Automated Macro-scale Causal Hypothesis Formation Based on Micro-scale Observation.

Thesis by

Krzysztof Chalupka

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2017
(Submitted November 30, 2016)

To my wife Lulu and my son Lech

Acknowledgements

From my mother Barbara I learned to always push myself to the limits, and about honesty and decency.

From my Polish friends Artur, Filip, Jagoda, Kuba, Marek, Ula, Wojtek, Zuza, that camaraderie transcends time and space.

From Master Ian Cameron of Edinburgh that training is its own reward, and so is mastery; and that the only thing that always goes with the flow is the dead fish; a set of exercises that saved my heart and stomach; to appreciate depth of character and resolution; and about music creative yet unpretentious.

From the I Ching, to seek out the subtleties of reality without resort to pompous spirituality.

From Kun Zhang, to carefully remove snails into safety from the walkway during rainy nights.

From my Chinese family, that family can and should be at the center of one's life; and to be generous.

From Alex, to think about the other's mind before approaching them. From Aram, to fight against a larger and stronger opponent. From Bo, that cross-cultural friendship can be European. From Ron, I learned that there are people whose problem-solving skills far surpass mine. From Steve, I received an education in American friendship and how to respect it. From Tristan, I learned integrity and liberty.

From professor Yisong Yue, to mellow down (and not to write angry peer reviews).

From professor Doris Tsao, a passion for science, and a single-minded determination to reach a goal; that nothing is true by default, but only after I understand myself why it should be so.

From professor Frederick Eberhardt, to distrust any idea, thought or assumption, especially my own, until it is analyzed and broken into pieces and rebuilt in my own mind many times over.

From professor Pietro Perona, I learned to believe that there are still pioneers in science, and to will to be one of them; to be patient and kind to those without my experience or knowledge; to think clearly, and in terms of principles, goals, and people.

From my time at Caltech, I learned when doing anything to do it right; and that it is a mistake to strive for a mind purely analytical.

Abstract

This book introduces new concepts at the intersection of machine learning, causal inference and philosophy of science: the macrovariable cause and effect. Methods for learning such from microvariable data are introduced. The learning process proposes a minimal number of guided experiments that recover the macrovariable cause from observational data.

Mathematical definitions of a micro- and macro- scale manipulation, an observational and causal partition, and a subsidiary variable are given. These concepts provide a link to previous work in causal inference and machine learning.

The main theoretical result is the Causal Coarsening Theorem, a new insight into the measure-theoretic structure of probability spaces and structural equation models. The theorem provides grounds for automatic causal hypothesis formation from data. Other results concern the minimality and sufficiency of representations created in accordance with the theorem.

Finally, this book proposes the first algorithms for supervised and unsupervised causal macrovariable discovery. These algorithms bridge large-scale, multidimensional machine learning and causal inference. In an application to climate science, the algorithms re-discover a known causal mechanism as a viable causal hypothesis. In a psychophysical experiment, the algorithms learn to minimally change visual stimuli to achieve a desired effect on human perception.

Contents

List of Theorems

Acronyms

CCT Causal Coarsening Theorem

CFL Causal Feature Learning

eda electrodermal activity

HDL high-density lipoprotein

LDL low-density lipoprotein

SST sea surface temperature

ZW zonal wind strength

Chapter 1

Introduction

During my time at Caltech, I developed — together with Pietro Perona and Frederick Eberhardt — the theory and algorithms that aim at solving a previously open problem in machine learning and causal inference. Concisely stated, the problem is as follows. Learn, by looking at low-level measurements, a maximally compressed representation of the causal mechanisms underlying these measurements. Part of this book is dedicated to defining the mathematical apparatus necessary to approach this question. This leads to the first algorithms solving the task, both in settings with and without supervision. The algorithms extract features of the data that are, in a well-defined sense, *causal* features: changing the underlying data has an effect on a system of interest only inasmuch as the value of the causal features changes. Hence the name of the framework, CFL.

Much of this book contains results and in some cases text passages from four peer-reviewed publications I am a co-author of (Chalupka et al., 2015, 2016b,a,c). Chapter 6 contains new material available online but not peer-reviewed (Chalupka et al., 2016d). Computer programs that implement our algorithms and reproduce some of the experimental results presented in this book is available online at `http://vision.caltech.edu/~kchalupk/code.html`.

1.1 Causal Feature Learning

CFL is a machine learning and causal inference framework with two goals:

1. Formation of high-level causal hypotheses using low-level input data, and
2. Efficient testing of these hypotheses.

As a motivation, consider the following archetypical research situation (Fig. 1.1C): a neuroscientist notices that a specific neuron responds preferentially to some images containing humans. The scientist progressively refines and tests this hypothesis by exploring painstakingly the effect of different poses and occlusions of a large number of human shapes on the neuron. These experiments suggest that the neuron responds specifically to images of female faces. This conclusion is based on alternating three main steps: (a) formulating hypotheses through modeling and intuition, (b) designing experiments to test such hypotheses and (c) col-

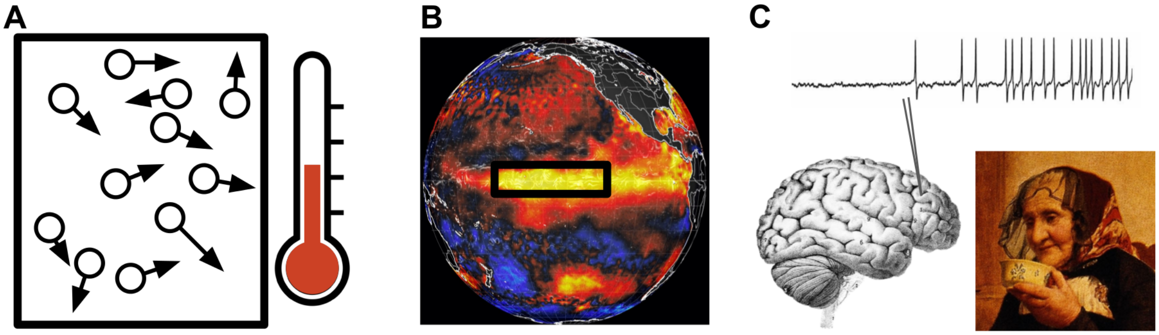


Figure 1.1: **Causal Macrovariables**. Macrovariables in science are functions of the underlying microvariable space. Each such function f corresponds to a *partition* on the microvariable state space, defined by the equivalence relation $x_1 \sim x_2 \iff f(x_1) = f(x_2)$. (A) Temperature may be defined as the mean kinetic energy of a system of particles. It is a one-dimensional function of a high-dimensional system consisting of a large number of particles, each one with a mass and velocity. (B) El Niño is defined as the sea surface temperature (SST) anomaly in a specific region of the Pacific Ocean exceeding 0.5^0C. It is a binary function of the high-dimensional sea surface temperature (SST) map. (C) Primate brains are thought to have areas specialized for face detection (see Tsao et al. (2006) for direct evidence in the macaque cortex). "Presence of a face" is a binary function on the space of all images.

lecting evidence from such experiments.

Steps (a) and (b) are guided by prior knowledge, intuition and formal reasoning. CFL aims to augment or fully automate this process in situations where observational data is plentiful, reducing the bias resulting from pre-conceived ideas of the scientist. The method is predicated on the idea that if the data in fact contains high-level features (such as female faces) that are causal, then these ought to be detectable by a learning algorithm. In addition, CFL distinguishes between features that are related by direct *causation* from features that are related through common causes. For example, atmospheric pressure causes the needle of the barometer the change, but the needle's position is neither a cause nor an effect of rainfall, with which it nevertheless strongly correlates.

1.2 Macrovariables in Science

In CFL, the distinction between high-level and low-level features is framed in terms of macrovariables and microvariables, terms often used in physical sciences. The semantics of these terms as used in science provide direct inspiration for our methods.

Just about any scientific discipline is concerned with developing 'macrovariables' that summarize an underlying finer-scale structure of 'microvariables' (see Fig. 1.1A-C). Temperature and pressure summarize the particles' masses and velocities in a gas at equilibrium; large-scale climate phenomena, such as El Niño, supervene on the geographical and temporal distribution of sea surface temperature (SST) and wind speed. Similarly, for the human sciences: Macro-economics supervenes on the economic activities of individuals, which in turn presumably summarize the psychological processes of each person, which are aggregates of neural states.

These abstractions are particularly useful when one can establish causal relations amongst macrovariables that hold independent of the micro-variable instantiations of the macrostates. For example, it is useful to propose that “El Niño is caused by strong westerly winds”. CFL is motivated by the need to automate the process of developing such hierarchical descriptions starting from the less-constrained space of microvariables. The key insight is that it is best to discover simultaneously the macrovariables and their causal relations. CFL thus searches for the macrovariable cause/effect hypotheses starting from microvariable data. Any random variable with a large, possibly infinite, number of states may be considered a microvariable. Continuous variables, as well as discrete variables with unmanageable numbers of states (such as digital images or spin configurations in the Ising model) are microvariables.

In science, macrovariables often correspond to equivalence relations on the microvariable state-space. For example, all the particle ensembles with the same mean kinetic energy correspond to the same temperature. Similarly, all the SST maps where the temperature anomaly in a specified region of the Pacific Ocean exceeds 0.5°C correspond to El Niño. Following this intuition, CFL defines the relation between micro- and macrovariables in terms of an equivalence relation, which we review formally in Chapters 2 and 3.

The learning task of CFL may be framed in terms of the micro- and macrovariable distinction:

1. Take two observational — that is, “sampled by nature”, not-experimental (Pearl, 2000, 2010)— microvariable datasets $\mathcal{L}$ and $\mathcal{R}$ as input, with the task of discovering “what in $\mathcal{L}$ causes what in $\mathcal{R}$.”
2. Search the space of all macrovariables (equivalence relations) on $\mathcal{L}$ and retain only those that could be causes of $\mathcal{R}$.
3. Search the space of all the macrovariables that supervene on $\mathcal{R}$ and retain only those that could be effects of $\mathcal{L}$.
4. Propose an efficient procedure that picks out the (unique) macrovariable cause and effect among the retained macrovariable pairs.

In general, there is an infinite number of macrovariables defined over two given microvariable spaces. However, not every random variable can function as a causal variable. First of all, causal variables cannot stand in logical or definitional relations to one another – X does not cause $2X$. Furthermore, causal variables should permit well-defined experimental interventions. This latter point raises a subtle but important issue for CFL: ambiguous manipulations.

1.2.1 Ambiguous Manipulation and Causal Macrovariables

Figure 1.2 illustrates a case of an unfortunate choice of a macrovariable. Total cholesterol used to be considered a risk factor for heart disease. However, further analysis revealed that ‘total cholesterol’ is not a good causal variable, since it is a sum of cholesterol carried by low-density lipoprotein (LDL) and high-density lipoprotein (HDL), commonly called “good” and “bad” cholesterol, which have different effects on heart disease (see Spirtes and Scheines (2004) for an in-depth discussion of this case.)

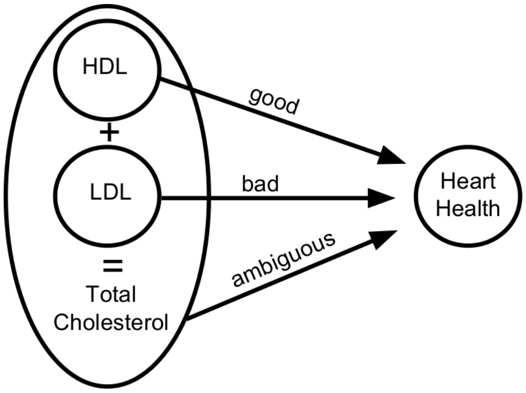


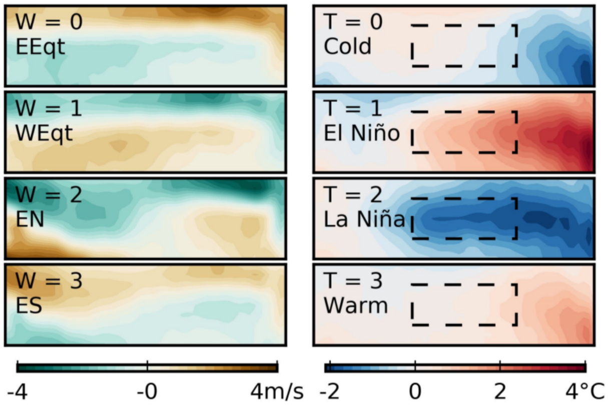
Figure 1.2: **Ambiguous Manipulation.** Total cholesterol is the sum of LDL and HDL. Suppose that LDL causes heart disease and HDL prevents it. The effect of total cholesterol on heart disease is then ambiguous as it depends on the proportion of HDL vs. LDL (see Sec.1.2.1). Experimental procedures based on adjusting total cholesterol only can give inconsistent results.

Consequently, to recommend a "low-cholesterol diet" is to prescribe an *ambiguous manipulation*: "low-cholesterol" could mean low in LDL, HDL or both, but each would have very different consequences for the heart. Unless the proportions of LDL vs. HDL are known in advance, this makes a proper experimental verification of the causal link between total cholesterol and heart disease impossible. The example illustrates that there exists an appropriate "ground-truth" level of aggregation to describe the causal relation, and "total cholesterol" is too high-level. The challenge is to identify when one has reached the correct level.

CFL addresses this concern, and requires causal variables to be unambiguous: Each macrovariable state must have a consistent, well-defined causal effect. This effect can be probabilistic and highly variable, but must not depend on the microvariable instantiation of the macrovariable. For just like the specifics of gas molecule momenta do not change the effects of temperature, as long as their mean is equal. In this way, CFL abstracts microscopic details of the problem away, allowing the scientist to focus on all the relevant macroscopic details. This is analogous to the role of the macrovariables in Fig. 1.1 (A–C).

1.2.2 Macrovariables Are Task-Specific

Although pre-theoretic intuition may suggest that there is some uniquely true taxonomy to the variables describing the world, we reject this view and propose that macrovariables should be thought of as task-specific. For example, there is evidence that the human visual system parses the image in terms of macrovariables that (among other things) track the location, shape and appearance of faces in the scene. However, there is no *a priori* reason that these are 'optimal' visual variables. To other creatures, occupying a different ecological niche and animated by different behavioral goals, a different grouping of visual information may be relevant. For example, an insect might be far more concerned about luminance, edges and motion flow in its visual input than about objects and faces. Thus, the appropriate equivalence relation on the microvariable state space is driven by the *relation* between $\mathcal{L}$ and $\mathcal{R}$ (for example, the statistics of the environment as imaged by the



P(T | W)	Cold	El Niño	La Niña	Warm
EEqt	2/3	0	1/3	0
WEqt	0	1/4	0	3/4
EN	∼1/10	0	1/4	∼2/3
ES	3/4	0	0	1/4

Figure 1.3: **Macrovariables of Pacific Weather Patterns**. A preview of the results of applying CFL to climate data in Chapter 5. The microvariables consist of zonal (East-West) wind strength over the equatorial Pacific and SST maps over the same region. The figure shows the causal hypothesis discovered by CFL. Each image represents one macrovariable state, the average over one cluster of wind W (left) or temperature T (right). The conditional probability table shows $P(T \mid W)$, the probability of the hypothesized SST macrovariable given the hypothesized wind macrovariable. It shows that CFL learned at least two relations that, causally interpreted, are consistent with current climate science: 'Westerly winds' (W=1) cause El Niño (T=1) and 'Easterly winds' (W=0) cause La Niña (T=2).

optic array, and the desired behavior), rather than by one or the other of the spaces considered individually.

As an example, take 'wind strength map over the Pacific Ocean' as the input space, and 'SST' as the output space. Applying CFL to this task yields a discrete division of each space into a set of wind pattern classes ('Westerly Winds', 'Easterly Winds' etc) and SST pattern classes ('El Niño', 'La Niña' etc) – Chapter 5 describes the experiment in detail. Knowing which class a wind pattern belongs to then gives all the useful information about its *possible* effects on SST[1]. For a different output space – say "average US income" – the input macrovariable would change, unless the causal consequences are entirely mediated by the same wind patterns.

1.3 Example Microvariable Cause-Effect Systems

The current section describes four example CFL problems. Further chapters develop the theory and algorithms necessary to solve these problems, and present experimental results. This section has two goals:

1. Clarify, by example, when CFL is useful, and
2. provide a guide to the contents of this book.

By necessity, most of our experiments are done on simulated systems. The reason is that the "causal ground-truth" can only be obtained either by definition of the system, or through thorough experimentation.

[1] As discussed throughout the book, a causal interpretation of purely observational data is not possible without further assumptions.

The latter is an expensive and time-consuming in most interesting cases. Nevertheless, Chapter 5 contains a limited application of the framework to real data where the ground-truth is given by expert opinion.

1.3.1 Images and a Spiking Neuron

Fig. 1.4 presents a cartoon of a paradigmatic case study in visual CFL. The contents of an image I are caused by external, non-visual binary hidden variables H_1 and H_2 such that if H_1 is on, I contains a vertical bar (v-bar[2]) at a random position, and if H_2 is on, I contains a horizontal bar (h-bar) at a random position. A target behavior $T \in \{0, 1\}$ is caused by H_1 and I, such that $T = 1$ is more likely whenever $H_1 = 1$ and whenever the image contains an h-bar. T could indicate, for example, whether a particular neuron in the human brain significantly exceeds its baseline spiking rate within 500ms after viewing the image.

This example is deliberately constructed such that the visual cause is clearly identifiable: manipulating the presence of an h-bar in the image will influence the distribution of T. Thus, we can call the following function $C\colon \mathcal{I} \to \{0, 1\}$ the *causal feature* of I or the *macrovariable cause* of T:

$$C(I) = \begin{cases} 1 & \text{if } I \text{ contains an h-bar} \\ 0 & \text{otherwise.} \end{cases}$$

The presence of a v-bar, on the other hand, is not a causal feature. Manipulating the presence of a v-bar in the image has no effect on H_1 or T. Still, the presence of a v-bar is as strongly correlated with the value of T (via the common cause H_1) as the presence of an h-bar is. Call the following function $S\colon \mathcal{I} \to \{0, 1\}$ the *spurious correlate* of T in I:

$$S(I) = \begin{cases} 1 & \text{if } I \text{ contains a v-bar} \\ 0 & \text{otherwise.} \end{cases}$$

Both the presence of h-bars and the presence of v-bars are good individual (and even better joint) predictors of the target variable, but only one of them is a cause. Identifying the visual cause from the image thus requires the ability to distinguish among the correlates of the target variables those that are actually causal, even if the non-causal correlates are (possibly more strongly) correlated with the target.

Chapter 2 defines rigorously what it means to be a macrovariable cause and spurious correlate in a general setting. It provides theory and algorithms for optimal experimental design to differentiate the two. Chapter 4 describes a method to learn a manipulator function. The manipulator takes a microvariable input (for example, an image with a horizontal and vertical bar in it, as well as other, causally irrelevant, structure) and constructs the *closest possible* (according to some metric) image that has a different causal effect. In our example, a perfect manipulator would remove only one pixel from the (causally relevant) h-bar to remove this feature of the image, but would leave any v-bars intact (since they are not causal features). As discussed in Chapter 4, manipulator functions make CFL useful in contexts where the goal is not only to understand

[2] We take a v-bar (h-bar) to consist of a complete column (row) of black pixels.

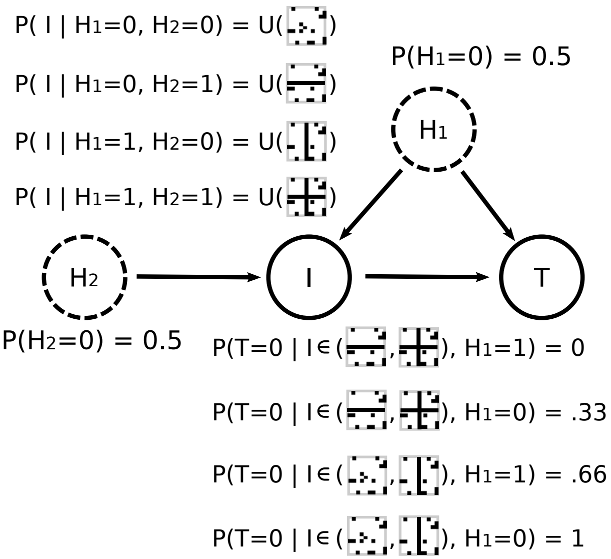

P(I | H1=0, H2=0) = U( )
P(I | H1=0, H2=1) = U( )
P(I | H1=1, H2=0) = U( )
P(I | H1=1, H2=1) = U( )
P(H1=0) = 0.5
H1
H2
I
T
P(H2=0) = 0.5
P(T=0 | I∈(,), H1=1) = 0
P(T=0 | I∈(,), H1=0) = .33
P(T=0 | I∈(,), H1=1) = .66
P(T=0 | I∈(,), H1=0) = 1

Figure 1.4: **A Toy Causal Model of Visual Features Activating a Single Neuron.** Two binary hidden (non-visual) variables H_1 and H_2 toss unbiased coins. These variables represent random events in the world, e.g. H_1 could mean “There is a tree nearby”. The content of the image I depends on these variables as follows. If $H_1 = H_2 = 0$, I is chosen uniformly at random from all the images containing no v-bars and no h-bars. If $H_1 = 0$ and $H_2 = 1$, I is chosen uniformly at random from all images containing at least one h-bar but no v-bars. If $H_1 = 1$ and $H_2 = 0$, I is chosen uniformly at random from all the images containing at least one v-bar but no h-bars. Finally, if $H_1 = H_2 = 1$, I is chosen from images containing at least one v-bar and at least one h-bar. The distribution of the binary behavior T depends only on the presence of an h-bar in I and the value of H_1. In observational studies, $H_1 = 1$ iff I contains a v-bar. However, a *manipulation* of any specific image $I = i$ that introduces a v-bar (without changing H_1) will in general not change the probability of T occurring. Thus, T does *not* depend causally on the presence of v-bars in I.

causal mechanisms of a system, but also manipulate the system efficiently. For example in healthcare, the desire to understand the relationship between the human body and its environment is driven by the underlying goal of *intervening* on the environment in order to improve health.

1.3.2 Hue and Skin Conductance

In the previous example, input images consisted of microvariables, but the output was a binary macrovariable. Sec. 1.2, however, motivated this work with many examples in which both the cause and the effect supervene on microvariables. In such cases it is not possible to "anchor" the discovery of causal features in a well-defined effect – both the cause and the effect have to be learned jointly. I call this the unsupervised CFL setting, in contrast to the supervised CFL described above.

A simple toy example will visualize the definitions and main algorithmic steps involved in unsupervised CFL (we return to this example in Chapter 3). Take a fictitious study on the influence of color on the electrodermal activity (eda) (also known as the skin conductance). In a fictitious experiment, the electrodermal response to a (constant but unspecified) stimulus is recorded in varying environments. At the same time, the predominant hue of the environment is recorded. Our simulated system is pictured in Fig. 1.5. In the system, "Red" hues increase eda (a perhaps controversial but plausible response, see e.g. Jacobs and Hustmyer (1974)). In addition, living in warmer climates increases eda, but also increases the chance of observing "Warm" colors in the environment. Our imaginary study consists of picking humans from diverse populations at random, and measuring their eda as well as the predominant hue in their environment. The example is set up to exhibit three characteristics:

1. The microvariables (hue and eda) are one-dimensional. Although this makes the example rather contrived, the visualizations of the algorithms and definitions are much simpler and more illuminating than in higher-dimensional cases.
2. Microvariable hue gives rise to intuitive macrovariables: color classes. "Red" colors, "Natural" colors or "Warm" colors are (subjective) partitions of the hue space, and clearly *supervene* on hue. For example, "Red" is not *caused* by hue, it is simply a range in the hue space.
3. The cause (hue) influences the effect (eda) by direct causation, but they are at the same time confounded by geographic location. The goal of CFL is to separate the causal information and the purely-confounded information, and compress each into a separate macrovariable.

In our model, eda (in units normalized to $(0, 1)$ where $.5$ is the global average) is causally influenced by hue (represented in degrees, with 0 being the red hue, see Fig. 1.5) and lat (geographic latitude, a one-dimensional proxy for "climate" for ease of visualization). Among these microvariables, only hue and eda are observed and lat is latent. Thus, hue *causes* eda and at the same time the two variables are *confounded*, as illustrated in Fig. 1.5. Assuming that lat fully captures the causal confounding between hue and eda, their joint distribution $p(hue, eda)$ factorizes as

$$p(hue, eda) = \sum_{lat} p(eda \mid hue, lat) p(eda \mid lat) p(lat). \tag{1.1}$$

The probability tables for these factors are shown in Fig. 1.5. We purposefully constructed the conditional distribution $p(eda \mid hue, lat)$ to take a special form: there are four ranges of hue within which $p(eda \mid hue)$ is constant. For example, the conditional is the same for any $hue \in (0, 90)$. This construction indicates that there are *macrovariables* driving the relation between hue and eda: to a good approximation, any hue within a given range has the same effect on eda. The situation is analogous to that of the temperature macrovariable driving the relation between, say, water and human pain receptors. The probability and intensity of experiencing pain is roughly the same upon touching any body of water with the same temperature – given that all the other relevant variables, such as the individual experiencing pain, remain unchanged.

Chapter 3 shows that in unsupervised CFL tasks such as the one described in this section, causal macrovariables are unique and can be extracted automatically from the data. For now, we propose a "ground truth" macrovariable model that agrees with the microvariable distribution shown in Fig. 1.5. Chapter 3 shows that this model is in fact *the* macrovariable structure that supervenes on hue and eda and can be automatically discovered using CFL.

The macrovariables are all binary. A supervenes (is a function of) eda, with $A = 1$ if and only if $eda > .5$. That is, A represents an "Above-average" skin conductance. A is caused by R ("Redness") which supervenes on hue, $R = 1 \iff hue \in (0, 90) \cup (270, 360)$. In addition, A *correlates with*, but is not *caused by*, W – another variable that supervenes on hue, $W = 1 \iff hue \in (0, 180)$. W represents "Warm" hues.

The causal graph of R, W and A, shown in Fig. 1.5, is determined by the variables' supervenience on hue and eda. Similarly, the joint probability distribution $P(R, W, A)$ is fully determined by $p(hue, eda)$. Algorithm 3 shows how to recover the variables R and A – the causally relevant variables – through data-driven experimental design.

1.3.3 Images and Neural Populations

Whereas the above simulated dataset is simple and low-dimensional, unsupervised CFL can be applied to very high-dimensional and complex data. The current example is partially inspired by a problem at the core of much of modern neuroscience that is a generalization of the problem presented in Sec. 1.3.1: Can we detect which features of a visual stimulus result in particular responses of *neural populations* without pre-defining the stimulus features or the types of population response we are interested in?

For example, Rutishauser et al. (2011) analyze data from multiple electrodes implanted in the human amygdala. The patient is asked to look at images containing either whole human faces, faces randomly occluded with Gaussian "bubbles", or images of specific regions of interest in the face—say the eye or the mouth. The neurons are then sorted according to whether they are full-face selective or not, and the response properties of the neurons are analyzed in the two populations. This set-up is an instance of a widely used

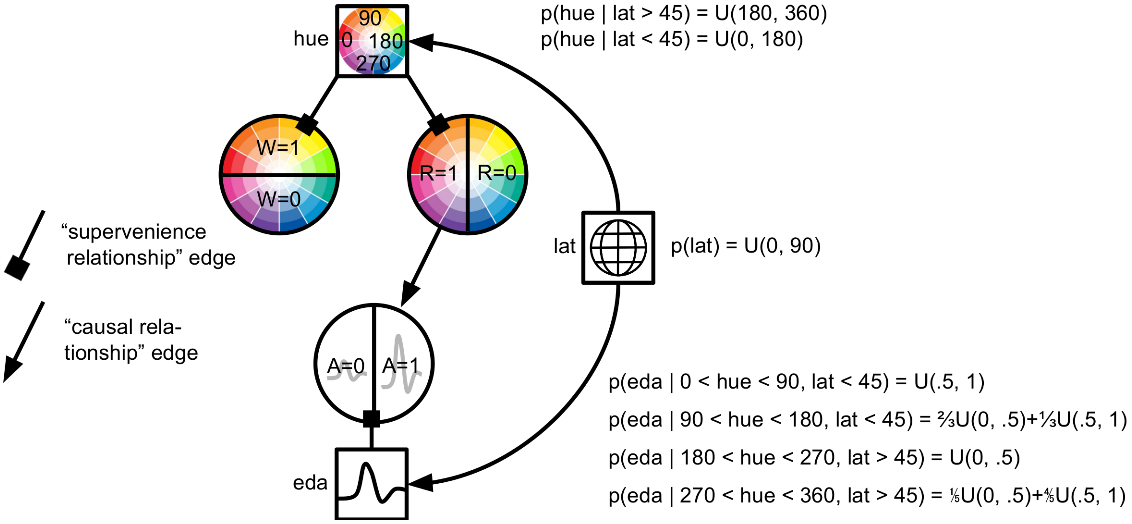

hue
90
0
180
270
p(hue | lat > 45) = U(180, 360)
p(hue | lat < 45) = U(0, 180)
W=1
W=0
R=1
R=0
"supervenience relationship" edge
"causal relationship" edge
lat
p(lat) = U(0, 90)
A=0
A=1
eda
p(eda | 0 < hue < 90, lat < 45) = U(.5, 1)
p(eda | 90 < hue < 180, lat < 45) = ⅔U(0, .5)+⅓U(.5, 1)
p(eda | 180 < hue < 270, lat > 45) = U(0, .5)
p(eda | 270 < hue < 360, lat > 45) = ⅕U(0, .5)+⅘U(.5, 1)

Figure 1.5: **Causal Graphical Model of Color Influencing eda.** In the simulated study, the predominant hue of the environment causes changes in the electrodermal response: red hues increase eda beyond the average, whereas non-red hues tend to decrease it. In addition, the latitude of the experiment influences eda: lower latitudes, close to the equator, cause higher absolute eda due to predominantly warm climate. Lower latitude visual environments tend also to have visually warmer hues, whereas higher latitude environments often have cooler hues. The probability tables show generative probabilities for our data, where $U(a, b)$ is the uniform distribution between a and b. For example, if $lat > 45$ and $270 < hue < 360$, then $p(eda) = .2U(0, .5) + .8U(.5, 1)$ – a mixture of two uniform distributions that indicates that most likely, eda is above the average in this situation.

experimental protocol in the field: prepare stimuli that represent various hypotheses about what the neurons respond to; record from single or multiple units; and analyze the responses with respect to the candidate hypotheses.

But what if the candidate hypotheses are wrong? Or if they do not line up cleanly with the actually relevant features? CFL offers a less biased and more automatized process of experimentation: Record neural population responses to a broad set of stimuli. Then, jointly analyze what features of the stimuli modify responses of the neural population *and* what features of neural activity are changing in response to the stimuli. To our knowledge, such joint cause-and-effect learning is a novel contribution not only in the neuroscientific setting, but to a whole array of other scientific disciplines.

A simple neural population simulation provides a motivating example (again, we resort to simulation to be able to compare our results to the *ground-truth* causal mechanisms). This example differs from the two above in that *there is no confounding* in the system: we will assume that visual stimuli influence neural behavior directly, and there are no common causes between the two. Fig. 1.6 illustrates a simulation of a population of 100 neurons whose dynamics follow Izhikevich's equations (Izhikevich, 2003). The equations are designed to mimic the behavior of human cortical neurons. As the ground-truth structures of interest, we define simple macro-level causes and effects: Presented with an image containing a horizontal bar (h-bar), the "top half" of the neural population produces a pulse of joint activity after about 100ms. When presented with a vertical

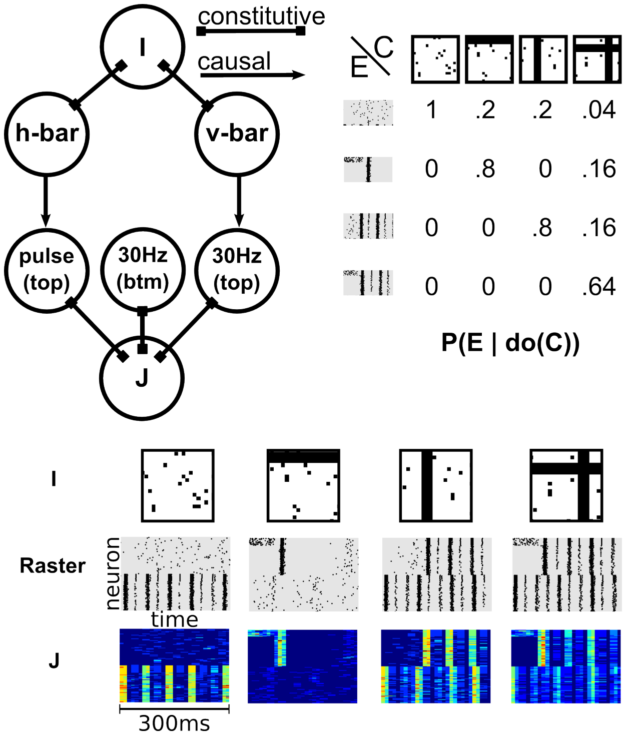


Figure 1.6: A simulated neuroscience experiment. A stimulus image I can contain a horizontal bar (h-bar), a vertical bar (v-bar), neither, or both (plus uniform pixel noise). In response to an image, a simulated population of neurons (the "top" population) can produce a single pulse of joint activity, a 30 Hz rhythm, both, or neither, with probabilities $P(\text{pulse} \mid \text{h-bar}) = 0.8$ and $P(30\text{Hz} \mid \text{v-bar}) = 0.8$. These two causal mechanisms compose to yield the full response probability table shown in top right. In addition, another ("bottom") population of neurons can exhibit a rhythmic activity independent of the stimulus image. The system's output J is a 10ms-window running average of the neural rasters, with the neuron indices shuffled (as a neuroscientist has no a-priori knowledge of how to order neurons). Here we show example J's sorted by neuron id; we use the shuffled version in our experiments.

bar (v-bar), the same population synchronizes in a 30Hz rhythm after roughly the same delay. The remaining ("bottom half") population acts independently of the visual stimuli (perhaps the experimenter unwittingly planted some of the electrodes in a non-visual brain area). Half the time these "distractor neurons" follow their spontaneous noisy dynamics, and half the time they synchronize to produce a rhythmic activity. One can think of this activity as being caused by internal network dynamics, extra-visual stimuli or any other cause, as long as it is independent of the image presented by the experimenter.

The example is made up of deliberately simple features for ease of illustration and interpretation. Nevertheless, it hints at what makes similar problems non-trivial to solve. The causal features can be convoluted with salient probabilistic structure (such as the rhythmic behaviors generated in the "bottom" neuronal population). Moreover, the data and its features can be difficult to interpret directly "by looking": after reshuffling the neural indices, the raster plots are hardly distinguishable by the human eye. In many domains (e.g. in finance) the data have no special spatial structure in the first place, since they can consist simply of rows of numbers. Chapter 3 shows how CFL can be applied in such domains, and how it solves this simulated problem.

1.4 Related Work

Our framework draws heavily on ideas developed in computational mechanics (Shalizi, 2001; Shalizi and Crutchfield, 2001; Shalizi and Moore, 2003) and connects them with the framework of causal graphical models (Spirtes et al., 2000; Pearl, 2000). After discussing these two frameworks, this section briefly indicates several other related areas of machine learning and information theory.

1.4.1 Computational Mechanics

Our approach derives its theoretical underpinnings from the theory of computational mechanics (Shalizi, 2001; Shalizi and Crutchfield, 2001). In particular, computational mechanics defines macrovariable states in terms of equivalence classes of conditional probabilities. Definition 5 from Cosma Shalizi's PhD dissertation (Shalizi, 2001) is in fact equivalent to our definition of the observational state (Definition 1 in this book).

However, in computational mechanics macrovariables stop at the level of conditional probabilities and are meant to 'summarize' the phenomena rather than to support causal reasoning. Our work supports an explicitly causal interpretation by incorporating the possibility of confounding and interventions. We take the distinction between interventional and observational distributions to be one of the key features of a causal analysis. Thus, our Def. 1 is just a first step, leading later to the development of the Causal Class and the Causal Coarsening Theorem that relates observation and intervention.

1.4.2 Causal Graphical Models

The framework of causal graphical models (Spirtes et al., 2000; Pearl, 2000) provides the grounds for our understanding of causality. In this framework, X causes Y if $P(Y \mid \mathrm{do}(X)) \neq P(Y)$, where do() is an operator representing a randomized experiment. That is, one variable causes another *if and only if* after intervening on the first we see a (probabilistic) change in the value of the latter – while all the other relevant factors are kept constant. This definition captures a wide range of intuitions about causality, for example:

1. Atmospheric pressure *causes* the position of the needle of a barometer. This is because changing the atmospheric pressure would have an effect on the needle. However, the needle's position is not a cause of atmospheric pressure. Tampering with the barometer will never cause a change in air pressure.
2. Whether it rains or not *correlates with*[3] the position of the barometer's needle, but there is no direct causal relation between the two. This is because they have a *common cause*, the atmospheric pressure.

[3] Throughout this book we will often use the expression "x correlates with y" to mean that the two variables are probabilistically dependent, that is $P(x, y) = P(x)P(y)$.

3. Position of a barometer's needle is neither a cause nor effect of a probabilistically unrelated variable such as the outcome of a coin toss.

The field of causal graphical models concerns itself mainly with two tasks: 1) discovering the causal relationships between probabilistic variables (also called "learning the causal graph"), and 2) given a causal graph, inferring what effect a specific set of interventions will have on a chosen set of variables (classical approaches to these two problems are discussed broadly by Spirtes et al. (2000) and Pearl (2000)).

For our purposes, the above definition of causality is sufficient and we need not discuss causal discovery and inference further. We wish to remark, however, that the standard causal graphical models setting presupposes that the relevant (macro)variables are given together with the problem specification. In contrast, in our setting the causal variables have to be constructed from the micro-variables they supervene on, before any causal relations can be established. This work is, as far as we know, the first attempt to construct meaningful causal variables from scratch, within the causal graphical models framework. We emphasize the difference between our method of causal feature *learning* and methods for causal feature *selection* (Guyon et al., 2007; Pellet and Elisseeff, 2008). The latter choose the best (under some causal criterion) features from a restricted set of plausible macro-variable candidates. In contrast, our framework efficiently searches the whole space of all the possible macro-variables that can be constructed from an image.

1.4.3 Machine Learning and Artificial Intelligence

David MacKay ties together the fields of machine learning and information theory as follows (MacKay, 2003):

> Why unify information theory and machine learning? Because they are two sides of the same coin. In the 1960s, a single field, cybernetics, was populated by information theorists, computer scientists, and neuroscientists, all studying common problems. Information theory and machine learning still belong together. Brains are the ultimate compression and communication systems.

If machine learning is concerned with compression, the field of Artificial Intelligence (AI) is about how to use compression to act. Archetypical machine learning algorithms of the past two decades (neural networks (Bishop, 1995), Support Vector Machines (Schölkopf and Smola, 2001), Gaussian processes (Williams and Rasmussen, 2006)) can compress complex data all the way to the level of discrete labels or rid time-series of information irrelevant to a given predictive task. Thanks largely to the flexibility of neural networks (Schmidhuber, 2015) and their commercialization however, we now see a revival of the desire to build agents endowed with the ability to act, at least in games (Silver et al., 2016; Mnih et al., 2013) – though the idea is certainly not new (Russell et al., 2003). Modern intelligent agents depend on machine learning when compressing information about their environment before acting.

The field of causal inference is to a large degree inspired by the desire to create intelligent acting systems (Pearl, 1995). Clearly, an agent would be well-advised to try and predict the results of its own interventions in the world, as well as the interventions of other agents. We see our work as bridging causal inference and modern machine learning. Our work is all about how to *compress* stimuli in order to get the most concise representation of the *causal mechanisms* relevant to a given task, and how to *act* or *intervene* in the world in an optimal manner, given that only low-level direct measurements are available. We use machine learning to create concise causal representations that can be directly used by intelligent agents to act in the world.

Active learning and automatic experimental design (see for example (Chaloner and Verdinelli, 1995; Tong and Koller, 2001; Srinivas et al., 2010; Snoek et al., 2012)) share CFL's goal of decreasing experimental effort in discovering causal mechanisms in the world. CFL and active learning are applicable in complementary situations. CFL serves its purpose best if microvariable observational data is easy to obtain and/or we suspect the presence of macrovariables driving the system. Active learning is applicable to macro-level, experimental data.

Chapter 2

Supervised Causal Feature Learning

This chapter develops the theory of supervised CFL within the context of *visual* causes. This setting makes the definitions most intuitive and is itself of significant practical interest. A visual cause is defined (more formally below) as a function (or *feature*) of raw image pixels that has a *causal effect* on a well-defined target behavior of a perceiving system of interest. However, the framework and results can be equally well applied to extract causal information from any aggregate of micro-variables on which manipulations are possible. Examples include auditory, olfactory and other sensory stimuli; high-dimensional neural recordings; market data in finance; consumer data in marketing. There, causal feature learning is both of theoretical ("What is the cause?") and practical ("Can we automatically manipulate it?") importance.

Visual perception is an important trigger of human and animal behavior. The visual cause of a behavior can be easy to define, say, when a traffic light turns green, or quite subtle: apparently it is the increased symmetry of features that leads people to judge faces more attractive than others (Grammer and Thornhill, 1994). Significant scientific and economic effort is focused on visual causes in advertising, entertainment, communication, design, medicine, robotics and the study of human and animal cognition. Visual causes profoundly influence our daily activity, yet our understanding of what constitutes a visual cause lacks a theoretical basis. In practice, it is well-known that images are composed of millions of variables (the pixels) but it is functions of the pixels (often called 'features') that have meaning, rather than the pixels themselves.

2.1 Advances in This Chapter

This chapter presents the following advances in machine learning and causal inference:

- A definition of the visual cause of a target behavior as a macro-variable that is constructed from the micro-variables (pixels) that make up the image space. The visual cause is distinguished from other macro-variables in that it contains all the causal information about the target behavior that is available in the image. The visual cause is defined within the standard framework of causal graphical models (Spirtes et al., 2000; Pearl, 2000), thereby contributing to an account of how to construct causal variables.

- The Causal Coarsening Theorem (CCT), which shows how observational data can be used to learn the visual cause with minimal experimental effort. CCT provides a connection between state-of-the-art machine learning methods for classification and causal discovery and experimental design.
- Algorithms to learn the visual cause from data and with minimal resort to experimentation.

2.2 Theory

Consider again the example from Sec. 1.3.1. There, the visual cause is identified with the presence of an h-bar. But the example does not provide a theoretical account of what it takes to be a visual cause in the general case when we do not know what the causally relevant pixel configurations are. In this section, we provide a general account of how the visual cause is related to pixel data.

2.2.1 Visual Causes as Macro-variables

A visual cause is a high-level random variable that is a function (or feature) of the image, which in turn is defined by the random micro-variables that determine the pixel values. The functional relation between the image and the visual cause is, in general, surjective, though in principle it could be bijective. While we are interested in identifying the visual causes of a target behavior, the functional relation between the image pixels and the visual cause should not itself be interpreted as causal. Pixels do not *cause* the features of an image, they *constitute* them, just as the atoms of a table constitute the table (and its features). The difference between the causal and the constitutive relation is that the former requires the possibility of independent manipulation (at least to some extent), whereas by definition one cannot manipulate the visual cause without manipulating the image pixels.

The probability distribution over the visual cause is induced by the probability distribution over the pixels in the image and the functional mapping from the image to the visual cause. But since a visual cause stands in a constitutive relation with the image, we cannot without further explanation describe interventions on the visual cause in terms of the standard do-operation (Pearl, 2000). Our goal will be to define a macro-variable C, which contains all the causal information available in an image about a given behavior T, and define its manipulation.

To make the problem approachable, we introduce two (natural) assumptions about the causal relation between the image and the behavior: (i) The value of the target behavior T is determined subsequently to the image in time, and (ii) the variable T is in no way represented in the image. These assumptions exclude the possibility that T is a cause of features in the image or that T can be seen as causing itself.

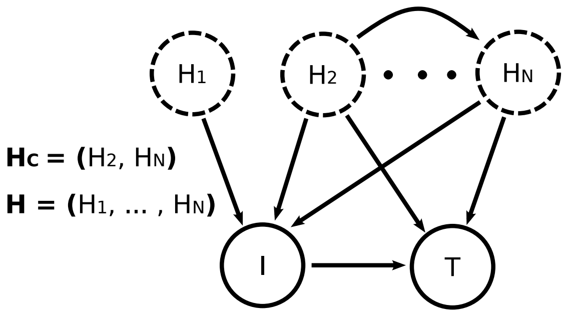


Figure 2.1: A general model of visual causation. In our model each image I is caused by a number of hidden non-visual variables H_i, which need not be independent. The image itself is the only observed cause of a target behavior T. In addition, a (not necessarily proper) subset of the hidden variables H_C can be a cause of the target behavior. These confounders create visual "spurious correlates" of the behavior in I.

2.2.2 From Micro- to Macro-variables

Let $T \in \{0, 1\}$ represent a target behavior.[1] Let $\mathcal{I}$ be a discrete space of all the images that can influence the target behavior (in our experiments in Section 2.4, $\mathcal{I}$ is the space of n-dimensional black-and-white images). We use the following generative model to describe the relation between the images and the target behavior: An image is generated by a finite set of unobserved discrete variables $H_1, \ldots, H_m$ (we write $\mathbf{H}$ for short). The target behavior is then determined by the image and possibly a subset of variables $\mathbf{H}_c \subseteq \mathbf{H}$ that are confounders of the image and the target behavior:

$$\begin{aligned} P(T, I) &= \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}) P(I \mid \mathbf{H}) P(\mathbf{H}) \\ &= \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}_c) P(I \mid \mathbf{H}) P(\mathbf{H}). \qquad (2.1) \end{aligned}$$

Independent noise that may contribute to the target behavior is marginalized and omitted for the sake of simplicity in the above equation. The noise term incorporates any hidden variables which influence the behavior but stand in no causal relation to the image. Such variables are not directly relevant to the problem. Fig. 2.1 shows this generative model.

Under this model, we can define an *observational partition* of the space of images $\mathcal{I}$ that groups images into classes that have the same conditional probability $P(T \mid I)$:

Definition 1 (Observational Partition, Observational Class)**.** *The* observational partition $\Pi_o(T, \mathcal{I})$ *of the set of images $\mathcal{I}$ w.r.t. behavior T is the partition induced by the equivalence relation $\sim$ such that $i \sim j$ if and only if $P(T \mid I = i) = P(T \mid I = j)$. We will denote it as Π_o when the context is clear. A cell of an observational partition is called an* observational class.

In standard classification tasks in machine learning, the observational partition is associated with class

[1] An extension of the framework to non-binary, discrete T is easy but complicates the notation significantly. An extension to the continuous case is beyond the scope of this book.

labels. In our case, two images that belong to the same cell of the observational partition assign equal *predictive* probability to the target behavior. Thus, knowing the observational class of an image allows us to predict the value of T. However, the predictive probability assigned to an image does not tell us the *causal* effect of the image on T. For example, a barometer is widely taken to be an excellent predictor of the weather. But changing the barometer needle does not cause an improvement of the weather. It is not a (visual or otherwise) cause of the weather. In contrast, seeing a particular barometer reading may well be a *visual cause* of whether we pack an umbrella.

Our notion of a visual cause depends on the ability to manipulate the image.

Definition 2 (Visual Manipulation)**.** *A* visual manipulation *is the operation* $man(I = i)$ *that changes (the pixels of) the image to image* $i \in \mathcal{I}$*, while not affecting any other variables (such as* $\mathbf{H}$ *or* T*). That is, the manipulated probability distribution of the generative model in Eq.* (2.1) *is given by* $P(T \mid man(I = i)) = \sum_{\mathbf{H}_c} P(T \mid I = i, \mathbf{H}_c)P(\mathbf{H}_c)$ *(see Pearl (2000) for a detailed discussion of the probabilistic interpretation of causal manipulation).*

The manipulation changes the values of the image pixels, but does not change the underlying “world”, represented in our model by the H_i that generated the image. Formally, the manipulation is similar to the do-operator for standard causal models. However, in this book we reserve the do-operation for interventions on causal *macro*-variables, such as the visual cause of T. We discuss the distinction in more detail below.

We can now define the *causal partition* of the image space (with respect to the target behavior T) as:

Definition 3 (Causal Partition, Causal Class)**.** *The causal partition* $\Pi_c(T, \mathcal{I})$ *of the set* $\mathcal{I}$ *w.r.t. behavior* T *is the partition induced by the equivalence relation* $\sim$ *defined on* $\mathcal{I}$ *such that* $i \sim j$ *if and only if* $P(T \mid man(I = i)) = P(T \mid man(I = j))$ *for* $i, j \in \mathcal{I}$*. When the image space and the target behavior are clear from the context, we will indicate the causal partition by* Π_c*. A cell of a causal partition is called a* causal class*.*

The underlying idea is that images are considered causally equivalent with respect to T if they have the same causal effect on T. Given a causal partition of the image space, we can now define the visual cause of T:

Definition 4 (Visual Cause)**.** *The* visual cause C *of a target behavior* T *is a random variable whose value stands in a bijective relation to the causal class of* I*.*

The visual cause is thus a function over $\mathcal{I}$, whose values correspond to the post-manipulation distributions $C(i) = P(T \mid \text{man}(I = i))$. We will write $C(i) = c$ to indicate that the causal class of image $i \in \mathcal{I}$ is c, or in other words, that in image i, the visual cause C takes value c. Knowing C allows us to predict the effects of a visual manipulation $P(T \mid \text{man}(I = i))$, as long as we have estimated $P(T \mid \text{man}(I = i^*_k))$ for one representative i^*_k of each causal class k.

2.2.3 The Causal Coarsening Theorem

The main theorem of this book relates the causal and observational partitions for a given $\mathcal{I}$ and T. It turns out that under appropriate, intuitive assumptions, the causal partition is a coarsening of the observational partition. That is, the causal partition aligns with the observational partition, but the observational partition may subdivide some of the causal classes.

2.2.4 Set-up and Definitions

For simplicity, consider a causal system between three discrete variables H, I, T in which I and H are both causes of T, and H is in addition a cause of I – equivalent to the setup in Fig. 2.1 but with all the confounders collapsed into one variable for simplicity of notation. We assume that these three variables fully describe the causal system, that is, with respect to these three variables the system is causally sufficient. (In fact, since we treat H as an unobserved common cause of I and T, H can be thought of as a catch-all for any confounding between I and T.) The parameterization of this causal system is given by

$$P(H, I, T) = P(H)P(I \mid H)P(T \mid I, H). \tag{2.2}$$

We define partitions of the micro-variable space $\mathcal{I}$.

Definition 5 (partition $\Pi_f(\mathcal{I})$)**.** *Let* $\Pi_f(\mathcal{I})$ *to be the partition on* $\mathcal{I}$ *induced by the relationship* $i_1 \sim i_2 \Leftrightarrow f(i_1) = f(i_2)$ *for any* $i_1, i_2 \in \mathcal{I}$*.*

Here f stands for any function whose domain contains $\mathcal{I}$. For example $P(H \mid I)$ or $P(I)$ are such functions, where $i_1 \sim i_2$ means that $P(H \mid i_1) = P(H \mid i_2)$ for any value of H. Thus the causal and observational partition above can be rewritten as, respectively

$$\Pi_c(\mathcal{I}) = \Pi_{P(T \mid \text{man}(I))}(\mathcal{I}) \tag{2.3}$$

$$\Pi_o(\mathcal{I}) = \Pi_{P(T \mid I)}(\mathcal{I}) \tag{2.4}$$

We write $C(i)$ to denote the causal class of i in $\Pi_c(\mathcal{I})$ and $O(i)$ to denote the observational class of i in $\Pi_o(\mathcal{I})$.

In addition, we will make use below of a partition $\Pi_{P(I \mid H)}(\mathcal{I})$, that we refer to as the confounding partition:

$$i_1 \sim i_2 \quad \Leftrightarrow \quad P(i_1 \mid H) = P(i_2 \mid H) \quad \forall h \in H.$$

We are now ready to state the theorem:

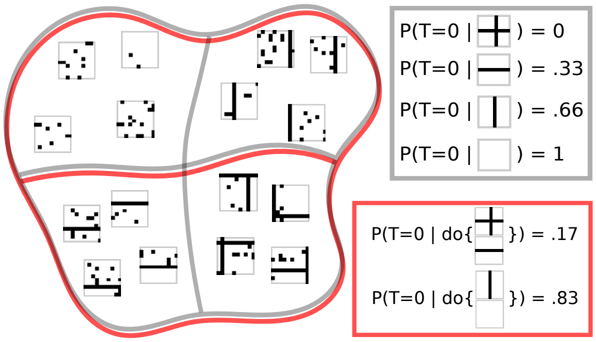


Figure 2.2: The Causal Coarsening Theorem. The observational probabilities of T given I (gray frame) induce an observational partition on the space of all the images (left, observational partition in gray). The causal probabilities (red frame) induce a causal partition, indicated on the left in red. The CCT allows us to expect that the causal partition is a coarsening of the observational partition. The observational and causal probabilities correspond to the generative model shown in Fig. 1.4.

Theorem 6 (Causal Coarsening Theorem)**.** *Among all the joint distributions $P(T, H, I)$ over discrete variables T, H, I, consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{I})$ and a fixed confounding partition $\Pi_{P(T|I)}(\mathcal{I})$. Within this subset, the set of distributions whose causal partition $\Pi_c(\mathcal{I})$ is not a coarsening of the observational partition $\Pi_o(\mathcal{I})$ is a set of measure zero.*

Fig. 2.2 illustrates the relation between the causal and the observational partition implied by the theorem. We prove the CCT in Sec. 2.5.

The motivation for the CCT is to establish a connection between the observational partition $\Pi_o(\mathcal{I})$ and the causal partition $\Pi_c(\mathcal{I})$ such that minimal experimental effort is required to learn the causal partition given an observational partition. In particular, if the observational partition already constitutes a coarsening of the micro-variable space $\mathcal{I}$, then the hope was to leverage this coarse observational partition to learn the causal partition. Consequently, in order to obtain any experimental savings from the developed algorithms we require a theorem that establishes a connection between an observational partition that is itself already a coarsening of the micro-variable space $\mathcal{I}$, and the the causal partition.

An observational partition that is a coarsening of the micro-variable space $\mathcal{I}$ can arise for several reasons, To have such a coarsening, the following equation must be satisfied for at least two distinct $i_1, i_2 \in \mathcal{I}$:

$$
\begin{aligned}
&P(T \mid i_1) = P(T \mid i_2) && (2.5)\\
\Leftrightarrow &\sum_H P(T \mid i_1, H)P(H \mid i_1) - P(T \mid i_2, H)P(H \mid i_2) = 0\\
\Leftrightarrow &\sum_H P(H)(P(T \mid i_1, H)P(i_1 \mid H) - P(T \mid i_2, H)P(i_2 \mid H)) = 0\\
\Leftrightarrow &\sum_H P(T \mid i_1, H)P(i_1 \mid H) - P(T \mid i_2, H)P(i_2 \mid H) = 0 && (2.6)\\
&\text{if } P(H) \neq 0 \quad \forall h \in H
\end{aligned}
$$

Since H is assumed to be a hidden variable there is no significance to states that have zero probability, so the assumption on the last line is innocuous. Note that equation 2.6 is stated entirely in terms of the fundamental parameters in equation 2.2.

Consequently, for an observational partition to be a coarsening of the micro-variable space, the fundamental parameters must combine in just such a way that equation 2.6 is satisfied. However, there is an important subclass of such combinations that satisfy the equation due to the fact that the corresponding fundamental parameters for i_1 and i_2 are equal, i.e. when

$$P(T \mid i_1, H) = P(T \mid i_2, H) \quad \forall h \in H$$
$$P(i_1 \mid H) = P(i_2 \mid H) \quad \forall h \in H$$

It is these cases that are of interest to the discovery of causal macro-variables, since – intuitively – the coarseness of the observational partition arises from causal effects that are invariant across distinctions at the micro-level – this is the case in all the simulated examples enumerated in Chapter 1. In other cases that satisfy equation 2.6, the parameters just happen to combine in such a way as to result in a coarse observational partition.

The CCT shows that no matter what partitions we fix Π_c and $\Pi_{P(I|H)}$ to, the set of distributions consistent with these partitions has the property that the causal partition will be a coarsening of the observational partition except for a set of distributions that has measure zero.

In particular, if we assume that the observational partition is a coarsening of $\mathcal{I}$ *only because* both the confounding partition $\Pi_{P(I|H)}$ and the causal partition $\Pi_{P(T|\text{man}(I))}$ are each coarsenings of $\mathcal{I}$, then the theorem justifies the application of the algorithms developed in the following section to problems where the observational partition is itself already a coarsening of the micro-variable space of $\mathcal{I}$. In other words, when using CFL we assume away cases where a coarse observational partition arises due to "coincidental" combinations of the fundamental parameters that satisfy Equation 2.6. Finally, the notion of coincidence here is not measure-theoretic in the standard sense, since for two fundamental parameters to be equal carries in a standard measure-theoretic analysis the same amount of measure as the event that a combination of parameters satisfy a particular algebraic constraint. However, our set-up takes as starting point the assumption that there exist causal macro-variables in nature. In that case, the equality of two fundamental parameters $P(T \mid h, i_1) = P(T \mid h, i_2)$ is not coincidental but a result of a macro-variable, whereas the satisfaction of some algebraic constraint such as Eq.(2.6) without equalities in the fundamental parameters is a rare event.

Two points are worth noting here: First, the CCT is interesting inasmuch as the visual causes of a behavior do not contain all the information in the image that predict the behavior. Such information, though not itself a cause of the behavior, can be informative about the state of other non-visual causes of the target behavior. Second, the CCT allows us to take any classification problem in which the data is divided into observational classes, and assume that the causal labels do not change within each observational class.

2.2.5 The Complete Macro-variable Description Theorem

Recall the example from Sec. 1.3.1, where the visual presence of an h-bar causes a neuron to spike, and the presence of a v-bar correlates with the spiking only through a confounder. In this section, we formalize the intuition that the v-bar is a visual *spurious correlate* of neural spiking.

Assume that the causal partition Π_c^T is a coarsening of the observational partition Π_o^T, in accordance with the CCT. Each of the causal classes $c_1, \cdots, c_K$ delineates a region in the image space $\mathcal{I}$ such that all the images belonging to that region induce the same $P(T \mid \text{man}(I))$. Each of those regions—say, the k-th one—can be further partitioned into sub-regions $s_1^k, \cdots, s_{M_k}^k$ such that all the images in the m-th sub-region of the k-th causal region induce the same observational probability $P(T \mid I)$. By assumption, the observational partition has a finite number of classes, and we can arbitrarily order the observational classes within each causal class. Once such an ordering is fixed, we can assign an integer $m \in \{1, 2, \cdots, M_k\}$ to each image i belonging to the k-th causal class such that i belongs to the m-th observational class among the M_k observational classes contained in c_k. By construction, this integer explains all the variation of the observational class within a given causal class. This suggests the following definition:

Definition 7 (Spurious Correlate)**.** *The* spurious correlate S *is a discrete random variable whose value differentiates between the observational classes contained in any causal class.*

The spurious correlate is a well-defined function on $\mathcal{I}$, whose value ranges between 1 and $\max_k M_k$. Like C, the spurious correlate S is a macro-variable constructed from the pixels that make up the image. C and S together contain all and only the visual information in I relevant to T, but only C contains the causal information:

Theorem 8 (Complete Macro-variable Description)**.** *The following two statements hold for C and S as defined above:*

1. $P(T \mid I) = P(T \mid C, S)$.
2. *Any other variable X such that $P(T \mid I) = P(T \mid X)$ has entropy $H(X) \geq H(C, S)$.*

We prove the theorem in Sec. 2.5. It guarantees that C and S constitute the smallest-entropy macro-variables that encompass all the information about the relationship between T and I. Fig. 2.3 shows the relationship between C, S and T, the image space $\mathcal{I}$ and the observational and causal partitions schematically. C is now a cause of T, S correlates with T due to the unobserved common causes $\mathbf{H}_C$, and any information irrelevant to T is pushed into the independent noise variables (commonly not shown in graphical representations of structural equation models).

The macro-variable model lends itself to the standard treatment of causal graphical models described in Pearl (2000). We can define interventions on the causal variables $\{C, S, T\}$ using the standard do-operation. The do-operator sets the value of the intervened variable to the desired value, making it independent of its

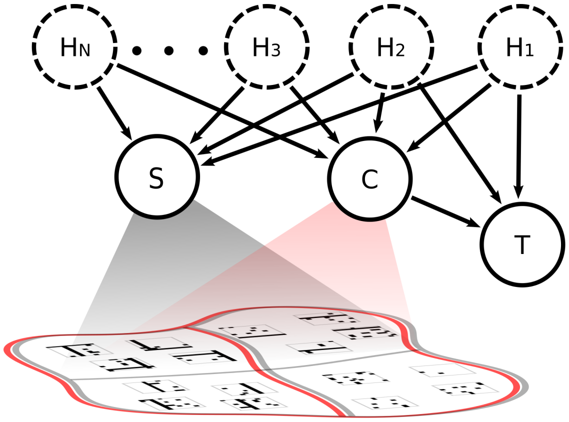


Figure 2.3: A macro-variable model of visual causation. Using our theory of visual causation we can aggregate the information present in visual micro-variables (image pixels) into the visual cause C and spurious correlate S. According to Theorem 8, C and S contain all the information about T available in I.

causes, but it does not (directly) affect the other variables in the system or the relationships between them (see the *modularity assumption* in Pearl (2000)). However, unlike the standard case where causal variables are separated in location (e.g. *smoking* and *lung cancer*), the causal variables in an image may involve the same pixels: C may be the average brightness of the image, whereas S may indicate the presence or absence of particular shapes in the image. An intervention on a causal variable using the do-operator thus requires that the underlying manipulation of the image respects the state of the other causal variables:

Definition 9 (Causal Intervention on Macro-variables)**.** *Given the set of macro-variables $\{C, S\}$ that take on values $\{c, s\}$ for an image $i \in \mathcal{I}$, an intervention $do(C = c')$ on the macro-variable C is given by the manipulation of the image $man(I = i')$ such that $C(i') = c'$ and $S(i') = s$. The intervention $do(S = s')$ is defined analogously as the change of the underlying image that keeps the value of C constant.*

In some cases it can be impossible to manipulate C to a desired value without changing S. We do not take this to be a problem special to our case. In fact, in the standard macro-variable setting of causal analysis we would expect interventions to be much more restricted by physical constraints than we are with our interventions in the image space. This issue is ultimately quite subtle both from the philosophical and practical point of view. We do not discuss it in full detail here, as the details of the discussion may vary significantly between various domains.

2.2.6 Predictive Non-causal Information in the Macro-variable Cause

In some cases C retains predictive information that is not causal. Consider the following example: We have a causal graph consisting of three variables $\{I, T, H\}$ where the causal relations are $I \to T$ and $I \leftarrow H \to T$. All three variables are binary and we have a positive distribution over the variables. In the general case, distributions over this graph satisfy

1. $P(T|do(I = 1)) \neq P(T|do(I = 0))$

2. $P(T|I=1) \neq P(T|I=0)$, and importantly

3. $P(T|I) \neq P(T|do(I))$.

If we view I as an image (which can either be all black or all white), T as the target behavior and H as a hidden confounder, analogous to the set-up in the main article, then the observational partition Π_o has just two classes, namely $\{1, 0\}$. But in this case the observational partition *is the same* as the causal partition: $\Pi_o = \Pi_c$. So by our definition of a spurious correlate, S is a constant, since there are no further distinctions to be made within any of the causal classes. S would be omitted from any standard causal model. Nevertheless, we have in our model still that $P(T|C) \neq P(T|do(C))$, i.e. the causal variable C still contains predictive information that is not causal. Given that there is by construction no other than the causal and the trivial partition in this example, it must be the case that C retains predictive non-causal information. It follows that in our definitions of C and S, it is not the case that the predictive non-causal components of an image can always be completely separated from the causal features. However, any distinction we make in C does make a causal difference.

2.3 Algorithms

The theoretical advances of the previous section allows us to develop algorithms to learn C, the visual cause of a behavior. In addition, knowledge of C will allow us to specify a *manipulator function* which we discuss separately in Chapter 4.

2.3.1 Predicting Macro-variable Intervention Results

A standard machine learning approach to learning the relation between I and T would be to take an *observational dataset* $\mathcal{D}_{obs} = \{(i_k, P(T \mid i_k))\}_{k=1,\cdots,N}$ and learn a predictor f whose training performance guarantees a low test error (so that $f(i^*) \approx P(T \mid i^*)$ for a test image i^*). In causal feature learning, low test error on observational data is insufficient; it is entirely possible that $\mathcal{D}$ contains spurious information useful in predicting test labels which is nevertheless not causal. That is, the prediction may be highly accurate for observational data, but completely inaccurate for a prediction of the effect of a manipulation of the image (recall the barometer example). However, we can use the CCT to obtain a causal dataset from the observational data, and then train a predictor on that dataset. Algorithm 1 uses this strategy to learn a function C that, presented with any image $i \in \mathcal{I}$, returns $C(i) \approx P(T \mid \text{man}(I = i))$. We use a fixed neural network architecture to learn C, but any differentiable hypothesis class could be substituted instead. Differentiability of C is necessary in Section 4.3 in order to learn the manipulator function.

In Step 1 the algorithm picks a representative member of each observational class. The CCT tells us that the causal partition coarsens the observational one. That is, in principle (ignoring sampling issues) it is sufficient to estimate $\hat{C}_m = P(T \mid \text{man}(I = i_{k_m}))$ for just one image in an observational class m in order

```
Algorithm 1: Causal Predictor Training
    input  : $\mathcal{D}_{obs} = \{(i_1, p_1 = p(T \mid i_1)), \cdots, (i_N, p_N = p(T \mid i_N))\}$ – observational data
             $\mathcal{P} = \{P_1, \cdots, P_M\}$ – the set of observational classes (so that $\forall k, p_k \in \mathcal{P}, 1 \leq k \leq N$)
             Train – a neural net training algorithm
    output: $C \colon \mathcal{I} \to [0, 1]$ – the causal variable
  1 Pick $\{i_{k_1}, \cdots, i_{k_M}\} \subset \{i_1, \cdots, i_N\}$ s.t. $p_{k_m} = P_m$;
  2 Estimate $\hat{C}_m \leftarrow P(T \mid \text{man}(I = i_{k_m}))$ for each $m$;
  3 For all $k$ let $\hat{C}(i_k) \leftarrow \hat{C}_m$ if $p_k = P_m$;
  4 $\mathcal{D}_{csl} \leftarrow \{(i_1, \hat{C}(i_1)), \cdots, (i_N, \hat{C}(i_N))\}$;
  5 $C \leftarrow \texttt{Train}(\mathcal{D}_{csl})$;
```

to know that $P(T \mid \text{man}(I = i)) = \hat{C}_m$ for any other i in the same observational class. The choice of the experimental method of estimating the causal class in Step 2 is left to the user and depends on the behaving agent and the behavior in question. If, for example, T represents whether the spiking rate of a recorded neuron is above a fixed threshold, estimating $P(T \mid \text{man}(I = i))$ could consist of recording the neuron's response to i in a laboratory setting multiple times, and then calculating the probability of spiking from the finite sample. The causal dataset created in Step 4 consists of the observational inputs and their causal classes. The causal dataset is acquired through $\mathcal{O}(N)$ experiments, where N is the number of observational classes. The final step of the algorithm trains a neural network that predicts the causal labels on unseen images. The choice of the method of training is again left to the user.

2.4 Experiments

Section 4.4 contains experiments shared between this chapter and Chapter 4.

2.5 Proofs

Before proving the CCT, we prove a useful lemma.

Lemma 10. *Let $S_{P(H)}$ denote the simplex of multinomial distributions over the values of H. For fixed $P(T \mid H, I)$, the subset of $S_{P(H)}$ for which Π_c is not equal to $\Pi_{P(T|H,I)}(\mathcal{I})$ is measure zero.*

Proof. We want to show that the subset of $S_{P(H)}$ for which, for any $i_1, i_2 \in \mathcal{I}$ and $h \in H$

$$P(T \mid H = h, i_1) \neq P(T \mid H = h, i_2), \text{ and} \tag{2.7}$$

$$P(T \mid \text{man}(i_1)) = P(T \mid \text{man}(i_2)), \tag{2.8}$$

is measure zero. (Note that if $P(T \mid H, I)$ is the same for all i, equality of Π_c and $\Pi_{P(T|H,I)}$ follows directly from their definitions).

Eq. 2.8 is equivalent to $\sum_h P(H = h)[P(T \mid H = h, i_1) - P(T \mid H = h, i_2)] = 0$. Since this is a linear constraint on $S_{P(H)}$, in order to show that it is satisfied on a measure-zero subset we only need to show that there is at least one point which does not satisfy it.

First, set $P(H = h) = 1/K$, where K is the number of states of H, for all h. If the equation is not satisfied, we are done. If it is satisfied, it must be for some h_1 that $P(T \mid H = h_1, i_1) - P(T \mid H = h_1, i_2) > 0$ and for some h_2, we have $P(T \mid H = h_2, i_1) - P(T \mid H = h_2, i_2) < 0$. Pick any $0 < \epsilon < min(1/K, 1 - 1/K)$. Set $P(H = h_1) = 1/K + \epsilon$ and $P(H = h_2) = 1/K - \epsilon$, and $P(H = h) = 1/K$ for other h. Then Eq. (2.8) does not hold. □

Theorem (Causal Coarsening) Among all the joint distributions $P(T, H, I)$ over discrete variables T, H, I, consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{I})$ and a fixed confounding partition $\Pi_{P(T|I)}(\mathcal{I})$. Within this subset, the set of distributions whose causal partition $\Pi_c(\mathcal{I})$ is not a coarsening of the observational partition $\Pi_o(\mathcal{I})$ is a set of measure zero.

Proof. (i) We first set up the notation. Assume that T is binary, and that H and I are discrete variables (say $|H| = K, |I| = N$, though N can be very large). $P(T \mid H, I)$ requires $K \times N$ parameters, $P(I \mid H)$ requires $(N - 1) \times K$ parameters, and $P(H)$ requires another $K - 1$ parameters. Call the parameters, respectively,

$$\alpha_{h,i} \triangleq P(T = 0 \mid H = h, I = i)$$
$$\beta_{i,h} \triangleq P(I = i \mid H = h)$$
$$\gamma_h \triangleq P(H = h)$$

We will denote parameter vectors as

$$\alpha = (\alpha_{h_1,i_1}, \cdots, \alpha_{h_K,i_N}) \in \mathbb{R}^{K \times N}$$
$$\beta = (\beta_{i_1,h_1}, \cdots, \beta_{i_{N-1},h_K}) \in \mathbb{R}^{(N-1) \times K}$$
$$\gamma = (\gamma_{h_1}, \cdots, \gamma_{h_K}) \in \mathbb{R}^{K-1},$$

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution $P(T, H, I)$ with a point $(\alpha, \beta, \gamma) \in P[\alpha, \beta, \gamma] \subset \mathbb{R}^{K^2 \times (K-1) \times N \times (N-1)}$.

(ii) Show that for any α, β consistent with Π_c and $\Pi_{P(I|H)}$, the causal partition and the confounding partition are, in general, fixed.

To proceed with the proof, pick any point in the $P(T \mid H, I) \times P(I \mid H)$ space – that is, fix α and β. The only remaining free parameters are now in γ. Varying these values creates a subset of the space of all joints isometric to the $(K - 1)$-dimensional simplex of multinomial distributions over K states (call the simplex S_{K-1}):

$$P[\gamma; \alpha, \beta] = \{(\alpha, \beta, \gamma) \mid \gamma \in S_{K-1}\} \subset [0, 1]^{(K-1)}.$$

Note that fixing β directly fixes $\Pi_{P(I|H)}$. Fixing α doesn't directly fix Π_c. But by Lemma 10, for *almost all* distributions in $P[\gamma; \alpha, \beta]$ the causal partition Π_c equals the partition $\Pi_{P(T|H,I)}$, which is directly fixed by α. Let $P'[\gamma; \alpha, \beta]$ be $P[\gamma; \alpha, \beta]$ minus this measure zero subset.

The statement of the theorem fixes Π_c and $\Pi_{P(I|H)}$. If the α, β we picked are consistent with these partitions within $P'[\gamma; \alpha, \beta]$, continue with the proof. Otherwise, choose other α, β.

We now prove that within $P'[\gamma; \alpha, \beta]$ the set of γ for which the causal partition Π_c is not a coarsening of the observational partition Π_o is of measure zero. Later in (iv) we integrate the result over all α, β.

(iii) Let the causal coarsening constraint be that for $i_1, i_2 \in \mathcal{I}$ we have

$$O(i_1) = O(i_2) \quad \Rightarrow \quad C(i_1) = C(i_2). \tag{2.9}$$

That is, it is not the case that two members of $\mathcal{I}$ are observationally equivalent but have causally different effects.

We show that the causal coarsening constraint holds for each pair $i_1, i_2 \in \mathcal{I}$: Pick any $i_1, i_2 \in \mathcal{I}$. If $C(i_1) = C(i_2)$, then we are done with this pair. So assume that there is a causal difference, i.e. $C(i_1) \neq C(i_2)$. Our goal is now to show that then only a measure-zero subset of $P'[\gamma; \alpha, \beta]$ allows for $O(i_1) = O(i_2)$.

We first show that $O(i_1) = O(i_2)$ places a polynomial constraint on $P'[\gamma; \alpha, \beta]$. We have

$$\begin{aligned} O(i_1) &= \frac{1}{P(i_1)} \sum_h \alpha_{h,i_1} \beta_{i_1,h} \gamma_h, \\ O(i_2) &= \frac{1}{P(i_2)} \sum_h \alpha_{h,i_2} \beta_{i_2,h} \gamma_h. \end{aligned}$$

After expanding in terms of α, β, γ, we have

$$\begin{gathered} O(i_1) = O(i_2) \quad \Leftrightarrow \\ \sum_{h_k, h_l} \gamma_{h_k} \gamma_{h_l} [\beta_{i_2,h_k} \beta_{i_1,h_l} \alpha_{h_l,i_1} - \beta_{i_1,h_k} \beta_{i_2,h_l} \alpha_{h_l,i_2}] = 0. \end{gathered} \tag{2.10}$$

We have thus shown that, for fixed α, β and i_1, i_2, the violation of the causal coarsening constraint (2.9), is a polynomial constraint on $P'[\gamma; \alpha, \beta]$. By an algebraic lemma (proven by Okamoto, 1973), the subset on which the constraint holds is measure zero *if the constraint is not trivial*. That is, we only need to find one γ for which Eq. (2.10) does not hold to prove that it almost never holds.

To find such γ, let $\gamma_h = 1/K$ for all h. If for this γ Eq. (2.10) does not hold, we are done. If it does hold, since we know α is not all 0, there must be in the sum of the equation at least one factor $[\beta_{i_2,h_k} \beta_{i_1,h_l} \alpha_{h_l,i_1} - \beta_{i_1,h_k} \beta_{i_2,h_l} \alpha_{h_l,i_2}]$ which is positive, and one that is negative. Call the h_k, h_l corresponding to the positive element h_{k^+}, h_{l^+} and to the negative element h_{k^-}, h_{l^-}. Since the factors are different, we must have either $k^+ \neq k^-$ or $l^+ \neq l^-$ (or both). Assume $k^+ \neq k^-$. Now, pick any positive $\epsilon < min(1/K, 1 - 1/K)$. Set $\gamma_h = 1/K$ for all $h \neq h_{k^+}, h_{k^-}$ and set $\gamma_{h_{k^+}} = \frac{1}{K} + \epsilon$ and $\gamma_{h_{k^-}} = \frac{1}{K} - \epsilon$. In this way, we keep $\sum_h \gamma$

unchanged, and are guaranteed that Eq. (2.10) does not hold. That is, for this γ we have $O(i_1) \neq O(i_2)^2$.

(iv) Show that the theorem holds over the space of all distributions.

To reiterate proof progress thus far:

1. We fixed the macro-scale causal partition Π_c and the confounding partition $\Pi_{P(I|H)}$ and picked arbitrary α and β compatible with these partitions.
2. We picked two points i_1, i_2 for which $C(i_1) \neq C(i_2)$.
3. We showed that for any such two points, the subset of $P'[\gamma; \alpha, \beta]$ for which $O(i_1) = O(i_2)$ is measure zero.

Since there are only finitely many points in $\mathcal{I}$, it follows that for the fixed α, β, the subset of $P'[\gamma; \alpha, \beta]$ on which the coarsening constraint (2.9 does not hold for at least one pair of points is also measure zero. Since $P[\gamma; \alpha, \beta] - P'[\gamma; \alpha, \beta]$ is a set of measure zero, the subset of $P[\gamma; \alpha, \beta]$ on which the causal coarsening constraint does not hold is also measure zero.

Now, call the set of all joint distributions that agree with Π_c and $\Pi_{P(I|H)}$ the admissible set, and denote it with $P[\alpha, \beta, \gamma]_A$. For each α, β consistent with the two partitions, call the (measure zero) subset of $P[\gamma; \alpha, \beta]_A$ that violates the causal coarsening constraint $z[\alpha, \beta]$. Let $Z = \cup_{\alpha,\beta} z[\alpha, \beta] \subset P[\alpha, \beta, \gamma]_A$ be the set of all the admissible joint distributions which violate the causal coarsening constraint. We want to prove that $\mu(Z) = 0$, where μ is the Lebesgue measure. To show this, we will use the indicator function

$$\hat{z}(\alpha, \beta, \gamma) = \begin{cases} 1 & \text{if } \gamma \in z[\alpha, \beta], \\ 0 & \text{otherwise.} \end{cases}$$

By basic properties of positive measures we have

$$\mu(Z) = \int_{P[\alpha,\beta,\gamma]_A} \hat{z} \; d\mu.$$

For simplicity of notation, let

1. $\mathcal{A} \subset \mathbb{R}^{K \times N}$ be the set of all possible α's (a Cartesian product of $K \times N$ 1-d simplexes);
2. $\mathcal{B} \subset \mathbb{R}^{N \times K}$ be the set of all possible β's (a Cartesian product of K simplexes, each $N-1$ dimensional);
3. $\mathcal{G} \subset \mathbb{R}^{K}$ be the set of all possible γ's (a $K-1$-dimensional simplex).

Note that each set has, in its respective Euclidean space, a non-empty interior, and comes equipped with the Lebesgue measure.

Finally, let $I_A(\alpha, \beta)$ be the indicator function that evaluates to 1 if α, β are admissible and evaluates to 0 otherwise. We have

[2] It is possible that this γ is not in $P'[\gamma; \alpha, \beta]$. However, it is guaranteed to be in $P[\gamma; \alpha, \beta]$. Since a subset of measure zero in $P[\gamma; \alpha, \beta]$ is also measure zero in $P'[\gamma; \alpha, \beta]$, this does not influence the proof.

$$\begin{aligned}
\int_{P[\alpha,\beta,\gamma]_A} \hat{z}\, d\mu &= \int_{\mathcal{A}\times\mathcal{B}\times\mathcal{G}} \hat{z}(\alpha,\beta,\gamma) I_A(\alpha,\beta)\, d(\gamma,\beta,\alpha) \\
&= \int_{\mathcal{A}\times\mathcal{B}} \int_{\mathcal{G}} \hat{z}(\alpha,\beta,\gamma)\, d(\gamma)\; I_{\Pi_c}(\beta,\alpha)\, d(\beta,\alpha) \\
&= \int_{\mathcal{A}\times\mathcal{B}} \mu(z[\alpha,\beta])\; I_A(\alpha,\beta)\, d(\beta,\alpha) \qquad (2.11) \\
&= \int_{\mathcal{A}\times\mathcal{B}} 0\, I_A(\alpha,\beta)\, d(\beta,\alpha) \\
&= 0.
\end{aligned}$$

Equation (2.11) follows as $\hat{z}$ restricted to $P[\gamma;\alpha,\beta]$ is the indicator function of $z[\alpha,\beta]$.

This completes the proof that Z, the set of joint distributions over T, H and I that violate the causal coarsening constraint (2.9) is measure zero. □

Theorem (Complete Macro-variable Description) *The following two statements hold for C and S as defined in Sec. 2.2.5:*

1. *$P(T \mid I) = P(T \mid C, S)$.*
2. *Any other variable X such that $P(T \mid I) = P(T \mid X)$ has Shannon entropy $H(X) \geq H(C, S)$.*

Proof. The first part follows by construction of S. For the second part, note that by the CCT there is a bijective correspondence between the pairs of values (c, s) and the observational probabilities $P(T \mid I)$. Call this correspondence f, that is $f(c, s) = P(T \mid c, s)$ and $f^{-1}(p) = \{c, s \mid P(T|c, s) = p\}$. Further, define g as the function on X such that $g\colon x \mapsto P(T \mid x)$. But since $P(T \mid X) = P(T \mid I)$, we have $(c, s) = f^{-1}(g(x))$. That is, the value of C and S is a function of the value of X, and thus the entropy of C and S is smaller than or equal to the entropy of X. □

2.6 Additional Acknowledgement

I wish to thank professor Jiji Zhang for pointing to an error in an earlier version of the Causal Coarsening Theorem. Without his attentive review of this work, the main theorem of this thesis would likely still be wrong.

Chapter 3

Unsupervised Causal Feature Learning

The previous chapter develops a method to discover from micro-variable data the macro-variable cause of a pre-defined macro-variable "target behavior". In this chapter, we do not assume that the macro-level effect is already specified. Instead, in a generalization of the CFL framework, we simultaneously recover the macro-level cause C and macro-level effect E from micro-variable data. We will use the name Causal Feature Learning to refer to both frameworks. When ambiguous, we will refer to the first as supervised, and the current chapter's as unsupervised CFL.

3.1 Advances in This Chapter

This chapter presents the following advances in machine learning and causal inference:

- An extension of the CFL framework of Chapter 2 to the scenario in which all the observed variables are micro-variables.
- An extension of the CCT to this case.
- A definition of the subsidiary variable, which makes mathematical sense of "micro-to-macro" hierarchies of variables (recall the example of macro-economic processes supervening on individual activities that in turn supervene on personal psychological processes and finally neural state aggregates).
- New algorithms that generalize algorithms from Chapter 2 to new situations.
- Algorithms to detect hierarchies of causal variables in data.
- The Sufficient Causal Description Theorem, which shows that our causal macro-variables are minimal sufficient statistics of causal interactions of a causal systems.

Some of the definitions of this chapter are similar or identical to those of Chapter 2. This is because this chapter extends the CFL framework and generalizes the previous chapter. We provide the older definitions for completeness, in the context of notation developed here.

3.2 Theory

CFL takes microvariable data and produces macrovariable causal hypotheses. Throughout this chapter, we will use the example from Sec. 1.3.2 to illustrate the definitions and algorithms. Recalling the example, let $\mathcal{L} = (0, 360)$ denote our input microvariable space (range of the hue variable), and $\mathcal{R} = (0, 1)$ the output microvariable space (range of eda). We will denote the random variables defined over these spaces as hue and eda, and their specific instantiations as h and e – for example, we will write $p(eda = e \mid hue = h)$ for some $e \in \mathcal{R}, h \in \mathcal{L}$. Note that the framework applies to general microvariables, not only this specific case used for illustration. For example, in Sec. 3.4 below we apply the framework to the high-dimensional example of images causing changes in neural populations.

3.2.1 Learning the Causal Hypothesis

Fig. 3.1A shows 1,000 samples from $p(hue, eda)$ together with the ground-truth conditional distribution $p(hue \mid eda)$. These observations are generated from the probabilistic model shown in Fig. 1.5, where hue and eda are confounded by the unobserved lat.

The empirical distribution shown in the figure indicates that $p(eda \mid hue)$ is constant for any $h \in (0, 90)$ as well as for $h \in (90, 180)$, $h \in (180, 270)$ and $h \in (270, 360)$. Fig. 1.5 shows that indeed, this *partition* of hue into four classes captures all the combinations of macrovariables supervening on hue. For example, $h \in (0, 90)$ if and only if $W = 1$ and $R = 1$, $h \in (90, 180)$ if and only if $W = 1$ and $R = 0$, and so on. Such partitioning of a microvariable space into the coarsest cells that retain all the observational distinctions is the key element of CFL. This construction, called the (supervised) *Observational Partition*, abstracts away all the irrelevant micro-level details:

Definition 11 (Unsupervised Observational Partition, Unsupervised Observational Class)**.** *The* unsupervised observational partition of $\mathcal{L}$*, denoted by* $\Pi_o(\mathcal{L})$*, is the partition induced by the equivalence relation* $\sim_h$ *such that*

$$h_1 \sim_h h_2 \quad \Leftrightarrow \quad \forall_{e \in \mathcal{R}} p(e \mid h_1) = p(e \mid h_2).$$

The unsupervised observational partition of $\mathcal{R}$*, denoted by* $\Pi_o(\mathcal{R})$*, is the partition induced by the equivalence relation* $\sim_e$ *such that*

$$e_1 \sim_e e_2 \quad \Leftrightarrow \quad \forall_{h \in \mathcal{L}}\ p(e_1 \mid h) = p(e_2 \mid h).$$

A cell of an observational partition is called an unsupervised observational class *(of* $\mathcal{L}$ *or* $\mathcal{R}$*).*

Whenever context allows, we will call the unsupervised observational partition and class simply the observational partition and class. The observational partition of $\mathcal{R}$ is easily discerned from Fig. 3.1: $e \in (0, .5)$

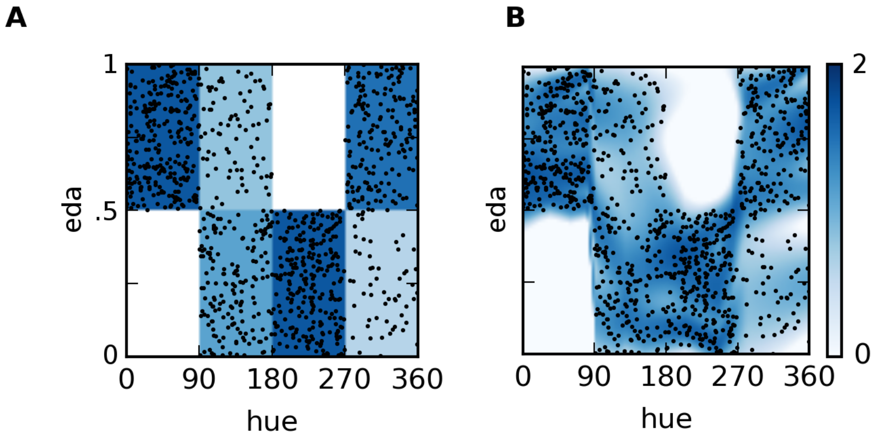


Figure 3.1: **Model Samples and pdf**. A) Black dots are samples from the joint $p(hue, eda)$, background color shows the ground-truth value of $p(eda \mid hue)$. B) The result of conditional density learning of $p(eda \mid hue)$ using a Mixture Density Network (see Sec. 3.3).

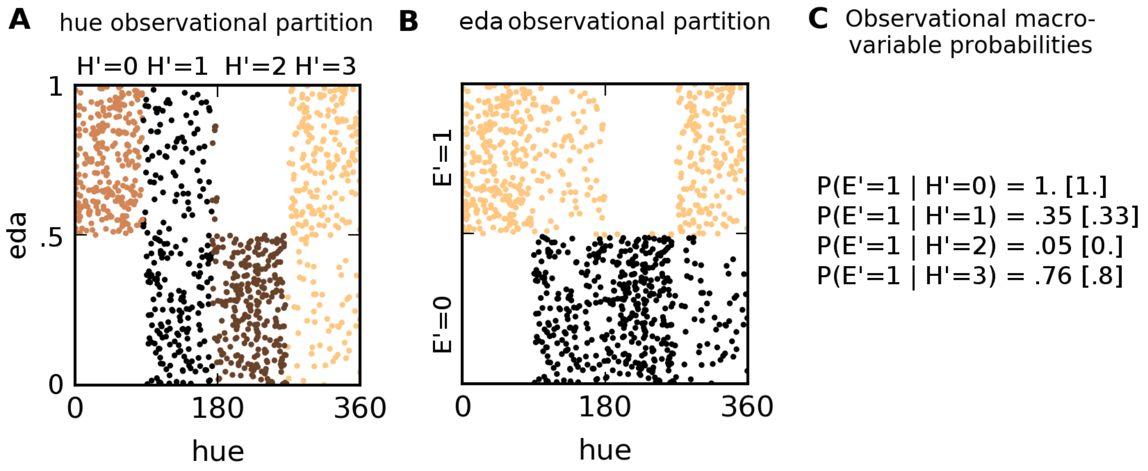


Figure 3.2: **Learning the Observational Partition**. A) The observational partition learned on $\mathcal{L}$ results from clustering the samples' h coordinate with respect to the inferred $p(eda \mid hue)$ shown in Fig. 3.1B. We indicate the learned partitions with an apostrophe, H' and E' in contrast with the ground-truth H and E. B) The observational partition of $\mathcal{R}$, with two cells, results from clustering the samples' eda-coordinate with respect to the inferred $p(eda \mid hue)$. C) The observational partitions are endowed with probability densities simply by counting the histogram of the microvariable samples in each (conditional) macrovariable state. The ground truth values (see Fig. 1.5) are given in square brackets.

has the same $P(e \mid h)$ for any h. Let us index the observational classes on $\mathcal{L}$ as $H = 0, 1, 2, 3$ if $h \in (0, 90)$, $(90, 180)$, $(180, 270)$, $(270, 360)$ respectively, and $E = 0, 1$ if $e \in (0, .5)$ and $(.5, 1)$ respectively. We can then compress $p(eda \mid hue)$ to only four numbers without losing any information:

$$\begin{aligned}
&P(E = 1 \mid H = 0) = 1,\\
&P(E = 1 \mid H = 1) = 1/3,\\
&P(E = 1 \mid H = 2) = 0,\\
&P(E = 1 \mid H = 3) = 4/5.
\end{aligned}$$

Note that this corresponds to $p(A \mid R, W)$ in Fig. 1.5. However, whereas R truly is a cause of A, the non-causal dependence of A on W results from the confounder lat. The observational partition can be seen as a *macrovariable causal hypothesis* for the causal effect of hue on eda. However, the observational partition of hue does not necessarily characterize the *cause* of the observational class of eda.

3.2.2 Weeding Out the Spurious Correlates

Our notion of causality is rooted in the framework of Pearl (2000) and Spirtes et al. (2000). Intuitively, X causes Y if *intervening* on (or manipulating) X, without influencing any other variables in the system, changes the distribution of Y. That is, $P(Y \mid \text{do}(X))$ is not constant. But as is well-known, the conditional probability distribution $P(Y \mid X)$ for any two variables X and Y does not fix the causal effect $P(Y \mid \text{do}(X))$. For example, the barometer's needle *predicts* rain, but manipulating the needle will not *cause* the weather to change.

The observational partition can be used as a basis for an efficient testing procedure of causal hypotheses. To distinguish interventions in the microvariable space from those on the macrovariable space, we denote the manipulation operation in the microvariable space with the operator man() and reserve the standard do() operator for causal macrovariables:

Definition 12 (Microvariable Manipulation)**.** *A* microvariable manipulation *is the operation man*$(hue = h)$ *(we will often simply write man*(h) *for a specific manipulation) that changes the microvariable hue to h* $\in \mathcal{L}$*, while not (directly) affecting any other variables (such as* lat *or* eda*). That is, the manipulated probability distribution of the generative model is given by*

$$P(eda \mid man(hue = h)) = \sum_{l} P(eda \mid hue = h, lat = l)P(lat = l).$$

In contrast to the conditional distribution $p(eda \mid hue = h)$, the dependency between lat and hue is removed in the manipulated probability $p(eda \mid \text{man}(hue = h))$. This is because the latter equation models an *intervention*, where the value $hue = h$ is set in a controlled setting. For example, placing a subject in a room with wallpapers of a particular hue is a micro-level manipulation.

A macrovariable intervention do$(X = x)$ amounts to setting the underlying microvariable to *any* value within the specified partition cell x. The value of the underlying microvariable need not be fully determined by the intervention. For example, do$(R = 1)$ in our toy model would mean that the subject is placed in a room colored with any hue belonging to the $R = 1$ range as indicated in Fig. 1.5B. Note that any such experiment would, according to our model, have the same effect on eda (and A).

In our model, $P(A \mid \text{do}(R = r)) \neq P(A)$ for any r, and $P(A \mid \text{do}(W = w)) = P(A)$ for any w, which confirms the intuition that R is a cause of A, but W is not. However, the unsupervised observational partition $\Pi_o(\mathcal{L})$ contains information about both R and W.

We can discover which cells of the observational partition are causally relevant using a simple experimental procedure, illustrated in Fig. 3.3. Pick one representative h_i from each observational class i and perform the intervention man$(hue = h_i)$. Then, merge those cells of the observational partition whose representatives induced the same $p(eda \mid \text{man}(h_i))$ (see Algorithm 3). The resulting *causal partition* retains only the distinction between $hue \in (90, 270)$ and $hue \in (0, 90) \cup (270, 360)$ — which is our "Red" variable, the true cause of A.

This procedure can be applied in the general setting. Let us first define the causal partition, which corresponds to the macrovariable true cause. We will then show that the causal partition is almost always a *coarsening* of the observational partition, just like in our toy model.

Definition 13 (Unsupervised Causal Partition, Causal Class)**.** *The* unsupervised causal partition of $\mathcal{L}$*, denoted by* $\Pi_c(\mathcal{L})$ *is the partition induced by the equivalence relation* $\sim_h$ *such that*

$$h_1 \sim_h h_2 \quad \Leftrightarrow \quad \forall_{e \in \mathcal{R}} p(e \mid \textit{man}(h_1)) = p(e \mid \textit{man}(h_2)).$$

Similarly, the unsupervised causal partition of $\mathcal{R}$*, denoted by* $\Pi_c(\mathcal{R})$*, is the partition induced by the equivalence relation* $\sim_e$ *such that*

$$e_1 \sim_e e_2 \quad \Leftrightarrow \quad \forall_{h \in \mathcal{L}}\ p(e_1 \mid \textit{man}(h)) = p(e_2 \mid \textit{man}(h)).$$

We call a cell of a causal partition a causal class *of hue or eda.*

That is, two microvariable states $h_1, h_2 \in \mathcal{L}$ belong to the same causal class if they have the same exact effect on the microvariable eda. This implies that switching between the causal classes of hue is the only way to change $p(eda \mid \text{man}(hue))$. The causal class is precisely the value of the macrovariable cause.

Definition 14 (Macrovariable Cause and Effect)**.** *The* unsupervised cause C *is a random variable whose value stands in a bijective relation to the causal class of* $\mathcal{L}$*. The* unsupervised effect S *is a random variable whose value stands in a bijective relation to the causal class of* $\mathcal{R}$*. We will also use* C *and* S *to denote the* *functions that map each h and e, respectively, to its causal class. We will thus write, for example, $C(h) = c$ to indicate that the causal cell of h is c.*

The standard do()-operator is now simply defined as an intervention on such a causal macrovariable. But note that a macrovariable intervention, while well-defined in the macrovariable space, in general has multiple instantiations in the microvariable space. In our simplified example, "do(R=0)" can be realized by man$(hue = h)$ for any $h \in (90, 270)$. Macrovariables treat such distinctions as irrelevant *because they make no causal difference*.

Definition 15 (Macrovariable Manipulation)**.** *The operation do$(X = x)$ on a macrovariable is given by a manipulation of the underlying microvariable man$(hue = h)$ to some value h such that $X(h) = x$.*

We are now ready to state our main theorem, which connects microvariable observations to macrovariable causal relations.

Theorem 16 (Unsupervised Causal Coarsening Theorem)**.** *Among all the joint distributions $P(T, H, I)$ over discrete variables T, H, I, consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{I})$ and a fixed confounding partition $\Pi_{P(T|I)}(\mathcal{I})$. Within this subset, the following two statements hold:*

1. *The subset of distributions for which $\Pi_c(\mathcal{I})$ is* not *a coarsening of the observational partition $\Pi_o(\mathcal{I})$ is Lebesgue measure zero, and*
2. *The subset of distributions for which $\Pi_c(\mathcal{J})$ is* not *a coarsening of the observational partition $\Pi_o(\mathcal{J})$ is Lebesgue measure zero.*

3.2.3 Subsidiary Variables and the Sufficient Causal Description Theorem

Consider the example of a neural population whose behavior is influenced by visual stimuli (Sec. 1.3.3). This system contains no confounder – the behavior of the simulated neural population is directly affected by two independent causal mechanisms: the presence of a v-bar can create a neural pulse, and the presence of an h-bar can induce a 30Hz neural rhythm. We wrote "$P(\text{30Hz} = 1 \mid \text{do(v-bar} = 1)) = .8$ and $P(\text{pulse} = 1 \mid \text{do(h-bar} = 1)) = .8$", and said that these two mechanisms compose to bring about the observed effects. We now formalize under what conditions higher-level variables, such as "30Hz" or "v-bar", can arise from the unsupervised causal partition.

Definition 17 (Subsidiary Causal Variables)**.** *Let C and E be the unsupervised cause and effect of a causal system. Let $\bar{C}$ and $\bar{E}$ be strict coarsenings of C and E. Denote by $c_1(l), \cdots, c_{N_l}(l)$ the cells of C that belong to the l-th cell of $\bar{C}$. We say that $\bar{C}$ and $\bar{E}$ are* subsidiary causal variables*, and that $\bar{C}$ is a* subsidiary cause *of the* subsidiary effect *$\bar{E}$ if (i) $\forall_l P(\bar{E} \mid do(C = c_1(l))) = \cdots = P(\bar{E} \mid do(C = c_{N_l}(l)))$, and (ii) $P(\bar{E} \mid do(\bar{C} = \bar{c}_1)) \neq P(\bar{E} \mid do(\bar{C} = \bar{c}_2))$ for any distinct $\bar{c}_1$ and $\bar{c}_2$ in the range of $\bar{C}$.*

According to the definition, any coarsening of C and E that aspires to be a subsidiary cause-effect pair has to satisfy two conditions. First, manipulations on the subsidiary cause $\bar{C}$ have to be well-defined. The

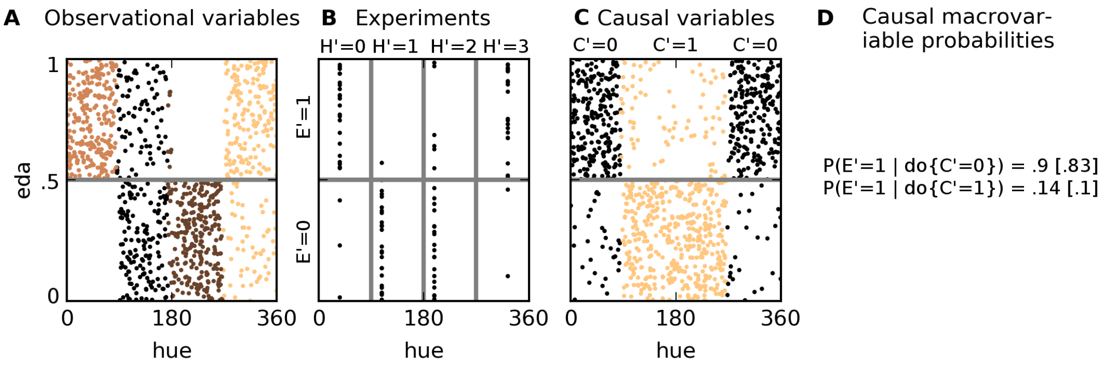


Figure 3.3: **Learning the Causal Partition**. A) The observational partition, obtained from the empirical distribution $p(eda \mid hue)$ (Eq. 2.1), is a causal hypothesis: each cell of $\Pi_o(\mathcal{L})$ could have a different effect on the probability of cells of $\Pi_o(\mathcal{R})$. B) Conducting experiments to estimate $P(E' \mid \text{do}(H' = h')$ amounts to estimating $P(E' \mid \text{man}(hue = h))$ for any $h \in h'$. In this case, we arbitrarily chose $hue = 36, 100, 195, 320$ as representatives of the observational cells. By experimental estimate, $P(E' = 1 \mid \text{man}(hue = 36)) = 22/25$ [the ground-truth is .83], $P(E' = 1 \mid \text{man}(hue = 100)) = 1/25[.1]$, $P(E' = 1 \mid \text{man}(hue = 195)) = 6/25[.1]$, $P(E' = 1 \mid \text{man}(hue = 320)) = 23/25[.83]$. C) The causal partition on $\mathcal{L}$ results from merging the observational cells whose representatives induce similar $P(E' \mid \text{man}(hue))$. Here, we show both the causal partition (in color) and 1000 samples from the causal density $p(eda \mid \text{man}(hue))$. As expected, the sampled structure is homogeneous within each causal class. It is also different from the observational density, because the man() operator removes confounding. D) Estimates of the macrovariable causal probabilities, obtained from experiments shown in B) – ground-truth values in square brackets. Note that a close prediction of the behavior shown in C) was obtained from the few samples in B).

definition guarantees that any two i_1, i_2 for which $\bar{C}(i_1) = \bar{C}(i_2)$ generate the same distribution over the subsidiary effect: For such i_1, i_2 we have $P(\bar{E} \mid \text{do}(\bar{C} = \bar{C}(i_1))) = P(\bar{E} \mid \text{do}(\bar{C} = \bar{C}(i_2)))$. In our example, producing an image with an h-bar induces the neural pulse with probability .8. The probability of the pulse is indifferent to the presence/absence of a v-bar (or any other structure) in the image (see also Fig. 3.4a,b). On the other hand, we claimed that v-bars cause rhythms, not pulses (see Fig. 3.4c). What shows formally that v-bars do not cause pulses? Producing an image i with a v-bar but no h-bar gives us $P(\text{pulse} \mid \text{man}(i)) = 0$, but if i contains both h- and v-bars, we have $P(\text{pulse} \mid \text{man}(i)) = .8$. This disagrees with our definition of what it takes to be a causal variable: the manipulation on the macro-cause v-bar is not well-defined with respect to the macro-effect pulse, as the effects of micro-variables belonging to the same macro-variable causal class are not the same. We have what Spirtes and Scheines (2004) call an "ambiguous manipulation" of v-bar with respect to the pulse.

The second condition in the definition ensures that the values of subsidiary causes are only distinct when they have distinct effects. A succinct answer to the question "what causes the neural pulse?" is "the presence of a horizontal bar" — not "two states: one corresponding to the presence of a horizontal bar along with the presence of a vertical bar; the other corresponding to the presence of a horizontal bar without the presence of a vertical bar". Two states with the same probabilistic effect should be combined.

Together, the two conditions ensure that subsidiary causes and effects allow for well-defined, parsimonious manipulations. Equipped with the notion of subsidiary causal variables and an understanding of what it takes to define $P(\bar{E} \mid \mathrm{do}(\bar{C}))$, we can complete our Unsupervised Sufficient Causal Description theorem:

Theorem 18 (Unsupervised Sufficient Causal Description)**.** *Let $(\mathcal{I}, \mathcal{J})$ be a causal system and let C and E be its cause and effect. Let $\mathbf{E}$ be E applied sample-wise to a sample from the system (so that e.g. $\mathbf{E}(j_1, \cdots, j_k) = (E(j_1), \cdots, E(j_k))$). Then:*

1. *Among all the partitions of $\mathcal{J}$, $\mathbf{E}$ is the* minimal *sufficient statistic for $P(J \mid man(i))$ for any $i \in \mathcal{I}$, and*
2. *C and E losslessly recover $P(j \mid man(i))$. No other (subsidiary) causal variable losslessly recovers $P(j \mid man(i))$. Any other partition is either finer than C, E or does not define unambiguous manipulations. In this sense, the unsupervised causal partition corresponds to the coarsest partition that losslessly recovers $P(j \mid man(i))$.*

The proof is provided in Sec. 3.5. The theorem suggests that the use of subsidiary variables is to *ignore* causal information that is not of interest. For example, having discovered the unsupervised effects of images on a brain region the neuroscientist might want to focus on the subsidiary effects whose analogues were observed in other brain regions, or in other animals. Alg. 4 shows a simple (but combinatorially expensive) procedure to discover the full set of subsidiary causes and effects. The algorithm iterates over all the possible coarsenings of E, the unsupervised effect, and computes, for each, the corresponding coarsening (not necessarily strict) of the unsupervised cause that adheres to Def. 17.

To complete the picture of how the unsupervised and subsidiary variables relate to each other, we formalize the intuition that the unsupervised causal partition can be a product of its subsidiary variables. Recall that we have defined causal macro-variables as partitions of sets of values of random micro-variables. The composition of causal variables is defined in terms of the product of partitions.

Definition 19 (Partition Product, Macro-Variable Composition)**.** *Let Π_1 and Π_2 be partitions of the same set X. The product of the partitions, denoted $\Pi_1 \otimes \Pi_2$, is the coarsest partition of X that is a refinement of both Π_1 and Π_2. The set of partitions of X forms a commutative monoid under $\otimes$. The composition C of two causal macro-variables C_1 and C_2 is defined as the product of the corresponding partitions. In this case, we will use the $\otimes$ operator to write $C = C_1 \otimes C_2$.*

Finally, we describe a special class of subsidiary variables to gain additional insight into the unsupervised causal structure of causal systems.

Definition 20 (Non-Interacting Subsidiary Variables)**.** *Let C^1, C^2 be subsidiary causes with respective subsidiary effects E^1, E^2. Denote by (e_1, e_2) the cell of $E^1 \otimes E^2$ that corresponds to the intersection of a cell e^1 of E^1 and cell e_2 of E^2, and analogously for (c_1, c_2). C^1 and C^2 are* non-interactive *if for any non-empty (c_1, c_2) and (e_1, e_2) we have $P(E^1 \otimes E^2 = (e_1, e_2) \mid do(C^1 \otimes C^2 = (c_1, c_2))) = P(E^1 = e_1 \mid do(C^1 = c_1)) \times P(E^2 = e_2 \mid do(C^2 = c_2))$.*

The unsupervised causal partition gives rise to *no subsidiary causes* in almost all the cases. The presence of coarse, non-interacting subsidiary causes (such as the h-bar and the v-bar in our example) can be assumed a strong indicator of independent physical mechanisms that produce symmetries in the unsupervised causal structure of the system. Our framework enables the scientist to automatically detect such independent mechanisms from data.

For example, let C^1= "presence of h-bar", C^2= "presence of v-bar", E^1= "presence of pulse", E^2= "presence of rhythm (top)". We can discover these variables from data using Alg. 4, and check that indeed they are non-interacting. In fact, these two subsidiary variables compose to yield the unsupervised causal partition and its probability table – we can write $C = C^1 \otimes C^2$ and $E = E^1 \otimes E^2$ (see Fig. 3.4d).

3.3 Algorithms

Learning the observational partition amounts to clustering $\mathcal{L}$ such that all the h belonging to one cluster induce the same $p(eda \mid hue = h)$, and clustering $\mathcal{R}$ such that all the e in one cluster have the same likelihood $p(eda = e \mid hue)$ for any value of hue. We outline the procedure in Algorithm 2. Its most involved component is the density learning subroutine used in Line 1. Fortunately, we only need to estimate the conditional density well enough to discover its equivalence classes.

In Fig. 3.1B, the learned density differs from the ground truth. Nevertheless, we used this learned density to perform clustering on the $\mathcal{L}$ and $\mathcal{R}$ spaces into the ground-truth number of clusters (4 and 2, respectively). Fig. 3.2 shows that simple K-means clustering of the density vectors accurately discovers the observational class boundaries in both $\mathcal{L}$ and $\mathcal{R}$. Sec. 3.3.1 discusses in detail observational partition learning in the more realistic situation where the ground-truth number of macrovariable states (clusters) is unknown.

To estimate the conditional density, we used a Mixture Density Network (MDN) (Bishop, 1995) with three hidden layers of 64, 64 and 32 units and four mixture components. MDNs can be relatively easily applied to high-dimensional conditional density learning problems with large datasets, even in the online setting where new data is arriving continuously. In very high-dimensional problems, an MDN might be unable to learn the true density accurately. Nevertheless, if the ground-truth generative model has a discrete macrovariable structure, we can expect the mixture coefficients to have similar values within each observational class as long as the number of components is not significantly smaller than the number of observational classes.

3.3.1 Choosing the Number of States

In Fig. 3.2 we provided the algorithm with the ground-truth number of observational states. In practice we want to *learn* the variables starting only from continuous microvariable data – their *a priori* unkown cardinalities must also be discovered. A solution we propose is to run Alg. 2 with N_h and N_e (the target number of observational classes for $\mathcal{L}$ and $\mathcal{R}$) slightly larger than our best guess. Steps 8-17 then *merge* the appropriate classes to obtain the observational partition.

This procedure is based on the assumption that the density learning and clustering steps return to a good approximation a *refinement* of the observational partition. In the limit of infinite samples and a good density

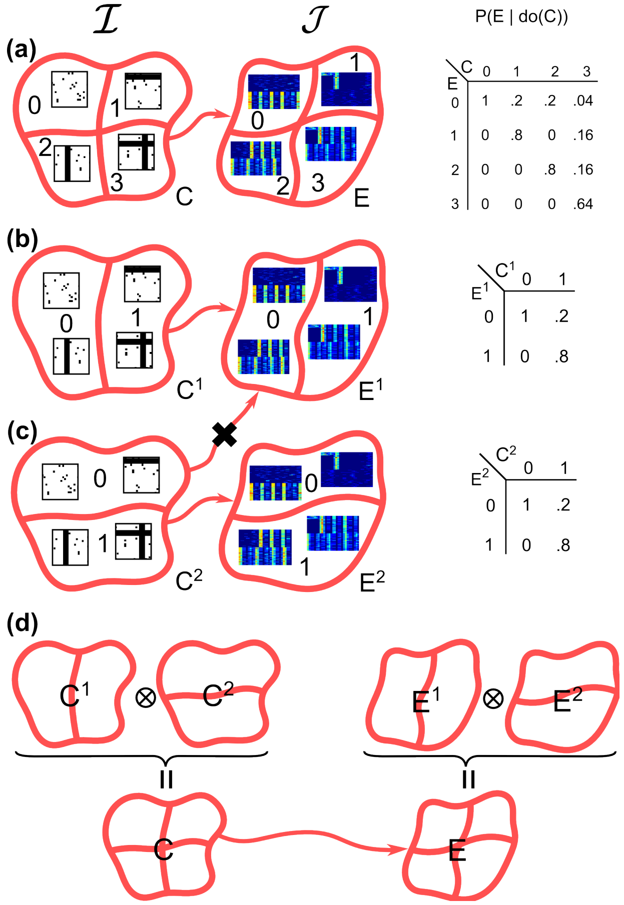


Figure 3.4: **Subsidiary Causal Variables**. **(a)** The unsupervised cause and effect of our neuroscience example. **(b)** The subsidiary cause C^1, "presence of an h-bar". The corresponding coarsening of C groups together the images which contain no h-bars ($C^1 = 0$) and the images which contain an h-bar ($C^1 = 1$). Similarly, the subsidiary effect of C^1 groups together raster plots with and without the "pulse" behavior. **(c)** The subsidiary cause C^2, "presence of a v-bar" and its effect E^2. Note that E^1, for example, is *not* an effect of C^2. If it was, the effects of manipulations $\text{do}(C^2 = 0)$ as well as $\text{do}(C^2 = 1)$ would be ambiguous: $P(E^1 = 1 \mid \text{do}(C^2 = 1))$ could be either .8 or 0, depending on whether the manipulated micro-variable contains an h-bar or not. **(d)** C^1 and C^2 are non-interacting subsidiary causes. The effect of their product is the product of their effects.

Algorithm 2: Learning the Unsupervised Observational Partition

```
input  : $\{(h_1, e_1), \cdots, (h_N, e_N)\}$ – observational microvariable data.
         $N_h, N_e$ – number of observational classes to learn.
         DensityLearning – a conditional density learning routine.
         Cluster – a clustering routine.
         EMD – earth mover's distance routine.
         $\theta_{emd}$ – EMD histogram similarity threshold.
output : $H' : \mathcal{L} \rightarrow \{1, \cdots, N_h\}$ – the $\mathcal{L}$ observational partition.
         $E' : \mathcal{R} \rightarrow \{1, \cdots, N_e\}$ – the $\mathcal{R}$ observational partition.

$p_{e|h} \leftarrow$ DensityLearning($\mathcal{D}_{csl}$);
$Eft_{mic} \leftarrow \{[p_{e|h}(h, e_1), \cdots, p_{e|h}(h, e_N)] \mid h \in \mathcal{L}\}$;
$Cs_{mic} \leftarrow \{[p_{e|h}(h_1, e), \cdots, p_{e|h}(h_N, e)] \mid e \in \mathcal{R}\}$;
$H' \leftarrow$ Cluster($Eft_{mic}$);                                   // range($H'$) $= \{1, \cdots, N_h\}$
$E' \leftarrow$ Cluster($Cs_{mic}$);                                    // range(E')$=\{1, \cdots, N_e\}$
$Eft_{mac} \leftarrow \{[P(e'_1|h'), ..., P(e'_{N_e}|h')] \mid h' = 1, ..., N_h\}$;
$Cs_{mac} \leftarrow \{[P(e'|h'_1), ..., P(e'|h'_{N_h})]/(P(e'|h'_1) + ... + P(e'|h'_{N_h})) \mid e' = 1, ..., N_e\}$;
for $h'_i, h'_j \in H' \times H'$ do
    if EMD $(Eft_{mac}(h'_i), Eft_{mac}(h'_j)) < \theta_{emd}$ then
        Merge $H'$ clusters $h'_i$ and $h'_j$;
    end
end
for $e'_i, e'_j \in E' \times E'$ do
    if EMD $(Cs_{mac}(e'_i), Cs_{mac}(e'_j)) < \theta_{emd}$ then
        Merge $E'$ clusters $e'_i$ and $e'_j$;
    end
end
```

Algorithm 3: Learning the Unsupervised Macrovariable Cause

```
input  : $D_{obs} = \{(h_1, e_1), \cdots, (h_N, e_N)\}$ – observational microvariable data.
         EMD – Earth Mover's Distance routine.
         $\theta_{emd}$ – threshold on Earth Mover's Distance similarity.
output : $C : \mathcal{L} \rightarrow \{1, \cdots, N_c\}$ – the $\mathcal{L}$ causal partition.

$H', E' \leftarrow$ Run Algorithm 2 on $D_{obs}$ to obtain the observational partitions on $\mathcal{L}$ and $\mathcal{R}$;
$h_1, \cdots, h_{N_{H'}} \leftarrow$ Pick one representative for each $h'_i \in range(H')$ s.t. $H'(h_i) = h'_i$;
$e_1, \cdots, e_{N_{E'}} \leftarrow$ Pick one representative for each $e'_i \in range(E')$ s.t. $E'(e_i) = e'_i$;
Estimate $P(E' \mid \text{man}(h_i))$ for each representative $h_i$;
for $h'_i, h'_j \in range(H') \times range(H')$ do
    if EMD $(P(E' \mid man(h_i)), P(E' \mid man(h_j))) < \theta_{emd}$ then
        Merge $H'$ clusters $h'_i$ and $h'_j$;
    end
end
$C = \{c_1, \cdots, c_{N_c}\} \leftarrow$ merged H';
```

learning and clustering algorithm this should always be true.

Figure 3.5A illustrates the result of running our algorithm on toy data with $N_h = N_e = 6$ (as opposed to the ground-truth $N_h = 4, N_e = 2$). The algorithm divided $\mathcal{L}$ into six groups, which are close to a refinement of the true observational partition. The pink group (spanning about $hue \in (160, 190)$ crosses

Algorithm 4: Finding Subsidiary Variables

```
input  : $C, E$ – the unsupervised cause and effect (and the corresponding partitions).
output: $\mathcal{S} = (C^1, E^1), \cdots, (C^N, E^N)$ – subsidiary variables of the system.
$\mathcal{S} \leftarrow \emptyset$;
$c_1, \cdots, c_m \leftarrow \texttt{range}(C)$;
$e_1, \cdots, e_n \leftarrow \texttt{range}(E)$;
for $\bar{E} \in \texttt{Partitions}(E)$ do
    for $\bar{e} \in \texttt{range}(\bar{E})$ do
        $P(\bar{e} \mid \text{do}(C = c_k)) \leftarrow \sum_{e_l \in \bar{e}} P(e_l \mid \text{do}(C = c_k))$;
    end
    Define $\texttt{effect}: c_k \mapsto P(\bar{E} \mid \text{do}(C = c_k))$;
    Let $c_i \sim_{\bar{C}} c_j \Leftrightarrow \texttt{effect}(c_i) = \texttt{effect}(c_j)$;
    $\Pi_{\bar{C}} \leftarrow$ partition of $\texttt{range(}C\texttt{)}$ induced by $\sim_{\bar{C}}$;
    $\bar{C} \leftarrow$ random variable corresponding to $\Pi_{\bar{C}}$;
    $\mathcal{S} \leftarrow \mathcal{S} \cup (\bar{C}, \bar{E})$;
end
```

the true observational boundary at $hue = 180$. This error type can be ascribed to low sample numbers and clustering mistakes and is hard to avoid given finite sampling.

Given a refinement of the observational partition on $\mathcal{L}$, it is easy to recover the true observational partition. If any two clusters h'_i and h'_j are subsets of the same observational state, then $P(E' \mid h'_i)$ should be similar to $P(E' \mid h'_j)$, where E' is (the refinement of) the observational partition on $\mathcal{R}$. In Fig. 3.5B, the i-th column corresponds to the empirical $P(E' \mid h'_i)$ where E' and H' are the 6-state observational variables. Fig. 3.5C shows the result of merging these E' states whose corresponding columns in Fig. 3.5B have Earth Mover's Distance (Levina and Bickel, 2001) smaller than .2. Due to sampling errors, it deviates from the ground truth slightly, but contains four states as expected.

Figure 3.5D-F shows the merging process for $\mathcal{R}$. The end result is almost exactly the ground truth. The merging process for $\mathcal{R}$ requires a slight modification, since one cannot simply merge e'_i and e'_j when $P(e'_i \mid h') = P(e'_j \mid h')$ for any h'. To see why, consider clusters e'_0 and e'_1 in Fig. 3.5D (counting from the top of the plot, the dark-green and the brown clusters). Both clusters are subsets of the same ground-truth observational cell, but e'_0 consists of significantly fewer samples. As a result, the vector $[P(e'_0 \mid h'_0), \cdots, P(e'_0 \mid h'_5)]$ is a *scaled version* of $[P(e'_1 \mid h'_0), \cdots, P(e'_1 \mid h'_5)]$ (see the first and second *rows* in Fig. 3.5B). Normalizing the likelihood vectors such that they sum to 1 solves the problem. It is always true that if two clusters are subsets of the same observational class, their normalized likelihood vectors are (in the limit of infinite sample size) the same.

3.4 Experiments

Consider a dataset $\{(i, j)\}$ of size N generated experimentally from a causal system with input and output spaces $\mathcal{I}$ and $\mathcal{J}$: each i is chosen by the experimenter arbitrarily, and each j is generated from $P(J \mid \text{man}(i))$.

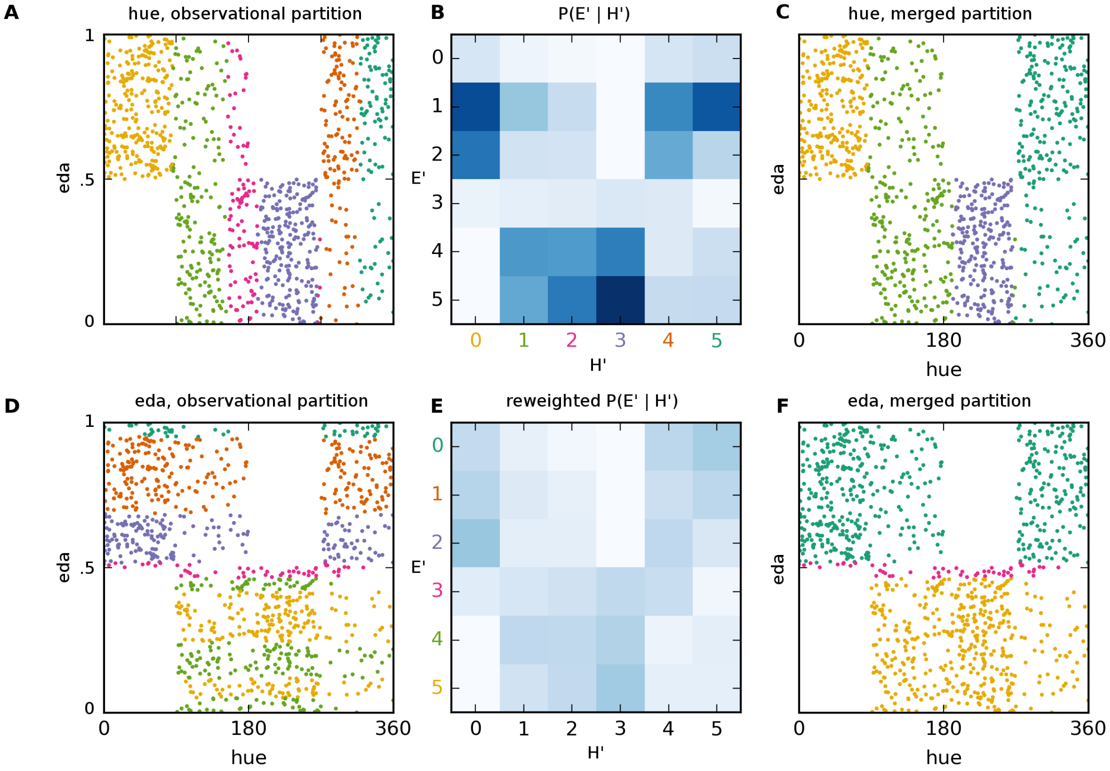


Figure 3.5: **Observational Partition Learning — Unknown Number of States**. A) We clustered hue w.r.t. $p(eda \mid hue)$ into six (as opposed to the ground truth four) clusters $H' = 0, \cdots, 5$. With enough samples and good density learning, the overclustering should *refine* the true observational partition. We repeated the procedure in $\mathcal{R}$, clustering reaction states into six (instead of two) clusters E'. B) Computing empirical $P(E' \mid H')$ shows that $H' = 1, 2$ induce similar conditionals on E'. $H' = 4, 5$ also induce similar probabilities. C) Merging clusters with similar conditionals brings us close to the ground truth observational partition (compare with Fig. 3.2). We merged clusters whose Earth Mover's Distance is less than .1. D-F) A similar merging procedure is repeated for E' clusters. In this case, we renormalized the likelihoods to account for different sample counts in clusters with the same $P(E' \mid H')$ (see text for details). Two of the merged clusters correspond well to the ground-truth. Because of sample-size issues a small additional cluster a the boundary of the true classes was detected (colored pink).

Algorithm 3 can take such data as input, and compute the unsupervised cause and effect of the system. Here we provide a step-by-step illustration of the algorithm's application to the simulated neuroscience problem from Sec. 1.3.3.

We generated 10000 images i similar to those shown in Fig. 1.6: 2500 h-bar images (with varying h-bar locations and uniform pixel noise), 2500 v-bar images, 2500 "h-bar + v-bar" images and 2500 uniform noise images. Of course, this is an ideal dataset that we can only design because we know the ground-truth causal features. In practice, the experimenter would want to choose as broad a class of stimuli as reasonable. Next, for each image we generated a corresponding time-averaged, neuron-index-shuffled raster plot j according to $P(J \mid \text{man}(i))$. We then applied Alg. 3 to this experimental data. The output is for each image i an estimate of its causal class $C(i)$, and for each raster j an estimate of its effect class $E(j)$, as defined in Fig. 1.6.

Figure 3.6 shows how Alg. 3 recovers the macro-variable causal mechanism of our simulated single-unit-recording experiment. Two remarks are in order:

1. For purposes of illustration, the macro-level causal variables are very simple. Nevertheless, the procedure is completely general and could be applied to detect causal macro-variables that do not admit such a simple description. We believe the method holds promise for applications in a broad set of scientific domains.
2. The algorithm does not simply cluster $\mathcal{I}$ and $\mathcal{J}$. Instead, it clusters the probabilistic effects of points in $\mathcal{I}$, and the probabilities of causation for points in $\mathcal{J}$. Its crucial function is to ignore any structures that are not related to the causal effect of I on J. In our example, the raster plots contain salient structure that is causally irrelevant: With probability 0.5, the "bottom" subpopulation of neurons spikes in a synchronized rhythm. Simply clustering $\mathcal{J}$ would sub-divide the true causal classes in half. Fig. 3.6e shows that the algorithm finds the correct solution.

3.5 Proofs

Theorem (Unsupervised Causal Coarsening) *Among all the joint distributions $P(T, H, I)$ over discrete variables T, H, I, consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{I})$ and a fixed confounding partition $\Pi_{P(T|I)}(\mathcal{I})$. Within this subset, the following two statements hold:*

1. *The subset of distributions for which $\Pi_c(\mathcal{I})$ is* not *a coarsening of the observational partition $\Pi_o(\mathcal{I})$ is Lebesgue measure zero, and*
2. *The subset of distributions for which $\Pi_c(\mathcal{J})$ is* not *a coarsening of the observational partition $\Pi_o(\mathcal{J})$ is Lebesgue measure zero.*

Proof. (1) $\Pi_c(\mathcal{I})$, $\Pi_o(\mathcal{I})$ and $\mathcal{J}$ can be treated as the causal partition, observational partition and target variable as defined in Chapter 3. Thus we can directly use the proof of Theorem 6 to prove (1).

(2) We cannot use Theorem 6 to prove (2), but we use the same strategy with some differences in the

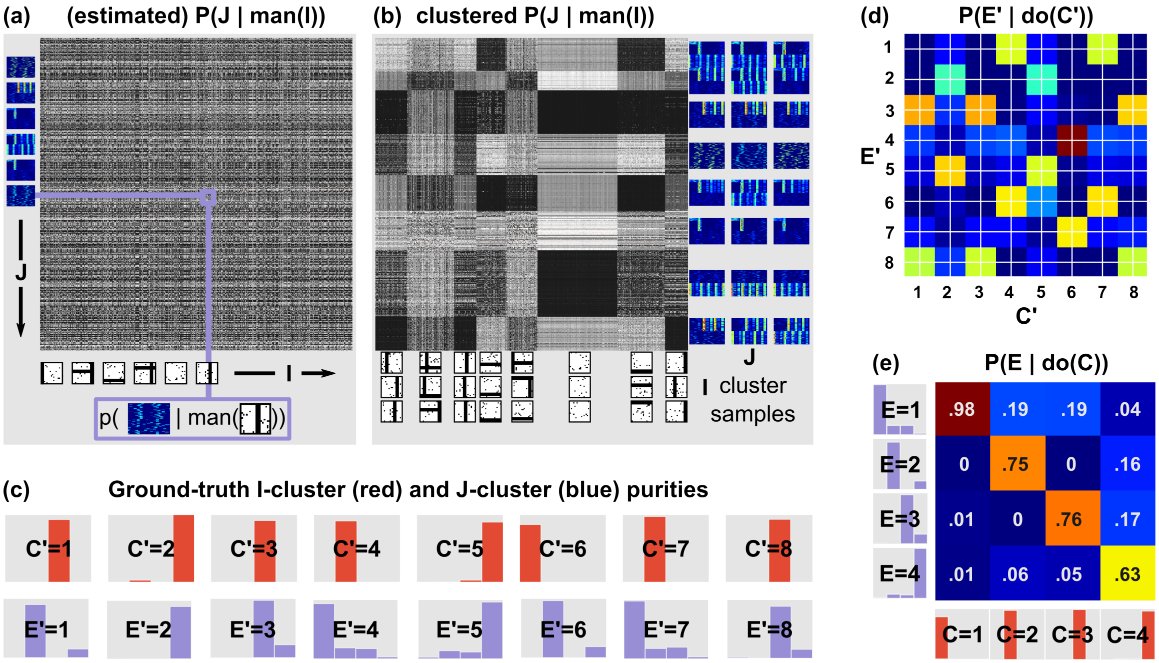


Figure 3.6: **Learning the unsupervised causal partition**. The figure demonstrates Algorithm 3 applied to the example from Fig. 1.6. **(a)** Given a dataset $\{(i_k, j_k)\}_{k=1\dots N}$, the algorithm learns data density $P(j \mid \text{man}(i))$ and forms a matrix in which the kl-th entry is the estimated $P(j_k \mid \text{man}(i_l))$. **(b)** The rows and columns of the matrix are clustered. Each cluster of columns corresponds to a cell of C', the proposed unsupervised partition of $\mathcal{I}$, and each cluster of rows corresponds to a cell of E', the proposed unsupervised partition of $\mathcal{J}$. **(c)** The histograms show the ground-truth causal class of the points in each cluster (this ground truth is unknown to the algorithm). For example, the cell $E' = 8$ contains a majority of raster plots that contain the "30Hz (top)" causal structure; it also contains some "30Hz (top) + pulse" rasters. **(d)** The algorithm computes the probability table $P(E' \mid \text{do}(C'))$ by counting the co-occurrences of the cluster labels. **(e)** Finally, the columns of this table are merged according to their similarity to form the unsupervised partition Π_C, and the rows are merged to form Π_E. For example, columns $C' = 1$ and $C' = 3$ of the table in (d) are similar—indeed, the cluster purity histograms indicate that both rows correspond to sets of images with a vertical bar. $P(E \mid \text{do}(C))$ is very similar to the ground-truth table (see Fig. 1.6), and the final C, E clusters are pure (as shown along the axes of the table).

details of the algebra.

(i) We first set up the notation. Let H be the hidden variable of the system, with cardinality K; let J have cardinality N and I cardinality M. We can factorize the joint on I, J, H as $P(J, I, H) = P(J \mid H, I)P(I \mid H)P(H)$. $P(J \mid H, I)$ can be parametrized by $(N-1) \times K \times M$ parameters, $P(I \mid H)$ by $(M-1) \times K$ parameters, and $P(H)$ by $K-1$ parameters, all of which are independent.

Call the parameters, respectively,

$$\alpha_{j,h,i} \triangleq P(J = j \mid H = h, I = i)$$

$$\beta_{i,h} \triangleq P(I = i \mid H = h)$$

$$\gamma_h \triangleq P(H = h)$$

We will denote parameter vectors as

$$\begin{aligned}\alpha &= (\alpha_{j_1,h_1,i_1}, \cdots, \alpha_{j_{N-1},h_K,i_M}) \in \mathbb{R}^{(N-1)\times K\times M}\\ \beta &= (\beta_{i_1,h_1}, \cdots, \beta_{i_{N-1},h_K}) \in \mathbb{R}^{(M-1)\times K}\\ \gamma &= (\gamma_{h_1}, \cdots, \gamma_{h_{K-1}}) \in \mathbb{R}^{K-1},\end{aligned}$$

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution $P(J, H, I)$ with a point $(\alpha, \beta, \gamma) \in P[\alpha, \beta, \gamma] \subset \mathbb{R}^{(N-1)\times K^2(K-1)\times M(M-1)}$.

(ii) Show that for any α, β consistent with Π_c and $\Pi_{P(I|H)}$, the causal partition and the confounding partition are, in general, fixed.

To proceed with the proof, pick any point in the $P(J \mid H, I) \times P(I \mid H)$ space – that is, fix α and β. The only remaining free parameters are now in γ. Varying these values creates a subset of the space of all joints isometric to the $(K-1)$-dimensional simplex of multinomial distributions over K states (call the simplex S_{K-1}):

$$P[\gamma; \alpha, \beta] = \{(\alpha, \beta, \gamma) \mid \gamma \in S_{K-1}\} \subset [0, 1]^{(K-1)}.$$

Note that fixing β directly fixes $\Pi_{P(I|H)}$. Fixing α doesn't directly fix Π_c. But by Lemma 10, for *almost all* distributions in $P[\gamma; \alpha, \beta]$ the causal partition Π_c equals the partition $\Pi_{P(T|H,I)}$, which is directly fixed by α. Let $P'[\gamma; \alpha, \beta]$ be $P[\gamma; \alpha, \beta]$ minus this measure zero subset.

The statement of the theorem fixes Π_c and $\Pi_{P(I|H)}$. If the α, β we picked are consistent with these partitions within $P'[\gamma; \alpha, \beta]$, continue with the proof. Otherwise, choose other α, β.

We now prove that within $P'[\gamma; \alpha, \beta]$ the set of γ for which the causal partition $\Pi_c(\mathcal{J})$ is not a coarsening of the observational partition $\Pi_o(\mathcal{J})$ is of measure zero. Later in (iv) we integrate the result over all α, β.

(iii) Let the causal coarsening constraint be that for $j_1, j_2 \in \mathcal{J}$ we have

$$O(j_1) = O(j_2) \quad \Rightarrow \quad C(j_1) = C(j_2). \tag{3.1}$$

That is, it is not the case that two members of $\mathcal{J}$ are observationally equivalent but have different likelihoods of causation.

We show that the causal coarsening constraint holds for each pair $j_1, j_2 \in \mathcal{J}$: Pick any $j_1, j_2 \in \mathcal{J}$. If $C(j_1) = C(j_2)$, then we are done with this pair. So assume that there is a causal difference, i.e. $C(j_1) \neq C(j_2)$. Our goal is now to show that then only a measure-zero subset of $P'[\gamma; \alpha, \beta]$ allows for $O(j_1) = O(j_2)$.

We first show that equivalence of the observational classes ($O(j_1) = O(j_2)$) places a polynomial constraint on $P'[\gamma; \alpha, \beta]$. By definition, we have

$$\begin{aligned} O(j_1) = O(j_2) \quad &\Leftrightarrow \\ \forall_i \sum_h \alpha_{j_1,h,i}\beta_{i,h}\gamma_h = &\sum_h \alpha_{j_1,h,i}\beta_{i,h}\gamma_h \end{aligned}$$

Pick an arbitrary i. The above equation places the following polynomial constraint, for this i, on $P'[\gamma; \alpha, \beta]$:

$$\sum_h \gamma_h(\alpha_{j_1,h,i}\beta_{i,h} - \alpha_{j_2,h,i}\beta_{i,h}) = 0. \tag{3.2}$$

We have thus shown that, for fixed α, β and j_1, j_2, the violation of the causal coarsening constraint (3.1) places a polynomial constraint on $P'[\gamma; \alpha, \beta]$. By an algebraic lemma (proven by Okamoto, 1973), the subset on which the constraint holds is measure zero *if the constraint is not trivial*. That is, we only need to find one γ for which Eq. (3.2) does not hold to prove that it almost never holds.

To find such γ, let $\gamma_h = 1/K$ for all h. If for this γ Eq. (3.2) does not hold, we are done. If it does hold, since we know α is not all 0, there must be in the sum of the equation at least one factor $[\alpha_{j_1,h,i}\beta_{i,h} - \alpha_{j_1,h,i}\beta_{i,h}]$ which is positive, and one that is negative. Call the h corresponding to the positive element h_+ and to the negative element h_-. Pick any positive $\epsilon < min(1/K, 1 - 1/K)$. Set $\gamma_h = 1/K$ for all $h \neq h_+, h_-$ and set $\gamma_{h_+} = \frac{1}{K} + \epsilon$ and $\gamma_{h_-} = \frac{1}{K} - \epsilon$. In this way, we keep $\sum_h \gamma$ unchanged, and are guaranteed that Eq. (3.2) does not hold. That is, for this γ we have $O(j_1) \neq O(j_2)$[1].

The rest of the proof follows exactly as the proof of Thm. 2.5. □

Theorem (Unsupervised Sufficient Causal Description) *Let $(\mathcal{I}, \mathcal{J})$ be a causal system let C and E be its unsupervised cause and effect. Let $\mathbf{E}$ be E applied sample-wise to a sample from the system (so that e.g. $\mathbf{E}(j_1, \cdots, j_k) = (E(j_1), \cdots, E(j_k))$). Then:*

1. *Among all the partitions of $\mathcal{J}$, $\mathbf{E}$ is the* minimal *sufficient statistic for $P(J \mid man(i))$ for any $i \in \mathcal{I}$, and*
2. *C and E losslessly recover $P(j \mid man(i))$. No other (subsidiary) causal variable losslessly recovers $P(j \mid man(i))$. Any other partition is either finer than C, E or does not define unambiguous manipulations. In this sense, the unsupervised causal partition corresponds to the coarsest partition that losslessly recovers $P(j \mid man(i))$.*

Proof. 1. We first prove that $\mathbf{E}$ is a sufficient statistic. Recall that we assumed $\mathcal{J}$ to be discrete, although possibly of vast cardinality. For any $j_k \in \mathcal{J}$, write $P(j_k \mid \text{man}(i)) = p_{j_k}$ for the corresponding categorical distribution parameter. Let $\texttt{range}(E) = \{E_1, \cdots, E_M\}$ be the set of causal classes of J. By Definition 3

[1] It is possible that this γ is not in $P'[\gamma; \alpha, \beta]$. However, it is guaranteed to be in $P[\gamma; \alpha, \beta]$. Since a subset of measure zero in $P[\gamma; \alpha, \beta]$ is also measure zero in $P'[\gamma; \alpha, \beta]$, this does not influence the proof.

there is a number of “template” probabilities $p_{E_1}, \cdots, p_{E_M}$ such that $p_{j_k} = p_{E_k}$ if and only if $E(j_k) = E_k$. Consider an i.i.d. sample $\mathbf{j} = j_1, \cdots, j_l$ from $P(J \mid \text{man}(i))$. Then

$$\begin{aligned} P(j_1, \cdots, j_l \mid \text{man}(i)) &= \Pi_{k=1}^{l} p_{j_k} \\ &= \Pi_{m=1}^{M} p_{E_m}^{\#(E_m)}, \end{aligned}$$

where $\#(E_m) \triangleq \Sigma_{k=1}^{l} \mathbb{1}\{E(j_k) == E_m\}$ is the number of samples with causal class E_m. Since the sample density depends on the samples only through C and E it follows from Fisher's factorization theorem that $\mathbf{E}$ is a sufficient statistic for $P(J \mid \text{man}(i))$ for any $i \in \mathcal{I}$.

Now, we prove the minimality of E among all the partitions of $\mathcal{J}$. Consider first any refinement of E. One can directly apply the reasoning above to show that the cell assignment in such a partition is also a sufficient statistic. However, any refinement is not the *minimal* sufficient statistic, as the unsupervised causal partition is its coarsening— and thus also its function. Now, consider any partition that is not the unsupervised causal partition, and is not its refinement. Call it E'. Assume, for contradiction, that $\mathbf{E}'$ is a sufficient statistic for $P(J \mid \text{man}(i))$. Then, by the factorization theorem, $P(j_1, \cdots, j_k \mid \text{man}(i))$ would factorize as $h(j_1, \cdots, j_k)g(E'(j_1), \cdots, E'(j_k))$, where h does not depend on the parameters p_{j_l}. Now, take some j_1^1, j_1^2 such that $E(j_1^1) \neq E(j_1^2)$ but $E'(j_1^1) = E'(j_1^2)$ (such a pair must exists since E' is not a refinement of E and is not equal to it). Then

$$\frac{P(j_1^1, j_2, \cdots, j_k \mid \text{man}(i))}{P(j_1^2, j_2, \cdots, j_k \mid \text{man}(i))} = \frac{p_{E(j_1^1)}}{p_{E(j_1^2)}},$$

$$\frac{P(j_1^1, j_2, \cdots, j_k \mid \text{man}(i))}{P(j_1^2, j_2, \cdots, j_k \mid \text{man}(i))} = \frac{h(j_1^1, \cdots, j_k)g(E'(j_1^1), \cdots, E'(j_k))}{h(j_1^2, \cdots, j_k)g(E'(j_1^2), \cdots, E'(j_k))} = \frac{h(j_1^1, \cdots, j_k)}{h(j_1^2, \cdots, j_k)}$$

which, as already stated, does not depend on the parameters of the distribution – a contradiction.

2. That $P(J \mid \text{man}(i))$ can be recovered from C and E follows directly from the definition of an unsupervised causal partition. That it cannot be recovered losslessly from any partition that is not a refinement of C and E follows again from the fact that for any such partitions C' and E' there must be is at least one pair $(i_1, j_1), (i_2, j_2)$ for which $p(E'(j_1) \mid \text{do}(C'(i_1))) = p(E'(j_2) \mid \text{do}(C'(i_2)))$ even though $p(j_1 \mid \text{man}(i_1)) \neq p(j_2 \mid \text{man}(i_2))$. □

We note that the first part of the Sufficient Causal Description Theorem indicates that $\mathbf{E}$ is only a minimal sufficient statistic among all partitions of $\mathcal{J}$, i.e. among the set of possible causal variables. It is not the minimal sufficient statistic over all possible sufficient statistics for $P(J \mid \text{man}(i))$. In particular, a histogram is a minimal sufficient statistic for the multinomial distribution and is a function of $\mathbf{E}$, but a histogram does not correspond to a partition of $\mathcal{J}$.

Chapter 4

Learning Optimal Interventions

Chapters 2 and 3 developed theory and algorithms to learn causal features. Sometimes, knowledge of the causal features can be of interest by itself – for example in neuroscience. At other times, one wants to know causal mechanisms of a system in order to intervene on the causes to achieve desired results. In health science, for example, the goal is to understand what causes good health (or disease) in order to intervene on the causes to increase health and decrease disease.

Knowledge of C, the macrovariable cause of a system, is a starting point to learning a *manipulator function*: a function that, given any microvariable instance, constructs the smallest perturbation to the instance that has the desired causal effect.

There are several reasons why we might want such a manipulator function:

- If our goal is to perform causal manipulations on the system, the manipulator function offers an automated solution.
- A manipulator that uses a given C and produces the desired causal effect provides strong evidence that C is indeed the causal macrovariable.
- The manipulator function can enrich the dataset, in hope of achieving better generalization on both the causal and predictive learning tasks.

4.1 Advances in This Chapter

This chapter shows how the concepts of Chapters 2 and 3 can be used to automatically design optimal control of a causal system.

4.2 Theory

We develop the manipulator function within the supervised CFL framework (and to easier refer to ideas in Chapter 2, use the visual causes example). Extension to the unsupervised case is trivial.

Definition 21 (Manipulator Function)**.** *Let C be the causal macrovariable of T and d a metric on $\mathcal{I}$. The* manipulator function *of C is a function $M_C\colon \mathcal{I} \times \mathcal{C} \to \mathcal{I}$ such that $M_C(i,k) = \arg\min_{\hat{i}\in C^{-1}(k)} d(i,\hat{i})$ for any $i \in \mathcal{I}, k \in \mathcal{C}$. In case $d(i,.)$ has multiple minima, we group them together into one equivalence class and leave the choice of the representative to the manipulator function.*

The manipulator searches for an image closest to I among all the images with the desired causal effect k. The meaning of "closest" depends on the metric d. The choice of d is task-specific and crucial to the quality of the manipulations. In our experiments, we use a metric induced by an L_2 norm. Alternatives include other L_p-induced metrics, distances in implicit feature spaces induced by image kernels (Harchaoui and Bach, 2007; Grauman and Darrell, 2007; Bosch et al., 2007; Vishwanathan, 2010) and distances in learned representation spaces (Bengio et al., 2013).

Note that the manipulator function can find candidates for the image manipulation underlying the desired causal manipulation do($C = c$), but it does not check whether other variables in the system (in particular, the spurious correlate) remain in fact unchanged. Using the closest possible image with the desired causal effect is a heuristic approach to fulfilling that requirement.

4.3 Algorithms

Algorithm 5 proposes one way to learn the manipulator function using a simple manipulation procedure that approximates the requirements of Definition 21 up to local minima.

Algorithm 5: Manipulator Function Learning

```
input : $d\colon \mathcal{I} \times \mathcal{I} \to \mathbb{R}_+$ – a metric on the image space
        $\mathcal{D}_{csl} = \{(i_1, c_1), \cdots (i_N, c_N)\}$ – causal data
        $\mathcal{C} = \{C_1, \cdots, C_M\}$ – the set of causal classes (so that $\forall i, c_i \in \mathcal{C}$)
        Train – a neural net training algorithm
        nIters – number of experiment iterations
        Q – number of queries per iteration
        $\alpha$ – manipulation tuning parameter
        A: $\mathcal{I} \to \mathcal{C}$ – an oracle for $P(T \mid \text{do}(I))$
output: $M_C\colon \mathcal{I} \times \mathcal{C} \to \mathcal{I}$ – the manipulator function
1 for $l \leftarrow 1$ to nIters do
2     $C \leftarrow$ Train($\mathcal{D}_{csl}$);
3     Choose manipulation starting points $\{i_{l,1}, \cdots, i_{l,\mathsf{Q}}\}$ at random from $\mathcal{D}_{csl}$;
4     Choose manipulation targets $\{\hat{c}_{l,1}, \cdots, \hat{c}_{l,\mathsf{Q}}\}$such that $\hat{c}_{l,k} \neq c_{l,k}$;
5     for $k \leftarrow 1$ to Q do
6         $\hat{i}_{l,k} \leftarrow \underset{j\in\mathcal{I}}{\text{argmin}}\ (1-\alpha)|C(j) - \hat{c}_{l,k}| + \alpha\, d(j, i_{l,k})$;
7     end
8     $\mathcal{D}_{csl} \leftarrow \mathcal{D}_{csl} \cup \{(\hat{i}_{l,1}, \mathtt{A}(\hat{i}_{l,1})), \cdots, (\hat{i}_{l,\mathsf{Q}}, \mathtt{A}(\hat{i}_{l,\mathsf{Q}}))\}$;
9 end
```

The algorithm, inspired by the active learning techniques of uncertainty sampling (Lewis and Gale, 1994) and density weighting (Settles and Craven, 2008), starts off by training a causal neural network in Step 2. If only observational data is available, this can be achieved using algorithms of Chapters 2 and/or 3. Next, it randomly chooses a set of images to be manipulated, and their target post-manipulation causal labels. The loop that starts in Step 6 then takes each of those images and searches for the image that, among the images with the same desired causal class, is closest to the original image. Note that the causal class boundaries are defined by the current causal neural net C. Since C is in general a highly nonlinear function and it can be hard to find its inverse sets, we use an approximate solution. The algorithm thus finds the minimum of a weighted sum of $|C(j) - \hat{c}_{l,k}|$ (the difference of the output image j's label and the desired label $\hat{c}_{l,k}$) and $d(i_{l,k}, j)$ (the distance of the output image j from the original image $i_{l,k}$).

At each iteration, the algorithm performs Q manipulations and the same number of causal queries to the agent, which result in new datapoints $(\hat{\imath}_{l,1}, A(\hat{\imath}_{l,1})), \cdots, (\hat{\imath}_{l,Q}, A(\hat{\imath}_{l,Q}))$. It is natural to claim that the manipulator performs well if $A(\hat{\imath}_{l,k}) \approx \hat{c}_{l,k}$ for many k, which means the target causal labels agree with the true causal labels. We thus define the *manipulation error* of the lth iteration $MErr_l$ as

$$MErr_l = \frac{1}{Q}\sum_{k=1}^{Q} |A(\hat{\imath}_{l,k}) - \hat{c}_{l,k}|. \tag{4.1}$$

While it is important that our manipulations are accurate, we also want them to be minimal. Another measure of interest is thus the *average manipulation distance*

$$MDist_l = \frac{1}{Q}\sum_{k=1}^{Q} d(I_{l,k}, \hat{\imath}_{l,k}). \tag{4.2}$$

A natural variant of Algorithm 5 is to set $nIters$ to a large integer and break the loop when one or both of these performance criteria reaches a desired value.

4.4 Experiments

In order to illustrate the concept of learning a manipulator we perform two causal feature learning experiments. The first experiment, called GRATING, uses observational and causal data generated by the model defined in Sec 1.3.1. The GRATING experiment confirms that our system can learn the ground truth cause and ignore the spurious correlates of a behavior. The second experiment, MNIST, uses images of hand-written digits (LeCun et al., 1998) to exemplify the use of the manipulator function on slightly more realistic data: in this example, we transform an image into a maximally similar image with another class label.

We chose problems that are simple from the computer vision point of view. Our goal is to develop the theory of visual causal feature learning and show that it has feasible algorithmic solutions; we are at this point not engineering advanced computer vision systems.

4.4.1 The GRATING Experiment

In this experiment we generate data using the model of Sec. 1.3.1, with two minor differences: H_1 and H_2 only induce one v-bar or h-bar in the image and we restrict our observational dataset to images with only about 3% of the pixels filled with random noise (see Fig. 4.1). Both restrictions increase the clarity of presentation. We use Algorithms 1 and 5 (with minor modifications imposed by the binary nature of the images) to learn the visual cause of behavior T.

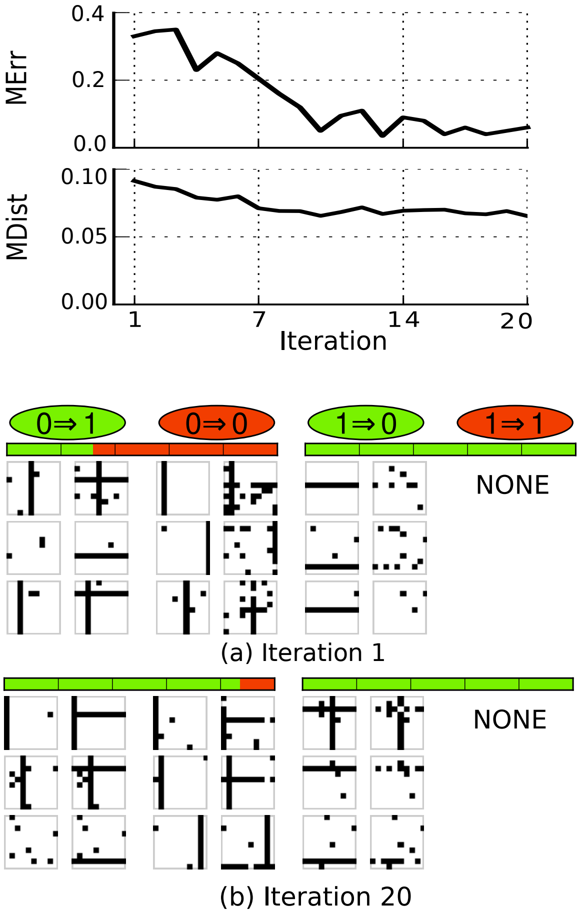
Figure 4.1 (top) shows the progress of the training process. The first step (not shown in the figure) uses the CCT to learn the causal labels on the observational data. We then train a simple neural network (a fully connected network with one hidden layer of 100 units) on this data. The same network is used on Iteration 1 to create new manipulated exemplars. We then follow Algorithm 5 to train the manipulator iteratively. Fig. 4.1 (bottom) illustrates the difference between the manipulator on Iteration 1 (which fails almost 40% of the time) and Iteration 20, where the error is about 6%. Each column shows example manipulations of a particular kind. Columns with green labels indicate successful manipulations of which there are two kinds: switching the causal variable on ($0 \Rightarrow 1$, "adding the h-bar"), or switching it off ($1 \Rightarrow 0$, "removing the h-bar"). Red-labeled columns show cases in which the manipulator failed to influence the cause: That is, each red column shows an original image and its manipulated version which the manipulator believes should cause a change in T, but which does not induce such change. The red/green horizontal bars show the percentage of success/error for each manipulation direction. Fig. 4.1 (bottom, a) shows that after training on the causally-coarsened observational dataset, the manipulator fails about 40% of the time. In Fig. 4.1 (b), after twenty manipulator learning iterations, only six manipulations out of a hundred are unsuccessful. Furthermore, the causally irrelevant image pixels are also much better preserved than at iteration 1. The fully-trained manipulator correctly learned to manipulate the presence of the h-bar to cause changes in T, and ignores the v-bar that is strongly correlated with the behavior but does not cause it.

4.4.2 The MNIST ON MTURK Experiment

In this experiment we start with the MNIST dataset of handwritten digits. In our terminology, this – as well as any standard vision dataset – is already causal data: the labels are assigned in an experimental setting, not "in nature".

Consider the following binary human behavior: $T = 1$ if a human observer answers affirmatively to the question "Does this image contain the digit '7'?", while $T = 0$ if the observer judges that the image does not contain the digit '7'. For simplicity we will assume that for any image either $P(T = 1 \mid \text{man}(I)) = 0$ or $P(T = 1 \mid \text{man}(I)) = 1$. Our task is to learn the manipulator function that will take any image and modify it minimally such that it will become a '7' if it was not before, or will stop resembling a '7' if it did originally.

We conduct the manipulator training separately for all the ten MNIST digits using human annotators on Amazon Mechanical Turk. The exact training procedure is described below. Fig. 4.2 (top) shows training



(a) Iteration 1

(b) Iteration 20

Figure 4.1: Manipulator learning for GRATING. **Top.** The plots show the progress of our manipulator function learning algorithm over twenty iterations of experiments for the GRATING problem. The manipulation error decreases quickly with progressing iterations, whereas the manipulation distance stays close to constant. **Bottom.** Original and manipulated GRATING images. See text for the details.

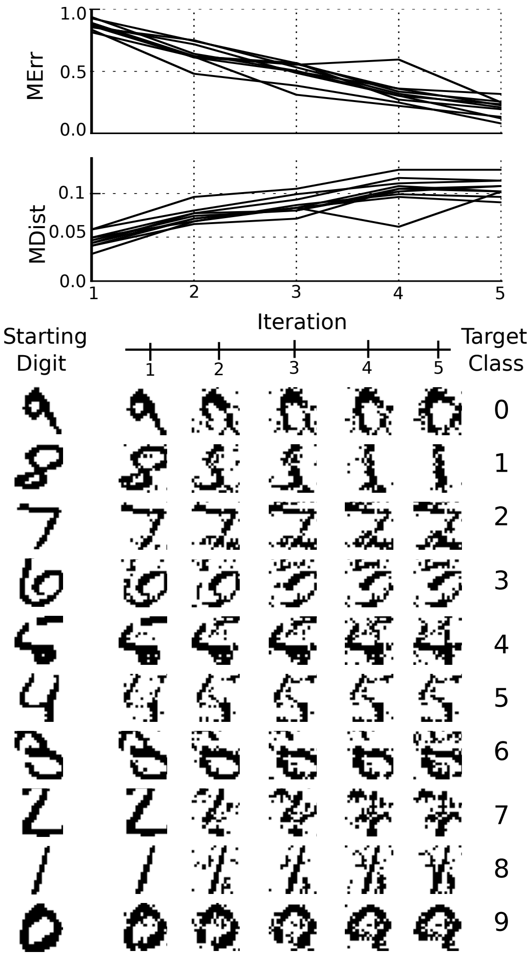


Figure 4.2: Manipulator Learning for MNIST ON MTURK. **Top.** In contrast to the GRATING experiment, here the manipulation distance grows as the manipulation error decreases. This is because a successful manipulator needs to change significant parts of each image (such as continuous strokes). **Bottom.** Visualization of manipulator training on randomly selected (not cherry-picked) MNIST digits. See text for the details.

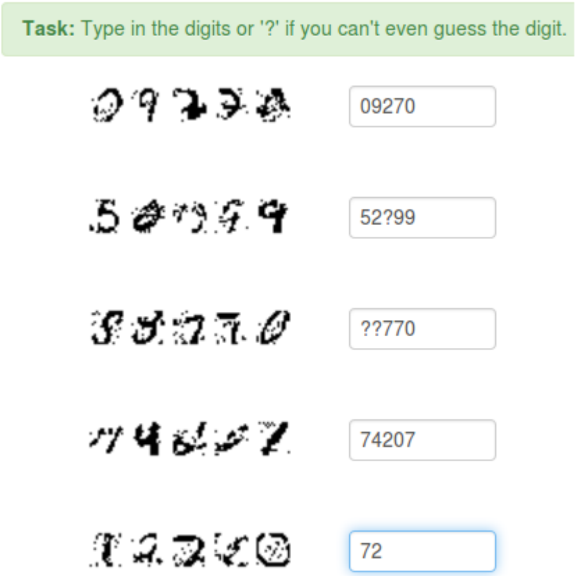
progress. As in Fig. 4.1, the manipulation error decreases with training. Fig. 4.2 (bottom) visualizes the manipulator training progress. In the first row we see a randomly chosen MNIST "9" being manipulated to resemble a "0", pushed through successive "0-vs-all" manipulators trained at iterations 0, 1, ..., 5 (iteration 1 shows what the neural net takes to be the closest manipulation to change the "9" to a "0" purely on the basis of the non-manipulated data). Further rows perform similar experiments for the other digits. The plots show how successive manipulators progressively remove the original digits' features and add target class features to the image.

For this experiment, we started off by training ten one-vs-all neural nets. We used cross-validation to choose among the following architectures: 100 hidden units (h.u.), 300 h.u. (one layer), 100-100 h.u (two layers), 300-300 h.u. (two layers). We used maxout (Goodfellow et al., 2013) activations (each of which computed the max of 5 linear functions). For training we used stochastic gradient descent in batches of 50 with 50% dropout (Hinton and Srivastava, 2012) on the hidden units, momentum adjustment from 0.5 to 0.99 at iteration 100, learning rate decaying from 0.1 to 0.0001 with exponential coefficient of 1/0.9998, no weight decay, and we enforced the maximum norm of a column of hidden units to 5. The training stopped after 1000 iterations and the iteration with best validation error was chosen. We used the Pylearn2 package (Goodfellow et al., 2013) to train the networks.

This initial training was done on 5000 training points and 1250 validation points (both of which come from the MNIST dataset) for each machine. The training points were chosen at random to include 2500 images of a specific digit class (that is, 2500 zeros for the first machine, 2500 ones for the second machine and so on), and 2500 images of random other digits for each machine. The validation sets were composed similarly. Each machine then used Algorithm 2 to transform 1000 images of digits *from its training set* into maximally similar images of the opposing class.

We thus started off with ten manipulated datasets of 1000 images each. The first dataset contained images of zeros manipulated to be non-zeros, and all the other digits manipulated to be zeros. The tenth dataset contained images of nines manipulated to be non-nines and the other digits manipulated to be nines. We then used Amazon Mechanical Turk to present all those images to human annotators, using the interface shown in Fig. 4.3. The images created by all the manipulator networks were mixed at random together, so that each single annotator (annotating 250 images in one task) would see some images created by each machine. Finally, each of the 10000 images was shown to five annotators; we used $5 \times 40 = 200$ annotators total on each iteration. The annotators labeled the images as either one of the ten digits, or the question mark '?' if there was no recognizable digit in an image. The final label ("target digit" or "not target digit") was chosen using majority of the annotators' votes.

The annotated manipulated digits were then added to the datasets which their respective original images belonged to. We then proceeded to train the next iteration of neural network manipulators on the updated datasets, and so on until completion of the manipulator training.



Task: Type in the digits or '?' if you can't even guess the digit.

09270

52?99

??770

74207

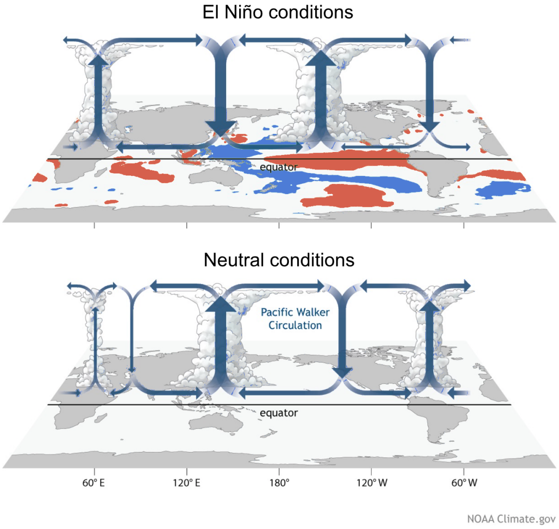
72

Figure 4.3: The Amazon Mechanical Turk interface we used to query online annotators. An annotator is shown five rows of five manipulated digit images, and is requested to type the digit labels (or '?') into the input boxes. Each annotator goes through ten similar screens, annotating a total of 250 digits.

Chapter 5

Application to Climate Science

The accurate characterization of macro-level climate phenomena is crucial to an understanding of climate dynamics, long term climate evolution and forecasting. Modern climate science models, despite their complexity, rely on an accurate and valid aggregation of micro-level measurements into macro-phenomena. While many aspects of climate may indeed be subject fundamentally to chaotic dynamics, many large scale phenomena are deemed amenable to precise modeling. The El Niño–Southern Oscillation (ENSO) is arguably the most studied climate phenomenon at the inter-annual time scale, but much about its dynamics relating zonal winds (zonal wind strength (ZW)) and sea surface temperatures (SST) remains poorly understood.



NOAA Climate.gov

Figure 5.1: El Niño vs. neutral conditions from Di Liberto (2014). Top: An illustration of the state of the atmosphere and surface during typical El Niño conditions. Here, the colors indicate SST deviations from the neutral state with red being a positive and blue being a negative deviation. Bottom: Similar to the top panel but now showing neutral conditions of the Walker circulation (neither El Niño nor La Niña).

From the climate-science point of view, our research shows that CFL can be successfully used for an unbiased automated extraction of climate macro-variables, which would otherwise require tedious hand-crafting by domain experts. Moreover, the framework can directly suggest (computationally) expensive climate ex-

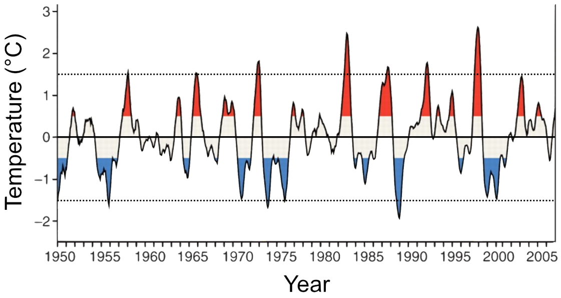


Figure 5.2: Niño 3.4 SST anomalies for the time period 1950–2005. The figure was adapted from McPhaden et al. (2006). Red shadings indicate El Niño years and blue shadings indicate La Niña years. The two dashed lines indicate the threshold for strong El Niño or La Niña events.

periments (for example, through climate simulations) that could differentiate between true causes and mere correlations efficiently. Closer inspection of the output of CFL can also yield insights about new climate macro-phenomena (or important variants of existing ones) that inspire new physical models of the climate. Python code that reproduces our results and figures is available online at `http://vision.caltech.edu/~kchalupk/code.html`.

5.1 Advances in This Chapter

This chapter uses CFL to learn causal macro-variables from equatorial Pacific climate data. It shows that CFL can:

- be applied to real-world data, and
- learn, without supervision, the causal hypothesis that El Niño is an important macro-variable state in the ZW-SST system's dynamics.

5.2 El Niño–Southern Oscillation

El Niño is a weather pattern that is principally characterized by the state of eastern Pacific near-surface winds, sea surface temperature patterns, and the associated state of the atmospheric Walker circulation (see for example, Holton et al., 1989; Trenberth, 1997). The Walker circulation (see Fig. 5.1) is characterized by warm air rising over Indonesia and Papua New Guinea and cooler subsiding air over the eastern Pacific cold tongue region just west of equatorial South America (Lau and Yang, 2003). Near the surface, easterly winds (winds blowing from the east) drive water from east to west resulting in oceanic upwelling near the coast of equatorial South America (and downwelling east of Indonesia), that brings with it cold and nutrient rich waters from the deep oceans. During the ENSO warm phase, commonly referred to as El Niño (because it often occurs around and after Christmas), the Walker circulation weakens, ultimately resulting in weaker upwelling in the Eastern Pacific and thus in positive SST anomalies. Fig. 5.1 illustrates these phenomena.

ENSO-related weather in the tropics includes droughts, flooding, and may have direct impact on fisheries through reduced nutrient upwelling (e.g., Glantz, 2001). Atmospheric waves (ripples in wind, SST and rainfall patterns) generated by the change in circulation and SST anomalies in the tropics, make their way across the planet with dramatic impact (e.g, Ropelewski and Halpert, 1987; Changnon, 1999). Cashin et al. (2015) show that the economic impact of El Niño varies across regions. Economic activity may decline briefly in Australia, Chile, Indonesia, India, Japan, New Zealand, and South Africa after an El Niño event. Enhanced growth may be registered in other countries, such as the United States.

The ENSO cold phase, usually referred to as La Niña, is the opposing phase of El Niño with enhanced upwelling and colder SSTs in the eastern Pacific. Currently, predicting the strength of El Niño and La Niña events remains a difficult challenge for climate scientists as the period may vary between 3 and 7 years (see Fig. 5.2); as a consequence accurate forecasts are only possible less than a year in advance (e.g., Landsea and Knaff, 2000).

The National Oceanic and Atmospheric Administration (NOAA) defines El Niño as a positive three-month running mean SST anomaly of more than $0.5°$C from normal (for the 1971–2000 base period) in the Niño 3.4 region ($120°$W–$170°$W, $5°$N–$5°$S, see also Fig. 5.3). Similarly, La Niña conditions are defined as negative anomalies of more than $-0.5°$ C. Conditions in between $-0.5°$C and $0.5°$C are called neutral. This is illustrated using red and blue shadings in Fig. 5.2. Strong El Niño/La Niña events are defined as SST-anomalies greater than $1.5°$C. However, the definitions for El Niño and La Niña have evolved over time. For example, other regions than the Niño 3.4 region or other averaging conventions have been used in the specification of the SST anomalies.

5.3 Experiment: Learning Pacific Macro-variables

Climate experts view zonal winds as drivers of SST patterns. We take the view that if El Niño and La Niña are indeed genuine macro-level climate phenomena in their own right (and not just arbitrary quantities defined by convention) then they must consist of macro-level features of the relation between the high-dimensional micro-level ZW and SST patterns that can be detected by an unsupervised method. That is, it must be possible to identify El Niño and La Niña from a mass of air pressure and sea temperature readings, using a method that has no independent information about when such periods occurred.

Chapters 2 and 3 develop a theoretically precise account of causal relations of macro-variables that supervene on micro-variables and proposed an unsupervised method for their discovery. This chapter adopts the framework with a few interpretational adjustments for our climate setting.

The input micro-variable X is, in this case, the ZW map. The output micro-variable Y takes values in the high-dimensional domain $\mathcal{Y}$ (SST patterns). The basic idea underlying our set-up is that the causal macro-variable relation is defined in terms of the *coarsest* aggregation of the micro-level spaces that preserves the probabilistic relations under intervention (hence, causal) between the micro-level spaces. Conceptually,

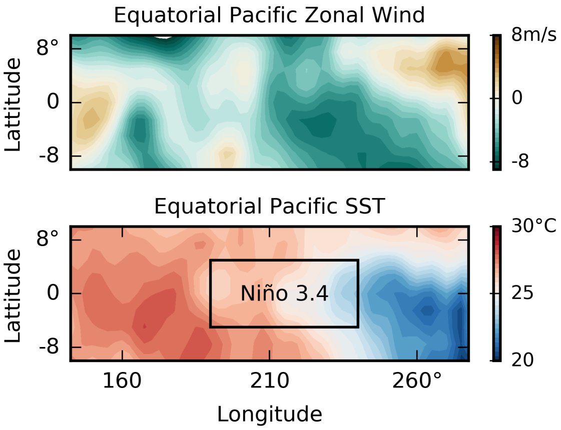


Figure 5.3: A micro-variable climate dataset. Top: A week's average ZW field. Bottom: A week's average SST field over the same region. In addition, the Niño 3.4 region is marked. Our dataset comprises 36 years' worth of overlapping weekly averages over the presented region.

macro-level causal variables group together micro-level states that make no causal difference.

In the present context, our climate data consisting of ZW and SST measurements (Sec. 5.3.1 below describes the dataset in detail) is entirely observational. That is, the data is naturally sampled from $P(\text{SST},\ \text{ZW})$ and not created by a (hypothetical) experimentalist from $P(\text{SST} \mid \text{man}(\text{ZW} = z))$ for different values of z. Nevertheless, we can identify the *observational* macro-variables that characterize the probabilistic relation between ZW and SST.

In Chapter 3 we showed that the fundamental *causal* partition is almost always a *coarsening* of the corresponding fundamental *observational* partition. We thus have some reason to expect that any macro-variables we do identify from our observational climate data will capture all the distinctions that are causal, but may in addition make some distinctions that do not support a causal inference. We return to this point in Section 5.3.5, where we discuss in more detail what causal insights can be drawn from this work. Our results should be seen as a step towards a characterization of macro-level causal variables for climate science, but we fully acknowledge that a complete causal characterization of the equatorial Pacific climate dynamics is beyond the scope of this book.

5.3.1 Dataset

The data used for this study is based on the daily-averaged version of the NCEP-DOE Reanalysis 2 product for the time period 1979–2014 inclusive (Kanamitsu et al., 2002), a data product provided by the US National Centers for Environmental Protection (NCEP) and the Department of Energy (DOE). Reanalysis data sets are generated by fitting a complex climate model to all available data for a given period of time, thus generating estimates for times and locations that were not originally observed. In addition, we used the Geophysical Observational Analysis Tool (http://www.goat-geo.org) to interpolate the SST and zonal wind fields onto a

2.5° × 2.5° spatial grid for easier analysis. We chose to focus on the (140°, 280°)E×(-10°, +10°)N equatorial band of the Pacific Ocean. From the raw dataset, we extracted the zonal (west-to-east) wind component and SST data in this region (specifically, we extracted the fields at the 1000 hPa level near the surface). Finally, we smoothed the data by computing a running weekly average in each domain. The resulting dataset contains 13140 zonal wind and 13140 corresponding SST maps, each a 9×55 matrix. Fig. 5.3 shows sample data points.

5.3.2 Pacific Macro-Variables

To apply CFL in practice, we applied the algorithms of Chapter 3 to our dataset[1]. The algorithms extracted in an unsupervised manner the SST and ZW macrovariables. We start with the description of the results.

We will refer to zonal wind *macro-variables* as W, and to temperature *macro-variables* as T. We first chose to search for four-state macro-variables (though we experiment with varying this number in Sec. 5.3.3) and considered a zero-time delay[2] between W and T. In the CFL framework, each macro-variable state corresponds to a cell of a partition of the respective micro-variable input space. Fig. 5.4 visualizes the W and T we learned by plotting the difference between each macro-variable cell's mean and the ZW (SST) mean across the whole dataset. The visualized states are easy to describe: For example, when W=WEqt there is a larger-than-average westerly wind component in the west-equatorial region, a feature often associated with the causes of El Niño (see Fig. 5.1). Indeed, Table 5.1 shows that the El Niño cell of T only arises in connection with W=WEqt. In addition, WEqt is often positively correlated with the T=Warm. Throughout the rest of the article, we will mostly focus on the T macro-variable. Our first goal is to quantitatively justify calling T=1 “El Niño” and calling T=2 “La Niña”. Qualitatively, the warm and cold water tongues that reach westward across the Pacific and that are often used to describe the two phenomena, are evident in the image.

Following the standard definition of El Niño (see Section 5.2), we use the SST anomaly in the Niño 3.4 region to detect its presence (Trenberth, 1997). The anomaly is computed with respect to the climatological mean, that is the mean temperature *during the same week of the year* over all the weeks in our dataset. We will call a weekly average anomaly exceeding +.5°C a mild episode, and an anomaly exceeding +1.5°C a strong episode. The definition of La Niña is analogous, with negative thresholds. Fig. 5.5 shows that in the T=1 and T=2 cells, over 75% of all the points exceed the threshold for a mild (positive and negative, respectively) anomaly, and over 50% of the points exceed the strong threshold. The situation is different in the Warm and Cold cells, where almost no points exceed the strong threshold while the number of points falling in these non-anomalous cells is about 30% of the total. Since this macro-variable contains a state capturing a high

[1] We actually used a slightly modified version of the algorithm simply because we hadn't found the best machine learning ingredients while doing this experiment. The original work published in the proceedings of the conference on Uncertainty in Artificial Intelligence contains the details.

[2] A zero time delay implies that CFL will attempt to relate the weekly moving ZW average to the weekly moving SST average. The question of different time delays turns out to be a very subtle issue in the study of El Niño as El Niño is not a periodic event, nor does it have a fixed duration (see Fig. 5.2). We chose not to discuss other delays here and the zero-time delay was deemed a reasonable starting point by domain experts we consulted.

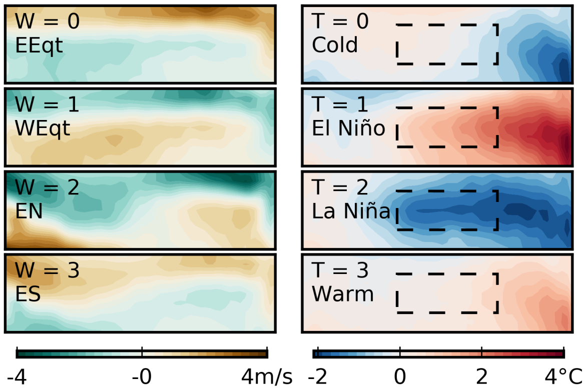


Figure 5.4: Macro-variables discovered by our algorithms. For each state, the average difference from the dataset mean is shown. Left: Four states of W, the zonal wind macro-variable. We named the states "Easterly Equatorial" (EEqt),"Westerly Equatorial" (WEqt), "Easterly North of Equator" (EN) and "Easterly South of Equator" (ES). Right: Four states of T, the SST macro-variable. We named the states "Cold [American Coastal Waters]", "El Niño", "La Niña" and "Warm [American Coastal Waters]". The main text provides additional justification for calling T=1 and T=2 "El Niño" and "La Niña", respectively.

proportion of El Niño-like patterns, we will say that this state has a "high precision" of detecting El Niño, while similarly, state T=2 has a high La Niña precision. Formally, we define the precision of a macro-variable state as follows:

Definition 22 (Precision)**.** *Let $T = \{T_1, \cdots, T_K\}$ be a partition of the set of all the SST maps used in our experiments. Let $n34 : SST \to \mathbb{R}$ be the function that computes the Niño 3.4 anomaly for a given map. Then, let*

$$c_\theta(T_k) = \begin{cases} \frac{1}{|T_k|}|\{t \in T_k \ s.t. \ n34(t) > \theta\}| & \textit{if } \theta > 0 \\ \\ \frac{1}{|T_k|}|\{t \in T_k \ s.t. \ n34(t) < \theta\}| & \textit{if } \theta < 0 \end{cases}$$

be the function that computes for, a given cell T_k of the partition, the fraction of its members whose anomaly is greater than (if $\theta > 0$) or lesser than (if $\theta < 0$) a given threshold θ. Finally, call the four numbers $\max_k c_{.5}(T_k)$, $\max_k c_{1.5}(T_k)$, $\max_k c_{(-.5)}(T_k)$, $\max_k c_{(-1.5)}(T_k)$ the mild/strong-El Niño and mild/strong-La Niña precision of the macro-variable T.

Together, the precisions indicate how well the partition T separates the mild and strong El Niño and La Niña anomalies from other structures in the data. In Fig. 5.5, for example, $c_{.5}(T) \approx .75$ and $c_{1.5}(T) \approx .25$ (both because of T=1), $c_{(-.5)}(T) \approx .85$ and $c_{(-1.5)}(T) \approx .5$ (both because of T=2). Thus, T has high mild-El Niño precision, and high mild-La Niña precision.

As further evidence that we recovered El Niño and La Niña, we show minimal state-to-state manipulations in Fig. 5.5. Take the La Niña→El Niño plot as an example. To compute it, we took all the SST maps for which T=La Niña, and for each found *the closest* (in the Euclidean space) map for which T=El Niño. We then averaged these differences. One of the insights the figure offers is that low SSTs in the Niño 3.4 region

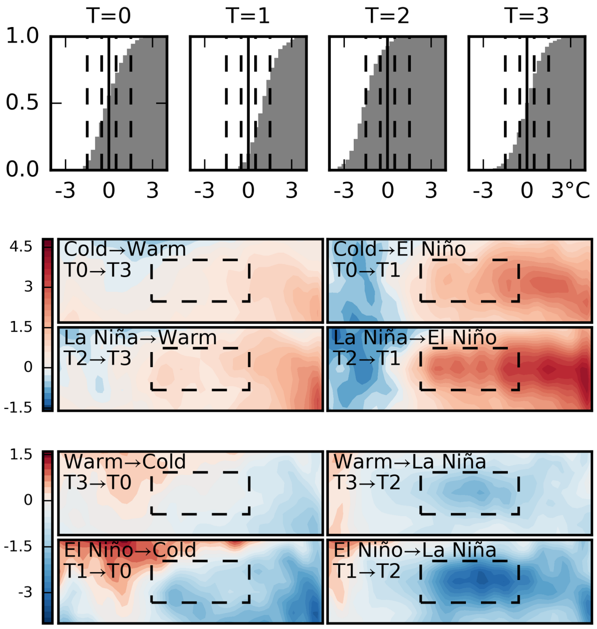


Figure 5.5: T=1 and T=2 are El Niño and La Niña. Top: Each plot shows the cumulative histogram of the Niño 3.4 anomalies, computed over all the weekly SST averages that belong to the given state of T. The dashed lines show the +/-0.5 and +/-1.5 "mild" and "strong" anomaly thresholds. Bottom: The minimal manipulations needed to transition from a given T-state into another (the exact procedure to obtain the plots is described in the text).

really are the distinguishing feature of T=La Niña. Similarly, an important difference between the T=Warm and T=El Niño is the characteristic tongue of warm water extending into the Niño 3.4 region. Adding this tongue is necessary to switch from T=Cold to T=El Niño, but not to switch from T=Cold or T=La Niña to T=Warm.

The CFL framework allows us to interpret W and T as standard probabilistic random variables with distribution we can estimate. Table 5.1 offers a probabilistic description of the system we learned. "When the equatorial zonal wind is unusually westerly, there is a 75% chance that the eastern Pacific is warm, and a 25% chance that El Niño arises." and "When the North-equatorial zonal wind is predominantly westerly, but the South-equatorial easterly, then the Eastern Pacific is most likely to be cold."—are example insights about the equatorial Pacific wind-SST system offered by CFL. We emphasize that both the macro-variables and the probabilities are learned from the data in an entirely unsupervised manner, without any a priori input about what constitutes ENSO events (except the fact that we restrict the SST and ZW fields to the equatorial Pacific region).

5.3.3 Varying the Number of States

Our choice of discovering four-state macro-variables was rather arbitrary. To check how varying the number of states changes the macro-variable precision (Def. 22), we repeated our experimental procedure, varying

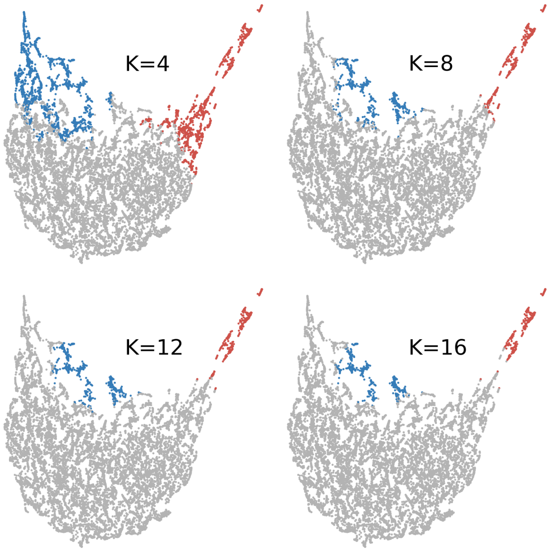


Figure 5.6: t-SNE (Van der Maaten and Hinton, 2008) embedding of the k-nn representation of SST data. The blue dots show, for varying K, the state of T with largest $c_{(-.5)}$ precision (see Def. 22). The red dots show the state with largest $c_{.5}$. Thus, the blue dots are "the" La Niña cluster for each K, and the red dots "the" El Niño cluster.

the number of states K from 2 to 16 (both in the ZW and SST space). Fig. 7.1 (relegated to the last chapter, as it provides basis for general discussion on our methods) shows the precisions for each case. As expected, a low number of states (K=2, 3) doesn't allow the algorithm to precisely detect El Niño and La Niña. With K $>$ 4 however, a slowly growing trend persists at high precision values. El Niño and La Niña remain important features as K changes.

There are several possible behaviors of the algorithm given the slowly growing precision of the macro-variables with growing K: (1) The El Niño and La Niña states remain roughly constant, (2) CFL sub-divides the El Niño and La Niña states, (3) CFL finds better El Niño and La Niña regions, (3) A mix of the above. Fig. 5.6 suggests that (2) is true. As K grows, the clusters that most precisely detect the mild El Niño and mild La Niña phenomena form a chain of strict subsets.

	Cold	El Niño	La Niña	Warm
EEqt	2/3	0	1/3	0
WEqt	0	1/4	0	3/4
EN	∼1/10	0	1/4	∼2/3
ES	3/4	0	0	1/4

Table 5.1: Each row shows $P(T \mid W = w)$ for a given w.

	T1	T2	T3	T4
W1	.075	.40	.25	.27
W2	.083	.39	.25	.27
W3	.084	.39	.26	.27
W4	.080	.40	.24	.27

Table 5.2: Conditional probabilities $P(T \mid W)$ when CFL is applied to randomly (in time) reshuffled ZW and SST data.

5.3.4 Reshuffled Data

As a sanity check, we ran our algorithms on randomly reshuffled (across the time dimension) ZW and SST data. We asked the algorithm to find K=4, $\ldots$, 16-state ZW and SST macro-variables. Table 5.2 shows $P(T \mid W)$, where W and T are the input and output macro-variables discovered in the randomized dataset with $K = 4$. Note that $P(T \mid W = W1)$, $P(T \mid W = W2)$, $P(T \mid W = W3)$ and $P(T \mid W = W4)$ are all equal. This is exactly as expected, since by reshuffling the data we removed any probabilistic dependence between the inputs and the outputs.

Applying Def. 11 to this data indicates that the algorithm implicitly only discovered one true ZW state, even though we explicitly asked it to look for a four-state macro-variable. The cardinality of the output macro-variable is three or four states, depending on whether .24–.26 is close enough to .27 to apply Def. 11 to merge the last two columns. We performed the same reshuffled analysis for each K and computed as before the precision for the weak and strong El Niño and the weak and strong La Niña. Fig. 7.1, large dotted lines, shows that in each case none of the clusters contains a significant proportion of either El Niño or La Niña patterns. This experiment shows that CFL passes the sanity check. When the inputs and outputs are independent, the input macro-variable is trivial, it has a single state.

5.3.5 Challenges to Establishing Causality

The CFL framework aspires to solve an important problem in causal reasoning: how to automatically form macro-level variables from micro-level observations. In this work we have shown, for the first time, that these algorithms can be successfully applied to real-life data. We have recovered well-known, complex climate phenomena (El Niño, La Niña) as macro-variable states directly from climate data, in an entirely unsupervised manner.

We emphasize that our experiments use *observational* climate data, and we have to be cautious about causal conclusions. It is not even clear *a priori* whether the $ZW \rightarrow SST$ causal direction is a reasonable choice: it is known that wind patterns cause changes in SST and it in turn affects the wind by changing the atmospheric pressure. Feedback loops are commonplace in climate dynamics.

The Causal Coarsening Theorem (Theorem 16) provides the basis for an efficient learning of causal relationships based on observational macro-variables – but some experiments are required. In addition, the theorems were only shown to hold for variables that are not subject to feedback. However, we are hopeful that an extension accounting for feedback can be proven. While real climate experiments are generally not feasible, such a theorem would provide the basis to perform large-scale climate experiments with detailed climate models, for example, to check whether *interventionally* shifting from the $W = 0$ zonal wind state to $W = 1$ in the climate model increases the likelihood of El Niño (i.e. of SST ending up in state T=1). Connecting the CFL framework with such experiments is an exciting future direction as it would also enable the possibility of using the macro-variables we have found to inform policy that aims to influence climate phenomena.

Even when working with purely observational data, CFL offers an important causal insight not revealed by clustering methods. It guards against learning variables with ambiguous manipulation effects (Spirtes and Scheines, 2004). An illustrative example of an ambiguous macro-variable is total cholesterol. Low density lipids (LDL, commonly called "bad cholesterol") and high density lipids (HDL, "good cholesterol") can be aggregated together to count total cholesterol (TC), but TC has an ambiguous effect on heart disease because effects of LDL and HDL differ. The Causal Coarsening Theorem guarantees that each state of the observational macro-variable is causally unambiguous: no mixing of HDL and LDL can occur. In case of our El Niño setup, this means that two ZW states within the same cell are guaranteed to have the same effect on the SST macro-variable.

Finally, we note that there still is significant debate among climate scientists about what exactly constitutes El Niño and what its causes are. For example, recent research has shown that there may be multiple different types of El Niño states (Kao and Yu, 2009; Johnson, 2013) that all fall under NOAA's definition. Our results suggest that the current definition described in Section 5.2 coincides well with states of the probabilistic macro-variable discovered by CFL. In addition, Sec. 5.3.3 indicates that finer-grained structure does exist within the El Niño and La Niña clusters when they are analyzed from the relational-probabilistic standpoint. We leave this line of research as an important future direction.

Chapter 6

Causation without Intervention

Take two discrete variables X and Y that are probabilistically dependent. Assume there is no feedback between the variables: it is not the case that both X causes Y and Y causes X. Further assume (Reichenbach, 1991) that all probabilistic dependence always arises due to causation. The fundamental causal question is then to assess three hypotheses: 1) Does X cause Y, 2) Does Y cause X, and 3) Do X and Y have a common cause H? Since we assumed no feedback in the system, hypotheses 1) and 2) are mutually exclusive. Each of them, however, can occur together with hypothesis 3). Fig. 6.1 enumerates the possibilities.

Within the causal graphical models framework (Pearl, 2000; Spirtes et al., 2000), differentiating between any two of the causally interesting possibilities (shown in Fig. 6.1B-F) is in general only possible if one has the ability to intervene on the system. For example, to differentiate between the pure-confounding and the direct-causal case (Fig. 6.1B and C), one can intervene on X and observe whether that has an effect on the distribution of Y. Given only observations of X and Y and no ability to intervene on the system however, the problem is in general not identifiable. Roughly speaking, the reason is simply that any joint $P(X, Y)$ can be factorized as $P(X)P(Y \mid X)$ and $P(Y)P(X \mid Y)$, and the hidden confounder H can easily be endowed with a distribution that can give the marginal $\sum_H P(X, Y, H)$ any desired form.

6.1 Advances in This Chapter

In this chapter we design a novel method to establish the likelihood of each possible causal graph given samples of two discrete variables X, Y. The method is entirely observational, based only on looking at the joint probability $P(X, Y)$ – without resorting to intervention.

Previous chapters showed that, given an observational partition, *one of its coarsenings* is the causal partition. Thus, in principle, it should be possible to iterate through all the coarsenings and pick out the "right" or "most-causal" one using methods inspired by this section. Whereas the specifics of this algorithm are still unclear, in this chapter outline a possible solution.

6.2 Related Work

There are two common remedies to the fundamental unidentifiability of the two-variable causal system: 1) Resort to interventions or 2) Introduce additional assumptions about the system and derive a solution that works under these assumptions.

Whereas the first solution is straightforward, research in the second direction is a more recent and exciting enterprise.

6.2.1 Additive Noise Models

A recent body of work attacks the problem of establishing whether $x \to y$ or $y \to x$ when specific assumptions with respect to the functional form of the causal relationship hold. Shimizu et al. (2006) showed that when the effect is a linear function of the cause, with *non-Gaussian* noise, then the casual direction can be identified in the limit of infinite sample size.

This inspired further work on the so called "additive noise models". Hoyer et al. (2009) extended Shimizu's idea to the case when the effect is any (except for a small enumerated set) nonlinear function of the cause, and the noise is additive – even Gaussian. Zhang and Hyvärinen (2009) showed that a *post*nonlinear model – the case where $y = f(g(x) + \epsilon)$ with f an invertible function and ϵ a noise term – is identifiable. The nonlinear noise models framework was applied to discrete variables by Peters et al. (2011). Janzing et al. (2009) showed that the additive noise assumption can be used to detect confounding with some success.

Unfortunately, the additive noise assumption is rather stringent. Some extensions of the additive noise framework (such as the post-nonlinear model) do not apply in the discrete case.

6.2.2 Bayesian Causal Model Selection

In a classic work on Bayesian Network learning (then called Belief Net learning), Heckermann and Chickering develop a Bayesian scoring criterion that allows them to assess the likelihood of each possible Bayesian network given a dataset. This work introduces five assumptions that together define which networks are more and less likely. Their Assumptions 1 (Multinomial Probabilities), 2 (Parameter Independence) and 5 (Multinomial Hyperpriors) can be used to define the likelihood of the structures shown here in Fig. 6.1. We do not repeat the assumptions here, as we propose their modified versions in Sec. 6.3.

The crucial difference between the work of Heckermann and ours is that their goal is to find *Markov Equivalence Classes* of Bayesian Networks. That is, to them two networks that encode the same independence assumptions are equivalent. This, however, renders our task impossible: all the possibilities enumerated in Fig. 6.1B-F are Markov-equivalent. Our contribution is thus to use assumptions similar and assess the likelihood of fundamentally unidentifiable causal structures over two variables, bearing in mind that there is no "right" structure for any observed joint, but there are "more likely" and "less likely" structures.

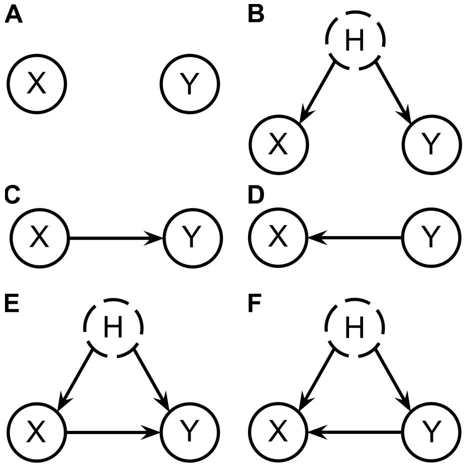


Figure 6.1: Possible causal structures linking X and Y. Assume X, Y and H are all discrete but H is unobserved. In principle, it is impossible to identify the correct causal structure given only X and Y samples. In this report, we will tackle this problem using a minimalistic set of assumptions. Our final result is a classifier that differentiates between these six cases – the confusion matrices are shown in Fig. 6.9.

Our idea is most similar in spirit to the work of Sun et al. (2006). Sun puts an explicit Bayesian prior on what a likely causal system is: if X causes Y, then the conditionals $p(Y \mid X = x)$ are *less complex* than the reverse conditionals $P(X \mid Y = y)$, where complexity is measured by the Hilbert space norm of the conditional density functions. This formulation is plausible and easily applicable to discrete systems (by defining the complexity of discrete probability tables by their entropy).

6.2.3 Desiderata

Our contribution is to create the first algorithm with the following properties:

1. Works for *discrete* variables X and Y.
2. Decides between all the six possible graphs shown in Fig. 6.1.
3. Does not make any functional assumptions about the functional form of the discrete parametrization (e.g. additive noise).

In a recent review, Mooij et al. (2014) compares a range of methods that decide the causal direction between two variables, including the methods discussed above. To our knowledge, none of these methods attempt to distinguish between the pure-causal, the confounded, and the causal+confounded case.

6.3 Assumptions

We take an approach inspired by the Bayesian methods discussed in Sec. 6.2. Consider the Bayesian model in which $P(X, Y)$ is sampled from a hyperprior. Our method is to make this hyperprior as weak or uninformative as possible while retaining the property that distribution of the cause is independent of the distribution of the effect conditioned on the cause:

1. Assume that $P(effect \mid cause) \perp\!\!\!\perp P(cause)$.
2. Assume that $P(effect \mid cause = c)$ is sampled from the uninformative hyperprior for each c.
3. Assume that $P(cause)$ is sampled from the uninformative hyperprior.

Since all the distributions under considerations are multinomial, the "uninformative hyperprior" is the Dirichlet distribution with parameters all equal to 1 (which we will denote as $Dir(1)$, remembering that 1 is actually a vector whose dimensionality will be clear from context). What $cause$ and $effect$ are depends on which causal system is sampled. For example, if $X \to Y$ and there is also confounding $X \leftarrow h \to Y$ (Fig. 6.1D), then our assumptions set

$$\begin{aligned} P(X) &\sim Dir(1) \\ \forall_x P(Y \mid X = x) &\sim Dir(1) \\ P(H) &\sim Dir(1) \\ \forall_h P(X \mid H = h) &\sim Dir(1) \\ \forall_h P(Y \mid H = h) &\sim Dir(1) \end{aligned}$$

6.4 An Analytical Solution: Causal Direction

Consider first the problem of identifying the causal direction. That is, assume that either $X \to Y$ or $Y \to X$, and there is no confounding. The assumptions of Sec. 6.3 then allow us to compute, for any given joint $P(X, Y)$ (which we will from now on denote P_{XY} to simplify notation), the likelihood $p(X \to Y \mid P_{XY})$ and the likelihood $p(Y \to X \mid P_{XY})$. The likelihood ratio allows us to decide which causal direction P_{XY} more likely represents.

We first derive and visualize the likelihood for the case of X and Y both binary variables. Next, we generalize the result to general X and Y. Finally, we analyze experimentally how sensitive such causal direction classifier is to breaking the assumption of uninformative Dirichlet hyperpriors (but keeping the independent mechanisms assumption).

6.4.1 Optimal Classifier for Binary X and Y

Consider first the binary case. Let $P_X = \begin{bmatrix} a \\ 1-a \end{bmatrix}$ and $P_{Y|X} = \begin{bmatrix} b & 1-b \\ c & 1-c \end{bmatrix}$. Assume P_X is sampled independently from $P_{Y|X}$, and that the densities (parameterized by a and b, c) are $\mathcal{F}_a, \mathcal{F}_b, \mathcal{F}_c \colon (0,1) \to \mathbb{R}$. This defines a density over (a, b, c), the three-dimensional parameterization of an $x \to y$ system, as $\mathcal{F}(a, b, c) = \mathcal{F}_a(a)\mathcal{F}_b(b)\mathcal{F}_c(c) \colon (0,1)^3 \to \mathbb{R}$.

Now, consider $P_{XY} = \begin{bmatrix} d & e \\ f & 1-(d+e+f) \end{bmatrix}$ – a three-dimensional parameterization of the joint. If we assume that P_{XY} is sampled according to the $X \to Y$ sampling procedure, we can compute its density $\mathcal{H}_{XY}\colon (0,1)^3 \to \mathbb{R}$ as a function of $\mathcal{F}$ using the multivariate change of variables formula. We have

$$\begin{bmatrix} d \\ e \\ f \end{bmatrix} = \begin{bmatrix} ab \\ a(1-b) \\ (1-a)c \end{bmatrix}$$

and the inverse transformation is

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} d+e \\ \frac{d}{d+e} \\ \frac{f}{1-d-e} \end{bmatrix} \tag{6.1}$$

The Jacobian of the inverse transformation is

$$\frac{d(a,b,c)}{d(d,e,f)} = \begin{bmatrix} 1 & 1 & 0 \\ \frac{e}{(d+e)^2} & \frac{-d}{(d+e)^2} & 0 \\ \frac{f}{(1-d-e)^2} & \frac{f}{(1-d-e)^2} & \frac{1}{1-d-e}, \end{bmatrix}$$

its determinant $\det\left(\frac{d(a,b,c)}{d(d,e,f)}\right) = \frac{-1}{(d+e)-(d+e)^2}$. The change of variables formula then gives us

$$\mathcal{H}_{XY}(d,e,f) = \frac{\mathcal{F}(d+e, \frac{d}{d+e}, \frac{f}{1-d-e})}{(d+e)-(d+e)^2},$$

where a, b, c are obtained from Eq. (6.1).

We can repeat the same reasoning for the inverse causal direction, $Y \to X$. In this case, we obtain

$$\mathcal{H}_{YX}(d,e,f) = \frac{\mathcal{F}(d+f, \frac{d}{d+f}, \frac{e}{1-d-f})}{(d+f)-(d+f)^2}.$$

Given P_{XY} and the hyperpriors $\mathcal{F}$, we can now test which causal direction P_{XY} most likely corresponds to. Assuming equal priors on both causal directions, we have

$$\begin{aligned} \frac{p(X \to Y \mid (d,e,f))}{p(Y \to X \mid (d,e,f))} &= \frac{\mathcal{H}_{xy}(d,e,f)}{\mathcal{H}_{yx}(d,e,f)} \\ &= \frac{\mathcal{F}\left(d+e, \frac{d}{d+e}, \frac{f}{1-d-e}\right)}{\mathcal{F}\left(d+f, \frac{d}{d+f}, \frac{e}{1-d-f}\right)} \frac{(d+f)-(d+f)^2}{(d+e)-(d+e)^2} \end{aligned}$$

Only the first factor in the likelihood ratio depends on the hyperprior $\mathcal{F}$. If we fix $\mathcal{F}_a, \mathcal{F}_b, \mathcal{F}_c$ to all be $Dir(1)$, the factor reduces to 1 and the likelihood ratio becomes

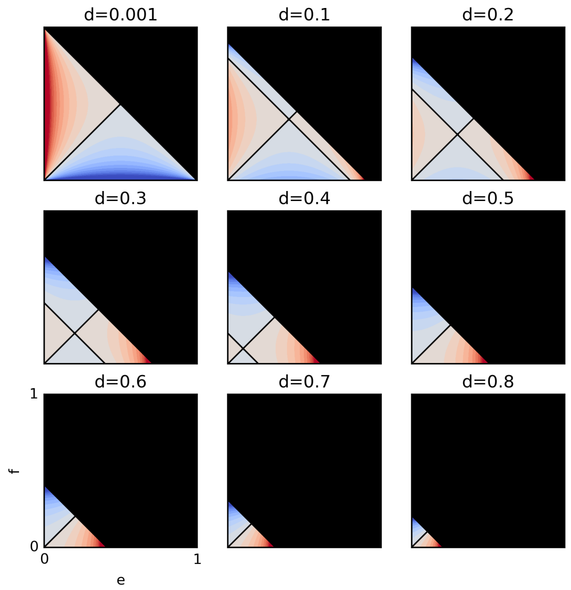


Figure 6.2: Log likelihood-ratio $\log\left(\frac{P(X\to Y|(d,e,f))}{P(Y\to X|(d,e,f))}\right)$ as a function of e, f for nine different values of d. Red corresponds to values larger than 0 — that is, $X \to Y$ is more likely than the opposite causal direction in the red regions. Blue signifies the opposite. The decision boundary is shown in black. It is a union of two orthogonal planes that cut the (d, e, f) simplex into four connected components along a skewed axis.

$$\frac{p(X \to Y \mid (d,e,f))}{p(Y \to X \mid (d,e,f))} = \frac{(d+f)-(d+f)^2}{(d+e)-(d+e)^2}.$$

Denote the "uninformative-hyperprior likelihood ratio" function

$$LR\colon P_{XY}(d,e,f) \mapsto \frac{(d+f)-(d+f)^2}{(d+e)-(d+e)^2}.$$

The classifier that assigns the $X \to Y$ class to P_{XY} with $LR(P_{XY}) > 1$, and the $Y \to X$ class otherwise is the optimal classifier under our assumptions. Fig. 6.2 shows LR across the three-dimensional P_{XY} simplex. The figure shows nine slices of this simplex for different values of the d coordinate.

6.4.2 Optimal Classifier for Arbitrary X and Y

Deriving the optimal classifier for the case where X and Y are not binary is analogous to the binary derivation. The resulting likelihood ratio is

$$\frac{p(X \to Y \mid P_{XY})}{p(Y \to X \mid P_{XY})} = \tag{6.2}$$

$$= \frac{\mathcal{F}\left(P_X, P_{Y|X}\right)}{\mathcal{F}\left(P_Y, P_{X|Y}\right)} \frac{|\det J_{XY}|^{-1}}{|\det J_{YX}|^{-1}}, \tag{6.3}$$

where J_{XY} is the Jacobian of the linear transformation $(P_X, P_{Y|X}) \mapsto P_{XY}$ and J_{YX} is the Jacobian of the transformation $(P_Y, P_{X|Y}) \mapsto P_{XY}$. The transformation, its determinant and Jacobian are readily computable on paper or using computer algebra systems. In our implementation, we used Theano (Theano Development Team, 2016) to perform the computation for us. Note that if X has cardinality k_X and Y has cardinality k_Y, the Jacobians have $(k_X k_Y - 1)^2$ entries. Computing their determinants has complexity $\mathcal{O}((k_X k_Y - 1)^6)$ or, if we assume $k_X = k_Y = k$, $\mathcal{O}(k^{12})$ – it grows rather quickly with growing cardinality.

If $\mathcal{F}$ is flat, that is all the priors are $Dir(1)$, we will call the causal direction classifier that follows Eq. (6.3) the LR classifier. That is, the LR classifier outputs $X \to Y$ if the uninformative-hyperprior likelihood ratio is larger than 1, and outputs $Y \to X$ otherwise.

Note that the *optimal* classifier is not *perfect* – there is a baseline error that the optimal classifier has under the assumptions it is built on. This error is

$$\begin{aligned} E_{LR} = \int p(Y \to X \mid P_{XY}) I_{[LR(P_{XY})>1]} + \\ p(X \to Y \mid P_{XY}) I_{[LR(P_{XY})<1]} dP_{XY}, \end{aligned}$$

where the integral varies over all the possible joints P_{XY} with uniform measure, and $I_{[LR(P_{XY})<>1]}$ is the indicator function that evaluates to 1 if its subscript condition holds, and to 0 otherwise.

That is, assuming that each P_{XY} is sampled from the uninformative Dirichlet prior given that either $X \to Y$ or $Y \to X$ with given probability, in the limit of infinite classification trials the error rate of the LR classifier is E_{LR}. Whereas this integral is not analytically computable (at least neither by the authors nor by available computer algebra systems), we can estimate it using Monte Carlo methods in the following sections. In Fig. 6.6, the leftmost entry on each curve corresponds to E_{LR} for various cardinalities of X and Y. For example, for $|X| = |Y| = 2$, $E_{LR} \approx .4$ but already for $|X| = |Y| = 10$, $E_{LR} < .001$.

6.4.3 Robustness: Changing the Hyperprior $\mathcal{F}$

What if we use the LR classifier, but our assumptions do not match reality? Namely, what if $\mathcal{F}$ is *not* $Dir(1)$? For example, what if $\mathcal{F}$ is a mixture of ten Dirichlet distributions[1]?

We will draw $\mathcal{F}$ from mixtures with fixed "$|\log_2(\alpha_{max})|$". Let the k-th component of the mixture have parameter $\alpha^k = (\alpha_1^k, \cdots, \alpha_N^k)$ where N is the cardinality of X or Y. Then fixed α_{max} means that we drew each α_i^k uniformly at random from the interval $2^{-\alpha_{max}}, 2^{\alpha_{max}}$. Fig. 6.4 shows samples from such mixtures with growing α_{max}. The figure shows that increasing the parameter allows the distributions to grow in complexity.

Note that if $\alpha_{max} = 0$, we recover the noninformative prior case. How does the likelihood ratio and the causal direction decision boundary change as we allow α_{max} to depart from 0? For binary X and Y, Fig. 6.4 illustrates the change. Comparing with Fig. 6.2, we see that as α_{max} grows, the likelihood ratios become

[1] A mixture of Dirichlet distributions with arbitrary many components can approximate any distribution over the simplex.

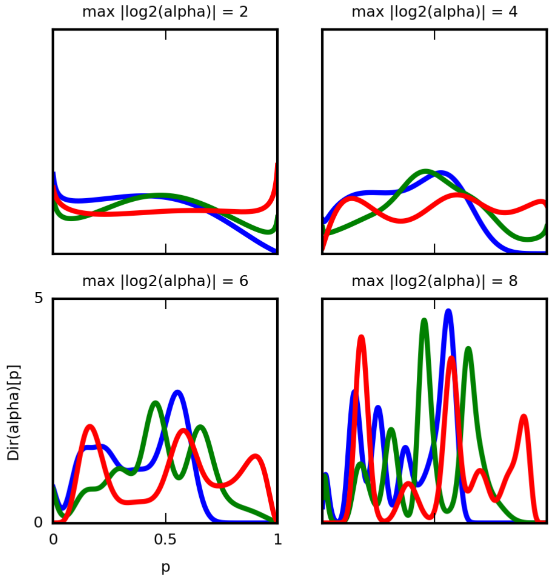


Figure 6.3: Samples from Dirichlet mixtures. Each plot shows three random samples from a ten-component mixture of Dirichlet distributions over the 1D simplex. Each mixture component has a different, random parameter α. For each plot we fixed a different $|log_2(\alpha_{max})|$, a parameter which limits both the smallest and largest value of any of the two α coordinates that define each mixture component.

more extreme, and the decision boundaries become more complex. Fig. 6.5 makes it clear that a fixed α_{max} allows for the decision boundary to vary significantly.

That the "independent mechanisms" assumption as we framed it is not sufficient to provide identifiability of the causal direction was clear from the outset (since each joint can be factorized as $P(X)P(Y \mid X)$ and $P(Y)P(X \mid Y)$). However, the above considerations suggest that the assumption of noninformative hyperpriors is rather strong: In fact, it is possible to show that the decision surface can be precisely flipped with appropriate adjustment of $\mathcal{F}$, making the LR classifier's error precisely 100%.

Our experiments, however, suggest that using the LR classifier is a reasonable choice in a wide range of circumstances, *especially as the cardinality of X and Y grows*. In our experiments, we checked how the error changes as we allow the α_{max} parameter of all the hyperpriors to grow. Our experimental procedure is as follows:

1. **Fix the dimensionality** of X and Y, and fix α_{max}.
2. **Sample 100 hyperpriors for each dimensionality and** α_{max}. Sample α parameters for $\mathcal{F}$ within given α_{max} bounds, where $\mathcal{F}$ consists of Dirichlet mixtures (with 10 components), as described above.
3. **Sample 100 priors for each hyperprior**. Sample $P(cause)$ and $P(effect \mid cause)$ 100 times for each hyperpriors (that is, for each α setting).
4. **Sample the causal label uniformly**. If chose $X \rightarrow Y$ then let $P_{XY} = P(cause)P(effect \mid cause)$. If chose $Y \rightarrow X$, let $P_{XY} = transpose[P(cause)P(effect \mid cause)]$.
5. **Classify**. Use the LR classifier to classify P_{XY}'s causal direction and record "error" if the causal label

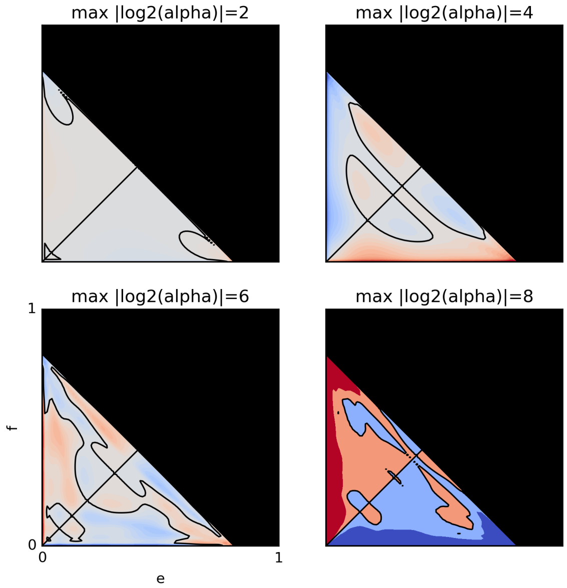


Figure 6.4: Log-likelihood ratios for the causal direction when $\mathcal{F}$ is a mixture of ten Dirichlet distributions with growing α_{max} (see Fig. 6.3).

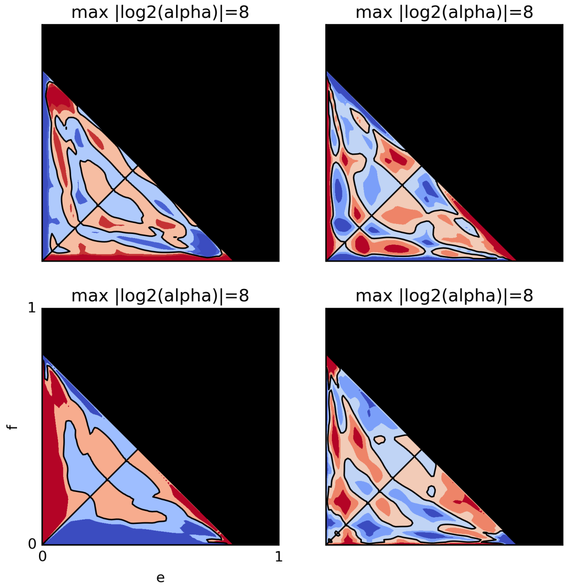


Figure 6.5: Log-likelihood ratios for the causal direction when $\mathcal{F}$ is a mixture of ten Dirichlet distributions with $|\alpha_{max}| = 2^8$ (see Fig. 6.3) – each plot corresponds to different, randomly sampled α.

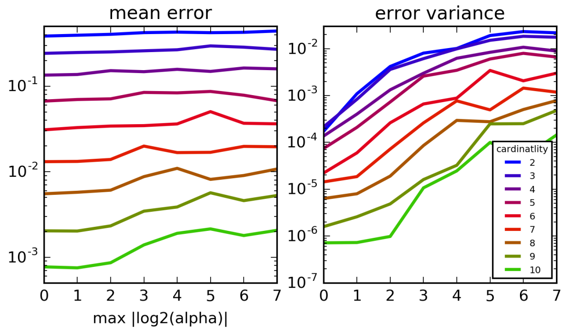


Figure 6.6: Results of the direction-classification experiment. We varied the cardinality of X, Y as well as α_{max} of the mixture of Dirichlets $\mathcal{F}$. For each setting, we sampled 100 P_{XY} distributions according to our causal model and recorded the classification error of the simple LR classifier. The results show that, as cardinality of X and Y grows, the LR classifier's accuracy increases.

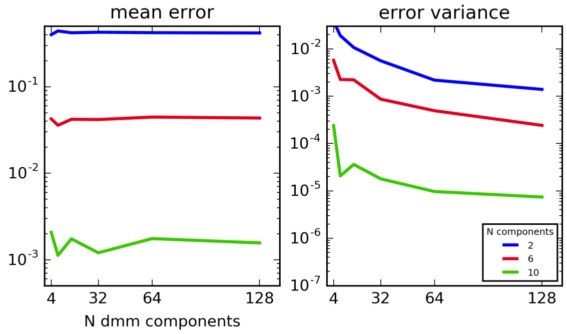


Figure 6.7: Results of the direction-classification experiment when the number of Dirichlet mixture model hyperprior components varies. We fixed α to vary between 2^{-7} and 2^{7}. The results show that the max-likelihood classifier that assumes the noninformative priors is not sensitive to the number of Dirichlet mixture components that the test data is sampled from.

disagrees with the classifier.

Figure 6.6 shows the results. As the cardinality of the system grows, the LR classifier's decision boundary approximates the decision boundary for most Dirichlet mixtures. Another trend is that as α_{max} grows, the variance of the error grows, but there is only a small growing trend in the error itself. In addition, Fig. 6.7 shows that the error does not increase as we allow more mixture components, up to 128 components, while holding α_{max} at the large value of 7. Thus, the LR classifier performs well even for extremely complex hyperpriors, at least on average.

6.5 A Black-box Solution: Detecting Confounding

Consider now the question of whether $X \to Y$ or $X \leftarrow H \to Y$, where H is a latent variable (a confounder). In this section we present a solution to this problem, under assumptions from Sec. 6.3.

Unfortunately, deriving the optimal classifier for this case is difficult without additional assumptions on

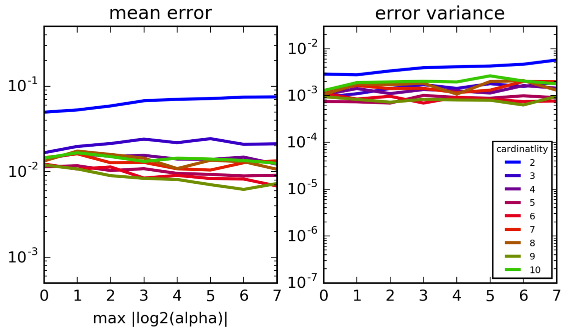


Figure 6.8: Results of the black-box confounding detector. We varied cardinality of X, Y as well as α_{max} of the mixture of Dirichlets $\mathcal{F}$. For each setting, we sampled 1000 P_{XY} distributions according to our causal model and recorded the classification error of a neural net classifier trained on noninformative Dirichlet hyperprior data. The results show that, as cardinality of X and Y grows, the LR classifier's accuracy increases.

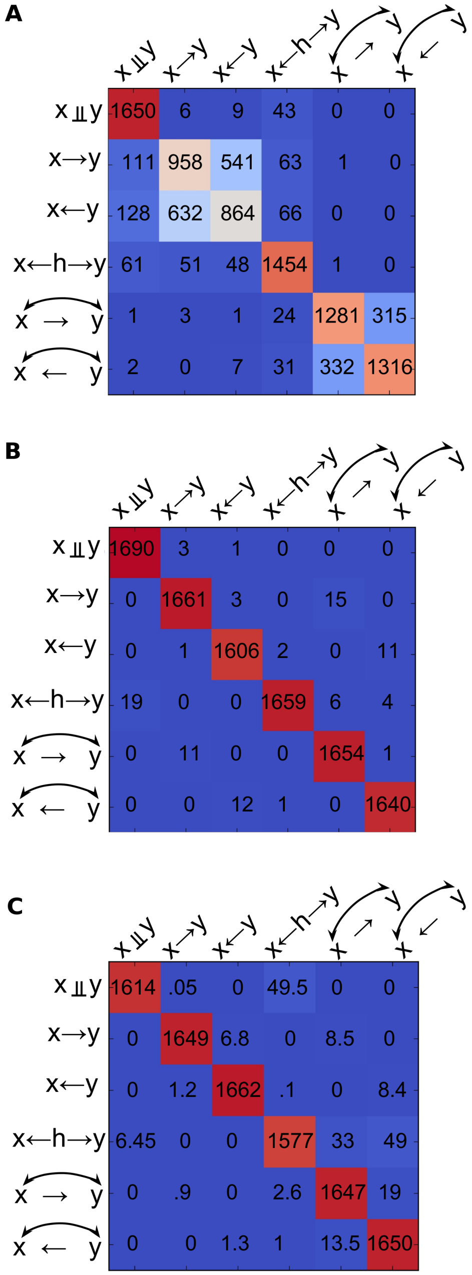
the latent H. Instead, we propose a black-box classifier. We created a dataset of distributions from both the direct-causal and confounded case, using the uninformative Dirichlet prior on either $P(X)$ and $P(Y \mid X)$ (the direct-causal case) or $P(H)$, $P(X \mid H)$ and $P(Y \mid H)$ in the confounded case. For each confounded distribution, we chose the cardinality of H, the hidden confounder, uniformly at random between 2 and 100. Next, we trained a neural network to classify the causal structure (Python code that reproduces the experiment is available at `vision.caltech.edu/~kchalupk/code.html`). We then checked how well this classifier performs as we vary the cardinality of the variables, and as we allow the true hyperprior to be a mixture of 10 Dirichlets, analogously to the experiment from Sec. 6.4.

Fig. 6.8 shows the results. Note that the classification errors are much lower than for the "deciding causal direction" case. Both problems (deciding causal direction and detecting confounding) are in principle unidentifiable, but it appears the latter is inherently easier. The neural net classifier seems to be little bothered by growing α_{max}. The largest source of error, for cardinality of X and Y larger than 3, seems to be neural network training rather than anything else.

6.6 A Black-Box Solution to the General Problem

Finally, we present a solution to the general causal discovery problem over the two variables X, Y: deciding between the six alternatives shown in Fig. 6.1. The idea is a natural extension of the black-box classifier from Sec. 6.5. We created a dataset containing all the six cases, sampled under the assumptions of Sec. 6.3. We then trained a neural network on this dataset (the neural network architecture, as well as the details of the training procedure, are available in the accompanying Python code).

Figure 6.9 shows the results of applying the classifier to distributions sampled from flat hyperpriors (that is, from a test set with statistics identical to the training set), for cardinalities $|X| = |Y| = 2$ and $|X| = |Y| = 10$. As expected, the number of errors is much lower for the higher cardinality. For the cardinality of



A

	$x \perp\!\!\!\perp y$	$x \to y$	$x \leftarrow y$	$x \leftarrow h \to y$	$\overset{\frown}{x \to y}$	$\overset{\frown}{x \leftarrow y}$
$x \perp\!\!\!\perp y$	1650	6	9	43	0	0
$x \to y$	111	958	541	63	1	0
$x \leftarrow y$	128	632	864	66	0	0
$x \leftarrow h \to y$	61	51	48	1454	1	0
$\overset{\frown}{x \to y}$	1	3	1	24	1281	315
$\overset{\frown}{x \leftarrow y}$	2	0	7	31	332	1316

B

	$x \perp\!\!\!\perp y$	$x \to y$	$x \leftarrow y$	$x \leftarrow h \to y$	$\overset{\frown}{x \to y}$	$\overset{\frown}{x \leftarrow y}$
$x \perp\!\!\!\perp y$	1690	3	1	0	0	0
$x \to y$	0	1661	3	0	15	0
$x \leftarrow y$	0	1	1606	2	0	11
$x \leftarrow h \to y$	19	0	0	1659	6	4
$\overset{\frown}{x \to y}$	0	11	0	0	1654	1
$\overset{\frown}{x \leftarrow y}$	0	0	12	1	0	1640

C

	$x \perp\!\!\!\perp y$	$x \to y$	$x \leftarrow y$	$x \leftarrow h \to y$	$\overset{\frown}{x \to y}$	$\overset{\frown}{x \leftarrow y}$
$x \perp\!\!\!\perp y$	1614	.05	0	49.5	0	0
$x \to y$	0	1649	6.8	0	8.5	0
$x \leftarrow y$	0	1.2	1662	.1	0	8.4
$x \leftarrow h \to y$	6.45	0	0	1577	33	49
$\overset{\frown}{x \to y}$	0	.9	0	2.6	1647	19
$\overset{\frown}{x \leftarrow y}$	0	0	1.3	1	13.5	1650

Figure 6.9: Confusion matrices for the all-causal-classes classification task. The test set consists of distributions sampled from uniform hyperpriors – that is, sampled from the same statistics as the training data (equivalent to $\alpha_{max} = 0$ in previous sections). A) Results for $|X| = |Y| = 2$. Total number of errors=2477. B) Results for $|X| = |Y| = 10$, total errors=85. C) Average results for $|X| = |Y| = 10$, same classifier as in B) but test set sampled with non-uniform hyperpriors with $\alpha_{max} = 7$ (see text). 201 errors on average. In each case, the test set contains 10000 distributions, with all the classes sampled with an equal chance.

2, the confusion matrix shows that the neural networks:

1. easily learn to classify independent vs dependent variables,
2. confuse the $X \to Y$ and $Y \to X$ cases, and
3. confuse the two "directed-causal plus confounding" cases (Fig. 6.1E,F).

However, all these are insignificant issues when $|X| = 10$, where the total error is 85 out of 10000 testpoints. For $|X| = 2$, the error is 25.7%. We remark again that the problem is not identifiable – that is, there is no "true causal class" for any point in our training or test dataset. Each distribution *could* arise from any of the possible five causal systems in which X and Y are not independent. The fact that the error nears 0 in the high-cardinality case indicates that the likelihoods under our assumptions grow very peaked as the cardinality grows. Thus, the *optimal* decision can quite safely be called the *true* decision. In addition, Fig. 6.9C shows the average confusion table for a hundred trials in which our classifier was applied to distributions over X and Y with cardinality 10, corresponding to all the possible six causal structures, but sampled from non-uniform hyperpriors with $\alpha_{max} = 7$. The performance drop is not drastic compared to Fig. 6.9B.

6.7 Discussion

We developed a neural network that determines the causal structure that links two discrete variables. We allow for confounding between the two variables, but assumed acyclicity. The classifier takes as input a joint probability table P_{XY} between the two variables and outputs the *most likely* causal graph that corresponds to this joint. The possible causal graphs span the range shown in Fig. 6.1 - from independence to confounding co-occurring with direct causation. We emphasize two limitations of the classifier:

1. Since the classifier makes a forced choice between the six acyclic alternatives, it will necessarily produce 100% error on P_{XY}'s generated from cyclic systems.
2. Our goal was not, and can not be, to achieve 100% accuracy. For example, error in Fig. 6.9A is about 25%. However, this is not necessarily a "bad" result. Our considerations in Sec. 6.3 and 6.4 show that even when all our assumptions hold, the *optimal* classifier has a non-zero error.

The latter is a consequence of the non-identifiability of the problem: it is not possible, in general, to identify the causal structure between two variables by looking at the joint distribution and without intervention. Our goal was to introduce a minimal set of assumptions that, while acknowledging the nonidentifiability, enable us to make useful inferences.

We noted that as the cardinality of the variables grows, the task becomes more and more "identifiable" in the sense that, for each given P_{XY}, one out of the possible six causal graphs strongly dominates the others with respect to its likelihood. In this situation, the *most likely* causal structure becomes essentially the only possible one, barring a small error, and the problem becomes *practically* identifiable.

All of the above applies assuming that our generative model corresponds to reality. The assumptions, discussed in Sec. 6.3, boil down to two ideas: 1) The world creates causes independently of causal mechanisms and 2) Causes are random variables whose distributions are sampled from flat Dirichlet hyperpriors. Causal mechanisms are conditional distributions of effects given causes, and are also sampled from flat Dirichlet hyperpriors. Whether these assumptions are realistic or not is an open question. Nevertheless, through a series of simple experiments (Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig. 6.9) we showed that the assumption of flat hyperpriors is not essential – our classifiers' average performance does not decrease significantly as we allow the hyperpriors to vary, although the variance of the performance grows. In future work, we will carefully analyze under what conditions the flat-hyperprior classifier performs well even if the hyperpriors are not flat. The current working hypothesis is that as long as the hyperprior on $P(cause)$ is the same as the hyperprior on $P(effect \mid cause)$, the classification performance doesn't change significantly *on average*, but –as seen in our experiments – it will have increased variance.

Shohei Shimizu explained our task (for the case of continuous variables) as: "Under what circumstances and in what way can one determine causal structure based on data which is not obtained by controlled experiments but by passive observation only?" (Shimizu et al., 2006). Our answer is, "For high-cardinality discrete variables, it seems enough to assume independence of $P(cause)$ from $P(effect \mid cause)$, and train a neural network that learns the black-box mapping between observations and their causal generative mechanism."

6.8 Conjectured Application to CFL

Throughout this section, we freely use terminology, definitions and algorithms of Chapters 2 and 3. We propose the following procedure to discover the causal partition directly from observational data:

1. Take as input an *unsupervised observational dataset* $(x_1, y_1), \cdots, (x_N, y_N)$.
2. Use Alg. 2 to learn the observational partition of the data.
3. Iterate through each pair of coarsenings of the observational partitions on $\mathcal{X}$ and $\mathcal{Y}$. Name the i-th coarsening pair $(C', E')_i$.
4. Each C', E' pair is a *causal hypothesis*. There are now four possibilities: 1) The pair constitutes the causal partition, 2) The pair is a coarsening of the causal partition, or 3) The pair is a spurious correlate (see Chapter 2) or its coarsening, and finally 4) The pair is neither of the three.
5. Using methods of this chapter, among all the C', E' pairs pick the *coarsest one* classified as $P(C'_i \rightarrow E'_i)$.

In a nutshell, the algorithm first learns the observational partition. In previous chapters, we suggested efficient algorithms to pick out the causal partition based on the observational partition using experimentation. This algorithm substitutes the interventional experiments with methods of this chapter.

In Step 4, we enumerated four possible causal meanings for each coarsening of the observational partition. These claims rely on the Causal Coarsening Theorem. First of all, we know that the causal partition is one of the coarsenings of the observational one. Note that, in the terminology of this chapter, the causal partition C, E stands in the direct causal reaction: $C \rightarrow E$. We also know that the spurious correlate pair S, E_S stands in the relation $S \leftarrow H \rightarrow E_S$ – by definition, the spurious correlate is confounded and not causal. Note that all the coarsenings that are not the causal partition are probabilistic variables that *are not well-defined causal variables*, because they do not support unambiguous manipulation (see Chapter 1). In particular, any coarsening that contains microvariables from two distinct causal classes is not a causal variable.

Nevertheless, since all the coarsenings are probabilistic variables, the methods of this chapter can in principle be applied to them. Unfortunately, at the time of writing of this book the efficacy of this approach has not been researched.

Note that the algorithm iterates over *all the possible partitions* of the observational partition. This means its runtime is super-exponential in the number of observational classes. The methods of this chapter however work best for high-cardinality variables. This means that the algorithm of this section works best when it is slowest – that is, when the observational partition has a large number of cells.

Chapter 7

Discussion

CFL can learn high-level causal knowledge from low-level data in an automatic, unbiased manner. Given that discussion of macro-causal relations is commonplace in scientific discourse, we take the scientific endeavors mentioned in the first chapter to be predicated on the assumption that micro-level descriptions are not all there is to the phenomena under investigation. Whether or not there in fact are macro-level causes that justify such an assumption is, in light of our theoretical account, an empirical question. Taking the definitions literally, macro-causes cannot be defined arbitrarily.

Throughout this book, we outlined the theory of causal macrovariables and proposed algorithms for their learning. We cleanly accounted for the interventional/observational distinction that is central to most analyses of causation. This distinction is entirely lost in heuristic approaches, such as that of Hoel et al. (2013).

Altogether, we have an account of how causal variables can be identified that does not rely on a definition obtained from domain experts. Given its theoretical generality, we expect our method to be useful in many domains where micro-level data is readily available, but where the relevant causal macro-level factors are still poorly understood.

Our contribution is most directly to the field of causal discovery. Modern causal discovery algorithms presuppose that the set of causal variables is well-defined and meaningful. What exactly this presupposition entails is unclear, but there are clear counter-examples: x and $2x$ cannot be two distinct causal variables. There are also well understood problems when causal variables are aggregates of other variables (Chu et al., 2003; Spirtes and Scheines, 2004). We provide an account of how causal macro-variables can supervene on micro-variables.

In general it is possible that macro-variable causes C and effects E are barely coarser (if at all) than the corresponding micro-variables. The hope that C and E have a “manageable” cardinality is similar in spirit to standard assumptions in both supervised and unsupervised learning. There, a set of continuous data is clustered into a discrete number of subsets according to some feature of interest. Here the “feature of interest” is the causal relationship between C and E.

In this final chapter, we make explicit several assumptions that our methods presuppose, and discuss their significance to real-world applications. In our minds, the largest contribution of this thesis is the theory of causal macrovariables and their learning. At this moment in time, compelling applications of the framework to science or industry are missing – but are the next, and perhaps the most important, step in the evolution of the framework.

7.1 CFL in the Real World: Assumptions and Challenges

Chapters 2 and 3 proposed methods to learn causal macrovariables from observational microvariable data. Applications of CFL presented in this work (enumerated in Chapter 1) range from almost abstract toy problems to attempts to gain knowledge about the mechanisms driving Earth climate.

CFL makes a set of assumptions that do not necessarily hold in all real-world settings. Assessing to what degree violations of these assumptions decrease usefulness of the framework is an open issue, but we can at least lay out and discuss some of the caveats.

7.1.1 Discreteness of Macrovariables

The essential assumption of CFL in its current form is that the macrovariables are *discrete* – that is, the statistics of the system, while supervening on continuous microvariables, can be captured by discrete variables with manageable cardinalities.

In all our toy examples (Sec. 1.3), the micro-variable spaces all collapse to an observational partition with a small number of cells. Each input cell consists of microvariable states that share *exactly the same* conditional probabilities w.r.t. each target cell. Many real-world phenomena, however, are thought to have continuous probabilistic structure.

In Fig. 1.1A for example, temperature is a continuous macrovariable. The observational partition (with respect to any variable that is not independent of temperature) still exists, but it divides the state space of particle masses and velocities into uncountably many cells. Two states belong to the same cell of that partition if and only if the average kinetic energy of its particles is equal. This observational partition corresponds precisely to the ‘temperature’ macrovariable. Unfortunately, an equivalent of the Causal Coarsening Theorem for uncountable partitions does not currently exist, so the value of such a partition for causal discovery is unclear.

Consequently, before applying CFL it is essential to establish whether the probabilistic structure of the problem can possibly be captured or at least well-approximated by discrete variables. In low-dimensional domains visualization of the data can provide guidance. Expert knowledge or physical intuition can also justify the discreteness assumption. We discuss extensions of CCT to the continuous case in Sec. 7.2.2.

7.1.2 Smoothness Assumptions During Learning

While the theory of CFL assumes a discrete macrovariable structure, our learning algorithms make the contrary assumption. Recall the example of hue influencing eda, from Sec. 1.3.2. Consider $p(eda \mid hue)$ evaluated on a fixed set of h samples as a vector-valued function of hue. Fig. 3.1A shows that this function is discontinuous at the boundaries of the observational states (namely at $hue = 0, 90, 180, 270$). However, the learned density shown in Fig. 3.1B varies continuously with hue. We chose to learn a continuous density mainly because there are good and flexible neural-net-based algorithms for learning continuous conditionals. As Fig. 3.1B suggests, these algorithms can take sharp boundaries into account. Nevertheless, mistakes at the boundaries are a likely artifact of the learning method.

A similar situation is encountered in neural network classification (Rumelhart et al., 1985; Bishop, 1995; Krizhevsky et al., 2012): an essentially discrete problem (dividing the feature space into a discrete number of classes) is solved using a continuous algorithm and appropriate thresholding of the final output. The success of neural networks in machine learning tasks proves that this strategy can yield good results.

7.1.3 Why Not Naive Clustering?

It is instructive to compare our results with unsupervised clustering. Recall our results of learning the observational partition on climate science data in Chapter 5. We used the "precision coefficients" (Definition 22) to measure the degree to which each cell of the observational partition corresponds to El Niño. Fig. 7.1 shows the precision coefficients for k-means clustering with k=4, $\dots$, 16 (small dotted line), alongside our CFL results. Whereas CFL detects both El Niño and La Niña with high precision using only four states, k-means struggles to achieve a similar result even for larger K.

Barring particularities of the data, there is in general no reason for CFL to give the same results as clustering. Consider the example in Fig. 7.2. Arguably, a reasonable clustering algorithm should find four linearly separable clusters in the joint $\mathcal{X}, \mathcal{Y}$ space, and two clusters in the $\mathcal{X}$ and $\mathcal{Y}$ space each. However, the variables are probabilistically independent. In contrast, CFL would only find a one-state input variable, since all values of X (in non-zero density regions) imply the same distribution over Y. Additionally, since $P(Y \mid X) = P(Y)$ is constant across all the Y samples, CFL would also only find a one-state output variable.

Our experiments that compare CFL with clustering showed that, as the number of clusters grows, k-means approaches never exceed CFL's precision in detecting El Niño and La Niña. One explanation for this finding is that while clustering looks for *spatial features* in the data, CFL looks for *relational probabilistic features*. Fig. 7.1 suggests that when the number of clusters is small there are strong spatial features in the data that supersede El Niño and La Niña in their distinctiveness. In contrast, CFL already detects El Niño with high precision with only four clusters. This indicates that either (1) There is something unique about $P(\text{El Niño} \mid \text{W})$ and $P(\text{La Niña} \mid \text{W})$, or (2) There is something unique about $P(\text{El Niño})$ and $P(\text{La Niña})$.

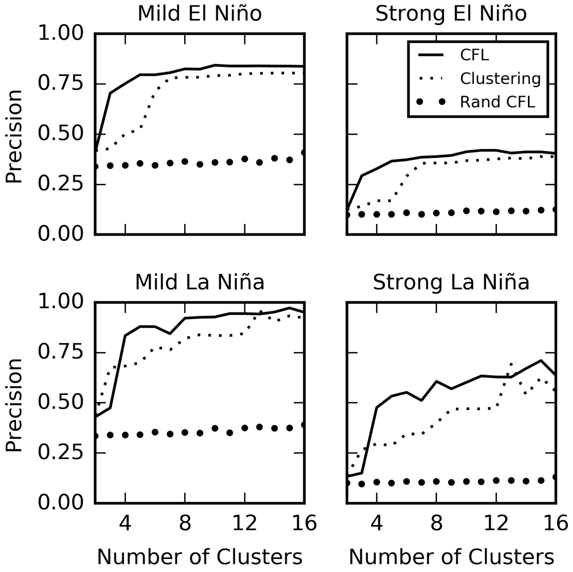


Figure 7.1: Changes in macro-variable precision as we vary the number of states in CFL, clustering, and CFL on reshuffled data (“Rand CFL”). With two states, it is impossible to differentiate El Niño and La Niña from other weather features, be it dynamic (CFL) or spatio-structural (clustering). Increasing the number of states reveals differences between the algorithms.

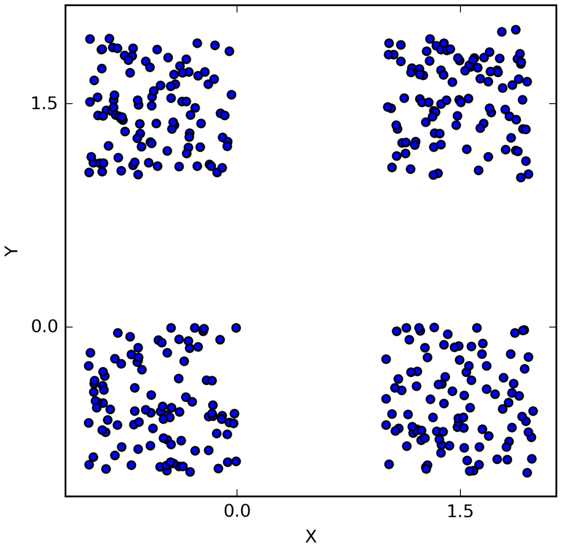


Figure 7.2: Samples from a two-dimensional distribution uniform over four square space regions. A reasonable clustering algorithm would divide this space into four regions. However, CFL sees only one observational class both in the input and the output space, as all the regions of non-zero density have the same $p(Y \mid X)$ (and the conditional density is not defined over other regions).

Since we disproved the second hypothesis in Sec. 5.3.4, our results overall indicate that the El Niño and La Niña phenomena do not only constitute interesting spatial features of the SST map, but are also crucially characterized by the dynamic aspect of the interplay between zonal winds and sea surface temperatures.

7.2 Open Problems

The following problems point to future work that would significantly extend the range of domains CFL can be applied to.

7.2.1 Learning Without Experimentation

Currently, transitioning from the observational testing the causal hypothesis requires intervention – one experiment per each observational class. However, in the field of causal discovery there are methods to reject causal hypotheses based on observational data only. These are either based on the independence structure of the generative distributions (Spirtes et al., 2000; Chickering, 2002; Silander and Myllymäki, 2006; Claassen and Heskes, 2012; Hyttinen et al., 2014) or assumptions about the functional form of the structural equations that govern the system (Shimizu et al., 2006; Hoyer et al., 2009; Mooij et al., 2011). None of these methods can be directly applied to the formation of causal macrovariables – they all assume the causal variables of interested are given. Extending these ideas to the CFL framework would make it useful in domains where direct experimentation is expensive (medicine) or impossible (climate science).

In Chapter 6, we developed a novel method to establish the likelihood of different causal structures based on observational data only. The development of this method was motivated by the needs of CFL. We also proposed an algorithm that – we conjecture – can pick out the causal partition from observational data only. How well and under what conditions the algorithm works however, is at the moment entirely unclear.

7.2.2 Continuous Macrovariables

Section 7.1 discussed the discreteness assumption in CFL. A logical next step is to extend the framework to systems where the macrovariables are continuous, or hybrid systems. Whereas definitions of continuous macrovariables do not pose a challenge, extending the Causal Coarsening Theorem — which makes the framework useful — to the continuous case appears non-trivial.

7.2.3 Cyclic Microvariable Graphs

Like the majority of work in causal inference and discovery — notable exceptions being (Richardson, 1996; Mooij et al., 2011; Lacerda et al., 2008; Hyttinen et al., 2012, 2014) — CFL assumes that the microvariable system is acyclic: in our toy example, hue has causal influence on eda, but we assumed (quite plausibly) that eda is not a cause of hue. This assumption is not always warranted. For example, in the wind-temperature climate case we worked with in Chapter 5, the system definitely experiences feedback (over time). While cyclicity may not break the CCT or our algorithms, there is currently no proof either way.

7.3 Brief Philosophical Considerations

Our goal has been to provide a rigorous and objective account of causal macrovariables as they occur in the sciences. Motivation has come from examples, such as *temperature* that supervenes on the kinetic energy of particles. Just as it is the room temperature – the *mean* kinetic energy of the particles – that triggers the air conditioning, rather than the exact distribution of particle velocities in the room, we have defined causal macrovariables as aggregates of those microvariables that have the same causal consequences.

Non-trivial macrovariables exist to the extent that there are such equivalent microstates. There is a sense in which the occurrence of macrovariables is a measure-zero event. Whether or not this licenses inferences to the existence of macrovariables in practice depends on the appropriateness of the measure for the description of our world. But there is a further consideration worth noting: We claimed that it was in fact the mean kinetic energy and not the exact distribution of kinetic energies of the particles that determined whether the air conditioning was triggered.

But perhaps that is not quite right. After all, it is the specific movement of the particles close to the sensor that triggers the air conditioning. We could maintain the mean kinetic energy of the particles in the room overall constant, while significantly changing the velocities of the particles close to the sensor. In that case one may argue that temperature is not a causal macrovariable in this system. Another view is to say that these sorts of microstates are extraordinarily improbable and therefore can be neglected. Assuming that such a view can be properly formalized, the macrovariable *temperature* then does not have a completely clean delineation in terms of its causal consequences. There will be a few micro-states within each of its macro-states that have very different causal consequences from the other micro-states within the same macro-state. Metaphorically, the macrovariable is a little bit "fuzzy around the edges". Such a metaphysical account of macrovariables may be anathema to many, but we note that our epistemology – our learning method – is unable to distinguish between these and sharply delineated macrovariables since we will in practice never be able to investigate all possible micro-states.

With these definitions there is no reason *a priori* to think that macro-variables are common phenomena. In fact quite the opposite: The conditions that the probability distributions over X and Y must satisfy to give rise to non-trivial macro-variables C and E can easily be described as a measure-zero event when taken in their strict form. Consequently, our view is that to the extent that macro-variables are discussed in a scientific domain, there must be a pre-supposition that such strong conditions are satisfied at least approximately.

Bibliography

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *6th ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.

P. A. Cashin, K. Mohaddes, and M. Raissi. Fair weather or foul? The macroeconomic effects of El Niño. 2015.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 181–190. AUAI Press, 2015.

K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt. Unsupervised discovery of El Niño using causal feature learning on microlevel climate data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016a.

K. Chalupka, P. Perona, and F. Eberhardt. Multi-level cause-effect systems. In *19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016*, 2016b.

K. Chalupka, P. Perona, and F. Eberhardt. Introduction to Causal Feature Learning. *Behaviormetrika*, 44(1), 2016c.

K. Chalupka, P. Perona, and F. Eberhardt. Black-box Identification of Unidentifiable Causal Relationships. *arXiv*, 2016d.

S. A. Changnon. Impacts of 1997-98 El Niño-generated weather in the United States. *Bulletin of the American Meteorological Society*, 80(9):1819, 1999.

D. M. Chickering. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2(Feb):445–498, 2002.

T. Chu, C. Glymour, R. Scheines, and P. Spirtes. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152, 2003.

T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of UAI*, page 207–216. AUAI Press, 2012.

T. Di Liberto. The Walker Circulation: ENSO's atmospheric buddy, 2014.

M. H. Glantz. *Currents of change: impacts of El Niño and La Niña on climate and society*. Cambridge University Press, 2001.

I.J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.

K. Grammer and R. Thornhill. Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, 108(3):233–242, 1994.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–260, 2007.

I. Guyon, A. Elisseeff, and C. Aliferis. Causal feature selection. In *Computational Methods of Feature Selection Data Mining and Knowledge Discovery Series*, pages 63–85. Chapman and Hall/CRC, 2007.

Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

G. E. Hinton and N. Srivastava. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

J. R. Holton, R. Dmowska, and S. G. Philander. *El Niño, La Niña, and the southern oscillation*, volume 46. Academic Press, 1989.

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

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. *arXiv preprint arXiv:1210.4879*, 2012.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with Answer Set Programming. In *Thirtieth Conference on Uncertainty in Articial Intelligence*, 2014.

E. M. Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

K. W. Jacobs and F. E. Hustmyer. Effects of four psychological primary colors on gsr, heart rate and respiration rate. *Perceptual and motor skills*, 38(3):763–766, 1974.

D. Janzing, J. Peters, J. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 249–257. AUAI Press, 2009.

N. C. Johnson. How many ENSO flavors can we distinguish? *Journal of Climate*, 26(13):4816–4827, 2013.

M. Kanamitsu, W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter. NCEP-DOE AMIP-II reanalysis (r-2). *Bulletin of the American Meteorological Society*, 83(11):1631–1643, 2002.

H.-Y. Kao and J.-Y. Yu. Contrasting eastern-Pacific and central-Pacific types of ENSO. *Journal of Climate*, 22(3):615–632, 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

G. Lacerda, P. L. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Twenty-fourth Conference on Uncertainty in Artificial Intelligence*, pages 366–374, 2008.

C. W. Landsea and J. A. Knaff. How much skill was there in forecasting the very strong 1997-98 El Niño? *Bulletin of the American Meteorological Society*, 81(9):2107–2119, 2000.

K. M. Lau and S. Yang. Walker circulation. *Encyclopedia of atmospheric sciences*, pages 2505–2510, 2003.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *Eighth IEEE International Conference on Computer Vision*, volume 2, pages 251–256. IEEE, 2001.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *ACM SIGIR Seventeenth Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

M. J. McPhaden, S. E. Zebiak, and M. H. Glantz. ENSO as an integrating concept in Earth science. *Science*, 314(5806):1740–1745, 2006.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, pages 639–647, 2011.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.

J. Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Synthese-Dordrecht*, 104(1):161, 1995.

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

J. Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010.

J. P. Pellet and A. Elisseeff. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.

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(12):2436–2450, 2011.

H. Reichenbach. *The direction of time*, volume 65. University of California Press, 1991.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461. Morgan Kaufmann Publishers Inc., 1996.

C. F. Ropelewski and M. S. Halpert. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, 115(8):1606–1626, 1987.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, No. ICS-8506. California University of San Diego La Jolla Institute for Cognitive Science, 1985.

S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

U. Rutishauser, O. Tudusciuc, D. Neumann, N. Adam, A. Mamelak, C. Heller, I. B. Ross, L. Philpott, W. W. Sutherling, and R. Adolphs. Single-unit responses selective for whole faces in the human amygdala. *Current Biology*, 21(19):1654–1660, 2011.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Langauge Processing*, pages 1070–1079, 2008.

C. R. Shalizi. *Causal architecture, complexity and self-organization in the time series and cellular automata*. PhD thesis, University of Wisconsin at Madison, 2001.

C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, 2001.

C. R. Shalizi and C. Moore. What is a macrostate? Subjective observations and objective dynamics. *arXiv preprint cond-mat/0303625*, 2003.

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

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of UAI*, page 445–452. AUAI Press, 2006.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

J. Snoek, J. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

P. Spirtes and R. Scheines. Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5): 833–845, 2004.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

N. Srinivas, A. Krause, M. Seeger, and S. M. Kakade. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible markov kernels. In *ISAIM*, 2006.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL `http://arxiv.org/abs/1605.02688`.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001.

K. E. Trenberth. The definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2777, 1997.

D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, and M. S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (2579-2605):85, 2008.

S. V. N. Vishwanathan. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

C. K. I. Williams and C. E. Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-fifth conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.