of the regression method and the independence test, both chosen by the user.

Algorithm 4 TiMINo causality

1: **Input:** Samples from a p-dimensional time series of length T: $(\mathbf{X}_1, \ldots, \mathbf{X}_T)$, maximal order p
2: $S := (1, \ldots, p)$
3: **repeat**
4: **for** k in S **do**
5: Fit TiMINo for X_t^k using
$X_{t-\pi}^k, \ldots, X_{t-1}^k, X_{t-\pi}^i, \ldots, X_{t-1}^i, X_t^i$ for all $i \in S \setminus \{k\}$.
6: Test if residuals are indep. of $X^i, i \in S$.
7: **end for**
8: Choose k^* to be the k with the weakest dependence. (If there is no k with independence, break and output: "I do not know - bad model fit").
9: $S := S \setminus \{k^*\}$
10: $\mathbf{PA}^{k^*} := S$
11: **until** length(S)=1
12: For all k remove all unnecessary parents.
13: **Output:** $(\mathbf{PA}^1, \ldots, \mathbf{PA}^p)$

Depending on the assumed model class, TiMINo causality has to be provided with a fitting method. Here, we chose `ar`, `gam` and `gptk` in R (`http://www.r-project.org/`) for linear models, generalized

additive models, and GP regression, We call the methods TiMINo-linear, TiMINo-gam and TiMINo-GP, respectively. For the first two AIC determines the order of the process. All fitting methods are used in a "standard way". For `gam` we used the built-in nonparametric smoothing splines. For the GP we used zero mean, squared exponential covariance function and Gaussian likelihood. The hyper-parameters are automatically chosen by marginal likelihood optimization.
To test for independence between a residual time series N_t^k and another time series $X_t^i, i \in S$, we shift the latter time series up to the maximal order $\pm\pi$ (but at least up to ± 4); for each of those combinations we perform HSIC [Gretton et al., 2008], an independence test for iid data. One could also use a test based on cross-correlation that can be derived from Thm 11.2.3. in [Brockwell and Davis, 1991]. This is related to what is done in transfer function modeling [e.g. §13.1 in Brockwell and Davis, 1991], which is restricted to two time series and linear functions. But testing for cross-correlation is often not enough: if no time structure is present (iid data), it is obvious that correlation tests are most often insufficient. Also, Experiments 1 and 5 in Section 11.4 describe situations, in which cross-correlations fail. To reduce the running time, however, one can use cross-correlation to determine the graph structure and use HSIC as a final model check. For HSIC we used a Gaussian kernel; as in [Gretton et al., 2008], the bandwidth is chosen to be the median distance of the input data. This is a heuristic but well-established choice.
Note that any other fitting method and independence test can be used as well. Although they work well in practice, we do not claim that our choices are optimal.

Partial causal discovery

Let $\mathbf{X}_t$ "almost" satisfy a TiMINo model, that is some time series are unobserved or some functional relationships are not included in the model. We expect that the full discovery method remains undecided. One can modify the method such that it tries to discover parts of the causal graph: Whenever no k with independent residuals is found in line 8 of Algorithm 4 one subtracts a subset S_0 from the current version of S (first subtract one element, then any combination of two etc.) and repeat. If the method is able to fit a TiMINo model using only the

remaining set $S \setminus S_0$, output this solution and S_0, which has been excluded. Since there are $2^{\#S}$ subsets, this is only feasible for small S (see Exp. 6 in Section 11.4). This method may also be useful for the i.i.d. case; its theoretical properties remain to be investigated.

Weaknesses

(i) In principle, it may happen that the model assumption are violated, but one can nevertheless fit a model in the wrong direction (that is why we wrote "remaining *mostly* undecided"). Here, we refer to the argumentation in Section 2.7.2. Also, (i) is relevant only when there are time series without time structure or the data are non-faithful (see Theorem 8.2). We do not provide a precise analysis of the case with confounders, but analyze this situation empirically in Experiment 1 in Section 11.4. (ii) The null hypothesis of the independence test represents independence, although the scientific discovery of a causal relationship should rather be the alternative hypothesis. This fact may lead to wrong causal conclusions (instead of "I do not know") on small data sets since we cannot reject independence for the wrong direction. This effect is strengthened by the Bonferroni correction of the HSIC based independence test. This may require modifications, when the number of time series is very high. It is thus useful to develop heuristics for "minimal" sample sizes. (iii) For large sample sizes, even smallest differences between the true data generating process and the model may lead to rejected independence tests [discussed by Peters et al., 2011a].

10.5. Confounders

In this section we propose an algorithm (ICAN) that is able to identify a confounder in CAN models. While we only addressed the low noise regime in the previous theoretical section, the practical method we propose here should work even for strong noise, although in that case more data points are needed.
Assume that X, Y are distributed according to the CAN model (9.2). We write $\mathbf{s}(t) := \big(u(t), v(t)\big)$ for the "true" curve of the confounder in $\mathbb{R}^2$. A scatter plot of the samples (X, Y) (right panel of Figure 9.1, for example) suggests a simplistic method for detecting the confounder:

for every curve $\mathbf{s} : [0,1] \to \mathbb{R}^2$ project the data points (X_k, Y_k) onto this curve $\mathbf{s}$, such that the Euclidean distance is minimized:

$$\hat{T}_k = \underset{t \in [0,1]}{\operatorname{argmin}} \|(X_k, Y_k) - \mathbf{s}(t)\|_2 \,.$$

From a set of all possible paths $\mathcal{S}$ now choose the $\hat{\mathbf{s}}$ that minimizes the global ℓ_2 distance $\sum_{k=1}^{n} \|(X_k, Y_k) - \mathbf{s}(\hat{T}_k)\|_2$ (dimensionality reduction) and propose $\hat{T}_k$ to be the confounder for X_k and Y_k. This results in the estimated residuals $(\hat{N}_{X,k}, \hat{N}_{Y,k}) = (X_k, Y_k) - \hat{\mathbf{s}}(\hat{T}_k)$. If the hypotheses $\hat{T} \perp\!\!\!\perp \hat{N}_X, \hat{T} \perp\!\!\!\perp \hat{N}_Y, \hat{N}_X \perp\!\!\!\perp \hat{N}_Y$ cannot be rejected, propose that there is a confounder whose values are given by $\hat{T}_k$.
This idea turns out to be too naive: even if the data have been generated according to the model (9.2), the procedure results in dependent residuals. As an example, consider a data set simulated from the following model:

$$\begin{aligned} X &= 4 \cdot \varphi_{-0.1}(T) + 4 \cdot \varphi_{1.1}(T) + N_X \\ Y &= 1 \cdot \varphi_{-0.1}(T) - 1 \cdot \varphi_{1.1}(T) + N_Y \end{aligned}$$

where φ_μ is the probability density of a $\mathcal{N}(\mu, 0.1^2)$ distributed random variable and $N_X, N_Y \sim U([-0.1, 0.1])$ and $T \sim U([0,1])$ are jointly independent. We now minimize the global ℓ_2 distance over the set of functions

$$\mathcal{S} = \left\{ \mathbf{s} \,:\, \mathbf{s}_i(t) = \alpha_i \cdot \varphi_{-0.1}(t) + \beta_i \cdot \varphi_{1.1}(t); i = 1, 2 \right\}.$$

Since there are only four parameters to fit, the problem is numerically solvable and gives the following optimal solution: $\alpha_1 = 3.9216, \beta_1 = 4.0112, \alpha_2 = 0.9776, \beta_2 = -0.9911$. The ℓ_2 projections $\hat{T}$ result in a lower global ℓ_2 distance (6.92) than the true values of T (11.87).
Figure 10.5 shows the true function $\mathbf{s}$ (black line), a scatter plot of (X, Y) (black circles) and the computed curve $\hat{\mathbf{s}}$ that minimizes the global ℓ_2 distance (dashed red line). Additionally, for some data points projections onto $\mathbf{s}$ and $\hat{\mathbf{s}}$ are shown, too: black crosses correspond to the "true projections" (i.e., the points without noise) onto $\mathbf{s}$ and red crosses correspond to projections onto the estimated function $\hat{\mathbf{s}}$ minimizing the ℓ_2 distance. The latter result in the proposed residuals, which are shown together with the estimated values of the confounder on the

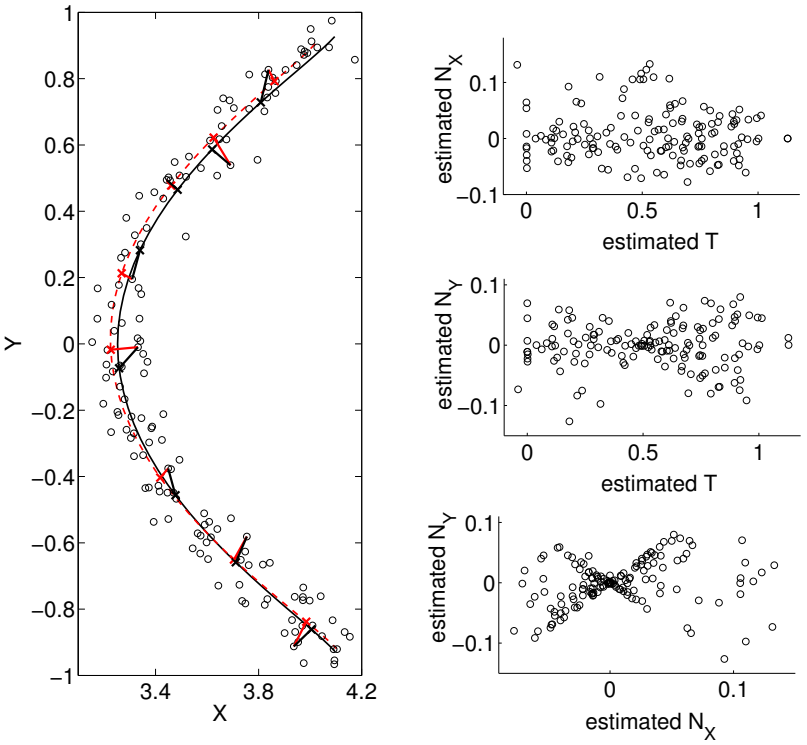


Figure 10.1.: Left: a scatter plot of the data, true path $\mathbf{s}$ and projections (black and solid), estimated path $\hat{\mathbf{s}}$ and projections (red and dashed). Right: residuals plotted against each other and estimated confounder.

right side of Figure 10.1. Estimated residuals and confounder values clearly depend on each other. Also independence tests like the Hilbert-Schmidt Independence Criterion (see below) reject the hypotheses of independence: p-values of 5×10^{-2}, 7×10^{-5} and 1×10^{-4} are computed corresponding to the right plots in Figure 10.1 from top to bottom. These dependencies often occur when the projections onto the curve $\hat{\mathbf{s}}$ are chosen to minimize the global ℓ_2 distance, which can be seen as follows: in our example $\frac{\partial \mathbf{s}_1}{\partial t}$ is small for $T \approx 0.5$ or $Y \approx 0$. Since the points are projected onto the curve orthogonally, the projection results in very small residuals $\hat{N}_Y$. This introduces a dependency between $\hat{N}_Y$ and $\hat{T}$. Dependency between the residuals $\hat{N}_Y$ and $\hat{N}_X$ can arise from regions, where $\hat{\mathbf{s}}$ is approximately linear, like in the bottom right part of Figure 10.1: positive residuals N_Y lead to positive residuals N_X and vice versa.

Summarizing, projecting the pairs (X, Y) onto the path $(\hat{\mathbf{s}}(t))$ by minimizing the ℓ_2 distance to the path is the wrong approach for our purpose. Instead, the data (X, Y) should be projected in a way that minimizes the dependence between residuals and confounder ($\hat{N}_X, \hat{T}$ and $\hat{N}_Y, \hat{T}$) and between the residuals itself ($\hat{N}_X, \hat{N}_Y$).

Let $\mathrm{DEP}(W, Z)$ denote any non-negative dependence measure between random variables W and Z, which is zero if and only if W and Z are independent (we later suggest to use the Hilbert-Schmidt Independence Criterion). In the example above we can use the red curve as an initial guess, but choosing the projections by minimizing the sum of the three dependence measures instead of ℓ_2 distances. In our example this indeed leads to residuals that fulfill the independence constraints (p-values of 1.00, 0.65, 0.80). For the general case, we propose Algorithm 5 as a method for identifying the hidden confounder T given an i.i.d. sample of (X, Y).

If a CAN model can be found we interpret the outcome of our algorithm as $X \to Y$ if $\frac{\mathbf{var}\hat{N}_X}{\mathbf{var}\hat{N}_Y} \ll 1$ and $\hat{u}$ invertible and as $Y \to X$ if $\frac{\mathbf{var}\hat{N}_X}{\mathbf{var}\hat{N}_Y} \gg 1$ and $\hat{v}$ invertible. There is no mathematical rule that tells whether one should identify a variable X and its (possibly noisy) measurement $\tilde{X}$ or consider them as separate variables instead. Thus we cannot be more explicit about the threshold for the factor between $\mathbf{var}\hat{N}_X$ and $\mathbf{var}\hat{N}_Y$ that tells us when to accept $X \to Y$ or $Y \to X$ or $X \leftarrow T \to Y$.

To implement the method we still need an algorithm for the initial

Algorithm 5 Identifying Confounders using Additive Noise Models (ICAN)

Input: $(X_1, Y_1), \ldots, (X_n, Y_n)$ (normalized)

Initialization:

Fit a curve $\hat{\mathbf{s}}$ to the data that minimizes ℓ_2 distance: $\hat{\mathbf{s}} := \operatorname{argmin}_{\mathbf{s}\in\mathcal{S}} \sum_{k=1}^{n} \operatorname{dist}\big(\mathbf{s}, (X_k, Y_k)\big)$.

repeat

  Projection:

  $\hat{T} := \operatorname{argmin}_T \operatorname{DEP}(\hat{N}_X, \hat{N}_Y) + \operatorname{DEP}(\hat{N}_X, T) + \operatorname{DEP}(\hat{N}_Y, T)$ with $(\hat{N}_{X,k}, \hat{N}_{Y,k}) = (X_k, Y_k) - \hat{\mathbf{s}}(T_k)$

  if $\hat{N}_X \perp\!\!\!\perp \hat{N}_Y$ and $\hat{N}_X \perp\!\!\!\perp \hat{T}$ and $\hat{N}_Y \perp\!\!\!\perp \hat{T}$ **then**

    Output: $(\hat{T}_1, \ldots, \hat{T}_n)$, $\hat{u} = \hat{\mathbf{s}}_1$, $\hat{v} = \hat{\mathbf{s}}_2$, and $\frac{\mathbf{var}\hat{N}_X}{\mathbf{var}\hat{N}_Y}$.

    Break.

  end if

  Regression:

  Estimate $\hat{\mathbf{s}}$ by regression $(X, Y) = \hat{\mathbf{s}}(\hat{T}) + \hat{\mathbf{N}}$. Set $\hat{u} = \hat{\mathbf{s}}_1, \hat{v} = \hat{\mathbf{s}}_2$.

until K iterations

Output: Data cannot be fitted by a CAN model.

dimensionality reduction, a dependence criterion DEP, a way to minimize it and an algorithm for non-linear regression. Surely, many different choices are possible. We will now briefly justify our choices for the implementation.

Initial Dimensionality Reduction

It is difficult to solve the optimization problem (line 3 of the algorithm) for a big function class $\mathcal{S}$. Our approach thus separates the problem into two parts: we start with an initial guess for the projection values $\hat{T}_k$ (this is chosen using an implementation of the Isomap algorithm [Tenenbaum et al., 2000] by van der Maaten [2007]) and then iterate between two steps: In the first step we keep the projection values $\hat{T}_k$ fixed and choose a new function $\hat{\mathbf{s}} = (\hat{u}, \hat{v})$, where $\hat{u}$ and $\hat{v}$ are chosen by regression from X on $\hat{T}$ and Y on $\hat{T}$, respectively. To this end we used Gaussian Process Regression [Rasmussen and Williams, 2006], using the implementation of Rasmussen and Nickisch [2007], with hyperparameters set by maximizing marginal likelihood. In the second step the curve is fixed and each data point (X_k, Y_k) is projected to the nearest point of the curve: T_k is chosen such that $\|\hat{\mathbf{s}}(T_k) - (X_k, Y_k)\|_{\ell_2}$ is minimized. We then perform the first step again. A similar iterative procedure for dimensionality reduction has been proposed by Hastie and Stuetzle [1989].
This initial step of the algorithm is used for stabilization. Although the true curve $\mathbf{s}$ may differ from the ℓ_2 minimizer $\hat{\mathbf{s}}$, the difference is not expected to be very large. Minimizing dependence criteria from the beginning often results in very bad fits.

Dependence Criterion and its Minimization

There are various choices of dependence criteria that can be used for the algorithm. Notice, however, that they should be able both to deal with continuous data and to detect non-linear dependencies. Since there is no canonical way of discretizing continuous variables, methods that work for discrete data (like a χ^2 test) are not suitable for our purpose. In our method we chose the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2008]. It can be defined as the distance between the joint distribution and the product of the marginal distribution

represented in a Reproducing Kernel Hilbert Space. For specific choices of the kernel (e.g., a Gaussian kernel) it has been shown that HSIC is zero if and only if the two distributions are independent. Furthermore the distribution of HSIC under the hypothesis of independence can be approximated by a Gamma distribution [Kankainen, 1995]. Thus we can construct a statistical test for the null hypothesis of independence. In our experiments we used Gaussian kernels and chose their kernel sizes to be the median distances between the points [Schölkopf and Smola, 2002]: e.g. $2\sigma^2 = \text{median}\{\|X_k - X_l\|^2 : k < l\}$. We will use the term HSIC for the value of the Hilbert-Schmidt norm and p_{HSIC} for the corresponding p-value. For a small p-value (< 0.05, say) the hypothesis of independence is rejected.
For the projection step we now minimize

$$\text{HSIC}(\hat{N}_X, \hat{N}_Y) + \text{HSIC}(\hat{N}_X, \hat{T}) + \text{HSIC}(\hat{N}_Y, \hat{T})$$

with respect to $\hat{T}$. Note that at this part of the algorithm the function $\hat{\mathbf{s}}$ (and thus $\hat{u}$ and $\hat{v}$) remain fixed and the residuals are computed according to $N_X = X - \hat{u}(\hat{T})$ and $N_Y = Y - \hat{v}(\hat{T})$. We used a standard optimization algorithm for this task (`fminsearch` in MatLab) initializing it with the values of $\hat{T}$ obtained in the previous iteration. Instead of the sum of the three dependence criterion the maximum can be used, too. This is theoretically possible, but complicates the optimization problem since it introduces non-differentiability.
It should be mentioned that sometimes (not for all data sets though) a regularization for the T values may be needed. Even for dependent noise very positive (or negative) values of T result in large residuals, which may be regarded as independent. In our implementation we used a heuristic and just performed 5000 iterations of the minimization algorithm, which proved to work well in practice.

Non-linear Regression

Here, again, we used Gaussian process regression for both variables separately. Since the confounder values $\hat{T}$ are fixed we can fit $X = \hat{u}(\hat{T}) + \hat{N}_X$ and $Y = \hat{v}(\hat{T}) + \hat{N}_Y$ to obtain $\hat{\mathbf{s}} = (\hat{u}, \hat{v})$.
In the experiments this step was mostly not necessary: whenever the algorithm was able to find a solution with independent residuals, it did so in the first or second iteration after optimizing the projections

according to the dependence measures. We still think that this step can be useful for difficult data sets, where the curve that minimizes the ℓ_2 distance is very different from the ground truth.

Chapter 11.

Experiments

11.1. Continuous Bivariate Models

To show the ability of our method to find the correct model when all the assumptions hold we have applied our implementation to a variety of simulated and real data.
For the regression, we used the GPML code from Rasmussen and Nickisch [2007] corresponding to Rasmussen and Williams [2006], using a Gaussian kernel and independent Gaussian noise, optimizing the hyperparameters for each regression individually.[1] In principle, any regression method can be used; we have verified that our results do not depend significantly on the choice of the regression method by comparing with ν-SVR Schölkopf et al. [1999] and with thin-plate spline kernel regression Wahba [1990]. For the independence test, we implemented the HSIC Gretton et al. [2005] with a Gaussian kernel, where we used the gamma distribution as an approximation for the distribution of the HSIC under the null hypothesis of independence in order to calculate the p-value of the test result.

Simulations. The main results for the two-variable case are shown in Figure 11.1. We simulated data using the model $Y = X + bX^3 + N$; the random variables X and N were sampled from a Gaussian distribution and their absolute values were raised to the power q while keeping the

[1] The assumption of Gaussian noise is somewhat inconsistent with our general setting where the residuals are allowed to have any distribution (we even prefer the noise to be non-Gaussian); in practice however, the regression yields acceptable results as long as the noise is sufficiently similar to Gaussian noise. In case of significant outliers, other regression methods may yield better results.

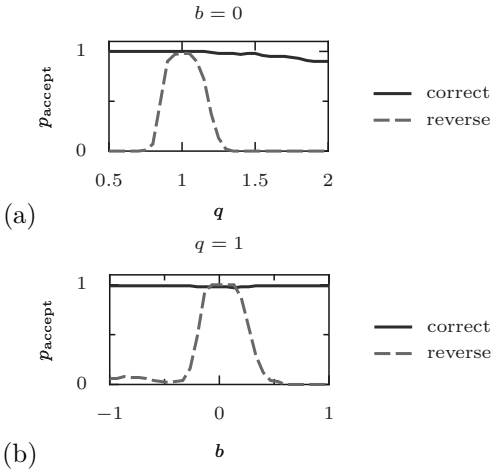


Figure 11.1.: Results of simulations (see main text for details): **(a)** The proportion of times the forward and the reverse model were accepted, p_{accept}, as a function of the non-Gaussianity parameter q (for $b = 0$), and **(b)** as a function of the nonlinearity parameter b (for $q = 1$).

original sign. The parameter b controls the strength of the nonlinearity of the function, $b = 0$ corresponding to the linear case. The parameter q controls the non-Gaussianity of the noise: $q = 1$ gives a Gaussian, while $q > 1$ and $q < 1$ produces super-Gaussian and sub-Gaussian distributions respectively. We used 300 (X, Y) samples for each trial and used a significance level of 2% for rejecting the null hypothesis of independence of residuals and cause. For each b value (or q value) we repeated the experiment 100 times in order to estimate the acceptance probabilities. Panel **(a)** shows that our method can solve the well-known linear but non-Gaussian special case Shimizu et al. [2006]. By plotting the acceptance probability of the correct and the reverse model as a function of non-Gaussianity we can see that when the distributions are sufficiently non-Gaussian the method is able to infer the correct causal direction. Then, in panel **(b)** we similarly demonstrate that we can identify the correct direction for the Gaussian marginal and Gaussian noise model when the functional relationship is sufficiently nonlinear. Note in particular, that the model is identifiable also for positive b in which case the function is invertible, indicating that non-invertibility is not a necessary condition for identification.

Real-world data. Of particular empirical interest is how well the proposed method performs on real world datasets for which the assumptions of our method might only hold approximately. Due to space constraints we only discuss three real world datasets here.

The first dataset, the "Old Faithful" dataset Azzalini and Bowman [1990] contains data about the duration of an eruption and the time interval between subsequent eruptions of the Old Faithful geyser in Yellowstone National Park, USA. Our method obtains a p-value of 0.5 for the (forward) model "current duration causes next interval length" and a p-value of 4.4×10^{-9} for the (backward) model "next interval length causes current duration". Thus, we accept the model where the time interval between the current and the next eruption is a function of the duration of the current eruption, but reject the reverse model. This is in line with the chronological ordering of these events. Figure 11.2 illustrates the data, the forward and backward fit and the residuals for both fits. Note that for the forward model, the residuals seem to be independent of the duration, whereas for the backward model, the

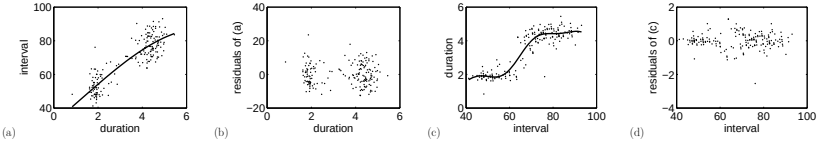


Figure 11.2.: The Old Faithful Geyser data: (a) forward fit corresponding to "current duration causes next interval length"; (b) residuals for forward fit; (c) backward fit corresponding to "next interval length causes current duration"; (d) residuals for backward fit.

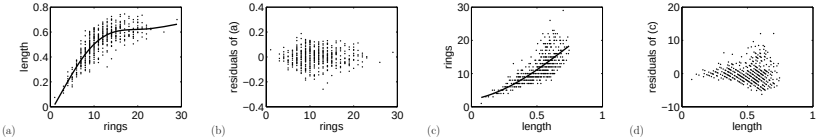


Figure 11.3.: Abalone data: (a) forward fit corresponding to "age (rings) causes length"; (b) residuals for forward fit; (c) backward fit corresponding to "length causes age (rings)"; (d) residuals for backward fit.

residuals are clearly dependent on the interval length. Time-shifting the data by one time step, we obtain for the (forward) model "current interval length causes next duration" a p-value smaller than 10^{-15} and for the (backward) model "next duration causes current interval length" we get a p-value of 1.8×10^{-8}. Hence, our simple nonlinear model with independent additive noise is not consistent with the data in either direction.

The second dataset, the "Abalone" dataset from the UCI ML repository Asuncion and Newman [2007], contains measurements of the number of rings in the shell of abalone (a group of shellfish), which indicate their age, and the length of the shell. Figure 11.3 shows the results for a subsample of 500 datapoints. The correct model "age causes length" leads to a p-value of 0.19, while the reverse model "length causes age"

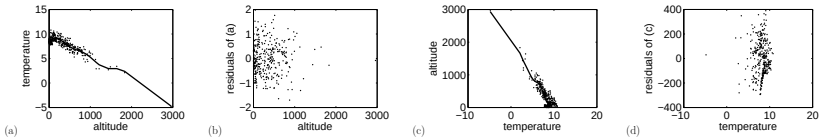


Figure 11.4.: Altitude–temperature data. (a) forward fit corresponding to "altitude causes temperature"; (b) residuals for forward fit; (c) backward fit corresponding to "temperature causes altitude"; (d) residuals for backward fit.

comes with $p < 10^{-15}$. This is in accordance with our intuition. Note that our method favors the correct direction although the assumption of independent additive noise is only approximately correct here; indeed, the variance of the length is dependent on age.

As a third data set, we assay the method on a simple example involving two observed variables: The altitude above sea level (in meters) and the local yearly average outdoor temperature in centigrade, for 349 weather stations in Germany, collected over the time period of 1961–1990 Deutscher Wetter Dienst [2008]. The correct model "altitude causes temperature" leads to $p = 0.017$, while "temperature causes altitude" can clearly be rejected ($p = 8 \times 10^{-15}$), in agreement with common understanding of causality in this case. The results are shown in Figure 11.4.

Janzing et al. [2012] compare the performance of different causal inference techniques on 70 cause-effect pairs from various domains. The three pairs described above are examples from this collection. All data sets and a detailed description can be found at `http://webdav.tuebingen.mpg.de/cause-effect`. They apply the methods on 500 randomly chosen points of the data points since some of the methods require an exhaustive amount of computing time. Figure 11.5 is taken from their paper and shows the results. The gray area is the interval of acceptance for a significance test for H_0 : probability of success is 50% versus H_1 : probability of success is different from 50%. Janzing et al. [2012] compare the additive noise approach presented here (AN) with LINGAM [Shimizu et al., 2006], the post-nonlinear model (PNL) [Zhang and Hyvärinen, 2009], and a recent non-parametric

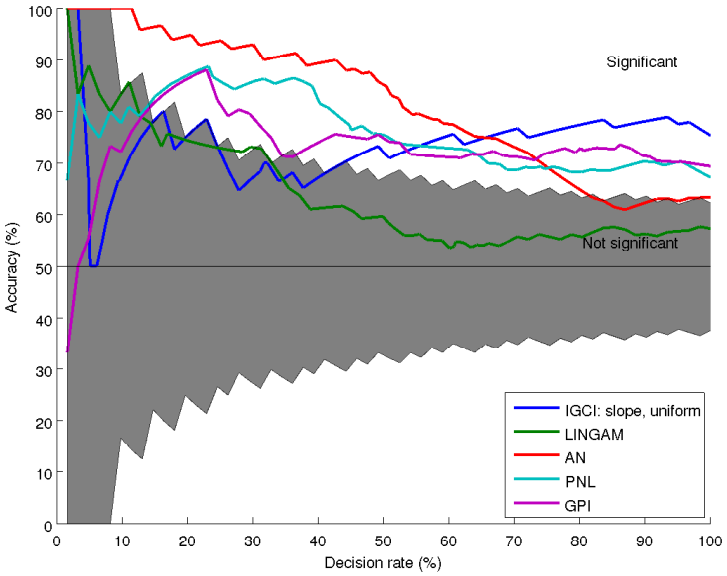


Figure 11.5.: The performance of various causal inference methods on 70 cause-effect pairs with known ground truth. This figure is taken from [Janzing et al., 2012].

method (GPI) [Mooij et al., 2010]. For all methods, except for GPI, they compute both p-values for the independence test of the residuals corresponding to each causal direction using the HSIC independence test [Gretton et al., 2008]. They take their maximum as a confidence estimate for accepting or rejecting the fitted models.

11.2. Discrete Bivariate Models

Simulated Data

We first investigate the performance of our method on synthetic data sets. Therefore we simulate data from ANMs and check whether the method is able to rediscover the true model. We showed in section 5.3 that only very few examples allow a reversible ANM. Experiments A1 and B5 support these theoretical results. We simulate data from many randomly chosen models. All models that allow an ANM in both directions are instances of our examples from above (without exception). Experiments A2 and B6 show how well our method performs for small data size and models that are close to non-identifiability. Experiment A3 empirically investigates the runtime performance of our regression method and compares it with a brute-force search. Experiment A4 show that two consecutive ANMs $Z = g(f(X) + N_1) + N_2$ do not necessarily follow a single ANM. Experiment B7 shows that the method does not favor one direction if the supports of X and Y are of different size. All experiments are available with the code.

Integer Models

Experiment A1 (identifiability). With equal probability we sample from a model with

(1) $\operatorname{supp} X \subset \{1, \ldots, 4\}$
(2) $\operatorname{supp} X \subset \{1, \ldots, 6\}$
(3) X binomial with parameters (n, p)
(4) X geometric with parameter p
(5) X hypergeometric with parameters (M, K, N)
(6) X Poisson with parameter λ or
(7) X negative binomial with parameters (n, p).

Table 11.1.: Experiment A1. The true direction is almost always identified.

correct dir.:	89.9%	both dir. poss.:	5.3%
wrong dir.:	0%	bad fit in both dir.:	4.8%

For each model the parameters of these distributions are chosen randomly (n, M, K, N uniformly between 1 and $40, 40, M, K$, respectively, p uniformly between 0.1 and 0.9 and λ uniformly between 1 and 10), the functions are random ($f(x) \sim U(\{-7, -6, \dots, 7\})$ is uniform for each $x \in \operatorname{supp} X$) and the noise distribution is random, too ($S \sim U(\{1, 2, 3, 4, 5\})$ determines the support $\operatorname{supp} N = \{-S, \dots, S\}$ and $\mathcal{L}(N)$ is chosen by drawing $\#\operatorname{supp} N - 1$ numbers in $[0, 1]$ and taking differences). This way we also construct $\mathcal{L}(X)$ in cases (1) and (2).
We then consider 1000 different models. For each model we sample 1000 data points and apply our algorithm with a signficance level of $\alpha = 0.05$ for the independence test. The results given in Table 11.1 show that the method works well on almost all simulated data sets. The algorithm outputs "bad fit in both directions" in roughly 5% of all cases, which corresponds to the chosen test level. The model is non-identifiable only in 5.3% of the cases, all of which are instances either with a constant function f (2.3%) and thus independent X and Y or with "non-overlapping noise" (3.0%), that is: $f(x) + \operatorname{supp} N$ are disjoint for $x \in X$, which means $\#C_i = 1$ (see Theorem 5.3). This empirically supports Corollary 5.4 and therefore our proposition that the model is identifiable in the generic case.

Experiment A2 (close to non-identifiable). For this data set we sample from the model $Y = f(X) + N$ with $n(-2) = 0.2$, $n(0) = 0.5$, $n(2) = 0.3$, and $f(-3) = f(1) = 1$, $f(-1) = f(3) = 2$. Depending on the parameter r we sample X from $p(-3) = 0.1 + r/2$, $p(-1) = 0.3 - r/2$, $p(1) = 0.15 - r/2$, $p(3) = 0.45 + r/2$.
For each value of the parameter r ranging between $-0.2 \leq r \leq 0.2$ we use 100 different data sets, each of which has the size 400. Theorem 5.3 shows that the ANM is reversible if and only if $r = 0$. Thus,

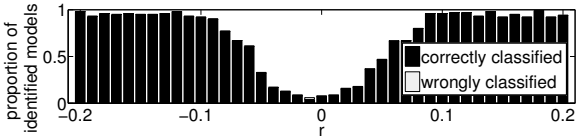


Figure 11.6.: Experiment A2. Proportion of correct and false results of the algorithm depending on the distribution of N. The model is not identifiable for $r = 0$. If r differs significantly from 0 almost all decisions are correct.

our algorithm does not decide when $r \approx 0$. Figure 11.6 shows that the algorithm identifies the correct direction for $r \neq 0$. Again, the test level of $\alpha = 5\%$ introduces indecisiveness of roughly the same size, which can be seen for $|r| \geq 0.15$.

Experiment A3 (fast regression). The space of all functions from the domain of X to the domain of Y is growing rapidly in their sizes: If $\#\operatorname{supp} X = m$ and $\#\operatorname{supp} Y = \tilde{m}$ then the space $\mathcal{F} := \{f : \operatorname{supp} X \to \operatorname{supp} Y\}$ has $\tilde{m}^m$ elements. If one of the variables has infinite support the set is even infinitely large (although this does not happen for any finite data set). It is clear that it is infeasible to optimize the regression criterion by trying every single function. As mentioned before one can argue that with high probability it is enough to only check the functions that correspond to an empirical mass that is greater than 0 (again assuming $n(0) > 0$): E.g. it is likely that $\mathbb{P}_n(X = -2, Y = f(-2)) > 0$. We call these functions "empirically supported". But even this approach is often infeasible. In this experiment we compare the number of possible functions (with values between $\min Y$ and $\max Y$), the number of empirically supported functions and the number of functions that were checked by the algorithm we proposed in section 10.2 in order to find the true function (which it always did).
We simulate from the model $Y = \operatorname{round}(0.5 \cdot X^2) + N$ for two different noise distributions: $n_1(-2) = n_1(2) = 0.05, n_1(k) = 0.3$ for $|k| \leq 1$ and $n_2(-3) = n_2(3) = 0.05, n_2(k) = 0.18$ for $|k| \leq 2$. Each time we simulate a uniformly distributed X with i values between $-\frac{i-1}{2}$

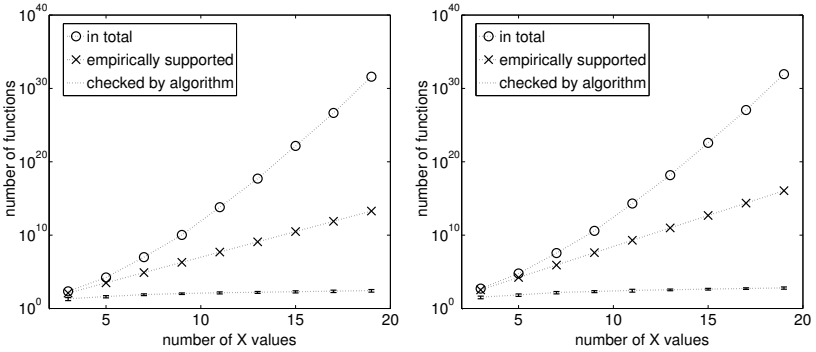


Figure 11.7.: Experiment A3. The size of the whole function space, the number of all functions with empirical support and the number of functions checked by our algorithm (including standard deviation) is shown for N_1 (left) and N_2 (right). An extensive search would be intractable in these cases. The proposed algorithm is very efficient and still finds the correct function for all data sets.

and $\frac{i-1}{2}$ for $i = 3, 5, \ldots, 19$. For each noise-regressor distribution we simulated 100 data sets. For N_1 and $i = 9$, for example, there are $(11 - (-2))^9 \approx 1.1 \cdot 10^{10}$ possible functions in total and $5^9 \approx 2.0 \cdot 10^6$ functions with positive empirical support. Our method only checked 107 ± 25 functions before termination. The highest number of functions checked by the algorithm is 645 ± 220. The full results are shown in Figure 11.7.

Experiment A4 (limitation of ANMs). One can imagine that (for a non-linear g) two consecutive ANMs $Z = g(f(X) + N_1) + N_2$ (which could come from a causal chain $X \to Y \to Z$ with unobserved Y) do not necessarily allow an ANM from X to Z. This means that if a relevant intermediate variable is missing, our method would output "I do not know (bad model fit)" and therefore does not propose a causal direction. We hope, however, that even in this situation the joint distribution is often reasonably "closer" to ANM in the correct

Table 11.2.: Experiment A4. Since the distribution does not allow an ANM, the method does not decide in most cases. Still, the method seems to prefer an ANM in the correct direction.

p-value	$5 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$
correct dir.:	18%	24%	34%	35%
wrong dir.:	5%	4%	2%	5%
both dir. poss.:	2%	18%	27%	36%
bad fit in both dir.:	75%	54%	37%	24%

direction than to an ANM in the wrong direction. We demonstrate this effect on simulated data: We use 300 samples, $\operatorname{supp} X \subset \{1, \ldots, 8\}$ and $\operatorname{supp} N \subset \{-3, \ldots, 3\}$ (the distributions are chosen as in Experiment A1), simulated 100 data sets and obtained the results in Table 11.2. Clearly, the effect vanishes if one either increase the sample size (to 2000, say) or one includes even more ANMs between X and Z (results not shown).

Cyclic Models

All experiments with a cyclic model are denoted with B.

Experiment B5 (identifiability). For three different combinations $(m, \tilde{m}) \in \{(3,3), (3,5), (5,3)\}$ we consider 1000 different models each: As in Experiment A1 we randomly choose a function $f \neq \text{const}$, $\mathcal{L}(X)$ and $\mathcal{L}(N)$. For each model we sample 2000 data points and apply our algorithm with a significance threshold of $\alpha = 0.05$. The results given in Table 11.3 show that the method works well on almost all simulated data sets. The algorithm outputs "bad fit in both directions" in roughly 5% of all cases, which corresponds to the chosen test level. The model is non-identifiable only in very few cases. All of these cases are instances of the counter examples 1(i), 1(ii) and 2 from above. Due to space limitations we only show 6 (out of 11) in Table 11.4. This experiment further supports our theoretical result that the model is identifiable in the generic case.

Table 11.3.: Experiment B5. The algorithm identifies the true causal direction in almost all cases. Only in few cases ANMs can be fit in both directions, which supports the results of section 5.3.

$(m, \tilde{m})$	$(3,3)$	$(3,5)$	$(5,3)$
correct dir.:	95.3%	94.8%	95.5%
wrong dir.:	0.0%	0.0%	0.0%
both dir. poss.:	0.8%	0.0%	0.3%
bad fit in both dir.:	3.9%	5.2%	4.2%

Experiment B6 (close to non-identifiable). For this data set let $m = \tilde{m} = 4$ and $f = \mathrm{id}$. The distribution of p is given by: $p(0) = 0.6, p(1) = 0.1, p(2) = 0.1, p(3) = 0.2$. Depending on the parameter $\frac{1}{2} \leq r \leq \frac{4}{5}$ we sample the noise N from the distribution $n(0) = n(1) = r/2, n(2) = n(3) = 1/2 - r/2$. That means N is uniformly distributed if and only if $r = \frac{1}{2}$. Thus, the model is not identifiable if and only if the noise distribution is uniform, i.e. if and only if $r = \frac{1}{2}$.
(This can be seen as follows: Since $\mathbb{P}(X = 0, Y = 0) > \mathbb{P}(X = k, Y = 0)$ and $\mathbb{P}(X = 0, Y = 1) > \mathbb{P}(X = k, Y = 1)$ for all $k \neq 0$ we have that $g(0) = 0 = g(1)$, still assuming $\mathbb{P}(\tilde{N} = 0) > \mathbb{P}(\tilde{N} = k)$ for all $k \neq 0$. Thus g is not injective. The special form of f leads to one cycle of length 4, which implies that uniformly distributed N is a necessary condition for a reversible ANM, see Proposition A.2 in Section A.4. Example 1(ii) shows that it is also sufficient.)
The further r is away from $\frac{1}{2}$, the easier it should be for our method to detect the true direction. For each value of the parameter r we use 100 different samples, each of which has size 200. This time we choose a significance level of 0.01, which still leads to no wrong decisions (see Figure 11.8).
For $r = 0.58$ and $r = 0.68$ (indicated by the arrows in Figure 11.8) we further investigate the dependence on the data size. Clearly, $r = 0.58$

Table 11.4.: Experiment B5. This table shows only some cases, where ANMs in both directions were possible. All cases (including the ones not shown) are instances of the examples given in section 5.3.

Function f	$p(1), \ldots, p(m)$	$n(1), \ldots, n(\tilde{m})$	Example
$0 \mapsto 0, 1 \mapsto 2, 2 \mapsto 0$	0.83, 0.00, 0.17	$0.15, 0.26, 0.58$	$1(i)$
$0 \mapsto 2, 1 \mapsto 0, 2 \mapsto 2$	0.34, 0.53, 0.14	$0.33, 0.34, 0.33$	$1(ii)$
$0 \mapsto 2, 1 \mapsto 1, 2 \mapsto 0$	0.33, 0.33, 0.34	$0.85, 0.14, 0.02$	2
$0 \mapsto 1, 1 \mapsto 0, 2 \mapsto 1, 3 \mapsto 0, 4 \mapsto 0$	0.20, 0.47, 0.14, 0.08, 0.12	$0.33, 0.33, 0.34$	$1(ii)$
$0 \mapsto 1, 1 \mapsto 0, 2 \mapsto 1, 3 \mapsto 1, 4 \mapsto 1$	0.55, 0.01, 0.03, 0.26, 0.14	$0.37, 0.32, 0.31$	$1(i)$
$0 \mapsto 0, 1 \mapsto 1, 2 \mapsto 0, 3 \mapsto 1, 4 \mapsto 2$	0.03, 0.71, 0.06, 0.10, 0.32	$0.32, 0.34, 0.34$	$1(ii)$

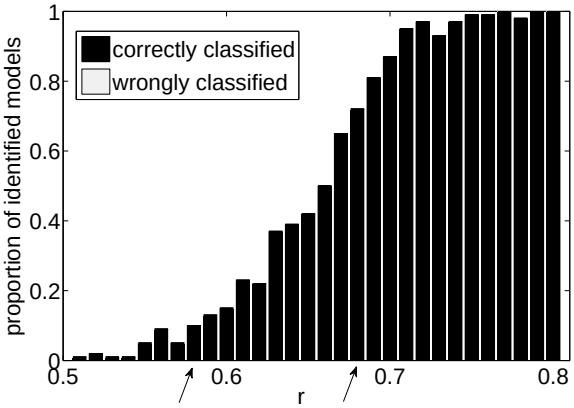


Figure 11.8.: Experiment B6. Proportion of correct results of the algorithm depending on the distribution of N. The model is not identifiable for $r = 0.5$. If r differs significantly from 0.5 the algorithm makes a decisions in almost all cases.

results in a model that is still very close to non-identifiability and thus we need more data to perform well, whereas for $r = 0.68$ the performance increase quickly with the sample size (see Figure 11.9). Note that non-identifiable models lead to very few, but not to wrong decisions.

Experiment B7 (no direction is favored a priori). Here, we consider two random variables, which supports are very unequal in size. If we choose $m := \#\mathcal{X} := \#\operatorname{supp} X = 2$ and $\tilde{m} := \#\mathcal{Y} := \#\operatorname{supp} Y = 10$, there are $2^{10} = 1024$ function from $\mathcal{Y}$ to $\mathcal{X}$, but only $10^2 = 100$ functions from $\mathcal{X}$ to $\mathcal{Y}$; one could expect the method to favor models from Y to X. We show that this is not the case.
For $m \neq \tilde{m} \in \{2, 10\}$ and $m \neq \tilde{m} \in \{3, 20\}$ we randomly choose distributions for X and N and a function f (as before) and sampled 500 data points from this forward ANM. Table 11.5 shows that the algorithm detects the true direction in almost all cases (except if the

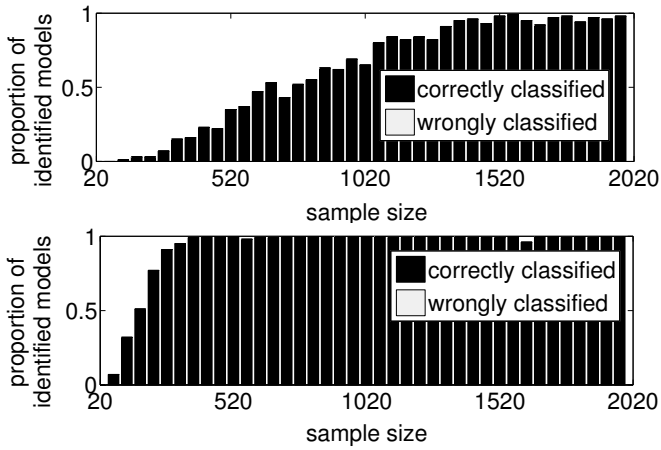


Figure 11.9.: Data Set B6. For $r = 0.58$ (top) and $r = 0.68$ (bottom) the performance depending on the data size is shown. More data is needed if the true model is close to non-identifiable (top). In both cases the performance clearly increases with the sample size.

Table 11.5.: Experiment B7. The algorithm identifies the true causal direction in almost all cases. There is no evidence that the algorithm always favors one direction.

m	$\tilde{m}$	cor. dir.	wrong dir.	both dir. poss.	both dir. bad fit
2	10	97.4%	0%	2.5%	0.1%
10	2	85.2%	0%	14.8%	0.0%
3	20	96.8%	0%	1.6%	1.6%
20	3	95.5%	0%	4.4%	0.1%

model is non-identifiable).

Real Data

Experiment 8 (abalone). We also applied our method to the data set `abalone` [Nash et al., 1994] from the UCI Machine Learning Repository [Asuncion and Newman, 2007]. We tested the sex X of the abalone (male (1), female (2) or infant (0)) against length Y_1, diameter Y_2 and height Y_3, which are all measured in mm, and have $70, 57$ and 28 different values, respectively. Compared to the number of samples (up to 4177) we treat this data as being discrete. Because we do not have information about the underlying continuous length we have to assume that the data structure has not been destroyed by the user-specific discretization. We regard $X \to Y_1$, $X \to Y_2$ and $X \to Y_3$ as being the ground truth, since the sex is probably causing the size of the abalone, but not vice versa.
Clearly, the Y variables do not have a cyclic structure. For the sex variable, however, the most natural model would be a structureless set which is contained in the cyclic constraints; for comparison we try both models for X. Our method is able to identify 2 out of 3 directions correctly and does not make a decision in one case. Except for this one exception all of the backward models are rejected (see Table 11.6 and Figure 11.10). We used the test level $\alpha = 5\%$ and the first 1000 samples of the data set.
For this data set the method proposed by [Sun et al., 2006] returns a

Table 11.6.: Experiment 8. The algorithm identifies the true causal direction in 2 cases. Also for Y_1 the p-value is higher for the correct direction, but formally the method does not make a decision. Here, we assumed a non-cyclic structure on Y and tried both cyclic and non-cyclic for X.

	Y_1	Y_2	Y_3
p-value$_{X \to Y}$	0.17	0.19	0.05
p-value$_{Y \to X}$ (non-cyclic)	$6 \cdot 10^{-12}$	$2 \cdot 10^{-14}$	$< 10^{-16}$
p-value$_{Y \to X}$ (cyclic)	0.06	$4 \cdot 10^{-3}$	$1 \cdot 10^{-8}$

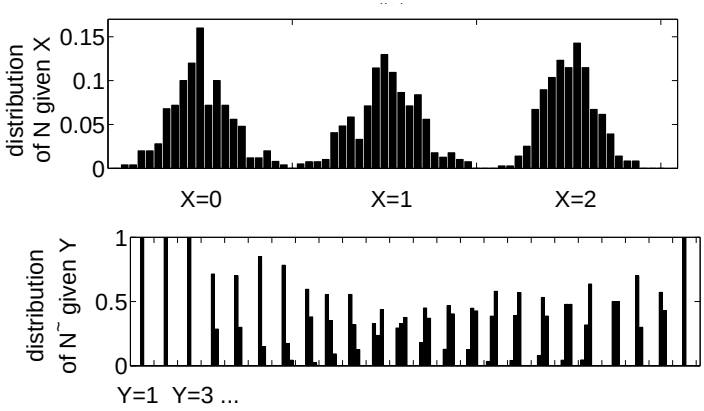


Figure 11.10.: Experiment 8. For Y_3 regressing on X (top) and vice versa (bottom) the plot shows the conditional distribution of the fitted noise given the regressor. If the noise is independent, then the distribution must not depend on the regressor state. Clearly, this is only the case for $X \to Y_3$ (top), which corresponds to the ground truth.

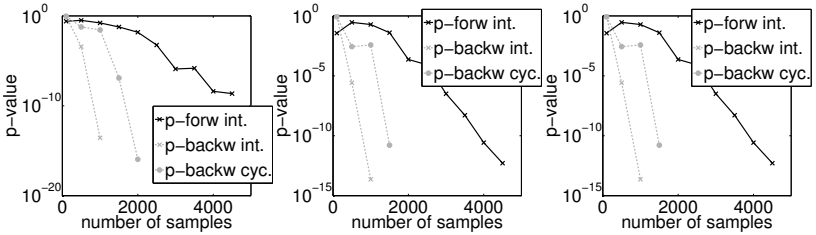


Figure 11.11.: Experiment 8. The plots show p-values of forward and backward direction depending on the number of samples we included (no data point means $p < 10^{-16}$). The p-value in the correct direction is eventually lower than any reasonable threshold. Nevertheless we prefer this direction since it is decreasing much more slowly than p-backward.

slightly higher likelihood for the true causal directions than for the false directions, but this difference is so small, that their algorithm does not consider it to be significant.
The abalone data set also shows that working with p-values requires some care. For synthetic data sets that we simulate from one fixed model the p-values do not depend on the data size. In real world data, however, this often is the case. If the data generating process does not exactly follow the model we assume, but is reasonable close to it, we get good fits for moderate data sizes. Only including more and more data reveals the small difference between process and model, which therefore leads to small p-values. Figure 11.11 shows how the p-values vary if we include the first n data points of the abalone data set (in total: 4177). One can see that although the p-values for the correct direction decrease they are clearly preferable to the p-values of the wrong direction. This is a well-known problem in applied statistics that also has to be considered using our method.

Experiment 9 (acute inflammations). The following data set `acute inflammations` [Czerniak and Zarzycki, 2003] can be found at the

UCI ML Repository [Asuncion and Newman, 2007] and consists of 120 patients. For each patient we have an indicator that tells us whether a specific symptom is present or absent, the temperature and the diagnosis of a medical expert, whether the patient suffers from acute inflammations of urinary bladder and/or whether he suffers from acute nephritises. In particular, we have binary indicators (yes or no) Y_1: occurrence of nausea, Y_2: lumbar pain, Y_3: urine pushing, Y_4: micturition pains and Y_5: burning of urethra, itch, swelling of urethra outlet. Furthermore, the temperature T is measured in $°C$ with $0.1°C$ accuracy. We denote the diagnosis by X_1 (inflammation of urinary bladder) and X_2 (nephritis of renal pelvis origin).
Since the medical expert's diagnosis is based only on the symptoms we expect $Y \to X_i$ and $T \to X_i$ for $i = 1, 2$ (precisely, we expect all Y's and T to be *common* causes for X_i, but here, we only consider the bivariate case and hope that the method still works). It is crucial that the variables X_i only indicate the *diagnosis* and not necessarily the truth. If the X_i corresponded to the true state, X_i would be regarded as the cause and Y as the effect. But in this data set we model the diagnosis behavior of doctors and not the disease process in the patients.
Note further that except for T all variables are binary and should be modeled as being cyclic. The results are presented in Table 11.7. Since T takes 44 different values and the sample size is only 120 we also introduce $T_* := \text{round}(T)$ that only takes 7 values. This is necessary in order to meet Cochran's condition and get reliable results from the independence test. (We are aware that on the other hand, this may introduce small changes in the data generating model, but we hope that this has no effect on the causal reasoning.) The method correctly identifies the causal links $Y_1 \to X_1$, $Y_2 \to X_1$, $T_* \to X_1$ and $T_* \to X_2$. In six more cases the method does not decide. This is relatively often and may be explained by the small data size, for which it is difficult to reject a null hypothesis. We therefore assign an asterisks to all further directions for which the corresponding p-value is at least 10 times larger than the one for the other direction. Furthermore, we checked that the method does not find any causal link between the symptom variables Y, as expected.
Here, the method from [Sun et al., 2006] does not find a significant result in 12 cases (8 cases: exactly the same likelihood for both directions, 3 cases: small favor of the wrong direction, 1 case: small favor

Table 11.7.: Experiment 9. The algorithm identifies the true causal direction in four cases (bold font). In all other cases the method does not decide. The asterisks indicate, where one p-value is at least 10 times larger than the other.

	$p\text{-val}_{X_1 \to Y}$	$p\text{-val}_{Y \to X_1}$	$p\text{-val}_{X_2 \to Y}$	$p\text{-val}_{Y \to X_2}$
Y_1	0.043	**0.368**	$2 \cdot 10^{-9}$	$0.004 *$
Y_2	0.043	**0.368**	$3 \cdot 10^{-5}$	$3 \cdot 10^{-5}$
Y_3	$7 \cdot 10^{-7}$	$4 \cdot 10^{-4} *$	0.009	0.009
Y_4	0.935	0.935	0.925	0.102
Y_5	0.102	0.925	0.190	0.190
T	0.556	1.000	0.080	$0.997 *$
T_*	0.013	**0.435**	0.005	**0.142**

for the correct direction) and it wrongly infers $X_2 \to T$ and $X_2 \to T_*$ as being significant.

Experiment 10 (temperature). We further applied our method to a data set consisting of 9162 daily values of temperature measured in Furtwangen (Germany)[2] using the variables temperature (T, in $^\circ C$) and month (M). Clearly M inherits a cyclic structure, whereas T does not. Since the month indicates the position of the earth relatively to the sun, which is causing the temperature on earth, we take $M \to T$ as the ground truth. Here, we aggregate states and use months instead of days. Again, this is done in order to meet Cochran's condition; it is not a scaling problem of our method (if we do not aggregate the method returns $p_{\text{days} \to T} = 0.9327$ and $p_{T \to \text{days}} = 1.0000$).
For 1000 data points both directions are rejected ($p\text{-value}_{M \to T} = 3e{-}4$, $p\text{-value}_{T \to M} = 1e - 13$). Figure 11.12 shows, however, that again the $p\text{-values}_{M \to T}$ are decreasing much more slowly than $p\text{-values}_{T \to M}$. Using other criteria than simple p-values we still may prefer this direction and propose it as the true one.

[2] B. Janzing contributed this data set. It is one pair on `https://webdav.tuebingen.mpg.de/cause-effect/`

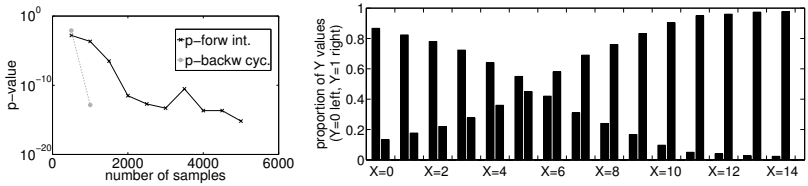


Figure 11.12.: Experiment 10 and Experiment 11. Left: The plot show p-values of forward and backward direction depending on the number of samples we included (no data point means $p < 10^{-16}$). Again we prefer the correct direction since the p-values are decreasing much more slowly than p-backward. Right: The data set does not allow an ANM in any of the two directions. Therefore the method does not propose an answer.

The method proposed in [Sun et al., 2006] finds a larger likelihood for the correct direction, but does not consider this difference as being significant.

Experiment 11 (faces). This data set [Armann and Bülthoff, 2010] (4499 instances) shows the limitations of ANMs. Here, X represents a parameter used to create pictures of artificial faces. X takes values between 0 and 14, where 0 corresponds to a female face, 14 corresponds to a face that is rather masculine. All other parameter values are interpolated. These faces were shown to some subjects who had to indicate whether they believe this is a male ($Y = 1$) or a female ($Y = 0$) face. In this example we regard X as being the cause of Y. However, the data do not admit an ANM in any direction ($p_{X\to Y} = 0$ and $p_{Y\to X} = 0$). Thus, the method does not make a mistake, but does not find the correct answer, either. On this data set the method in [Sun et al., 2006] again detects an insignificantly larger likelihood for the correct direction.

It is possible, however, to include generalizations of ANMs that are capable of modeling this data set. One possibility is to consider models of the form

$$Y = f(X + N), N \perp\!\!\!\perp X \quad \text{and} \quad X = g(Y + \tilde{N}), \tilde{N} \perp\!\!\!\perp Y \tag{11.1}$$

with some possibly non-invertible functions f and g (for continuous data, a similar model has been proposed by Zhang and Hyvärinen [2009]). In this model the function f does not only act on the support of X, but on an enlarged space. Using a method that is based on the same ideas described in section 10.2 one is able to fit this data set quite well ($p_{X\to Y} = 1.000$ and $p_{Y\to X} = 0$). However, we do not have any theoretical identifiability results and the method has one further drawback: Simulations show that it often prefers the variable with the smaller support as the effect.
In particular, we can indicate why the model class at the right hand side of equation (11.1) gets too large if X is a binary variable and Y is the discretization of a continuous variable: If one sets g to be the Heaviside step function defined by $g(w) = 1$ if $w \geq 0$ and $g(w) = 0$ otherwise, equation (11.1) leads to (with $m(t)$ the probability mass function of $M := -\tilde{N}$)

$$\mathbb{P}(X = 1|Y = y) = \mathbb{P}(y + \tilde{N} \geq 0) = \mathbb{P}(M \leq y) = \sum_{t=-\infty}^{y} m(t).$$

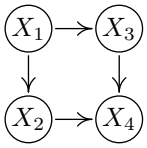
Hence, every conditional for which $\mathbb{P}(X = 1|Y = y)$ is monotonously increasing can be described by an ANM. But even some models that we regard as a natural examples for $X \to Y$ lead to such a monotonously increasing conditional: For example, when $\mathcal{L}(Y|X = 0)$ and $\mathcal{L}(Y|X = 1)$ are discretized Gaussians with equal variance and different means.

11.3. Multivariate Models

In this section, we apply the method described in Section 10.3.2 to simulated data sets. Recall that the method requires a regression function and an independence test. For regression we either use linear regression (IFMOC$_{\text{lin}}$) or Gaussian Processes as in [Hoyer et al., 2009]

(IFMOC_{GP}). To check whether the residuals are independent of the regressors we use HSIC [Gretton et al., 2008]. For the PC algorithm we used an implementation by Tillman et al. [2010] and as a test either partial correlation (PC_{corr}) or "conditional HSIC" (PC_{HSIC}) proposed by Fukumizu et al. [2008] with 500 bootstrap samples to generate the null distribution. Ignoring problems of multiple testing we always set the significance level of statistical tests to 5%.

Experiment 1: How often do we miss faithfulness? For sample sizes between 100 and 500,000 we simulate 500 times data from the following model:



$$
\begin{aligned}
X_1 &= \beta_1 N_1 \\
X_2 &= \alpha_{12} X_1 + \beta_2 N_2 \\
X_3 &= \alpha_{13} X_1 + \beta_3 N_3 \\
X_4 &= \alpha_{24} X_2 + \alpha_{34} X_3 + \beta_4 N_4
\end{aligned}
$$

with $N_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We regard the left DAG as ground truth and sample the coefficients α uniformly between -5 and 5 and β uniformly between 0 and 0.5. We expect the distribution to be non-faithful only on a subset of measure 0. Indeed, given the sampled coefficients we computed all (partial) correlations and verified that all distributions were faithful to the true graph. For finite sample size, however, we expect some cases, where the false hypothesis of zero partial correlation is not rejected. These type 2 errors lead to wrong conclusions about the underlying graph. Figure 11.13 shows how often they occur in the experiments. "due to partial corr. (given two variables)" means that in these proportion of cases there was a partial correlation given two variables wrongly accepted as zero. The number decreases slowly with the sample size, but even for a sample size of 500,000 they happen in more than 10% of the cases. This experiment gives some empirical insight, why strong faithfulness may be useful in order to prove consistency of the PC algorithm. Note that they would be even more frequent if one lowers the significance threshold of the test. In our experiments, other distributions for α and β lead to almost identical results (not shown).

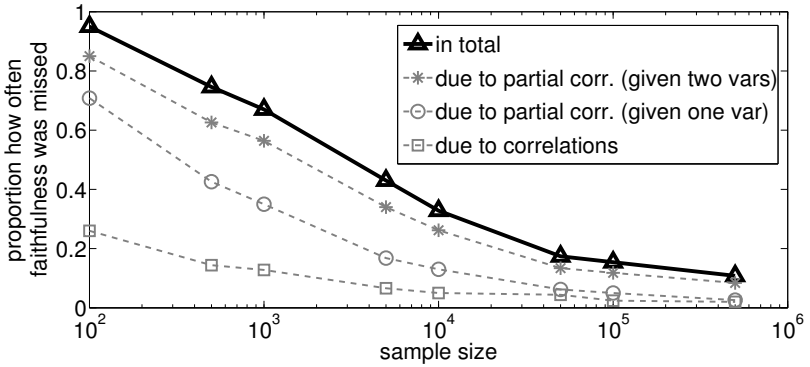
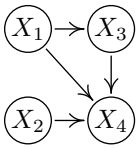


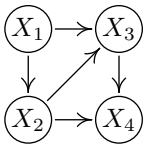
Figure 11.13.: Experiment 1. The graph shows the proportion of cases (out of 500), where at least one (partial) correlation was falsly regarded as zero. These errors lead to wrong causal conclusions.

Experiment 2: Both methods should work when both assumptions are met. We simulate 100 data sets (sample size 400) from two different structures:



linear1 and nonlinear1

$$X_1 = N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_1) + N_3$$
$$X_4 = f_4(X_1, X_2, X_3) + N_4$$



linear2 and nonlinear2

$$X_1 = N_1$$
$$X_2 = f_2(X_1) + N_2$$
$$X_3 = f_3(X_1, X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

	lin1	nonlin1	lin2	nonlin2
PC_{corr}	90/10/0	6/94/0	47/53/0	0/100/0
PC_{HSIC}	60/40/0	96/4/0	3/97/0	4/96/0
$\text{IFMOC}_{\text{lin}}$	82/0/18	0/0/100	86/0/14	0/0/100
IFMOC_{GP}	79/2/19	86/1/13	76/1/23	86/8/6

Table 11.8.: Experiment 2. correct/wrong/undecided (out of 100). The proposed method clearly makes the least mistakes and is not always forced to take a decision.

with $N_i \overset{\text{iid}}{\sim} \mathcal{U}([-0.5, 0.5])$. We regard the drawn graphs as the true causal DAGs.
In linear1 we choose $f_i(x) = a_i^t x$ and in nonlinear1

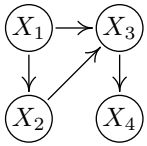
$$f_3(x_1) = a_3 \exp(-2x_1^2) - 1$$
$$f_4(x_1, x_2, x_3) = a_{41} (x_1 + 1)^2 + a_{42} x_2 + a_{43} x_3 .$$

For linear2 we have $f_i(x) = b_i^t x$ and for nonlinear2

$$f_2(x_1) = b_2 x_1, \; f_4(x_2, x_3) = b_{42} (x_2 + 1)^2 + b_{43} x_3$$
$$\text{and } f_3(x_1, x_2) = b_{31} \exp(-2x_1^2) + b_{32} x_2 ,$$

with $a_i, b_i \overset{\text{iid}}{\sim} \mathcal{U}([-2, -1] \cup [1, 2])$. Table 11.8 shows the results. PC_{part} fails for the nonlinear data sets, whereas $\text{IMFOC}_{\text{lin}}$ is undecided. The second setting is more difficult because X_1 and X_4 are only independent given X_2 and X_4 and not a single variable. Especially in this case, the proposed method seems to be more robust. Recall that for PC "correct" means having identified the Markov equivalence class containing the true graph (e.g., with an undirected arrow $X_1 - X_3$), whereas the IFMOC approach identifies the single correct DAG.

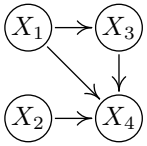
Experiment 3: If the distribution is not faithful, PC fails, IFMOC approach does not. We simulate 100 data sets (sample size 400) from



$$
\begin{aligned}
X_1 &= N_1 \\
X_2 &= 1.5X_1 + N_2 \\
X_3 &= 3X_1 - 2X_2 + N_3 \\
X_4 &= 1.8X_3 + N_4
\end{aligned}
$$

with $N_i \overset{\text{iid}}{\sim} \mathcal{U}([0, 0.5])$. The distribution is not faithful to the true graph (left) since $X_1 \perp\!\!\!\perp X_3$ is not entailed by the Markov condition. This instance of non-faithfulness (triangle-faithfulness) cannot be detected from the data, see Section 2.7.1. Out of the 100 data sets, both PC algorithms always return a wrong DAG that is not Markov equivalent to the true graph. $\text{IFMOC}_{\text{lin}}$ returns the correct DAG in 89 cases and no wrong graph.

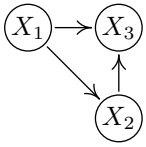
Experiment 4: If the data are induced by an FMOC, but not an IFMOC, both methods can return the Markov equivalence class. We simulate 100 data sets (sample size 400) from



$$
\begin{aligned}
X_1 &= 0.5N_1 \\
X_2 &= 0.5N_2 \\
X_3 &= -X_1 + 0.1N_3 \\
X_4 &= 1.5X_1 - 2X_2 + X_3 + N_4
\end{aligned}
$$

with $N_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The corresponding distribution is faithful to the true graph (left). Since the regime is Gaussian and linear, we use PC_{corr} that uses partial correlation to test for conditional independence. In principle, we expect IFMOC to successfully fit functional models from different structures and to output "I do not know". If one is willing to assume faithfulness, one can output all graphs with the minimal number of edges, which correspond to the true Markov equivalence class (Proposition A.10). Out of 100 data sets PC_{corr} recovers the true Markov equivalence class in 47 cases (the rest is incorrect); $\text{IFMOC}_{\text{lin}}$ in 94 cases and remains undecided 6 times.

Experiment 5: If the assumptions are violated, PC gives wrong results, IFMOC is undecided. We simulate 100 data sets (sample size is 400) from



$$X_1 = N_1$$
$$X_2 = X_1 + 0.5N_2$$
$$X_3 = (X_1 - X_2) \cdot 0.5N_3$$

with $N_i \overset{\text{iid}}{\sim} \mathcal{U}([-0.5, 0.5])$. The corresponding distribution is neither faithful to the true DAG (left) nor do we expect it to satisfy an ANM. Both PC algorithms always output wrong results, whereas both IFMOC methods always output "I do not know".

11.4. Time Series

Artificial Data

We always included instantaneous effects, fitted models up to order $p = 2$ or $p = 6$ and set $\alpha = 0.05$.

Experiment 1: Confounder with time lag. We simulate 100 data sets (length 1000) from

$$Z_t = a \cdot Z_{t-1} + N_{Z,t}\,,$$
$$X_t = 0.6 \cdot X_{t-1} + 0.5 \cdot Z_{t-1} + N_{X,t}\,,$$
$$Y_t = 0.6 \cdot Y_{t-1} + 0.5 \cdot Z_{t-2} + N_{Y,t}\,,$$

with a between 0 and 0.95 and $N_{\cdot,t} \sim 0.4 \cdot \mathcal{N}(0,1)^3$. Here, Z is a hidden common cause for X and Y. For all a, X_t contains information about Z_{t-1} and Y_{t+1} (see Figure 11.15); Granger causality and TS-LiNGAM wrongly infer $X \to Y$. For large a, Y_t contains additional information about X_{t+1}, exploiting Z_{t-2} and Z_t which leads to the wrong arrow $Y \to X$. TiMINo-linear causality does not decide for any a. We only show the linear methods, the nonlinear methods perform very similar (not shown). Note that for $a = 0$, a cross-correlation test is not enough

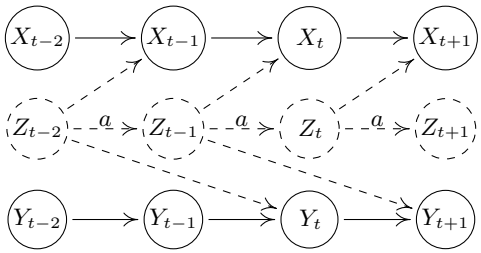


Figure 11.14.: Exp. 1: The figure shows a part of the causal full time graph with Z as a hidden common cause.

to reject $X \to Y$. Further, all methods fail for $a = 0$ and Gaussian noise.

Experiment 2: Linear, Gaussian with instantaneous effects. We sample 100 data sets (length 2000) from

$$\begin{aligned} X_t &= A_1 \cdot X_{t-1} + N_{X,t}\,, \\ W_t &= A_2 \cdot W_{t-1} + A_3 \cdot X_t + N_{W,t}\,, \\ Y_t &= A_4 \cdot Y_{t-1} + A_5 \cdot W_{t-1} + N_{Y,t}\,, \\ Z_t &= A_6 \cdot Z_{t-1} + A_7 \cdot W_t + A_8 \cdot Y_{t-1} + N_{Z,t} \end{aligned}$$

and $N_{\cdot,t} \sim 0.4 \cdot \mathcal{N}(0,1)$ and A_i iid from $\mathcal{U}([-0.8,-0.2] \cup [0.2,0.8])$. We regard the graph containing $X \to W \to Y \to Z$ and $W \to Z$ as correct. TS-LiNGAM and Granger causality are not able to recover the true structure (see Table 11.9).

Experiment 3: Nonlinear, non-Gaussian without instantaneous effects. We simulate 100 data sets (length 500) from

$$\begin{aligned} X_t &= 0.8X_{t-1} + 0.3N_{X,t}, \\ Y_t &= 0.4Y_{t-1} + (X_{t-1} - 1)^2 + 0.3N_{Y,t}, \\ Z_t &= 0.4Z_{t-1} + 0.5\cos(Y_{t-1}) + \sin(Y_{t-1}) + 0.3N_{Z,t}, \end{aligned}$$

with $N_{\cdot,t} \sim \mathcal{U}([-0.5,0.5])$ (similar results for other noise distributions, e.g. exponential). Thus, $X \to Y \to Z$ is the ground truth. Nonlinear

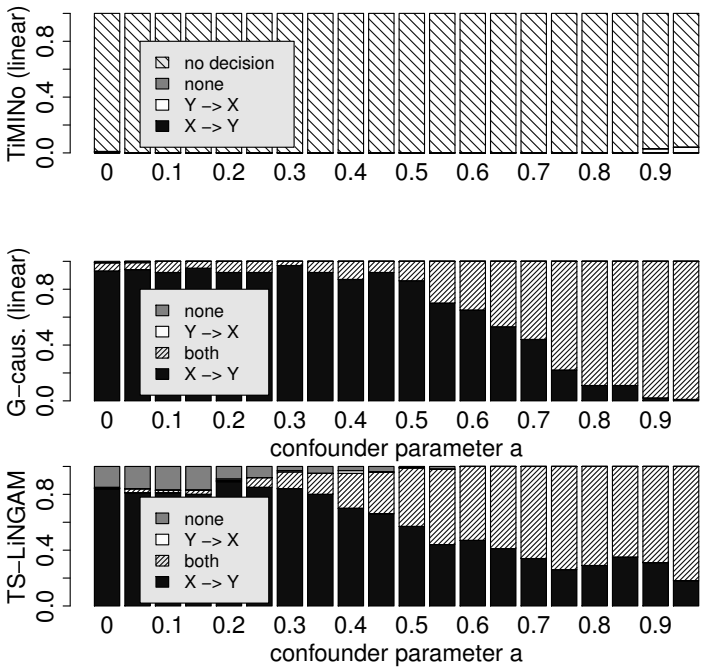


Figure 11.15.: Exp. 1: Because of the existence of a hidden common cause, Granger causality and TS-LiNGAM wrongly infer causal connections between X and Y (top), whereas TiMINo causality does not decide (bottom).

Table 11.9.: Exp. 2: Gaussian data and linear instantaneous effects: only TiMINo mostly discovers the correct DAG.

DAG	lin. Granger	TiMINo-lin	TS-LiNGAM
correct	13%	83%	19%
wrong	87%	7%	81%
no dec.	0%	10%	0%

Granger causality fails since the implementation is only pairwise and it thus always infers an effect from X to Z. Linear Granger causality cannot remove the nonlinear effect from X_{t-2} to Z_t by using Y_{t-1} and gives many wrong answers. Also TiMINo-linear assumes a wrong model, but does not make any decision. TiMINo-gam and TiMINo-GP work well on this data set (Table 11.10). This specific choice of parameters show that a significant difference in performance is possible. For other parameters (e.g. less impact of the nonlinearity), Granger causality and TS-LiNGAM still assume a wrong model but make fewer mistakes.

Experiment 4: Non-additive interaction. We simulate 100 data sets with different lengths from

$$\begin{aligned} X_t &= 0.2 \cdot X_{t-1} + 0.9 N_{X,t} \,, \\ Y_t &= -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.1 N_{Y,t} \,, \end{aligned}$$

with $N_{\cdot,t} \sim \mathcal{N}(0,1)$. Figure 2 shows that TiMINo-linear and TiMINo-gam remain mainly undecided, whereas TiMINo-GP performs well. For small sample sizes, one observes two effects: GP regression does not obtain accurate estimates for the residuals, these estimates are not independent and thus TiMINo-GP remains more often undecided. Also, TiMINo-gam makes more correct answers than one would expect due to more type II errors. Linear Granger causality and TS-LiNGAM give more than 90% incorrect answers, but non-linear Granger causality is most often correct (not shown). Bad model assumptions do not *always* lead to incorrect causal conclusions.

Experiment 5: Non-linear Dependence of Residuals. In Experiment 1, TiMINo equipped with a cross-correlation inferred a causal edge, although there were none. The opposite is also possible:

$$\begin{aligned} X_t &= -0.5 \cdot X_{t-1} + N_{X,t} \,, \\ Y_t &= -0.5 \cdot Y_{t-1} + X_{t-1}^2 + N_{Y,t} \end{aligned}$$

and $N_{\cdot,t} \sim 0.4 \cdot \mathcal{N}(0,1)$ (length 1000). TiMINo-gam with cross-correlation infers no causal link between X and Y, whereas TiMINo-gam with HSIC correctly identifies $X \to Y$.

Table 11.10.: Exp. 3: Since the data are nonlinear, linear Granger causality and TS-LiNGAM give wrong answers, TiMINo-lin does not decide. Nonlinear Granger causality fails because it analyzes the causal structure between pairs of time series.

DAG	Granger (linear)	Granger (nonlinear)	TiMINo-linear	TiMINo-gam	TiMINo-GP	TS-LiNGAM
correct	69%	0%	0%	95%	94%	12%
wrong	31%	100%	0%	1%	1%	88%
no dec.	0%	0%	100%	4%	5%	0%

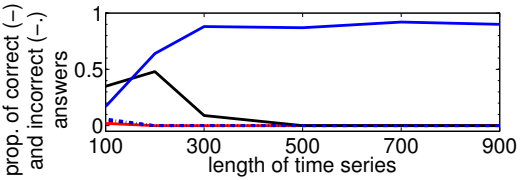


Figure 11.16.: Exp. 4: TiMINo-GP (blue) works reliably for long time series. TiMINo-linear (red) and TiMINo-gam (black) mostly remain undecided.

Experiment 6: Partial Causal Discovery. We sample 100 data sets (length 600) from

$$\begin{aligned}
X_t &= 0.5 \cdot X_{t-1} + N_{X,t}\,, \\
B_t &= 0.5 \cdot B_{t-1} + N_{B,t}\,, \\
A_t &= 0.5 \cdot A_{t-1} + 0.5 \cdot B_{t-1} + N_{A,t}\,, \\
Y_t &= 0.5 \cdot Y_{t-1} - 0.9 \cdot X_{t-1} + 0.8 \cdot B_{t-1} + N_{Y,t}\,, \\
W_t &= 0.5 \cdot W_{t-1} + 0.8 \cdot X_{t-1} + N_{W,t}
\end{aligned}$$

and $N_{\cdot,t} \sim 0.4 \cdot \mathcal{U}([-0.5, 0.5])$. Let X_t be latent. The standard method finds A_t as a "sink time series" and halts in iteration two (line 8 in Algorithm 1). Instead of outputting "I do not know", the partial discovery method described in Section 10.4 is able to infer a DAG (see Figure 11.17) in 82% of the cases (18% wrong answers). Here, a question mark on an edges does not encode a (conditional) independence, but rather allows for any direction or the absence of this edge. Granger causality and TS-LiNGAM give only wrong answers. This experiment should be interpreted as a proof of concept. It remains to be shown when it is possible to output a partial graph. Because of the independence between the time series W and B, it might further be possible to get rid of the edge between W and B.

Real Data

We fitted up to order 6 and included instantaneous effects. For TiMINo, "correct" means that TiMINo-gam makes the correct decision and

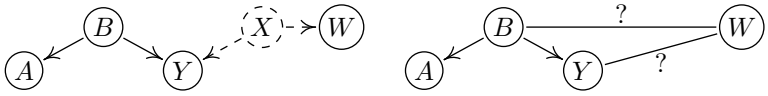


Figure 11.17.: Exp. 6: The true causal summary time graph (left) cannot be recovered because X_t is unobserved. TiMINo gives a partial result (right).

TiMINo-linear is correct or undecided. TiMINo-GP always remains undecided because there are too few data points to fit such a general model. Again, α is set to 0.05.

Experiment 7: Gas Furnace. [Box et al., 2008, length 296], X_t describes the input gas rate and Y_t the output CO_2. We regard $X \to Y$ as being true. TS-LiNGAM, Granger causality, TiMINo-lin and TiMINo-gam correctly infer $X \to Y$. Disregarding time information leads to a wrong causal conclusion: The method described by Section 10.1 leads to a p-value of 4.8% in the correct and 9.1% in the false direction.

Experiment 8: Old Faithful. [Azzalini and Bowman, 1990, length 194] X_t contains the duration of an eruption and Y_t the time interval to the next eruption of the Old Faithful geyser. We regard $X \to Y$ as the ground truth. Although the time intervals $[t, t+1]$ do not have the same length for all t, we model the data as two time series. TS-LiNGAM and TiMINo give correct answers, whereas linear Granger causality infers $X \to Y$, and nonlinear Granger causality infers $Y \to X$.

Experiment 9: Temperature. This data set is available at `https://webdav.tuebingen.mpg.de/cause-effect/`, length 16382. X_t are indoor and Y_t outdoor measurements (recorded every 5 minutes), we expect that there is a causal link $Y \to X$. TS-LiNGAM wrongly infers $X \to Y$ and both Granger causality methods infer a bidirected arrow. TiMINo remains undecided. Maybe, the data are causal insufficient: time may confound outdoor temperature and the usage of heating, the

latter is a direct cause for indoor temperature. Also, Y may cause heating. Such a model does not allow for a TiMINo from Y to X.

Experiment 10: Abalone (no time structure). The abalone data set [Asuncion and Newman, 2007] contains (among others that lead to similar results) age X_t and diameter Y_t of a certain shell fish. If we model 1000 randomly chosen samples as time series, Granger causality (both linear and nonlinear) infers no causal relation as expected. TS-LiNGAM wrongly infers $Y \to X$, which is probably due to the nonlinear relationship. TiMINo gives the correct result.

Experiment 11: Diary (confounder). Here, we consider ten years of weekly prices for butter X_t and cheddar cheese Y_t [Gould, 2007, length 522]. They are strongly correlated, but we expect this correlation to be due to the (hidden) milk price M_t: $X \leftarrow M \to Y$. TiMINo does not decide, whereas TS-LiNGAM and Granger causality wrongly infer $X \to Y$. This may be due to different time lags of the confounder (cheese has longer storing and maturing times than butter).

The phase slope index [Nolte et al., 2008] performed well only in Exp. 6, in all other experiments it either gave wrong results or did not decide.

11.5. Confounder

In this section we apply the method described in Section 10.5 to two simulated and one real data set. These experiments should be interpreted as a proof of concept. We describe some challenges one has to face in confounder detection and principles one could exploit to circumvent them.

Simulated data

Experiment 1. We show on a simulated data set that our algorithm finds the confounder if the data come from the model assumed in (9.2). We simulated 200 data points from a curve whose components u and v consist of a random linear combination of Gaussian bumps each. The noise is uniformly distributed on $[-0.035, 0.035]$. Note

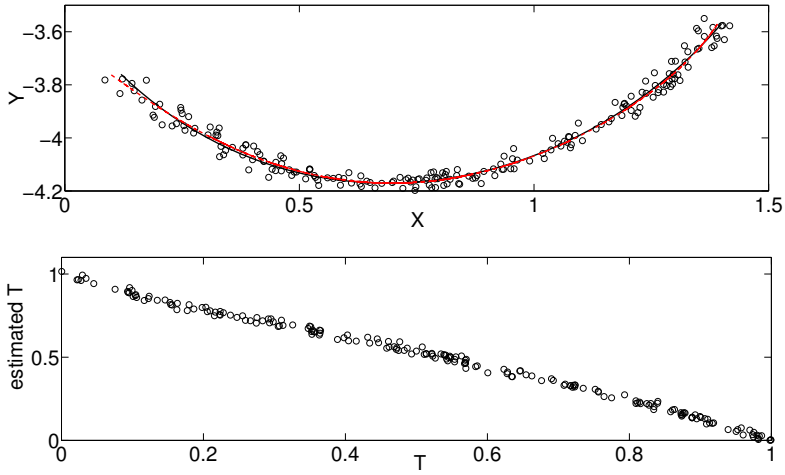


Figure 11.18.: Experiment 1. Top: true (black) and estimated (red) curve. Bottom: The estimated values of the confounder are plotted against the true values. Apart from the arbitrary reparameterization $t \mapsto -t$ the method inferred confounder values close to the true ones.

that in contrast to the example given in Section 10.5 we are now doing the regression using Gaussian processes. The algorithm finds a curve and corresponding projections of the data points onto this curve, such that $\hat{N}_X, \hat{N}_Y$ and $\hat{T}$ are pairwise independent, which can be seen from the p-values $p_{\text{HSIC}}(\hat{N}_X, \hat{N}_Y) = 0.94$, $p_{\text{HSIC}}(\hat{N}_X, \hat{T}) = 0.78$ and $p_{\text{HSIC}}(\hat{N}_Y, \hat{T}) = 0.23$. The top panel of Figure 11.18 shows the data and both true (black) and estimated (red) curve. The bottom panel shows estimated and true values of the confounder. Recall that the confounder can be estimated only up to an arbitrary reparameterization (e.g. $t \mapsto -t$).

In this example the empirical joint distribution of (X, Y) does not allow a simple direct causal relationship between X and Y. It is obvious that the data cannot be explained by $X = g(Y) + N$ with a noise N that is independent of Y. It turns out that also the model corresponding to

the other direction $X \to Y$ can be rejected since a regression of Y onto X leads to dependent residuals ($p_{\text{HSIC}}(X, Y - \hat{f}(X)) = 0.0015$).

Experiment 2. This data set is produced in the same way as data set 1, but this time using an invertible v and unequal scaled noises. We sampled $N_X \sim U([-0.008, 0.008])$ and $N_Y \sim U([-0.0015, 0])$. We argued above that for finite sample sizes this case should rather be regarded as $Y \to X$ and not as an example with a hidden common cause. The algorithm again identifies a curve and projections, such that the independence constraints are satisfied ($p_{\text{HSIC}}(\hat{N}_X, \hat{N}_Y) = 1.00$, $p_{\text{HSIC}}(\hat{N}_X, \hat{T}) = 1.00$ and $p_{\text{HSIC}}(\hat{N}_Y, \hat{T}) = 1.00$, see Figure 11.19), and it is important to note that the different scales of the variances are reduced, but still noticeable ($\frac{\mathbf{var}(\hat{N}_X)}{\mathbf{var}(\hat{N}_Y)} \approx 5$). In such a case we indeed interpret the outcome of our algorithm as "Y causes X".
Since the variances of N_X and N_Y differ significantly and the sample size is small, we can (as expected) even fit a direct causal relationship between X and Y: Assuming the model

$$X = g(Y) + N \tag{11.2}$$

and fitting the function $\hat{g}$ by Gaussian Process regression, for example, results in independent residuals: $p_{\text{HSIC}}(Y, X - \hat{g}(Y)) = 0.97$. Thus we regard the model (11.2) and thus $Y \to X$ to be true. This does not contradict the identifiability conjecture because the dependencies introduced by setting the noise $\hat{N}_Y$ mistakenly to zero are not detectable at this sample size.

Experiment 3. We also simulated a data set for which the noise terms N_X and N_Y clearly depend on T. Figure 11.20 shows a scatter plot of the data set, the outcome curve of the algorithm after $K = 5$ iterations (top) and a scatter plot between the estimated residuals $\hat{N}_Y$ and confounder values $\hat{T}$ (bottom). The method did not find a curve and corresponding projections for which the residuals were independent ($p_{\text{HSIC}}(\hat{N}_Y, \hat{T}) = 0.00$, for example), and thus results in "Data cannot be fitted by a CAN model". This makes sense since the data set was not simulated according to model (9.2).

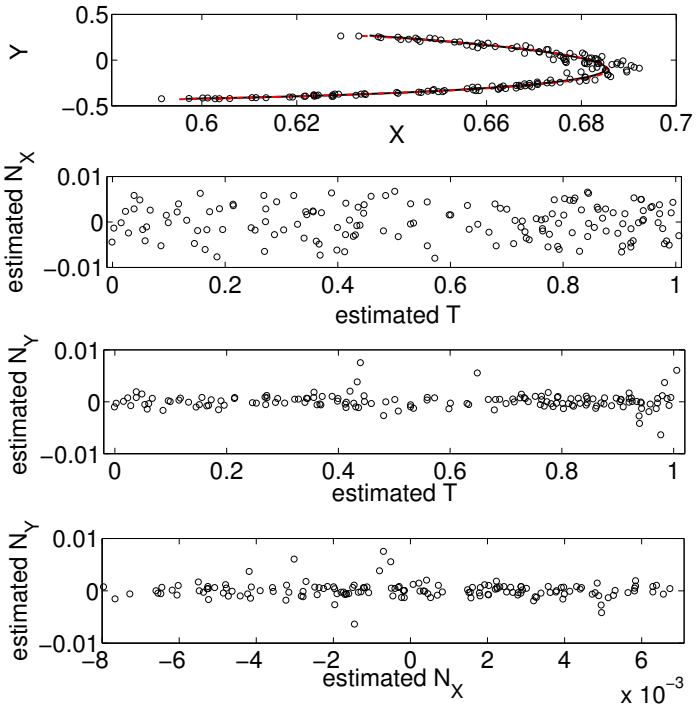


Figure 11.19.: Experiment 2. Top: true (black) and estimated (red) curve. Others: Scatter plots of the fitted residuals against each other and against estimated values for the confounder. The fact that the noise N_X has been sampled with a higher variance than N_Y can also be detected in the fitted residuals.

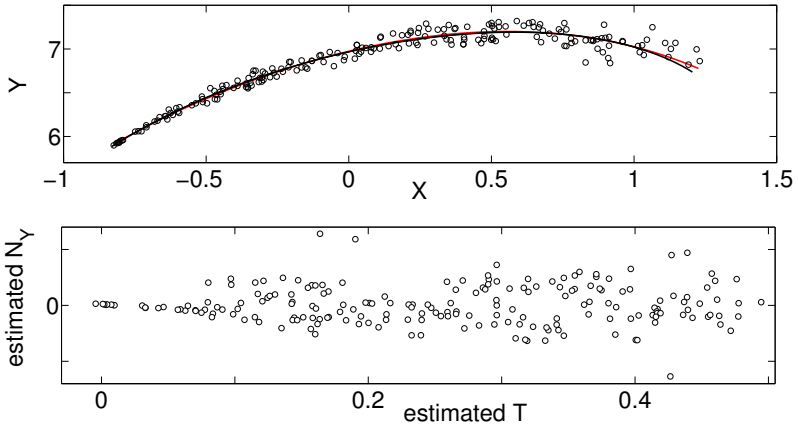


Figure 11.20.: Experiment 3. To check whether our method does not always find a confounder we simulated a data set where the noise clearly depends on T. Indeed the algorithm does not find an independent solution and stops after $K = 5$ iterations. Top: true (black) and estimated (red) curve. Bottom: the estimated residuals clearly depend on the estimated confounder.

Real data

ASOS data. The Automated Surface Observations Systems (ASOS) consists of several stations that automatically collect and transmit weather data every minute. We used 150 values for air pressure that were collected by stations KABE and KABI in January 2000 [NCDC, 2009]. We expect the time to be a confounder. As in the other experiments a projection minimizing the ℓ_2 distance would not be sufficient: after the initialization step we obtain p-values, which reject independence ($p_{\text{HSIC}}(\hat{N}_X, \hat{N}_Y) = 0.00$, $p_{\text{HSIC}}(\hat{N}_X, \hat{T}) = 0.00$, $p_{\text{HSIC}}(\hat{N}_Y, \hat{T}) = 0.02$). After the projection step minimizing the sum of HSICs the residuals are regarded as independent: $p_{\text{HSIC}}(\hat{N}_X, \hat{N}_Y) = 1.00$ and $p_{\text{HSIC}}(\hat{N}_X, \hat{T}) = 1.00$, as well as $p_{\text{HSIC}}(\hat{N}_Y, \hat{T}) = 0.16$. Figures 11.21 and 11.22 show the results. The confounder has been successfully identified.

11.6. Semi-Supervised Learning

Semi-supervised classification

We compare the performance of SSL algorithms with that of base classifiers using only labeled data. We always predict Y from $\mathbf{X}$, where for many examples $\mathbf{X} = (X_1, \ldots, X_p)$ is vector-valued. We assign each data set to one of three categories:

1. *Anticausal/confounded:* (a) data sets in which at least one feature X_i is an effect of the class Y to be predicted (anticausal) (includes also cyclic causal relations between X_i and Y) and (b) data sets in which at least one feature X_i has an unobserved common cause with Y (confounded). In both (a) and (b) the mechanism $\mathcal{L}(Y|X_i)$ can depend on $\mathcal{L}(X_i)$. For these data sets, additional data from $\mathcal{L}(X_i)$ may thus improve prediction.

2. *Causal:* data sets in which some features are causes of the class, and there is no feature which (a) is an effect of the class or (b) has a common cause with the class. If our assumption on independence of cause and mechanism holds, then SSL should be futile on these data sets.

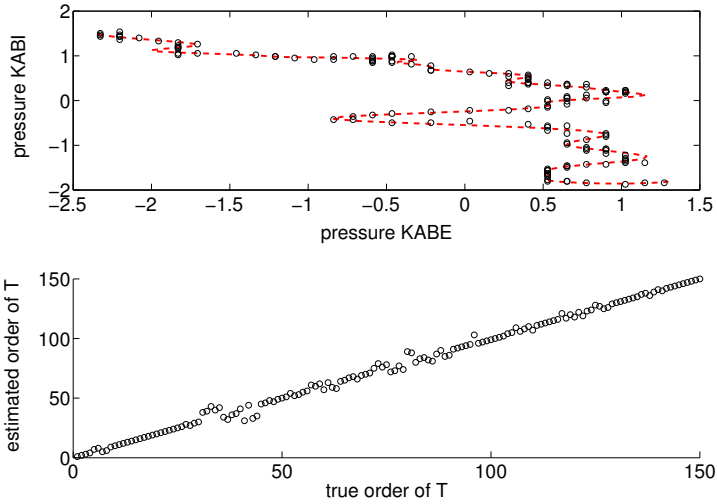


Figure 11.21.: ASOS data. Top: scatter plot of the data, together with the estimated path $\hat{\mathbf{s}}$ (note that it is not interpolating between the data points). Bottom: ordering of the estimated confounder values against the true ordering. The true ordering is almost completely recovered.

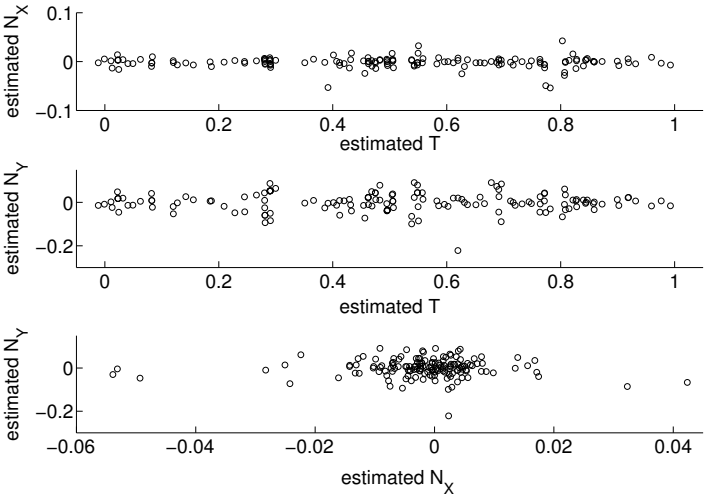


Figure 11.22.: ASOS data. Residuals plotted against each other and against the estimated confounder. The hypothesis of independence is not rejected, which means the method identified the confounder.

3. *Unclear:* data sets which were difficult to be categorized to one of the aforementioned categories. Some of the reasons for that are incomplete documentation or lack of domain knowledge.

In practice, we count a data set already as causal when we believe that the dependence between X and Y is *mainly* due to X causing Y, although additional confounding effects may be possible.

We first analyze the results in the benchmark chapter of a book on SSL (Tables 21.11 and 21.13 of Chapelle et al. [2006]), for the case of 100 labeled training points. The chapter compares eleven SSL methods to the base classifiers 1-NN and SVM[3]. In view of our hypothesis, it is encouraging to see (Figure 11.23) that SSL does not significantly improve the accuracy in the one causal data set, but it helps in most of the anticausal/confounded data sets. However, it is difficult to draw conclusions from this small collection of data sets. Moreover, two additional issues may confound things: (1) the experiments were carried out in a *transductive* setting. Inductive methods use unlabeled data to arrive at a classifier which is subsequently applied to an unknown test set; in contrast, transductive methods only try to make predictions on the test inputs. This could potentially allow performance improvements independent of whether a data set is causal or anticausal. And (2), the SSL methods cover a broad range, and were not extensions of the base classifiers; moreover, the results for the SecStr data set are based on a different set of methods than the rest of the benchmarks.
We next consider 26 UCI data sets and six different base classifiers[4]. The original results are from Tables III and IV in Guo et al. [2010], and are presently re-analyzed in terms of the above data set categories. The comprehensive results of Guo et al. [2010] allow us the luxury of (1) considering only self-training, which is an extension of supervised learning to unlabeled data in the sense that if the set of unlabeled data is empty, we recover the results of the base method (in this case, self-training would stop at the first iteration). This lets us compare an SSL method to its corresponding base algorithm. Moreover, (2) we included

[3] On `http://pl.is.tue.mpg.de/p/causal-anticausal`, we give details on our subjective categorization of the eight data sets used in the chapter.

[4] Again, the webpage `http://pl.is.tue.mpg.de/p/causal-anticausal` describes our subjective categorization of the 26 UCI data sets into "anticausal/confounded", "causal" or "unclear".

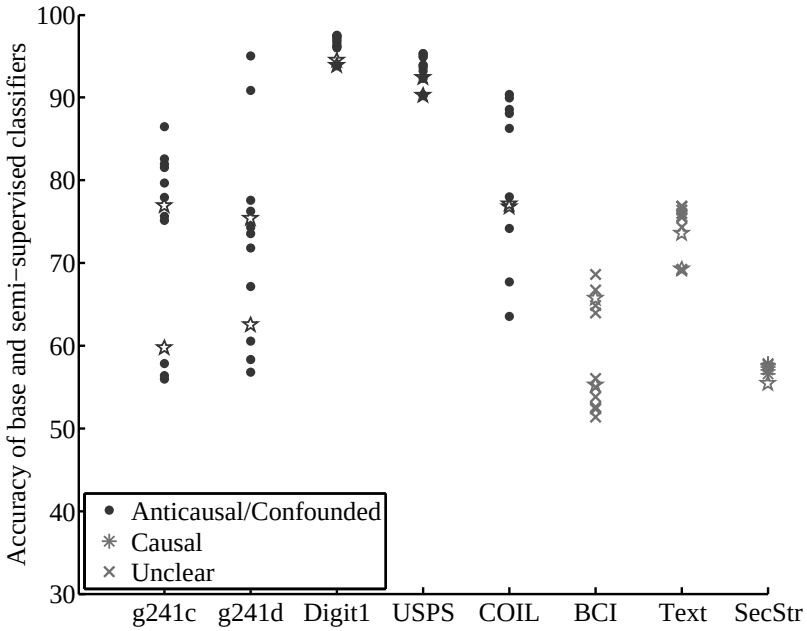


Figure 11.23.: Accuracy of base classifiers (star shape) and different SSL methods on eight benchmark data sets.

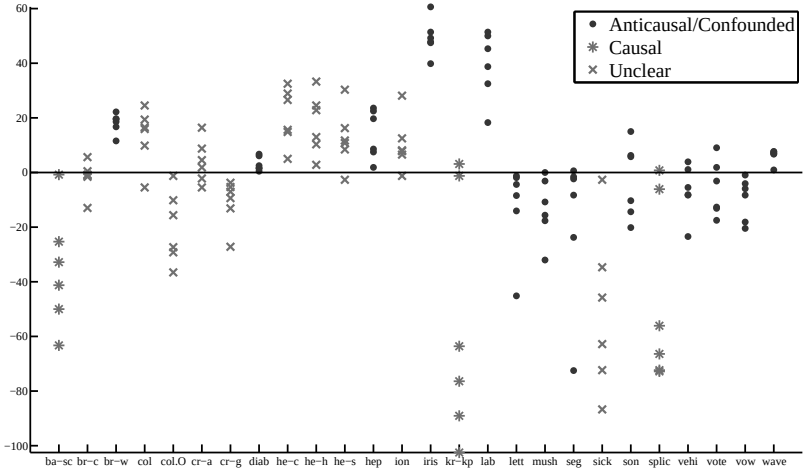


Figure 11.24.: Plot of the relative decrease of error when using self-training, for six base classifiers on 26 UCI data sets. Here, the relative decrease is defined as (error(base) – error(self-train)) / error(base). Self-training, a method for SSL, overall does not help for the causal data sets, but it does help for several of the anticausal/confounded data sets.

only the *inductive* methods considered by Guo et al. [2010], and not the *transductive* ones (cf. our discussion above).
In Figure 11.24, we observe that SSL does not significantly decrease the error rate in the three causal data sets, but it does increase the performance in several of the anticausal/confounded data sets. This is again consistent with our hypothesis that if mechanism and input are independent, SSL will not help for causal data sets.

Semi-supervised regression (SSR)

Classification problems are often inherently asymmetric in that the inputs are continuous and the outputs categorical. It is worth reassuring that we obtain similar results in the case of regression. To this end, we consider the co-regularized least squares regression (co-RLSR) algorithm, compared to regular RLSR on 32 real-world data sets by Brefeld et al. [2006] (two of which are identical, so 31 data sets were considered). We categorized them into causal/anticausal/unclear, prior to the subsequent analysis.
We deemed seven of the data sets anticausal, i.e., the target variable can be considered as the cause of (some of) the predictors; Figure 11.25 shows that SSR reduces the root mean square errors (RMSE) in all these cases. Nine of the remaining data sets can be considered causal, and Figure 11.26 shows that there is usually little performance improvement for those. As Brefeld et al. [2006], we used the Wilcoxon signed rank test to assess whether SSR outperforms supervised regression, in the anticausal and causal cases. The null hypothesis is that the distribution of the difference between the RMSE produced by SSR and that by supervised regression is symmetric around 0 (i.e., that SSR does not help). On the anticausal data sets, the p-value is 0.0156, while it is 0.6523 on the causal data sets. Therefore, we reject the null hypothesis in the anticausal case at a 5% significance level, but not in the causal case.

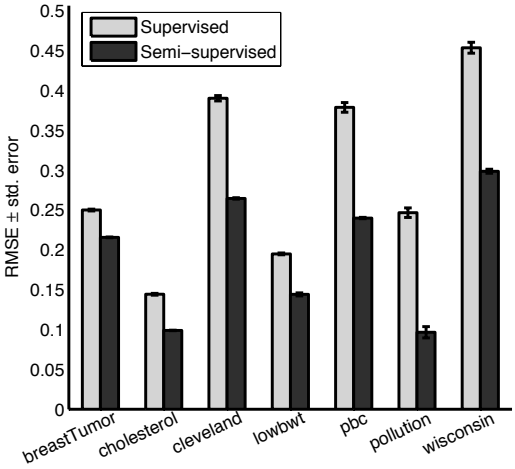


Figure 11.25.: RMSE for Anticausal/Confounded data sets.

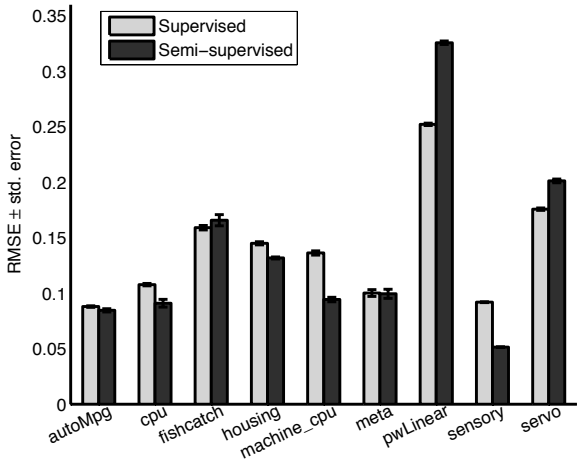


Figure 11.26.: RMSE for Causal data sets.

Chapter 12.

Conclusions and Future Work

12.1. Conclusions

We have shown how restricted structural equation models can be used for causal inference methdos.

Theoretical Findings For two continuous random variables we have shown that the linear non-Gaussian causal framework can be generalized to nonlinear functional dependencies as long as the noise on the variables remains additive. Apart from a few exceptions (e.g. the known linear Gaussian case) we have seen that the structure of the SEM (here: $X \to Y$ or $Y \to X$) can be inferred from the distribution.
Also for two discrete random variables we proposed a method that tries to infer the cause-effect relationship using the concept of additive noise models. We proved that for generic choices the direction of a discrete ANM is identifiable from the distribution.
We proved that for model classes that are able to distinguish between $X \to Y$ and $Y \to X$ (including the two mentioned above), the whole true causal graph is identifiable from the joint distribution. This result requires causal minimality, a weak form of faithfulness.
We have also shown that a Gaussian SEM with same error variances is identifiable from the distribution. The assumption of same error variances constitutes an alternative to the restrictions of non-linear functions and non-Gaussian noise.
We have applied restricted SEMs to time series data (TiMINo causality) and proved identifiability statements analogous to the i.i.d. case.
We have considered a model, in which two observed variables are functions of an unobserved confounder plus some independent additive

noise. We have provided a reference to a theoretical motivation for the question whether this model is distinguishable from an additive noise model between the two variables (without hidden variable). In this case, our findings should be regarded as initial results rather than a full theoretical answer.

Methods and Experiments Corresponding to all theoretical findings mentioned above we have developed algorithms that make the proposed inference principles applicable to a finite amount of data. The experiments support our theoretical results and show that restricted SEMs may have an advantage compared to independence-based methods. We have further succesfully applied the methods to real world data sets for two variables and time series data with known ground truth. On the data sets considered they are more reliable than existing methods.
Estimation of Gaussian structural equation models with same error variances can be done using maximum likelihood with the BIC-penalty, in analogy to Chickering [2002]. The search, however, should be done in the space of directed acyclic graphs rather than Markov equivalence classes. This method is not shown in this thesis, but has been analyzed empirically in a master thesis [Tanase, 2012].
We have pointed out in Section 11.2 that working with p-values for a model check on real world data sets introduces some problems. If we observe more and more data, even slightest deviations of the data generating process and the model class become apparent and we have to reject all structures.

Potential Benefits In our opinion, the SEM-based approach comes with the following benefits:

1. We can identify the true causal graph even within the Markov equivalence class.

2. We can use restricted SEMs to identify non-faithful causal models (even those "undetectable" versions of unfaithfulness mentioned in Section 2.7.1), for which conditional independence-based methods usually fail.

3. If the true data generating mechanism satisfies all assumptions, the SEM-based approach seems to perform better than standard

independence-based methods. This can only occur for finitely many samples and should be investigated further.

4. If the data do not allow for any representation as a restricted SEM, the methods are able to remain undecided instead of making a wrong decision.

5. An SEM contains more information than the corresponding causal Markov DAG: some counterfactual statements can only be deduced from the SEM, see Example 2.5. In this example, the two different SEMs give different answers to a counterfactual statement, but lead to the same observational and interventional distributions. It remains open, whether this point has practical implications.

6. Applied to time series data, the SEM-based method (TiMINo) lead to an identifiability result that is more general than existing results. The algorithm is applicable to multivariate, linear, nonlinear and instantaneous interactions and can also discover partial structures. It also allows for the ability to make no decision instead of a wrong one.

12.2. Future Work

High-dimensional Methods We do not regard the multivariate algorithm we presented as an optimal method. Even for data sets with a large sample size the methods we present are restricted to less than 10 variables. In many applications there are even more random variables (e.g. genes) than samples (e.g. replicates). Such a setting with $p \gg n$ is usually referred to as being high-dimensional. For independence-based methods, both theoretical [Kalisch and Bühlmann, 2007] and methodological problems have been addressed [e.g. IDA by Maathuis et al., 2009, 2010]. Restricted structural equation models, however, have only been applied to low-dimensional data sets. Developing methodology, algorithms and theoretical understanding constitute main goals of future research, see Figure 12.1. We believe that score-based methods may play a major role here and may even outperform all existing methods in a low-dimensional setting.

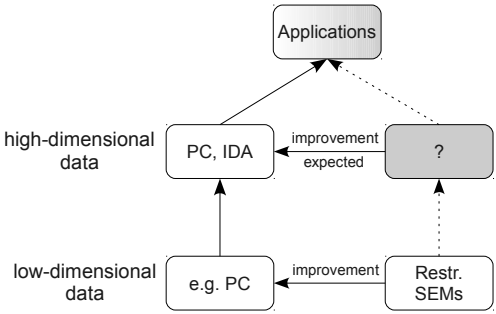


Figure 12.1.: It would be beneficial to "lift" the proposed methods to a high-dimensional setting with $p \gg n$.

Interventional Data In general, interventional data is hard to obtain, and we thus exploit methods that try to infer the causal structure from observational data only. For some applications, however, both observational and interventional data are available. Since interventional data often helps for identifying the causal structure, it is beneficial to make use of all available data. Hauser and Bühlmann [2012] propose a score-based inference with precisely that goal for Gaussian data. We believe that an analogous method for restricted SEMs could be developed.

Distance to Restricted SEMs Until now we decide for a causal structure if it is the only one, for which we can fit the structural equations that lead to independent residuals, and remain undecided if we cannot fit any restricted SEM. We believe that it is valuable to investigate the following principle for causal inference: we also decide for a graph structure if the observed empirical distribution is *reasonably closer* to the set of distributions that are generated by this structure than to the set of distributions for all other structures (e.g. the KL divergence to the subset $\mathrm{ANM}_{X\to Y}$ is smaller than to the subset $\mathrm{ANM}_{Y\to X}$, see Section 5.3.3). Clearly, computing the distances to those sets of distributions and quantifying what *reasonably* closer means are challenges that need to be addressed.

More Experiments with Ground Truth We believe that it is crucial to test causal inference methods on real data with known ground truth. Although we have obtained promising results on real world data sets in this thesis, an extensive evaluation of methods on even more real data sets is necessary (both for i.i.d and time series data). After all, exhaustive experiments may show that the assumptions current methods for causal inference are based on are most often not met in nature.
For the case of two variables there is a collection of data sets available at `https://webdav.tuebingen.mpg.de/cause-effect/`. We have used this data set in Figure 11.5. It would be beneficial to have a similar collection of benchmark data sets for more than two variables in order to be able to compare more causal inference methods (recall that independence- and score-based methods do not work on the two variable case). The same applies for time series data. We hope that the community can be enhanced by challenges as organized by `http://www.chalearn.org/`.

Discrete Variables We believe that for the cyclic case even stronger identifiability statements than the ones we proved may hold.
Further, our method can be developed in different directions: Since it is known that χ^2 fails for small data sizes, changing the independence test for those cases may lead to a higher performance of the algorithm. Handling more than two variables is straightforward and already implemented; one may have to introduce regularization to make the regression computationally feasible, though. Especially for discrete data it is necessary to work on practically feasible extensions of additive noise and come up with other ways of restricting SEMs. It could be investigated how the procedure can be applied to data sets that consist of discrete and continuous variables.

Confounder There are different ways of improving the proposed algorithm. The regression step can be changed such that the regression function is not chosen by minimizing an ℓ_2 loss, but rather by minimizing the dependencies of the residuals, which is somewhat more consistent in our case [Mooij et al., 2009]. Therefore a stable regression method is used that can also be based on HSIC. It may also be possible

to improve the performance by a better way of choosing the bandwidth of the Gaussian kernel. A complete identifiability result in the style of the other chapters, however, would clearly be desirable, along with further experimental evidence.

Time Series Regarding time series, we think that the following investigations would be worthwhile: In additive heteroscedastic models the innovation variance for a node may depend on the value of the node's parents

$$X_t^i = f_i\big(\mathbf{PA}(X_t^i)\big) + \sigma\big(\mathbf{PA}(X_t^i)\big) \cdot N_t^i \,.$$

A generalization to such models and preprocessing the data (removing trends, periodicities, etc.) may decrease the number of cases where TiMINo causality is undecided. Checking for autocorrelations in the residuals is another possible model check and not included yet. In the case of non-instantaneous feedback loops, one should find a method to fit the model structure that is faster than brute-force search. TiMINo causality evaluates a model fit by checking independence of the residuals. Again, one may make the independence of the residuals as a criterion for the fitting process or at least for order selection [Mooij et al., 2009].

Partial Inference We would like to analyze situations, where parts of the graph satisfy the assumptions (e.g. is generated by a restricted SEM) and other parts do not. Preliminary experiments (as shown in Section 11.4) show that some parts of the graph remain identifiable.

Although many questions remain open, we regard our work as a small step towards understanding and detecting the traces that the underlying causal structure of some data generating process leaves in the data.

Appendix A.

Proofs

A.1. Proofs of Chapter 1

A.1.1. Proof of Proposition 1.4

Proof. Assume there is another true causal DAG $\mathcal{G}_1$. Consider any node X_i and denote the $\mathcal{G}$-parents by $X_{j_1}, \ldots, X_{j_r}$ and the $\mathcal{G}_1$-parents by $X_{k_1}, \ldots, X_{k_s}$. We have for all $x_{j_1}, \ldots, x_{j_r}$ and $x_{k_1}, \ldots, x_{k_s}$ that

$$\begin{aligned} & p(X_i \,|\, X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r}) \\ &= p(X_i \,|\, do(X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r}, X_{k_1} = x_{k_1}, \ldots, X_{k_s} = x_{k_s}) \\ &= p(X_i \,|\, X_{k_1} = x_{k_1}, \ldots, X_{k_s} = x_{k_s}) \end{aligned}$$

Because of causal minimality this is only possible if the set of parents is exactly the same. □

A.2. Proofs of Chapter 2

A.2.1. Proof of Proposition 2.6

Proof. Let $N_1, \cdots, N_p$ be independent and uniformly distributed between 0 and 1. We then define $X_j = f_j(X_{\mathbf{PA}_j}, N_j)$ with

$$f_j(x_{\mathbf{PA}_j}, n) = F^{-1}_{X_j | X_{\mathbf{PA}_j} = x_{\mathbf{PA}_j}}(n)$$

where $F_{X_j | X_{\mathbf{PA}_j} = x_{\mathbf{PA}_j}}$ is the inverse cdf from X_j given $X_{\mathbf{PA}_j} = x_{\mathbf{PA}_j}$. □

A.3. Proofs of Chapter 4

A.3.1. Proof of Theorem 4.1

Proof. Set

$$\pi(x,y) := \log p(x,y) = \nu(y - f(x)) + \xi(x)\,, \tag{A.1}$$

and $\tilde{\nu} := \log p_{\tilde{n}}$, $\eta := \log p_y$. If Equation (4.4) holds, then $\pi(x,y) = \tilde{\nu}(x - g(y)) + \eta(y)$, implying

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\tilde{\nu}''(x - g(y))g'(y) \quad \text{and} \quad \frac{\partial^2 \pi}{\partial x^2} = \tilde{\nu}''(x - g(y))\,.$$

We conclude

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 \pi / \partial x^2}{\partial^2 \pi / (\partial x \partial y)}\right) = 0\,. \tag{A.2}$$

Using Equation (A.1) we obtain

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\nu''(y - f(x))f'(x)\,, \tag{A.3}$$

and

$$\frac{\partial^2 \pi}{\partial x^2} = \frac{\partial}{\partial x}\left(-\nu'(y - f(x))f'(x) + \xi'(x)\right) = \nu''(f')^2 - \nu' f'' + \xi''\,, \tag{A.4}$$

where we have dropped the arguments for convenience. Combining Equations (A.3) and (A.4) yields

$$\begin{aligned}\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}}\right) = &-2f'' + \frac{\nu' f'''}{\nu'' f'} - \xi''' \frac{1}{\nu'' f'} + \frac{\nu' \nu''' f''}{(\nu'')^2} \\ &- \frac{\nu'(f'')^2}{\nu''(f')^2} - \xi'' \frac{\nu'''}{(\nu'')^2} + \xi'' \frac{f''}{\nu''(f')^2}\,.\end{aligned}$$

Due to Equation (A.2) this expression must vanish and we obtain DE (4.5)

$$\begin{aligned}\xi''' = \xi''\left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu'' f'' f' \\ + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}\,,\end{aligned} \tag{A.5}$$

by term reordering. Given f, ν, we obtain for every fixed y a linear inhomogeneous DE for ξ:

$$\xi'''(x) = \xi''(x)G(x,y) + H(x,y)\,, \tag{A.6}$$

where G and H are defined by

$$G := -\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}$$

and

$$H := -2\nu'' f'' f' + \nu' f''' + \frac{\nu'\nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}\,.$$

Setting $z := \xi''$ we have $z'(x) = z(x)G(x,y) + H(x,y)$. Given that such a function z exists, it is given by

$$z(x) = z(x_0)e^{\int_{x_0}^{x} G(\tilde{x},y)d\tilde{x}} + \int_{x_0}^{x} e^{\int_{\hat{x}}^{x} G(\tilde{x},y)d\tilde{x}} H(\hat{x},y)d\hat{x}\,. \tag{A.7}$$

Let y be fixed such that $\nu''(y - f(x))f'(x) \neq 0$ holds for all but countably many x. Then z is determined by $z(x_0)$ since we can extend Equation (A.7) to the remaining points. The set of all functions ξ satisfying the linear inhomogenous DE (A.6) is a 3-dimensional affine space: Once we have fixed $\xi(x_0), \xi'(x_0), \xi''(x_0)$ for some arbitrary point x_0, ξ is completely determined. Given fixed f and ν, the set of all ξ admitting a backward model is contained in this subspace. □

A.3.2. Proof of Corollary 4.2

Proof. Similarly to how (A.2) was derived, under the assumption of the existence of a reverse model one can derive

$$\frac{\partial^2\pi}{\partial x\partial y} \cdot \frac{\partial}{\partial x}\left(\frac{\partial^2\pi}{\partial x^2}\right) = \frac{\partial^2\pi}{\partial x^2} \cdot \frac{\partial}{\partial x}\left(\frac{\partial^2\pi}{\partial x\partial y}\right)$$

Now using (A.3) and (A.4), we obtain

$$\begin{aligned}(-\nu'' f')\cdot\frac{\partial}{\partial x}\left(\nu''(f')^2 - \nu' f'' + \xi''\right)& \\ = (\nu''(f')^2 - \nu' f'' + \xi'') \cdot \frac{\partial}{\partial x}\left(-\nu'' f'\right)&\end{aligned}$$

which reduces to

$$\begin{aligned}&-2(\nu'' f')^2 f'' + \nu'' f' \nu' f''' - \nu'' f' \xi''' \\ &\quad = -\nu' f'' \nu''' (f')^2 + \xi'' \nu''' (f')^2 + \nu'' \nu' (f'')^2 - \nu'' f'' \xi'' \,.\end{aligned}$$

Substituting the assumptions $\xi''' = 0$ and $\nu''' = 0$ (and hence $\nu'' = C$ everywhere with $C \neq 0$ since otherwise ν cannot be a proper log-density) yields

$$\nu'\big(y - f(x)\big) \cdot \big(f' f''' - (f'')^2\big) = 2C(f')^2 f'' - f'' \xi''.$$

Since $C \neq 0$ there exists an α such that $\nu'(\alpha) = 0$. Then, restricting ourselves to the submanifold $\{(x, y) \in \mathbb{R}^2 : y - f(x) = \alpha\}$ on which $\nu' = 0$, we have

$$0 = f''(2C(f')^2 - \xi'').$$

Therefore, for all x in the open set $[f'' \neq 0]$, we have $(f'(x))^2 = \xi''/(2C)$ which is a constant, so $f'' = 0$ on $[f'' \neq 0]$: a contradiction. Therefore, $f'' = 0$ everywhere. □

A.4. Proofs of Chapter 5

A.4.1. Proof of Theorem 5.3

Proof.

⇒: First we assume $\operatorname{supp} Y = \{y_0, \ldots, y_m\}$ with $y_0 < y_1 < \ldots < y_m$. This implies that $N_{\max} := \min\{n \in \mathbb{N} \,|\, \mathbb{P}(N = n) > 0\}$ is finite. Define the non-empty sets $\tilde{C}_i := \operatorname{supp} X | Y = y_i$, for $i = 0, \ldots, m$. That means $\tilde{C}_0, \ldots, \tilde{C}_m \subset \operatorname{supp} X$ are the smallest sets satisfying $\mathbb{P}(X \in \tilde{C}_i \,|\, Y = y_i) = 1$. For all i, j it follows that

$$\tilde{C}_i = \tilde{C}_j \text{ or } \tilde{C}_i \cap \tilde{C}_j = \emptyset \text{ and } f\,|_{\tilde{C}_i} = \tilde{c}_i = \text{const.} \tag{A.8}$$

This is proved by an induction argument.

Base step: Consider $\tilde{C}_m$ corresponding to the largest value $y_m = \max\{f(x) \,|\, x \in X\} + N_{\max}$ of $\operatorname{supp} Y$. Assuming $f(x_1) < f(x_2)$ for $x_1, x_2 \in \tilde{C}_m$ leads to $y_m = f(x_1) + N_{\max} < f(x_2) + N_{\max} = y_m$ and therefore to a contradiction. Induction

step: Consider $\tilde{C}_k$ and assume properties (A.8) are satisfied for all $\tilde{C}_{\tilde{k}}$ with $k < \tilde{k} \leq m$. If $x \in \tilde{C}_k \cap \tilde{C}_{\tilde{k}}$ for some $\tilde{k}$

$$\Rightarrow \mathbb{P}(N = y_k - f(\tilde{x})) = \mathbb{P}(N = y_k - f(x)) > 0 \quad \forall \tilde{x} \in \tilde{C}_{\tilde{k}}$$
$$\Rightarrow \tilde{C}_{\tilde{k}} \subset \tilde{C}_k \quad \Rightarrow \tilde{C}_{\tilde{k}} = \tilde{C}_k \quad \Rightarrow f\,|_{\tilde{C}_k} = f\,|_{\tilde{C}_{\tilde{k}}} = \text{const}$$

Furthermore, if $\tilde{C}_k \cap \tilde{C}_{\tilde{k}} = \emptyset \, \forall k < \tilde{k} \leq m$, then $f\,|_{\tilde{C}_k} = \text{const}$ using the same argument as for C_m.

Thus we can choose some sets $C_0, \ldots, C_l$ from $\tilde{C}_0, \ldots, \tilde{C}_m$, where $l \leq m$, such that $C_0, \ldots, C_l$ are disjoint, and $c_k := f(C_k)$ are pairwise different values. Without loss of generality assume $C_0 = \tilde{C}_0$. Further, even the sets

$$c_k + \operatorname{supp} N := \{c_k + h \,:\, \mathbb{P}(N = h) > 0\}$$

are pairwise different: If $y_i = c_k + h_1 = c_l + h_2$ then $C_k \subset \operatorname{supp}(X|Y = y_i) = \tilde{C}_i$ and $C_l \subset \tilde{C}_i$, which implies $k = l$.

Now consider the case where Y has infinite and X finite support: $\operatorname{supp} X = \{x_0, \ldots, x_p\}$. Then we define $C_0, \ldots, C_l$ to be disjoint sets, such that f is constant on each of them: $c_i := f(C_i)$. This time, it does not matter which of these sets is called C_0. Again, we will deduce that the sets $c_k + \operatorname{supp} N$ are disjoint:

The sets $\tilde{D}_i := \operatorname{supp} Y|X = x_i$ fulfill

$$\tilde{D}_i = \tilde{D}_j \text{ or } \tilde{D}_i \cap \tilde{D}_j = \emptyset \text{ and } g\,|_{\tilde{D}_i} = \tilde{d}_i = \text{const}.$$

Thus we have $c_k + \operatorname{supp} N$ and $c_k + \operatorname{supp} N$ are either equal or disjoint. But if $c_k + \operatorname{supp} N = c_l + \operatorname{supp} N$ for $k \neq l$ it follows for $x_a \in C_k$, $x_b \in C_l$ and all $y \in c_k + \operatorname{supp} N$ (since there is a backward model $X = g(Y) + \tilde{N}$)

$$\begin{aligned} & \frac{\mathbb{P}(X = x_a, Y = y)}{\mathbb{P}(X = x_b, Y = y)} = const \\ \Rightarrow & \frac{\mathbb{P}(X = x_a) \cdot \mathbb{P}(N = y - f(x_a))}{\mathbb{P}(X = x_b) \cdot \mathbb{P}(N = y - f(x_b))} = const \\ \Rightarrow & \frac{\mathbb{P}(N = y - f(x_a))}{\mathbb{P}(N = y - f(x_b))} = const \end{aligned}$$

and thus $\mathbb{P}(N = 0)/\mathbb{P}(N = r) = const, \forall r \in \operatorname{supp} N$. This is only possible for a uniformly distributed N, which leads to a contradiction since Y has been assumed to have infinite support.

Thus we have proved condition c). For a) it remains to show that the sets C_i are shifted versions of each other. This part of the proof is valid for both cases (either X or Y has finite support): Consider C_i for any i. According to the assumption that an ANM $Y \to X$ holds we have

$$\begin{aligned} & \tilde{N}|Y = c_0 \stackrel{\mathcal{L}}{=} \tilde{N}|Y = c_i \\ \Leftrightarrow \quad & X - g(c_0)|Y = c_0 \stackrel{\mathcal{L}}{=} X - g(c_i)|Y = c_i \\ \Rightarrow \quad & X + d_i|Y = c_0 \stackrel{\mathcal{L}}{=} X|Y = c_i \qquad (*) \end{aligned}$$

with $d_i = g(c_i) - g(c_0)$. Thus $C_i = C_0 + d_i$ (including $d_0 = 0$), which completes conditions a).

To prove b) observe that we have for all $x \in C_i$

$$\begin{aligned} \frac{\mathbb{P}(X = x)}{\mathbb{P}(X \in C_i)} &= \frac{\mathbb{P}(X = x)\mathbb{P}(N = c_i - f(x))}{\sum_{\tilde{x} \in C_i} \mathbb{P}(X = \tilde{x})\mathbb{P}(N = c_i - f(\tilde{x}))} \\ &= \frac{\mathbb{P}(X = x, N = c_i - f(x))}{\mathbb{P}(Y = c_i)} = \mathbb{P}(X = x \,|\, Y = c_i) \\ &\stackrel{(*)}{=} \mathbb{P}(X = x - d_i \,|\, Y = c_0) \\ &= \frac{\mathbb{P}(X = x - d_i, N = c_0 - f(x - d_i))}{\mathbb{P}(Y = c_0)} \\ &= \frac{\mathbb{P}(X = x - d_i)}{\mathbb{P}(X \in C_0)} \end{aligned}$$

$\Leftarrow$: In order to show that we have a reversible ANM, we have to construct a g, such that $X = g(Y) + \tilde{N}$. Therefore define the function g as follows: $g(y) = 0, \forall y \in c_0 + \operatorname{supp} N$ and $g(y) = d_i, \forall y \in c_i + \operatorname{supp} N, i > 0$. (This is well-defined because of a) and c).) The noise $\tilde{N}$ is determined by the joint distribution $\mathcal{L}(X, Y)$, of course. It remains to check, whether the distribution of $\tilde{N}|Y = y$ is independent of y. Consider a fixed y

and choose i such that $y \in c_i + \operatorname{supp} N$. Since $C_i = C_0 + d_i$ the condition $g(y) + h \in C_i$ is satisfied for all $h \in C_0$ and therefore independently of y and c_i. Now, if $g(y) + h \in C_i$ we have

$$\begin{aligned}
\mathbb{P}(\tilde{N} = h \,|\, Y = y) &= \frac{\mathbb{P}(X = g(y) + h, Y = y)}{\mathbb{P}(Y = y)} \\
&= \frac{\mathbb{P}(X = g(y) + h, N = y - f(g(y) + h))}{\mathbb{P}(Y = y)} \\
&= \frac{\mathbb{P}(X = g(y) + h)\mathbb{P}(N = y - c_i)}{\sum_{\tilde{x} \in C_i} \mathbb{P}(X = \tilde{x})\mathbb{P}(N = y - f(\tilde{x}))} \\
&= \frac{\mathbb{P}(X = g(y) + h)}{\mathbb{P}(X \in C_i)} = \frac{\mathbb{P}(X = g(y) + h - d_i)}{\mathbb{P}(X \in C_0)} \\
&= \frac{\mathbb{P}(X = h)}{\mathbb{P}(X \in C_0)}
\end{aligned}$$

which does not depend on y. And if $g(y) + h \notin C_i$ then $\mathbb{P}(\tilde{N} = h \,|\, Y = y) = 0$, which does not depend on y either.

□

A.4.2. Proof of Theorem 5.5

Proof. We distinguish between two different cases:

a) $\mathbb{P}(N = k) > 0 \,\forall\, m \leq k \leq l$ and $\mathbb{P}(N = k) = 0$ for all other k.

⇒: Assume that there is an ANM in both directions $X \to Y$ and $Y \to X$. As mentioned above we have a freedom of choosing an additive constant for the regression function. In the remainder of this proof we require $\mathbb{P}(N = k) = \mathbb{P}(\tilde{N} = k) = 0 \,\forall k < 0$ and $\mathbb{P}(\tilde{N} = 0), \mathbb{P}(N = 0) > 0$. The largest k, such that $\mathbb{P}(N = k) > 0$ will be called $N_{\max}$. In analogy to the proof above we define $C_y := \operatorname{supp} X|Y = y$ for all $y \in \operatorname{supp} Y$.

At first we note that all C_y are shifted versions of each other (since there is a backward ANM) and additionally, they are finite sets (otherwise it follows from the compact support of N that there are infinitely many infinite sets $f^{-1}(f(x))$ on which f is constant, which contradicts the assumptions.)

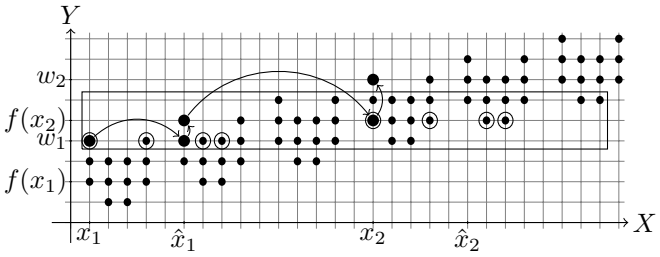


Figure A.1.: Visualization of the path from equation (A.9). Here, $w_2 = f(x_2) + N_{\max}$ and $w_1 = f(x_1) + N_{\max}$.

Start with any x_1 that satisfies $x_1 = \min\{f^{-1}(f(x_1))\}$ and define

$$\hat{x}_1 := \min\left\{x \in C_{f(x_1)+N_{\max}} \setminus f^{-1}(f(x_1))\right\}$$

This implies $f(\hat{x}_1) > f(x_1)$ and $x_1 \in C_{f(\hat{x}_1)}$.

If such an $\hat{x}_1$ does not exist because the set on the right hand side is empty, then it cannot exist for any choice of x_1: It is clear that $C_{f(x_1)+N_{\max}} = f^{-1}(f(x_1))$ and then we consider the first $C_{f(x_1)+N_{\max}+i}$ that is not empty. Then this set must be $f^{-1}(f(\hat{x}_1))$ for some $\hat{x}_1$. This leads to an iterative procedure and to the required decomposition of $\operatorname{supp} X$.

We have that either

$$\max\{f^{-1}(f(\hat{x}_1))\} > \max\{f^{-1}(f(x_1))\} \text{ or}$$
$$\min\{f^{-1}(f(\hat{x}_1))\} < \min\{f^{-1}(f(x_1))\} :$$

Otherwise $C_{f(\hat{x}_1)}$ and $C_{f(\hat{x}_1)-1}$ satisfy

$$\max C_{f(\hat{x}_1)-1} \geq \max C_{f(\hat{x}_1)} \qquad \text{and}$$
$$\min C_{f(\hat{x}_1)-1} \leq \min C_{f(\hat{x}_1)} .$$

Because of $\hat{x}_1 \in C_{f(\hat{x}_1)}, \hat{x}_1 \notin C_{f(\hat{x}_1)-1}$ this contradicts the existence of an backward ANM. We therefore assume without loss of generality $\max\{f^{-1}(f(\hat{x}_1))\} > \max\{f^{-1}(f(x_1))\}$.

Then we even have $\hat{x}_1 > x_1$, $x_1 = \min\{C_{f(x_1)+N_{\max}}\}$ and $\hat{x}_1 = \min\{C_{f(x_1)+N_{\max}+1}\}$. (Otherwise we use the same argument as above with $C_{f(x_1)+N_{\max}}$ and $C_{f(x_1)+N_{\max}+1}$.) Define further

$$x_2 := \min f^{-1}(f(x_1) + N_{\max} + 1)$$

Since

$$f^{-1}(f(x_1)) \subset C_{f(x_1)+N_{\max}}\,,$$

but

$$f^{-1}(f(x_1)) \cap C_{f(x_1)+N_{\max}+1} = \emptyset\,,$$

such a value must exist. Again, we can define $\hat{x}_2$ in the same way as above.

Set $y_1 := f(x_1) + N_{\max}$ and $z_1 := f(x_1) + 2 \cdot N_{\max}$ and consider the finite box from $(\min C_{y_1}, y_1)$ to $(\max C_{z_1}, z_1)$. This box contains all the support from $X \,|\, Y = f(x_1) + N_{\max} + i$, where $i = 0, \ldots, N_{\max}$. Assume we know the positions in this box, where $\mathcal{L}(X, Y)$ is larger than zero. Then this box determines the support of $X \,|\, Y = f(x_1) + 2 \cdot N_{\max} + 1$ (the line above the box) just using the support of N and $\tilde{N}$. Iterating gives us the whole support of $\mathcal{L}(X, Y)$ in the box above (from $y_2 = f(x_2) + N_{\max}$ to $z_2 = f(x_2) + 2 \cdot N_{\max}$). Since the width of the boxes are bounded by $3 \cdot \max C_{f(x_1)} - \min C_{f(x_1)}$, for example, at some point the box of x_n must have the same support as the one of x_1. Figure A.1 shows an example, in which $n = 2$. Using only the distributions of N and $\tilde{N}$ we can now determine a factor α for which $\mathbb{P}(X = x_1, Y = f(x_1) + N_{\max}) = \alpha \cdot \mathbb{P}(X = x_n, Y = f(x_n) + N_{\max})$ This is done by following a sequence between (x_1, y_1) and (x_n, y_n) using only horizontal and vertical steps:

$$\begin{aligned}(x_1, y_1), (\hat{x}_1, y_1), (\hat{x}_1, f(x_2)), (x_2, f(x_2)),\\ (x_2, y_2), (\hat{x}_2, y_2), \ldots, (x_n, y_n)\end{aligned} \tag{A.9}$$

(cf Figure A.1). Since this factor only depends on the distributions of N and $\tilde{N}$, the same α satisfies $\mathbb{P}(X =$

$x_n, Y = f(x_n) + N_{\max}) = \alpha \cdot \mathbb{P}(X = x_{2n-1}, Y = f(x_{2n-1}) + N_{\max})$ and therefore

$$\mathbb{P}(X = x_1, Y = f(x_1) + N_{\max}) = \alpha^k \cdot \mathbb{P}(X = x_{(k+1)n-k}, Y = f(x_{(k+1)n-k}) + N_{\max})$$

Note that a corresponding equation with the same constant α holds for the direction to the left of x_1. This leads to a contradiction, since there is no probability distribution for X with infinite support that can fulfill this condition (no matter if α is greater, equal or smaller than 1).

$\Leftarrow$: This direction is proved in exactly the same way as in Theorem 5.3.

b) $\mathbb{P}(N = k) > 0 \,\forall\, k \in \mathbb{Z}$.
Since X and Y are dependent there are y_1 and y_2, such that $g(y_1) \neq g(y_2)$ with g being the "backward function". Comparing $\{\mathbb{P}(X = k, Y = y_1), k \geq m\}$ and $\{\mathbb{P}(X = k, Y = y_2), k \geq m\}$ we can identify the difference $d := g(y_2) - g(y_1)$. Wlog consider $d > 0$. We use $\frac{\mathbb{P}(X=m-1,Y=y_1)}{\mathbb{P}(X=m,Y=y_1)} = \frac{\mathbb{P}(X=m+d-1,Y=y_2)}{\mathbb{P}(X=m+d,Y=y_2)}$ in order to determine $\mathbb{P}(X = m-1, Y = y_1)$ and then $\mathbb{P}(X = m-1)$ (using f and $\mathcal{L}(N)$). Iterations lead to all $\mathbb{P}(X = x)$.

□

A.4.3. Proof of Theorem 5.9

Each distribution $Y \,|\, X = x_j$ has to have the same support (up to an additive shift) and thus the same number of elements with probability larger than 0: $\#\text{supp}\, X \cdot \#\text{supp}\, N = k \cdot \#\text{supp}\, Y$. This proves (i). For (ii) we now consider 3 different cases: 1. f and g are bijective, 2. g is not injective and 3. f is not injective. These three cases are sufficient since f and g injective implies $n = m$ and f and g bijective. For each of those cases we show that a necessary condition for reversibility includes at least one additional equality constraint for $\mathcal{L}(X)$ or $\mathcal{L}(N)$.

1st case: f and g are bijective.

Proposition A.1 *Assume $Y = f(X) + N$, $N \perp\!\!\!\perp X$ for bijective f and $n(l) \neq 0, p(k) \neq 0 \,\forall k, l$. If the model is reversible with a bijective g, then X and Y are uniformly distributed.*

Proof. Since g is bijective we have that $\forall y \exists t_y : g(t_y) = g(y) - 1$. From (5.2) we can deduce

$$\frac{n\big(y - f(x+1)\big)p(x+1)}{n\big(t_y - f(x)\big)p(x)} = \frac{\tilde{n}\big(x + 1 - g(y)\big)q(y)}{\tilde{n}\big(x + 1 - g(y)\big)q(t_y)}$$

which implies

$$\frac{p(x+1)}{p(x)} = \frac{n\big(t_y - f(x)\big)q(y)}{n\big(y - f(x+1)\big)q(t_y)} \quad \text{and}$$

$$1 = \frac{p(x+m)}{p(x)} = \frac{\prod_{k=0}^{m-1} n\big(t_y - f(x+k)\big)q(y)^m}{\prod_{k=0}^{m-1} n\big(y - f(x+k+1)\big)q(t_y)^m}$$

Since f is bijective it follows that $q(y) = q(t_y)$. This holds for all y and thus Y and X are uniformly distributed. □

2nd case: g is not injective.
Assume $g(y_0) = g(y_1)$. From (5.2) it follows that

$$\frac{n\big(y_0 - f(x)\big)}{n\big(y_1 - f(x)\big)} = \frac{q(y_0)}{q(y_1)} \,\forall x$$

and thus

$$\frac{n\big(y_0 - f(x)\big)}{n\big(y_1 - f(x)\big)} = \frac{n\big(y_0 - f(\tilde{x})\big)}{n\big(y_1 - f(\tilde{x})\big)} \,\forall x, \tilde{x},$$

which imply equality constraints on n. To determine the number of constraints we define a function that maps the arguments of the numerator to those of the denominator

$$h_{y_0, y_1, f} : \begin{array}{ccc} \mathrm{Im}(y_0 - f) & \to & \mathbb{Z}/\tilde{m}\mathbb{Z} \\ y_0 - f(x) & \mapsto & y_1 - f(x) \end{array} .$$

We say h has a cycle if there is a $z \in \mathbb{N}$, s.t. $h^k(a) = (h \circ \ldots \circ h)(a) \in \mathrm{Im}(y_0 - f) \,\forall k \leq z$ and $h^z(a) = a$. For example: $2 \overset{h}{\mapsto} 4 \overset{h}{\mapsto} 6 \overset{h}{\mapsto} 0 \overset{h}{\mapsto} 2$.

Proposition A.2 *Assume $Y = f(X) + N$, $N \perp\!\!\!\perp X$ and $n(l) \neq 0, p(k) \neq 0 \,\forall k, l$. Assume further that the model is reversible with a non-injective g.*

- *If h has at least one cycle, $\#\mathrm{Im}f - \#cycles + 1$ parameters of n are determined by the others.*
- *If h has no cycles, $\#\mathrm{Im}f$ parameters of n are determined by the others.*

Proof. Assume h has a cycle of length r: $n_1 \overset{h}{\mapsto} n_2 \overset{h}{\mapsto} \ldots \overset{h}{\mapsto} n_r \overset{h}{\mapsto} n_1$ (here, $y_0 - n_1, \ldots, y_0 - n_r \in \mathrm{Im}f$), then $\frac{q(y_0)}{q(y_1)} = 1$ because $\frac{q(y_0)^r}{q(y_1)^r} = \frac{n(n_1)}{n(n_2)} \cdot \frac{n(n_2)}{n(n_3)} \cdots \frac{n(n_r)}{n(n_1)} = \frac{n(n_1)}{n(n_1)} = 1$ and $n\big(y_0 - f(x)\big) = n\big(y_1 - f(x)\big) \,\forall x$, that is $n(n_1) = n(n_2) = \ldots = n(n_r)$. Thus we get $r - 1$ equality constraints for each cycle of length r. For any (additional) non-cyclic structure of length r: $n_1 \mapsto n_2 \mapsto \ldots \mapsto n_r$ and $n_r \notin \mathrm{Im}(y_0 - f)$ (here, $y_0 - n_1, \ldots, y_0 - n_{r-1} \in \mathrm{Im}f$), we have $n(n_1) = \ldots = n(n_r)$ and thus $r - 1$ equality constraints. Together with the normalization these are $\#\mathrm{Im}f - \#\mathrm{cycles} + 1$ constraints.

If h has no cycle, we have $\#\mathrm{Im}f - 1$ independent equations plus the sum constraint. E.g.: $\frac{n(2)}{n(4)} = \frac{n(4)}{n(6)} = \frac{n(3)}{n(5)}$ implies $n(4) = n(6)\frac{n(3)}{n(5)}$ and $n(2) = \frac{n(4)^2}{n(6)}$. Further,

$$\frac{n\big(y_0 - f(x)\big)}{n\big(y_1 - f(x)\big)} = \frac{q(y_0)}{q(y_1)} = \frac{\sum_{\tilde{x}} p(\tilde{x}) n(y_0 - f(\tilde{x})}{\sum_{\tilde{x}} p(\tilde{x}) n(y_1 - f(\tilde{x}))}$$

introduces a functional relationship between p and n. □

Note that if $\tilde{m}$ does not have any divisors, there are no cycles and thus $\#\mathrm{Im}f$ parameters of n are determined. We have the following corollary

Corollary A.3 *In all cases the number of fixed parameters is lower bounded by $\lceil 1/2 \cdot \#\mathrm{Im}f \rceil + 1 \geq 2$.*

3rd case: f is not injective.

Assume $f(x_0) = f(x_1)$. In a slight abuse of notation we write

$$g - g : \begin{array}{ccc} \mathbb{Z}/\tilde{m}\mathbb{Z} \times \mathbb{Z}/\tilde{m}\mathbb{Z} & \to & \mathbb{Z}/m\mathbb{Z} \\ (y, \tilde{y}) & \mapsto & g(y) - g(\tilde{y}) \end{array} .$$

Similar as above, we define

$$h_{x_0,x_1,g} : \begin{array}{ccc} \mathrm{Im}\big(x_0 - (g-g)\big) & \to & \mathbb{Z}/m\mathbb{Z} \\ x_0 - g(y) + g(\tilde{y}) & \mapsto & x_1 - g(y) + g(\tilde{y}) \end{array} .$$

We say that h has a cycle if there is a $z \in \mathbb{N}$, s.t. $h^k(a) = (h \circ \ldots \circ h)(a) \in \mathrm{Im}\big(x_0 - (g-g)\big)\ \forall k \leq z$ and $h^z(a) = a$.

Proposition A.4 *Assume $Y = f(X) + N$, $N \perp\!\!\!\perp X$, f is not injective and $n(l) \neq 0, p(k) \neq 0\ \forall k, l$. Assume further that the model is reversible for a function g.*

- *If h has at least one cycle, $\#\mathrm{Im}(g-g) - \#cycles + 1$ parameters of p are determined by the others.*
- *If h has no cycles, $\#\mathrm{Im}(g-g)$ parameters of p are determined by the others.*

Proof. From (5.2) it follows that

$$\begin{aligned} \frac{p(x_0)}{p(x_1)} &= \frac{\tilde{n}\big(x_0 - g(y)\big)}{\tilde{n}\big(x_1 - g(y)\big)} \\ &= \frac{p\big(x_0 - g(y) + g(\tilde{y})\big) \cdot n\Big(\tilde{y} - f\big(x_0 - g(y) + g(\tilde{y})\big)\Big)}{p\big(x_1 - g(y) + g(\tilde{y})\big) \cdot n\Big(\tilde{y} - f\big(x_1 - g(y) + g(\tilde{y})\big)\Big)} \end{aligned}$$

for all $y, \tilde{y}$. The rest follows analogously to the proof of Proposition A.2. □

If $(x_1 - x_0)$ does not divide m, there are no cycles and thus $\#\mathrm{Im}(g-g)$ parameters of p are determined.

Corollary A.5 *In all cases the number of fixed parameters is lower bounded by $\lceil 1/2 \cdot \#\mathrm{Im}(g-g) \rceil + 1 \geq 2$.*

Remark A.6 Note that some of the constraints described above depend on the backward function g. This introduces no problems

because of the following reason: If we put any (prior) measure on the set of all possible parameters $p(0), p(1), \ldots, p(n-1)$ (or on $n(0), \ldots, n(m-1)$) that is absolutely continuous with respect to the Lebesgue measure, a single equality constraint reduces the set of possible parameters to a set of measure zero. There are only finitely many possibilities to choose the function g and thus even the union of all those parameter sets has measure zero.

A.5. Proofs of Chapter 6

Recall that we identify the node i with the variable X_i and the parents $\mathbf{PA}_i^{\mathcal{G}}$ with the variables $X_{\mathbf{PA}_i^{\mathcal{G}}}$. We further overload the notation of $\mathbf{X}$. It can either be the vector of random variables or the set.

A.5.1. Some Lemmata

We first provide and prove some lemmata.

Lemma A.7 *Let $Y \in \mathcal{Y}, N \in \mathcal{N}, Z \in \mathcal{Z}, S \in \mathcal{S}$ be random variables whose joint density is absolutely continuous with respect to some product measure (Z and S can be multivariate) and with density $p_{Y,Z,N,S}(y,z,n,s)$. Let $f : \mathcal{Y} \times \mathcal{Z} \times \mathcal{N} \to \mathbb{R}$ be a measurable function. If $N \perp\!\!\!\perp (Y,Z,S)$ then for all $z \in \mathcal{Z}, s \in S$ with $p_{Z,S}(z,s) > 0$:*

$$f(Y,Z,N)_{\mid Z=z,S=s} \stackrel{\mathcal{L}}{=} f(Y_{\mid Z=z,S=s}, z, N)$$

Proof of Lemma A.7 First, note that the joint of Y, Z, N, S satisfies:

$$p_{Y,Z,N,S}(y,z,n,s) = p_{Z,S}(z,s) p_{Y \mid Z=z,S=s}(y) p_N(n)$$

because $N \perp\!\!\!\perp (Y,Z,S)$. Consider $X := f(Y,Z,N)$. We have, for all $z \in \mathcal{Z}, s \in S$ with $p_{Z,S}(z,s) > 0$ and for all $x \in \mathcal{X}$:

$$\begin{aligned} p_{X \mid Z=z,S=s}(x) &= \frac{p_{X,Z,S}(x,z,s)}{p_{Z,S}(z,s)} \\ &= \frac{\int p_{Y,Z,N,S}(y,z,n,s) \delta\big(x - f(y,z,n)\big)\, dy\, dn}{p_{Z,S}(z,s)} \end{aligned}$$

$$= \int p_{Y \mid Z=z,S=s}(y) p_N(n) \delta\big(x - f(y,z,n)\big)\, dy\, dn$$
$$= p_{f(Y_{\mid Z=z,S=s},z,N)}(x)$$

Ergo, $X_{\mid Z=z,S=s} = f(Y_{\mid Z=z,S=s}, z, N)$ for all z, s with $p_{Z,S}(z,s) > 0$. □

Lemma A.8 *Let $\mathcal{L}(\mathbf{X})$ be generated by a functional model with corresponding DAG $\mathcal{G}$ and consider a random variable $X \in \mathbf{X}$. If $\mathbf{S} \subseteq \mathbf{ND}_X^{\mathcal{G}}$ then $N_X \perp\!\!\!\perp \mathbf{S}$.*

Proof of Lemma A.8 Write $\mathbf{S} = \{S_1, \ldots, S_k\}$. Then

$$\mathbf{S} = \big(f_{S_1}(\mathbf{PA}_{S_1}^{\mathcal{G}}, N_{S_1}), \ldots, f_{S_k}(\mathbf{PA}_{S_k}^{\mathcal{G}}, N_{S_k})\big)\,.$$

Again, one can substitute the parents of S_i by the corresponding functional equations and proceed recursively. After finitely many steps one obtains $\mathbf{S} = f(N_{T_1}, \ldots, N_{T_l})$, where $\{T_1, \ldots, T_l\}$ is the set of *all* ancestors of nodes in $\mathbf{S}$, which does not contain X. Since all noise variables are jointly independent we have $N_X \perp\!\!\!\perp \mathbf{S}$. □

To simplify notation, we restate Lemma 6.7.

Lemma A.9 *[same as Lemma 6.7] Consider an instance of an IFMOC with DAG $\mathcal{G}_0$, a variable B and one of its parents A. For all sets $\mathbf{S}$ with $\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}}$ we have*

$$B \not\perp\!\!\!\perp A \,|\, \mathbf{S} \tag{A.10}$$

Proof of Lemma A.9 According to Definition 6.4 we can choose $x_{\mathbf{S}}$, such that $p(x_{\mathbf{S}}) > 0$ and

$$\big(f_B(x_{\mathbf{PA}_B^{\mathcal{G}}\setminus\{A\}}, \underbrace{\cdot}_{A}, \underbrace{\cdot}_{N_B}), \mathcal{L}(A|_{X_{\mathbf{S}}=x_{\mathbf{S}}}), \mathcal{L}(N_B)\big) \in \mathcal{A}.$$

Because of $\mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}}$ and Lemma A.8 we can apply Lemma A.7, which gives $f_B(x_{\mathbf{PA}_B^{\mathcal{G}}\setminus\{A\}}, A|_{X_{\mathbf{S}}=x_{\mathbf{S}}}, N_B) = B|_{X_{\mathbf{S}}=x_{\mathbf{S}}}$.
But then (6.1) reads

$$A|_{X_{\mathbf{S}}=x_{\mathbf{S}}} \not\perp\!\!\!\perp f_B(x_{\mathbf{PA}_B^{\mathcal{G}}\setminus\{A\}}, A|_{X_{\mathbf{S}}=x_{\mathbf{S}}}, N_B) = B|_{X_{\mathbf{S}}=x_{\mathbf{S}}}$$

□

A.5.2. Some Propositions

Proposition A.10 *Assume $\mathcal{L}(\mathbf{X})$ is faithful and Markov with respect to $\mathcal{G} = (\mathbf{V}, \mathcal{E})$. If $\mathcal{L}(\mathbf{X})$ is induced by an SEM with corresponding DAG $\mathcal{G}' = (\mathbf{V}, \mathcal{E}')$, we have*

$$\#\{\text{edges in } \mathcal{G}\} \leq \#\{\text{edges in } \mathcal{G}'\}\,.$$

Further, when $\#\{\text{edges in } \mathcal{G}\} = \#\{\text{edges in } \mathcal{G}'\}$, $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent. (This result is used by [van de Geer and Bühlmann, 2012, Sec. 2.1.1].)

Proof of Proposition A.10 $\mathcal{L}(\mathbf{X})$ must be Markov with respect to $\mathcal{G}'$ and must thus satisfy $I_{\mathcal{G}'}$ (which stands for all (conditional) independences that are induced by the graph structure of $\mathcal{G}'$). $\mathcal{L}(\mathbf{X})$ must also satisfy $I_{\mathcal{G}}$ and since $\mathcal{L}(\mathbf{X})$ is faithful wrt $\mathcal{G}$, we have $I_{\mathcal{G}'} \subseteq I_{\mathcal{G}}$. Thus, $\{\text{missing edges in } \mathcal{G}'\} \supseteq \{\text{missing edges in } \mathcal{G}\}$ and therefore: $\#\{\text{edges in } \mathcal{G}\} \leq \#\{\text{edges in } \mathcal{G}'\}$. The rest follows immediately. □

Proof of Proposition 6.8 Suppose property (A.10) in Lemma A.9 does not hold. Then

$$\begin{aligned}
&\exists \mathbf{S} : \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}} \text{ and } B \perp\!\!\!\perp A \,|\, \mathbf{S} \\
\Rightarrow\ &\exists \tilde{\mathbf{S}} : B \perp\!\!\!\perp A \,|\, \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \cup \tilde{\mathbf{S}} \text{ and } B \perp\!\!\!\perp \tilde{\mathbf{S}} \,|\, \mathbf{PA}_B^{\mathcal{G}} \\
\overset{(*)}{\Rightarrow}\ &\exists \tilde{\mathbf{S}} : B \perp\!\!\!\perp (A, \tilde{\mathbf{S}}) \,|\, \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \\
\Rightarrow\ &B \perp\!\!\!\perp A \,|\, \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \\
\Rightarrow\ &P(X_{\mathbf{V}}) = P(B|\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\}) \prod_{X \neq B} P(X|\mathbf{PA}_X^{\mathcal{G}}) \\
\Rightarrow\ &P(X_{\mathbf{V}}) \text{ is Markov wrt to } \mathcal{G} \text{ without } A \to B \\
\Rightarrow\ &\text{Causal minimality is violated.} \\
\Rightarrow\ &\exists A, B : \text{ is Markov wrt to } \mathcal{G} \text{ without } A \to B \\
\Rightarrow\ &\exists A, B : A \perp\!\!\!\perp B \,|\, \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \\
\Rightarrow\ &\text{Lemma 6.7 is violated.}
\end{aligned}$$

$(*)$ is the "intersection" property of conditional independence [e.g. 1.1.5 in Pearl, 2009] and requires positivity of the densities. □

A.5.3. Proof of Theorem 6.6

Proof of Theorem 6.6 The idea of the proof is as follows: we assume there are two SEMs with graphs $\mathcal{G}$ and $\mathcal{G}'$ that lead to the same joint distribution and then deduce a contradiction. We first try to find variables L and Y that have the same set of parents $\mathbf{S} = \{S_1, S_2\}$ in both graphs, but reversed edges between each other in $\mathcal{G}$ and $\mathcal{G}'$ (as in Fig. A.5). This case is treated in part (ii)-2 and contains the main argument of the proof.

Figure A.2.: This situation is dealt with in part (ii)-2 of the proof (with $\mathbf{S} = \{S_1, S_2\}$ and $\mathbf{D} = \emptyset$). It contains the proof's main argument.

If we assumed faithfulness, $\mathcal{G}$ and $\mathcal{G}'$ could be supposed to be Markov equivalent, which itself implies the existence of such an L and Y [Chickering, 1995, Theorem 2]. Since we are not assuming faithfulness, proving existence of a situation similar as in Fig. A.2 requires more work.

We assume that there are two instances of an IFMOC that both induce $\mathcal{L}(\mathbf{X})$, one with graph $\mathcal{G}$, the other with graph $\mathcal{G}'$. We will show that $\mathcal{G} = \mathcal{G}'$. Since DAGs do not contain any cycles, we always find nodes that have no descendants (start a directed path at some node: after at most $\#\mathbf{X} - 1 = p - 1$ steps you reach a node without a child). Eliminating such a node from the graph leads to a DAG, again; we can discard further nodes without children in the new graph. We repeat this process for all nodes that have no children in both $\mathcal{G}$ and $\mathcal{G}'$ and have the same parents in both graphs. If we end up with no nodes left, the two graphs are identical and we are done. Otherwise, we end up with two smaller graphs that we again call $\mathcal{G}$ and $\mathcal{G}'$ and a node L that has no children in $\mathcal{G}$ and either $\mathbf{PA}_L^{\mathcal{G}} \neq \mathbf{PA}_L^{\mathcal{G}'}$ or $\mathbf{CH}_L^{\mathcal{G}'} \neq \emptyset$. We

Figure A.3.: Nodes adjacent to L in $\mathcal{G}$ and $\mathcal{G}'$.

will show that this leads to a contradiction. Importantly, because of the Markov property of $\mathcal{G}$, all other nodes are independent of L given $\mathbf{PA}_L^{\mathcal{G}}$:

$$L \perp\!\!\!\perp \mathbf{X} \setminus (\mathbf{PA}_L^{\mathcal{G}} \cup \{L\}) \mid \mathbf{PA}_L^{\mathcal{G}} \tag{A.11}$$

To make the arguments easier to understand, we introduce the following notation (see also Figure A.3): We partition $\mathcal{G}$-parents of L into $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$. Here, $\mathbf{Z}$ are also $\mathcal{G}'$-parents of L, $\mathbf{Y}$ are $\mathcal{G}'$-children of L and $\mathbf{W}$ are not adjacent to L in $\mathcal{G}'$. We denote with $\mathbf{D}$ the $\mathcal{G}'$-parents of L that are not adjacent to L in $\mathcal{G}$ and by $\mathbf{E}$ the $\mathcal{G}'$-children of L that are not adjacent to L in $\mathcal{G}$. Thus: $\mathbf{PA}_L^{\mathcal{G}} = \mathbf{Y} \cup \mathbf{Z} \cup \mathbf{W}$, $\mathbf{CH}_L^{\mathcal{G}} = \emptyset$, $\mathbf{PA}_L^{\mathcal{G}'} = \mathbf{Z} \cup \mathbf{D}$, $\mathbf{CH}_L^{\mathcal{G}'} = \mathbf{Y} \cup \mathbf{E}$.
Consider $\mathbf{T} := \mathbf{W} \cup \mathbf{Y}$. We distinguish two cases:

Case (i): $\mathbf{T} = \emptyset$.
Then there must be a node $D \in \mathbf{D}$ or a node $E \in \mathbf{E}$, otherwise L would have been discarded.

1. If there is a $D \in \mathbf{D}$ then (A.11) implies $L \perp\!\!\!\perp D \,|\, \mathbf{S}$ for $\mathbf{S} := \mathbf{Z} \cup \mathbf{D} \setminus \{D\}$, which contradicts Lemma 6.7 (applied to $\mathcal{G}'$).

2. If $\mathbf{D} = \emptyset$ and there is $E \in \mathbf{E}$ then $E \perp\!\!\!\perp L \,|\, \mathbf{S}$ holds for $\mathbf{S} := \mathbf{Z} \cup \mathbf{PA}_E^{\mathcal{G}'} \setminus \{L\}$, which also contradicts Lemma 6.7 (note that $\mathbf{Z} \subseteq \mathbf{ND}_E^{\mathcal{G}'}$ to avoid cycles).

Case (ii): $\mathbf{T} \neq \emptyset$.
Then $\mathbf{T}$ contains a "$\mathcal{G}'$-youngest" node with the property that there is no directed $\mathcal{G}'$-path from this node to any other node in $\mathbf{T}$. This node may not be unique.

1. Suppose that some $W \in \mathbf{W}$ is such a youngest node. Consider the DAG $\tilde{\mathcal{G}}'$ that equals $\mathcal{G}'$ with additional edges $Y \to W$ and $W' \to W$ for all $Y \in \mathbf{Y}$ and $W' \in \mathbf{W} \setminus \{W\}$. In $\tilde{\mathcal{G}}'$ L and W are not adjacent. Thus we find a set $\tilde{\mathbf{S}}$ such that $\tilde{\mathbf{S}}$ d-separates L and W in $\tilde{\mathcal{G}}'$; indeed, one can take[1] $\tilde{\mathbf{S}} := \big(\mathbf{CH}_L^{\tilde{\mathcal{G}}'} \cup \mathbf{PA}^{\tilde{\mathcal{G}}'}(\mathbf{CH}_L^{\tilde{\mathcal{G}}'})\big) \setminus \big(\mathbf{U} \cup \mathbf{DE}^{\tilde{\mathcal{G}}'}(\mathbf{U})\big)$ with $\mathbf{U} = \mathbf{CH}_L^{\tilde{\mathcal{G}}'} \cap \mathbf{CH}_W^{\tilde{\mathcal{G}}'}$. Then also $\mathbf{S} = \tilde{\mathbf{S}} \cup \{\mathbf{Y}, \mathbf{Z}, \mathbf{W} \setminus \{W\}\}$ d-separates L and W in $\tilde{\mathcal{G}}'$.

   Indeed: All $Y \in \mathbf{Y}$ are already in $\tilde{\mathbf{S}}$ in order to block $L \to Y \to W$. Suppose there is a $\tilde{\mathcal{G}}'$-path that is blocked by $\tilde{\mathbf{S}}$ and unblocked if we add Z and W' nodes to $\tilde{\mathbf{S}}$. How can we unblock a path by including more nodes? The path ($L \cdots V_1 \cdots U_1 \cdots W$ in Figure A.4) must contain a collider V_1 that is an ancestor of a Z with $V_1, \ldots, V_m, Z \notin \tilde{\mathbf{S}}$ and corresponding nodes U_i for a W' node. Choose V_1 and U_1 on the given path so close to each other such that there is no such a collider in between. If there is no V_1, choose U_1 close to L, if there is no U_1, choose V_1 close to W. Now the path $L \leftarrow Z \cdots V_1 \cdots U_1 \cdots W' \to W$ is unblocked given $\tilde{\mathbf{S}}$, which is a contradiction to $\tilde{\mathbf{S}}$ d-separates L and W.

   But then $\mathbf{S}$ d-separates L and W in $\mathcal{G}'$, too and we have $L \perp\!\!\!\perp W \,|\, \mathbf{S}$ which contradicts Lemma 6.7 (applied to $\mathcal{G}$).

2. Therefore, the $\mathcal{G}'$-youngest node in $\mathbf{T}$ must be some $Y \in \mathbf{Y}$. We define $\mathbf{S} := \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\} \cup \mathbf{PA}_Y^{\mathcal{G}'} \setminus \{L\}$. Clearly, $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$ since L does not have any descendants in $\mathcal{G}$. Further, $\mathbf{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}'}$ because Y is the youngest under all $W \in \mathbf{W}$ and $Y \in \mathbf{Y} \setminus \{Y\}$ by construction and any directed path from Y to $Z \in \mathbf{Z}$ would introduce a cycle in $\mathcal{G}'$. Ergo, $\{Y\} \cup \mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$ and $\{L\} \cup \mathbf{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}'}$. Lemma A.8 gives us $N_L \perp\!\!\!\perp (Y, \mathbf{S})$ and $N_Y \perp\!\!\!\perp (L, \mathbf{S})$ and we can thus apply Lemma A.7. From $\mathcal{G}$ we find

$$L_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}} = f_L(x_{\mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\}}, Y_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}}, N_L),$$
$$N_L \perp\!\!\!\perp Y_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}}$$

[1] By $\mathbf{PA}^{\mathcal{G}}(\mathbf{B})$ for some set $\mathbf{B} \subset \mathbf{X}$ we denote the union of all parents: $\bigcup_{B \in \mathbf{B}} \mathbf{PA}_B^{\mathcal{G}}$.

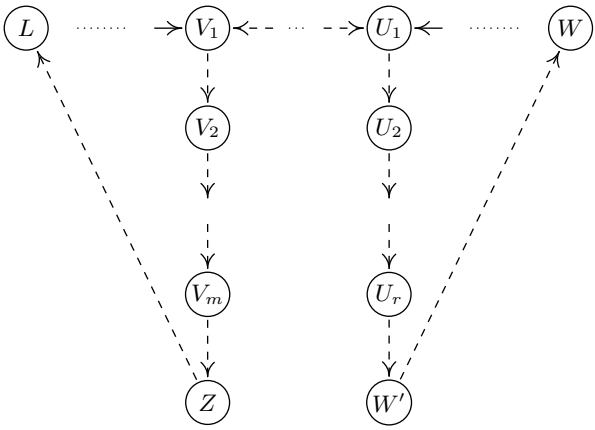


Figure A.4.: Assume the path $L \cdots V_1 \cdots U_1 \cdots W$ is blocked by $\tilde{\mathbf{S}}$, but unblocked if we include Z and W'. Then the dashed path is unblocked given $\tilde{\mathbf{S}}$.

and from $\mathcal{G}'$ we have

$$Y_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}} = g_Y(x_{\mathbf{PA}_Y^{\mathcal{G}'}\setminus\{L\}}, L_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}}, N_Y),$$
$$N_Y \perp\!\!\!\perp L_{\,|\,X_{\mathbf{S}}=x_{\mathbf{S}}}$$

This leads to a contradiction since according to Definition 6.4 we can choose $x_{\mathbf{S}}$ such that

$$(f_L(x_{\mathbf{PA}_L^{\mathcal{G}}\setminus\{Y\}}, \cdot, \cdot), \mathcal{L}(Y \,|\, X_{\mathbf{S}} = x_{\mathbf{S}}), \mathcal{L}(N_L)) \in \mathcal{B}\,,$$

and $g_Y(x_{\mathbf{PA}_Y^{\mathcal{G}'}\setminus\{L\}}, \cdot, \cdot) \in \mathcal{F}$.

□

A.6. Proofs of Chapter 7

A.6.1. Some Lemmata

In the following two sections we consider different subsets of the set of variables $\mathbf{X}$: to simplify notation we do not distinguish between indices

and variables anymore since the context should clarify the meaning. This way, we can also speak of the parents $\mathbf{PA}_B^{\mathcal{G}}$ of a variable $B \in \mathbf{X}$. We also consider sets of variables $\mathbf{S} \subset \mathbf{X}$ as a single multivariate variable.
The following four statements are all plausible and their proof is mostly about technicalities. The reader may skip to the next section and use the lemmata whenever needed.

Lemma A.11 *Let $(A_1, \ldots, A_m) \sim \mathcal{N}((\mu_1, \ldots, \mu_m)^T, \Sigma)$ with strictly positive definite Σ and define $A_1^* = A_{1\,|\,(A_2,\ldots,A_m)=(a_2,\ldots,a_m)}$. Then, for all $(a_2, \ldots, a_m) \in \mathbb{R}^{m-1}$ it holds*

$$\mathbf{var} A_1^* \leq \mathbf{var} A_1 \,.$$

Proof. Let us decompose Σ into

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12}^T \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

with Σ_{12} being an $(m-1) \times 1$ vector. Then

$$\mathbf{var} A_1^* = \sigma_1^2 - \Sigma_{12}^T \cdot \Sigma_{22}^{-1} \cdot \Sigma_{12} \leq \sigma_1^2$$

since Σ_{22}^{-1} is positive definite. □

Lemma A.12 *[same as Lemma A.7] Let $Y \in \mathcal{Y}, N \in \mathcal{N}, \mathbf{Q} \in \mathcal{Q}, \mathbf{R} \in \mathcal{R}$ be random variables whose joint distribution is absolutely continuous with respect to some product measure ($\mathbf{Q}$ and $\mathbf{R}$ can be multivariate) and with density $p_{Y,\mathbf{Q},\mathbf{R},N}(y, \mathbf{q}, \mathbf{r}, n)$. Let $f : \mathcal{Y} \times \mathcal{Q} \times \mathcal{N} \to \mathbb{R}$ be a measurable function. If $N \perp\!\!\!\perp (Y, \mathbf{Q}, \mathbf{R})$ then for all $\mathbf{q} \in \mathcal{Q}, \mathbf{r} \in \mathcal{R}$ with $p_{\mathbf{Q},\mathbf{R}}(\mathbf{q}, \mathbf{r}) > 0$:*

$$f(Y, \mathbf{Q}, N)\,|\,_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}} \stackrel{\mathcal{L}}{=} f(Y\,|\,_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}}, \mathbf{q}, N) \,.$$

Here, $\stackrel{\mathcal{L}}{=}$ means that both sides have the same distribution.

Lemma A.13 *[same as Lemma A.8] Let $\mathcal{L}(\mathbf{X})$ be generated according to an SEM as in (7.1) with corresponding DAG $\mathcal{G}$ and consider a variable $X \in \mathbf{X}$. If $\mathbf{S} \subseteq \mathbf{ND}_X^{\mathcal{G}}$ then $N_X \perp\!\!\!\perp \mathbf{S}$.*

Lemma A.14 *Let $\mathcal{L}(\mathbf{X})$ be generated from an SEM as in (7.1) with DAG $\mathcal{G}$. Consider a variable $B \in \mathbf{X}$ and one of its parents $A \in \mathbf{PA}_B^{\mathcal{G}}$. For all sets $\mathbf{S}$ with $\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}} \setminus \{A\}$ we have*

$$B \not\perp\!\!\!\perp A \mid \mathbf{S}\,.$$

Proof. Define $\mathbf{Q} = \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\}$ such that we have $\mathbf{S} = (\mathbf{Q}, \mathbf{R})$ for some $\mathbf{R}$. Using Lemma A.12 we have:

$$B|_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}} = f(\mathbf{q}) + \beta \cdot A|_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}} + N_B$$

with $N_B \perp\!\!\!\perp A|_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}}$. But since $\beta \neq 0$, it follows:

$$A|_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}} \not\perp\!\!\!\perp B|_{\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}}\,.$$

□

A.6.2. Proof of Theorem 7.1.

Proof of Theorem 7.1 The idea of the proof is as follows: we assume there are two SEMs with graphs $\mathcal{G}$ and $\mathcal{G}'$ that lead to the same joint distribution and then deduce a contradiction. We first try to find variables L and Y that have the same set of parents $\mathbf{S} = \{S_1, S_2\}$ in both graphs, but reversed edges between each other in $\mathcal{G}$ and $\mathcal{G}'$ (as in Fig. A.5). This case is treated in part (ii)-2 and contains the main argument of the proof.

Figure A.5.: This situation is dealt with in part (ii)-2 of the proof (with $\mathbf{S} = \{S_1, S_2\}$ and $\mathbf{D} = \emptyset$). It contains the proof's main argument.

If we assumed faithfulness, $\mathcal{G}$ and $\mathcal{G}'$ could be supposed to be Markov equivalent, which itself implies the existence of such an L and Y

[Chickering, 1995, Theorem 2]. Since we are not assuming faithfulness, proving existence of a situation similar as in Fig. A.5 requires more work. Note that this part of the proof (that is due to not assuming faithfulness) is taken from Peters et al. [2011b] and remains almost the same. It is given here for completeness. The difference to Peters et al. [2011b] is that we can prove causal minimality and do not have to assume it. New are also Lemmata A.11 and A.14, as well as the proof's main argument (ii)-2. We now give a formal proof.
We assume that there are two SEMs as in Equation (7.1) that both induce $\mathcal{L}(\mathbf{X})$, one with graph $\mathcal{G}$, the other with graph $\mathcal{G}'$. We will show that $\mathcal{G} = \mathcal{G}'$. Since DAGs do not contain any cycles, we always find nodes that have no descendants (start a directed path at some node: after at most $\#\mathbf{X} - 1$ steps we reach a node without a child). Eliminating such a node from the graph leads to a DAG, again; we can discard further nodes without children in the new graph. We repeat this process for all nodes that have no children in both $\mathcal{G}$ and $\mathcal{G}'$ and have the same parents in both graphs. If we end up with no nodes left, the two graphs are identical and we are done. Otherwise, the procedure results in two smaller graphs that we again call $\mathcal{G}$ and $\mathcal{G}'$ and a node L that has no children in $\mathcal{G}$ and either $\mathbf{PA}_L^{\mathcal{G}} \neq \mathbf{PA}_L^{\mathcal{G}'}$ or $\mathbf{CH}_L^{\mathcal{G}'} \neq \emptyset$. We will show that this leads to a contradiction. Importantly, because of the Markov property of $\mathcal{G}$, all other nodes are independent of L given $\mathbf{PA}_L^{\mathcal{G}}$:

$$L \perp\!\!\!\perp \mathbf{X} \setminus (\mathbf{PA}_L^{\mathcal{G}} \cup \{L\}) \mid \mathbf{PA}_L^{\mathcal{G}} . \tag{A.12}$$

To make the arguments easier to understand, we introduce the following notation (see also Fig. A.6): we partition $\mathcal{G}$-parents of L into $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$. Here, $\mathbf{Z}$ are also $\mathcal{G}'$-parents of L, $\mathbf{Y}$ are $\mathcal{G}'$-children of L and $\mathbf{W}$ are not adjacent to L in $\mathcal{G}'$. We denote with $\mathbf{D}$ the $\mathcal{G}'$-parents of L that are not adjacent to L in $\mathcal{G}$ and by $\mathbf{E}$ the $\mathcal{G}'$-children of L that are not adjacent to L in $\mathcal{G}$. Thus: $\mathbf{PA}_L^{\mathcal{G}} = \mathbf{Y} \cup \mathbf{Z} \cup \mathbf{W}$, $\mathbf{CH}_L^{\mathcal{G}} = \emptyset$, $\mathbf{PA}_L^{\mathcal{G}'} = \mathbf{Z} \cup \mathbf{D}$, $\mathbf{CH}_L^{\mathcal{G}'} = \mathbf{Y} \cup \mathbf{E}$. Consider $\mathbf{T} := \mathbf{W} \cup \mathbf{Y}$. We distinguish two cases:

Case (i): $\mathbf{T} = \emptyset$.
Then there must be a node $D \in \mathbf{D}$ or a node $E \in \mathbf{E}$, otherwise L would have been discarded.

1. If there is a $D \in \mathbf{D}$ then (A.12) implies $L \perp\!\!\!\perp D \,|\, \mathbf{S}$ for $\mathbf{S} :=$

Figure A.6.: Nodes adjacent to L in $\mathcal{G}$ and $\mathcal{G}'$

$\mathbf{Z} \cup \mathbf{D} \setminus \{D\}$, which contradicts Lemma A.14 (applied to $\mathcal{G}'$).

2. If $\mathbf{D} = \emptyset$ and there is $E \in \mathbf{E}$ then $E \perp\!\!\!\perp L \,|\, \mathbf{S}$ holds for $\mathbf{S} := \mathbf{Z} \cup \mathbf{PA}_E^{\mathcal{G}'} \setminus \{L\}$, which also contradicts Lemma A.14 (note that $\mathbf{Z} \subseteq \mathbf{ND}_E^{\mathcal{G}'}$ to avoid cycles).

Case (ii): $\mathbf{T} \neq \emptyset$.
Then $\mathbf{T}$ contains a "$\mathcal{G}'$-youngest" node with the property that there is no directed $\mathcal{G}'$-path from this node to any other node in $\mathbf{T}$. This node may not be unique.

1. Suppose that some $W \in \mathbf{W}$ is such a youngest node. Consider the DAG $\tilde{\mathcal{G}}'$ that equals $\mathcal{G}'$ with additional edges $Y \to W$ and $W' \to W$ for all $Y \in \mathbf{Y}$ and $W' \in \mathbf{W} \setminus \{W\}$. In $\tilde{\mathcal{G}}'$ L and W are not adjacent. Thus we find a set $\tilde{\mathbf{S}}$ such that $\tilde{\mathbf{S}}$ d-separates L and W in $\tilde{\mathcal{G}}'$; indeed, one can take[2] $\tilde{\mathbf{S}} := \big(\mathbf{CH}_L^{\tilde{\mathcal{G}}'} \cup \mathbf{PA}^{\tilde{\mathcal{G}}'}(\mathbf{CH}_L^{\tilde{\mathcal{G}}'})\big) \setminus \big(\mathbf{U} \cup \mathbf{DE}^{\tilde{\mathcal{G}}'}(\mathbf{U})\big)$ with $\mathbf{U} = \mathbf{CH}_L^{\tilde{\mathcal{G}}'} \cap \mathbf{CH}_W^{\tilde{\mathcal{G}}'}$. Then also $\mathbf{S} = \tilde{\mathbf{S}} \cup \{\mathbf{Y}, \mathbf{Z}, \mathbf{W} \setminus \{W\}\}$ d-separates L and W in $\tilde{\mathcal{G}}'$.

   Indeed, all $Y \in \mathbf{Y}$ are already in $\tilde{\mathbf{S}}$ in order to block $L \to Y \to W$. Suppose there is a $\tilde{\mathcal{G}}'$-path that is blocked by $\tilde{\mathbf{S}}$ and unblocked if we add Z and W' nodes to $\tilde{\mathbf{S}}$. How can we unblock a path by including more nodes? The path ($L \cdots V_1 \cdots U_1 \cdots W$ in Fig. A.7) must contain a collider V_1 that is an ancestor of a Z with $V_1, \ldots, V_m, Z \notin \tilde{\mathbf{S}}$ and corresponding nodes U_i for a W' node. Choose V_1 and U_1 on the given path so close to each other such that there is no such collider in between. If there is no V_1, choose U_1 close to L, if there is no U_1, choose V_1

[2] By $\mathbf{PA}^{\mathcal{G}}(\mathbf{B})$ for some set $\mathbf{B} \subset \mathbf{X}$ we denote the union of all parents: $\bigcup_{B \in \mathbf{B}} \mathbf{PA}_B^{\mathcal{G}}$.

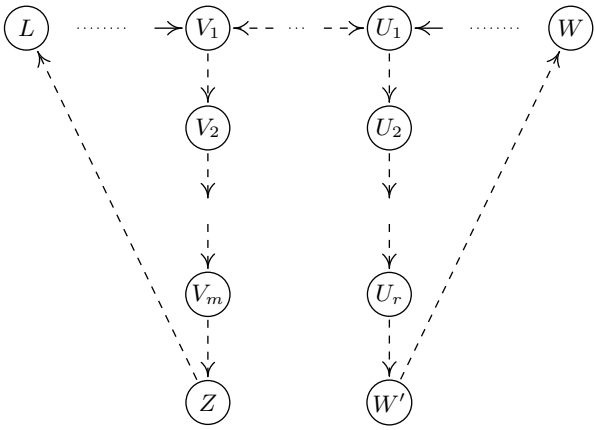


Figure A.7.: Assume the path $L \cdots V_1 \cdots U_1 \cdots W$ is blocked by $\tilde{\mathbf{S}}$, but unblocked if we include Z and W'. Then the dashed path is unblocked given $\tilde{\mathbf{S}}$.

close to W. Now the path $L \leftarrow Z \cdots V_1 \cdots U_1 \cdots W' \rightarrow W$ is unblocked given $\tilde{\mathbf{S}}$, which is a contradiction to the assumption $\tilde{\mathbf{S}}$ d-separates L and W.

But then $\mathbf{S}$ d-separates L and W in $\mathcal{G}'$, too (there are less paths), and we have $L \perp\!\!\!\perp W \mid \mathbf{S}$ which contradicts Lemma A.14 (applied to $\mathcal{G}$).

2. Therefore, the $\mathcal{G}'$-youngest node in $\mathbf{T}$ must be some $Y \in \mathbf{Y}$.

First, note that

$$\sigma^2_{\mathcal{G}} = \sigma^2_{\mathcal{G}'} = \min_{X \in \mathbf{X}} \mathbf{var} X = \sigma^2 \tag{A.13}$$

We define $\mathbf{S} := \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\} \cup \mathbf{D}$. Clearly, $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$ since L does not have any descendants in $\mathcal{G}$. Define $\mathbf{Q} := \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\}$ and take any $\mathbf{s} = (\mathbf{q}, \mathbf{d})$. Define

$$L^* := L \,|_{\mathbf{S}=\mathbf{s}} \qquad \text{and} \qquad Y^* := Y \,|_{\mathbf{S}=\mathbf{s}}$$

Then, from $\mathcal{G}$ and Lemma A.12 we find

$$\begin{aligned} L^* &= f_L(\mathbf{q}, Y^*) + N_L, & N_L \perp\!\!\!\perp Y \,|_{\mathbf{S}=\mathbf{s}} \\ &= f(\mathbf{q}) + \beta \cdot Y^* + N_L, & N_L \perp\!\!\!\perp Y \,|_{\mathbf{S}=\mathbf{s}} \end{aligned}$$

Note that the independence holds because of $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$. Then, we have

$$\mathbf{var} L^* = \beta^2 \mathbf{var} Y^* + \sigma^2 > \sigma^2 \,. \tag{A.14}$$

Since $\mathbf{PA}_L^{\mathcal{G}'} \subseteq \mathbf{S}$ we find from $\mathcal{G}'$ and Lemma A.11 that

$$\mathbf{var} L^* \leq \sigma^2 \,. \tag{A.15}$$

(Note that $\det(\mathrm{cov}(\mathbf{X})) \neq 0$.) Equations (A.14) and (A.15) contradict each other.

In order to prove Remark 7.3, replace $\mathbf{var} X$ by $\mathbf{var} X / \sigma_X^2$ in (A.13) and σ^2 by $\sigma^2 \cdot \sigma_X^2$ in Equations (A.14) and (A.15). □

A.7. Proofs of Chapter 8

A.7.1. A Lemma

Lemma A.15 *If $\mathbf{X}_t = (X_t^i)_{1 \leq i \leq p}$ satisfy a TiMINo model, each variable X_t^i is conditionally independent of each of its non-descendants given its parents.*

Proof. With $\mathcal{S} := \mathbf{PA}(X_t^i) = \bigcup_{k=0}^{\pi} (\mathbf{PA}_k^i)_{t-k}$ and Equation (8.4) we get $X_t^i|_{\mathcal{S}=s} = f_i(s, N_t^i)$ for an s with $p(s) > 0$. Any non-descendant of X_t^i can be written as a function of all noise variables from its ancestors and $\mathbf{X}_0, \ldots, \mathbf{X}_{\pi-1}$. It is therefore independent of X_t^i given $\mathcal{S} = s$. For this proof we consider time series for $t \in \mathbb{N}_0$. A corresponding statement holds for $t \in \mathbb{Z}$, when we assume causality (innovations are independent of the past). □

A.7.2. Proof of Theorem 8.2

Proof. Suppose that $\mathbf{X}_t$ allows two different representations of TiMINo that lead to two different full time graphs $\mathcal{G}$ and $\mathcal{G}'$.

(i) First we assume that $\mathcal{G}$ and $\mathcal{G}'$ do not differ in the instantaneous effects: $\mathbf{PA}_0^i(\text{in } \mathcal{G}) = \mathbf{PA}_0^i(\text{in } \mathcal{G}')\ \forall i$. Without loss of generality, there is some $k > 0$ and an edge $X_{t-k}^1 \to X_t^2$, say, that is in $\mathcal{G}$ but not in $\mathcal{G}'$. From $\mathcal{G}'$ and Lemma A.15 we have that $X_{t-k}^1 \perp\!\!\!\perp X_t^2 \,|\, \mathcal{S}$, where

$$\mathcal{S} = (\{X_{t-l}^i, 1 \le l \le \pi, 1 \le i \le p\} \cup \mathbf{ND}_t) \setminus \{X_{t-k}^1, X_t^2\},$$

and $\mathbf{ND}_t$ are all X_t^i that are non-descendants (wrt instantaneous effects) of X_t^2. Applied to $\mathcal{G}$, causal minimality leads to a contradiction:

$$X_{t-k}^1 \not\perp\!\!\!\perp X_t^2 \,|\, \mathcal{S}.$$

Now we suppose $\mathcal{G}$ and $\mathcal{G}'$ differ in the instantaneous effects. This time we choose $\mathcal{S} = \{X_{t-l}^i, 1 \le l \le \pi, 1 \le i \le p\}$. Then for each s and i we have:

$$X_t^i|_{\mathcal{S}=s} = f_i(s, (\tilde{\mathbf{PA}}_0^{\,i})_t),$$

where $\tilde{\mathbf{PA}}_0^{\,i}$ are all instantaneous parents of X_t^i conditioned on $\mathcal{S} = s$. All $X_t^i|_{\mathcal{S}=s}$ with the instantaneous effects describe two different structures of an IFMOC. This contradicts the identifiability results by Peters et al. [2011b].

(ii) Because of Lemma A.15 and faithfulness $\mathcal{G}$ and $\mathcal{G}'$ only differs in the instantaneous effects. But each instantaneous arrow $X_t^i \to X_t^j$ forms a v-structure together with $X_{t-k}^j \to X_t^j$; the latter exists because of the time structure and X_{t-k}^j cannot be connected with X_t^i since this introduces a cycle in the summary time graph.

□

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.

N. Ancona, D. Marinazzo, and S. Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev. E*, 70(5):056221, 2004.

S.A. Andersson, D. Madigan, and M.D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

R. Armann and I. Bülthoff. in preparation. Website, 2010. `https://webdav.tuebingen.mpg.de/cause-effect/`.

A. Asuncion and D.J. Newman. UCI machine learning repository. `http://archive.ics.uci.edu/ml/`, 2007.

A. Azzalini and A.W. Bowman. A look at some data on the Old Faithful Geyser. *Applied Statistics*, 39(3):357–365, 1990.

D. Bell, J. Kay, and J. Malley. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51(1):7–18, 1996.

W.P. Bergsma. *Testing conditional independence for continuous random variables*, 2004. EURANDOM-report 2004-049.

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

L. Bottou, J. Peters, J. Quiñonero-Candela, D.X. Charles, D.M. Chickering, E. Portugualy, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems. *ArXiv e-prints (1209.2355)*, 2012.

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: forecasting and control*. Wiley series in probability and statistics. John Wiley, 2008.

U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 137–144, 2006.

P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, second edition, 1991.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, 2006.

Y. Chen, G. Rangarajan, J. Feng, and M. Ding. Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, 324, 2004.

D.M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1995.

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

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

W.G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10:417–451, 1954.

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

J. Czerniak and H. Zarzycki. *Application of rough sets in the presumptive diagnosis of urinary system diseases*, pages 41–51. Kluwer Academic Publishers, 2003.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst.Internationale Statist.*, 21:2–8, 1953.

Deutscher Wetter Dienst. Climate data. `http://www.dwd.de/`, 2008.

M.J. Druzdzel and H. van Leijen. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):45–62, 2001.

M. Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, pages 1–36, 2011.

J.P. Florens and M. Mouchart. A note on noncausality. *Econometrica*, 50(3):583–591, 1982.

N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 489–496. MIT Press, 2008.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1994.

B.W. Gould. Diary data sets. `http://future.aae.wisc.edu/tab/prices.html`, 2007.

C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, July 1969.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

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

Y. Guo, X. Niu, and H. Zhang. An extensive empirical study on semi-supervised learning. In *ICDM*, 2010.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

D.M.A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

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

D. Heckerman. A Bayesian approach to causal discovery. Technical report, Microsoft Research (MSR-TR-97-05), 1997.

D. Heckerman and D. Geiger. Likelihoods and parameter priors for bayesian networks. Technical report, Microsoft Research (MSR-TR-95-54), 1995.

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

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

A. Hyvärinen, S. Shimizu, and P.O. Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 424–431, 2008.

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

D. Janzing and B. Steudel. Justifying additive-noise-model based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17:189–212, 2010.

D. Janzing, J. Peters, J.M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

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

D. Janzing, J.M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

A. Kankainen. Consistent testing of total independence based on the empirical characteristic function. *PhD Thesis, University of Jyväskylä*, 1995.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

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

S. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

J. Lemeire and E. Dirkx. Causal models as minimal descriptions of multivariate systems. `http://parallel.vub.ac.be/~jan/`, 2006.

L. Li, W. Chu, J. Langford, and R.E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on the World Wide Web (WWW 2010)*, pages 661–670. ACM, 2010.

M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

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

R. Matthews. Storks deliver babies ($p = 0.008$). *Teaching Statistics*, 22(2):36–38, 2000.

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

J.M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 1687–1695. MIT Press, 2010.

W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. The population biology of abalone (haliotis species) in Tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288), 1994.

NCDC. ASOS 1-minute data (DSI 6406/page 2 data), 2009. URL `http://www.ncdc.noaa.gov/oa/climate/climatedata.html`.

G. Nolte, A. Ziehe, V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Letters*, 100, 2008.

M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden, and C.E. Hanson, editors. *IPCC Fourth Assessment Report: Working*

Group II Report "Impacts, Adaptation and Vulnerability". Cambridge University Press, 2007. URL `http://www.ipcc.ch/ipccreports/ar4-wg2.htm`.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009.

J. Peters. Asymmetries of time series under inverting their direction. Diploma Thesis (University of Heidelberg), 2008. `http://stat.ethz.ch/people/jopeters`.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with same error variances. *ArXiv e-prints (1205.2536)*, 2012.

J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 801–808, 2009.

J. Peters, D. Janzing, and B. Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *AIStats 13*, volume 9, pages 597–604. JMLR: W&CP, 2010.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2436–2450, 2011a.

J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011b.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using structural equation models. *ArXiv e-prints (1207.5136)*, 2012.

C.E. Rasmussen and H. Nickisch. GPML code. Website, 2007. `http://www.gaussianprocess.org/gpml/code`.

C.E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.

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

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

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

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Massachusetts, 2002.

B. Schölkopf, A. J. Smola, and R. Williamson. Shrinking the tube: A new support vector regression algorithm. In *Advances in Neural Information Processing Systems 11 (NIPS)*, pages 330–336. MIT Press, 1999.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

R. Shachter and C. Kenley. Gaussian influence diagrams. *Management Science*, 35:527–550, 1989.

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

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

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

V.P. Skitovic. Linear forms in independent random variables and the normal distribution law (in Russian). *Izvestiia AN SSSR, Ser. Matem.*, 18:185–200, 1954.

V.P. Skitovic. Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab.*, 2:211–228, 1962.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

D.J. Stekhoven, I. Moraes, G. Sveinbjörnsson, L. Hennig, M.H. Maathuis, and P. Bühlmann. Causal stability ranking. *submitted*, 2012.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71:1248–1256, 2008.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

R.P. Tanase. Learning causal structures from Gaussian structural equation models. Master Thesis (ETH Zurich), 2012. available online.

J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290: 2319–2323, 2000.

R. Tillman, A. Gretton, and P. Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 1847–1855. MIT Press, 2010.

H. Tong. *Non-Linear Time Series: A Dynamical System Approach*. Oxford Statistical Science Series. Clarendon Press, 1990.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *ArXiv e-prints (1207.0547)*, 2012.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *submitted*, 2012.

L.J.P. van der Maaten. An introduction to dimensionality reduction using matlab. Technical Report MICC-IKAT 07-07, Maastricht University, Maastricht, 2007.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1991.

G. Wahba. *Spline Models for Observational Data*. Series in Applied Math., Vol. 59, SIAM, Philadelphia, 1990.

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

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. In *LSE-Pitt Conference: Confirmation, Induction and Science*, 2007.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

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

J. Zscheischler, D. Janzing, K. Zhang, and B. Schölkopf. Testing whether linear equations are causal: A free probability theory approach. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

Curriculum Vitae

Jonas Peters was born in Nordhorn, Germany, on 28.5.1984 and finished the Abitur at the Burg-Gymnasium Bad Bentheim in 2002. He performed his civilian service in the administration of a children's home between 2002 and 2003.

In 2003 he started studying mathematics with physics as a minor at the University of Heidelberg and was working as a teaching assistant from 2005 until 2006. Between 2006 and 2007 he obtained the Master of Advanced Study in Mathematics (Part III) from University of Cambridge (UK). Afterwards, he wrote his diploma thesis about detecting the direction of time at the Max-Planck-Institute of Intelligent Systems in Tübingen (former MPI for Biological Cybernetics). He obtained the Diplom in Mathematics from the University of Heidelberg in 2009. During his studies he has been supported by the Studienstiftung des deutschen Volkes, by the DAAD and by the Kurt-Hahn-Trust.

He wrote this thesis under the supervision of PD Dr. Dominik Janzing and Prof. Dr. Bernhard Schölkopf between 2009 and 2011 at the MPI for Intelligent Systems in Tübingen and under the supervision of Prof. Dr. Peter Bühlmann at the Seminar for Statistics at the ETH Zürich in 2012.

In 2011 he spent three months with Dr. Leon Bottou at Microsoft Research in Bellevue (USA). In 2001 he participated in the Deutsche SchülerAkademie, where he later taught courses between 2009 and 2012. He has been reviewing for IEEE Transactions of Pattern Analysis and Machine Intelligence and for the conferences ICONIP 2011, NIPS 2011, ICML 2012 and UAI 2012.