ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Learning of Causal Relations

John A. Quinn[1] and Joris Mooij[2] and Tom Heskes[2] and Michael Biehl[3]

[1]Faculty of Computing & IT, Makerere University
P.O. Box 7062, Kampala, Uganda

[2] Institute for Computing and Information Sciences
Radboud University Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

[3]Johann Bernoulli Institute for Mathematics and Computer Science
University of Groningen, P.O. Box 407, 9700AK Groningen, The Netherlands

Abstract. To learn about causal relations between variables just by observing samples from them, particular assumptions must be made about those variables' distributions. This article gives a practical description of how such a learning task can be undertaken based on different possible assumptions. Two categories of assumptions lead to different methods, constraint-based and Bayesian learning, and in each case we review both the basic ideas and some recent extensions and alternatives to them.

1 Introduction

Much of machine learning is concerned with modelling joint distributions of sets of variables, with the objective of reasoning about the likely values of some variables given observations of others. A different type of analysis is to make inferences about the values of some variables when other variables are *manipulated.* To do this requires an understanding of which variables are causes of other variables; when we say that variable X_1 is a "cause" of variable X_2, we mean that a physical intervention to change the value of variable X_1 produces a change in the distribution of X_2 regardless of the values of all other variables we observe. Gaining insight into such relationships between variables is clearly valuable—indeed, a major part of scientific discovery is to explain observed phenomena by finding their causes.

The most obvious way to learn such a relationship is to actually make a series of interventions on a suspected cause and observe the effects, while controlling for the possible confounding effects of other variables. However, for practical or ethical reasons it is not always possible to make interventions, and hence our interest in the field of inferring causal relationships from purely observational data. Given some data $\mathcal{D}$ which is a set of samples from variables $X_1, \ldots, X_n$, we wish to find a causal model M which describes the causal relationship, if any, between each pair of variables.

We can formulate this problem as the estimation of a set of structural equations. Given observed random variables $X_1, \ldots, X_n$ and unobserved (latent) random variables $E_1, \ldots, E_n$, we take n functions (which could be nonlinear) $f_1, \ldots, f_n$ that constitute the structural equation model

$$X_i = f_i(X_{\mathrm{pa(i)}}, E_i), \qquad i = 1 \ldots, n \tag{1}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

where $\mathrm{pa}(i)$ is a subset of $\{1, ..., n\}$. The key assumption to link such a probabilistic model to causality is that each functional relationship corresponds with a direct causal relationship: X_i is caused by its "parents" $X_{\mathrm{pa}(i)}$ and by the noise variable E_i. One can represent this structure by a graph with arrows from E_i and $X_{\mathrm{pa}(i)}$ to X_i, for each i.

Estimating the structure $\{\mathrm{pa}(i)|i = 1, \dots, n\}$ from data hinges on finding independencies between variables. To illustrate this, consider an example where there are three variables, X_1, X_2, X_3. Given many observations of these variables, say we test for independence between all pairs and find the independence $X_1 \perp\!\!\!\perp X_2$ but the dependencies $X_1 \not\perp\!\!\!\perp X_3$ and $X_2 \not\perp\!\!\!\perp X_3$. Furthermore, we test each pair of variables when conditioned on the third variable and find a dependency in every case: $X_1 \not\perp\!\!\!\perp X_2|X_3$, $X_1 \not\perp\!\!\!\perp X_3|X_2$ and $X_2 \not\perp\!\!\!\perp X_3|X_1$. Of the 27 causal configurations possible, only one is consistent with the single independence: $\mathrm{pa}(1) = \emptyset$, $\mathrm{pa}(2) = \emptyset$, $\mathrm{pa}(3) = \{1, 2\}$. This could be rendered graphically as $X_1 \to X_3 \leftarrow X_2$; such a structure is known as a *collider* or *V-structure*. To help see the link between the independence relation and the causal structure, imagine seeing someone enter a room with wet hair, who may have either taken a shower recently, or arrived by bicycle in the rain. Take X_1 to be "shower", X_2 to be "cycled in the rain" and X_3 to be "wet hair". If X_1 and X_2 are taken in isolation, their occurrences may be independent—knowing X_1 gives us no information about X_2. However, an observation of X_3 creates a dependency between X_1 and X_2: if we see that the person's hair is wet, and also we know that it is hot and sunny outside, then we have information about whether the person may have had a shower, i.e. that it is more likely. A similar structure arises if we imagine that X_1 and X_2 are the DNA information of a mother and father respectively, and X_3 is the DNA of their child.

In common with any type of inference, we cannot reach any conclusions about causal structure without making assumptions. The remainder of this article describes some different approaches to causal structure learning, from the perspective of which assumptions can be made and what can be learnt given these different assumptions. Note that these assumptions are typically made because they lead to the ability to make strong inferences. In practice they might hold to varying extents, and we highlight those that tend to be made for the sake of convenience at the expense of fidelity of the model.

A readable account of the assumptions made in causal inference are given by Scheines [1], with further details given in [2, 3]. Guyon et al [4] review causal learning methods from the perspective of causal feature selection; this task has many aspects in common with more general causal learning problems.

2 Faithfulness and sufficiency: constraint based learning

A central assumption made in causal learning is that of *faithfulness*, which is that each (conditional) independence relationship found in the joint distribution of $X_1, \dots, X_n$ is due to the causal structure, and not to the peculiarities of the probability distributions $p(E_i)$ or the functions $f_i(\cdot)$. Without a strong assumption like the faithfulness assumption, the inferences which can be made are weak: we cannot tell for certain whether the independencies found in the data are structural or coincidental. In our experience, when working on real-world datasets this assumption tends to be benign.

Another popular assumption, though somewhat less justified on a typical real-world dataset, is *sufficiency*: that is, that all source nodes of the graph (all X_i where $\text{pa}(i) = \emptyset$) are jointly independent random variables. In other words, there are no confounders (hidden common causes). This is an assumption about which variables were measured when the data was collected. It is always possible in principle for there to be an unobserved common cause that changes the independence relationships. It is still possible to make weak causal inferences when this assumption is removed, but only when more data is collected and additional assumptions are made, which are beyond the scope of this article.

Another assumption that is made in the great majority of causal learning literature is that the causal graph is acyclic. The difficulties in causal inference when there are cycles of causal influence is clear—we depend on identifying independencies, but there cannot be independence between any pair of variables in a cycle. Hence in such settings it is difficult to distinguish the direction of any causal relationships.

These three assumptions (all of which are necessary even to reach the conclusion for the simple example in the previous section) lead directly to constraint-based structure learning as a method for causal discovery. The idea is to detect a sufficient subset of the (conditional) independencies between the observed variables and from these to infer $\text{pa}(i)$ for each i.

Algorithm 1 is a prototypical method of doing this, which is explained in more detail by Pearl [2, §2.5], and which along with the similar PC algorithm [3] is the basis of much work on structure learning since both were proposed. The algorithm begins with identifying the colliders amongst all combinations of three variables, looking for sets in which two are independent when not conditioned, but dependent otherwise.

It is often the case that not all variables related to the causal system of interest are available for measurement. If the variables $X_1, \ldots, X_n$ are split into a set of observed variables O and hidden variables H, Algorithm 1 can be extended to infer causal relations in this setting [2, §2.6].

2.1 Testing for conditional independence

We implicitly assume above that we have a way of reliably assessing all (conditional) (in)dependencies, and generally some kind of hypothesis test is employed to assess this in each case. The type of independence test selected also brings in implicit assumptions about the form of the data. Fisher's Z-test can be used for multivariate Gaussian distributions, for which a p-value must be selected. If partial correlation is used as a conditional independence test for continuous data, this assumes that any relationships between variables are linear Gaussian. Mutual information can be used for discrete valued data. Another approach to quantifying dependence is to assess how well X_i predicts X_j given some condi-

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Algorithm 1: Prototypical constraint-based structure learning.

```
Input: a dataset $\mathcal{D}$ drawn from observed variables $X_1, \ldots, X_n$.
Output: a graph $G$ with arcs representing the causal influences between
        variables.
Construct a graph $G$ with $n$ vertices and no edges.
for every pair of variables $X_i$, $X_j$ where $i, j \in \{1, \ldots, n\}$ do
    for every conditioning set of variables $S \subset \{X_1, \ldots, X_n\} \setminus \{X_i, X_j\}$ do
        Test for independence $X_i \perp\!\!\!\perp X_j | S$ in the data $\mathcal{D}$
    end
    if no independence was found then
        Add undirected edge $(i, j)$ to $G$.
    end
end
for every i,j,k for which there are edges (i,k) and (j,k) in G but not (i,j)
do
    Orientate the edges $i \to k$ and $j \to k$ in $G$.
end
for every undirected edge $(i, j)$ in $G$ do
    Orientate edge $i \to j$ or $j \to i$ if both (i) no new colliders are created
    and (ii) no directed cycles are introduced in $G$.
end
```

tioning set of variables. This principle has been used in work to find colliders using a nonparametric prediction framework [5].

As the algorithm makes a hard (binary) decision about independencies, some significance threshold has to be applied. Finding the right threshold is non-trivial: if for some causal learning system we raise or lower the significance threshold, we find that the number of causes returned by the system increases and decreases respectively. In practice it may not be immediately obvious what the right threshold is. Too low, and dependencies will not be detected (false negatives); too high, and non-causes will be falsely reported as causes (false discoveries). This is studied further in [6], work which shows that common thresholds give a higher false discovery rate than might be expected, and suggests mechanisms for controlling this rate in the context of finding the undirected graph.

Note also that independencies have to be found for each possible conditioning set between a pair of variables, $2^{|V|-2}$, making a naïve version of the algorithm scale exponentially in the number of variables.

2.2 Extensions of and alternatives to the basic method

The methods described so far can only identify the causal model up to Markov equivalence. In other words, based on (conditional) (in)dependence statements between the variables alone, not all the edges can be oriented. The simplest causal discovery problem where the classical methods fail, is the bivariate case [7]. Indeed, a dependence between X_1 and X_2 does not yield any conclusions about the direction of the causal arrow between both variables, because there are no conditional (in)dependence statements to be made. Distinguishing between models in the Markov equivalence class therefore requires other assumptions.

The first approach in that direction was LiNGAM [8]. The assumption is that all functions $f_i(\cdot)$ are linear, and in addition, that all probability distributions of source nodes in the causal graph are non-Gaussian. Under these assumptions, the causal structure is identifiable, i.e., *all* edges can be oriented. The linearity assumption can be weakened: instead one could assume that the noise variables E_i are additive, so that each structural equation is of the form $X_i = f_i(X_{\mathrm{pa}(i)}) + E_i$. This leads to the additive noise method in [9], which does not require non-Gaussian probability distributions. Finally, the "postnonlinear" model [10] allows for an additional nonlinear function after the noise has been added: $X_i = g_i\big(f_i(X_{\mathrm{pa(i)}}) + E_i\big)$.

Another possibility is to look at information-theoretic "independence" between the distribution of the cause and the function: usually, the causal mechanism has been chosen independently from the distribution of the cause. Any "mutual information" between those two objects is suspect and indicates a wrong causal direction. This leads to the deterministic method in [11], a general non-deterministic method in [12] and the method described in [13]. All these methods prefer the models which are "simpler" in a information-theoretic sense and thereby allow us to resolve the causal structure to a finer scale than the Markov equivalence class.

The prototypical Algorithm 1 can also in fact be improved without requiring any additional assumptions. The steps on lines 11 and 14 can be completed by following a set of seven rules. However, these rules have recently been found to be special cases of just two underlying rules [14], by exploiting a notion of observed minimal conditional independence; the idea is to find all information which rules out particular causal influences between variables.

We may be interested in incorporating prior knowledge in such learning. An obvious way to do this is to take any other knowledge about whether certain edges must exist or cannot exist, and constrain $\mathrm{pa}(i)$ accordingly. We might have knowledge about causal influences that cannot exist when there is temporal information: a variable cannot usually have a causal effect on another variable which precedes it in time. One way of doing this is with an intermediate representation known as a Maximum Ancestral Graph (MAG), which can encode information about variable X_1 having a causative effect on X_2, possibly acting through a number of intermediate variables. Working out which possible causal configurations satisfy such prior constraints can then be framed as a satisfiability problem [15]. Prior information can also be incorporated in another framework using Bayesian principles, which we now describe.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

3 Priors over structures: Bayesian structure learning

Learning causal structure might be seen as an iterative process, in which we continually make observations of quantities of interest and use this information to update our model of causal interactions. Before performing statistical learning of causal relations, we might therefore have *a priori* knowledge of relationships between particular variables. We can formalise this by assuming prior knowledge $P(\Theta)$, where Θ is a structure containing both the model edges $\mathrm{pa}(i)$ and parameters. Given observed data $\mathcal{D}$, we can then obtain a posterior distribution $P(\Theta|\mathcal{D})$ via Bayes' rule.

Probabilistic graphical models have long been used as a way of representing joint probability distributions. In a directed graphical model, the presence or absence of an edge between two variables signifies an independence relationship. Although in general such models are merely a mathematical artifact which provides a compact representation of a joint probability distribution between a set of variables, with some extra assumptions (expounded in [16] and [3]) they can be thought of as causal models in which the parents of a variable are the direct causes of that variable.

In principle, the Bayesian approach to inference is to explore every possible model, calculate a posterior probability of each model given the prior and the data which has been observed, and answer inferential questions such as "is there a causal relationship $X_i \rightarrow X_j$?" with reference to a weighted average under all these possible models [17]. In the setting of causal structure learning, the model space is generally too large for this to be tractable with current methods, however, since the number of possible structures is super-exponential in the number of variables. A statistic which is easier to calculate is the maximum *a posteriori* (MAP) model, which is the single set of parameter settings that is most plausible given prior and data, $\Theta_{\mathrm{MAP}} = \arg\max_\Theta \left(P(\mathcal{D}|\Theta)P(\Theta)\right)$. In this approach, we find a single high scoring model and pretend it is the "right" one (this could be seen as an assumption that the model posterior distribution is sharply peaked around Θ_{MAP}).

The likelihood $P(\mathcal{D}|\Theta)$ can be calculated by making assumptions on the form of the variable distributions. Commonly these are discrete variables (in which case a number of conditional probability tables are estimated from $\mathcal{D}$) or for the case of continuous variables that the relationships are linear Gaussian. The prior $P(\Theta)$, as well as encoding any knowledge we have from the domain about which models are more likely, can also apply general principles that we believe to be true about models, such as that simpler models are more compelling (a minimum description length prior). A common choice in the absence of specific domain knowledge is the uniform Bayesian Dirichlet prior (BDeu) [16].

Given that the search space is so large and analytical approximations are difficult to make, the strategy usually employed here is known as "search-and-score". This approach requires three components. The first is a way of scoring a particular model, as discussed above. Many other scoring functions exist in the literature, such as the Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC) and Minimum Description Length (MDL). The second requirement is a way to iterate between models, for example by randomly introducing or removing arcs and reversing arc directions. Finally, we need some kind of search strategy, which could be a greedy hill climbing algorithm or a stochastic search method. With all these ingredients, we can iterate between models trying to explore high-scoring regions of the model space, and find the single highest-scoring model we can.

Some techniques use elements of both constraint based learning and search-and-score. MMHC [18] is a hybrid algorithm which uses the former for finding an undirected network, and the latter for searching for orientations. This increases the speed of processing, the possibilities being reduced to $2^{|\varepsilon|}$ configurations where $|\varepsilon|$ is the number of undirected edges found in the first stage. Another improvement to the basic method described here is to explore only the space of conditional independence equivalance classes, as in the Greedy Equivalent Search [19].

3.1 Induction with hidden variables

The preceeding method finds the best fitting model given some observations, where the data is fully observed. If there are latent variables, one approach to learn the model is by application of the Expectation Maximisation algorithm. This algorithm alternates between two stages. In (i) the expectation step (or "E-step"), expectations of the values of latent variables are sampled from the model. We use $\mathcal{D}^*$ to denote the values of variables which were unobserved, corresponding to the observed data $\mathcal{D}$. Given the data and current estimate of the parameters, we can calculate summary statistics of the distribution of unobserved variables $P\left(D^*|\mathcal{D},\Theta\right)$. (ii) The maximisation step (or "M-step") finds a model structure and parameters that maximise an auxilliary function as follows:

$$\hat{\Theta}_k \leftarrow \arg\max_{\Theta} \int \log P(\mathcal{D},\mathcal{D}^*|\Theta)P(\mathcal{D}^*|\mathcal{D},\hat{\Theta}_{k-1})\,\mathrm{d}\mathcal{D}^* \ . \qquad (2)$$

The M-step is composed of two parts in the case of causal structure learning: finding the structure which maximises the model score (which may involve adding or removing hidden variables, as well as changing edges), and finding the parameters that best fit the model. The updated model structure is found using the same principle as for fully observed data above, using $E\left[D^*|\mathcal{D},\Theta\right]$. In order to complete the M-step, we therefore only have to find the best fitting parameters given the new structure. This is straightforward given a particular parameterisation (such as that all relationships between variables are linear Gaussian).

Given some initial parameters and model structure Θ_0, we therefore begin by performing one E-step to estimate the most likely settings of the unobserved variables under the current model. Given this expectation over the complete data, we then update the model structure and parameters. These steps are alternated until convergence.

4 Future directions

There are many areas in causal learning which are the subject of further investigation. Part of this involves the development of methods to deal with hard cases, for example to deal with cyclic causal relationships or with high dimensional data with few observations. Datasets with these latter characteristics are commonly encountered in biology, for example in genetic research.

Another interesting area under study is reasoning when there is a combination of observational and experimental data. This is straightforward to deal with when dealing with "ideal" manipulations that affect only one variable; e.g., in Bayesian learning we simply do not update the parameters for any variables we know to have been intervened upon. However, there are many other types of intervention possible, for example which change the parameters of the causal relationship or the probability distributions of the source nodes. We may also have a situation in which the data we observe is a result of some manipulation, but we have some uncertainty about the nature of the manipulation [20].

The classical causal discovery methods are mostly incapable of distinguishing between models in the same conditional independence class, motivating further work on finding other principles with which to favour one causal structure over another, such as the simplicity of representation. Speed of execution is also an issue: given that the model space is large in Bayesian search-and-score, and constraint based learning has to search an exponential number of conditioning sets, causal learning algorithms tend to be slow when there are large numbers of variables.

The literature is divided on the claims which can be made regarding the strength of causal inferences. Induction of the type we have described here may be useful as a first step, to guide further research: if a causal learning method presents a number of possible causes of a target variable, this could be used as a way of prioritising potential causes for further investigation by experimentation.

Finally, if we want our causal discovery methods to become mainstream, what is still lacking as of today are convincing applications where causal discovery methods are the key to obtaining good results.

References

[1] R Scheines. An introduction to causal inference. In *Causality in Crisis?* University of Notre Dame Press, 1997.

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

[3] P Spirtes, C Glymour, and R Scheines. *Causation, Prediction and Search.* MIT Press, 2000.

[4] I Guyon, C Aliferis, and A Elisseeff. Causal Feature Selection. Technical report, Clopinet, 2007.

[5] E. Mwebaze, M. Biehl, and J.A. Quinn. Causal relevance learning for robust classification under interventions. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

[6] P.A. Armen and I. Tsamardinos. A unified approach to estimation and control of the false discovery rate in Bayesian network skeleton identification. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

[7] J. M. Mooij and D. Janzing. Distinguishing between cause and effect. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 6, pages 147–156, 2010.

[8] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear, non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[9] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf:. Nonlinear causal discovery with additive noise models. In *Proc. Neural Information Processing Systems (NIPS)*, 2008.

[10] K. Zhang and A. Hyvarinen. On the identifiability of the post-nonlinear causal model. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2009.

[11] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2010.

[12] J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Proc. Neural Information Processing Systems (NIPS)*, 2010.

[13] J. Lemeire, S. Meganck, F. Cartella, T. Liu, and A. Statnikov. Inferring the causal decomposition under the presence of deterministic relations. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

[14] T. Claassen and T. Heskes. A structure independent algorithm for causal discovery. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

[15] G. Borboudakis, S. Triantafillou, V. Lagani, and I. Tsamardinos. A constraint-based approach to incorporate prior knowledge in causal models. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

[16] D Heckerman. A Bayesian Approach to Learning Causal Networks. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 1998.

[17] N. Friedman and D. Koller. Being Bayesian about network structure. In *Proc. Sixteenth Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 201–210, 2000.

[18] I Tsamardinos, LE Brown, and CF Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[19] D. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

[20] D Eaton and K Murphy. Exact Bayesian Structure Learning from Uncertain Interventions. In *Proc. AI and Statistics (AISTATS)*, 2007.