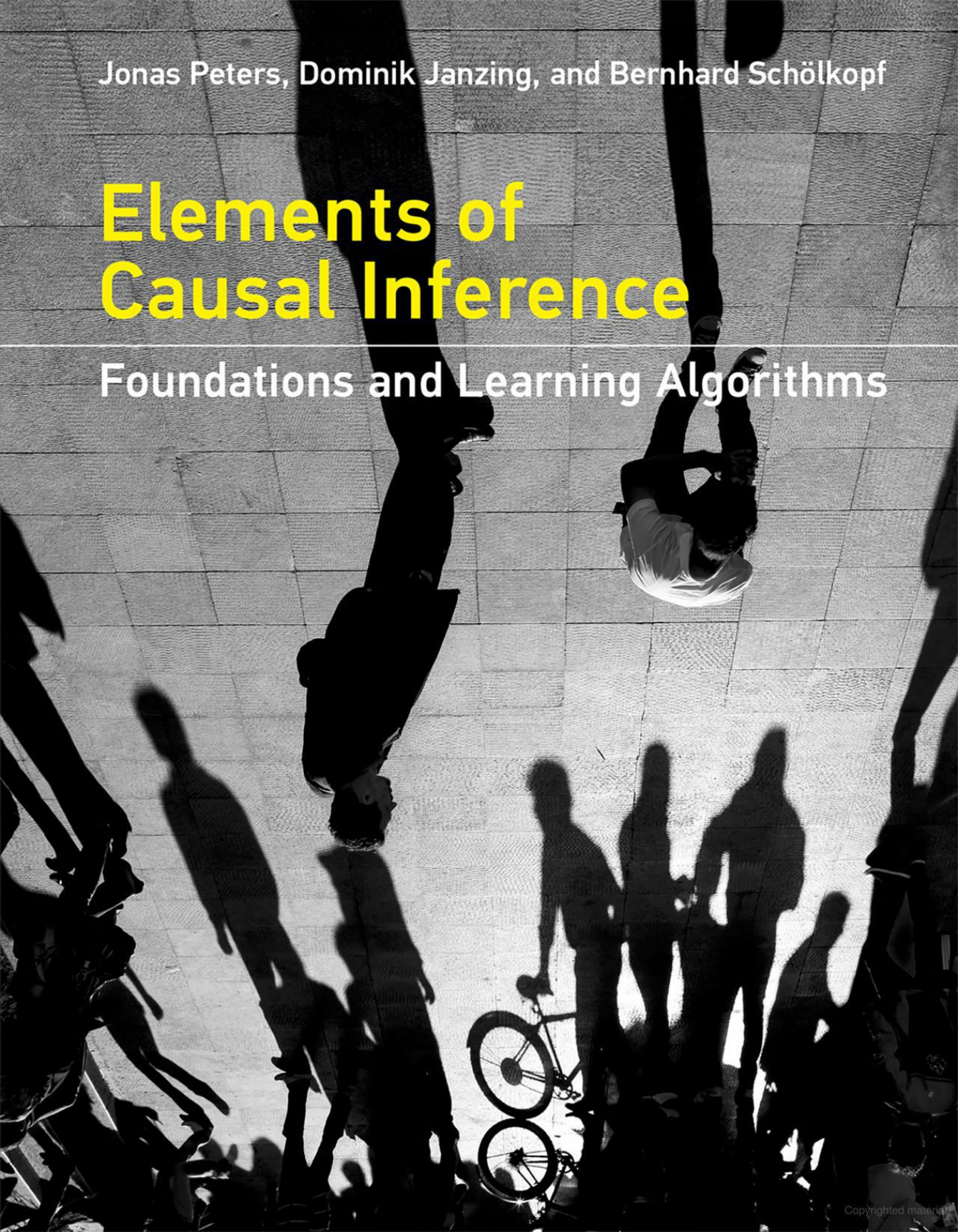
Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

Elements of Causal Inference

Foundations and Learning Algorithms



Elements of Causal Inference

Foundations and Learning Algorithms

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in LaTeX by the authors.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data
Names: Peters, Jonas. | Janzing, Dominik. | Schölkopf, Bernhard.
Title: Elements of causal inference : foundations and learning algorithms / Jonas Peters, Dominik Janzing, and Bernhard Schölkopf.
Description: Cambridge, MA : MIT Press, 2017. | Series: Adaptive computation and machine learning series | Includes bibliographical references and index.
Identifiers: LCCN 2017020087 | ISBN 9780262037310 (hardcover : alk. paper)
Subjects: LCSH: Machine learning. | Logic, Symbolic and mathematical. | Causation. | Inference. | Computer algorithms.
Classification: LCC Q325.5 .P48 2017 | DDC 006.3/1–dc23
LC record available at https://lccn.loc.gov/2017020087

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Causality is a fascinating topic of research. Its mathematization has only relatively recently started, and many conceptual problems are still being debated — often with considerable intensity.

While this book summarizes the results of spending a decade assaying causality, others have studied this problem much longer than we have, and there already exist books about causality, including the comprehensive treatments of Pearl [2009], Spirtes et al. [2000], and Imbens and Rubin [2015]. We hope that our book is able to complement existing work in two ways.

First, the present book represents a bias toward a subproblem of causality that may be considered both the most fundamental and the least realistic. This is the cause-effect problem, where the system under analysis contains only two observables. We have studied this problem in some detail during the last decade. We report much of this work, and try to embed it into a larger context of what we consider fundamental for gaining a selective but profound understanding of the issues of causality. Although it might be instructive to study the bivariate case first, following the sequential chapter order, it is also possible to directly start reading the multivariate chapters; see Figure I.

And second, our treatment is motivated and influenced by the fields of machine learning and computational statistics. We are interested in how methods thereof can help with the inference of causal structures, and even more so whether causal reasoning can inform the way we should be doing machine learning. Indeed, we feel that some of the most profound open issues of machine learning are best understood if we do not take a random experiment described by a probability distribution as our starting point, but instead we consider causal structures underlying the distribution.

We try to provide a systematic introduction into the topic that is accessible to readers familiar with the basics of probability theory and statistics or machine

learning (for completeness, the most important concepts are summarized in Appendices A.1 and A.2).

While we build on the graphical approach to causality as represented by the work of Pearl [2009] and Spirtes et al. [2000], our personal taste influenced the choice of topics. To keep the book accessible and focus on the conceptual issues, we were forced to devote regrettably little space to a number of significant issues in causality, be it advanced theoretical insights for particular settings or various methods of practical importance. We have tried to include references to the literature for some of the most glaring omissions, but we may have missed important topics.

Our book has a number of shortcomings. Some of them are inherited from the field, such as the tendency that theoretical results are often restricted to the case where we have infinite amounts of data. Although we do provide algorithms and methodology for the finite data case, we do not discuss statistical properties of such methods. Additionally, at some places we neglect measure theoretic issues, often by assuming the existence of densities. We find all of these questions both relevant and interesting but made these choices to keep the book short and accessible to a broad audience.

Another disclaimer is in order. Computational causality methods are still in their infancy, and in particular, learning causal structures from data is only doable in rather limited situations. We have tried to include concrete algorithms wherever possible, but we are acutely aware that many of the problems of causal inference are harder than typical machine learning problems, and we thus make no promises as to whether the algorithms will work on the reader's problems. Please do not feel discouraged by this remark — causal learning is a fascinating topic and we hope that reading this book may convince you to start working on it.

We would have not been able to finish this book without the support of various people.

We gratefully acknowledge support for a Research in Pairs stay of the three authors at the Mathematisches Forschungsinstitut Oberwolfach, during which a substantial part of this book was written.

We thank Michel Besserve, Peter Bühlmann, Rune Christiansen, Frederick Eberhardt, Jan Ernest, Philipp Geiger, Niels Richard Hansen, Alain Hauser, Biwei Huang, Marek Kaluba, Hansruedi Künsch, Steffen Lauritzen, Jan Lemeire, David Lopez-Paz, Marloes Maathuis, Nicolai Meinshausen, Søren Wengel Mogensen, Joris Mooij, Krikamol Muandet, Judea Pearl, Niklas Pfister, Thomas Richardson, Mateo Rojas-Carulla, Eleni Sgouritsa, Carl Johann Simon-Gabriel, Xiaohai Sun, Ilya Tolstikhin, Kun Zhang, and Jakob Zscheischler for many helpful comments and interesting discussions during the time this book was written. In particular,

Notation and Terminology

X, Y, Z	random variable; for noise variables, we use $N, N_X, N_j, \ldots$
x	value of a random variable X
P	probability measure
P_X	probability distribution of X
$X^1, \ldots, X^n \overset{\text{iid}}{\sim} P_X$	an i.i.d. sample of size n; sample index is usually i
$P_{Y\|X=x}$	conditional distribution of Y given $X = x$
$P_{Y\|X}$	collection of $P_{Y\|X=x}$ for all x; for short: conditional of Y given X
p	density (either probability mass function or probability density function)
p_X	density of P_X
$p(x)$	density of P_X evaluated at the point x
$p(y\|x)$	(conditional) density of $P_{Y\|X=x}$ evaluated at y
$\mathbb{E}[X]$	expectation of X
$\text{var}[X]$	variance of X
$\text{cov}[X, Y]$	covariance of X, Y
$X \perp\!\!\!\perp Y$	independence between random variables X and Y
$X \perp\!\!\!\perp Y \,\|\, Z$	conditional independence
$\mathbf{X} = (X_1, \ldots, X_d)$	random vector of length d; dimension index is usually j
$\mathfrak{C}$	structural causal model
$P_Y^{\mathfrak{C};do(X:=3)}$	intervention distribution
$P_Y^{\mathfrak{C}\|Z=2,X=1;do(X:=3)}$	counterfactual distribution
$\mathcal{G}$	graph
$\mathbf{PA}_X^{\mathcal{G}}$, $\mathbf{DE}_X^{\mathcal{G}}$, $\mathbf{AN}_X^{\mathcal{G}}$	parents, descendants, and ancestors of node X in graph $\mathcal{G}$

1

Statistical and Causal Models

Using statistical learning, we try to infer properties of the dependence among random variables from observational data. For instance, based on a joint sample of observations of two random variables, we might build a predictor that, given new values of only one of them, will provide a good estimate of the other one. The theory underlying such predictions is well developed, and — although it applies to simple settings — already provides profound insights into learning from data. For two reasons, we will describe some of these insights in the present chapter. First, this will help us appreciate how much harder the problems of *causal* inference are, where the underlying model is no longer a fixed joint distribution of random variables, but a structure that implies multiple such distributions. Second, although finite sample results for causal estimation are scarce, it is important to keep in mind that the basic statistical estimation problems do not go away when moving to the more complex causal setting, even if they seem small compared to the causal problems that do not appear in purely statistical learning. Building on the preceding groundwork, the chapter also provides a gentle introduction to the basic notions of causality, using two examples, one of which is well known from machine learning.

1.1 Probability Theory and Statistics

Probability theory and statistics are based on the model of a random experiment or probability space $(\Omega, \mathcal{F}, P)$. Here, Ω is a set (containing all possible outcomes), $\mathcal{F}$ is a collection of events $A \subseteq \Omega$, and P is a measure assigning a probability to each event. Probability theory allows us to reason about the outcomes of random experiments, given the preceding mathematical structure. Statistical learning, on

the other hand, essentially deals with the inverse problem: We are given the outcomes of experiments, and from this we want to infer properties of the underlying mathematical structure. For instance, suppose that we have observed data

$$(x_1, y_1), \dots, (x_n, y_n), \tag{1.1}$$

where $x_i \in \mathcal{X}$ are **inputs** (sometimes called **covariates** or **cases**) and $y_i \in \mathcal{Y}$ are **outputs** (sometimes called **targets** or **labels**). We may now assume that each (x_i, y_i), $i = 1, \dots, n$, has been generated independently by the same unknown random experiment. More precisely, such a model assumes that the observations $(x_1, y_1), \dots, (x_n, y_n)$ are realizations of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ that are **i.i.d. (independent and identically distributed)** with joint distribution $P_{X,Y}$. Here, X and Y are random variables taking values in metric spaces $\mathcal{X}$ and $\mathcal{Y}$.[1] Almost all of statistics and machine learning builds on i.i.d. data. In practice, the i.i.d. assumption can be violated in various ways, for instance if distributions shift or interventions in a system occur. As we shall see later, some of these are intricately linked to causality.

We may now be interested in certain properties of $P_{X,Y}$, such as:

(i) the expectation of the output given the input, $f(x) = \mathbb{E}[Y \,|\, X = x]$, called **regression**, where often $\mathcal{Y} = \mathbb{R}$,

(ii) a binary **classifier** assigning each x to the class that is more likely, $f(x) = \operatorname{argmax}_{y \in \mathcal{Y}} P(Y = y \,|\, X = x)$, where $\mathcal{Y} = \{\pm 1\}$,

(iii) the density $p_{X,Y}$ of $P_{X,Y}$ (assuming it exists).

In practice, we seek to estimate these properties from finite data sets, that is, based on the sample (1.1), or equivalently an empirical distribution $P^n_{X,Y}$ that puts a point mass of equal weight on each observation.

This constitutes an **inverse problem**: We want to estimate a property of an object we cannot observe (the underlying distribution), based on observations that are obtained by applying an operation (in the present case: sampling from the unknown distribution) to the underlying object.

[1] A random variable X is a measurable function $\Omega \to \mathcal{X}$, where the metric space $\mathcal{X}$ is equipped with the Borel σ-algebra. Its distribution P_X on $\mathcal{X}$ can be obtained from the measure P of the underlying probability space $(\Omega, \mathcal{F}, P)$. We need not worry about this underlying space, and instead we generally start directly with the distribution of the random variables, assuming the random experiment directly provides us with values sampled from that distribution.

1.2 Learning Theory

Now suppose that just like we can obtain f from $P_{X,Y}$, we use the empirical distribution to infer empirical estimates f^n. This turns out to be an **ill-posed problem** [e.g., Vapnik, 1998], since for any values of x that we have not seen in the sample $(x_1, y_1), \ldots, (x_n, y_n)$, the conditional expectation is undefined. We may, however, define the function f on the observed sample and extend it according to any fixed rule (e.g., setting f to $+1$ outside the sample or by choosing a continuous piecewise linear f). But for any such choice, small changes in the input, that is, in the empirical distribution, can lead to large changes in the output. No matter how many observations we have, the empirical distribution will usually not perfectly approximate the true distribution, and small errors in this approximation can then lead to large errors in the estimates. This implies that without additional assumptions about the class of functions from which we choose our empirical estimates f^n, we cannot guarantee that the estimates will approximate the optimal quantities f in a suitable sense. In statistical learning theory, these assumptions are formalized in terms of **capacity** measures. If we work with a function class that is so rich that it can fit most conceivable data sets, then it is not surprising if we can fit the data at hand. If, however, the function class is a priori restricted to have small capacity, then there are only a few data sets (out of the space of all possible data sets) that we can explain using a function from that class. If it turns out that nevertheless we can explain the data at hand, then we have reason to believe that we have found a regularity underlying the data. In that case, we can give probabilistic guarantees for the solution's accuracy on future data sampled from the same distribution $P_{X,Y}$.

Another way to think of this is that our function class has incorporated **a priori knowledge** (such as smoothness of functions) consistent with the regularity underlying the observed data. Such knowledge can be incorporated in various ways, and different approaches to machine learning differ in how they handle the issue. In Bayesian approaches, we specify prior distributions over function classes and noise models. In regularization theory, we construct suitable regularizers and incorporate them into optimization problems to bias our solutions.

The complexity of statistical learning arises primarily from the fact that we are trying to solve an inverse problem based on empirical data — if we were given the full probabilistic model, then all these problems go away. When we discuss causal models, we will see that in a sense, the causal learning problem is harder in that it is ill-posed *on two levels*. In addition to the statistical ill-posed-ness, which is essentially because a finite sample of arbitrary size will never contain all information about the underlying distribution, there is an ill-posed-ness due to the

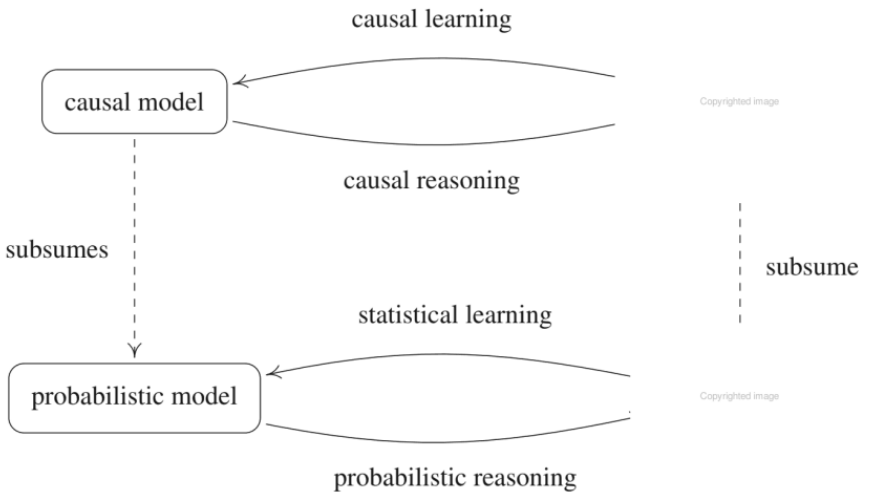


Figure 1.1: Terminology used by the present book for various **probabilistic inference** problems (bottom) and **causal inference** problems (top); see Section 1.3. Note that we use the term "inference" to include both learning and reasoning.

not distract us from the fact, however, that the ill-posed-ness of the usual statistical problems is still there (and thus it is important to worry about the capacity of function classes also in causality, such as by using additive noise models — see Section 4.1.4 below), only confounded by an additional difficulty arising from the fact that we are trying to estimate a richer structure than just a probabilistic one. We will refer to this overall problem as **causal learning**. Figure 1.1 summarizes the relationships between the preceding problems and models.

To learn causal structures from observational distributions, we need to understand how causal models and statistical models relate to each other. We will come back to this issue in Chapters 4 and 7 but provide an example now. A well-known topos holds that *correlation does not imply causation*; in other words, statistical properties alone do not determine causal structures. It is less well known that one may postulate that while we cannot infer a concrete causal structure, we may at least infer the existence of causal links from statistical dependences. This was first understood by Reichenbach [1956]; we now formulate his insight (see also Figure 1.2).[3]

[3]For clarity, we formulate some important assumptions as *principles*. We do not take them for granted throughout the book; in this sense, they are not axioms.

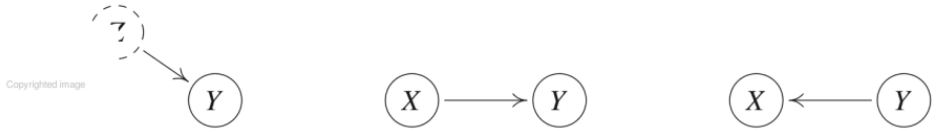


Figure 1.2: Reichenbach's common cause principle establishes a link between statistical properties and causal structures. A *statistical* dependence between two observables X and Y indicates that they are *caused* by a variable Z, often referred to as a **confounder** (left). Here, Z may coincide with either X or Y, in which case the figure simplifies (middle/right). The principle further argues that X and Y are statistically independent, conditional on Z. In this figure, direct causation is indicated by arrows; see Chapters 3 and 6.

Principle 1.1 (Reichenbach's common cause principle) *If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a third variable Z that* causally *influences both. (As a special case, Z may coincide with either X or Y.) Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp\!\!\!\perp Y \,|\, Z$.*

In practice, dependences may also arise for a reason different from the ones mentioned in the common cause principle, for instance: (1) The random variables we observe are conditioned on others (often implicitly by a selection bias). We shall return to this issue; see Remark 6.29. (2) The random variables only *appear* to be dependent. For example, they may be the result of a search procedure over a large number of pairs of random variables that was run without a multiple testing correction. In this case, inferring a dependence between the variables does not satisfy the desired type I error control; see Appendix A.2. (3) Similarly, both random variables may inherit a time dependence and follow a simple physical law, such as exponential growth. The variables then *look* as if they depend on each other, but because the i.i.d. assumption is violated, there is no justification of applying a standard independence test. In particular, arguments (2) and (3) should be kept in mind when reporting "spurious correlations" between random variables, as it is done on many popular websites.

1.4 Two Examples

1.4.1 Pattern Recognition

As the first example, we consider *optical character recognition*, a well-studied problem in machine learning. This is not a run-of-the-mill example of a causal structure, but it may be instructive for readers familiar with machine learning. We

describe two causal models giving rise to a dependence between two random variables, which we will assume to be handwritten digits X and class labels Y. The two models will lead to the same statistical structure, using distinct underlying causal structures.

Model (i) assumes we generate each pair of observations by providing a sequence of class labels y to a human writer, with the instruction to always produce a corresponding handwritten digit image x. We assume that the writer tries to do a good job, but there may be noise in perceiving the class label and executing the motor program to draw the image. We can model this process by writing the image X as a function (or mechanism) f of the class label Y (modeled as a random variable) and some independent noise N_X (see Figure 1.3, left). We can then compute $P_{X,Y}$ from P_Y, P_{N_X}, and f. This is referred to as the **observational distribution**, where the word "observational" refers to the fact that we are passively observing the system without intervening. X and Y will be dependent random variables, and we will be able to learn the mapping from x to y from observations and predict the correct label y from an image x better than chance.

There are two possible interventions in this causal structure, which lead to **intervention distributions**.[4] If we intervene on the resulting image X (by manipulating it, or exchanging it for another image after it has been produced), then this has no effect on the class labels that were provided to the writer and recorded in the data set. Formally, changing X has no effect on Y since $Y := N_Y$. Intervening on Y, on the other hand, amounts to changing the class labels provided to the writer. This will obviously have a strong effect on the produced images. Formally, changing Y has an effect on X since $X := f(Y, N_X)$. This directionality is visible in the arrow in the figure, and we think of this arrow as representing direct causation.

In alternative **model (ii)**, we assume that we *do not* provide class labels to the writer. Rather, the writer is asked to decide himself or herself which digits to write, and to record the class labels alongside. In this case, both the image X and the recorded class label Y are functions of the writer's intention (call it Z and think of it as a random variable). For generality, we assume that not only the process generating the image is noisy but also the one recording the class label, again with independent noise terms (see Figure 1.3, right). Note that if the functions and noise terms are chosen suitably, we can ensure that this model entails an observational distribution $P_{X,Y}$ that is identical to the one entailed by model (i).[5]

[4] We shall see in Section 6.3 that a more general way to think of interventions is that they change functions and random variables.

[5] Indeed, Proposition 4.1 implies that *any* joint distribution $P_{X,Y}$ can be entailed by both models.

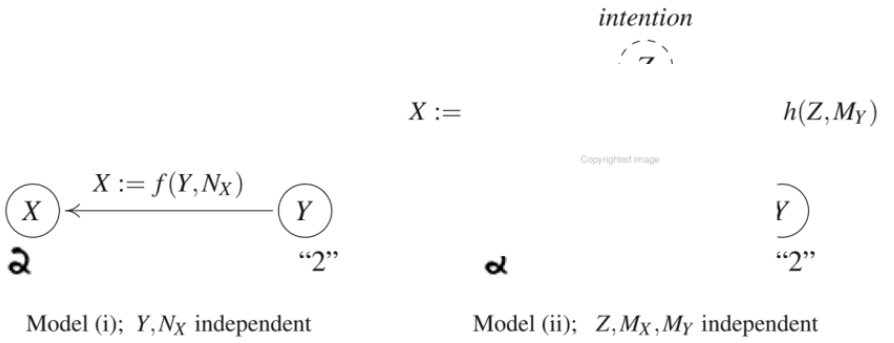


Figure 1.3: Two structural causal models of handwritten digit data sets. In the left model (i), a human is provided with class labels Y and produces images X. In the right model (ii), the human decides which class to write (Z) and produces both images and class labels. For suitable functions f, g, h and noise variables N_X, M_X, M_Y, Z, the two models produce the same observable distribution $P_{X,Y}$, yet they are interventionally different; see Section 1.4.1.

Let us now discuss possible interventions in model (ii). If we intervene on the image X, then things are as we just discussed and the class label Y is not affected. However, if we intervene on the class label Y (i.e., we change what the writer has recorded as the class label), then unlike before this will *not* affect the image.

In summary, without restricting the class of involved functions and distributions, the causal models described in (i) and (ii) induce the same observational distribution over X and Y, but different intervention distributions. This difference is not visible in a purely probabilistic description (where everything derives from $P_{X,Y}$). However, we were able to discuss it by incorporating structural knowledge about how $P_{X,Y}$ comes about, in particular graph structure, functions, and noise terms.

Models (i) and (ii) are examples of **structural causal models (SCMs)**, sometimes referred to as **structural equation models** [e.g., Aldrich, 1989, Hoover, 2008, Pearl, 2009, Pearl et al., 2016]. In an SCM, all dependences are generated by functions that compute variables from other variables. Crucially, these functions are to be read as assignments, that is, as functions as in computer science rather than as mathematical equations. We usually think of them as modeling physical mechanisms. An SCM entails a joint distribution over all observables. We have seen that the same distribution can be generated by different SCMs, and thus information about the effect of interventions (and, as we shall see in Section 6.4, information about counterfactuals) may be lost when we make the transition from an SCM to the corresponding probability model. In this book, we take SCMs as

our starting point and try to develop everything from there.

We conclude with two points connected to our example:

First, Figure 1.3 nicely illustrates Reichenbach's common cause principle. The dependence between X and Y admits several causal explanations, and X and Y become independent if we condition on Z in the right-hand figure: The image and the label share no information that is not contained in the intention.

Second, it is sometimes said that causality can only be discussed when taking into account the notion of **time**. Indeed, time does play a role in the preceding example, for instance by ruling out that an intervention on X will affect the class label. However, this is perfectly fine, and indeed it is quite common that a statistical data set is generated by a process taking place in time. For instance, in model (i), the underlying reason for the statistical dependence between X and Y is a dynamical process. The writer reads the label and plans a movement, entailing complicated processes in the brain, and finally executes the movement using muscles and a pen. This process is only partly understood, but it is a physical, dynamical process taking place in time whose end result leads to a non-trivial joint distribution of X and Y. When we perform statistical learning, we only care about the end result. Thus, not only causal structures, but also purely probabilistic structures may arise through processes taking place in time — indeed, one could hold that this is ultimately the only way they can come about. However, in both cases, it is often instructive to disregard time. In statistics, time is often not necessary to discuss concepts such as statistical dependence. In causal models, time is often not necessary to discuss the effect of interventions. But both levels of description can be thought of as abstractions of an underlying more accurate physical model that describes reality more fully than either; see Table 1.1. Moreover, note that variables in a model may not necessarily refer to well-defined time instances. If, for instance, a psychologist investigates the statistical or causal relation between the motivation and the performance of students, both variables cannot easily be assigned to specific time instances. Measurements that refer to well-defined time instances are rather typical for "hard" sciences like physics and chemistry.

1.4.2 Gene Perturbation

We have seen in Section 1.4.1 that different causal structures lead to different intervention distributions. Sometimes, we are indeed interested in predicting the outcome of a random variable under such an intervention. Consider the following, in some ways oversimplified, example from genetics. Assume that we are given activity data from gene A and, correspondingly, measurements of a phenotype; see

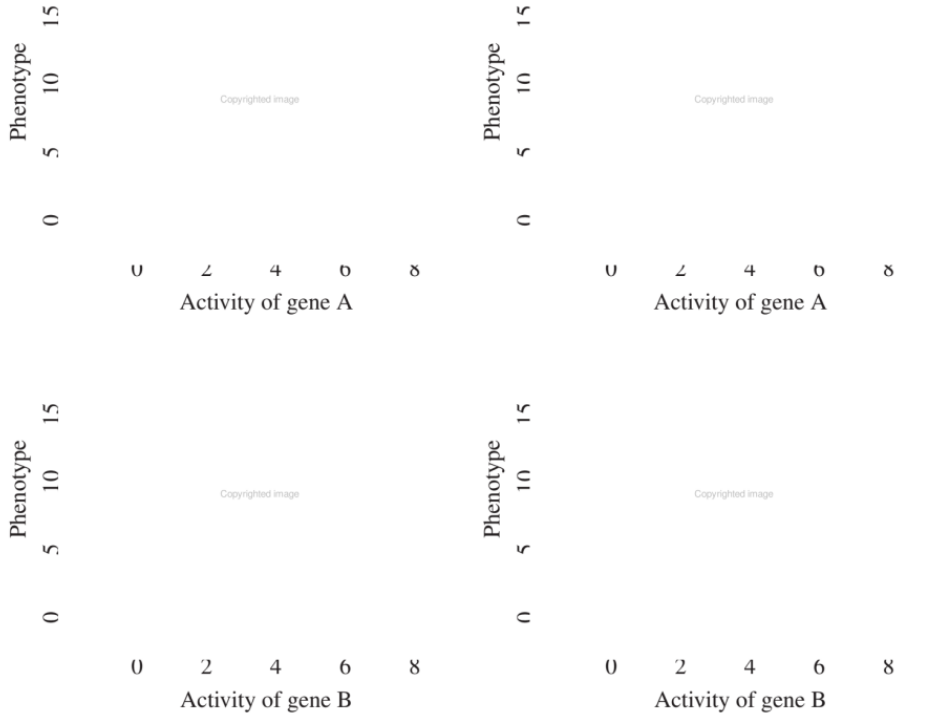


Figure 1.4: The activity of two genes (top: gene A; bottom: gene B) is strongly correlated with the phenotype (black dots). However, the best prediction for the phenotype when *deleting the gene*, that is, setting its activity to 0 (left), depends on the causal structure (right). If a common cause is responsible for the correlation between gene and phenotype, we expect the phenotype to behave under the intervention as it usually does (bottom right), whereas the intervention clearly changes the value of the phenotype if it is causally influenced by the gene (top right). The idea of this figure is based on Peters et al. [2016].

2

Assumptions for Causal Inference

Now that we have encountered the basic components of SCMs, it is a good time to pause and consider some of the assumptions we have seen, as well as what these assumptions imply for the purpose of causal reasoning and learning.

A crucial notion in our discussion will be a form of **independence**, and we can informally introduce it using an optical illusion known as the Beuchet chair. When we see an object such as the one on the left of Figure 2.1, our brain makes the assumption that the object and the mechanism by which the information contained in its light reaches our brain are independent. We can violate this assumption by looking at the object from a very specific viewpoint. If we do that, perception goes wrong: We perceive the three-dimensional structure of a chair, which in reality is not there. Most of the time, however, the independence assumption does hold. If we look at an object, our brain assumes that the object is independent from our vantage point and the illumination. So there should be no unlikely coincidences, no separate 3D structures lining up in two dimensions, or shadow boundaries coinciding with texture boundaries. This is called the *generic viewpoint assumption* in vision [Freeman, 1994].

The independence assumption is more general than this, though. We will see in Section 2.1 below that the causal generative process is composed of autonomous modules that do not inform or influence each other. As we shall describe below, this means that while one module's output may influence another module's input, the modules themselves are independent of each other.

In the preceding example, while the overall percept is a function of object, lighting, and viewpoint, the object and the lighting are not affected by us moving about — in other words, some components of the overall causal generative model remain *invariant*, and we can infer three-dimensional information from this invariance.

Figure 2.1: The left panel shows a generic view of the (separate) parts comprising a *Beuchet chair*. The right panel shows the illusory percept of a chair if the parts are viewed from a single, very special vantage point. From this *accidental viewpoint*, we perceive a chair. (Image courtesy of Markus Elsholz.)

This is the basic idea of *structure from motion* [Ullman, 1979], which plays a central role in both biological vision and computer vision.

2.1 The Principle of Independent Mechanisms

We now describe a simple cause-effect problem and point out several observations. Subsequently, we shall try to provide a unified view of how these observation relate to each other, arguing that they derive from a common independence principle.

Suppose we have estimated the joint density $p(a,t)$ of the altitude A and the average annual temperature T of a sample of cities in some country (see Figure 4.6 on page 65). Consider the following ways of expressing $p(a,t)$:

$$\begin{aligned} p(a,t) &= p(a|t)\; p(t) \\ &= p(t|a)\; p(a) \end{aligned} \tag{2.1}$$

The first decomposition describes T and the conditional $A|T$. It corresponds to a factorization of $p(a,t)$ according to the graph $T \to A$.[1] The second decomposition corresponds to a factorization according to $A \to T$ (cf. Definition 6.21). Can we

[1] Note that the conditional density $p(a|t)$ allows us to compute $p(a,t)$ (and thus also $p(a)$) from

decide which of the two structures is the *causal* one (i.e., in which case would we be able to think of the arrow as causal)?

A first idea (see Figure 2.2, left) is to consider the **effect of interventions**. Imagine we could change the altitude A of a city by some hypothetical mechanism that raises the grounds on which the city is built. Suppose that we find that the average temperature decreases. Let us next imagine that we devise another intervention experiment. This time, we do not change the altitude, but instead we build a massive heating system around the city that raises the average temperature by a few degrees. Suppose we find that the altitude of the city is unaffected. Intervening on A has changed T, but intervening on T has not changed A. We would thus reasonably prefer $A \to T$ as a description of the causal structure.

Why do we find this description of the effect of interventions plausible, even though the hypothetical intervention is hard or impossible to carry out in practice?

If we change the altitude A, then we assume that the *physical mechanism* $p(t|a)$ responsible for producing an average temperature (e.g., the chemical composition of the atmosphere, the physics of how pressure decreases with altitude, the meteorological mechanisms of winds) is still in place and leads to a changed T. This would hold true independent of the distribution from which we have sampled the cities, and thus independent of $p(a)$. Austrians may have founded their cities in locations subtly different from those of the Swiss, but the mechanism $p(t|a)$ would apply in both cases.[2]

If, on the other hand, we change T, then we have a hard time thinking of $p(a|t)$ as a mechanism that is still in place — we probably do not believe that such a mechanism exists in the first place. Given a set of different city distributions $p(a,t)$, while we could write them all as $p(a|t)\ p(t)$, we would find that it is impossible to explain them all using an invariant $p(a|t)$.

Our intuition can be rephrased and postulated in two ways: If $A \to T$ is the correct causal structure, then

(i) it is in principle **possible to perform a localized intervention** on A, in other words, to change $p(a)$ without changing $p(t|a)$, and

(ii) $p(a)$ and $p(t|a)$ are **autonomous**, **modular**, or **invariant** mechanisms or objects in the world.

$p(t)$, which may serve to motivate the direction of the arrow in $T \to A$ for the time being. This will be made precise in Definition 6.21.

[2]This is an idealized setting — no doubt counterexamples to these general remarks can be constructed.

We will presently argue that the principle is sufficiently broad to cover the main aspects of causal reasoning and causal learning (see Figure 2.2). Let us address three aspects, corresponding, from left to right, to the three branches of the tree in Figure 2.2.

1. One way to think of these modules is as physical machines that incorporate an input-output behavior. This assumption implies that we can **change one mechanism without affecting the others** — or, in causal terminology, we can *intervene* on one mechanism without affecting the others. Changing a mechanism will change its input-output behavior, and thus the inputs other mechanisms downstream might receive, but we are assuming that the physical mechanisms themselves are unaffected by this change. An assumption such as this one is often implicit to justify the possibility of interventions in the first place, but one can also view it as a more general basis for causal reasoning and causal learning. If a system allows such localized interventions, there is no physical pathway that would connect the mechanisms to each other in a directed way by "meta-mechanisms." The latter makes it plausible that we can also expect a tendency for mechanisms to remain *invariant* with respect to changes within the system under consideration and possibly also to some changes stemming from outside the system (see Section 7.1.6). This kind of *autonomy* of mechanisms can be expected to help with *transfer* of knowledge learned in one domain to a related one where some of the *modules* coincide with the source domain (see Sections 5.2 and 8.3).

2. While the discussion of the first aspect focused on the physical aspect of independence and its ramifications, there is also an information theoretic aspect that is implied by the above. A time evolution involving several coupled objects and mechanisms can generate statistical dependence. This is related to our discussion from page 10, where we considered the dependence between the class label and the image of a handwritten digit. Similarly, mechanisms that are physically coupled will tend to generate information that can be quantified in terms of statistical or algorithmic information measures (see Sections 4.1.9 and 6.10 below).

   Here, it is important to distinguish between two levels of information: obviously, an effect contains information about its cause, but — according to the independence principle — the mechanism that generates the effect from its cause contains no information about the mechanism generating the cause. For a causal structure with more than two nodes, the independence princi-

ple states that the mechanism generating every node from its direct causes contain no information about each other.[4]

3. Finally, we should discuss how the assumption of independent noise terms, commonly made in structural equation modeling, is connected to the principle of independent mechanism. This connection is less obvious. To this end, consider a variable $E := f(C,N)$ where the noise N is discrete. For each value s taken by N, the assignment $E := f(C,N)$ reduces to a deterministic mechanism $E := f^s(C)$ that turns an input C into an output E. Effectively, this means that the noise randomly chooses between a number of mechanisms f^s (where the number equals the cardinality of the range of the noise variable N). Now suppose the noise variables for two mechanisms at the vertices X_j and X_k were statistically dependent.[5] Such a dependence could ensure, for instance, that whenever one mechanism f_j^s is active at node j, we know which mechanism f_k^t is active at node k. This would violate our principle of independent mechanisms.

   The preceding paragraph uses the somewhat extreme view of noise variables as selectors between mechanisms (see also Section 3.4). In practice, the role of the noise might be less pronounced. For instance, if the noise is additive (i.e., $E := f(C) + N$), then its influence on the mechanism is restricted. In this case, it can only *shift* the output of the mechanism up or down, so it selects between a set of mechanisms that are very similar to each other. This is consistent with a view of the noise variables as variables outside the system that we are trying to describe, representing the fact that a system can never be totally isolated from its environment. In such a view, one would think that a weak dependence of noises may be possible without invalidating the principle of independent mechanisms.

All of the above-mentioned aspects of Principle 2.1 may help for the problem of causal learning, in other words, they may provide information about causal structures. It is conceivable, however, that this information may in cases be conflicting, depending on which assumptions hold true in any given situation.

---

[4] There is an intuitive relation between this aspect of independence and the one described under 1.: whenever the mechanisms change independently, the change of one mechanism does not provide information on how the others have changed. Despite this overlap, the second independence contains an aspect that is not strictly contained in the first one because it is also applicable to a scenario in which none of the mechanisms has changed; for example, it refers also to homogeneous data sets.

[5] Although we have so far focused on the two-variable case, we phrase this argument such that it also applies for causal structures with more than two variables.

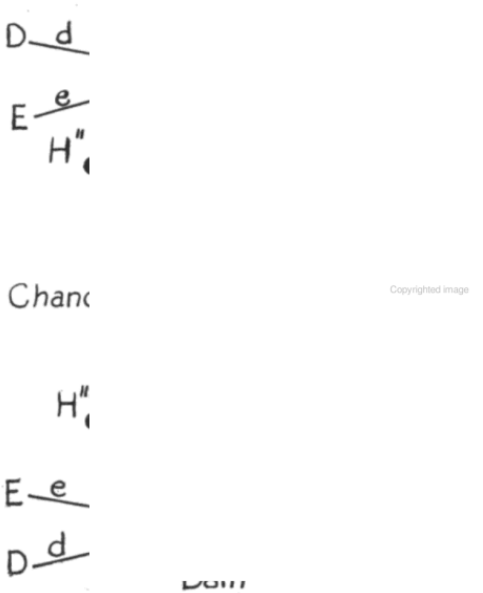


FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

Figure 2.3: Early path diagram; dam and sire are the female and male parents of a guinea pig, respectively. The path coefficients capture the importance of a given path, defined as *the ratio of the variability of the effect to be found when all causes are constant except the one in question, the variability of which is kept unchanged, to the total variability.* (Reproduced from Wright [1920].)

2.2 Historical Notes

The idea of autonomy and invariance is deeply engrained in the concept of structural equation models (SEMs) or SCMs. We prefer the latter term, since the term SEM has been used in a number of contexts where the structural assignments are used as algebraic equations rather than assignments. The literature is wide ranging, with overviews provided by Aldrich [1989], Hoover [2008], and Pearl [2009].

An intellectual antecedent to SEMs is the concept of a path model pioneered by Wright [1918, 1920, 1921] (see Figure 2.3). Although Wright was a biologist, SEMs are nowadays most strongly associated with econometrics. Following Hoover [2008], pioneering work on structural econometric models was done in the

1930s by Jan Tinbergen, and the conceptual foundations of probabilistic econometrics were laid in Trgyve Haavelmo's work [Haavelmo, 1944]. Early economists were trying to conceptualize the fact that unlike correlation, regression has a natural direction. The regression of Y on X leads to a solution that usually is *not* the inverse of the regression of X on Y.[6] But how would the data then tell us in which direction we should perform the regression? This is a problem of *observational equivalence*, and it is closely related to a problem econometricians call *identification*.

A number of early works saw a connection between what made a set of equations or relations *structural* [Frisch and Waugh, 1933], and properties of *invariance* and *autonomy* — according to Aldrich [1989], indeed the central notion in the pioneering work of Frisch et al. [1948]. Here, a *structural* relation was aiming for more than merely modeling an observed distribution of data — it was trying to capture an underlying structure connecting the variables of the model.

At the time, the Cowles Commission was a major economic research institute, instrumental in creating the field of econometrics. Its work related causality to the invariance properties of the structural econometric model [Hoover, 2008]. Pearl [2009] credits Marschak's opening chapter of a 1950 Cowles monograph with the idea that structural equations remain invariant to certain changes in the system [Marschak, 1950]. A crucial distinction emphasized by the Cowles work was the one between *endogenous* and *exogenous variables*. Endogeneous variables are those that the modeler tries to understand, while exogenous ones are determined by factors outside the model, and are taken as given. Koopmans [1950] assayed two principles for determining what should be treated as exogeneous. The *departmental principle* considers variables outside of the scope of the discipline as exogeneous (e.g., weather is exogeneous to economics). The (preferred) *causal principle* calls those variables exogenous that influence the remaining (endogeneous) variables, but are (almost) not influenced thereby.

Haavelmo [1943] interpreted structural equations as statements about hypothetical controlled experiments. He considered cyclic stochastic equation models and discussed the role of invariance as well as policy interventions. Pearl [2015] gives an appraisal of Haavelmo's role in the study of policy intervention questions and the development of the field of causal inference. In an account of causality in

[6] As an aside, while most of the early works were using linear equations only, there have also been attempts to generalize to nonlinear SEMs [Hoover, 2008].

economics and econometrics, Hoover [2008] discusses a system of the form

$$X^i := N_X^i$$
$$Y^i := \theta\, X^i + N_Y^i,$$

where the errors N_X^i, N_Y^i are i.i.d., and θ is a parameter. He attributes to Simon [1953] the view (which does not require any temporal order) that X^i may be referred to as causing Y^i since one knows all about X^i without knowing about Y^i, but not vice versa. The equations also allow us to predict the effect of interventions. Hoover goes on to argue that one can rewrite the system reversing the roles of X^i and Y^i while retaining the property that the error terms are uncorrelated.[7] He thus points out that we cannot infer the correct causal direction on the basis of a single set of data ("observational equivalence"). Experiments, either controlled or natural, could help us decide. If, for example, an experiment can change the conditional distribution of Y^i given X^i, without altering the marginal distribution of X^i, then it must be that X^i causes Y^i. Hoover refers to this as *Simon's invariance criterion*: the true causal order is the one that is invariant under the right sort of intervention.[8] Hurwicz [1962] argues that an equation system becomes *structural* by virtue of invariance to a domain of modifications. Such a system then bears resemblance to a natural law. Hurwicz recognized that one can use such modifications to determine structure, and that while structure is necessary for causality, it is not for prediction.

Aldrich [1989] provides an account of the role of autonomy in structural equation modeling. He argues that autonomous relations are likely to be more stable than others. He equates Haavelmo's *autonomous variables* with what subsequently became known as exogeneous variables. Autonomous variables are parameters fixed by external forces, or treated as stochastically independent.[9] Following Aldrich [1989, page 30], "the use of the qualifier *autonomous* and the phrase *forces external to the sector under consideration* suggest that ... the parameters of that model would be invariant to changes in the sectoral parameters." He also relates invariance to a notion termed *super-exogeneity* [Engle et al., 1983].

While the early proponents of structural equation modeling already had some profound insights in their causal underpinnings, the developments in computer sci-

[7] We shall revisit this topic in more detail in Section 4.1.3.

[8] We would argue that this may not hold true if interventions are coupled to each other, for example, to keep the *anticausal* conditional (which describes the cause, given its effect) invariant. This could be seen as a violation of Principle 2.1 *on the level of interventions*. We return to this point in Section 2.3.4.

[9] This is akin to the independence of noise terms we use in SCMs.

case, time evolution (assuming it is sufficiently ergodic) will tend to increase complexity. In the other way, we assume that we are considering open systems. Even if the time evolution for a closed system is invertible (e.g., in quantum mechanics, a unitary time evolution), the time evolution of an open subsystem (which interacts with its environment) in the generic case need not be invertible.

2.3.2 Physical Laws

An often discussed causal question can be addressed with the following example. The ideal gas law stipulates that pressure p, volume V, amount of substance n, and absolute temperature T satisfy the equation

$$p \cdot V = n \cdot R \cdot T, \tag{2.2}$$

where R is the ideal gas constant. If we, for instance, change the volume V allocated to a given amount of gas, then pressure p and/or temperature T will change, and the specifics will depend on the exact setup of the intervention. If, on the other hand, we change T, then V and/or p will change. If we keep p constant, then we can, at least approximately, construct a cycle involving T and V. So what causes what? It is sometimes argued that such laws show that it does not make sense to talk about causality unless the system is temporal. In the next paragraph, we argue that this is misleading. The gas law (2.2) refers to an *equilibrium state* of an underlying dynamical system, and writing it as a simple equation does not provide enough information about what interventions are in principle possible and what is their effect. SCMs and their corresponding directed acyclic graphs do provide us with this information, but in the general case of non-equilibrium systems, it is a hard problem whether and how a given dynamical systems leads to an SCM.

2.3.3 Cyclic Assignments

We think of SCMs as abstractions of underlying processes that take place in time. For these underlying processes, there is no problem with feedback loops, since at a sufficiently fast time scale, those loops will be unfolded in time, assuming there are no instantaneous interactions, which are arguably excluded by the finiteness of the speed of light.

Even though the time-dependent processes do not have cycles, it is possible that an SCM derived from such processes (for instance, by methods mentioned below in Remarks 6.5 and 6.7), involving only quantities that no longer depend on time, does have cycles. It becomes a little harder to define general interventions in such

systems, but certain types of interventions should still be doable. For instance, a hard intervention where we set the value of one variable to a fixed value may be possible (and realizable physically by a forcing term in an underlying set of differential equations; see Remark 6.7). This cuts the cycle, and we can then derive the entailed *intervention* distribution.

However, it may be impossible to derive an entailed *observational* distribution from a cyclic set of structural assignments. Let us consider the two assignments

$$X := f_X(Y, N_X)$$
$$Y := f_Y(X, N_Y)$$

and noise variables $N_X \perp\!\!\!\perp N_Y$. Just like in the case of acyclic models, we consider the noises and functions as given and seek to compute the entailed joint distribution of X and Y. To this end, let us start with the first assignment $X := f_X(Y, N_X)$, and substitute some initial Y into it. This yields an X, which we can then substitute into the other assignment. Suppose we iterate the two assignments and converge to some fixed point. This point would then correspond to a joint distribution of X, Y simultaneously satisfying both structural assignments as equalities of random variables.[12] Note that we have here assumed that the same N_X, N_Y are used at every step, rather than independent copies thereof.

However, such an equilibrium for X, Y need not always exist, and even if it does, it need not be the case that it can be found using the iteration. In the linear case, this has been analyzed by Lacerda et al. [2008] and Hyttinen et al. [2012]; see also Lauritzen and Richardson [2002]. For further details see Remark 6.5.

This observation that one may not always be able to get an entailed distribution satisfying two cyclic structural assignments is consistent with the view of SCMs as *abstractions* of underlying physical processes — abstractions whose domain of validity as causal models is limited. If we want to understand general cyclic systems, it may be unavoidable to study systems of differential equations rather than SCMs. For certain restricted settings, on the other hand, it can still make sense to stay on the phenomenologically more superficial level of SCMs; see, for example, Mooij et al. [2013]. One may speculate that this difficulty inherent to SCMs (or SEMs) is part of the reason why the econometrics community started off viewing SEMs as

[12] The fact that the assignments are satisfied as equalities of random variables means that we are considering an ensemble of systems that differ in the realizations of the noise variables. Each realization leads to a (possibly different) realization for X, Y, and thus the distribution of the noises implies a distribution over X, Y.

causal models, but later on parts of the community decided to forgo this interpretation in favor of a view of structural equations as purely algebraic equations.

2.3.4 Feasibility of Interventions

We have used the principle of independent mechanisms to motivate interventions that only affect one mechanism (or structural assignment) at a time. While real systems may admit such kind of interventions, there will also be interventions that replace several assignments at the same time. The former type of interventions may be considered more elementary in an intuitive physical sense. If multiple elementary interventions are combined, then this may in principle happen in a way such that they tuned to each other, and we would view this as violating a form of our independence Principle 2.1; see footnote 8 on page 24. One may hope that combined interventions that are "natural" will not violate independence. However, to tell whether an intervention is "natural" in this sense requires knowledge of the causal structure, which we do not have when trying to use such principles to perform causal learning in the first place. Ultimately, one can try to resort to physics to assay what is elementary or natural.

The questions of which operations on a physical system are elementary plays a crucial role in modern quantum information theory. There, the question is closely related to analyzing the structure of physical interactions.[13] Likewise, we believe that understanding physical mechanisms underlying causal relations may sometimes explain why some interventions are natural and others are complex, which essentially defines the "modules" given by the different structural equations.

2.3.5 Independence of Cause and Mechanism and the Thermodynamic Arrow of Time

We provide a discussion as well as a toy model illustrating how the principle of independent mechanisms can be viewed as a principle of physics. To this end, we

[13] For the interested reader: A system consisting of n two-level quantum systems is described by the 2^n-dimensional Hilbert space $\mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2$. Unitary operators acting on this Hilbert space correspond to physical processes. For several such systems, researchers have shown how to implement "basic" unitaries that act on at most two of the n tensor components [Nielsen and Chuang, 2000] and act trivially on the remaining $n-2$ ones. Then one can generate any other unitary [DiVincenzo, 1995] approximately by concatenation. Although this is by no means the only possible choice for the set of "basic" unitary operations, the choice seems natural given the structure of physical interactions.

Figure 2.4: Simple example of the independence of initial state and dynamical law: beam of particles that are scattered at an object. The outgoing particles contain information about the object while the incoming do not.

consider the special case of two variables and postulate the following as a specialization of Principle 2.1:

Principle 2.2 (Initial state and dynamical law) *If s is the initial state of a physical system and M a map describing the effect of applying the system dynamics for some fixed time, then s and M are independent. Here, we assume that the initial state, by definition, is a state that has not interacted with the dynamics before.*

Here, the "initial" state s and "final" state $M(s)$ are considered as "cause" and "effect." Accordingly, M is the mechanism relating cause and effect. The last sentence of Principle 2.2 requires some explanation to avoid erroneous conclusions. We now discuss its meaning for an intuitive example.

Figure 2.4 shows a scenario where the independence of initial state and dynamics is so natural that we take it for granted: a beam of n particles propagating in exactly the same direction are approaching some object, where they are scattered in various directions. The directions of the outgoing particles contain information about the object, while the beam of *incoming* particles does not contain information about it. The assumption that the particles initially propagate exactly in the same direction can certainly be weakened. Even if there is some disorder in the incoming beam, the outgoing beam can still contain information about the object. Indeed, vision and photography are only possible because photons contain information about the objects at which they were scattered.

We can easily time-reverse the scenario by "hand-designing" an incoming beam for which all particles propagate in the same direction *after* the scattering process. We now argue how to make sense of Principle 2.2 in this case. Certainly, such a