Causality

Lecture Notes
Version: May 10, 2018

Spring Semester 2018, ETH Zurich

Nicolai Meinshausen,
based on a previous script by Jonas Peters
which developed into an open-access book:
https://mitpress.mit.edu/books/elements-causal-inference

Contents

1 Introduction **5**
- 1.1 Motivation . . . 5
- 1.2 Some bits of probability and statistics . . . 10
- 1.3 Graphs . . . 12
- 1.4 Exercises . . . 14

2 Structural equation models **17**
- 2.1 Definitions and first properties . . . 17
- 2.2 Interventions . . . 19
- 2.3 Counterfactuals . . . 23
- 2.4 Markov property, faithfulness and causal minimality . . . 26
  - 2.4.1 Markov property . . . 26
  - 2.4.2 Faithfulness and causal minimality . . . 29
- 2.5 Some more properties of SEMs . . . 31
- 2.6 Exercises . . . 32

3 Using the known underlying causal structure **33**
- 3.1 Adjustment formulas . . . 33
  - 3.1.1 Truncated factorization, G-computation formula or manipulation theorem . . . 33
  - 3.1.2 Invariances and adjusting . . . 34
- 3.2 Alternative identification of interventional distributions . . . 39
- 3.3 Instrumental variables . . . 40
- 3.4 Exercises . . . 41

4 Causal structure learning **43**
- 4.1 Structure identifiability . . . 43
  - 4.1.1 Faithfulness . . . 43
  - 4.1.2 Additive noise models . . . 44
  - 4.1.3 Linear non-Gaussian acyclic models . . . 46
  - 4.1.4 Nonlinear Gaussian additive noise models . . . 47
  - 4.1.5 Modularity and Independence of cause and mechanism (bivariate case) 49
- 4.2 Independence-based methods . . . 49

4.3 Score-based methods 51
4.4 Data from different environments (not only observational data) 53
4.5 Exercises 54

A Proofs **55**
A.1 Proofs from Chapter 1 55
A.2 Proofs from Chapter 2 55
A.2.1 Proof of Proposition 2.2.4 55
A.2.2 Proof of Proposition 2.2.9 56
A.2.3 Proof of Proposition 2.5.2 56
A.2.4 Proof of Theorem 2.4.2 56
A.2.5 Proof of Proposition 2.4.13 56
A.3 Proofs from Chapter 3 57
A.4 Proofs from Chapter 4 57
A.4.1 Proof of Proposition 4.1.3 57
A.4.2 Proof of Proposition 4.1.6 57

Chapter 1

Introduction

1.1 Motivation

In statistics, we often deal with properties of a joint distribution $\mathbb{P}^{\mathbf{X}}$ of some p-dimensional random vector $\mathbf{X}$. In many situations, however, we are interested in another distribution $\tilde{\mathbb{P}}^{\mathbf{X}}$ that differs from the observed distribution, $\tilde{\mathbb{P}}^{\mathbf{X}} \neq \mathbb{P}^{\mathbf{X}}$. We are trying to support this claim by the following three illustrative examples.

Example 1.1.1 [Chocolate - Nobel Prizes] Messerli [2012] reports that there is a significant correlation between a country's chocolate consumption (per capita) and the number of Nobel prizes awarded to its citizens (also per capita), see Figure 1.1. These correlations are properties of some observational distribution $\mathbb{P}^{\mathbf{X}}$. We must be careful with drawing conclusions like "Eating chocolate produces Nobel prize." or "Geniuses are more likely to eat lots of chocolate", see Figure 1.2 because these statements are "causal". We will see later (Definition 2.2.1) that they concern different distributions $\tilde{\mathbb{P}}^{\mathbf{X}}$: The first statement suggests, for example, that in a distribution, where each country dictates its citizen to eat a randomly chosen amount of chocolate (same for all citizens), there is still a dependence between chocolate consumption and Nobel prizes: more chocolate means more Nobel prizes. Taking our background knowledge into account, however, we do not expect this to happen. We might rather think that the correlation stems from some hidden variables like economic strength of a country, for example.

In this sense, the famous sentence "Correlation does not imply causation" can also be understood as: properties in $\mathbb{P}^{\mathbf{X}}$ do not necessarily tell you anything about properties in $\tilde{\mathbb{P}}^{\mathbf{X}}$. We will see in Section 2.2 how causal language helps us to formulate relations between those distributions.

This data set comes with many difficulties: the variables are averaged quantities, for example, and the observations for different countries are not independent (e.g. there are not arbitrary many Nobel prizes). We nevertheless hope that the reader can still filter out the relevant causal deliberations.

Example 1.1.2 [Myopia] Only very few people infer a direct causal relationship between

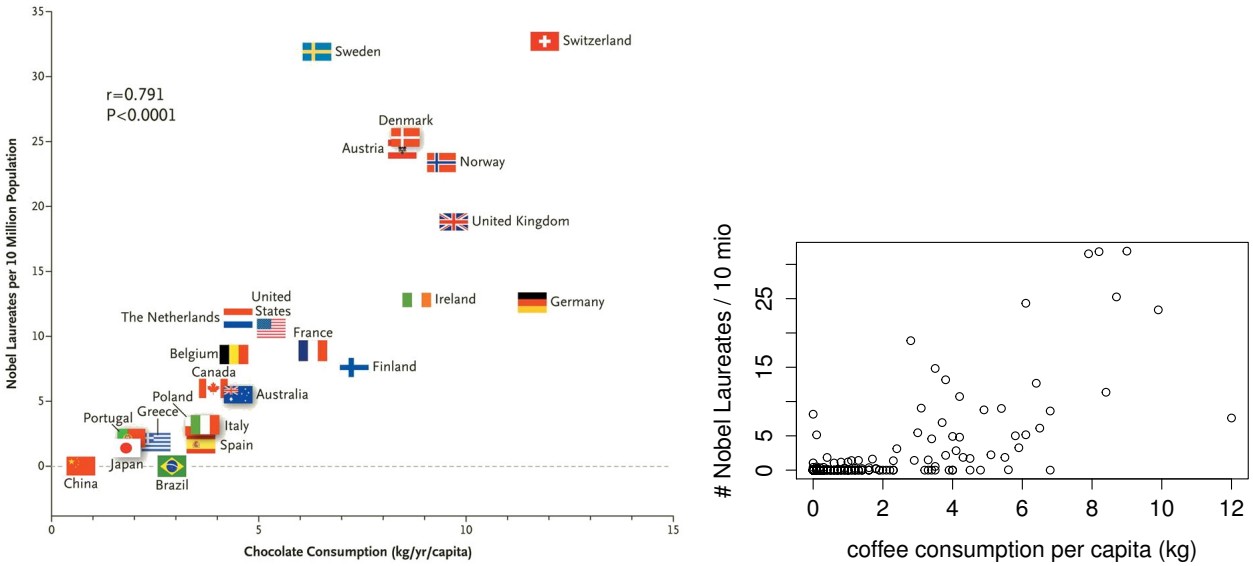


Figure 1.1: The left figure is slightly modified from [Messerli, 2012], it shows a significant correlation between a country's consumption of chocolate and the number of Nobel prizes (averaged per person). The right figure shows a similar result for coffee consumption; the data are based on [Wikipedia, 2013b,a].

Figure 1.2: Two online articles (downloaded from confectionarynews.com and forbes.com on Jan 29th 2013) drawing causal conclusions from the observed correlation between chocolate consumption and Nobel prizes, see Figure 1.1.

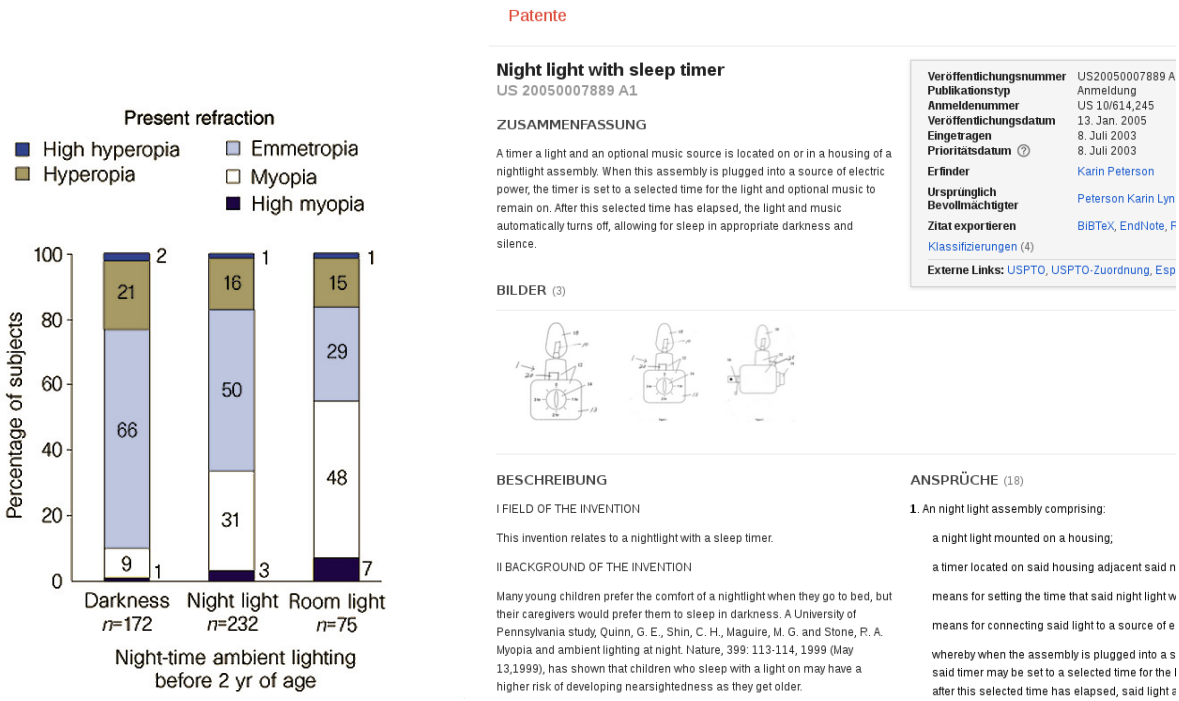


Figure 1.3: The plot on the left shows a (significant) dependence between lighting conditions in a child's bedroom and the development of myopia (shortsightedness). The right figure shows a patent for a night light with timer indicating that enforcing dark rooms decreases the risk of myopia.

Nobel prize winners and chocolate consumption when looking at Figure 1.1. Most people realize that the dependence must be due to "some latent factors". There is an increased risk of false inference when less background knowledge is available. Figure 1.3 (left) shows an example, where people have falsely drawn causal conclusions from observational data. The data set shows a dependence between the usage of a night light in a child's room and the occurrence of myopia [Quinn et al., 1999]. While the authors are cautious enough to say that the study "does not establish a causal link", they add that "the strength of the association [...] does suggest that the absence of a daily period of darkness during childhood is a potential precipitating factor in the development of myopia. Later Gwiazda et al. [2000], Zadnik et al. [2000] found that the correlation is due to whether the child's parents have myopia. If they have, they are more likely to put a night light in their child's room and at the same time, the child has an increased risk of inheriting the disease from its parents. In the meantime, there was a patent filed, see Figure 1.3 (right).

Example 1.1.3 [Kidney Stones] Table 1.1 shows a famous data set from kidney stone recovery [Charig et al., 1986]. Out of 700 patients, one half has been treated with open surgery (78% recovery rate) the other with percutaneous nephrolithotomy (treatment B, with 83% success), a surgical procedure to remove kidney stones by a small

Table 1.1: A classic example of Simpson's paradox. The table reports the success rates of two treatments for kidney stones [Charig et al., 1986, tables I and II] and [Bottou et al., 2013]. Although the overall success rate of treatment B seems better, treatment B performs worse than treatment A on both patients with small kidney stones and patients with large kidney stones, see Examples 3.1.1 and 3.1.7.

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

puncture wound. If we do not know anything else than the overall recovery rates, many people would prefer treatment B if they had to decide. Observing the data in more detail, however, we realize that the open surgery performs better on both small and large kidney stones. How do we deal with this inversion of conclusion? The answer is to concentrate on the precise question we are interested in. This is not whether treatment A or treatment B was more successful in this particular study but how the treatments compare when we force all patients to take treatment A or B, respectively; alternatively, we can compare them only on large stones or small stones, of course. Again, these questions concern some distribution $\tilde{\mathbb{P}}^{\mathbf{X}}$ different from the observational distribution $\mathbb{P}^{\mathbf{X}}$. We will see in Example 3.1.1 why we should prefer treatment A over treatment B. This data set is a famous example for Simpson's paradox [Simpson, 1951], see Example 3.1.7. In fact, it is much less a paradox than the result of the influence of a confounder (i.e. hidden common cause).

If you perform a significance test on the data (e.g. using a proportion test or χ^2 independence test) it turns out that the difference in methods is not significant on a 5% significance level. Note, however, this is not the point of this example. By multiplying each entry in Table 1.1 by a factor of ten, the results would become statistically significant.

Example 1.1.4 [Genetic Data] Causal questions also appear in biological data sets, where we try to predict the effect of interventions (e.g. gene knock-outs). Kemmeren et al. [2014] measures genome-wide mRNA expression levels in yeast, we therefore have data for $p = 6170$ genes. There are $n_{obs} = 160$ "observational" samples of wild-types and $n_{int} = 1479$ data points for the "interventional" setting where each of them corresponds to a strain for which a single gene $k \in K := \{k_1, \dots, k_{1479}\} \subset \{1, \dots, 6170\}$ has been deleted. The data may therefore be interpreted as coming from an observational distribution $\mathbb{P}^{\mathbf{X}}$ and then from 1479 other distributions $\mathbb{P}^{\mathbf{X}}_1, \dots, \mathbb{P}^{\mathbf{X}}_{1479}$. And we are interested in yet other distributions $\tilde{\mathbb{P}}^{\mathbf{X}}$ that tell us how the system reacts after deleting other genes or any combination of genes. Figure 1.4 shows a small subset of the data.

Figure 1.4: The plot on the left shows the observational data (log expression level) for two of the 6170 genes. The middle plot shows 1478 out of the 1479 interventional data points for the same two genes; only the data point that corresponds to a deletion of gene 5954 is omitted. It is shown as the red point in the right plot. Because gene 4710 shows reduced activity after we have intervened on gene 5954, we can infer that 5954 has a (possibly indirect) causal influence on gene 4710. This way, we can use (part of the data) as ground truth for evaluating causal inference methods, that try to infer causal statements either from observational data or from a combination of observational and interventional data. The black lines indicate that the expression levels of both genes are correlated.

Example 1.1.4 is taken from [Peters et al., 2015].

Example 1.1.5 [Advertising placement]

The system Figure 1.5 shows a (heavily) simplified version of an advertisement system that is implemented on a search website. In a nutshell, advertisers can bid on a combination of advertisements and search queries hoping that their ad will be placed in a good location: either on the top of the "sidebar" or even above the search results, i.e. in the "mainline". Only if the user clicks on one of the ads, the advertiser pays money to the publisher according to some (rather involved) pricing system. When the user enters the site, he has some intention (e.g. to buy some organic fruits) and puts a query into the search mask. While the intention usually remains hidden, the publisher does have access to some user data as search query, time of the year or location. Based on this information he chooses the number and kind of ads that are chosen. In particular, we are concentrating now on a parameter that is called the main line reserve which determines the number of ads shown in the mainline.

Making money In practice, the publisher can control the edge "user data → main line reserve", that is he can decide which conditional $p(\text{main line reserve} \,|\, \text{user data})$ to use. Assume that the publisher lets the system run for a while and observes data

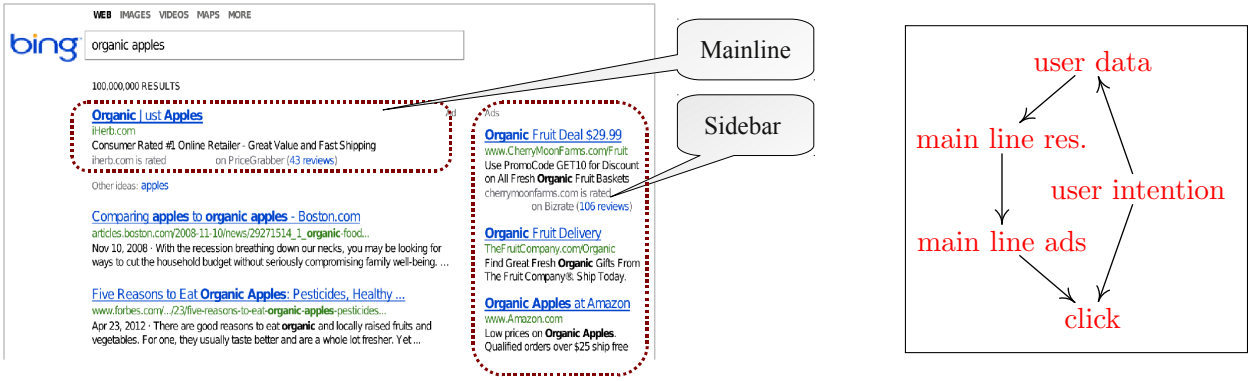


Figure 1.5: Search results (left) and simplified version of an advertisement system (right)

from this system. He would then like to know whether he could perform even better. That is, would a different parameter setting $p(\text{main line reserve} \,|\, \text{user data})$ lead to a higher expected number of clicks? Again, we are interested in the system's behavior under a different distribution $\tilde{\mathbb{P}}^{\mathbf{X}} \neq \mathbb{P}^{\mathbf{X}}$.

Disclaimer In practice the system is more complicated since one may want to take into account the bids of the advertiser. Also, the publisher has to take care of some long-term goals: showing too many or misleading ads, may lead to more clicks but may also annoy users which then decide to use another search website or install an adblock system (which, by the way, is available for free and very easy to install).

1.2 Some bits of probability and statistics

Throughout the lecture we use the following notation.

- $(\Omega, \mathcal{F}, \mathbb{P})$: probability space, where Ω, $\mathcal{F}$ and $\mathbb{P}$ are set, σ-algebra and probability measure, respectively.
- We use capital letters for real-valued random variables. E.g., $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a measurable function, with respect to the Borel σ-algebra.
- We usually denote vectors with bold letters.
- $\mathbb{P}^{\mathbf{X}}$ is the distribution of the p-dimensional random vector $\mathbf{X}$, i.e. a probability measure on $(\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p})$.
- We write $x \mapsto p_X(x)$ or simply $x \mapsto p(x)$ for the Radon-Nikodym derivative of $\mathbb{P}^X$ either with respect to the Lebesgue or the counting measure. We (sometimes implicitly) assume its existence or continuity.
- We call X **independent** of Y and write $X \perp\!\!\!\perp Y$ if and only if

$$p(x, y) = p(x)p(y) \tag{1.1}$$

for all x, y. Otherwise, X and Y are **dependent** and we write $X \not\perp\!\!\!\perp Y$.

- We call $X_1, \ldots, X_p$ **jointly (or mutually) independent** if and only if

$$p(x_1, \ldots, x_p) = p(x_1) \cdot \ldots \cdot p(x_p) \tag{1.2}$$

  for all $x_1, \ldots, x_p$.
- We call $\mathbf{X}$ **independent** of $\mathbf{Y}$ **conditional on** $\mathbf{Z}$ and write $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}$ if and only if

$$p(\mathbf{x}, \mathbf{y} \,|\, \mathbf{z}) = p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{y} \,|\, \mathbf{z}) \tag{1.3}$$

  for all $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that $p(\mathbf{z}) > 0$. Otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are dependent conditional on $\mathbf{Z}$ and we write $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}$. Here, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are allowed to be random vectors.
- The **variance** of a random variable X is defined as

$$\mathbf{var} X := \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2$$

  if $\mathbf{E}X^2 < \infty$.
- We call X and Y **uncorrelated** if $\mathbf{E}X^2, \mathbf{E}Y^2 < \infty$ and

$$\rho_{X,Y} := \frac{\mathbf{E}XY - \mathbf{E}X\mathbf{E}Y}{\sqrt{\mathbf{var}X \mathbf{var}Y}} = 0\,.$$

  Otherwise, that is if $\rho_{X,Y} \neq 0$, X and Y are correlated. $\rho_{X,Y}$ is called the correlation coefficient between X and Y. If X and Y are independent, then they are uncorrelated.
- We say that X and Y are **partially uncorrelated given** Z if

$$\rho_{X,Y \,|\, Z} := \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Z,Y}}{\sqrt{(1 - \rho_{X,Z}^2)(1 - \rho_{Z,Y}^2)}} = 0\,.$$

  The following interpretation of partial correlation is important: $\rho_{X,Y \,|\, Z}$ equals the correlation between residuals after linearly regressing X on Z and Y on Z.
- In general, we have

$$\begin{aligned} \rho_{X,Y \,|\, Z} = 0 \quad &\not\Rightarrow \quad X \perp\!\!\!\perp Y \,|\, Z \qquad \text{and} \\ \rho_{X,Y \,|\, Z} = 0 \quad &\not\Leftarrow \quad X \perp\!\!\!\perp Y \,|\, Z\,. \end{aligned}$$

  The latter holds because a linear regression does not necessarily remove all the dependence from Z in X: after linearly regressing X on Z, there might still be dependence between the residuals and Z.
- Given finitely many data we do not expect the empirical correlation (or any independence measure) to be exactly zero. We therefore make use of statistical hypothesis tests. To test for vanishing correlation, we can use the empirical correlation coefficient and a t-test (for Gaussian variables) or Fisher's z-transform [e.g. `cor.test` in RProject, 2015].
  As an independence test, we can use a test of vanishing partial correlation in case of multivariate normal data or we may use a χ^2-test for discrete or discretized data

or the Hilbert-Schmidt Independence Criterion (HSIC), see [Gretton et al., 2008]. As usual, the null hypothesis is chosen to be vanishing correlation or independence of the variables. Note, however, that in causal inference we do not necessarily want to treat type I error and type II error equally. We will see in Section 4 that some methods for causal structure learning make use of both independences and dependences.

- In a slight abuse of notation we consider sets of variables $\mathbf{B} \subseteq \mathbf{X}$ as a single multivariate variable.

For an introduction to measure theory, see for example [Dudley, 2002].

1.3 Graphs

We start with some basic notation for graphs. Consider finitely many random variables $\mathbf{X} = (X_1, \ldots, X_p)$ with index set $\mathbf{V} := \{1, \ldots, p\}$, joint distribution $\mathbb{P}^{\mathbf{X}}$ and density $p(\mathbf{x})$.

Definition 1.3.1 A **graph** $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of (finitely many) **nodes** or **vertices** $\mathbf{V}$ and **edges** $\mathcal{E} \subseteq \mathbf{V}^2$ with $(v, v) \notin \mathcal{E}$ for any $v \in \mathbf{V}$.

We now introduce graph terminology that we require later. Most of the definitions can be found in Spirtes et al. [2000], Koller and Friedman [2009] and Lauritzen [1996], for example. The terminology is meant to be self-explanatory, it is widely used. When reading papers it usually suffices to check some details in the definitions; e.g, is a node descendant of itself?

- Let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be a graph with $\mathbf{V} := \{1, \ldots, p\}$ and corresponding random variables $\mathbf{X} = (X_1, \ldots, X_p)$. A graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a **subgraph** of $\mathcal{G}$ if $\mathbf{V}_1 = \mathbf{V}$ and $\mathcal{E}_1 \subseteq \mathcal{E}$; we then write $\mathcal{G}_1 \leq \mathcal{G}$. If additionally, $\mathcal{E}_1 \neq \mathcal{E}$, $\mathcal{G}_1$ is a **proper subgraph** of $\mathcal{G}$.
- A node i is called a **parent** of j if $(i, j) \in \mathcal{E}$ and $(j, i) \notin \mathcal{E}$ and a **child** if $(j, i) \in \mathcal{E}$ and $(i, j) \notin \mathcal{E}$. The set of parents of j is denoted by $\mathbf{PA}_j^{\mathcal{G}}$, the set of its children by $\mathbf{CH}_j^{\mathcal{G}}$. Two nodes i and j are **adjacent** 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. We say that there is an **undirected edge** between two adjacent nodes i and j if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$. An edge between two adjacent nodes is **directed** if it is not undirected. We then write $i \to j$ for $(i, j) \in \mathcal{E}$. Three nodes are called an **immorality** or a **v-structure** if one node is a child of the two others that themselves are not adjacent. The **skeleton** of $\mathcal{G}$ does not take the directions of the edges into account: it is the graph $(\mathbf{V}, \tilde{\mathcal{E}})$ with $(i, j) \in \tilde{\mathcal{E}}$, if $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$.
- A **path** in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, \ldots, i_n$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \ldots, n-1$. If $i_k \to i_{k+1}$ for all k we speak of a **directed path** from i_1 to i_n and call i_n a **descendant** of i_1. In this work, i is neither a descendant nor a non-descendant of itself. We denote all descendants of i by $\mathbf{DE}_i^{\mathcal{G}}$ and all non-descendants of i, excluding i, by $\mathbf{ND}_i^{\mathcal{G}}$. If $i_{k-1} \to i_k$ and $i_{k+1} \to i_k$, i_k is called a **collider relative to this path**.
- $\mathcal{G}$ is called a **partially directed acyclic graph (PDAG)** if there is no directed cycle, i.e., if there is no pair $(j,\ k)$ with directed paths from j to k and from k to j. $\mathcal{G}$ is called a **directed acyclic graph (DAG)** if it is a PDAG and all edges are directed.

- In a DAG, a path between i_1 and i_n is **blocked by a set S** (with neither i_1 nor i_n in S) whenever there is a node i_k, $1 < k < n$, such that one of the following two possibilities holds:

  1. $i_k \in \mathbf{S}$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$
  2. $i_{k-1} \to i_k \leftarrow i_{k+1}$ and neither 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 $\mathbf{S}$; otherwise, $\mathbf{A}$ and $\mathbf{B}$ are *d*-**connected** by $\mathbf{S}$.
- Given a DAG $\mathcal{G}$, we obtain the undirected **moralized graph** $\mathcal{G}^m$ of $\mathcal{G}$ by connecting the parents of each node and removing the directions of the edges.
-
- $\mathbf{An}(A)$ is the smallest **ancestral set** of A, where a set $U \subseteq V$ is ancestral for A if and only if $an_j \subseteq U$ for all $j \in A$.
- $G_{V'}$ is the graph induced by keeping a subset $V' \subseteq V$ of nodes.
- In a slight abuse of notation we identify the nodes $j \in \mathbf{V}$ with variables X_j from a random vector $\mathbf{X} = (X_1, \ldots, X_p)$, see Section 1.2, the context should clarify the meaning.

Proposition 1.3.2 *Sets* $\mathbf{A}$ *and* $\mathbf{B}$ *are d-separated by* $\mathbf{S}$ *if and only if* $\mathbf{A}$ *and* $\mathbf{B}$ *are separated in the undirected graph* $(\mathcal{G}_{\mathbf{An}(A \cup B \cup S)})^m$*, that is there is no path in* $(\mathcal{G}_{\mathbf{An}(A \cup B \cup S)})^m$ *between a node in* $\mathbf{A}$ *and a node in* $\mathbf{B}$ *that is not passing through a node in* $\mathbf{S}$

For a proof, see Lauritzen Proposition 3.25.

Definition 1.3.3 Given a DAG $\mathcal{G}$, we say that a $\pi \in S_p$, that is a bijective mapping

$$\pi : \{1, \ldots, p\} \to \{1, \ldots, p\},$$

is a *topological (or causal) ordering* of the variables if it satisfies

$$\pi(i) < \pi(j) \quad \text{if} \quad j \in \mathbf{DE}_i^{\mathcal{G}}.$$

Because of the acyclic structure of the DAG, there is always a topological ordering (see below). But this order does not have to be unique. The node $\pi^{-1}(1)$ is a source node, $\pi^{-1}(p)$ a sink node.

Proposition 1.3.4 *For each DAG there is a topological ordering.*

Proof. We need to show that each DAG has a node without any ancestors: start with any node and move to one of its parents (if there are any). You will never visit a parent that you have seen before (if you did there had been a directed cycle). At latest after $p - 1$ steps you reach a node without any parent. □

Definition 1.3.5 We can represent a DAG $\mathcal{G} = (V, \mathcal{E})$ over p nodes with a binary $p \times p$ matrix A (taking values 0 or 1):

$$A_{i,j} = 1 \quad \Leftrightarrow \quad (i,j) \in \mathcal{E}\,.$$

A is called the **adjacency matrix** of $\mathcal{G}$.

Remark 1.3.6 (i) Let A be the adjacency matrix for DAG $\mathcal{G}$. The entry (i,j) of A^2 equals the number of directed paths of length 2 from i to j because of

$$A^2_{i,j} = \sum_k A_{ik} A_{kj}\,.$$

(ii) In general, we have

$$A^k_{ij} = \#\text{ directed paths of length } k \text{ from } i \text{ to } j$$

(iii) If there is a DAG with the identity map is in causal order, its adjacency matrix is upper triangular, i.e., only the upper-right half of the matrix contains non-zeros.

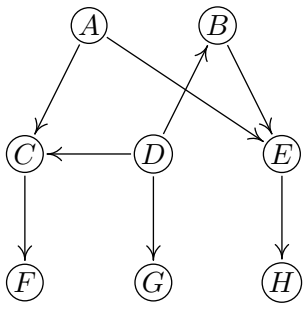
(iv) We may want to use sparse matrices when the graph is sparse in order to save space and/or computation time.

The number of DAGs with p nodes have been studied by Robinson [1970, 1973], and independently by Stanley [1973]. The number of such matrices (or DAGs) is growing very quickly in p, see Table 1.3. McKay [2004] proves the following equivalent description of DAGs which had been conjectured by Eric W. Weisstein.

Theorem 1.3.7 *The matrix A is an adjacency matrix of a DAG $\mathcal{G}$ if and only if $A + Id$ is a $0-1$ matrix with all eigenvalues being real and strictly greater than zero.*

1.4 Exercises

Exercise 1.4.1 *For the following graph $\mathcal{G}$*



write down

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	3160345939641891760 7425
12	521939651343829405020504063
13	186766007444320351866648169 26721
14	143942814104439833494179071 9839535103
15	237725265553410354992180218286376719253505
16	837566707737333202876993030479964122352231383 03
17	627079211969238898994464526024949219069635514826 75201
18	994211953221595158952289145923545245165550268785883050147 83
19	332771901227107591736177573311261125883583076258421902583546773505
20	234488045105108898815255985522909918889908119223429129879580323606849 1263

Table 1.2: The number of DAGs depending on the number p of nodes, taken from `http://oeis.org/A003024` (Feb 2015).

a) the non-descendants of D,

b) all variables that are d-separated from A given F, D.

c) all sets of variables that you can condition on in order to d-separate A and D.

Exercise 1.4.2 *Which graphs satisfy the following d-separation statements? (Assume, that these are <u>all</u> d-separations that can be found in the graphs.)*

a) Consider graphs with three nodes A, B and C such that

$\cdot$	*AND*	$\cdot$	*d-separated by*
A		C	$\{B\}$

b) Consider graphs with four nodes A, B, C and D such that

$\cdot$	*AND*	$\cdot$	*d-separated by*
A		C	$\emptyset$
A		D	$\{B\}$
A		D	$\{B, C\}$
D		C	$\{B\}$
D		C	$\{B, A\}$

Chapter 2

Structural equation models

Structural equation models have been used for a long time in fields like agriculture or social sciences [e.g., Wright, 1921a, 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. In Section 4, we consider the question of identifiability.

2.1 Definitions and first properties

Definition 2.1.1 A *structural equation model (SEM)* (also called a functional model) is defined as a tuple $\mathcal{S} := (\mathbf{S}, \mathbb{P}^{\mathbf{N}})$, where $\mathbf{S} = (S_1, \ldots, S_p)$ is a collection of p equations

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

where $\mathbf{PA}_j \subseteq \{X_1, \ldots, X_p\} \setminus \{X_j\}$ are called *parents of* X_j and $\mathbb{P}^{\mathbf{N}} = \mathbb{P}^{N_1,\ldots,N_p}$ is the joint distribution of the noise variables, which we require to be jointly independent, i.e., $\mathbb{P}^{\mathbf{N}}$ is a product distribution. The graph of a structural equation model is obtained simply by drawing direct edges from each parent to its direct effects, i.e., from each variable X_k occurring on the right-hand side of equation (2.1) to X_j, see Figure 2.1. We henceforth assume this graph to be acyclic. According to the notation defined in Section 1.3, $\mathbf{PA}_j$ are the parents of X_j.

Proposition 2.1.2 *Because of the acyclic structure an SEM defines a unique distribution over the variables* $(X_1, \ldots, X_p)$ *such that* $X_j \stackrel{d}{=} f_j(\mathbf{PA}_j, N_j)$ *for* $j = 1, \ldots, p$.

Proof. Using a topological ordering π we can write each node j as a function of the noise terms N_k with $\pi(k) \leq \pi(j)$ (use the structural equations iteratively). That is,

$$X_j = g_j((N_k)_{k\,:\,\pi(k)\leq\pi(j)})\,.$$

□

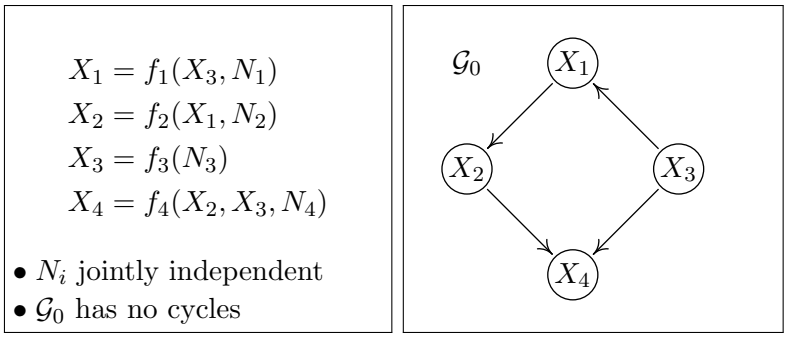


Figure 2.1: Example of a structural equation model (SEM) (left) with corresponding graph (right). There is only one topological ordering π (that satisfies $3 \mapsto 1$, $1 \mapsto 2$, $2 \mapsto 3$, $4 \mapsto 4$).

We use the SEM to define not only the distribution of observed data but also so-called interventional distributions (see Remark 2.2.5, for example). These are formally defined in Definition 2.2.1.

Remark 2.1.3 (i) It may be helpful to think about generating n samples from this distribution: one first samples $(\mathbf{N}_1, \dots, \mathbf{N}_n) \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{N}}$ and then subsequently uses the structural equations (starting from a source node $\pi^{-1}(1)$, then $\pi^{-1}(2)$, and so on) to generate samples from the X_j.

(ii) Definition 2.1.1 is purely mathematical, we relate SEMs to reality in Remark 2.2.5. The parents $\mathbf{PA}_j$ may then be thought of as the direct causes of X_j. An SEM specifies how the $\mathbf{PA}_j$ affect X_j. Note that for many authors, SEMs already have a causal meaning. In this script, we try to separate mathematical from the causal language.

(iii) In physics (chemistry, biology, ...), we would usually expect that such causal relationships occur in time, and are governed by sets of coupled differential equations. Under certain assumptions such as stable equilibria, one can derive an SEM that describes how the equilibrium states of such a dynamical system will react to physical interventions on the observables involved [Mooij et al., 2013]. In this lecture, we do not deal with these issues but take the SEM as our starting point instead.

(iv) The model class of SEMs, i.e. the set of distributions that can be generated by an SEM, is the set of *all* distributions. We will see later (Proposition 2.5.2) that each distribution can be generated by many SEM's with a fully connected graph, for example.

(v) It seems surprising that the two SEMs $\mathbf{S}_1 : X = N_X, Y = N_Y$ and $\mathbf{S}_2 : X = N_X, Y = 0 \cdot X + N_Y$ correspond to different graphs; see also causal minimality (Definition 2.4.10).

(vi) This is one of the reasons why we should not use the structural equations (2.1) as usual equations. They should be thought of as a tool that tells us how to generate a distribution (see Proposition 2.1.2) and the intervention distributions (see Section 2.2).

(vii) The goal in Chapter 4 will be to estimate the causal structure from the joint distribution. Remark (iv) shows that we will need additional assumptions. It turns out that finding a causal order π is difficult. Assume that π is given, i.e. we have:

$$\begin{aligned} X &= N_X \\ Y &= f(X, N_Y) \\ Z &= g(X, Y, N_Z) \end{aligned}$$

with unknown f, g, N_X, N_Y, N_Z. Deciding whether f depends on X, and g depends on X and/or Y is a well-studied significance problem in "traditional" statistics (herefore, one often assumes an easier model class, e.g. linear functions and additive noise).

2.2 Interventions

We are now ready to use the structure of SEMs to construct the "other distributions" $\tilde{\mathbb{P}}^{\mathbf{X}}$ from $\mathbb{P}^{\mathbf{X}}$.

Definition 2.2.1 [Intervention Distribution] Consider a distribution $\mathbb{P}^{\mathbf{X}}$ that has been generated from an SEM $\mathcal{S} := (\mathcal{S}, \mathbb{P}^{\mathbf{N}})$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{\mathcal{S}}$. We call the distributions in the new SEM *intervention distributions* and say that the variables whose structural equation we have replaced have been "intervened on". We denote the new distribution by[1]

$$\mathbb{P}^{\mathbf{X}}_{\tilde{\mathcal{S}}} = \mathbb{P}^{\mathbf{X} \,|\, do\left(X_j = \tilde{f}(\widetilde{\mathbf{PA}}_j, \tilde{N}_j)\right)}_{\mathcal{S}}.$$

The set of noise variables in $\tilde{\mathcal{S}}$ now contains both some "new" $\tilde{N}$'s and some "old" N's and is required to be mutually independent.

When $\tilde{f}(\widetilde{\mathbf{PA}}_j, \tilde{N}_j)$ puts a point mass on a real value a, we simply write $\mathbb{P}^{\mathbf{X} \,|\, do(X_j = a)}_{\mathcal{S}}$ and call this a **perfect** intervention[2]. An intervention with $\widetilde{\mathbf{PA}}_j = \mathbf{PA}_j$ is called

[1] Although the set of parents can change arbitrarily (as long as they are not introducing cycles), we mainly consider interventions, for which the new set of parents $\widetilde{\mathbf{PA}}_j$ is either empty or equals $\mathbf{PA}_j$.

[2] This is also referred to as an **ideal**, **structural** [Eberhardt and Scheines, 2007], **surgical** [Pearl, 2009], **independent** or **deterministic** [Korb et al., 2004] intervention.

imperfect[3]. It's a special case of a **stochastic** intervention [Korb et al., 2004], in which the marginal distribution of the intervened variable has positive variance.

(Because of acyclicity the set of allowed interventions depends on the graph induced by $\mathcal{S}$.) It turns out that this simple concept is a powerful tool to model differences in distributions and to understand causal relationships. We try to illustrate this with a couple of examples.

Example 2.2.2 ["Cause-Effect"] Suppose that $\mathbb{P}^{(X,Y)}$ is induced by a structural equation model $\mathcal{S}$

$$X = N_X \tag{2.2}$$
$$Y = 4 \cdot X + N_Y \tag{2.3}$$

with $N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and graph $X \to Y$. Then,

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}^{Y} = \mathcal{N}(0,17) &\neq \mathcal{N}(8,1) = \mathbb{P}_{\mathcal{S}}^{Y\,|\,do(X=2)} = \mathbb{P}_{\mathcal{S}}^{Y\,|\,X=2} \\ &\neq \mathcal{N}(12,1) = \mathbb{P}_{\mathcal{S}}^{Y\,|\,do(X=3)} = \mathbb{P}_{\mathcal{S}}^{Y\,|\,X=3}.\end{aligned}$$

Intervening on X changes the distribution of Y.

But on the other hand,

$$\mathbb{P}_{\mathcal{S}}^{X\,|\,do(Y=2)} = \mathcal{N}(0,1) = \mathbb{P}_{\mathcal{S}}^{X} = \mathbb{P}_{\mathcal{S}}^{X\,|\,do(Y=314159265)} \neq \mathbb{P}_{\mathcal{S}}^{X\,|\,Y=2}. \tag{2.4}$$

No matter how strongly we intervene on Y, the distribution of X remains what it was before. This model behavior corresponds well to our intuition of X is "causing" Y: no matter how much we whiten someone's teeth, this will not have any effect on his smoking habits. Equation (2.4) further shows that intervening is different from conditioning, see also Example 3.1.1.

The asymmetry between cause and effect can also be formulated as an independence statement: When we replace the structural equation for Y with $Y = \tilde{N}_Y$, we break the dependence between X and Y: in $\mathbb{P}_{\mathcal{S}}^{X,Y\,|\,do\left(Y=\tilde{N}_Y\right)}$ we find $X \perp\!\!\!\perp Y$. This does not hold for $\mathbb{P}_{\mathcal{S}}^{X,Y\,|\,do\left(X=\tilde{N}_X\right)}$ as long as $\mathbf{var}(\tilde{N}_X) \neq 0$: the correlation between X and Y is non-zero.

We use the latter statement in the preceding Example 2.2.2 for defining the existence of a (total) causal effect.

[3] This has also been referred to as a **parametric** [Eberhardt and Scheines, 2007] or **dependent** intervention [Korb et al., 2004] or simply as a **mechanism change** [Tian and Pearl, 2001]. Unfortunately, the term **soft** intervention can either mean the same thing [Eberhardt and Scheines, 2007] and is also used for an intervention that increases the chances that a node takes a particular value [Eaton and Murphy, 2007, Markowetz et al., 2005]

Definition 2.2.3 [total causal effect] Given an SEM $\mathcal{S}$, there is a **(total) causal effect** from X to Y if and only if

$$X \not\perp\!\!\!\perp Y \quad \text{in } \mathbb{P}_{\mathcal{S}}^{\mathbf{X} \mid do\left(X=\tilde{N}_X\right)}$$

for some variable $\tilde{N}_X$.

There are several equivalent statements.

Proposition 2.2.4 *Given an SEM S, the following statements are equivalent*

(i) There is a causal effect from X to Y.

(ii) There are $x^{\triangle}$ and $x^{\square}$, such that $\mathbb{P}_{\mathcal{S}}^{Y \mid do\left(X=x^{\triangle}\right)} \neq \mathbb{P}_{\mathcal{S}}^{Y \mid do\left(X=x^{\square}\right)}$.

(iii) There is $x^{\triangle}$, such that $\mathbb{P}_{\mathcal{S}}^{Y \mid do\left(X=x^{\triangle}\right)} \neq \mathbb{P}_{\mathcal{S}}^{Y}$.

(iv) $X \not\perp\!\!\!\perp Y$ in $\mathbb{P}_{\mathcal{S}}^{X,Y \mid do\left(X=\tilde{N}_X\right)}$ for any $\tilde{N}_X$ whose distribution has full support.

The proof can be found in Appendix A.2.1.

Remark 2.2.5 [the "correct" SEM] So far SEMs are mathematical objects. We regard them as models for a data generating process both with and without interventions in real life. It is a complicated model though. Instead of modeling "just" a joint distribution (as we can model a physical process with a Poisson process, for example) we now model the system in an observational state and under perturbations at the same time.

Formally, we say that an SEM $\mathcal{S}$ over $\mathbf{X} = (X_1, \ldots, X_p)$ is a correct model (the "correct SEM") for the underlying data generating process if the observational distribution is correct and all interventional distributions $\mathbb{P}_{\mathcal{S}}^{\mathbf{X} \mid do\left(X_j=\tilde{N}_j\right)}$ correspond to distributions that we obtain from randomized experiments[4]. Importantly, an SEM is therefore falsifiable (if we can do the randomized experiments).

For the rest of this section we usually provide the correct SEM. Under what kind of assumptions we can obtain the SEM from real data is the question of Chapter 4.

Example 2.2.6 [Randomized trials] In randomized trials we randomly assign the treatment T according to $\tilde{N}_T$ to a patient (this may include a placebo). In the SEM, this is modeled with observing data from the distribution $\mathbb{P}_{\mathcal{S}}^{\mathbf{X} \mid do\left(T=\tilde{N}_T\right)}$. If we then still find a dependence between the treatment and recovery, for example, we conclude that T has a total causal effect on the recovery.

The idea of using randomized trials for causal inference was described (using different mathematical language) by C.S. Peirce [Peirce, 1883, Peirce and Jastrow, 1885] and

[4]This includes the assumption that there is an agreement about what a randomized experiment should look like.

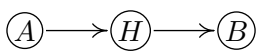
later by J. Neyman [Splawa-Neyman et al., 1990, a translated and edited version of the original article] and R.A. Fisher [Fisher, 1925], for applications in agriculture.

One of the first examples of a randomized experiment was performed by James Lind. During the 18th century Great Britain lost more soldiers due to scurvy than to enemy action. James Lind thought that scurvy is a putrefaction of the body and expected acids to be helpful. In 1747, he treated 12 sailors who caught the disease in 6 different ways: with apple cider, drops of sulfuric acid, vinegar, sea water, two oranges and one lemon and barley water respectively. After a couple of days the two subjects treated with citrus fruits had recovered and the two people drinking cider showed first signs of recovery [Wikipedia, 2015].

Example 2.2.7 Consider the following SEM[5]:

$$\mathbf{S}: \quad \begin{aligned} A &= N_A \\ H &= A \oplus N_H \\ B &= H \oplus N_B \end{aligned}$$

with graph



where $N_A \sim \text{Ber}(1/2)$, $N_H \sim \text{Ber}(1/3)$ and $N_B \sim \text{Ber}(1/20)$ are independent. The symbol $\oplus$ denotes addition modulo 2 (i.e. $1 \oplus 1 = 0$). Although B is in some sense a better predictor for H than A (e.g. the mutual information between B and H is larger than the mutual information between A and H), an intervention on A has a larger influence on H than intervening on B. More precisely, we have that

$$\mathbb{P}_{\mathcal{S}}^{H \,|\, do(B=1)} = \mathbb{P}_{\mathcal{S}}^{H} \qquad \text{(forcing } B \text{ to be one)}$$

and

$$\mathbb{P}_{\mathcal{S}}^{H \,|\, do(A=1)} = \text{Ber}(2/3) \neq \text{Ber}(1/2) = \mathbb{P}_{\mathcal{S}}^{H} \qquad \text{(forcing } A \text{ to be one)}$$

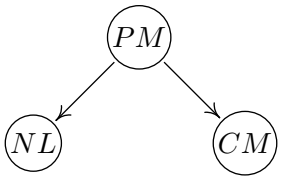
We now revisit the example about myopia (the example about chocolate and Nobel prizes works analogously).

Example 2.2.8 [Myopia, cont.] Assume that the underlying ("correct") SEM is of the form

$$\mathbf{S}: \quad \begin{aligned} PM &= N_{PM} \\ NL &= f(PM, N_{NL}) \\ CM &= g(PM, N_{CM}) \end{aligned}$$

where PM stands for parent myopia, NL for night light and CM for child myopia. The corresponding graph is

[5]This example was provided by Nicolai Meinshausen.



Quinn et al. [1999] found that $NL \not\perp\!\!\!\perp CM$ but if we replace the structural equation of NL with $NL = \tilde{N}_{NL}$, we have $NL \perp\!\!\!\perp CM$ in the intervention distribution (since $CM = g(N_{PM}, N_{CM})$). This holds for any variable $\tilde{N}_{NL}$, in particular for variables with full support. Thus, there is no causal effect from NL to CM.

In general, we have that

Proposition 2.2.9 *(i) If there is no directed path from X to Y, then there is no causal effect.*

(ii) Sometimes there is a directed path but no causal effect.

The proof can be found in Appendix A.2.2.

2.3 Counterfactuals

The definition and interpretation of counterfactuals has received a lot of attention in literature. They concern the following situation: assume you are playing poker and as a starting hand you have ♣J and ♣3 (sometimes called a "lumberjack" - tree and a jack); you fold because you estimate the probability of winning not to be high enough. The flop, however, turns out to be ♣4, ♣Q and ♣2. The reaction is a typical counterfactual statement: "If I had stayed in the game, my chances would have been good.".

Definition 2.3.1 Consider an SEM $\mathcal{S} := (\mathbf{S}, \mathbb{P}^{\mathbf{N}})$ over nodes $\mathbf{X}$. Given some observations $\mathbf{x}$, we define a *counterfactual SEM* by replacing the distribution of noise variables:

$$\mathcal{S}_{\mathbf{X}=\mathbf{x}} := (\mathbf{S}, \mathbb{P}^{\mathbf{N}}_{\mathcal{S},\mathbf{X}=\mathbf{x}}),$$

where $\mathbb{P}^{\mathbf{N}}_{\mathcal{S},\mathbf{X}=\mathbf{x}} := \mathbb{P}^{\mathbf{N}\,|\,\mathbf{X}=\mathbf{x}}$. The new set of noise variables need not be mutually independent anymore. *Counterfactual statements* can now be seen as *do*-statements in the new counterfactual SEM[6].

This definition can be generalized such that we observe not the full vector $\mathbf{X} = \mathbf{x}$ but only some of the variables.

[6]for simplicity, we consider only *do*-statements, for which the replaced structural equation contains a new noise variable that is independent of all other noise variables

Example 2.3.2 Consider the following SEM

$$\begin{aligned} X &= N_X \\ Y &= X^2 + N_Y \\ Z &= 2 \cdot Y + X + N_Z \end{aligned}$$

with $N_X, N_Y, N_Z \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Now, assume that we observe $\mathbf{X} = (X, Y, Z) = (1, 2, 4)$. Then $\mathbb{P}^{\mathbf{N}}_{\mathcal{S},\mathbf{X}=\mathbf{x}}$ puts a point mass on $(N_X, N_Y, N_Z) = (1, 1, -1)$. We therefore have the counterfactual statement (in the context of $(X, Y, Z) = (1, 2, 4)$): "Z would have been 11, had X been 2." Mathematically, this means that $\mathbb{P}^{Z \,|\, do(X=2)}_{\mathcal{S},\mathbf{X}=\mathbf{x}}$ has a point mass on 11.

In the same way, we obtain "Y would have been 5, had X been 2." and "Z would have been 10, had Y been 5."

Example 2.3.3 Consider the following made up scenario: a patient with poor eyesight comes to the hospital and goes blind ($B = 1$) after the doctor suggests the treatment $T = 1$. Let us assume that the correct SEM has the form

$$S: \begin{array}{rcl} T &=& N_T \\ B &=& T \cdot N_B + (1 - T) \cdot (1 - N_B) \end{array}$$

with $N_B \sim \text{Ber}(0.01)$ and corresponding graph $T \to B$. The question: "What would have happened had the doctor decided to give treatment $T = 0$?" can be answered with

$$\mathbb{P}^{B \,|\, do(T=0)}_{\mathcal{S},B=1,T=1} = \text{Ber}(0)\,,$$

i.e.,

$$\mathbb{P}_{\mathcal{S},B=1,T=1}(B = 0 \,|\, do\,(T = 0)) = 1\,,$$

the patient would have been cured ($B = 0$) if the doctor had given him treatment $T = 0$. Because of

$$\begin{aligned} &\mathbb{P}_{\mathcal{S}}(B = 0 \,|\, do\,(T = 1)) = 0.99 \quad \text{and} \\ &\mathbb{P}_{\mathcal{S}}(B = 0 \,|\, do\,(T = 0)) = 0.01\,, \end{aligned}$$

however, we can still argue that the doctor acted optimally (according to his knowledge).

Counterfactual statements depend strongly on the structure of the SEM. The following example shows two SEMs that agree on all observational and interventional statements but predict different counterfactual statements.

Example 2.3.4 Let $N_1, N_2 \sim \text{Ber}(0.5)$ and $N_3 \sim \text{U}(\{0, 1, 2\})$, such that the three variables are jointly independent. That is, N_1, N_2 have a Bernoulli distribution with parameter

0.5 and N_3 is uniformly distributed on $\{0, 1, 2\}$. We define two different SEMs, first consider $\mathcal{S}_A$:

$$\begin{aligned} X_1 &= N_1 \\ X_2 &= N_2 \\ X_3 &= (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} + N_3 \cdot 1_{X_1=X_2} \,. \end{aligned}$$

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:

$$\begin{aligned} X_1 &= N_1 \\ X_2 &= N_2 \\ X_3 &= (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} + (2 - N_3) \cdot 1_{X_1=X_2} \,. \end{aligned}$$

It can be checked that both SEMs generate the same observational distribution, which satisfies causal minimality (see Definition 2.4.10) with respect to the graph $X_1 \to X_3 \leftarrow X_2$. They also generate the same intervention distributions, for any possible intervention. But the two models differ in a counterfactual statement. 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 SEMs it follows that $N_3 = 0$, and thus the two SEMs $\mathcal{S}_A$ and $\mathcal{S}_B$ "predict" different values for X_3 under a counterfactual change of X_1 (namely 0 and 2 respectively).

If we want to use an estimated SEM to predict counterfactual questions, this example shows that we require assumptions that let us distinguish between $\mathcal{S}_A$ or $\mathcal{S}_B$.

We now summarize some properties of counterfactuals.

Remark 2.3.5 (i) Counterfactual statements are not transitive. In Example 2.3.2 we found that given the observation $(X, Y, Z) = (1, 2, 4)$, Y would have been 5, had X been 2 and Z would have been 10, had Y been 5 but Z would have not been 10 had X been 2.

(ii) Humans often think in counterfactuals: "I should have taken the train.", "Do you remember our flight to New York on Sep 11th 2000? Imagine we would have taken the flight one year later!" and "Imagine we would have invested in CHF last year." are only few examples. Interestingly, this sometimes even concerns situations in which we made optimal decisions (based on the available information). Assume, someone offers you $\$10,000$ if you predict the result of a coin flip, you guess 'heads' and lose. How many people would think: "Why didn't I say 'tails'?" Discussing whether counterfactual statements contain any information that can help us making better decisions in future is interesting but lies beyond this work.

(iii) Similarly, we cannot provide details about the role of counterfactuals in our law system. The question whether counterfactuals should be taken as a basis of verdicts, for example, seems interesting to us though (see Example 2.3.3).

(iv) Thinking about counterfactuals has been done since a long time; it is a popular tool of historians. Titus Livius, for example, discusses in 25BC what would have happened if Alexander the Great had not died in Asia and had attacked Rome [Geradin and Girgenson, 2011].

(v) We can think of interventional statements as a mathematical construct for (randomized) experiments. For counterfactual statements, there is no apparent correspondence in the real world. But if there is none, these statements may be considered as being not falsifiable and therefore as non-scientific according to Popper [e.g. Popper, 2002].

In summary, a SEM and a causal graphical model can be used to describe either one of the following:

1. the observational distribution
2. all interventional distributions (how does the distribution change if I randomize the treatment variable?)
3. counterfactual statements (what would have happened if I had intervened in a specific way?)

SEM are sometimes meant to imply counterfactual statements even though there is no reason to do so. Counterfactual statements in the example above imply that if I had changed an intervention, the realizations of the noise variables would not change. Whereas interventional distributions do not make such an assumption: if I change an intervention, I get a new realization of the noise variables. This might be correlated (or in fact identical) to the realization of the noise under a different intervention but I can never observe these two outcomes simultaneously and thus it is in most settings safer to just work with the first and second implication of the models and not the counterfactual implications.

2.4 Markov property, faithfulness and causal minimality

We now develop some language that helps us to formalize some intuition we discussed in the preceding sections.

2.4.1 Markov property

The Markov property is a commonly used assumption that is on the basis of graphical modeling. When a distribution is Markov with respect to a graph, this graph encodes certain independencies in the distribution that we can exploit for efficient computation or data storage. The Markov property exists for both directed and undirected graphs and it is well known that these two classes encode different sets of independencies. In causal inference,

however, we are mainly interested in directed graphs. While many introductions to causal inference start with the Markov property as the underlying assumption, we will derive it as a property of SEMs.

Definition 2.4.1 [Markov property] Given a DAG $\mathcal{G}$ and a joint distribution $\mathbb{P}^{\mathbf{X}}$, this distribution is said to satisfy

(i) the **global Markov property** with respect to the DAG $\mathcal{G}$ if

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

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

(ii) the **local Markov property** with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and

(iii) the **Markov factorization property** with respect to the DAG $\mathcal{G}$ if

$$p(\mathbf{x}) = p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p(x_j \,|\, x_{\mathbf{PA}_j^{\mathcal{G}}})$$

(here, we have to assume that $\mathbb{P}^{\mathbf{X}}$ has a density p).

It turns out that as long as the joint distribution has a density[7] these three definitions are equivalent.

Theorem 2.4.2 *If $\mathbb{P}^{\mathbf{X}}$ has a density p (with respect to a product measure), then all Markov properties in Definition 2.4.1 are equivalent.*

The proof can be found as Theorem 3.27 in [Lauritzen, 1996], for example.

Example 2.4.3 A distribution $\mathbb{P}^{X_1,X_2,X_3,X_4}$ is Markov with respect to the graph $\mathcal{G}_0$ shown in Figure 2.1 if, according to (i) or (ii), $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$ and $X_1 \perp\!\!\!\perp X_4 \,|\, X_2, X_3$, or, according to (iii),

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 \,|\, x_3)p(x_2 \,|\, x_1)p(x_4 \,|\, x_2, x_3)\,.$$

We will see later in Proposition 2.5.1 that the distribution generated from the SEM shown on the left hand side in Figure 2.1 on page 18 is Markov w.r.t. $\mathcal{G}_0$.

Definition 2.4.4 [Markov equivalence class of graphs] We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov with respect to $\mathcal{G}$:

$$\mathcal{M}(\mathcal{G}) := \{\mathbb{P} \,:\, \mathbb{P} \text{ satisfies the global (or local) Markov property w.r.t. } \mathcal{G}\}\,.$$

Two DAGs $\mathcal{G}_1$ 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. The set of all DAGs that are Markov equivalent to some DAG (a so-called Markov equivalence class) can be represented by a **completed PDAG** $\mathrm{CPDAG}(\mathcal{G}) = (V, \mathcal{E})$. This graph satisfies $(i, j) \in \mathcal{E}$ if and only if one member of the Markov equivalence class does.

[7] In this script, we always consider densities with respect to Lebesgue or counting measure. For this theorem it suffices if the distribution is absolutely continuous w.r.t. a product measure.

Verma and Pearl [1991] showed that:

Lemma 2.4.5 *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

The following Figure 2.2 shows an example of two Markov equivalent graphs. The graphs share the same skeleton and both of them have the immorality $Z \to V \leftarrow U$.

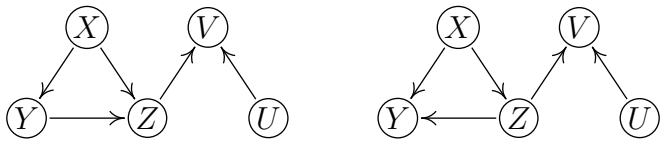


Figure 2.2: Two Markov-equivalent DAGs.

Remark 2.4.6 Consider a graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ and a target node Y. The *Markov blanket of* Y is the smallest set M such that

$$Y \text{ } d\text{-sep. } \mathbf{V} \setminus (\{Y\} \cup M) \text{ by } M\,.$$

If $\mathbb{P}^{\mathbf{X}}$ is Markov w.r.t. $\mathcal{G}$, then

$$Y \perp\!\!\!\perp \mathbf{V} \setminus (\{Y\} \cup M) \text{ given } M\,.$$

If we have a powerful regression technique, we only need to include the variables in M for predicting Y. Given the Markov blanket, the other variables do not provide any further information about Y.

Remark 2.4.7 [Reichenbach's common cause principle] Reichenbach's common cause principle [Reichenbach, 1956] states that when the random variables X and Y are dependent, there must be a "causal explanation" for this dependence:

- X is (possibly indirectly) causing Y or
- Y is (possibly indirectly) causing X or
- there is a (possibly unobserved) confounder T that (possibly indirectly) causes both X and Y.

Here, we do not further specify the meaning of the word "causing".

Proposition 2.4.8 *Assume that any pair of variables X and Y can be embedded into a larger system in the following sense: there exists a correct SEM over the collection $\mathbf{X}$ of random variables that contains X and Y with graph $\mathcal{G}$. Then the Reichenbach's common cause principle follows from the Markov property in the following sense: If X and Y are dependent, then there is*

- *either a directed path from X to Y*
- *or from Y to X*
- *or there is a node T with a directed path from T to X and from T to Y.*

Proof. The proof is immediate: Given dependent variables X and Y we embed them into a larger system of random variables with graph $\mathcal{G}$. Because of the Markov property, $\mathcal{G}$ contains an unblocked path between X and Y. □

In Reichenbach's principle, we start with two dependent random variables and obtain a valid statement. In real applications, however, it might be that we have implicitly conditioned on a third variable ("selection bias"). As the following example shows[8], this may lead to a dependence between X and Y, although there none of the three conditions hold.

Example 2.4.9 Let us assume that whether you study engineering in Zurich ($Z = 1$) is determined only by the fact whether you like nature ($N = 1$) and whether you think ETH is a great university ($U = 1$). More precisely, assume that the correct SEM has the form:

$$\begin{aligned} N &= N_N \,, \\ U &= N_U \,, \\ Z &= \mathrm{OR}(N, U) \oplus N_Z \,, \end{aligned}$$

where $N_N, N_U \overset{\text{iid}}{\sim} \mathrm{Ber}(0.5)$, $N_Z \sim \mathrm{Ber}(0.1)$ and $\mathrm{OR}(N, U)$ equals one if $N = 1$ or $U = 1$ and zero otherwise. Again, $\oplus$ is addition modulo 2, see Example 2.2.7. As we can see from the SEM, N and U are assumed to be independent. If you ask engineering students *in Zurich*, however, i.e. you condition on $Z = 1$, the answers to whether they like nature or ETH become anti-correlated: if someone is not a fan of nature, he probably likes ETH and vice versa (otherwise he would have not studied at ETH). We have that

$$N \not\perp\!\!\!\perp U \,|\, Z = 1 \,.$$

The Markov assumption enables us to read off independencies from the graph structure (i.e. from d-separations). Faithfulness, defined in the following section, allows us to infer dependencies from the graph structure (i.e. from d-connections), see Example 2.4.9.

2.4.2 Faithfulness and causal minimality

Definition 2.4.10 (i) $\mathbb{P}^{\mathbf{X}}$ 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 $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (compare this to the global Markov condition).

[8]The author thanks Marloes Maathuis for pointing out this comment and Dominik Janzing for the example.

(ii) 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 $\mathcal{G}$.

Faithfulness is not very intuitive at first glance. We now give an example of a distribution that is Markov but not faithful with respect to some DAG $\mathcal{G}_1$. This is achieved by making two paths cancel each other and creating an independence that is not implied by the graph structure.

Example 2.4.11 Consider the two graphs in the following figure.

We first look at a linear Gaussian SEM that corresponds to the left graph $\mathcal{G}_1$.

$$\begin{aligned} X &= N_X \,, \\ Y &= aX + N_Y \,, \\ Z &= bY + cX + N_Z \,, \end{aligned}$$

with normally distributed noise variables $N_X \sim \mathcal{N}(0, \sigma_X^2)$, $N_Y \sim \mathcal{N}(0, \sigma_Y^2)$ and $N_Z \sim \mathcal{N}(0, \sigma_Z^2)$ that are jointly independent. This is an example of a linear Gaussian structural equation model with graph $\mathcal{G}_1$, see Definition 2.1.1. Now, if $a \cdot b + c = 0$, the distribution is not faithful with respect to $\mathcal{G}_1$ since we obtain $X \perp\!\!\!\perp Z$; more precisely, it is not triangle-faithful [Zhang and Spirtes, 2008].

Correspondingly, we consider an SEM that corresponds 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 all $\tilde{N}_A \sim \mathcal{N}(0, \tau_A^2)$, $A \in \{X, Y, Z\}$, jointly independent. If we choose $\tau_X^2 = \sigma_X^2$, $\tilde{a} = a$, $\tau_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2$, $\tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2)$ and $\tau_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2)$, 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 observational distribution. It can be checked that the distribution is faithful with respect to $\mathcal{G}_2$ if $\tilde{a}, \tilde{b} \neq 0$ and all $\tilde{\tau_A} > 0$.

The distribution from Example 2.4.11 is faithful with respect to $\mathcal{G}_2$, but not with respect to $\mathcal{G}_1$. Nevertheless, for both models, causal minimality is satisfied if none of the parameters vanishes: the distribution is not Markov to any proper subgraph of $\mathcal{G}_1$ or $\mathcal{G}_2$ since removing an arrow 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 $\mathcal{G}_1$. In general, causal minimality is weaker than faithfulness:

Remark 2.4.12 If $\mathbb{P}^{\mathbf{X}}$ is faithful and Markov with respect to $\mathcal{G}$, then causal minimality is satisfied.

This is due to the fact that any two nodes that are not directly connected by an edge can be d-separated, see Exercise 2.6.2.

It turns out that in most model classes, identifiability is impossible to obtain without causal minimality: we cannot distinguish between $Y = f(X) + N_Y$ and $Y = c + N_Y$, for example, if f is allowed to be constant. At first, we therefore look at an equivalent formulation of causal minimality in the case of SEMs.

Proposition 2.4.13 *Consider the random vector $\mathbf{X} = (X_1, \dots, X_p)$ and assume that the joint distribution has a density with respect to a product measure. Suppose that $\mathbb{P}^{\mathbf{X}}$ is Markov with respect to $\mathcal{G}$. Then $\mathbb{P}^{\mathbf{X}}$ satisfies causal minimality with respect to $\mathcal{G}$ if and only if $\forall X_j \; \forall Y \in \mathbf{PA}_j^{\mathcal{G}}$ we have that $X_j \not\perp\!\!\!\perp Y \,|\, \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}$.*

Proof. See Appendix A.2.5. □

2.5 Some more properties of SEMs

Pearl [2009] shows in Theorem 1.4.1 that the law $\mathbb{P}^{\mathbf{X}}$ generated by an SEM is Markov with respect to its graph.

Proposition 2.5.1 *Assume that $\mathbb{P}^{\mathbf{X}}$ is generated by an SEM with graph $\mathcal{G}$. Then, $\mathbb{P}^{\mathbf{X}}$ is Markov with respect to $\mathcal{G}$.*

We can now come back to the question how large the class of SEMs is. More precisely, we are interested in the question: "Given a distribution $\mathbb{P}^{\mathbf{X}}$, which SEMs can generate this distribution? This can be answered with the following proposition[9].

Proposition 2.5.2 *Consider $X_1, \dots, X_p$ and let $\mathbb{P}^{\mathbf{X}}$ have a strictly positive density with respect to Lebesgue measure and assume it is Markov with respect to $\mathcal{G}$. Then there exists an SEM $(\mathcal{S}, \mathbb{P}^{\mathbf{N}})$ with DAG $\mathcal{G}$ that generates the distribution $\mathbb{P}^{\mathbf{X}}$. In particular, this holds for all fully connected graphs $\mathcal{G}$.*

Proof. See Appendix A.2.3. □

[9]Similar but weaker statements than Proposition 2.5.2 can be found in Druzdzel and Simon [1993], Druzdzel and van Leijen [2001], Janzing and Schölkopf [2010].

Remark 2.5.3 Why do we often work with SEMs and not just with graphs and the Markov condition (i.e. graphical models)? These two are equivalent and can be used interchangeably. Counterfactual statements can be implied (or not) in both formulations, even though most people interpret SEMs as implying counterfactual statements (for historical reasons).

2.6 Exercises

Exercise 2.6.1 *Consider the following structural equation model* $\mathcal{S}$

$$\begin{aligned} V &= N_V \\ W &= -2V + 3Y + 5Z + N_W \\ X &= 2V + N_X \\ Y &= -X + N_Y \\ Z &= \alpha X + N_Z \end{aligned}$$

with $N_V, N_W, N_X, N_Y, N_Z \overset{iid}{\sim} \mathcal{N}(0,1)$.

a) Draw the graph corresponding to the SEM.

b) Set $\alpha = 2$ *and simulate 200 i.i.d. data points from the joint distribution; plot the values of* X *and* W *in order to visualize the distribution* $\mathbb{P}_{\mathcal{S}}^{(X,W)}$.

c) Again, set $\alpha = 2$ *and sample 200 i.i.d. data points from the interventional distribution*

$$\mathbb{P}_{\mathcal{S}}^{(X,W)\,|\,do(Z=1)},$$

in which we have intervened on Z. *Again, plot the samples and compare with the plot from exercise 2.6.1b).*

d) A directed path from one node to another does not necessarily imply that the former node has a causal effect on the latter. Choose a value of α *and prove that for this value* X *has no causal effect on* W.

e) For any given α, *compute*

$$\frac{\partial}{\partial x}\mathbb{E}[W \,|\, do(X = x)]\,.$$

Exercise 2.6.2 *Prove that one can d-separate any two nodes in a DAG* $\mathcal{G}$ *that are not directly connected by an edge. Use this statement to prove Remark 2.4.12.*

Chapter 3

Using the known underlying causal structure

In the following chapters we will make use of an invariance statement. We first state it as a tautology in the hope that this helps the reader to remember it:

> "If we replace only the structural equation for X_j,
> we replace only the structural equation for X_j."

More precisely, we mean that given an SEM $\mathcal{S}$, we have

$$p_{\tilde{\mathcal{S}}}(x_k \,|\, x_{pa(k)}) = p_{\mathcal{S}}(x_k \,|\, x_{pa(k)}) \tag{3.1}$$

for any SEM $\tilde{\mathcal{S}}$ that is constructed from $\mathcal{S}$ by replacing the structural equation(s) for (some) X_j but not the one for X_k. Equation (3.1) shows that causal relationships are autonomous under interventions, it is therefore sometimes called "autonomy", but also "structural invariance" or "separability". Aldrich [1989] provides a brief overview of the historical development in economy. Interestingly, Aldrich [1989] argues that the "'most basic' question one can ask about a relation should be: How autonomous is it?" [Frisch et al., 1948, preface]. Other relevant references include work from Frisch's assistant Trygve Haavelmo [Haavelmo, 1944, Girshick and Haavelmo, 1947]. For a discussion and more references see also [Pearl, 2009, chapter 1.4]. Schölkopf et al. [2012] discusses the potential relevance of autonomy for machine learning.

3.1 Adjustment formulas

3.1.1 Truncated factorization, G-computation formula or manipulation theorem

We deduce a formula from (3.1) that became known under three different names: "truncated factorization" [Pearl, 1993a], "G-computation formula" [Robins, 1986] and "manipulation

theorem" [Spirtes et al., 1993]. Its importance stems from the fact that it allows us to compute statements about distributions that we have never seen data from.

Consider an SEM $\mathcal{S}$ with structural equations

$$X_j = f_j(X_{pa(j)}, N_j)$$

and density $p_{\mathcal{S}}$. Because of the Markov property we have

$$p_{\mathcal{S}}(x_1, \ldots, x_p) = \prod_{j=1}^{p} p_{\mathcal{S}}(x_j \,|\, x_{pa(j)})\,.$$

Now consider the SEM $\tilde{\mathcal{S}}$ which evolves from $\mathcal{S}$ after $do(X_k = \tilde{N}_k)$, where $\tilde{N}_k$ allows for the density $\tilde{p}$. Again, it follows from the Markov assumption that

$$p_{\mathcal{S},do(X_k=\tilde{N}_k)}(x_1, \ldots, x_p) = \prod_{j=1}^{p} p_{\mathcal{S},do(X_j=\tilde{N}_j)}(x_j \,|\, x_{pa(j)}) = \prod_{j\neq k} p_{\mathcal{S}}(x_j \,|\, x_{pa(j)})\tilde{p}(x_k)\,. \tag{3.2}$$

As a special case we obtain

$$p_{\mathcal{S},do(X_k=a)}(x_1, \ldots, x_p) = \begin{cases} \prod_{j\neq k} p_{\mathcal{S}}(x_j \,|\, x_{pa(j)}) & \text{if } x_k = a \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

It immediately follows that conditioning and intervening with $do\,()$ becomes equivalent for any variable that does not have any parents (w.l.o.g. let X_1 be such a source node):

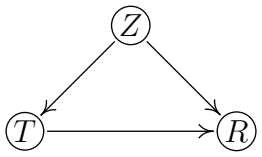
$$p_{\mathcal{S}}(x_2, \ldots, x_p \,|\, x_1 = a) = \frac{p(x_1 = a)\prod_{j=2}^{p} p_{\mathcal{S}}(x_j \,|\, x_{pa(j)})}{p(x_1 = a)} = p_{\mathcal{S},do(X_1=a)}(x_2, \ldots, x_p)\,. \tag{3.4}$$

In general, however, intervening and conditioning are usually two different things.

3.1.2 Invariances and adjusting

Equations (3.2) and (3.3) are widely applicable but sometimes a bit cumbersome to use. We will now learn about some practical alternatives. Therefore, we recall the kidney stone Example 1.1.3 that we will be able to generalize.

Example 3.1.1 [kidney stones, cont.] Assume that the true underlying SEM allows for the graph



Here, Z is the size of the stone ($0 \triangleq$ small, $1 \triangleq$ large), T the treatment and R the recovery (all binary). Consider further the two SEMs $\mathcal{S}_A$ and $\mathcal{S}_B$ that we obtain after replacing the structural equation for T with $T = A$ and $T = B$ respectively. Let us call the corresponding resulting probability distributions $\mathbb{P}_{\mathcal{S}_A}$ and $\mathbb{P}_{\mathcal{S}_B}$. Given that we are diagnosed with a kidney stone *without knowing its size*, we should base our choice of treatment on a comparison between

$$\mathbf{E}_{\mathcal{S}_A} R = \mathbb{P}_{\mathcal{S}_A}(R = 1) = \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = A))$$

and

$$\mathbf{E}_{\mathcal{S}_B} R = \mathbb{P}_{\mathcal{S}_B}(R = 1) = \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = B))\,.$$

Given that we have observed data from $\mathcal{S}$, how can we estimate these quantities? Consider the following computation

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}_A}(R = 1) &= \sum_{z=0}^{1} \mathbb{P}_{\mathcal{S}_A}(R = 1, T = A, Z = z) && (3.5)\\
&= \sum_{z=0}^{1} \mathbb{P}_{\mathcal{S}_A}(R = 1 \,|\, T = A, Z = z)\, \mathbb{P}_{\mathcal{S}_A}(T = A, Z = z) && (3.6)\\
&= \sum_{z=0}^{1} \mathbb{P}_{\mathcal{S}_A}(R = 1 \,|\, T = A, Z = z)\, \mathbb{P}_{\mathcal{S}_A}(Z = z) && (3.7)\\
&\overset{(3.1)}{=} \sum_{z=0}^{1} \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, T = A, Z = z)\, \mathbb{P}_{\mathcal{S}}(Z = z)\,. && (3.8)
\end{aligned}$$

The last step contains the key idea: again, we have made use of (3.1). We can estimate $\mathbb{P}_{\mathcal{S}_A}(R = 1)$ from the empirical data shown in Table 1.1 and obtain

$$\mathbb{P}_{\mathcal{S}_A}(R = 1) \approx 0.93 \times \frac{357}{700} + 0.73 \times \frac{343}{700} = 0.832\,.$$

It is important to realize that this is different from $\mathbb{P}_{\mathcal{S}}(R = 1 \,|\, T = 1) = 0.78$. Analogously, we obtain

$$\mathbb{P}_{\mathcal{S}_B}(R = 1) \approx 0.87 \times \frac{357}{700} + 0.69 \times \frac{343}{700} \approx 0.782\,,$$

and we conclude that we rather go for treatment A. (We have not checked whether there is a statistically significance difference between the treatments but from a decision theoretic point of view we do not need to do so.)

The deriviation above could also be seen as an implication from (3.3) but we will see in Proposition 3.1.4 that the idea of this alternative computation carries over to more complicated settings.

Definition 3.1.2 [valid adjustment set] Consider an SEM $\mathcal{S}$ over nodes $\mathbf{V}$ and let $Y \notin \mathbf{PA}_X$ (otherwise we have $p_{\mathcal{S},do(X=x)}(y) = p_{\mathcal{S}}(y)$). We call a set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ a **valid adjustment set** for the ordered pair (X, Y) if

$$p_{\mathcal{S},do(X=x)}(y) = \sum_{\mathbf{z}} p_{\mathcal{S}}(y \,|\, x, \mathbf{z})\, p_{\mathcal{S}}(\mathbf{z})\,. \tag{3.9}$$

Here, the sum (could also be an integral) is over the range of $\mathbf{Z}$, i.e., over all values $\mathbf{z}$ that $\mathbf{Z}$ can take.

In Example 3.1.1 above, $\mathbf{Z} = \{Z\}$ is a valid adjustment set for (T, R). We will now investigate which sets we can use for adjusting. We use the same idea as in Example 3.1.1 and write (for any set $\mathbf{Z}$)

$$\begin{aligned} p_{\mathcal{S},do(X=x)}(y) &= \sum_{\mathbf{z}} p_{\mathcal{S},do(X=x)}(y, \mathbf{z}) \\ &= \sum_{\mathbf{z}} p_{\mathcal{S},do(X=x)}(y \,|\, x, \mathbf{z})\, p_{\mathcal{S},do(X=x)}(\mathbf{z})\,. \end{aligned}$$

If these conditionals are invariant, i.e.,

$$p_{\mathcal{S},do(X=x)}(y \,|\, x, \mathbf{z}) = p_{\mathcal{S}}(y \,|\, x, \mathbf{z}) \;\text{ and }\; p_{\mathcal{S},do(X=x)}(\mathbf{z}) = p_{\mathcal{S}}(\mathbf{z})\,, \tag{3.10}$$

we can deduce (as above) that $\mathbf{Z}$ is a valid adjustment set. We therefore address the question, which conditionals remain invariant under the intervention $do\,(X = x)$.

Remark 3.1.3 [Characterization of invariant conditionals] Consider an SEM $\mathcal{S}$ with structural equations

$$X_j = f_j(\mathbf{PA}_j, N_j)$$

and an intervention $do\,(X_k = x_k)$. Analogously to what is done in [Pearl, 2009, Chapter 3.2.2], for example, we can now construct a new SEM $\mathcal{S}^*$ that equals $\mathcal{S}$ but has one more variable I that indicates whether the intervention took place or not. More precisely, I is a parent of X_k and does not have any other neighbors. The corresponding structural equations are

$$\begin{aligned} I &= N_I \\ X_j &= f_j(\mathbf{PA}_j, N_j) \quad \text{for } j \neq k \\ X_k &= \begin{cases} f_k(\mathbf{PA}_k, N_k) & \text{if } I = 0 \\ x_k & \text{otherwise} \end{cases}, \end{aligned}$$

where $N_I \sim \mathrm{Ber}(0.5)$. Thus, $I = 0$ corresponds to the observational setting and $I = 1$ to the interventional setting. More precisely, using (3.4), we obtain

$$\begin{aligned} p_{\mathcal{S}^*}(x_1, \ldots, x_p \,|\, I = 0) &= p_{\mathcal{S}^*,do(I=0)}(x_1, \ldots, x_p) \\ &= p_{\mathcal{S}}(x_1, \ldots, x_p) \end{aligned}$$

and similarly

$$p_{\mathcal{S}^*}(x_1, \ldots, x_p \,|\, I = 1) = p_{\mathcal{S}, do(X_k = x_k)}(x_1, \ldots, x_p)\,. \tag{3.11}$$

Using the Markov condition for $\mathcal{S}^*$ it thus follows for variables A and a set of variables $\mathbf{B}$ that

$$\begin{aligned} A \text{ } d\text{-sep. } I \,|\, \mathbf{B} \quad \text{in } \mathcal{G}^* \quad &\Longrightarrow \quad p_{\mathcal{S}^*}(a \,|\, \mathbf{b}, I = 0) = p_{\mathcal{S}^*}(a \,|\, \mathbf{b}, I = 1) \\ &\Longrightarrow \quad p_{\mathcal{S}}(a \,|\, \mathbf{b}) = p_{\mathcal{S}, do(X_k = x_k)}(a \,|\, \mathbf{b})\,. \end{aligned}$$

We are now able to continue the argument from before. Equation (3.10) is satisfied for sets $\mathbf{Z}$, for which we have

$$Y \text{ } d\text{-sep.}_{\mathcal{G}^*} \, I \,|\, X, \mathbf{Z} \qquad \text{and} \qquad \mathbf{Z} \text{ } d\text{-sep.}_{\mathcal{G}^*} \, I\,.$$

The subscript $\mathcal{G}^*$ means that the d-separation statement is required to hold in $\mathcal{G}^*$. This immediately implies the first two statements of the following proposition.

Proposition 3.1.4 *(i)* **"parent adjustment"***:*

$$\mathbf{Z} := \mathbf{PA}_X$$

is a valid adjustment set for (X, Y) for any $Y \notin \{X, \mathbf{PA}_X\}$.

(ii) **"backdoor-criterion"***: Any* $\mathbf{Z}$ *with*

- $\mathbf{Z}$ *contains no descendant of X AND*
- $\mathbf{Z}$ *blocks all paths from X to Y entering X through the backdoor ($X \leftarrow \ldots$, see Figure 3.1)*

is a valid adjustment set for (X, Y) for any $Y \notin \{X, \mathbf{PA}_X\}$.

(iii) **"towards necessity"***: Any* $\mathbf{Z}$ *with*

- $\mathbf{Z}$ *contains no descendant of any node on a directed path from X to Y (except for descendants of X that are not on a directed path from X to Y) AND*
- $\mathbf{Z}$ *blocks all non-directed paths from X to Y*

is a valid adjustment set for (X, Y).

Only the third statement [Shpitser et al., 2010] requires some explanation: we can add any node Z_0 to a valid adjustment set that satisfies $Z_0 \perp\!\!\!\perp Y \,|\, X$ because then

$$\begin{aligned} \sum_{\mathbf{z}, z_0} p(y \,|\, x, \mathbf{z}, z_0) p(\mathbf{z}, z_0) &= \sum_{\mathbf{z}} p(y \,|\, x, \mathbf{z}) \sum_{z_0} p(\mathbf{z}, z_0) \\ &= \sum_{\mathbf{z}} p(y \,|\, x, \mathbf{z}) p(\mathbf{z})\,. \end{aligned}$$

In fact, *all* valid adjustment sets can be characterized by Proposition 3.1.4 (iii) [Shpitser et al., 2010].

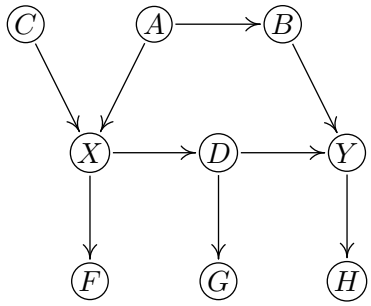


Figure 3.1: Only the path $X \leftarrow A \rightarrow B \rightarrow Y$ is a "backdoor path" from X to Y.

Example 3.1.5 [Adjustment in linear Gaussian systems] Consider an SEM $\mathcal{S}$ over variables $\mathbf{V}$ with $\{X, Y\}, \mathbf{Z} \subseteq \mathbf{V}$. Sometimes, we want to summarize a causal effect from X to Y by a single real number instead of looking at $p_{\mathcal{S},do(X=x)}(y)$ for all x. As a first approximation we may look at the expectation of this distribution and then take the derivative with respect to x (this works whenever X is continuous):

$$\frac{\partial}{\partial x} \mathbf{E}_{\mathcal{S},do(X=x)} Y \,.$$

In general, this is still a function of x. In linear Gaussian systems, however, this function turns out to be constant. Assume that $\mathbf{Z}$ is a valid adjustment set for (X, Y). The Gaussian distribution of $\mathbf{V}$ implies that $Y \,|\, X, \mathbf{Z}$ follows a Gaussian distribution, too; its mean is

$$aX + \mathbf{b}^t \mathbf{Z}$$

for some a and $\mathbf{b}$. If there is exactly one directed path from X to Y, then a equals the product of the path coefficients. If there is no directed path, then $a = 0$ and if there are different paths, a can be computed using the Wright's formula [Wright, 1921b]. It follows from (3.9) that

$$\frac{\partial}{\partial x} \mathbf{E}_{\mathcal{S},do(X=x)} Y \;=\; a \,. \tag{3.12}$$

Remark 3.1.6 It is not the case that all sets are valid adjustment sets. Therefore, it is not always a good idea to adjust for as many variables as possible, for example, cf. Berkson's paradox [Berkson, 1946].

Example 3.1.7 [Simpson's Paradox] Example 1.1.3 on page 7 is well-known for the following reason: we have

$$\begin{aligned} \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, T = A) < \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, T = B) \qquad &\text{but} \\ \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = A)) > \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = B))\,, \end{aligned} \tag{3.13}$$

see Example 3.1.1. Suppose that we have not measured the confounder Z (size of the stone) and furthermore that we do not even know about its existence. We might then

hypothesize that $T \to R$ is the correct graph. If we denote this (wrong) SEM by $\tilde{\mathcal{S}}$, we can rewrite (3.13) as

$$\begin{aligned} &\mathbb{P}_{\tilde{\mathcal{S}}}(R = 1 \,|\, do\,(T = A)) < \mathbb{P}_{\tilde{\mathcal{S}}}(R = 1 \,|\, do\,(T = B)) \qquad \text{but} \\ &\mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = A)) > \mathbb{P}_{\mathcal{S}}(R = 1 \,|\, do\,(T = B))\,. \end{aligned} \tag{3.14}$$

Due to the model misspecification, the causal inference statement gets reversed! Although A is the more effective drug, we propose to use B. What happens if there is yet another confounder that we did not correct for? If we are unlucky, it could be that we have to reverse the conclusion once more if we include this variable. In principle, this could lead to an arbitrarily long sequence of reversed causal conclusions (see Exercises).

This means that we have to be really careful when writing down the underlying graph. In some situations, we know the DAG from the protocol how the data have been recorded. If the medical doctors assigning the treatments, for example, did not have any knowledge about the patient other than the size of the kidney stone, there cannot be any other confounder than the size of the stone. Recent work investigates, whether we can check for confounders if we are willing to make further assumptions on the data generating process [e.g. Janzing et al., 2009, Sgouritsa et al., 2013].

Summarizing, the Simpson's paradox is not so much of a paradox but rather an example of how sensitive causal analysis could be with respect to model misspecifications.

3.2 Alternative identification of interventional distributions

Again, consider an SEM over variables $\mathbf{V}$. Sometimes, we can compute interventional distributions $p_{\mathcal{S},do(X=x)}$ in other ways than the adjustment formula (3.9). Let us therefore call an interventional distribution $p_{\mathcal{S},do(X=x)}(y)$ *identifiable* if it can be computed from the observational distribution and the graph structure. If there is a valid adjustment set for (X, Y), for example, $p_{\mathcal{S},do(X=x)}(y)$ is certainly identifiable. Judea Pearl has developed the so-called *do*-calculus that consists of three rules [Pearl, 2009, Theorem 3.4.1]. Given a graph $\mathcal{G}$ and disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$, we have

1. "Insertion/deletion of observations":

   $$p_{\mathcal{S},do(\mathbf{X}=\mathbf{x})}(\mathbf{y} \,|\, \mathbf{z}, \mathbf{w}) = p_{\mathcal{S},do(\mathbf{X}=\mathbf{x})}(\mathbf{y} \,|\, \mathbf{w})$$

   if $\mathbf{Y}$ d-separates $\mathbf{Z}$ given $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ have been removed.

2. "Action/observation exchange":

   $$p_{\mathcal{S},do(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}(\mathbf{y} \,|\, \mathbf{w}) = p_{\mathcal{S},do(\mathbf{X}=\mathbf{x})}(\mathbf{y} \,|\, \mathbf{z}, \mathbf{w})$$

   if $\mathbf{Y}$ d-separates $\mathbf{Z}$ given $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ and outgoing edges from $\mathbf{Z}$ have been removed.

3. "Insertion/deletion of actions":

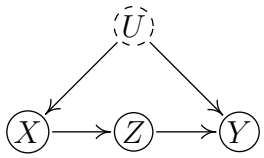
$$p_{\mathcal{S},do(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}(\mathbf{y} \mid \mathbf{w}) = p_{\mathcal{S},do(\mathbf{X}=\mathbf{x})}(\mathbf{y} \mid \mathbf{w})$$

if $\mathbf{Y}$ d-separates $\mathbf{Z}$ given $\mathbf{X}, \mathbf{W}$ in a graph where incoming edges in $\mathbf{X}$ and $\mathbf{Z}(\mathbf{W})$ have been removed. Here, $\mathbf{Z}(\mathbf{W})$ is the subset of nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$ in a graph that is obtained from $\mathcal{G}$ after removing all edges into $\mathbf{X}$.

Theorem 3.2.1 *The following statements can be proved*

- *The rules are complete [Huang and Valtorta, 2006, Shpitser and Pearl, 2006], that is all identifiable intervention distributions can be computed by an iterative application of these three rules.*
- *In fact, there is an algorithm, proposed by Tian [2002] that is guaranteed [Huang and Valtorta, 2006, Shpitser and Pearl, 2006] to find all identifiable interventional distributions.*

Example 3.2.2 [Front-door adjustment] Let $\mathcal{S}$ be an SEM with corresponding graph

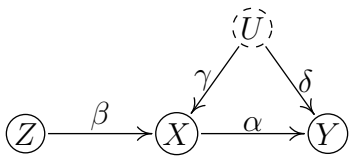


If we do not observe U, we cannot apply the backdoor criterion. In fact, there is no valid adjustment set. But still, provided that $p_{\mathcal{S}}(x, z) > 0$, the *do*-calculus provides us with

$$p_{\mathcal{S},do(X=x)}(y) \;=\; \sum_{z} p_{\mathcal{S}}(z \mid x) \sum_{\tilde{x}} p_{\mathcal{S}}(y \mid \tilde{x}, z)\, p_{\mathcal{S}}(\tilde{x})\,. \tag{3.15}$$

3.3 Instrumental variables

Instrumental variables date back to the 1920s [Wright, 1928] and are widely used in practice [e.g. Imbens and Angrist, 1994, Bowden and Turkington, 1990]. Although there exist numerous extensions and alternative methods, here, we focus on the essential idea. Consider a linear Gaussian SEM with the following corresponding graph



Here, the coefficient α is the quantity of interest (see Example 3.1.5) but not directly accessible because of the hidden common cause U. Because (U, N_X) is independent of Z, we can regard $\gamma U + N_X$ in

$$X = \beta Z + \gamma U + N_X$$

as noise. It becomes apparent that we can therefore consistently estimate the coefficient β and therefore have access to βZ. From

$$Y = \alpha X + \delta U + N_Y = \alpha\beta Z + (\alpha\gamma + \delta)U + N_Y$$

it is clear that we can then consistently estimate α. Thus, we first regress X on Z and then regress Y on the predicted values of X (predicted from the first regression). This method is commonly referred to as "two-stage-least-squares". It makes heavy use of the following assumptions

- linear SEMs,
- non-zero β (in the case of small or vanishing β, Z is often called a "weak instrument"),
- the independence between U and Z, and
- the absence of a direct influence from Z to Y.

3.4 Exercises

Exercise 3.4.1 *Prove the backdoor criterion Proposition 3.1.4 (ii).*

Exercise 3.4.2 *Prove the frontdoor criterion* (3.15) *starting with*

$$p_{\mathcal{S},do(X=x)}(y) = \sum_z p_{\mathcal{S},do(X=x)}(y \mid z, x) p_{\mathcal{S},do(X=x)}(z)$$

and then using rules 2 and 3 from the do-calculus.

Chapter 4

Causal structure learning

In this chapter, we first state some known identifiability results and then briefly introduce causal discovery methods (e.g. independence-based and score-based methods).

4.1 Structure identifiability

We have seen in Proposition 2.5.2 that any distribution could have been generated from many SEMs with different graphs. We therefore require further assumptions in order to obtain identifiability results. We discuss some of those assumptions in the following subsections.

4.1.1 Faithfulness

If the distribution $\mathbb{P}^{\mathbf{X}}$ is Markov and faithful with respect to the underlying DAG $\mathcal{G}^0$, we have a one-to-one correspondence between d-separation statements in the graph $\mathcal{G}^0$ and the corresponding conditional independence statements in the distribution. All graphs outside the correct Markov equivalence class of $\mathcal{G}^0$ can therefore be rejected because they impose conditional independences that do not hold in $\mathbb{P}^{\mathbf{X}}$. Since both the Markov condition and faithfulness put restrictions *only* on the conditional independences in the joint distribution, it is also clear that we are not able to distinguish between two Markov equivalent graphs, i.e. between two graphs that entail exactly the same set of (conditional) independences (see for example Figure 2.2 on page 28). More precisely, the Markov equivalence class of $\mathcal{G}^0$, represented by $\mathrm{CPDAG}(\mathcal{G}^0)$ is identifiable from $\mathbb{P}^{\mathbf{X}}$.

Lemma 4.1.1 *Assume that $\mathbb{P}^{\mathbf{X}}$ is Markov and faithful with respect to $\mathcal{G}^0$. Then, for each graph $\mathcal{G} \in CPDAG(\mathcal{G}^0)$, we find an SEM that generates the distribution $\mathbb{P}^{\mathbf{X}}$. Furthermore, the distribution $\mathbb{P}^{\mathbf{X}}$ is not Markov and faithful to any graph $\mathcal{G} \notin CPDAG(\mathcal{G}^0)$.*

Proof. The first statement follows directly from Proposition 2.5.2 and the second statement is a reformulation of Definition 2.4.4. □

The key idea of independence- (or constraint-)based methods (Section 4.2) is to assume faithfulness and then to estimate the correct Markov equivalence class of graphs.

4.1.2 Additive noise models

Proposition 2.5.2 shows that any distribution could have been generated from many SEMs with different graphs. For many distributions, however, the functions f_j appearing in the proof are rather complicated. It turns out that we can obtain identifiability results if we do not allow for arbitrary complex functions, i.e. if we restrict the function class. In the following subsections 4.1.3 and 4.1.4 we will assume that the noise acts in an additive way.

Definition 4.1.2 [Additive Noise Model] We call an SEM $\mathcal{S}$ an Additive Noise Model if the structural equations are of the form

$$X_j = f_j(\mathbf{PA}_j) + N_j\,, \tag{4.1}$$

that is, if the noise acts additively. For simplicity, let us further assume that the functions f_j are continuous and the noise variables N_j have a strictly positive density.

For these models causal minimality (Section 2.4.2) reduces to the condition that each function f_j is not constant in any of its arguments:

Proposition 4.1.3 *Consider a distribution generated by a model* (4.1) *and assume that the functions f_j are not constant in any of its arguments, i.e., for all j and $i \in \mathbf{PA}_j$ there are some $x_{\mathbf{PA}_j\setminus\{i\}}$ and some $x_i \neq x_i'$ such that*

$$f_j(x_{\mathbf{PA}_j\setminus\{i\}}, x_i) \neq f_j(x_{\mathbf{PA}_j\setminus\{i\}}, x_i')\,.$$

Then the joint distribution satisfies causal minimality with respect to the corresponding graph. Conversely, if there is a j and i such that $f_j(x_{\mathbf{PA}_j\setminus\{i\}}, \cdot)$ is constant, causal minimality is violated.

Proof. See Appendix A.4.1 □

Some of the following results assume causal minimality. This seems a plausible assumption since we will in general not be able to detect whether a variable depends on another variable in a constant way. Intuitively, we require that a function really "depends" on its arguments.

Given the restricted class of SEMs described in (4.1), what can we say about identifiability? Again, the answer is negative because the linear Gaussian SEMs, for example, is not identifiable, see Example 4.1.5 and Exercise 4.5.2. It turns out, however, that this case is exceptional in the following sense. For almost all other combinations of functions and distributions, we obtain identifiability. All the nonidentifiable cases have been characterized [Zhang and Hyvärinen, 2009, Peters et al., 2014]. Another non-identifiable example different from the linear Gaussian case is shown in the right plot in Figure 4.1. Its details can be found in Example 25 in [Peters et al., 2014]. Table 4.1.2 shows some of the known identifiability results.

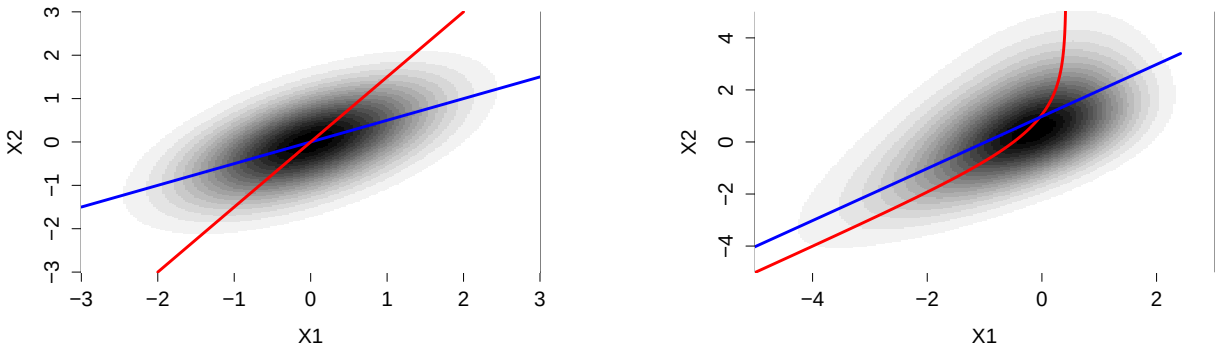


Figure 4.1: Joint density over X_1 and X_2 for two non-identifiable examples. The left panel shows Example 4.1.5 (linear Gaussian case) and the right panel shows a slightly more complicated example, with "fine-tuned" parameters for function, input and noise distribution (the latter plot is based on kernel density estimation). The blue function corresponds to the forward model $X_2 = f_2(X_1) + N_2$, the red function to the backward model $X_1 = \tilde{f}_1(X_2) + \tilde{N}_1$.

type of structural equation		conditions	DAG identif.	see
general SEM:	$X_i = f_i(X_{\mathbf{PA}_i}, N_i)$	-	✗	Prop. 2.5.2
additive noise model:	$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$	nonlin. fct.	✓	Thm 4.1.9(i)
causal additive model:	$X_i = \sum_{k \in \mathbf{PA}_i} f_{ik}(X_k) + N_i$	nonlin. fct.	✓	Thm 4.1.9(ii)
linear Gaussian:	$X_i = \sum_{k \in \mathbf{PA}_i} \beta_{ik} X_k + N_i$	linear fct.	✗	Exerc. 4.5.2

Table 4.1: Summary of some known identifiability results for Gaussian noise

Remark 4.1.4 There have been several extensions to the framework of additive noise models (4.1). For example, Zhang and Hyvärinen [2009] allow for a post-nonlinear transformation of the variables. Peters et al. [2011] consider additive noise models for discrete variables. Janzing et al. [2009] investigate what happens if there exists a hidden common cause.

In the following two subsections, we will look at two specific identifiable examples in more detail: the linear non-Gaussian case (Section 4.1.3) and the nonlinear Gaussian case (Section 4.1.4). Although more general results are available [Peters et al., 2014], we concentrate on those two examples because for them, precise conditions can be stated easily.

4.1.3 Linear non-Gaussian acyclic models

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:

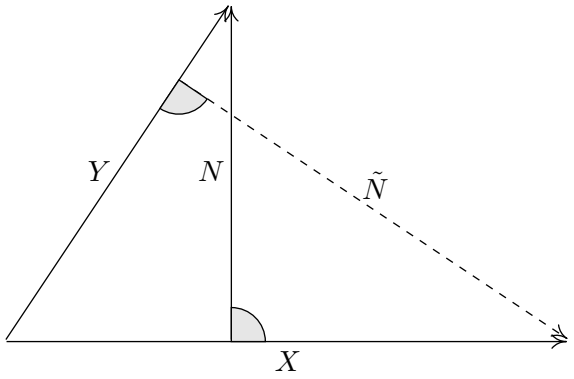
Example 4.1.5

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

where X and N are normally distributed with mean zero. It can be checked 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 $\tilde{N} = X - \tilde{\phi} Y$. The following figure depicts this example in $\mathcal{L}_2$, [e.g. Peters, 2008] with the dot product representing the covariance.



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

Proposition 4.1.6 *Let X and Y be two random variables, for which*

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

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 (Appendix A.4.2) is based on a characterization of the Gaussian distribution that was proved independently by Skitovič and Darmois [Skitovič, 1954, 1962, Darmois, 1953].

Theorem 4.1.7 *[Darmois-Skitovič] Let $X_1, \dots, X_d$ be independent, non-degenerate random variables. If there are non-vanishing coefficients $a_1, \dots, a_d$ and $b_1, \dots, b_d$ (that is, $a_i \neq 0 \neq b_i$ for all i) such that the two linear combinations*

$$l_1 = a_1 X_1 + \ldots + a_d X_d \,,$$
$$l_2 = b_1 X_1 + \ldots + b_d X_d$$

are independent, each X_i is normally distributed.

This result holds in the multivariate case, too. Shimizu et al. [2006] prove it using Independent Component Analysis (ICA) [Comon, 1994, Theorem 11], which itself is proved using the Darmois-Skitovič theorem.

Theorem 4.1.8 *[Shimizu et al. [2006]] Assume an SEM with graph $\mathcal{G}_0$*

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j \,, \qquad j = 1, \dots, p \tag{4.2}$$

where all N_j are jointly independent and non-Gaussian distributed with strictly positive density[1]. Additionally, for each $j \in \{1, \dots, p\}$ 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].

Interestingly, there is an alternative proof for Theorem 4.1.8: Theorem 28 in [Peters et al., 2014] extends bivariate identifiability results as Proposition 4.1.6 to the multivariate case. This trick will also be used for nonlinear additive models.

4.1.4 Nonlinear Gaussian additive noise models

We have seen that the graph structure of an additive noise model becomes identifiable if we assume the function to be linear and the noise to be non-Gaussian. Alternatively, we can exploit the nonlinearity of functions. The result is easiest to state with Gaussian noise:

Theorem 4.1.9 *(i) Let $\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X_1, \dots, X_p}$ be generated by an SEM with*

$$X_j = f_j(\mathbf{PA}_j) + N_j \,,$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable functions f_j that are not linear in any component: denote the parents $\mathbf{PA}_j$ of X_j by $X_{k_1}, \dots, X_{k_\ell}$, then the function $f_j(x_{k_1}, \dots, x_{k_{a-1}}, \cdot, x_{k_{a+1}}, \dots, x_{k_\ell})$ is assumed to be nonlinear for all a and some $x_{k_1}, \dots, x_{k_{a-1}}, x_{k_{a+1}}, \dots, x_{k_\ell} \in \mathbb{R}^{\ell-1}$.

[1] The condition of a strictly positive density was missing in the original version of this thesis. This condition is necessary although this might not be apparent on first sight of the original paper [Shimizu et al., 2006].

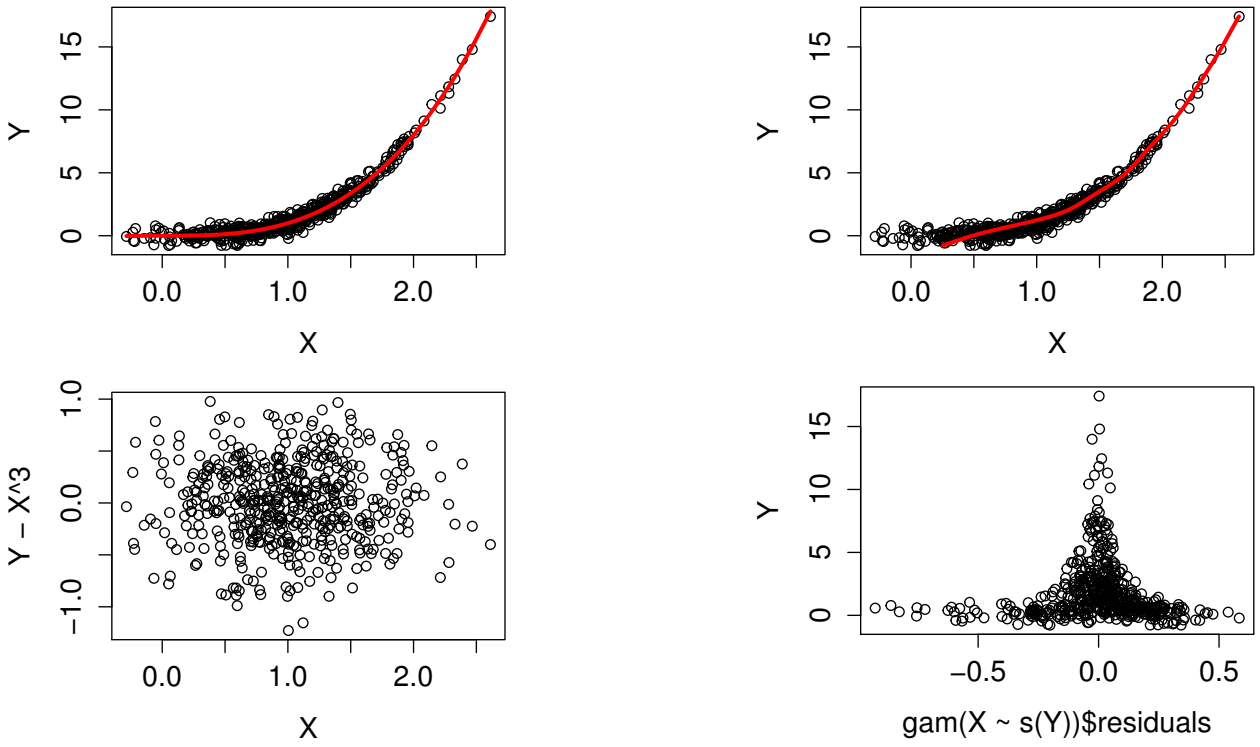


Figure 4.2: The data set contains i.i.d. data points from a distribution $\mathbb{P}^{(X,Y)}$ that has been generated from an additive noise model $Y = X^3 + N_Y$ with normally distributed noise N_Y. The left plots show the correct model and the independent residuals. Fitting a model in the backward direction $X = g(Y) + M_X$ leads to residuals that are dependent on the input (right hand side). (Here, regression is performed with `gam` from the R-package `mgcv` [Wood, 2011].) This corresponds to the identifiability proved in Theorem 4.1.9.

(ii) As a special case, let $\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X_1,\dots,X_p}$ be generated by an SEM with

$$X_j = \sum_{k \in \mathbf{PA}_j} f_{j,k}(X_k) + N_j \,, \tag{4.3}$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable, nonlinear functions $f_{j,k}$. This model is known as a causal additive model (CAM).

In both cases (i) and (ii), we can identify the corresponding graph $\mathcal{G}_0$ from the distribution $\mathbb{P}^{\mathbf{X}}$. The statements remain true if the noise distributions for source nodes, i.e., nodes with no parents, are allowed to have a non-Gaussian density with full support on the real line $\mathbb{R}$ (the proof remains identical).

The proof is omitted. The statement can be found as Corollary 31 in [Peters et al., 2014].

4.1.5 Modularity and Independence of cause and mechanism (bivariate case)

For two variables the difficulty of causal discovery can be seen from the following symmetric equation

$$p(x_2 \,|\, x_1)p(x_1) = p(x_1 \,|\, x_2)p(x_2)\,, \tag{4.4}$$

where the left (or right) hand side corresponds to the Markov factorization of $p(x_1, x_2)$ if the distribution is Markov w.r.t $X_1 \to X_2$ (or $X_2 \to X_1$).

Modularity [Pearl, 2009, and references therein] or autonomy [Haavelmo, 1944, Aldrich, 1989] describe the assumption that changing one of the structural equations leaves the other structural equations invariant, see the invariance principle described in Section 3.1.2. This leads to an asymmetry in Equation (4.4): intervening on the cause C changes its distribution $p(c)$ but not the conditional distribution $p(e \,|\, c)$ of the effect E given cause C. Intervening on E, however, is expected to change both $p(e)$ and $p(c \,|\, e)$. Hoover [1990] uses this for identification of cause and effects in economics.

Another related way to break the symmetry in (4.4) is by assuming that $p(e \,|\, c)$ is in some sense "independent" of $p(c)$. The hope is that this "independence" will not hold between $p(c \,|\, e)$ and $p(e)$.

Different formalizations of this idea, in particular formalizations of "independence", are given by Janzing et al. [2012], Sgouritsa et al. [2015], Zscheischler et al. [2011].

4.2 Independence-based methods

Independence-based methods assume that the distribution is faithful to the underlying DAG and therefore estimate the underlying CPDAG from conditional independences in $\mathbb{P}^{\mathbf{X}}$.

Estimation of skeleton Most methods first concentrate on estimating the skeleton and only later try to orient as many edges as possible. For the skeleton search it is useful to know that

Lemma 4.2.1 *(i) Two nodes X, Y in a DAG $(\mathbf{X}, \mathcal{E})$ are adjacent if and only if they cannot be d-separated by any subset $S \subseteq \mathbf{V} \setminus \{X, Y\}$.*

(ii) If two nodes X, Y in a DAG $(\mathbf{X}, \mathcal{E})$ are not adjacent, then they are d-separated by either $\mathbf{PA}_X$ or $\mathbf{PA}_Y$.

Using Lemma 4.2.1(i), we have that if two variables are always dependent, no matter what other variables one conditions on, these two variables must be adjacent. This reasoning is used in the **IC algorithm** (Inductive Causation) [Pearl, 2009] or in the **SGS algorithm** (after its inventors Spirtes, Glymour and Scheines) [Spirtes et al., 2000]; it is an example of

how properties of the joint distribution can help to infer parts of the graph structure. The algorithm starts with a completely undirected graph. Then for each pair of variables, all possible conditioning sets are examined in turn. For each such triple (of two variables and one conditioning set) we test the null hypothesis of conditional independence. If we accept the null, the edge between the two variables is removed. If we have an "oracle" test (in the sense that true null hypotheses are always accepted and false nulls always rejected by the test), then the output will be the skeleton of the true DAG and hence the skeleton of all DAGs in the equivalence class of the true DAG (since all members of an equivalence class share the same skeleton). As the test is not perfect in practice, we just have an estimate of the sketon (which can be shown to be pointwise consistent in the sense that the probability of it being identical to the true skeleton converges to 1 as $n \to \infty$ for any fixed graph.) Uniform consistency is impossible to achieve without stronger assumptions.

The **PC algorithm** (after its inventors Peter and Clark) [Spirtes et al., 2000] tries to avoid conditioning on all possible subsets and therefore improve the computation time. It starts by testing all pairs of variables with the empty conditioning set (just testing independence), removing any edge of the nitially as well fully connected undirected graph if the null of independence is accepted by the test. Thereafter it tries to condition on sets of size 1 but just for edges that have not been removed already. Then for sets of size 2 and so on. The graph gets sparse quite quickly and so many tests do not have to performed. From a statistical perspective, a further advantage is that we start with conditioning on small subsets for which conditional independence tests are more powerful and reliable (although all conditional independence tests have their own pitfalls).

Orientation of edges According to Lemma 2.4.5, we might be able to orient the immoralities (or v-structures) in the graph. If two nodes are not directly connected in the obtained skeleton, there must be a set that d-separates these nodes. Suppose that the skeleton contains the structure $X - Y - Z$ with no direct edge between X and Z; let further S denote the corresponding d-separation set S. The structure $X - Y - Z$ is an immorality and can therefore be oriented as $X \to Y \leftarrow Z$ if and only if $Y \notin S$. After the orientation of immoralities, we may be able to orient some further edges in order to avoid cycles, for example. One set of such orientation rules has been shown to be complete and is known as Meek's orientation rules [Meek, 1995].

Conditional independence tests In the two preceding paragraphs we have assumed the existence of an independence oracle that tells us whether a specific (conditional) independence is or is not present in the distribution. In practice, however, we have to infer this statement from a finite amount of data. There is some recent work on kernel-based tests [Fukumizu et al., 2008, Tillman et al., 2010, Zhang et al., 2011] but in general, conditional independence tests are difficult to perform in practice [e.g. Bergsma, 2004] if one does not restrict the variables to follow a Gaussian distribution, for example. In the latter case, we can test for vanishing partial correlation, see Section 1.2 for more details.

In the multivariate Gaussian setting, a test for conditional independence is equivalent to

testing for zero partial correlation:

$$H_0 : \rho_{X,Y|\mathbf{S}} = 0 \text{ versus } H_a : \rho_{X,Y|\mathbf{S}} \neq 0.$$

Partial correlations can be computed via regression, inversion of parts of the covariance matrix, or a recursive formula For testing, it is helpful to use Fisher's Z-transform:

$$\hat{z}_{X,Y|\mathbf{S}} = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{X,Y|\mathbf{S}}}{1 - \hat{\rho}_{X,Y|\mathbf{S}}} \right).$$

Under H_0, $\sqrt{n - |\mathbf{S}| - 3}\hat{z}_{X,Y|\mathbf{S}} \sim N(0,1)$. Hence, we reject H_0 versus H_a if $\sqrt{n - |\mathbf{S}| - 3}|\hat{z}_{X,Y|\mathbf{S}}| > \Phi^{-1}(1 - \alpha/2)$ The significance level α serves as a tuning parameter for the PC algorithm and determines the sparsity of the graph.

The partial correlation $\hat{\rho}_{X,Y|\mathbf{S}}$ can be computed in different ways. One option is to regress X on all variables in S and store the residuals in R_x. Then regress Y on S and store the residuals in R_y. The empirical partial correlation is then the standard correlation $\hat{\rho}(R_x, R_y)$ between the residuals. Another option is to invert the empirical correlation or covariance matrix $\hat{\Sigma}_{Z,Z}$ where $Z = (X, Y, S)$. Let $\hat{K} = (\hat{\Sigma}_{Z,Z})^{-1}$. Then

$$\hat{\rho}_{X,Y|\mathbf{S}} = -\frac{\hat{K}_{X,Y}}{\sqrt{\hat{K}_{X,X}\hat{K}_{Y,Y}}}$$

and we can use the test based on the z-transform to test the null hypothesis.

4.3 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, Cooper et al., 1997, Chickering, 2002] and references therein.

Best scoring graph 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.

$$\hat{\mathcal{G}} := \underset{\mathcal{G} \text{ DAG over } \mathbf{X}}{\operatorname{argmax}} S(\mathcal{D}, \mathcal{G}) \tag{4.5}$$

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 $\theta \in \mathbf{\Theta}$.

(Penalized) likelihood 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)

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

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 (4.5): 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.

Bayesian formalization 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):

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

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

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

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, and vice versa. It can be shown (see Exercise 4.5.1) 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.

Exact Methods There is a lot of interesting research that tries to scale up exact methods. Here, "exact" means that they aim at finding (one of) the best scoring graphs for a given finite data sets. Greedy search techniques are often heuristic and have guarantees only in the limit of infinite data.

In the Bayesian setting, Koivisto and Sood [2004], Koivisto [2006] compute marginal probabilities over edges.

The integer linear programming framework (probably added later) is studied by [De Campos and Ji, 2011, Cussens, 2011, Studený and Haws, 2014, Jaakkola et al., 2010, Sheehan et al., 2014, and others].

For a dynamic programming approach consider the work by [Silander and Myllymak, 2006, and references therein].

4.4 Data from different environments (not only observational data)

We now assume that we observe data from different environments $e \in \mathcal{E}$. We model this with

$$\mathbf{X}^e \sim \mathbb{P}^e \,,$$

where each variable X_j^e for different e denotes the same (physical) quantity, measured in different environments. We will talk about a variable X in different environments, which is a slight abuse of notation. From each of the environments, we assume to observe n^e i.i.d. samples.

Known intervention targets A first type of methods assumes that the different environments are generated from different interventional settings. In the case that the intervention targets $\mathcal{I}^e \subseteq \{1, \ldots, p\}$ are known, several methods have been proposed. Assuming faithfulness and a specific type of intervention, Tian and Pearl [2001], Hauser and Bühlmann [2012] define and characterize the interventional equivalence classes of graphs; that is the class of graphs that can explain the observed distributions. Eberhardt et al. [2005] investigate how many intervention experiments are necessary (in the worst case) in order to identify the graph.

Unknown intervention targets Let us now consider a slightly different setting. Instead of learning the whole causal structure, we may consider a target variable Y and try to learn its causal parents. That is, we have

$$(\mathbf{X}^e, Y^e) \sim \mathbb{P}^e \,.$$

for $e \in \mathcal{E}$. We may then assume that there is a set $\mathbf{PA}_Y$ such that the conditional

$$\mathbb{P}^{Y^e \,|\, \mathbf{PA}_Y^e} = \mathbb{P}^{Y^f \,|\, \mathbf{PA}_Y^f} \,,$$

for all $e, f \in \mathcal{E}$. This assumption is satisfied if the distributions are generated by an underlying SEM and the different environments correspond to different intervention distributions, for which Y has not been intervened on [Peters et al., 2015]. Having said that, the assumption is more general and does not require an underlying SEM. One can consider the collection $\mathcal{A}$ of *all* sets $\mathbf{A}$ of variables that lead to "invariant prediction", i.e., we have

$$\mathbb{P}^{Y^e \,|\, \mathbf{A}^e} = \mathbb{P}^{Y^f \,|\, \mathbf{A}^f} \,,$$

for all $e, f \in \mathcal{E}$ and for all $\mathbf{A} \in \mathcal{A}$. It is not difficult to see (Exercise 4.5.3) that the variables appearing in all those sets must be direct causes of Y:

$$\bigcap_{\mathbf{A} \in \mathcal{A}} \mathbf{A} \;\subseteq\; \mathbf{PA}_Y \,. \tag{4.6}$$

In the case of SEMs and interventions, it is further possible to write down sufficient conditions for the identifiability of the set of direct causes [Peters et al., 2015].

Tian and Pearl [2001] also address the question of identifiability with unknown intervention targets. They do not specify a target variable and focus on changes in marginal distributions rather than conditionals.

4.5 Exercises

Exercise 4.5.1 *Prove that for linear Gaussian SEMs, two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are distribution equivalent if and only if they are Markov equivalent.*

Exercise 4.5.2 *Consider a distribution $\mathbb{P}^{\mathbf{X}}$ that has been generated from a linear Gaussian SEM $\mathcal{S}$. Prove that for any DAG $\mathcal{G}$ such that $\mathbb{P}^{\mathbf{X}}$ is Markov w.r.t. $\mathcal{G}$ there is a corresponding SEM $\mathcal{S}_{\mathcal{G}}$ generating $\mathbb{P}^{\mathbf{X}}$.*

Exercise 4.5.3 *Prove Equation* (4.6).

Appendix A

Proofs

A.1 Proofs from Chapter 1

A.2 Proofs from Chapter 2

A.2.1 Proof of Proposition 2.2.4

Proof. In order to simplify notation we write X_1 instead of X and X_2 instead of Y. First, the truncated factorization formula (3.3) implies

$$\begin{aligned} p_{\mathcal{S}}^{X_2 \,|\, do(X_1 = x_1)}(x_2) &= \int \prod_{j \neq 1} p_j(x_j \,|\, x_{pa(j)})\, dx_3 \cdots dx_p \\ &= \int \prod_{j \neq 1} p_j(x_j \,|\, x_{pa(j)}) \frac{\tilde{p}(x_1)}{\tilde{p}(x_1)}\, dx_3 \cdots dx_p \\ &= p_{\mathcal{S}}^{X_2 \,|\, X_1 = x_1, do\left(X_1 = \tilde{N}_1\right)}(x_2) \end{aligned} \tag{A.1}$$

if $\tilde{N}_1$ puts positive mass on x_1, i.e., $\tilde{p}(x_1) > 0$. The other statement that we need is

$$X_2 \not\perp\!\!\!\perp X_1 \text{ in } \mathbb{Q} \iff \exists x_1^{\triangle}, x_1^{\square} \text{ with } q(x_1^{\triangle}), q(x_1^{\square}) > 0 \text{ and } \mathbb{Q}^{X_2 \,|\, X_1 = x_1^{\triangle}} \neq \mathbb{Q}^{X_2 \,|\, X_1 = x_1^{\square}} \tag{A.2}$$

and

$$X_2 \not\perp\!\!\!\perp X_1 \text{ in } \mathbb{Q} \iff \exists x_1^{\triangle} \text{ with } q(x_1^{\triangle}) > 0 \text{ and } \mathbb{Q}^{X_2 \,|\, X_1 = x_1^{\triangle}} \neq \mathbb{Q}^{X_2}\,. \tag{A.3}$$

We then have for any $\hat{N}_1$ with full support

$(i) \overset{\text{(A.2)}}{\Longrightarrow} \exists x_1^{\triangle}, x_1^{\square}$ with pos. density under $\tilde{N}_1$ s.t. $\mathbb{P}_{\mathcal{S}}^{X_2 \,|\, X_1 = x_1^{\triangle}, do\left(X_1 = \tilde{N}_1\right)} \neq \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, X_1 = x_1^{\square}, do\left(X_1 = \tilde{N}_1\right)}$

$\overset{\text{(A.1)}}{\Longrightarrow} (ii)$

$\overset{\text{(A.1)}}{\Longrightarrow} \exists x_1^{\triangle}, x_1^{\square}$ with pos. density under $\hat{N}_1$ s.t. $\mathbb{P}_{\mathcal{S}}^{X_2 \,|\, X_1 = x_1^{\triangle}, do\left(X_1 = \hat{N}_1\right)} \neq \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, X_1 = x_1^{\square}, do\left(X_1 = \hat{N}_1\right)}$

$\overset{\text{(A.2)}}{\Longrightarrow} (iv)$

$\overset{\text{(trivial)}}{\Longrightarrow} (i)$

We further have that $(ii) \overset{\text{(trivial)}}{\Longrightarrow} (iii)$ and that $\mathbb{P}_{\mathcal{S}}^{X_2} = \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, do\left(X_1 = N_1^*\right)}$ with N_1^* having the distribution $\mathbb{P}_{\mathcal{S}}^{X_1}$. The latter implies

$$
\begin{aligned}
\neg(i) &\implies X_2 \perp\!\!\!\perp X_1 \text{ in } \mathbb{P}_{\mathcal{S}}^{\mathbf{X} \,|\, do\left(X_1 = N_1^*\right)} \\
&\overset{\text{(A.3)}}{\Longrightarrow} \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, X_1 = x^{\triangle} \,|\, do\left(X_1 = N_1^*\right)} = \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, do\left(X_1 = N_1^*\right)} \text{ for all } x^{\triangle} \text{ with } p_1(x^{\triangle}) > 0 \\
&\overset{\text{(A.1)}}{\Longrightarrow} \mathbb{P}_{\mathcal{S}}^{X_2 \,|\, do\left(X_1 = x^{\triangle}\right)} = \mathbb{P}_{\mathcal{S}}^{X_2} \text{ for all } x^{\triangle} \text{ with } p_1(x^{\triangle}) > 0 \\
&\overset{\neg(ii)}{\Longrightarrow} \neg(iii)
\end{aligned}
$$

□

A.2.2 Proof of Proposition 2.2.9

Proof. (i) follows directly from the Markov property of the interventional SEM: after removing the incoming edges into X, X and Y are d-separated if there is no direct path from X to Y.

(ii) can be proved by counter example: e.g.

$$
\begin{aligned}
X &= N_X \\
Z &= 2X + N_Z \\
Y &= 4X - 2Z + N_Y
\end{aligned}
$$

Because $Y = -2N_Z + N_Y$, we have $X \perp\!\!\!\perp Y$ for all N_X. □

A.2.3 Proof of Proposition 2.5.2

Proof. Let $N_1, \ldots, N_p$ 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.2.4 Proof of Theorem 2.4.2

Proof. proof sketch for equiv. of markov properties □

A.2.5 Proof of Proposition 2.4.13

Proof. "if": Assume that causal minimality is not satisfied. Then, there is an X_j and a $Y \in \mathbf{PA}_j^{\mathcal{G}}$, such that $\mathbb{P}^{\mathbf{X}}$ is also Markov with respect to the graph obtained when removing the edge $Y \to X_j$ from $\mathcal{G}$.

"only if": If $\mathbb{P}^{\mathbf{X}}$ has a density, the Markov condition is equivalent to the Markov factorization [Lauritzen, 1996, Theorem 3.27]. Assume that $Y \in \mathbf{PA}_j^{\mathcal{G}}$ and $X_j \perp\!\!\!\perp Y \,|\, \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}$. Then $P(\mathbf{X}) = P(X_j|\mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}) \prod_{k \neq j} P(X_k|\mathbf{PA}_k^{\mathcal{G}})$, which implies that $\mathbb{P}^{\mathbf{X}}$ is Markov w.r.t. $\mathcal{G}$ without $Y \to X_j$. □

A.3 Proofs from Chapter 3

A.4 Proofs from Chapter 4

A.4.1 Proof of Proposition 4.1.3

Proof. Assume causal minimality is not satisfied. We can then find a j and $i \in \mathbf{PA}_j$ with $X_j = f_j(X_{\mathbf{PA}_j \setminus \{i\}}, X_i) + N_j$ that does not depend on X_i if we condition on all other parents $\mathbf{PA}_j \setminus \{i\}$ (Proposition 2.4.13). Let us denote $\mathbf{PA}_j \setminus \{X_i\}$ by X_A. For the function f_j it follows that $f_j(x_A, x_i) = c_{x_A}$ for $\mathbb{P}^{X_A, X_i}$-almost all (x_A, x_i). Indeed, assume without loss of generality that $\mathbf{E}N_j = 0$, take the mean of $X_j \,|\, \mathbf{PA}_j^{\mathcal{G}_0} = (x_A, x_i)$ and use e.g. (2b) from Dawid [1979]. The continuity of f_j implies that f_j is constant in its last argument.

The converse statement follows from Proposition 2.4.13, too. □

A.4.2 Proof of Proposition 4.1.6

We first prove the following lemma, which should be clear intuitively.

Lemma A.1 *Let X and ϵ be two independent variables and assume ϵ to be non-deterministic. Then*

$$\epsilon \not\perp\!\!\!\perp (X + \epsilon)\,.$$

Proof. Of course the proof becomes trivial if the variables have finite variance. Then $\mathbf{cov}(X, X + \epsilon) = \mathbf{var}(X) > 0$. For the general case, however, the argumentation is a bit more complex. Assume $N \perp\!\!\!\perp (X + \epsilon)$. Then for every $u, v \in \mathbb{R}$:

$$\begin{aligned}
\varphi_{(\epsilon, X+\epsilon)}(u, v) &= \mathcal{E}\left[\exp(iu\epsilon + iv\epsilon + ivX)\right] \\
&= \mathcal{E}\left[\exp(iu\epsilon + iv\epsilon) \cdot \exp(ivX)\right] \\
&= \mathcal{E}\left[\exp(iu\epsilon + iv\epsilon)\right] \cdot \mathcal{E}\left[\exp(ivX)\right] \\
&= \varphi_\epsilon(u + v) \cdot \varphi_X(v)\,.
\end{aligned}$$

We also have

$$\begin{aligned}
\varphi_{(\epsilon, X+\epsilon)}(u, v) &= \mathcal{E}\left[\exp(iu\epsilon + iv\epsilon + ivX)\right] \\
&= \mathcal{E}\left[\exp(iu\epsilon) \cdot \exp(iv\epsilon + ivX)\right] \\
&= \mathcal{E}\left[\exp(iu\epsilon)\right] \cdot \mathcal{E}\left[\exp(iv\epsilon + ivX)\right] \\
&= \varphi_\epsilon(u) \cdot \varphi_{(\epsilon+X)}(v) \\
&= \varphi_\epsilon(u) \cdot \varphi_\epsilon(v) \cdot \varphi_X(v)\,.
\end{aligned}$$

We know that $\varphi_X(0) = 1$ and that characteristic functions are continuous. Thus there exists a non-empty open interval $V = (-r, r) \subset \mathbb{R}$, such that $|\varphi_X(v)| > 0 \; \forall v \in V$. Thus we have for all $u \in \mathbb{R}$ and $v \in V$:

$$\varphi_\epsilon(u + v) = \varphi_\epsilon(u) \cdot \varphi_\epsilon(v) \,.$$

Note that this is still true for an arbitrary $v \in \mathbb{R}$: Choose $n \in \mathbb{N}$, such that $\|v/n\| \leq r$. It follows

$$\begin{aligned}
\varphi_\epsilon(u + v) &= \varphi_\epsilon\left(u + (n-1)\frac{v}{n} + \frac{v}{n}\right) \\
&= \varphi_\epsilon\left(u + (n-1)\frac{v}{n}\right) \cdot \varphi_\epsilon\left(\frac{v}{n}\right) \\
&\vdots \\
&= \varphi_\epsilon(u) \cdot \varphi_\epsilon\left(\frac{v}{n}\right)^n = \varphi_\epsilon(u) \cdot \varphi_\epsilon(v)
\end{aligned}$$

Then we know

$$\varphi_\epsilon(u) = z^u \qquad \text{for some } z \in \setminus\{c \in :\text{Im } c = 0, \text{Re } c < 0\} \,.$$

We can write $z = \exp(a + ib)$ and since $\|\varphi_\epsilon\|_\infty \leq 1$ we deduce that $a = 0$. It follows

$$\varphi_\epsilon(u) = \exp(ib \cdot u) \,.$$

Because of the uniqueness of characteristic functions this implies $\mathbb{P}(\epsilon = b) = 1$ and ϵ is degenerate. □

Proof of Proposition 4.1.6 If X and N are Gaussian distributed, the statement follows from Example 4.1.5. Conversely, we assume that

$$\begin{aligned}
Y &= \phi X + N \\
\text{and} \quad \tilde{N} &= (1 - \phi\psi)X - \psi N
\end{aligned}$$

are independent. Distinguish between the following cases:

1. $(1 - \phi\psi) \neq 0$ and $\psi \neq 0$
   Here, Theorem 4.1.7 implies that X, N and thus also $Y, \tilde{N}$ are normally distributed.

2. $\psi = 0$
   We have $(1 - \phi\psi)X \perp\!\!\!\perp \phi X + N$. $\psi = 0$ implies
   $$X \perp\!\!\!\perp \phi X + N,$$
   which is a contradiction to Lemma A.1.

3. $(1 - \phi\psi) = 0$
   It follows $-\psi N \perp\!\!\!\perp \phi X + N$. Thus
   $$N \perp\!\!\!\perp \phi X + N$$
   and we can apply Lemma A.1 again.

□

Bibliography

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

W. P. Bergsma. *Testing Conditional Independence for Continuous Random Variables*, 2004. EURANDOM-report 2004-049.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*, 2:47–53, 1946.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, New York, USA, 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: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

R. J. Bowden and D. A. Turkington. *Instrumental Variables*. Econometric Society Monographs. Cambridge University Press, New York, USA, 1990.

C. R. Charig, D. R. Webb, S. R. Payne, and J. E. A. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clin Res Ed)*, 292:879–882, 1986.

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

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann.

G. Cooper, D. Heckerman, and C. Meek. A Bayesian approach to causal discovery. Technical report, Microsoft Research (MSR-TR-97-05), 1997.

J. Cussens. Bayesian network learning with cutting planes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 153–160, Corvallis, OR, USA, 2011. AUAI Press.

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

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

C. P. De Campos and Q. Ji. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689, 2011.

M. Druzdzel and H. Simon. Causality in Bayesian belief networks. In *In Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 3–11, San Francisco, CA, USA, 1993. Morgan Kaufmann.

M. J. Druzdzel and H. van Leijen. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 13:45–62, 2001.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, USA, 2002.

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

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

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 178–184, Corvallis, OR, USA, 2005. AUAI Press.

R. A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

R. Frisch, T. Haavelmo, T.C. Koopmans, and J. Tinbergen. *Autonomy of economic relations*. Series: Memorandum fra Universitets Socialøkonomiske Institutt. Universitets Socialøkonomiske Institutt, Oslo, 1948.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 235–243, San Francisco, CA, USA, 1994. Morgan Kaufmann.

D. Geradin and I. Girgenson. The counterfactual method in EU competition law: The cornerstone of the effects-based approach. Available at SSRN: `http://ssrn.com/abstract=1970917`, 2011.

M. A. Girshick and T. Haavelmo. Statistical analysis of the demand for food: Examples of simultaneous estimation of structural equations. *Econometrica*, 2:79–110, 1947.

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, Cambridge, MA, USA, 2008. MIT Press.

J. Gwiazda, E. Ong, R. Held, and F. Thorn. Vision: Myopia and ambient night-time lighting. *Nature*, 404:144, 2000.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

D. M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16:342–355, 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.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318, 2015.

D. Heckerman and D. Geiger. Likelihoods and parameter priors for Bayesian networks. Technical report, Microsoft Research (MSR-TR-95-54), 1995.

K. D. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, OR, USA, 2006. AUAI Press.

G. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–75, 1994.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian network structure using LP relaxations. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 358–365, 2010.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 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)*, pages 249–257, Corvallis, OR, USA, 2009. AUAI Press.

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.

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*, Tokyo, Japan, 2003.

P. Kemmeren, K. Sameith, L. A. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, S. van Heesch, M. M.. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. Brabers, A. J. Miles, D. Bouwmeester, S. R. van Hooff, H. van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. van Leenen, M. J. Groot Koerkamp, and F. C. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157:740–752, 2014.

M. Koivisto. Advances in exact Bayesian structure discovery in Bayesian networks. In *Proceedings of the 22nd Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 241–248, Corvallis, OR, USA, 2006. AUAI Press.

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

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

K. Korb, L. Hope, A. Nicholson, and K. Axnick. Varieties of causal intervention. In *Proceedings of the Pacific Rim Conference on AI*, pages 322–331, 2004.

S.L. Lauritzen. *Graphical Models*. Oxford University Press, New York, USA, 1996.

F. Markowetz, S. Grossmann, and R. Spang. Probabilistic soft interventions in conditional Gaussian networks. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 214–221, 2005.

B. D. McKay. Acyclic digraphs and eigenvalues of (0, 1)–matrices. *Journal of Integer Sequences*, 7:1–5, 2004.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 403–441, San Francisco, CA, USA, 1995. Morgan Kaufmann.

F. H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine*, 367:1562–1564, 2012.

J. M. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 440–448, Corvallis, OR, USA, 2013. AUAI Press.

J. Pearl. Belief networks revisited. *Artificial Intelligence*, 59:49–56, 1993a.

J. Pearl. Graphical models, causality and interventions. *Statistical Science*, 8:266–269, 1993b.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

C. S. Peirce. A theory of probable inference. In Charles S. Peirce, editor, *Studies in Logic by Members of the Johns Hopkins University*, pages 126–181. Little, Brown, and Company, 1883.

C. S. Peirce and J. Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885.

J. Peters. Asymmetries of time series under inverting their direction. Diploma Thesis, University of Heidelberg, 2008. `http://stat.ethz.ch/people/jopeters`.

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, 2011.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *ArXiv e-prints (1501.01332)*, 2015.

K. R. Popper. *The Logic of Scientific Discovery*. Routledge, 2002. ISBN 0-415-27844-9. 1st English Edition:1959.

G. E. Quinn, C. H. Shin, M. G. Maguire, and R. A. Stone. Myopia and ambient lighting at night. *Nature*, 399:113–114, 1999.

H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.

J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

R. W. Robinson. Enumeration of acyclic digraphs. In *Proceedings of the 2nd Chapel Hill Conference on Combinatorial Mathematics and its Applications (University of North Carolina)*, pages 391–399, 1970.

R. W. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New Directions in the Theory of Graphs*, pages 239–273. Academic Press, NY, 1973.

RProject. The R project for statistical computing, 2015. `http://www.r-project.org/`.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 556–565, Corvallis, OR, USA, 2013. AUAI Press.

E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

N. A. Sheehan, M. B., and J. Cussens. Improved maximum likelihood reconstruction of complex multi-generational pedigrees. *Theoretical Population Biology*, 97:11 – 19, 2014.

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.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semimarkovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI) - Volume 2*, pages 1219–1226. AAAI Press, 2006.

I. Shpitser, T. J. Van der Weele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 527–536, Corvallis, OR, USA, 2010. AUAI Press.

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

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241, 1951.

V. P. Skitovič. Linear forms in independent random variables and the normal distribution law (in Russian). *Izvestiia AN SSSR, Ser. Matem.*, 18:185–200, 1954.

V. P. Skitovič. 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 (Lecture notes in statistics)*. Springer-Verlag, New York, NY, 1993.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA, 2nd edition, 2000.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5:465–472, 1990.

R. P. Stanley. Acyclic orientations of graphs. *Discrete Mathematics*, 7:171–178, 1973.

M. Studený and D. Haws. Learning Bayesian network structure: Towards the essential graph by integer linear programming tools. *International Journal of Approximate Reasoning*, 55 (4):1043 – 1071, 2014.

J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Department of Computer Science, University of California, Los Angeles, 2002.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17thnd Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–522, San Francisco, CA, USA, 2001. Morgan Kaufmann.

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)*, 2010.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In P. B. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, San Francisco, CA, USA, 1991. Morgan Kaufmann.

Wikipedia. List of countries by coffee consumption per capita. Website, 29.01.2013, 6:15 p.m., GMT+1, 2013a. `http://en.wikipedia.org/wiki/List_of_countries_by_coffee_consumption_per_capita`.

Wikipedia. List of countries by nobel laureates per capita. Website, 29.01.2013, 6:15 p.m., GMT+1, 2013b. `http://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita`.

Wikipedia. James Lind. Website, 10.3.2015, 2015. http://en.wikipedia.org/wiki/James_Lind.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B*, 73:3–36, 2011.

P. G. Wright. *The Tariff on Animal and Vegetable Oils*. Investigations in international commercial policies. Macmillan, 1928.

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

S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1921b.

K. Zadnik, L. A. Jones, B. C. Irvin, R. N. Kleinstein, R. E. Manny, J. A. Shin, and D. O. Mutti. Vision: Myopia and ambient night-time lighting. *Nature*, 404:143–144, 2000.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 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)*, pages 647–655, Corvallis, OR, USA, 2009. AUAI Press.

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)*, pages 804–813, Corvallis, OR, USA, 2011. AUAI Press.

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)*, Corvallis, OR, USA, 2011. AUAI Press.