Chapter 6

Beyond cause-effect pairs

Frederick Eberhardt

Abstract The cause-effect pair challenges focused on the development of inference methods to determine the causal relation between two variables. It is natural to then ask how such methods could generalize beyond the two variable case to settings that either involve more variables – such as is the case in graph learning – or to settings where the relationship between the candidate variables does not fall into one of the classes defined by the challenges. This chapter explores the extension of the proposed methods to such cases. It comes to the conclusion that such extensions are not likely to naturally evolve from the approaches that won the pair challenge.

Key words: graph learning, structure learning, confounding, feedback cycles, variable construction

6.1 Introduction

This volume has focused on the identification of cause-effect pairs. The original cause-effect pair challenge at the NIPS 2008 Causality workshop considered the specific case of determining the edge orientation between two causal variables: $X \rightarrow Y$ vs. $X \leftarrow Y$ (that is, the blue partition in Figure 6.1). The unconnected case $X \quad Y$ (case (c)) can be easily excluded with a suitable independence test. The NIPS 2013 challenge extended this setting to three classes, $X \rightarrow Y$, $X \leftarrow Y$ or the null class, which, for the purposes of that challenge, consisted of the unconnected case $X \quad Y$ or the case of pure confounding $X \leftrightarrow Y$ (this notation is used as shorthand for $X \leftarrow H \rightarrow Y$, where H is unobserved, but where neither X or Y cause each other).

Frederick Eberhardt
Caltech, USA e-mail: fde@caltech.edu

That is, it combined cases (c) and (d) in Figure 6.1 into one class, which had to be distinguished from (a) and (c), resulting in the red partition.

As Dominik Janzing already noted in Chapter 1, the identification of the causal relation between two variables can, of course, be more complex than just determining whether one causes the other. The variables that cause each other may in addition be confounded (cases (e) and (f)), they might stand in a feedback relation, where each causes the other ($X \leftrightarrows Y$), or a combination of feedback and confounding. Moreover, dependencies between the two variables may arise for other reasons, such as sample selection bias, ill-defined variables or other non-causal relations, such as e.g. logical relations.

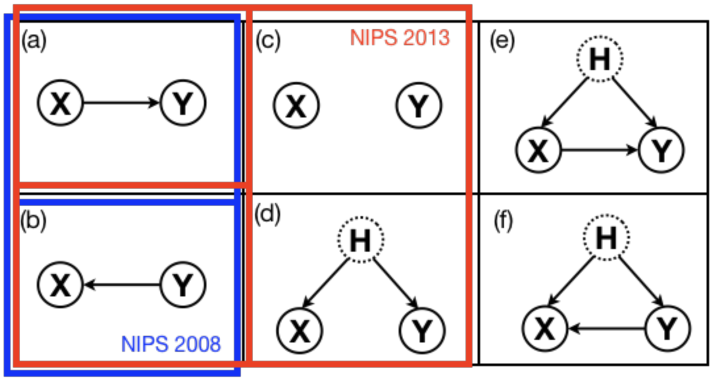


Fig. 6.1: The Cause-Effect Pair challenge and its extensions. The original 2008 proposal focused on the distinction between X causing Y and Y causing X, cases (a) and (b), the blue partition. The NIPS 2013 challenge had three output classes, (a) vs. (b) vs. {(c), (d)}, i.e. the red partition, H is assumed to be unobserved. The six depicted causal structures linking X and Y were distinguished by the classification method developed in Chalupka et al. [2016b], but they assumed that X, Y and H are all discrete. Of course, there are other possible causal structures among the two variables, such as a feedback loop (alone or in combination with unobserved confounding), and observed dependencies between two variables may have other non-causal sources, such as sample selection bias or ill-defined variables.

This chapter reverses the specific focus on determining the cause among a pair of variables by connecting the ideas that have come out of the NIPS 2013 challenge to the broader question of how to learn causal graph structures over multiple variables, and by placing the cause-effect pair challenge in the context of a variety of other causal discovery challenges.

In many ways, the generalization to learning the causal structure over a set of variables is anachronistic, as the development of methods for learning causal *graphs* from observational data preceded most of the methods for causal *pairs* discussed in this volume. Prior to the focus on causal pairs, causal discovery algorithms primarily used the independence structure over a set of variables to infer something about the underlying causal structure (see, e.g. the discovery algorithms discussed in Spirtes et al. [2000]). The independence structure often (though not always!) underdetermines the orientations of edges and consequently two or more different causal structures cannot be distinguished. These are then said to be Markov equivalent, as are, in particular, $X \rightarrow Y$ and $X \leftarrow Y$. Consequently, a significant motivation for studying causal pairs came from a desire to improve the number of edges that could be oriented for a discovery algorithm that outputs equivalence classes of causal graphs. This challenge then resulted in several methods to solve the Cause-Effect-Pair problem, distinguishing whether $X \rightarrow Y$, $X \leftarrow Y$ or neither. However, as is evident from the possible output classes of the NIPS 2013 challenge, the solution proposals were developed under the explicit assumption that confounding does not co-occur with a direct causal relation between the variables: Either one of the variables causes the other, or (exclusive or!) there is confounding or (inclusive or!) independence. There was no need to be sensitive to the distinction between cases (a) and (e) or (b) and (f) in Figure 6.1. My understanding is that training data was only generated from these three classes. Consequently, it is by no means obvious that methods that are successful at solving the cause-effect pair challenge would similarly work as a post-processing step to orient causal edges in a structure learning graph that returns unoriented edges that may, in addition, be confounded.

This chapter then aims to achieve the following: In Section 6.2 we explore whether and how the winning methods from the NIPS 2013 cause-effect pair challenge could be generalized to consider more classes. Section 6.3 briefly discusses the existing approaches (and their challenges) that apply to the cause-effect pair challenge, but have in fact been generalized to graph learning methods. Section 6.4 considers the generalization of the Cause-Effect Pair challenge in a different direction – not as an expansion to more variables, but posing it as a challenge in new types of discovery scenarios over a pair of variables. And finally, Section 6.5 considers an inverted version of the cause-effect pair challenge: how can one construct a cause-effect pair if in fact one has a large number of individual variables that one would like to aggregate into a pair of cause and effect variables?

6.2 How to extend the winning methods beyond the cause-effect pair case?

Figure 6.1 illustrates in red the partition of the output that was considered in the NIPS 2013 challenge. A natural initial extension is to ask whether there are straightforward adaptations of the methods that succeeded at that challenge to address, for example, the distinction among all six possible causal structures. That is, in particular, are these methods extendable to distinguish cases that have unobserved confounding in combination with a causal relation? I omit the case involving feedback here primarily because the presence of feedback raises separate questions about how exactly the data was sampled (e.g. as a time series or in equilibrium?) and what exactly the feedback graph means. Chapter 5 considers these issues in more detail.

6.2.1 Classification-based causal discovery

Remarkably, all the winning or highly ranked methods of the actual competition discussed in this volume treated the challenge as a pure classification task [Moitinho de Almeida, Fonollosa, 2016, Samothrakis et al., Minnaert]. In part, as Janzing describes in Chapter 1, converting the causal discovery problem to a classification task was a deliberate aim of the challenge, since it vastly simplified the comparative evaluation between methods.

The highly ranked approaches applied relatively standard machine learning methods of the time to generate and select between 20 and 20,000 features using the training data, which were then in turn applied to classifying the test data. The winning method [Moitinho de Almeida] deliberately did not select features associated with well-understood justifications for their relevance to causal inference, but instead used feature patterns, which generated features from a set of simple measures of the data (that is, features were generated from e.g. means, correlations, various loss functions etc.). In contrast, the second placed method [Fonollosa, 2016] at least used some features that were based on the assumption that when in fact $X \to Y$, then the conditional distribution $P(Y|X)$ is simpler than the conditional distribution $P(X|Y)$. Such an assumption derives from the "independent mechanisms" assumption discussed in Chapter 1. This approach is echoed, though with different features, in a subsequent paper [Hernandez-Lobato et al., 2016].

The extension of these pure classification-based approaches to the problem when the pair of causal variables may (also) be confounded is conceptually trivial. It merely extends the classification problem from two (or three) classes (the blue or red partition in Figure 6.1) to however many more constellations of two variables one intends to distinguish. One would expect that with more classes, more features might have to be generated, but to the extent that there is any marker in the data that provides a basis for distinctions between the underlying causal structures, there will be suitable classifiers (if not the present ones) that distinguish the classes.

This, however, is the crux of this approach: We know from basic results about the identifiability of causal models that for linear Gaussian and multinomial parameterizations, the underlying causal structure remains underdetermined by the Markov equivalence class, i.e. by the set of causal models that share the same independence structure. Under the assumption that the causal model is acyclic and that there are no unmeasured common causes, Geiger and Pearl [1988] and Meek [1995] proved the completeness of independence based methods for continuous and discrete causal models, respectively:

Theorem 6.1 (Markov completeness). *For linear Gaussian and multinomial causal relations, an algorithm that identifies the set of causal graphs with the same independence structure is complete.*

That is, if the value of each variable in the causal graph is determined by a linear function of its parents in the graph plus a Gaussian error term, or if the model is multinomial, then the independence structure contains all the information about the causal structure that there is. So, in particular, for these parameterizations, $X \to Y$ and $X \leftarrow Y$ cannot be distinguished in principle. In fact, this underdetermination is worsened if there can be unmeasured common causes. So for a causal pair, no matter whether in fact there is an edge one way or the other, or confounding, within the class of linear Gaussian or multinomial models, any of those structures can be fit to the data. As we will see in Section 6.3 this is not true for other specific model classes (such as e.g. for additive noise models). For those model classes, the underlying causal model can be uniquely (or close to uniquely) identified. For many other classes of models the identifiability is simply unknown.

What do these types of identifiability results imply for classification based causal discovery methods? — Most obviously, unless some additional assumption about, for example, the parameterization of the causal relations is made, such classifiers will exhibit an irreducible baseline error, a misclassification error that cannot be avoided. In the strict sense of the statistical term then, these methods cannot be consistent, since the models are not uniquely identifiable.

In classification tasks in machine learning, such a baseline error is standard and widely accepted as an unavoidable difficulty of the problem to be solved – after all, not all images of dogs can be distinguished from all images of cats. But for causal discovery, the challenge for any extension along these lines is that if there is no theoretical justification for the features being used, it remains unclear what the magnitude of the irreducible error of the method is. The estimation of such an error hinges on the appropriateness of the assumption that the training and test data accurately represent the manifold of causal models that the algorithm is subsequently applied to. Of course, the appropriateness of the training and test data is an issue every classification algorithm faces for any domain (e.g. are 20% of the images I will have to classify really going to be dogs, as my training/test datasets suggest?). But in the causal case this problem is exacerbated due to the dearth of real data for which the true causal model is known. For dogs and cats we have a sense of the manifold that any image of them will lie on. The manifold of causal models we encounter in our data is much more elusive, not least because for many datasets we have no idea what the ground truth is. Unlike for images of cats and dogs where a simple inspection of the image can generally determine the label, the true causal structure is not written into the data in any obvious way. This is why the Tübingen causal pairs data set [Mooij et al., 2016] proves so useful – it starts to address the question of what real data looks like in cases where we know (or have good reason to think we know) the causal ground truth.

Thus, while traditional causal discovery methods use background assumptions about the underlying causal model (say, linearity, Gaussianity etc.) as basis for the identifiability results, classification-based approaches to causal discovery have to replace these with an assumption that the training/test data is appropriate for the actual application of the algorithm and have to hope that the classes (the different underlying causal models) can indeed be distinguished (with a low misclassification error).

The assumption of independent mechanisms (discussed in Chapter 1) provides a basis for this latter hope, as it offers a reason to expect detectable features in the data that indicate what the underlying causal model is. These features might track the complexities of the conditional distributions or identify particular independencies between residuals in the data. Nevertheless, even under this assumption it remains quite unclear how to obtain well-justified estimates of the inevitable misclassification error, since the notion of independence in the assumption of independent mechanisms, or the notion of complexity for the conditionals, is only understood either in very abstract form (in terms of Kolmogorov complexity) or for very restricted settings (with a specific computable measure). Chapter 1 discusses some of these issues and provides useful references.

6.2.2 Extensions

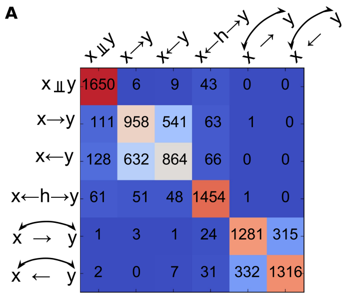
In light of the previous considerations, what can we say about the extension of classification based causal discovery algorithms beyond the cause-effect pair challenge?

Obviously, one can build a classifier with more classes and generate training and test data for which one knows that the classes are identifiable, thereby guaranteeing in principle a zero missclassification error. Alternatively, one could train a classification algorithm on a dataset with more varied ground truth causal structures and hope for the best that (i) they are distinguishable, and (ii) that the examples are representative of the domain of application. This latter approach is essentially the route Lopez-Paz et al. [2015] took when they trained a neural net on the features of a kernel mean embedding of the distribution of a pair of variables to address the causal pair task. They suggest that their approach can be extended to the more general case of also distinguishing confounding by simply adding more classes, but they do not actually show any results for the confounded case.

In Chalupka et al. [2016b] we took a somewhat different approach in trying to classifying the six cases shown in Figure 6.1, for discrete data. Obviously, for general multinomial distributions, these models are not identifiable (except for case (c)). Motivated by the assumption of "independent mechanisms", we made the following assumptions:

1. $P(\text{effect} \mid \text{cause}) \perp\!\!\!\perp P(\text{cause})$
2. $P(\text{effect} \mid \text{cause} = c)$ is sampled from an uninformative hyperprior for each c.
3. $P(\text{cause})$ is sampled from an uninformative hyperprior

In fact, (2) is often not taken to be part of the independent mechanisms assumption, instead allowing for additional structure within the generating conditional distribution. In the case of finite discrete variables, the uninformative hyperprior is given by the Dirichlet distribution with all parameters set to 1. Under these generating assumptions the causal models (a) vs. (b) in Figure 6.1 are not strictly identifiable, but we were able to derive an analytic classification boundary with a well-defined irreducible misclassification error. Without putting significant restrictions on the nature of the confounder, we could not derive analytical classification boundaries for all 6 cases in Figure 6.1, but instead used a neural net trained on distributions generated under the assumptions given above.



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

Fig. 6.2: Confusion matrix for the classification method developed in Chalupka et al. [2016b] (their figure) to address the causal discovery task of distinguishing the 6 causal structures in Figure 6.1. The test set contained 10,000 distributions, with all the classes sampled with equal probability. For the confusion matrix presented here X and Y are binary discrete variables. There is significantly less confusion when the cardinality of the variables is increased. When each variable has more than 10 states, there is hardly any misclassification error – see the reference for details.

The approach we took is intermediate between the case of making sufficient background assumptions that guarantee full identifiability, and one where the irreducible misclassification error is completely unknown. We leveraged the fact that the training and test data is known to contain features that distinguish the classes, even if a baseline error remains. As can be seen in Figure 6.2, the biggest "confusion" for the classifier arises, unsurprisingly, in distinguishing the orientation. But as noted above, while we can estimate the inevitable misclassification error in this way for the general problem of classifying the six causal structures in Figure 6.1, it remains unclear how reasonable these data generating assumptions are for any domain of application.

The only other approach that I am aware of (thanks to an anonymous reviewer) that attempts to tackle the full set of structures shown in Figure 6.1 is in Janzing and Schoelkopf [2017]. Here the authors take X, Y and H to be high-dimensional continuous variables, and apply spectral analysis. But rather than a lack of effort, I suspect that the dearth of attempts to extend the classification-based approach likely indicates the following realization: While these classification approaches are straightforwardly generalizable to include more possibilities of the causal relation between two variables (confounding, feedback, selection bias etc.), we generally have little insight about the manifold that causal structures in any domain may be described by. Consequently, our training data for these causal classification algorithms is a somewhat arbitrary guess about the distribution of causal models we expect to encounter. In contrast, in the case of the standard domains of application of classification algorithms, such as image or text classification, we have a relatively good understanding of the manifold that our samples come from, and we can more easily explore the classification boundaries actively.

On the other hand, if there is a justification for the features that the classification is based on – generally these are motivated by some version of the independent mechanisms assumption – then for simple cases the irreducible misclassification error can be quantified (such as in Hernandez-Lobato et al. [2016] and Chalupka et al. [2016b]), but these analytic results are often not easily generalized beyond the very simple cases.

6.3 Established Extensions to Graph Search

The previous section considered the generalization of the cause-effect pair challenge to more than just the two possible relationships between two variables $X \rightarrow Y$ and $X \leftarrow Y$. One may think that the search for causal graphs, i.e. a structure over a set of variables, is then just a repeated classification problem of the relationship between any two variables in the graph. I am not aware of any such approach using the types of feature-based classification algorithms suggested by the winning methods of the NIPS challenge, but the recent publication of Goudet et al. [2018] goes in this direction.

6.3.1 Additive Noise Models

There are a variety of methods based on the Additive Noise Model (see Peters et al. [2017] for an overview of ANMs) that both apply to the cause-effect pair challenge *and* have been extended to the general graph search. In fact, in the case of the "LiNGAM" discovery method for Linear non-Gaussian Models (a subclass of the additive noise models), the method was developed for graph search from the outset [Shimizu et al., 2006].

One way of looking at the ANM-based methods is to return to Geiger & Pearl's limiting result (Theorem 6.1 above) that indicates that for *linear Gaussian* models the Markov equivalence class of the true model is the best one can hope for from a causal discovery algorithm. This limiting result says nothing about the case when the causal relations are *non*-linear or *non*-Gaussian.

Shimizu et al. [2006] considered precisely one of these cases and showed that if the functional relation still remains linear, but the error terms are anything but Gaussian (LiNGAM), then the causal graph is uniquely identifiable. That is, the causal graph is identifiable if for each variable $y \in \mathbf{V}$, y is given by

$$y = \sum_{x_i \in pa(y)} a_i x_i + \varepsilon_y \quad \text{with} \quad \varepsilon_y \sim NonGauss(\theta),$$

where $pa(y)$ are the parents of y in the graph and $NonGauss(\theta)$ is some non-degenerate distribution that is not Gaussian. So, in particular, $X \rightarrow Y$ and $X \leftarrow Y$ can be distinguished in this model class.

With slightly weaker identifiability results, LiNGAM has been extended to acyclic causal structures with latent variables [Hoyer et al., 2008b] and to causal structures with cycles (but without latent confounding) [Lacerda et al., 2008]. So in many ways the LiNGAM method is precisely what one would have hoped to discover in the cause-effect pair challenge, since it addresses that particular challenge, but could also be extended usefully to more general scenarios. But it preceded the challenge.

The identifiability results for the non-linear ANMs take the other alternative of avoiding Theorem 6.1 by exploring the role of the function. They assume that for each variable $y \in \mathbf{V}$, y is given by

$$y = f_y(pa(y)) + \varepsilon_y$$

where $f_y(.)$ is a continuous function and ε_y is an additive error term with some positive distribution. Hoyer et al. [2008a] then show that in general (i.e. except for very special cases, that include the linear Gaussian case) $X \rightarrow Y$ can be distinguished from $X \leftarrow Y$. Peters et al. [2014] extended the identifiability result to graph structures.

It is worth noting an important aspect to the non-linear ANMs that is rarely discussed in any detail: Unlike the LiNGAM model or traditional linear Gaussian or multinomial causal models, non-linear ANMs are not (in general) closed under marginalization. This effectively makes them inapplicable to scenarios with unmeasured confounding, or, for that matter, any unobserved variable. If the true model is a non-linear ANM, the marginalized observable model is, in general, not. This feature then places a strong demand on having exactly the right set of variables: if the world is indeed well-described by non-linear ANMs, then there is exactly one level of correct causal description. Note, for example that the approach to confounder detection considered by Janzing et al. [2009] only considers non-linear ANMs where there is no causal connection between the observed variables, and therefore the marginalization problem does not arise. Of course, one might take the unique level of causal description implied by non-linear ANMs as a virtue, useful to detect truly direct causal relations, rather than as problem of this model class. In that case, it would be of interest to develop an argument why we should expect the world to be structured in this way.

While the LiNGAM methods in their original incarnation identified the graph directly on the basis of matrix operations on the data, the extensions of non-linear ANMs to identifying the graph are really just an iterative procedure of applying the pairwise result, taking into account that any edge might now also be subject to confounding from a variable higher up in the graph structure. In the Causal Additive Model (CAM) approach of Bühlmann et al. [2014], which considers non-linear ANMs with Gaussian errors, the search method first determines the (partial) order of the variables in the causal graph using maximum likelihood estimation, and then subsequently the specific parents of each variable are determined using sparse regression.

Other approaches based on the LiNGAM model echo this division of labor and also outsource the search for graph structure to methods that use the independence structure (such as the PC algorithm) and then only attempt to resolve the orientations. For example, Zhang and Chan [2006] and Zhang and Hyvärinen [2009] consider the case where the data is generated by a linear non-Gaussian model and then subject to a non-linear invertible transformation. This post-nonlinear model is in general identifiable. The generalization from the pairwise case to a graph is simply done by using another method like the PC-algorithm to search for the adjacency structure among the variables and then applying the post-non-linear test to each undirected edge. A similar approach is taken by various other methods implemented in the Tetrad code package [TET].

While these identifiability results are remarkable and constitute significant theoretical advances, the success of these methods at orienting edges in practice remains unclear. Perhaps the most thorough empirical investigation of the Additive Noise

Methods was done in Mooij et al. [2016] with a specific focus on the cause-effect pair challenge. The results were mixed, with the method described in Hoyer et al. [2008a] obtaining the best results. I am not aware of investigations that systematically considered graph search.

The LiNGAM methods and variations of them have been applied in a variety of real-world settings or realistic simulations and the authors report results on edge orientations that are consistent with background knowledge (see e.g. Shimizu and Bollen [2014] and Ramsey et al. [2011]). However, I am not aware of any application of these methods where an edge orientation that was not determinable from the independence structure, was subsequently confirmed, e.g. by experimentation. It remains an open question just how good these orientation methods are in practice.

6.4 More Causal Challenges for Pairs of Variables

Section 6.2 discussed the possibility of extending the methods developed for the cause-effect pair challenge to other causal scenarios among pairs of variables. Although we did not discuss cases of selection bias or feedback cycles in any detail, these are all cases that can be well described within the framework of causal graphical models. This section considers two cases of the search for cause-effect pairs that do not neatly fit this framework, but are still of significant interest to causal discovery.

6.4.1 Discovery of dynamical causal relations

Chapter 5 already discussed causal discovery in time series data. Time series data has the advantage of providing a time order over the samples and therefore somewhat restricts the possible causal influences among variables. But this time order usually comes in fixed discrete intervals and in general there is no reason to think that the measurement interval has anything to do with the speed of the causal process. As a result, even if the causal inference algorithm works well, the discovered causal effects should be thought of as causal effects relative to the particular sampling rate. Various attempts have then been made to determine what the actual causal process looks like if the time series subsamples the causal process, i.e. when the causal effects occur faster than the sampling rate [Gong et al., 2015, Hyttinen et al., 2017].

In the extreme of infinitesimal time delay between cause and effect, the system can be described as a dynamical causal process. Time is continuous, and – at least in principle – one could obtain measurements at any time granularity. However, the data is not independent and identically distributed (i.i.d.), there is no stationary distribution or any of the other niceties that come with (or are generally assumed for) standard causal (even time series) data sets. Nevertheless, the cause-effect pair challenge still remains: How can we learn the causal relation between X and Y (if any) when they both have continuous-in-time trajectories?

As with the standard cause-effect pair challenge, one might define the ground truth in terms of the results of interventions: X is a cause of Y if an intervention that sets X to a particular value results in a change in Y. But can such a causal relation be learned from just observing a suitably long trajectory? If so, how?

Despite the fact that dynamical models are ubiquitous in the natural sciences, there are only very few approaches to causal discovery in dynamical systems. Roy and Jantzen [2018] provide an explicitly causal treatment of the problem for the case of first order differential equation models. The challenge they pose can be easily stated: Suppose one has measurements in (continuous) time of two variables x and y, that may be unidirectionally coupled by a first order autonomous system, such as:

$$\begin{aligned}\dot{x} &= \alpha(x, \dot{y}, y)\\ \dot{y} &= \beta(y)\end{aligned}$$

Clearly, y has an influence on x but not vice-versa (hence, unidirectional coupling). Given measurements of x and y in continuous time, and the assumption that we are only considering first order autonomous systems (i.e. no unmeasured variables), how can one determine that $y \to x$ and not vice versa? The authors propose a method based on symmetry transformations and compare their approach to methods based on Transfer Entropy and the Convergent Cross Map. This opens the door to address much more general questions of causal discovery in dynamical systems.

6.4.2 Discovery of relational causes

The second setting also concerns a non-i.i.d. scenario for the data collection, but the issue is rather different from the case of dynamical systems. Consider a relational database that contains individuals that have properties, e.g. whether they smoke or not, and relations, e.g. which other individuals they are friends with. A causal question about an individul I that one may hope to address with such a database is: Is it the friends that I has that cause I to smoke, or does I's smoking cause I's friends to smoke, or is there a common cause of I's own and I's friends' smoking, for example, the friendship relation itself? There may of course be other, even causal, explanations, but we can start with these.

The challenge in addressing these types of relational causal discovery questions arises from the fact that in addition to the causal relations, there are logical relations between the individuals, in this case the friendship relation, that need to be taken into account. These logical relations can introduce dependencies in the data that are not in fact causal. Similar such constraints may arise from boundary conditions or conservation laws, such as that the total energy in a physical system is constant or that there are resource constraints in economic models.

With a few very notable exceptions [Schulte and Khosravi, 2012, Maier, 2014], this task of relational causal learning has been completely neglected, even though it might be one of the most important causal questions when it comes to social science and social network data. The most thorough investigation in this direction has been done in a variety of publications Maier, Marazopolou and Jensen (see e.g. Maier et al. [2013]. They have extended the PC algorithm to relational causal models. To my knowledge there has been no attempt to consider any of the insights from the cause-effect pair approaches in the relational setting.

6.5 Construction of cause effect pairs

Finally, I will invert the cause-effect pair challenge to ask how we obtain our cause and effect pair variables in the first place. This question is motivated by a concern that I think has been neglected in the discussion of causal models:

What makes a random variable causal?

The previous section already suggested that there can be logical relations among variables in addition to causal ones. It follows from the definition of a random variable that every function of a random variable is a random variable. But the same is not true in the same way for causal variables: We do not consider X and the variable Y that is *defined* as $Y = 2X$ as two distinct causal variables. We might consider them to be two descriptions of the same variable, or a translation of one another, but we do not have two separate causal variables. In addition to the features of ordinary random variables, causal variables play a role in supporting interventions and counterfactuals. One cannot intervene on X without intervening on Y (as defined above), nor are the counterfactuals between definitionally related variables analogous to the counterfactuals between causally related variables. These points are generally emphasized when structural equations are introduced to describe causal relations: The authors generally point out that these equations should be understood as *assignments*, often marked by the symbol ":=", rather than as mathematical equations where quantities can be exchanged between sides of the equal sign. The reason, though not always explicitly stated, is that the causal relations permit interventions and counterfactual statements that are, in general, not symmetric.

Obviously, any dependence between X and Y (as $2X$) should be attributed to their mathematical relation, and not to any causal connection. But this realization leads us to a concern for causal discovery: Before applying any causal discovery methods to a dataset, we need to ensure that the variables are indeed all distinct and appropriate causal variables to be combined in a model, they should not be definitionally related. The approaches to relational causal learning in the previous section provide one avenue to address this challenge. I will here consider a different approach motivated not by relational databases, but by the challenge of constructing causal variables.

Consider the following example: There have been several studies exploring which features of a face lead to judging that face to be attractive. Subjects are shown a variety of portraits and asked to rate them on a scale of how attractive they consider the faces to be. What is the (visual) cause of such an attractiveness judgment? In this case we have an effect variable (the judgment), but it is unclear what the cause is. Is each pixel of the image a cause? Is the presence of a smile in the picture the cause? Is there a correct level of description at which to identify the cause? — Commonly, the symmetry of facial features is cited as a candidate cause of attractiveness judgments. (See e.g. the overview here: https://en.wikipedia.org/wiki/Facial_symmetry. The proposal is supported by evidence that changes in the symmetry of the depicted faces lead to changes in the attractiveness judgment.

One may well wonder whether symmetry tells the whole story. There is evidence that perfect symmetry appears uncanny and that slight asymmetries in the face score higher on attractiveness. For our purposes here, the key question is about how this search for causes should be approached in the first place. If we just consider candidate causal hypotheses that are easily described in a few words, then even if we find that they have some effect on the attractiveness judgment, they might only describe *aspects* or *indicators* of the full (visual) cause of the attractiveness judgment.

In current machine learning circles this concern would be approached using a deep neural net to identify possibly very complex features in the portrait images that do strongly predict the attractiveness judgment, even if the features themselves do not lend themselves to a simple description in natural language. Since the studies use an experimental set-up in which the evaluating subjects were shown portrait images in a lab setting that minimizes any confounding, the predictive features such a deep neural net detects can be considered to be causes of the attractiveness judgments, not merely predictive features. The appropriate description of the cause of the attractiveness judgment is not at the level of the pixels of the image, but at the level of the features of the image. The pixels *constitute* the features in the images, but the individual pixels are not causes of the attractiveness judgment. Manipulating an individual pixel will not change the attractiveness judgment (unless it is a very coarse image).

Here then is an attempt to give a coherent causal account of how we might identify and construct the cause of the attractiveness judgment from a low level pixel space: The pixels of the image containing the face describe a high dimensional state space $\mathscr{I}$. An image of a face specifies a state $I = i$ in the pixel state space $\mathscr{I}$, where I is the variable ranging over the state space. We can use a neural net to identify the features in pixel space that predict the attractiveness judgment E. The features identified by the neural net then provide a partition Π of the state space of the pixels $\mathscr{I}$. The cause C of the attractiveness judgment E is then a variable, whose state-space stands in a bijection to the labels of the cells of the partition Π identified by the neural net. So the cause C of the attractiveness judgment E *supervenes* on the pixel state space $\mathscr{I}$, in the sense that any change in the cause C necessitates a change of at least one pixel, but there may be changes of the pixels that do not affect the cause of the attractiveness judgment E. Importantly, the relationship between the pixels and the cause of the attractiveness judgment is a mathematical one, not a causal one. That is, the variable I ranging over the pixel state space $\mathscr{I}$ is related to C, the cause of the attractiveness judgment, by a mathematical (definitional) relation, not a causal one: $C = f(I)$. Figure 6.3 illustrates the relations, including the possibility the effect may also supervene on a much higher dimensional space $\mathscr{J}$. For example, the (mental) attractiveness judgment is presumably defined in terms of the underlying neural activity. In general, there may also be unobserved confounding between C and E, which complicates the inference from I to C, since some features of I now may be merely predictive of E, but no longer causes of E.

In the scenario of judging the attractiveness of images of faces shown in a lab setting, the features that a neural net identifies are obviously causal because the set-up is experimental: the pictures of faces are shown to the judge in experimental conditions. There is no confounding between content in the picture and the attractiveness judgment. In these conditions, a simple application of a neural net an appropriate way to identify causal relations, since the predictive features correspond to the causal ones. — But what if the data were non-experimental? For example, suppose we have a temperature map and a wind map over a specific geographical region that specify the temperature/windspeed for each location in the region. We obtained this data by observation, not experimentation, and for all we know, the two maps may be heavily confounded by the location of the sun or other geological activity. Can we still analyze the causal effect of the wind on the temperature?

One approach would be to consider the causal relations at the "pixel" level. Each individual "pixel" of the wind map would be a candidate cause for any of the "pixels" of the temperature map. If we allow for feedback, then the reverse relations would also have to be considered, and there could be confounding. But even leaving feedback relations and confounding aside, this approach would be equivalent to claiming in our earlier example of attractiveness judgments that each pixel of the image of the face individually may be a cause of the attractiveness judgment. While it is possible that our judgments of attractiveness are that sensitive, it seems implausible. The cause of the attractiveness judgment is a feature that *supervenes* on the pixels, it is a function of the pixels, the pixels themselves are not the relevant causal variables. Analogously then, a causal analysis of the relationship between wind and temperature over a geographical area may well exist at a coarser level of description than the

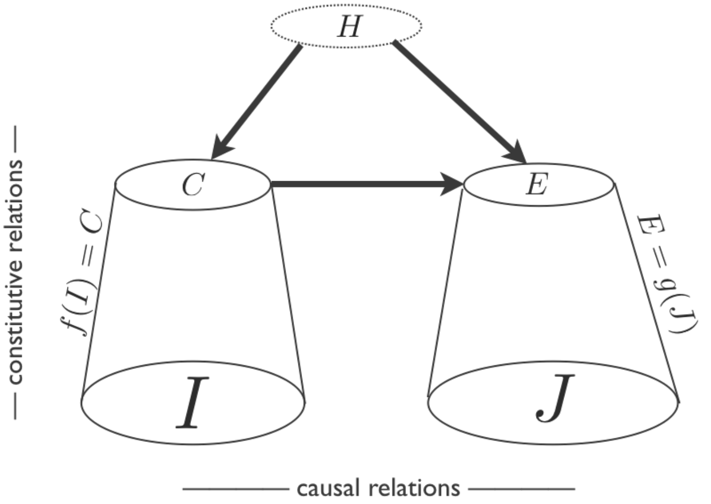


Fig. 6.3: Construction of cause effect pairs: Consider I and J to be very high-dimensional measurement variables, such as, for example images, or temperature maps, or neural recordings. In many cases only coarser descriptions of the underlying system are relevant to the causal questions we are interested in. For example, for the causal question "What makes a face look attractive?" we are not interested in the pixel values of the image depicting the face, but in candidate higher level features. Similarly, on the effect side, we are interested in the attractiveness judgment, not necessarily in the neural details that implement the attractiveness judgment. So we have a cause C of an effect E (which may in principle also be confounded by an unobserved H), but C and E supervene on the low level measurement spaces I and J, respectively. While the relations among C and E (and H) are causal, the relation between I and C is not causal, but constitutive (similarly, for J and E). Any intervention on C necessarily is an intervention on I. Given I and J, can we nevertheless construct a causal pair C and E?

"pixel level" of measurement. How then can we construct a cause-effect pair from two high-dimensional spaces of observational low level measurement variables?

In Chalupka et al. [2015] we addressed this problem in the following way: Given high-dimensional measurement variables I and J, we wanted to find a method that could determine whether there are coarser variables C and E such that C is a cause of E, i.e. $C \to E$, and $C = f(I)$ and $E = g(J)$ for some surjective functions f and g. One of the criteria of identifying such a macro-level cause C of E is that we have to be able to define intervention distributions $P(E|do(C))$. That is, in order to define C, we have to be able to make sense of what it means to intervene on C and specify a well-defined effect for such an intervention.

We started with an approach similar to the one proposed above for the case of attractiveness judgments: Given the observational conditional probability distribution $P(J|I)$, we clustered states of I and J according to the following two equivalences:

$$\begin{aligned} i_1 \sim i_2 &\Longleftrightarrow \forall j \in J, \quad P(j|i_1) = P(j|i_2) \\ j_1 \sim j_2 &\Longleftrightarrow \forall i \in I, \quad P(j_1|i) = P(j_2|i) \end{aligned}$$

That is, we clustered states of I if they implied the same conditional probability distribution for J, and we clustered states of J if for any i they had the same conditional probability distribution. This clustering can be performed by a neural net, and results in a partition $\Pi_o(I)$ of the state space of I and a partition $\Pi_o(J)$ of the state space of J. If the probabilities in the above equivalences had been interventional probabilities, i.e. $P(J|do(I))$, then the discovered partitions would already describe the states of C and E, respectively. But so far, the partitions only describe the dependencies between I and J in a maximally succinct form. These dependencies could still be entirely due to confounding between I and J.

So the questions is, how do the observational partitions Π_o defined by the equivalences above, differ from the causal partitions Π_c of the state spaces of I and J defined by:

$$\begin{aligned} i_1 \sim i_2 &\Longleftrightarrow \forall j \in J, \quad P(j|do(i_1)) = P(j|do(i_2)) \\ j_1 \sim j_2 &\Longleftrightarrow \forall i \in I, \quad P(j_1|do(i)) = P(j_2|do(i)) \end{aligned}$$

The state space of C is defined as a bijection to the labels of the cells of the partition $\Pi_c(I)$ implied by the first equivalence: While there may be several different states of I that map to the same state of C, those differences are causally irrelevant to E.

We showed in the Causal Coarsening Theorem (see proof in Chalupka et al. [2017]) that under relatively weak assumptions, if indeed there are descriptions of the causal system at a coarser level, then the causal partitions Π_c are a *coarsening* of the observational partitions Π_o. That is, the distinctions in the state space of I that are found by clustering on the basis of $P(J|I)$ are a superset of the distinctions in the state space of I that have a causal influence on J.

This should not come as a surprise: all dependencies between two variables, whether due to a causal relation or confounding, can be useful for predicting one variable from another. This is the reason why we use a barometer to predict the weather tomorrow: The distinctions it makes (in tracking pressure) are useful for prediction. But we do not think that the barometer reading is a cause of the weather tomorrow. The observational partition of the readings of the barometer needle are very fine, but the causal partition of the barometer readings is maximally coarse, since every reading has the same causal effect on the weather tomorrow, namely none.

The Causal Coarsening Theorem provides the basis for an efficient experimental method to check which distinctions in the observational partition are actually causally relevant. One does not have to check every possible state of I, but only the different distinctions of the observational partition. In cases where one cannot intervene, the methods developed as a result of the cause-effect pair challenge provide a natural basis to start detecting which features are causal and which are due to confounding.

The approach provides an account of how to construct causal variables with well-defined intervention distributions from low level measurement data. In Chalupka et al. [2016a] we extend this approach to consider cases where there might be multiple levels of causal description at various levels of analysis. But these results by no means provide a full-fledged account of how to identify causal variables. As they stand, they are restricted to discrete variables without feedback.

6.6 Final Remarks

My view is that the cause-effect pair challenge at the NIPS 2013 workshop failed in an interesting way: It did not produce winning methods with interesting generalizable insights for the field of causal discovery. However, this failure highlighted an important point about the limits of existing causal discovery methods: that for specific settings with well-established training and test data, brute force black box machine learning methods will outperform causal discovery algorithms. What is one to make of that? — Unlike the case of image or text classification, we rarely have a good understanding of what the ground truth causal relations look like. So the success of black box machine learning methods on these sorts of challenges provides very limited assurance of their success in general. This raises the question of how to structure causal discovery challenges in future? The Tübingen test data set seems like a step in the right direction. But we now need to go from the "causal MNIST" dataset to the "causal ImageNet". So which domains could provide large and varied datasets with known causal ground truth?

With regard to the general question of where we stand with regard to handling causal structures that are observationally Markov equivalent, I think the field has made enormous steps forward, largely driven by the careful analysis of the Additive Noise Model framework (which I frame broadly to include the LiNGAM methods). But on that front, the advance on the theoretical side has not yet been matched by broad successes in application.

References

Tetrad code package. URL http://www.phil.cmu.edu/tetrad/about.html.

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *Proceedings of UAI*, 2015.

K. Chalupka, P. Perona, and F. Eberhardt. Multi-level cause-effect systems. In *Proceedings of AISTATS*, 2016a.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Estimating causal direction and confounding of two discrete variables. *arXiv preprint arXiv:1611.01504*, 2016b.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

José AR Fonollosa. Conditional distribution variability measures for causality detection. *arXiv preprint arXiv:1601.06680*, 2016.

D. Geiger and J. Pearl. On the logic of causal models. In *Proceedings of UAI*, 1988.

Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906, 2015.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

Daniel Hernandez-Lobato, Pablo Morales Mombiela, David Lopez-Paz, and Alberto Suarez. Nonlinear causal inference using gaussianity measures. *Journal of Machine Learning Research*, 2016.

P.O. Hoyer, D. Janzing, J.M. Mooij, J.R. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. 2008a.

P.O. Hoyer, S. Shimizu, A.J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008b.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, 90:208–225, 2017.

Dominik Janzing and Bernhard Schoelkopf. Detecting confounding in multivariate linear models via spectral analysis. *arXiv preprint arXiv:1704.01430*, 2017.

Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard 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.

G Lacerda, P Spirtes, J Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of UAI*, pages 366–374, 2008.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya O. Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461, 2015.

Marc Maier. *Causal Discovery for Relational Domains: Representation, Reasoning, and Learning*. PhD thesis, 2014.

Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. A sound and complete algorithm for learning causal models from relational data. *arXiv preprint arXiv:1309.6843*, 2013.

C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of UAI*, pages 411–418. Morgan Kaufmann Publishers Inc., 1995.

Bram Minnaert. Feature importance in causal inference for numerical and categorical variables.

Diogo Moitinho de Almeida. Pattern-based causal feature extraction.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Joseph D Ramsey, Stephen José Hanson, and Clark Glymour. Multi-subject search correctly identifies causal connections and most causal directions in the dcm models of the smith et al. simulation study. *NeuroImage*, 58(3):838–848, 2011.

Subhradeep Roy and Benjamin Jantzen. Detecting causality using symmetry transformations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075305, 2018.

Spyridon Samothrakis, Diego Perez, and Simon Lucas. Training gradient boosting machines using curve-fitting and information-theoretic features for causal direction detection.

Oliver Schulte and Hassan Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88(3):331–368, 2012.

Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *The Journal of Machine Learning Research*, 15(1):2629–2652, 2014.

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

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

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of UAI*, pages 647–655. AUAI Press, 2009.

Kun Zhang and Lai-Wan Chan. Extensions of ica for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pages 400–409. Springer, 2006.