Behaviormetrika (2017) 44:137–164
DOI 10.1007/s41237-016-0008-2

INVITED PAPER

Causal feature learning: an overview

Krzysztof Chalupka[1] · Frederick Eberhardt[2] ·
Pietro Perona[3]

Received: 29 August 2016 / Accepted: 29 November 2016 / Published online: 26 December 2016

Abstract Causal feature learning (CFL) (Chalupka et al., Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence. AUAI Press, Edinburgh, pp 181–190, 2015) is a causal inference framework rooted in the language of causal graphical models (Pearl J, Reasoning and inference. Cambridge University Press, Cambridge, 2009; Spirtes et al., Causation, Prediction, and Search. Massachusetts Institute of Technology, Massachusetts, 2000), and computational mechanics (Shalizi, PhD thesis, University of Wisconsin at Madison, 2001). CFL is aimed at discovering high-level causal relations from low-level data, and at reducing the experimental effort to understand confounding among the high-level variables. We first review the scientific motivation for CFL, then present a detailed introduction to the framework, laying out the definitions and algorithmic steps. A simple example illustrates the techniques involved in the learning steps and provides visual intuition. Finally, we discuss the limitations of the current framework and list a number of open problems.

Keywords Causal discovery · Causal inference · Graphical models · Bayesian networks · Macrovariables · Multiscale modeling

Communicated by Shohei Shimizu.

✉ Krzysztof Chalupka
kjchalup@caltech.edu

[1] Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

[2] Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA

[3] Electrical Engineering, California Institute of Technology, Pasadena, CA, USA

1 Introduction

Causal feature learning (CFL) is an unsupervised machine learning and causal inference framework with two goals: (1) the formation of high-level causal hypotheses using low-level input data, and (2) efficient testing of these hypotheses (Chalupka et al. 2015, 2016a, b). As a motivation, consider the following archetypical research situation (Fig. 1c): a neuroscientist notices that a specific neuron responds preferentially to some images containing humans. He 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. He concludes 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) collecting evidence from such experiments.

Steps (a) and (b) are guided by prior knowledge, intuition and formal reasoning. CFL aims to 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 faces) that are causal, then these ought to be detectable by an unsupervised learning algorithm.

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 algorithms.

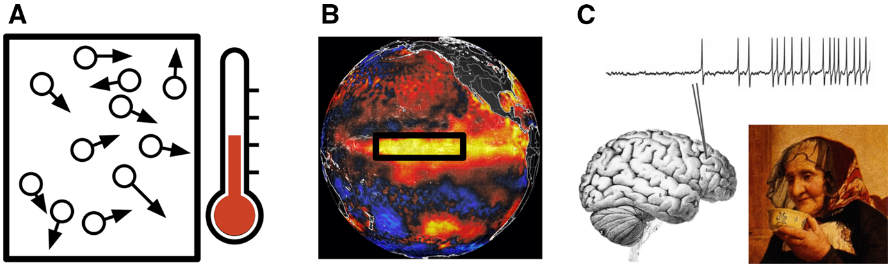


Fig. 1 Causal macrovariables. Macrovariables in science can often be thought of as functions of the underlying microvariable space. Each such function f corresponds to a *partition* on the microvariable state space, defined by the equivalence relation $x_1 \sim x_2 \iff f(x_1) = f(x_2)$. **a** Temperature 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 °C. It is a binary function of the high-dimensional 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 high-level feature of the space of all images

1.1 Macrovariables in science

Just about any scientific discipline is concerned with developing ‘macrovariables’ that summarize an underlying finer-scale structure of ‘microvariables’ (see Fig. 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 temperatures 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 among macrovariables that hold independent of the microvariable instantiations of the macrostates.

Causal feature learning 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. The framework draws heavily on ideas developed in computational mechanics (Shalizi and Crutchfield 2001; Shalizi 2001; Shalizi and Moore 2003) and connects them with the framework of causal graphical models (Spirtes et al. 2000; Pearl 2000).

CFL 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 exponentially many states (such as images) are microvariables.

One may think of macrovariables as equivalence relationships 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 sea surface temperature maps where the temperature anomaly in a specified region of the Pacific Ocean exceeds 0.5 °C correspond to El Niño. CFL also defines the relation between micro- and macrovariables in terms of an equivalence relation, which we review more formally in Sect. 2.1.

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 $\mathcal{H}$ and $\mathcal{E}$ as input, with the task of discovering “what in $\mathcal{H}$ causes what in $\mathcal{E}$.”
2. Search the space of all macrovariables (equivalence relationships) on $\mathcal{H}$ and retain only those that could be causes of $\mathcal{E}$.
3. Search the space of all the macrovariables that supervene on $\mathcal{E}$ and retain only those that could be effects of $\mathcal{H}$.
4. Propose an efficient experimental procedure that picks out the (unique) macrovariable cause and effect from among the retained macrovariable pairs.

In general, there is an infinite number of macrovariables one could define as supervening on the 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 the interplay between causality and aggregation: ambiguous manipulations.

1.2 Ambiguous manipulation and causal macrovariables

Figure 2 illustrates a case of an unfortunate choice of a macrovariable. Total cholesterol used to be considered as 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 and high-density lipoproteins (LDL and HDL, commonly called "bad" and "good" cholesterol), which have different effects on heart disease (see Spirtes and Scheines 2004 for an in-depth discussion of this case.)

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 is an appropriate 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 of aggregation.

Causal feature learning addresses this concern by learning unambiguous causal variables: each macrovariable state must have a consistent well-defined causal effect. This effect can be probabilistic and highly variable, but may not depend on the microvariable instantiation of the macrovariable—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. 1a–c.

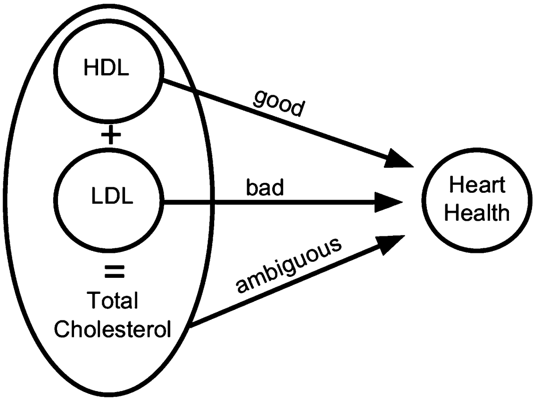
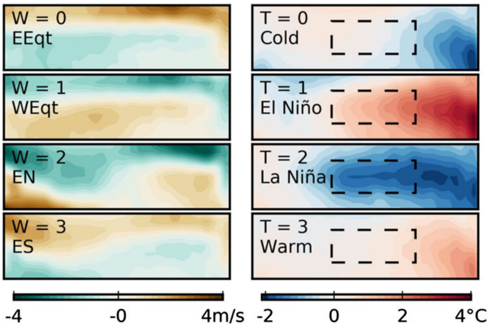


Fig. 2 Ambiguous macrovariables. Total cholesterol is the sum of low- and high-density lipids (LDL and HDL). Suppose that LDL causes heart disease and HDL prevents it. The effect of total cholesterol on heart disease is then ambiguous as it depends on the proportion of HDL vs. LDL (see Sect.1.2). Experimental procedures based on adjusting total cholesterol only can give inconsistent results



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

Fig. 3 Macrovariables of Pacific weather patterns. Chalupka et al. (2016a) applied CFL to climate data. The microvariables consisted of zonal (East–West) wind strength over the equatorial Pacific and sea surface temperatures (SST) over the same region of space. 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 $P(T|W)$, the probability of the hypothesized SST macrovariable given the hypothesized wind macrovariable. It shows that CFL learned at least two relations that causally interpreted, is consistent with current climate science: 'Westerly winds' ($W = 1$) cause El Niño ($T = 1$) and 'Easterly winds' ($W = 0$) cause La Niña ($T = 2$)

1.3 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 aggregation of pixel 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 "input" and "output" spaces (e.g., the statistics of the environment as imaged by the optic array, and the desired behavior), rather than by one or the other of the spaces considered individually.

Consequently, in the simplest case, CFL is concerned with the relationship between two microvariable spaces: the 'input' and 'output' space. The task is then to detect all the causes (in the input space) of all the effects (in the output space), in their simplest form. Any distinctions unimportant to the task at hand will be abstracted away.

For example, Chalupka et al. (2016a) take 'wind strength map over Pacific' as the input space, and 'sea surface temperature (SST)' as the output space. Applying CFL to this task yields a discrete division of the two spaces into a set of wind pattern classes ('Westerly Winds', 'Easterly Winds' etc) and SST pattern classes ('El Niño', 'La Niña' etc)—see Fig. 3. Knowing which class a wind pattern belongs to then gives all the useful information about its *possible* effects on SST[1]. However, given a different output space—say "average US income"—the input macrovariable would change, unless the causal consequences are entirely mediated by the same temperature macrovariable.

1.4 A toy example

To visualize the definitions and main algorithmic steps involved in CFL, we will resort to a simple toy example of a fictitious study on the influence of color on the electrodermal response (eda), also known as the skin conductance[2]. In the fictitious experiment, the eda 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. 4. In the system, "Red" hues increase eda (a perhaps controversial but plausible response, see 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 necessarily makes the example slightly contrived, the visualizations of the algorithms and definitions are much simpler and more illuminating than in the higher dimensional case.
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) through direct causation, but they are at the same time confounded by geographic location. As discussed in Sect. 4.1 in more detail, separating confounding from causation differentiates CFL from, for example, many other learning methods.

Whereas our simulated dataset is simple and low-dimensional, CFL, in the same form as presented here (Algorithms 1 and 2), can be applied to very high-dimensional and complex data (see Chalupka et al. 2015, 2016a, b for example complex applications).

In our model, *eda* [in units normalized to (0, 1) where 0.5 is the global average] is causally influenced by *hue* (represented in degrees, with 0 being the red hue, see Fig. 4) 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

[1] As discussed in some detail in Chalupka et al. (2016a), a causal interpretation of purely observational data is not possible without further assumptions.

[2] Python code that implements the learning algorithms and reproduces all the figures and experimental results is available online at http://vision.caltech.edu/~kchalupk/code.html.

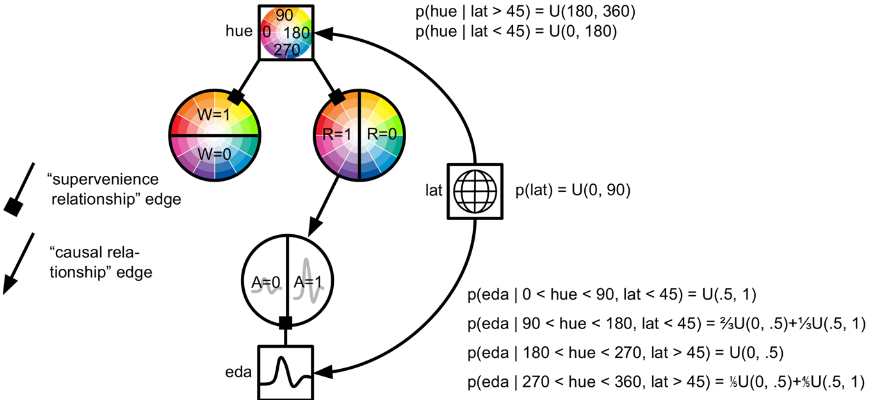

hue
90
0
180
270
p(hue | lat > 45) = U(180, 360)
p(hue | lat < 45) = U(0, 180)
W=1
W=0
R=1
R=0
lat
p(lat) = U(0, 90)
"supervenience relationship" edge
"causal relationship" edge
A=0
A=1
eda
p(eda | 0 < hue < 90, lat < 45) = U(.5, 1)
p(eda | 90 < hue < 180, lat < 45) = ⅔U(0, .5)+⅓U(.5, 1)
p(eda | 180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

Fig. 4 A toy causal model. In the simulated study, 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 warm climate predominant there. Finally, lower latitude environments tend 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 $p(eda) = 0.2\,U\,(0, 0.5) + 0.8\,U\,(0.5,\ 1)$—a mixture of two uniform distributions that indicates that most likely, eda is above the average in this situation

confounded, as illustrated in Fig. 4. Assuming that *lat* fully captures the causal confounding between *hue* and *eda*, their joint distribution $p(hue,\ eda)$ factorizes as

$$p(hue, eda) = \sum_{lat} p(eda|hue, lat)p(eda|lat)p(lat), \tag{1}$$

and can be described by the causal graphical model shown in Fig. 4.

The probability tables for these factors are shown in Fig. 4. We constructed the conditional $p(eda|hue, lat)$ to take a special form: there are four ranges of *hue* within which $p(eda|hue)$ is constant. For example, the conditional is the same for any $hue \in (0, 90)$. This construction indicates that there are *macrovariables* driving the relationship 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 relationship between, say, water and human pain receptors. To a good approximation, any body of water with the same temperature has the same effect on our pain receptors, no matter the individual velocities of specific water particles.

In Sect. 1.1, we will formalize the notion of a macrovariable and its relationship to the underlying microvariable. We will show that such 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. 4. Throughout the paper we will show that this model is in fact *the* macrovariable structure that supervenes on *hue* and *eda*.

Our macrovariables are all binary. A supervenes (is a function of) *eda*, with $A = 1$ if and only if $eda > 0.5$. That is, A represents an "Above-average" skin conductance. A is caused by R, 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*, $W = 1 \iff hue \in (0, 180)$.

The causal graph of R, W and A, shown in Fig. 4, is determined by the variables' supervenience on *hue* and *eda*. Similarly, the joint $P(R, W, H)$ is fully determined by $p(hue, eda)$. In the remainder of the article, we will show how CFL can recover such macrovariables, as well as the causal graph over these variables, given only observational samples of the microvariables and a few carefully selected interventions in the microvariable space.

2 Learning causal features

CFL takes microvariable inputs and produces macrovariable causal hypotheses. As discussed in Sect. 1.1, we will now define macrovariables as equivalence classes over our microvariables. Throughout this section, let $\mathcal{H} = (0, 360)$ denote our input microvariable space (range of the hue variable), and $\mathcal{E} = (0, 1)$ the output microvariable space (range of eda). We will denote the random variables defined over these spaces as *hue* and *eda* as before, and their specific instantiations as h and e, for example, we will write $p(eda = e|hue = h)$ for some $e \in \mathcal{E}, h \in \mathcal{H}$.

2.1 Learning the causal hypothesis

Fig. 5A shows 1,000 samples from $p(hue, eda)$ together with the ground-truth conditional distribution $p(hue|eda)$. These observations follow Eq. (1), where *hue* and *eda* are confounded by the unobserved *lat*.

The empirical distribution shown in the figure indicates that $p(eda|hue)$ is constant for any $h \in (0, 90)$ as well as for $h \in (90, 180)$, $h \in (180, 270)$ and $h \in (270, 360)$. Figure 4 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

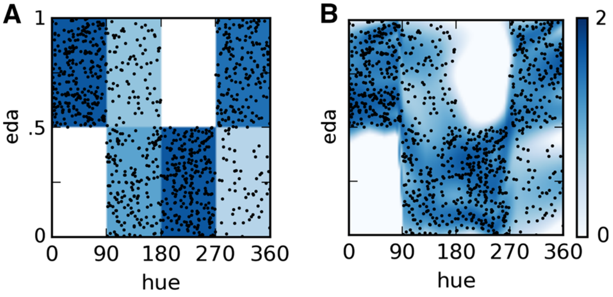


Fig. 5 Model samples and pdf. **a** *Black dots* are samples from the joint $p(hue, eda)$, *background color* shows the ground-truth value of $p(eda|hue)$. **b** The result of conditional density learning of $p(eda|hue)$ using a mixture density network (see Sect. 2.1)

observational distinctions is the key element of CFL. This construction, called the *Observational Partition*, abstracts away all the irrelevant micro-level details:

Definition 1 (*Observational partition, observational class*) The *observational partition of* $\mathcal{H}$, denoted by $\Pi_o(\mathcal{H})$, is the partition induced by the equivalence relation $\sim_h$ such that

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

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

$$e_1 \sim_e e_2 \quad \Leftrightarrow \quad \forall_{h\in\mathcal{H}}\, p(e_1|h) = p(e_2|h).$$

A cell of an observational partition is called an *observational class* (of $\mathcal{H}$ or $\mathcal{E}$).

The observational partition of $\mathcal{E}$ is also easily discerned from Fig. 5: $e \in (0, 0.5)$ has the same $P(e|h)$ for any h. Let us index the observational classes on $\mathcal{H}$ as $H = 0, 1, 2, 3$ if $h \in (0, 90)$, (90, 180), (180, 270), (270, 360), respectively, and $E = 0, 1$ if $e \in (0, 0.5)$ and (0.5, 1), respectively. We can then compress $p(eda|hue)$ to only four numbers without losing any information:

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

Note that this corresponds to $p(A|R, W)$ in Fig. 4. However, whereas R truly is a cause of A, the noncausal 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*. We will discuss the role of the observational partition in causal learning in detail in Sect. 2.2 below. Before that, let us see how to learn the observational class of *hue* and *eda* from data.

Learning the observational partition amounts to clustering $\mathcal{H}$ such that all the h belonging to one cluster induce the same $p(eda|hue = h)$, and clustering $\mathcal{E}$ such that all the e in one cluster have the same likelihood $p(eda = e|hue)$ for any value of *hue*. We outline the procedure in Algorithm 1. 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. 5b, the learned density differs from the ground-truth. Nevertheless, we used this learned density to perform clustering on the $\mathcal{H}$ and $\mathcal{E}$ spaces into the ground-truth number of clusters (4 and 2, respectively). Figure 6 shows that simple K-means clustering of the density vectors accurately discovers the observational class boundaries in both $\mathcal{H}$ and $\mathcal{E}$. Sect. 3.3 discusses in detail observational partition learning in the more realistic situation where the ground-truth number of macrovariable states (clusters) is unknown.

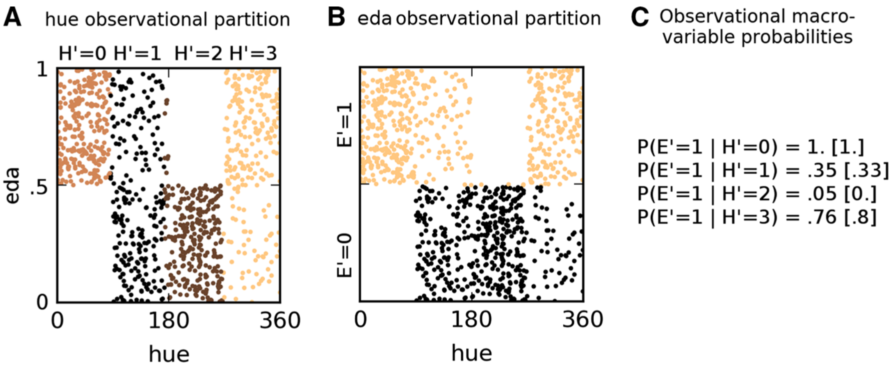


Fig. 6 Learning the observational partition. **a** The observational partition learned on $\mathcal{H}$ results from clustering the samples' h coordinate with respect to the inferred $p(eda|hue)$ shown in Fig. 5b. We indicate the learned partitions with an apostrophe, H' and E' in contrast with the ground-truth H and E. **b** The observational partition of $\mathcal{E}$, with two cells, results from clustering the samples' *eda*-coordinate with respect to the inferred $p(eda|hue)$. **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. 4) are given in square brackets

To estimate the conditional density, we used a mixture density network (MDN) (Bishop 1994) 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 with several mixture components 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.

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, $P(Y|do(X))$ is not constant. But as is well-known, the conditional probability distribution $P(Y|X)$ for any two variables X and Y does not fix the causal effect $P(Y|do(X))$. 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 2 (*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 $h \in \mathcal{H}$, while not (directly) affecting any other variables (such as *lat* or *eda*). That is, the manipulated probability distribution of the generative model in Eq. (1) is given by

$$P(eda|man(hue = h)) = \sum_{l} P(eda|hue = h, lat = l)P(lat = l). \tag{2}$$

Compare Eq. (2) in the definition with Eq. (1). In contrast to the conditional distribution $p(eda|hue = h) = \sum_l P(eda|hue = h, lat = l)P(lat = l|hue = h)$, the dependency between *lat* and *hue* is removed in the manipulated probability $p(eda|man(hue = h))$. 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 colored in 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. 4b. Note that any such experiment would, according to our model, have the same effect on *eda* (and A).

In our model, $P(A|do(R = r)) \neq P(A)$ for any r, and $P(A|do(W = w)) = P(A)$ for any w, which confirms the intuition that R is a cause of A, but W is not. However, the observational partition $\Pi_o(\mathcal{H})$ that we learned in Sect. 2.1 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. 7. 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 $p(eda|man(h_i))$ (see Algorithm 2). 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 and as illustrated in Fig. 7.

Definition 3 (*Fundamental causal partition, causal class*) The *fundamental causal partition of* $\mathcal{H}$, denoted by $\Pi_c(\mathcal{H})$ is the partition induced by the equivalence relation $\sim_h$ such that

$$h_1 \sim_h h_2 \quad \Leftrightarrow \quad \forall_{e \in \mathcal{E}} p(e|man(h_1)) = p(e|man(h_2)).$$

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

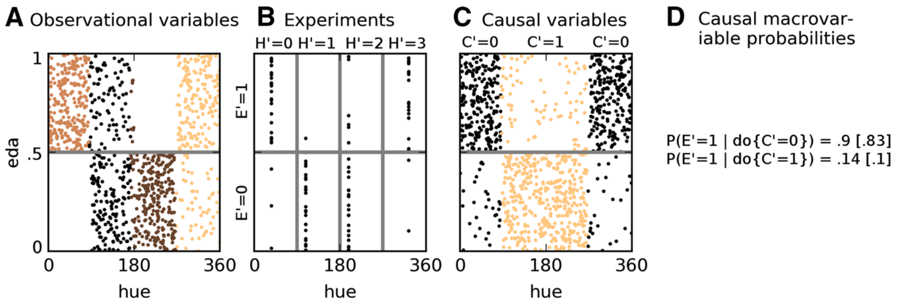


Fig. 7 Learning the causal partition. **a** The observational partition, obtained from the empirical distribution $p(eda|hue)$ (Eq. 1), is a causal hypothesis: each cell of $\Pi_o(\mathcal{H})$ could have a different effect on the probability of cells of $\Pi_o(\mathcal{E})$. **b** Conducting experiments to estimate $P(E'|do(H'=h')$ amounts to estimating $P(E'|man(hue=h))$ 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, $P(E'=1|man(hue=36)) = 22/25$ (the ground-truth is 0.83), $P(E'=1|man(hue=100)) = 1/25[0.1]$, $P(E'=1|man(hue=195)) = 6/25[0.1]$, $P(E'=1|man(hue=320)) = 23/25[0.83]$. **c** The causal partition on $\mathcal{H}$ results from merging the observational cells whose representatives induce similar $P(E'|man(hue))$. Here, we show both the causal partition (*in color*) and 1000 samples from the causal density $p(eda|man(hue))$. 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**

$$e_1 \sim_e e_2 \quad \Leftrightarrow \quad \forall_{h \in \mathcal{H}}\, p(e_1|man(h)) = p(e_2|man(h)).$$

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

That is, two microvariable states $h_1, h_2 \in \mathcal{H}$ 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 $p(eda|man(hue))$. The causal class is precisely the value of the macrovariable cause.

Definition 4 (*Macrovariable cause and effect*) The *fundamental cause* C is a random variable whose value stands in a bijective relation to the causal class of $\mathcal{H}$. The *fundamental effect* S is a random variable whose value stands in a bijective relation to the causal class of $\mathcal{E}$. 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 5 (*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 relates the causal and observational partitions. 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.

2.3 Set-up and definitions

We define general partitions of the microvariable space $\mathcal{H}$.

Definition 6 (*partition* $\Pi_f(\mathcal{H})$) Let $\Pi_f(\mathcal{H})$ be the partition on $\mathcal{H}$ induced by the relationship $h_1 \sim h_2 \Leftrightarrow f(h_1) = f(h_2)$ for any $h_1, h_2 \in \mathcal{H}$.

Here f stands for any function whose domain contains $\mathcal{H}$. For example, $P(lat|hue)$ or $P(hue)$ are such functions, where $h_1 \sim h_2$ means that $P(lat|h_1) = P(lat|h_2)$ for any value of lat. Thus, the causal and observational partition above can be rewritten as, respectively

$$\Pi_c(\mathcal{H}) = \Pi_{P(eda|man(hue))}(\mathcal{H}) \tag{3}$$

$$\Pi_o(\mathcal{H}) = \Pi_{P(eda|hue)}(\mathcal{H}) \tag{4}$$

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

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

$$h_1 \sim h_2 \quad \Leftrightarrow \quad P(h_1|lat) = P(h_2|lat) \quad \forall lat.$$

We are now ready to state the theorem:

Theorem 7 (Causal coarsening theorem) *Among all the joint distributions $P(hue, eda, lat)$, consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{H})$ and a fixed confounding partition $\Pi_{P(hue|lat)}(\mathcal{H})$. The following two statements hold within this subset:*

1. *The subset of distributions for which $\Pi_C(\mathcal{H})$ is not a coarsening of the observational partition $\Pi_O(\mathcal{H})$ is Lebesgue measure zero, and*
2. *The subset of distributions for which $\Pi_C(\mathcal{E})$ is not a coarsening of $\Pi_O(\mathcal{E})$ is Lebesgue measure zero.*

Proof of the theorem is provided in the mathematical Appendix.

An observational partition that is a coarsening of the microvariable space $\mathcal{H}$ (which is the only "interesting" kind) can arise for several reasons. To have such a coarsening, the following equation must be satisfied for at least two distinct $h_1, h_2 \in \mathcal{H}$:

$$P(eda|h_1) = P(eda|h_2) \tag{5}$$

$$\begin{aligned}
&\Leftrightarrow \sum_{lat} P(eda|h_1, lat)P(lat|h_1) - P(eda|h_2, lat)P(lat|h_2) = 0 \\
&\Leftrightarrow \sum_{lat} P(lat)(P(eda|h_1, lat)P(h_1|lat) - P(eda|h_2, lat)P(h_2|lat)) = 0 \\
&\Leftrightarrow \sum_{lat} P(eda|h_1, lat)P(h_1|lat) - P(eda|h_2, lat)P(h_2|lat) = 0 \\
&\quad \text{if P(lat)} \neq 0 \quad \forall \text{lat}
\end{aligned} \tag{6}$$

Since *lat* 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.

Consequently, for an observational partition to be a coarsening of the microvariable space, the fundamental parameters must combine in just such a way that Eq. 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 h_1 and h_2 are equal, i.e., when

$$\begin{gathered}
P(eda|h_1, lat) = P(eda|h_2, lat) \quad \forall lat \\
P(h_1|lat) = P(h_2|lat) \quad \forall lat.
\end{gathered}$$

It is these cases that are of interest to the discovery of causal macrovariables, since—intuitively—the coarseness of the observational partition arises from causal effects that are invariant across distinctions at the micro-level; as is the case in our simulated example. In other cases that satisfy Eq. 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 $\Pi_{P(hue|lat)}$ 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 $\mathcal{H}$ *only because* both the confounding partition $\Pi_{P(hue|lat)}$ and the causal partition $\Pi_{P(eda|man(hue))}$ are each coarsenings of $\mathcal{H}$, then the theorem justifies the application of the algorithms developed in this article to problems where the observational partition is itself already a coarsening of the microvariable space of $\mathcal{H}$. 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 Eq. 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 macrovariables in nature. In that case, the equality of two fundamental parameters $P(eda|lat, h_1) = P(eda|lat, h_2)$ is not coincidental but a result of a macrovariable, whereas the satisfaction of some algebraic constraint such as Eq. (6) without equalities in the fundamental parameters is a rare event.

The theorem allows us to use Algorithm 2 to find the macrovariable cause and effect with small experimental effort given microvariable data. The theorem is used in Line 2 of the algorithm. Estimating $P(E'|do(H' = h_i'))$ is equivalent to estimating $P(E'|man(hue = h_i))$ for *only one, arbitrary microvariable representative* of h_i'. This is made explicit in Fig. 7b, where we performed only four experiments (one for each observational class) to correctly learn the causal partition of $\mathcal{H}$.

3 CFL in the real world: assumptions and challenges

Sections 2.1 and 2.2 showed how to learn causal macrovariables from observational microvariable data. Our toy example demonstrates that the ideas work in a very simple setting. Chalupka et al. (2016b) used a simulated example where the input space consisted of all binary images, and the output space of all the possible activity patterns of a population of neurons over a 300 ms time window (see Fig. 8). In that case, too, CFL recovered the ground-truth macrovariable causal structure—namely that horizontal bars in an image cause a single spike in the activity of a subpopulation of neurons, and vertical bars trigger a rhythmic activation of the subpopulation. Finally, Chalupka et al. (2016a) showed that when CFL is applied to real-world climate data, it recovers a commonly believed hypothesis on Pacific wind-SST interactions: exceedingly strong westerly winds cause El Niño and strong easterly winds cause La Niña (see Fig. 3).

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

3.1 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 our toy example, the continuous space $\mathcal{H}$ collapses to an observational partition with four cells. Each cell consists of microvariable states that share *exactly the same* $p(eda|hue)$. Many real-world phenomena, however, are thought to have continuous probabilistic structure.

In Fig. 1a, for example, temperature is a continuous macrovariable. The observational partition 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

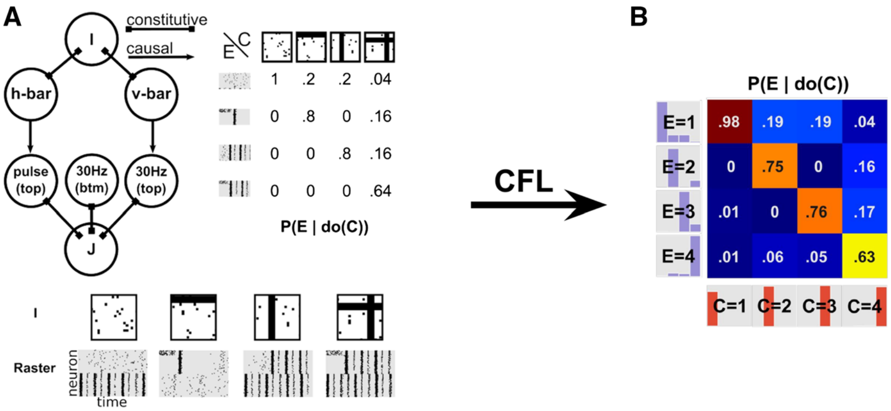


Fig. 8 CFL and simulated neuroscience. Chalupka et al. (2016b) applied CFL to a simulated neuroscience dataset. **a** The generative model of the data: images *I* constitute the microvariable input. The microvariable output consists of recordings of spikes in a population of a hundred neurons over a simulated time of 300 ms (*gray background plots*, *black dots* indicate spikes, each line corresponds to a neuron). The macrovariables in the image are the presence of a horizontal and the presence of a *vertical bar*. The first one causes neural macrovariable 'pulse', the second causes neural '30 Hz rhythm'. The conditional probability table of the macrovariables that is shown was used to generate 10,000 datapoints in the I–J space (altogether, hundreds of dimensions). **b** CFL recovered the macrovariables and the probabilistic structure of the system. The *histograms* show the counts of microvariable states in each cell of the recovered causal partition. For example, $C = 1$ corresponds to the 'first state of the macrovariable cause'. Its histogram—full mass in the first out of four possible *bars*—shows that this state contains almost exclusively images that belong to the ground-truth first causal class (that is, images with no bars)

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.

3.2 Smoothness assumptions during learning

While the theory of CFL assumes a discrete macrovariable structure, Algorithm 1 makes the contrary assumption. Consider $p(eda|hue)$ evaluated on a fixed set of h samples as a vector-valued function of hue. Figure 5a 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. 5b varies continuously with hue. We chose to learn a continuous density mainly because there are good algorithms for learning continuous conditionals (described in Sect. 2.1). As Fig. 5b 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 1994; 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.

3.3 Choosing the number of states

In Fig. 6, 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 unknown cardinalities must also be discovered. A solution we propose is to run Algorithm 1 with N_h and N_e (the target number of observational classes for $\mathcal{H}$ and $\mathcal{E}$) 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 learning and clustering algorithm this should always be true.

Figure 9a 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 $\mathcal{H}$ into six groups, which are close to a refinement of the true observational partition. The pink group (spanning about $hue \in (160, 190)$ crosses 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 $\mathcal{H}$, 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 $P(E'|h'_i)$ should be similar to $P(E'|h'_j)$, where E' is (the refinement of) the observational partition on $\mathcal{E}$. In Fig. 9b, the i-th column corresponds to the empirical $P(E'|h'_i)$ where E' and H' are the six-state observational variables. Fig. 9c shows the result of merging these E' states whose corresponding columns in Fig. 9b have earth mover's distance (Levina and Bickel 2001) smaller than 0.2. Due to sampling errors, it deviates from the ground-truth slightly, but contains four states as expected.

Figure 9d–f shows the merging process for $\mathcal{E}$. The end result is almost exactly the ground-truth. The merging process for $\mathcal{E}$ requires a slight modification, since one cannot simply merge e'_i and e'_j when $P(e'_i|h') = P(e'_j|h')$ for any h'. To see why, consider clusters e'_0 and e'_1 in Fig. 9d (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 $[P(e'_0|h'_0), \ldots, P(e'_0|h'_5)]$ is a *scaled version* of $[P(e'_1|h'_0), \ldots, P(e'_1|h'_5)]$ (see the first and second *rows* in Fig. 9b). 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.

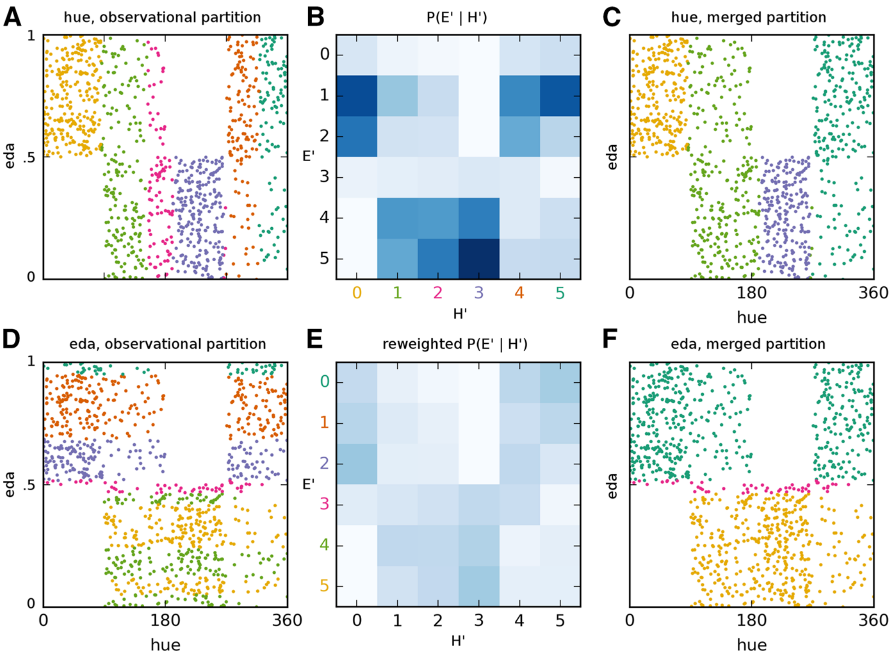


Fig. 9 Observational partition learning—unknown number of states. **a** We clustered hue w.r.t. $p(eda|hue)$ into six (as opposed to the ground-truth four) clusters $H' = 0, \ldots, 5$. With enough samples and good density learning, the overclustering should *refine* the true observational partition. We repeated the procedure in $\mathcal{E}$, clustering reaction states into six (instead of two) clusters E'. **b** Computing empirical $P(E'|H')$ 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. 6). 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 $P(E'|H')$ (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*)

4 Discussion

Causal feature learning can learn high-level causal knowledge from low-level data in an automatic, unbiased manner. We provided the necessary theoretical framework and demonstrated, on an easy-to-visualize example, how modern machine learning techniques can put the theory into practice. We pointed to applications of CFL to complex data where the method successfully recovered a known ground-truth. Finally, we discussed in detail the assumptions that CFL makes on the domain of application.

4.1 Related work

CFL is directly inspired by the theory of computational mechanics (Shalizi 2001; Shalizi and Crutchfield 2001; Shalizi and Moore 2003). Computational mechanics defines macrovariable states in time series in terms of equivalences of conditional probabilities. However, in computational mechanics macrovariables are meant to ‘summarize’ the phenomena rather than to support causal reasoning. The interventional/observational distinction, central to most analyses of causation, is also lost in heuristic approaches such as that of Hoel et al. (2013). We have adapted and developed the underlying theory to capture the more general causal setting.

Psychometric models (see Reise et al. 2010 for a good introduction) have a similar structure to the one presented here (see Fig. 10). There is, however, one crucial difference. In psychometric models, it is assumed that there are unobserved variables (such as IQ or some emotional or cognitive state; the dashed variables in Fig. 10) that stand in some unknown causal relation to one another. The only access the researcher has to these unobserved variables is through a set of measurement variables (the X_i), e.g., through questionnaire in which each question is supposed to provide an indication of the state of (one or more of) the unobserved psychological variable(s).

As the causal representations already indicate, the measurements are generally considered to be causal *effects* of the unobserved variables, in contrast to the *constitutive relation* we have considered in this article between the macro- and microvariables. That is, it is generally thought that specific answers to a questionnaire are *causal consequences* of IQ, rather than that having a high IQ *constitutes* answering a question in a particular way. In contrast, temperature is thought to be constituted by the mean kinetic energy of a particle ensemble, it is not a consequence of it. But as the example with IQ already suggests, the boundary between causal effects and constitutive relationships in the psychological setting may not be always as sharp. It suffices then to say here that the way we have defined interventions in this article requires that the relationship between the macro- and the microvariables is constitutive, since we define an intervention on a macrovariable as

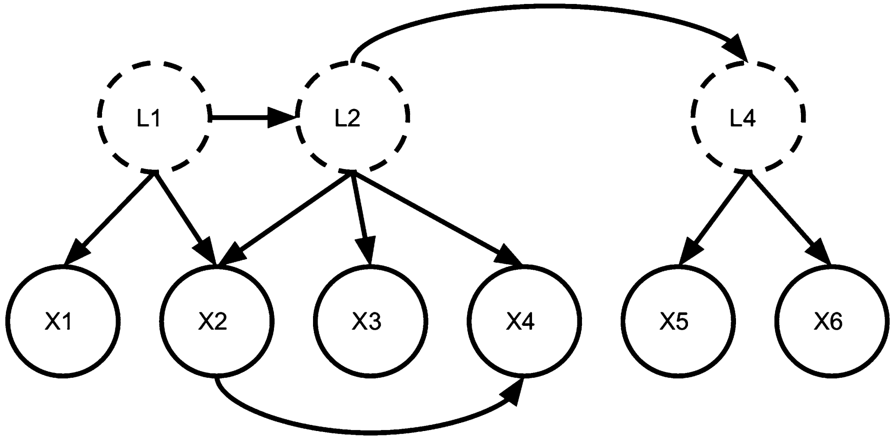


Fig. 10 Example of a psychometric model. Unobserved cognitive variables L_i stand in some unknown causal relation (Reise et al. 2010; Silva et al. 2006). The researcher can only learn about these causal relations via an analysis of the observed measurement variables X_i, which are thought to be causal effects of one or more of the unobserved variables. In general, the relation between the L_i and X_i is not thought to be constitutive, in contrast to the case of macro- and microvariables considered in this article

a class of interventions on the microvariable space. If the relations between unobserved variable and measurement variable in a psychometric system are causal rather than constitutive, our learning tools may nevertheless prove useful. It remains an area of application that we have yet to explore.

Work on learning causal graphs over high-dimensional variables (Entner and Hoyer 2012; Parviainen and Kaski 2015) is orthogonal to ours. The general idea of these approaches is to extend standard causal discovery methods (Spirtes et al. 2000) to the high-dimensional case. If the high-dimensional variables are interpreted as microvariables, CFL could be applied "on top" of each causal link in such a graph to compress it to retain only causal information.

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 scientific discovery. 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.

4.2 Open problems

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

1. **Learning without experimentation.** Currently, 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).
2. **Continuous macrovariables.** Section 3 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—appears nontrivial.
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. 2012; 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 system discussed in Chalupka et al. (2016a) (see also Sect. 3), the system definitely experiences feedback (over time). While cyclicity may not break the CCT or our algorithms, the proof of such a claim has not been completed.

4.3 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.

Nontrivial 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 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 microstates within each of its macrostates that have very different causal consequences from the other microstates within the same macrostate. 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 microstates.

5 Mathematical appendix

In this section, we prove the Causal Coarsening Theorem. To (significantly) simplify the notation and to easier relate this proof to earlier ones, we use the same variable names we used in Chalupka et al. (2016b):

1. The microvariable cause is $I \in \mathcal{I}$,

2. the microvariable effect is $J \in \mathcal{J}$,
3. the confounder is H,

These correspond to *hue*, *eda*, *lat*, respectively. Before we prove the CCT, we prove a simpler Lemma.

Lemma 8 *Let $S_{P(H)}$ denote the simplex of multinomial distributions over the values of H. For fixed $P(J|H,I)$, the subset of $S_{P(H)}$ for which Π_c is not equal to $\Pi_{P(J|H,I)}(\mathcal{I})$ 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 $h \in H$

$$P(J|H=h, i_1) \neq P(T|H=h, i_2), \text{and} \tag{7}$$

$$P(J|man(i_1)) = P(T|man(i_2)), \tag{8}$$

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

Equation 8 is equivalent to $\sum_h P(H=h)[P(J|H=h,i_1) - P(J|H=h,i_2)] = 0$. Since this is a linear constraint on $S_{P(H)}$, 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 $P(J|H=h_1,i_1) - P(J|H=h_1,i_2) > 0$ and for some h_2 $P(J|H=h_2,i_1) - P(J|H=h_2,i_2) < 0$. 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. (8) does not hold. □

Causal coarsening theorem *Among all the joint distributions P(J, H, I) consider the subset that induces any fixed causal partition $\Pi_c(\mathcal{I})$ and a fixed confounding partition $\Pi_{P(I|H)}(\mathcal{I})$. The following two statements hold within this subset:*

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

Proof 1. We first prove that $\Pi_c(\mathcal{I})$ is almost always a coarsening of $\Pi_o(\mathcal{I})$.

(i) 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 $P(J,I,H) = P(J|H,I)P(I|H)P(H)$. $P(J|H,I)$ can be parametrized by $(N-1) \times K \times M$ parameters, $P(I|H)$ by $(M-1) \times K$ parameters, and $P(H)$ by $K-1$ parameters, all of which are independent. Call the parameters, respectively,

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

We will denote parameter vectors as

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

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution *P*(*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 $P(J|H,I) \times P(I|H)$ 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 does not directly fix Π_c. But by Lemma 8, 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 $i_1, i_2 \in \mathcal{I}$, we have

$$O(i_1) = O(i_2) \quad \Rightarrow \quad C(i_1) = C(i_2). \tag{9}$$

That is, it is not the case that two members of $\mathcal{I}$ are observationally equivalent but have causally different effects. We will show that for any i_1, i_2 pair, the constraint holds for almost any γ. 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 $P'[\gamma; \alpha, \beta]$. By definition of the observational class, the equality implies that *for any j*,

$$\frac{1}{P(i_1)}\sum_h \alpha_{j,h,i_1}\beta_{i_1,h}\gamma_h,$$
$$=\frac{1}{P(i_2)}\sum_h \alpha_{j,h,i_2}\beta_{i_2,h}\gamma_h.$$

Picking an arbitrary j and expanding in terms of α, β, γ, we have

$$\begin{aligned} &O(i_1) = O(i_2) \\ &\Leftrightarrow \sum_{h_k,h_l} \gamma_{h_k}\gamma_{h_l}[\beta_{i_2,h_k}\beta_{i_1,h_l}\alpha_{j,h_l,i_1} - \beta_{i_1,h_k}\beta_{i_2,h_l}\alpha_{j,h_l,i_2}] = 0. \end{aligned} \tag{10}$$

We have thus shown that, for fixed α, β and i_1, i_2, the violation of the causal coarsening constraint (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. (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. (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 $[\beta_{i_2,h_k}\beta_{i_1,h_l}\alpha_{j,h_l,i_1} - \beta_{i_1,h_k}\beta_{i_2,h_l}\alpha_{j,h_l,i_2}]$ 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. (10) does not hold. That is, for this γ we have $O(i_1) \neq O(i_2)$[3].

(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 (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

[3] 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.

admittable 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 admittable 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

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

By basic properties of positive measures, we have

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

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 nonempty 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 admittable and evaluates to 0 otherwise. We have

$$\begin{aligned} \int_{P[\alpha,\beta,\gamma]_A} \hat{z} \, \mathrm{d}\mu &= \int_{\mathcal{A}\times\mathcal{B}\times\mathcal{G}} \hat{z}(\alpha, \beta, \gamma) I_A(\alpha, \beta) \, \mathrm{d}(\gamma, \beta, \alpha) \\ &= \int_{\mathcal{A}\times\mathcal{B}} \int_{\mathcal{G}} \hat{z}(\alpha, \beta, \gamma) \, \mathrm{d}(\gamma) \, I_{\Pi_c}(\beta, \alpha) \, \mathrm{d}(\beta, \alpha) \\ &= \int_{\mathcal{A}\times\mathcal{B}} \mu(z[\alpha, \beta]) \, I_A(\alpha, \beta) \, \mathrm{d}(\beta, \alpha) \\ &= \int_{\mathcal{A}\times\mathcal{B}} 0 \, I_A(\alpha, \beta) \, \mathrm{d}(\beta, \alpha) \\ &= 0. \end{aligned} \tag{11}$$

Equation (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 (9) is measure zero.

(2) We use the same strategy as above, with some differences in the details of the algebra of the polynomial constraint in step (iii):

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

$$O(j_1) = O(j_2) \quad \Rightarrow \quad C(j_1) = C(j_2). \tag{12}$$

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 $O(j_1) = O(j_2)$ 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 $P'[\gamma; \alpha, \beta]$:

$$\sum_h \gamma_h(\alpha_{j_1,h,i}\beta_{i,h} - \alpha_{j_2,h,i}\beta_{i,h}) = 0. \tag{13}$$

We have thus shown that for fixed α, β and j_1, j_2, the violation of the causal coarsening constraint (12) 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. (13) 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. (13) 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. (13) does not hold. That is, for this γ we have $O(j_1) \neq O(j_2)$[4]. □

Acknowledgements We thank an anonymous reviewer for pointing out an error in our original theorem. This work was supported by NSF Award #1564330.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Bishop CM (1994) Mixture density networks. Technical report

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 273–304

[4] 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.

Chalupka K, Perona P, Eberhardt F (2015) Visual causal feature learning. In: Proceedings of the thirty-first conference on uncertainty in artificial intelligence. AUAI Press, Corvallis, pp 181–190

Chalupka K, Bischoff T, Perona P, Eberhardt F (2016a) Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In: Proceedings of the thirty-second conference on uncertainty in artificial intelligence

Chalupka K, Perona P, Eberhardt F (2016b) Multi-level cause-effect systems. In: 19th international conference on artificial intelligence and statistics (AISTATS)

Chickering DM (2002) Learning equivalence classes of bayesian-network structures. J Mach Learn Res 2:445–498

Claassen T, Heskes T (2012) A bayesian approach to constraint based causal inference. In: Proceedings of UAI. AUAI Press, Corvallis, pp 207–216

Entner Doris, Hoyer Patrik O (2012) Estimating a causal order among groups of variables in linear models. Artif Neural Netw Mach Learn-ICANN 2012:84–91

Hoel Erik P, Albantakis L, Tononi G, Albantakis GT (2013) Quantifying causal emergence shows that macro can beat micro. Proc Natl Acad Sci 110(49):19790–19795

Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. In: Advances in neural information processing systems, pp 689–696

Hyttinen A, Eberhardt F, Hoyer PO (2012) Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. arXiv:1210.4879

Hyttinen A, Frederick E, Järvisalo M (2014) Conflict resolution with answer set programming. In: Proceedings of UAI, constraint-based causal discovery

Jacobs KW, Hustmyer FE (1974) Effects of four psychological primary colors on gsr, heart rate and respiration rate. Percept Motor Skills 38(3):763–766

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KO (eds). Advances in neural information processing systems, vol 25, pp 1097–1105

Lacerda G, Spirtes PL, Ramsey J, Hoyer PO (2012) Discovering cyclic causal models by independent components analysis. arXiv:1206.3273

Levina E, Bickel P (2001) The earth mover's distance is the mallows distance: some insights from statistics. In: Eighth IEEE international conference on computer vision, 2001. ICCV 2001. Proceedings, vol 2. IEEE, New York, pp 251–256

Mooij JM, Janzing D, Heskes T, Schölkopf B (2011) On causal discovery with cyclic additive noise models. In: Advances in neural information processing systems, pp 639–647

Okamoto M (1973) Distinctness of the eigenvalues of a quadratic form in a multivariate sample. Ann Stat 1(4):763–765

Parviainen P, Kaski S (2015) Bayesian networks for variable groups. arXiv:1508.07753

Pearl J (2000) Causality: models. Reasoning and inference. Cambridge University Press, Cambridge

Pearl J (2010) An introduction to causal inference. Int J Biostat 6(2)

Reise SP, Moore TM, Haviland MG (2010) Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. J Pers Assessm 92(6):544–559

Richardson T (1996) A discovery algorithm for directed cyclic graphs. In: Proceedings of the twelfth international conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., USA, pp 454–461

Rumelhart DE, Hinton GE, Williams RJ (1985) Learning internal representations by error propagation. Technical report, No. ICS-8506. California University of San Diego La Jolla Institute for Cognitive Science

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

Shalizi CR, Crutchfield JP (2001) Computational mechanics: pattern and prediction, structure and simplicity. J Stat Phys 104(3–4):817–879

Shalizi CR, Moore C (2003) What is a macrostate? Subjective observations and objective dynamics. arXiv:cond-mat/0303625

Shimizu Shohei, Hoyer Patrik O, Hyvärinen Aapo, Kerminen Antti (2006) A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res 7:2003–2030

Silander T, Myllymäki P (2006) A simple approach for finding the globally optimal bayesian network structure. In: Proc UAI. AUAI Press, Oregon, pp 445–452

Silva R, Scheines R, Glymour C, Spirtes P (2006) Learning the structure of linear latent variable models. J Mach Learn Res 7:191–246

Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems, pp 2951–2959

Spirtes Peter, Scheines Richard (2004) Causal inference of ambiguous manipulations. Philos Sci 71(5):833–845

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search, 2nd edn. Massachusetts Institute of Technology, Massachusetts

Srinivas N, Krause A, Seeger M, Kakade SM (2010) Gaussian process optimization in the bandit setting: no regret and experimental design. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 1015–1022

Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. J Mach Learn Res 2:45–66

Tsao Doris Y, Freiwald Winrich A, Tootell RBH, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. Science 311(5761):670–674