Causality: A Selective Review

Markus Kalisch and Peter Bühlmann
ETH Zürich

September 27, 2013

In this paper we give a review of recent causal inference methods. First, we discuss methods for causal structure learning from observational data when confounders are not present and have a close look at methods for exact identifiability. We then turn to methods which allow for a mix of observational and interventional data, where we also touch on active learning strategies. We also discuss methods which allow arbitrarily complex structures of hidden variables. Second, we present approaches for estimating the interventional distribution and causal effects given the (true or estimated) causal structure. We close with a note on available software and two examples on real data.

1 Introduction

A main goal of many scientific investigations is to establish the nature of dependencies. Statistics has a long tradition in dealing with events occurring together. For instance, in medical statistics, one might want to find a biomarker, say in the blood, that occurs together with a (perhaps early form of a) disease. The knowledge of this dependence could be used to predict the occurrence of the disease based on a blood test.

Suppose we detect such a dependence between the occurrence of the biomarker and the disease. We can then use the biomarker as a predictor for the disease. Beyond the use of the biomarker for predicting the disease, we might also be interested in a cure for the disease. In our search for a cure, we could argue along the following lines: If the biomarker would cause the disease, this would explain the observed dependence. Therefore, manipulating the biomarker in a proper way will cure the disease.

Of course, this conclusion based on an observed dependence can be wrong. The biomarker causing the disease is just one out of several scenarios in which a dependence between occurrence of biomarker and disease can be observed (see Fig. 1(a)). In an alternative scenario with the same observed dependence, the disease might be the cause of the biomarker (see Fig. 1(b)). Thus, manipulation of the biomarker would have no effect on

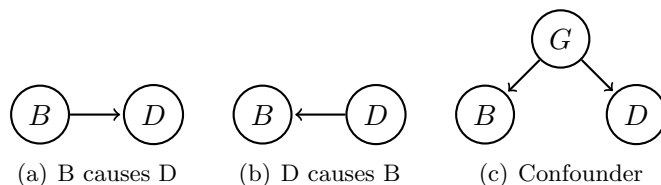(a) B causes D   (b) D causes B   (c) Confounder

Figure 1: Several causal models might give rise to a dependence between biomarker and disease. Therefore, it is not trivial to deduce the true causal model from the observed dependence.

the disease. As a final example, there might be a scenario, in which we did not observe all relevant variables. Thus, it might be the case, that there is an unobserved third variable (e.g. genetic disposition) which is the common cause of both the occurrence of the biomarker and the disease (see Fig. 1(c)). This would, too, give rise to a dependence between the occurrence of the biomarker and the disease, but manipulating the biomarker would have no effect on the disease. Hence, it is not trivial to infer causal relationships from observed dependencies.

For finding causal relationships, the gold standard are intervention experiments. In such an experiment, one not only observes nature as it is, but makes an intervention and observes the consequences of this intervention. The main principles for experimental design were developed in the context of agricultural research (Fisher, 1926; Cox, 1958). A key ingredient of a good experiment is randomization and control: Suppose in our example, that we have access to a set of patients with the disease. We select half of the patients at random and manipulate the biomarker (treatment group). The remaining patients are not treated (control group). After a certain time, we check how many patients were cured in the treatment group and compare with the control group. Suppose that many more patients were cured in the treatment group than in the control group. Since the treatment allocation was random, it is very unlikely that we (randomly) assigned treatment only to those patients, which would have become healthy anyway (e.g. because of genetic disposition). Thus, we have established, that the treatment has a causal effect on the disease.

In some situations, however, it is not feasible to do an intervention experiment for ethical, financial or other reasons. For example, it would be considered unethical to force a random selection of people to smoke in order to establish smoking as cause for lung cancer.

The question arises therefore, whether it is at all possible to identify cause and effect relationships by observation alone. The short answer is: No, not in the general case. However, if we make suitable assumptions, the answer becomes affirmative. In the past two decades, theories were developed which use suitable assumptions in such a way that causal relationships

indeed become identifiable from observational data.

In this article, we give a review of some of these methods. We focus on methods that aim at identifying the effect of an experimental intervention (as opposed to counterfactuals) given observational data or a mix of observational and interventional data. For alternative introductions to causality we point to Pearl (2009b), Spirtes *et al.* (2000), Spirtes (2010) and Pearl (2009a).

In chapter 2 we will give background information on concepts and notation. At some points, we will anticipate results from later chapters in order to put the definitions into context. In chapter 3 we will focus on methods for estimating the causal structure from observational data or a mix of observational and interventional data. In chapter 4 we will present approaches for estimating the interventional distribution (i.e., the probability distribution of a random variable in a causal system after some intervention was done at some other random variable in the causal system) based on observational data and additional information such as the causal structure. Finally, in chapter 6 we give some concluding remarks.

2 Background

We will focus on structural equation models, originally formulated by Wright (1921) and treated in detail by Pearl (2009b). A structural equation model describes a causal system by a set of equations. Each equation explains one variable of the system in terms of the variables, which are its direct causes. Moreover, some random noise might be involved. Finally, the equations are assumed to be *autonomous*: If the generating process of one variable (i.e. one equation) is changed, this has no effect on the generating process of the other variables (i.e., on the other equations).

As an example, we show a structural equation model on the three variables X, Y and Z and corresponding independent noise variables u_x, u_y and u_z:

$$
\begin{aligned}
X &\leftarrow f_1(u_x) \\
Y &\leftarrow f_2(X, u_y) \\
Z &\leftarrow f_3(X, Y, u_z)
\end{aligned}
\tag{1}
$$

In the equations, we use the symbol "$\leftarrow$" instead of "$=$" in order to indicate, that the equation should be interpreted in an *asymmetric* way. For example, $Y \leftarrow X + 1$ should be interpreted as "Y is generated by taking X and adding 1". Note, that we cannot invert this relationship (e.g. "X is generated by taking Y and subtracting 1") as we could do in algebraic equations.

The structure of model (1) can be visualized by a *causal graph* (also called *causal structure*): If X is a direct cause of Y (i.e., X is on the right

hand side of the equation defining Y), then there is an arrow from X to Y. By convention, noise variables are not added in this graph. Almost all of the theory presented in this review deals with causal systems without feedback. Thus, a variable cannot be a cause of itself (directly or indirectly) and therefore, in the corresponding causal graph, it is not possible to trace a circle when following the direction of the arrows. The resulting graph is therefore called a directed acyclic graph (DAG). The graph visualizing the structural equation model given in 1 is shown in 2.
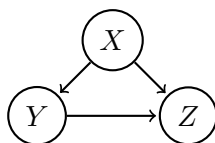

Figure 2: The causal structure visualizes the relationship between effects and direct causes of structural equation model (1) with a DAG.

We will now introduce some formalism for the causal graph. Two nodes are *adjacent*, if there is an edge between them. The *parents* of a node X in a DAG are all nodes from which an arrow points to X. Correspondingly, all nodes to which an arrow points from X are called *children* of X. A *path* is a sequence of distinct, adjacent vertices. A path is *directed*, if all edges on the path are directed in the same direction. The *ancestors* of X are all nodes from which a directed path leads to X. Correspondingly, the *descendants* of X are all nodes to which a directed path leads from X. The *skeleton* of a DAG is the undirected graph that is obtained when ignoring the directions of the arrows. The path $X - Y - Z$ where X and Z are not adjacent is called *unshielded triple*. The node Y on the path $X \rightarrow Y \leftarrow Z$ is called *collider*. If X and Z are not adjacent, Y is called an *unshielded collider*.

For example, in Fig. 2, X and Y are parents of Z, Y is a child of X and a directed path is $X \rightarrow Y \rightarrow Z$. X and Y would be the ancestors of Z even if the edge $X \rightarrow Z$ was removed. Finally, $X \rightarrow Z \leftarrow Y$ is a collider, but it is not an unshielded collider, since X and Y are connected.

Given data and some further assumptions we will then try to solve one of the two following problems:

Problem 1 (Causal Structure): Given observational data or a mix of observational and interventional data, find the DAG representing the causal structure, or, if this is not possible, give a class of DAGs among which the true DAG is.

Problem 2 (Interventional Distribution): Given observational data or a mix of observational and interventional data, find the interventional distribution of some variable in the structural equation model. The *interventional distribution* is the probability distribution of a random

variable Y in a causal system after some intervention was done at some other random variable X in the causal system setting the variable X to a certain value. Thus, using the interventional distribution, one can make quantitative predictions on the effect of interventions.

2.1 Without hidden variables

We first assume, that all variables of the underlying causal system have been observed and that the error variables are jointly independent among each other.

It turns out that each structural equation model has certain *invariance properties*, which only depend on the causal structure (but not on the functional details of the structural equation model). The most prominent example for invariance properties are conditional independencies. Given a causal structure, the implied conditional independencies can be read off by assuming the local Markov property: Every node is independent of its non-descendants given its parents (see chapter 3 of Lauritzen (1996) for much more details). For example, in Fig. 3(a), the local Markov property tells us that Z is independent of Y given A and B (conditioning only on B would be enough in this example). When applying the local Markov property on several different nodes, the generated list of conditional independencies might imply further (implicit) conditional independencies, which might be easy to miss. Therefore, in order to explicitly generate a list of all conditional independencies implied by the local Markov property, the (graphical) concept of *d-separation* was introduced (Pearl, 2009b, e.g. Definition 1.2.3). Two nodes X and Y are d-separated by the (possibly empty) set S, if on every path between X and Y at least (i) one non-collider is in S or (ii) one collider is not in S nor has a descendant in S. For example, in Fig. 3(a) we see that X and Y are d-separated by A, A and B are d-separated by the empty set, but A and B are not d-separated by Y. A distribution is said to be *Markov* with respect to a DAG, if all d-separation statements in the DAG hold as conditional independence statements in the distribution. Thus, using d-separation, it is easy to check whether a causal structure implies a given conditional independence statement. Note, that, assuming the Markov property holds, the conditional independencies found with d-separation hold for every structural equation model with this given causal structure (e.g. no matter what the details of the functional relationships or error distributions are). However, specific structural equation models might imply even more conditional independence statements. As a simplifying assumption, this is in practice often ruled out. Thus, one assumes, that the conditional independence assumptions implied by applying d-separation on the causal structure are exactly the same as the ones in the distribution described by the structural equation model. This assumption is known as *faithfulness*. A stronger version of faithfulness (called *strong faithfulness*) requires addi-

tionally that the strength of all non-zero conditional dependencies exceed a certain threshold.

It turns out that distinguishing between causal structures based on observational data is problematic: There might be several causal structures which are compatible with the list of invariance properties that can be computed from given observational data. This problem remains, even if we remove sampling issues from the data (i.e. replacing the data by an oracle). Based on this observation, DAGs are grouped into *equivalence classes* (Andersson *et al.*, 1997). DAGs within this class are called 'Markov-equivalent' and share the same skeleton and the same unshielded colliders (Verma and Pearl, 1990). Thus, in general, Markov-equivalent (or simply equivalent) DAGs cannot be distinguished based on observational data. Therefore, the answer to Problem 1 will, in general, be a Markov-equivalence class of DAGs, rather than a single DAG. We visualize the Markov-equivalence class with a 'completed partially directed acyclic graph' (*CPDAG*). A CPDAG consists of directed and undirected edges. All DAGs in the corresponding Markov-equivalence class have the same skeleton and the same directed edges as the CPDAG. However, for every undirected edge, there are at least two DAGs in the Markov-equivalence class whose corresponding edges point into opposite directions. For example, the DAG in Fig. 3(a) is in the equivalence class represented by the CPDAG in Fig. 3(b).

The existence of an equivalence class reflects our previous statement, that in general, we cannot infer causal effects from observational data. One way of dealing with this issue is to report suitable summary statistics of all DAGs within an equivalence class. Apart from that, there are several approaches to reduce the size of the equivalence classes and ideally obtain a unique DAG:

- Include knowledge on background, such as time, for further orienting some edges

- Make further assumptions on the structural equation model (see Sec. 3.1.1)

- In addition to observational data, acquire information on (perhaps optimally planned) intervention experiments (see Sec. 3.1.2)

2.2   With hidden variables

Up to now, we have assumed that all variables of the underlying causal system have been observed and that the error variables are independent. In real life, however, we often observe only part of all variables. I.e., some variables are observed, while others are hidden (or latent). Moreover, we often even don't know if or how many further latent variables in the causal

system exist or not. In this case, we still might be interested in causal statements among the observed variables.

An initial idea for solving this problem might be this: First, we assume that the true causal structure (including all latent variables) is a DAG G. Then, we try to marginalize out the latent variables. I.e., we try to come up with a DAG $\tilde{G}$ only on the observed variables, so that the invariance properties of $\tilde{G}$ coincide with the invariance properties on G restricted to only the observed variables. Our new goal would then be to estimate $\tilde{G}$ from data. Unfortunately, it turns out, that DAGs are not closed under marginalization. I.e., in general, there is no DAG $\tilde{G}$ which would match all invariance properties of DAG G when restricted to only the observed variables. To solve this problem, the broader class of ancestral graphs was developed, of which DAGs are a subset (Richardson and Spirtes, 2002). Of special importance in this class are the maximal ancestral graphs (MAGs). A MAG on the observed variables represents a class of infinitely many DAGs (on observed and arbitrarily many hidden variables) that have the same d-separation and ancestral relationships among the observed variables.

Although the theory of MAGs is useful for all kinds of hidden variables, a simplifying assumption often only allows hidden confounders, i.e., hidden variables which are a common cause of at least two observed variables (hidden variables that are cause of just one observed variables can be trivially solved by extending the noise variables). In contrast to this, there might also be hidden selection variables, i.e., hidden variables that are caused by at least two observed variables (hidden variables that are caused by only one observed variable can be dismissed, since they have no effect on the causal structure among the observed variables). Although some work has been done on structure learning with arbitrary hidden variables (Spirtes *et al.*, 1995), we will mainly focus on assuming only *hidden confounders*.

Graphically, a MAG consists of nodes representing (observed) variables and edges, which might be directed, undirected or bidirected. Thus, each end of an edge can have the edgemark 'tail' or 'head'. If on the edge between X and Y there is a head mark at X, then X is no ancestor of Y (or any selection variable) in the underlying DAG. If, on the other hand there is a tail mark at X, X is an ancestor of Y (or any selection variable). In Fig. 3(c) we show the MAG representing the DAG from Fig. 3(a), where we assume that A and B are latent. We can read off, that, for example, X is no ancestor of Y.

In the presence of latent variables, we reformulate problem 1 in the following way:

Problem 1': Given observational data, find the MAG representing the causal structure on the observed variables, or, if this is not possible, give a range of MAGs among which the true MAG is.

MAGs encode conditional independence information using *m-separation*, a

graphical criterion which is very closely related to d-separation for DAGs. As with DAGs, we usually cannot identify a single MAG given observational data. Rather, several MAGs which encode the same set of conditional independence statements form a Markov *equivalence class* represented by a partial ancestral graph (*PAG*; see Zhang (2008a)). A PAG is like a MAG with the only difference that some edge marks are unknown (and usually represented by circles). All MAGs in the equivalence class share the same skeleton and also all edge marks which are given in the PAG. For each unknown edge mark in the PAG, there is at least one MAG in the equivalence class with a tail mark and at least one MAG in the equivalence class with a head mark. The MAG shown in Fig. 3(c) is in the equivalence class that is represented by the PAG in Fig. 3(d). From this PAG, we can read off, that, for example, Y is no ancestor of X.

Finally, we'd like to mention an alternative way of dealing with hidden common causes in DAGs. We have seen in the previous paragraph that MAGs on the variables O can be used to represent the conditional independence constraints among the variables in O (using m-separation) of an underlying DAG with the variables O and an arbitrary set of hidden confounders. Thus, the MAG represents some aspects (conditional independencies) of the marginalized DAG. However, it is known, that there are more constraints on observable distributions that are implied by DAGs with hidden confounders, but cannot be formulated in terms of conditional independence. Examples of these additional constrains are Verma constraints (Verma and Pearl, 1991; Shpitser and Pearl, 2008b) and inequality constraints (see for example Evans (2012)). An object to deal with these additional constraints is an Acyclic Directed Mixed Graph (*ADMG*), which consists of directed and bidirected edges and has no directed cycles. These graphs are also called *latent projections* of the underlying DAG onto the observed variables. Richardson *et al.* (2012) propose a nested Markov theory which allows to read off both conditional independence and Verma constraints from ADMGs.

Note that although a MAG and a ADMG may look alike, they are interpreted in a different way.

2.3 Solving Problem 1 and Problem 2

There are two common approaches for solving Problem 1 and 1'.

In one approach ('search-and-score approach'), one first assumes a causal structure and specific functional restrictions (e.g. linear relations and independent Gaussian noise). Then, one optimizes some *score* (e.g. likelihood or BIC) given these restrictions. In the next step, the causal structure is changed and the new optimal score value is computed. This procedure is repeated for many causal structures. Finally, the causal structure with the best (optimized) score is returned.

The other approach focuses on constraints implied by a causal structure

('constraint based approach'). Constraint based learning starts with finding a list of invariance properties (e.g. conditional independencies) implied in the given observational data. Then, it rules out all causal structures which are incompatible with the list of invariance properties implied in the data. Ideally, one is left with only one compatible DAG, the true causal structure.

These (and further) approaches are discussed in more detail in chapter 3.

Approaches for solving Problem 2 are based on the *do-calculus*. When the true causal structure is given, the do-calculus can transform interventional distributions into expressions involving only standard statistical distributions. These approaches will be discussed in more detail in chapter 4.
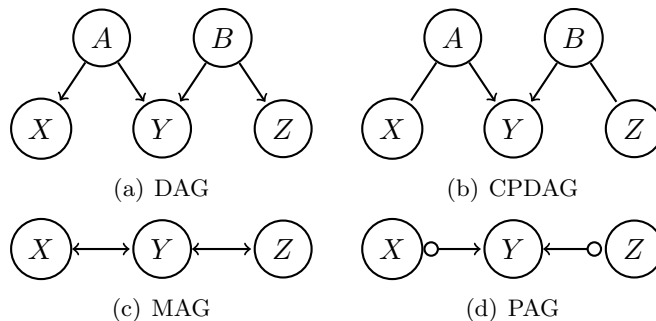

Figure 3: The DAG shown in (a) is in the equivalence class represented by the CPDAG shown in (b). The MAG shown in (c) represents the marginalized DAG from (a). Finally, the MAG shown in (c) is in the equivalence class represented by the PAG shown in (d).

3   Estimating Causal Structure

We will first assume no hidden confounders and will then relax this assumption.

3.1   No hidden confounders

A broad range of methods has been developed for estimating causal structures assuming no hidden confounders. We will first discuss methods using only observational data. Then, we will discuss methods which can handle both observational and interventional data.

3.1.1   Using observational data

PC-algorithm
The PC-algorithm (Spirtes *et al.*, 2000) is a popular algorithm for constraint

based structure learning. Its can be outlined in three steps. In the first step, the skeleton of the DAG is estimated. For doing this, the algorithm starts with a complete undirected graph (i.e., every node is connected to every other node). Then, for each edge (say, between X and Y) it is tested, whether there is any conditioning set S, so that X and Y are conditional independent given S (i.e. $X \perp\!\!\!\perp Y|S$). If such a set (called a *separation set*) is found, the edge between X and Y is deleted. The idea behind this step is that, if the dependence between X and Y can be explained away, there cannot be a direct causal connection between them and thus, there will be no edge between them in the true causal structure. The crucial idea of the PC-algorithm for this step is to arrange the tests in increasing size of adjacency sets of X and Y.

The separation sets are current adjacency sets of pairs of nodes and one tests sequentially, starting with small sets and gradually increasing their sizes. This ensures good statistical properties and efficient computation if the underlying DAG is large but sparse. In the second step, the information on separation sets is used to orient unshielded colliders. To understand why unshielded colliders can be oriented given a separation set (i.e., conditional independence) information, consider the following example. Suppose, we have a skeleton of the form $X - Z - Y$ where X and Y are not connected (i.e., the skeleton is an unshielded triplet). Thus, the true causal structure might be of the form $X \leftarrow Z \leftarrow Y$, $X \leftarrow Z \rightarrow Y$, $X \rightarrow Z \rightarrow Y$ or $X \rightarrow Z \leftarrow Y$. By applying d-separation, we see that $X \leftarrow Z \leftarrow Y$, $X \leftarrow Z \rightarrow Y$, $X \rightarrow Z \rightarrow Y$ imply that $X \perp\!\!\!\perp Y|Z$, whereas $X \rightarrow Z \leftarrow Y$ implies $X \not\!\perp\!\!\!\perp Y|Z$. Thus, if Z is not in the separation set of X and Y, we know that $X \not\!\perp\!\!\!\perp Y|Z$ holds and therefore the causal structure must be $X \rightarrow Z \leftarrow Y$, which is an unshielded collider. This example can be generalized to the rule, that whenever Z is not in the separation set of X and Y in an unshielded triple $X - Z - Y$, the unshielded triple must be directed into an unshielded collider. In the third step, all (still) undirected edges are checked, if one of the two possible directions would lead to a new v-structure or a cycle. If yes, the undirected edge is directed into the other direction to rule out a new v-structure (all of them were already found in step two) or a cycle (which is forbidden in a DAG).

While in the worst case the runtime of the PC algorithm is exponential in the number of variables, for sparse graphs it can handle up to thousands of variables in an acceptable amount of time. Robins *et al.* (2003) report on fundamental issues with uniform consistency in causal inference. However, Kalisch and Bühlmann (2007) show uniform consistency of the PC-algorithm under the strong faithfulness condition for high-dimensional but sparse graphs for linear structural equation models with Gaussian error terms. This result was extended to Gaussian copula models by Harris and Drton (2012). The assumption of faithfulness in this context is discussed in detail in Lin *et al.* (2012) and Uhler *et al.* (2013).

Score based methods

The Greedy-Equivalent-Search (GES) algorithm (Chickering, 2002, 2003) is a prominent example of a search-and-score algorithm. This algorithm scores the causal structure using a score-equivalent and decomposable score, such as the regularized log-likelihood, e.g. the BIC (Schwarz, 1978). A score is score-equivalent, if it assigns the same value to all DAGs within the same Markov equivalence class. This is a desirable property, since otherwise a haphazard DAG out of a selection of statistically indistinguishable DAGs would be preferred. A score is decomposable, if it can be computed as a sum of terms depending only on local features (typically for each node) of the causal structure. A decomposable score helps to speed up the computation of scores during the search process. A naive approach would now score every DAG on the given p variables and pick the DAG with the best score. However, since the number of DAGs on p variables grows super-exponentially in p (Robinson, 1977), this approach is not feasible. Silander and Myllymaki (2006) provide a dynamic programming approach which reduces the complexity to exponential time.

An improvement might be achieved by not searching over individual DAGs but equivalence classes of DAGs. Since the average number of DAGs per equivalence class seems to be 3.7 (Gillispie and Perlman, 2002; Garrido, 2009), such an improvement seems to be conceptually limited.

In view of the huge search space, greedy algorithms have been proposed. The idea of GES is to traverse the space of Markov equivalence classes step by step and can be outlined in two (or three) steps. The GES algorithm starts with an empty graph and then adds, in a first step, edges in a greedy way until an optimum is reached (called 'forward phase'). In each step, the edges are added in a way, so that the resulting DAG is a representative of a new equivalence class. Then, in a second step, edges are greedily removed until, again, an optimum is reached (called 'backward phase'). As before, edges are removed in a way, so that the resulting DAG is a representative of a new equivalence class. The resulting graph, representing the resulting equivalence class, is the final output of GES. The algorithm can be improved by including a turning phase of edges as well (Hauser and Bühlmann, 2012a).

The GES algorithm does not assume the faithfulness condition and consistency of GES in the low-dimensional case is given in Chickering (2003). An analysis of the penalized maximum likelihood estimator for sparse DAGs in high dimensions which also avoids the faithfulness assumption is presented in van de Geer and Bühlmann (2013).

The constraint-based and search-and-score methods mentioned in the previous paragraphs can be combined to form hybrid methods. A prominent example is the Max-Min Hill-Climbing (MMHC) algorithm of Tsamardinos *et al.* (2006). The idea of MMHC is to find the skeleton based on a constraint-based search and then use a search-and-score approach to orient the edges. In empirical studies, MMHC seems to outperform both PC

and GES (Tsamardinos *et al.*, 2006). In their empirical study, the authors used the Structural Hamming Distance (SHD) to assess the performance of several methods for structure learning. Recently, the Structural Intervention Distance (SID) was proposed by Peters and Bühlmann (2013b). When comparing causal structures in terms of their corresponding causal inference statements, the SID is more appropriate than SHD.

Methods for fully identifiable models
In general, the algorithms in the above mentioned classes (such as PC, GES and MMHC) return estimates of the equivalence class of the true causal structure based on observational data. It turns out, however, that under certain assumptions, the true causal structure can be recovered uniquely.

Peters *et al.* (2013) show that if the underlying causal structural equations are from an Additive Noise Model (ANM) with nonlinear functions and some technical assumptions hold, the true causal structure G can be identified uniquely from the probability distribution of the observational data. More precisely, an Additive Noise Model is of the form

$$x_i = f_i(pa_i) + n_i, \tag{2}$$

where pa_i are the parents of x_i in G and n_i is an independent, additive error variable. In particular, if the error variables are Gaussian, the technical assumptions rule out linear functions in the structural equation model. A notable improvement over previous assumptions is the fact, that the technical assumptions do not include faithfulness (but a weaker condition called causal minimality; a distribution is *causal minimal* with respect to a DAG, if it is Markov to the DAG but not to any subgraph). While the technical assumptions are weak in the two-dimensional case (Peters *et al.*, 2013, Proposition 15), it is not clear how restrictive they are when more than two variables are involved.

Shimizu *et al.* (2006) propose a method for uniquely identifying causal structures with the assumptions that the structural equation model is linear with non-Gaussian error terms (*LiNGAM*). Shimizu *et al.* (2011) propose an algorithmic improvement of their original method which was based on independent component analysis.

From the previous two paragraphs we see that unique identification from observational data is indeed possible. Given technical assumptions, the central theme is that as long as the data generating structural equation model is non-linear or the error distributions are non-Gaussian, unique identifiability is possible. This is in contrast to the situation where the structural equation model is linear with Gaussian errors. In this situation unique identifiability becomes in general impossible and only equivalence classes can be found. But even for linear Gaussian structural equation models, unique identifiability is possible when assuming equal error variances (Peters and Bühlmann, 2013a).

3.1.2 Mixed observational and interventional data

In the previous section, we considered approaches for finding the causal structure given observational data. It turned out that unique identification is possible given some assumptions. In practice, however, assumptions are hard to verify. To avoid assumptions as much as possible, one could therefore, on the other extreme, perform fully controlled intervention experiments. In an intervention experiment, fixed or random values are assigned to a single or several variables. Making, for example, a single variable intervention at each variable once would guarantee to find the causal structure.

A challenge, however, lies in combining both observational data and data from intervention experiments on parts of the causal system. The main reason for this is the fact that interventional samples from different interventions (or no intervention) are not identically distributed. A thorough framework dealing with this situation in the linear Gaussian case has been developed in Hauser and Bühlmann (2012a) and Hauser and Bühlmann (2013a). (While there exist also different approaches (Tian and Pearl, 2001), we will focus on the approach by Hauser and Bühlmann (2012a).) Note that it is precisely the linear Gaussian case in which unique identification is most difficult, which suggests, that the use of additional data from interventional experiments might bring the largest benefit here. In their work, the authors point out that even given interventional data, the causal structure is generally not fully identifiable. However, the interventional data help identifying at least parts of the causal structure, while other parts of the structure might remain ambiguous (and could be uniquely identified if results on different intervention were available). Therefore, the concept of Markov equivalence class must be extended to capture DAGs which are equivalent given observational data and data from specific interventions. The authors provide this extension in the form of the *interventional essential graph*, generalizing the CPDAG. Estimation and structure learning can be done with the maximum likelihood principle. Similar to GES for CPDAGs, they propose a greedy learning algorithm for a mix of observational and interventional data and call it 'greedy interventional equivalence search' (GIES). Like GES, GIES is based on a greedy optimization of the BIC-score greedily moving through the space of essential interventional graphs (encoding the interventional Markov equivalence classes) with repeated forward, backward and turning phases.

With GIES, the problem of combining observational and interventional data is solved satisfactorily for the linear Gaussian case. However, it might seem plausible to approach the identification problem sequentially: First, we acquire observational data. If the true causal structure is not uniquely identifiable, we could invest in specifically targeted intervention experiments. This approach is also known as 'active learning'. Hauser and Bühlmann (2012b) addresses this problem (allowing also for the situation with no observational data) and propose two solutions for any given interventional essential graph

that is not fully directed. In the first solution, the single variable intervention is found which, when analyzed in an intervention experiment, will direct as many undirected edges as possible. (The graph might still not be fully directed afterwards.) In the second solution, the multi variable intervention is found, that contains the minimal amount of variables necessary to achieve full identification when performing intervention experiments on these variables. This second solution also contains a proof of the Eberhardt conjecture dealing with the number of unbounded intervention targets which is sufficient and in the worst case necessary for full identifiability. (Eberhardt, 2008). Hauser and Bühlmann (2012b) demonstrate impressive performance of their algorithms in the case of negligible sampling error (e.g. large sample limit). However, when the sampling errors become large, the advantages of the proposed methods get blurred. He and Geng (2008) also propose single and multi variable intervention schemes which allow for full identification. However, their method uses (only) the observational data for estimating the CPDAG and (only) the interventional data for orienting edges. In contrast to this, Hauser and Bühlmann (2013b) make a more efficient use of the data, which leads to increased empirical performance.

Eberhardt *et al.* (2010) propose an algorithm for identifying cyclic linear causal models (in equilibrium) from a mix of observational and interventional data, which may also include latent variables. Although the authors propose a necessary and sufficient condition for the identifiability of the set of all total effects, they don't propose an optimal (in terms of minimal number of interventions) procedure for finding a set of variables to intervene on that guarantees full identifiability.

3.2   Hidden confounders

In the previous section, we have assumed that there are no hidden confounders or selection variables. In this section, we will discuss approaches for structural learning when hidden confounders and selection variables are assumed to be present.

Under the faithfulness assumption, the PAG on the observed variables representing a Markov equivalence class of DAGs with latent and selection variables can be learned from the observed variables using the FCI (Fast Causal Inference) algorithm (Spirtes *et al.*, 2000, 1995).

The idea of FCI can be outlined in five steps. In the first and second step, an initial skeleton and unshielded colliders are found as in the PC-algorithm. In the third step, a set called 'Possible-D-SEP' is computed. The edges of the initial skeleton are then tested for conditional independence given subsets of Possible-D-SEP. If conditional independencies are found, the conditioning sets are recorded as separation sets and the corresponding edges are removed (as in step 1 of PC-algorithm). Thus, edges of the initial skeleton might be removed and the list of all separation sets might get

extended. Since Possible-D-SEP can get very large even in sparse graphs, testing all subsets might be very time consuming. In step four, unshielded colliders in the updated skeleton are oriented based on the updated list of separation sets. In step five, further orientation rules are applied in order to avoid contradictions in the resulting PAG.

Originally, the output of FCI was a different object than a PAG. However, Zhang (2008a) showed that it can be interpreted as a PAG as well. Furthermore, Zhang (2008b) introduced extra orientation rules in FCI which make the algorithm complete, i.e., maximally informative.

One major drawback of FCI is that, despite its name, it suffers from an exponential time complexity (even when the underlying DAG is sparse). Therefore, a version called 'AnytimeFCI' was developed by Spirtes (2001), which allows a trade-off between computational speed and informativeness by setting an additional tuning parameter. Although the output of AnytimeFCI is no PAG anymore, it can still be interpreted causally.

Another improvement of FCI, the RFCI algorithm, was proposed by Colombo *et al.* (2012). The idea of RFCI is to avoid computing the potentially costly Possible-D-SEP and follows the same three steps as the PC-algorithm. However, some modifications of step two (unshielded colliders) and step three (further orientation) avoid some of the erroneous causal conclusions PC could make in the presence of hidden variables. The causal interpretation of RFCI is sound and consistent in high-dimensional, sparse settings, but slightly less informative than that of FCI. However, the computational speed is dramatically increased. Colombo *et al.* (2012) show in simulations that the estimation performances of FCI and RFCI are very similar.

An alternative improvement of FCI was given in Claassen *et al.* (2013). They propose the FCI+ algorithm, which is sound and complete and has polynomial complexity for underlying sparse DAGs. Note that the complexity of FCI is exponential even for sparse graphs.

We now turn away from estimating PAGs and discuss the estimation of ADMGs. A parameter fitting algorithm, and a search and score structure learning algorithm for these nested Markov models is given in Shpitser *et al.* (2012). As with DAGs, different ADMGs can be equivalent but have different causal interpretation. However, in contrast to DAGs and MAGs, there is currently no characterization of equivalent ADMGs known. Recall that with GES, we could find the equivalence class of all best-scoring DAGs. Since these DAGs have usually different causal interpretation, one could e.g. report summary statistics representing all DAGs in the equivalence class. However, while the search and score algorithm of Shpitser *et al.* (2012) might find several top-scoring ADMGs (which are equivalent, since they have the same score), it will in general not find *all* ADMGs which are equivalent (exhaustive search over all ADMGs would be computationally too expensive). Therefore, when giving summary statistics of the ADMGs

found, we can never be sure if we missed an equivalent ADMG with a completely different causal interpretation.

4   Estimating Causal Effect Strength

In the previous section, we have discussed methods for estimating the causal structure. The causal structure indicates which variable has a direct effect on which other variable. However, it does not indicate, how strong this effect is (or, for example, if it is positive or negative). In this section, we will review methods for estimating the strength of causal effects given data *and* a causal structure. In practice, the causal structure might be considered to be the true causal structure when, for example, sufficient background knowledge is available. However, oftentimes, the true causal structure is not known and it is estimated by methods discussed in the previous section.

The basic technical tool for quantifying the causal effect in structural equation models is the do-calculus (Pearl, 2009b). We denote the probability distribution of variable Y when an external intervention fixes variable X to the value x as $P(Y|do(X = x))$. While this expression looks very much like the conditional probability $P(Y|X = x)$, it is fundamentally different: Typically, $P(Y|do(X = x))$ corresponds to doing randomized experiments by intervening at X and then observing Y. However, $P(Y|X = x)$ has the interpretation of observing the system (without doing any intervention at all) and just noting the behavior of Y whenever we observe that X takes the value x. These two distributions can be fundamentally different, as is illustrated in the following example.

Suppose there is a street in a desert. The binary variable X indicates whether it is raining ($X = 1$ with $P(X = 1) = 0.01$, so rain is a rare event) or not ($X = 0$). The binary variable Y indicates whether the street is wet ($Y = 1$) or not ($Y = 0$). We suppose that the street gets wet when it is raining ($P(Y = 1|X = 1) = 1$) and that other reasons for a wet street are rare ($P(Y = 1|X = 0) = 0.001$). Now, let's suppose we observe that the street is wet. What is the probability that it is raining conditional on this observation? Using Bayes Theorem, we arrive at $P(X = 1|Y = 1) = 0.91$. So, if we see that the street is wet, we can be rather sure that it is raining. However, we know from background knowledge that if we intervene to make the street wet (e.g. with a water hose), this will have no effect on whether it is raining or not. Thus $P(X = 1|do(Y = 1))$ should be independent of the intervention: The probability of rain given we make the street wet is just the probability that it rained in that moment anyway. Thus, the probabilities under intervention and the conditional probability are indeed very different.

The previous example shows a very attractive goal. If we know the causal structure, it might be possible to transform interventional distributions into expressions involving standard statistical distributions. In the example, we

achieved this by setting $P(X = 1|do(Y = 1)) = P(X = 1)$. Thus, we are able to predict the distribution under intervention without doing any interventions at all but just evaluating certain standard statistical probabilities, which can be found by observation alone.

In the following, we will review ways of transforming interventional distributions into expressions involving standard statistical distributions.

4.1 No hidden confounders

Pearl (1995) proposed a set of three inference rules, which form the basis of the do-calculus (see also Pearl (2009b, Theorem 3.4.1)). These rules and any combination of them allow transformations of distributions under intervention into standard statistical distributions given the true causal structure. The set of rules was shown to be complete (Shpitser and Pearl, 2008a, Theorem 23), i.e., if an intervention distribution can be transformed into an expression containing only standard statistical distributions, there will be a sequence of these three rules that will make this transition.

In the case without hidden confounders, the rules of the do-calculus justify a particularly simple rule that is complete for transforming the interventional distribution into an expression of standard statistical distributions: The back-door criterion (Pearl, 2009b, Chapter 3.3.1). A set of variables B satisfies the back-door criterion relative to cause X and effect Y if in the causal structure B blocks every path between X and Y that is pointing into X and no node in B is a descendant of X. For example, in Fig. 2, the variable X satisfies the back-door criterion relative to Y and Z. With this definition, the interventional distribution can be transformed into an expression using only standard statistical distributions using the formula:

$$P(Y = y|do(X = x)) = \sum_B P(Y = y|X = x, B = b)P(B = b) \qquad (3)$$

Note, that in equation (3) on the left hand side is an interventional distribution while on the right hand side are only standard statistical distributions. Maathuis *et al.* (2009) propose a method that combines the estimation of the causal structure and the interventional distribution in the linear Gaussian case. They concentrate on the effect strength measured as the expected change in the response if the cause is changed by one unit. Using a local algorithm, the authors are able to compute the multiset of expected changes of all DAGs in the equivalence class. A summary statistics of this multiset can then be reported. The method, termed *IDA* for **I**ntervention calculus when the **D**AG is **A**bsent, was shown to perform well when compared with intervention experiments (Maathuis *et al.*, 2010). Further developments of this method were shown in Stekhoven *et al.* (2012) and Colombo and Maathuis (2012).

4.2 With hidden confounders

When hidden confounders are assumed to be present and the true causal structure is given in the form of an ADMG, Shpitser and Pearl (2008a) show that the do-calculus is still complete. The authors even give an algorithm (ID-algorithm) to find the explicit reformulation of the interventional distribution in terms of standard statistical distribution if such a transformation exists. An algorithmic improvement (EID-algorithm) was proposed in Shpitser *et al.* (2011).

If the true causal structure is given in the form of a MAG or PAG, the generalized backdoor criterion (Maathuis and Colombo, 2013) offers a sufficient (but not necessary) criterion for transforming the interventional distribution into an expression involving standard statistical distributions.

5 Software and Application

For the statistical software R (R Core Team, 2013) the package pcalg is available for causal inference (Kalisch *et al.*, 2012) (see also the vignette of the package for a detailed documentation of recent extensions of the package). The package contains implementations of PC, FCI, RFCI, GES, GIES and IDA. A detailed documentation (also of recent updates) can be found in the document accompanying the package ("vignette"). We demonstrate the application of this software in two recent examples.

In the first example, Bühlmann *et al.* (2013) make public and analyze a data set on vitamin (riboflavin) production in bacteria. There is a single real-valued response variable which is the logarithm of the riboflavin production rate; furthermore, there are $p = 4088$ (co-)variables measuring the logarithm of the expression level of 4088 genes. Using the R-package pcalg, the authors illustrate in detail how IDA can be applied to estimate bounds on the causal effect of a gene (in the example, the gene YCIC_at was chosen) on riboflavin production. It was found that an increase of the expression of YCIC_at by one unit leads to a change of the riboflavin production rate of at least 0.08 units.

In a second example, Colombo and Maathuis (2012) discuss the analysis of a data set on gene expression in yeast (extending previous analyses in Maathuis *et al.* (2010) and Stekhoven *et al.* (2012)) for experimental validation of causal inference methods. Observational data contained expression measurements of 5361 genes for 63 wild-type cultures. In addition to the purely observational data, the authors used interventional data on expression measurements of the same 5361 genes for 234 single-gene deletion mutant strains. The interventional data was used as the gold standard for estimating the total causal effect of the 234 deleted genes on the remaining genes (giving a total of $234 \cdot 5360$ effects) using the observational data. Then, the top 10% of these effects were defined as target set and it was evaluated
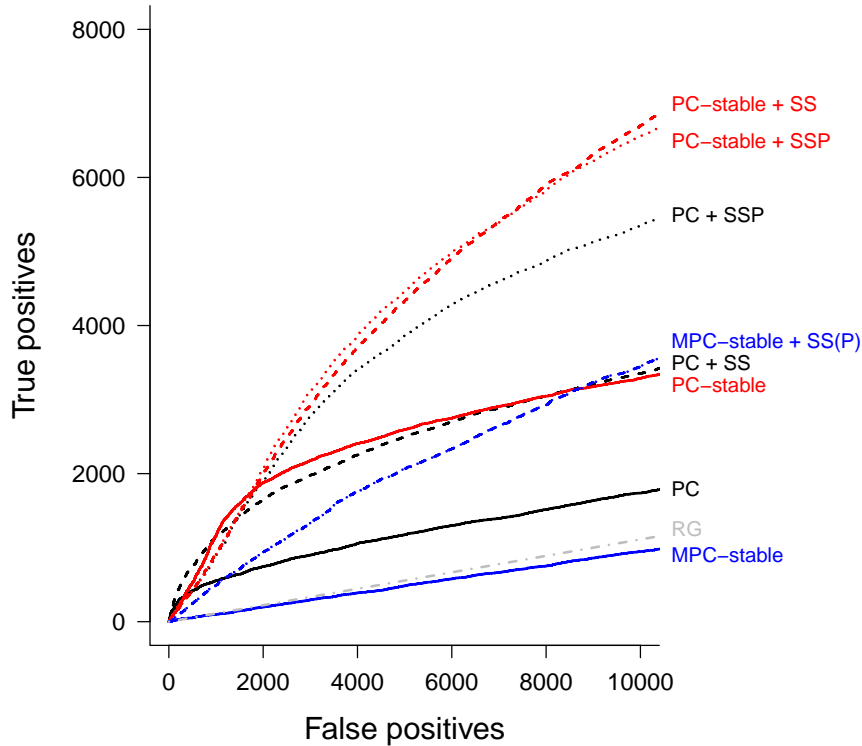
Figure 4: Experimental validation of causal inference methods using gene expression data in yeast. The dash-dotted line labelled with "RG" indicates the performance of random guessing. It is clear that random guessing is dramatically outperformed by a wide range of proposed approaches. This figure is taken from Colombo and Maathuis (2012) with the permission of the authors.

how well IDA could identify these effects from the observational data. For the structure learning in IDA, different variations of PC were proposed by the authors. The results are shown in a ROC curve in Fig. 4. The grey dash-dotted line labelled with "RG" indicates the performance of random guessing. The black solid line labelled with "PC" represents the method of Maathuis *et al.* (2010). The black dashed line labelled "PC + SS" represents the method of Stekhoven *et al.* (2012). The remaining lines correspond to variations of the PC algorithm proposed in Colombo and Maathuis (2012). It is clear that random guessing is dramatically outperformed by a wide range of proposed approaches.

6 Conclusion

In this paper we reviewed substantial progress on causal inference in the past two decades. Under some assumptions it is indeed possible to estimate the causal structure given observational data. While in some settings the result may not be unique (Markov equivalence), we discussed methods for full identifiability and also incorporation of interventional data (perhaps in an active way) to achieve full identifiability. We also showed that the interventional distribution can be computed when the true causal structure is known or estimated. This also holds true if we allow for hidden variables.

We close by highlighting three open problems for future research, whose solutions would enable to generalize IDA to other scenarios relevant for many practical applications.

The generalized backdoor criterion only gives a sufficient criterion for identifying the interventional distribution. Thus, if the criterion fails to identify we don't know whether there is no translation of the interventional distribution into an expression involving only standard statistical distributions or if we used the wrong criterion to find it. It would therefore be desirable to extend the generalized backdoor criterion so that it is also necessary.

The approach to causal structure learning based on ADMGs seems to be attractive: One can incorporate Verma and inequality constraints and if the structure was known, one could use complete identification methods (ID) for expressing interventional distributions in terms of standard statistical distributions. However, the estimation of ADMGs is not easy at the moment. This seems to be in part because it is not clear how to characterize equivalent ADMGs. It would therefore be desirable to develop a characterization of equivalent ADMGs.

The use of interventional data in order to support observational data helps to identify unique causal structures. However, methods with this intent only exist when no hidden variables are present. It would therefore be desirable to develop methods for mixed data (including active learning) when hidden variables may be present.

References

Andersson, S. A., Madigan, D. and Perlman, M. D. (1997) A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, **25**, 505–541.

Bühlmann, P., Kalisch, M. and Meier, L. (2013) High-dimensional statistics with a view towards applications in biology. *To appear in Annual Review of Statistics and its Applications.*

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

Chickering, D. M. (2003) Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, **3**, 507–554.

Claassen, T., Mooij, J. M. and Heskes, T. (2013) Learning sparse causal models is not NP-hard. In *Proceedings of the 29th conference on Uncertainty in artificial intelligence*, 172–181.

Colombo, D. and Maathuis, M. H. (2012) A modification of the PC algorithm yielding order-independent skeletons. *arXiv preprint arXiv:1211.3295*.

Colombo, D., Maathuis, M. H., Kalisch, M. and Richardson, T. S. (2012) Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, **40**, 294–321.

Cox, D. R. (1958) *Planning of experiments.* Wiley.

Eberhardt, F. (2008) Almost optimal intervention sets for causal discovery. In *Proceedings of the 24th conference on Uncertainty in artificial intelligence*, 161–168.

Eberhardt, F., Hoyer, P. O. and Scheines, R. (2010) Combining experiments to discover linear cyclic models with latent variables. In *International Conference on Artificial Intelligence and Statistics*, 185–192.

Evans, R. J. (2012) Graphical methods for inequality constraints in marginalized DAGs. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, 1–6. IEEE.

Fisher, R. (1926) The arrangement of field experiments. *Journal of Ministry of Agriculture of Great Britain*, **33**, 503–513.

Garrido, A. (2009) Asymptotic behaviour of essential graphs. *Advanced Modeling and Optimization*, **11**.

Gillispie, S. B. and Perlman, M. D. (2002) The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, **141**, 137–155.

Harris, N. and Drton, M. (2012) PC algorithm for Gaussian copula graphical models. *arXiv preprint arXiv:1207.0242*.

Hauser, A. and Bühlmann, P. (2012a) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, **13**, 2409–2464.

Hauser, A. and Bühlmann, P. (2012b) Two optimal strategies for active learning of causal models from interventions. In *Proc. of the 6th European Workshop on Probabilistic Graphical Models (PGM 2012)*, 123–130.

Hauser, A. and Bühlmann, P. (2013a) Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *arXiv preprint arXiv:1303.3216*.

Hauser, A. and Bühlmann, P. (2013b) Two optimal strategies for active learning of causal models from interventional data. *submitted*.

He, Y.-B. and Geng, Z. (2008) Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, **9**, 2523–2547.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, **8**, 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. and Bühlmann, P. (2012) Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47**, 1–26. URL http://www.jstatsoft.org/v47/i11/.

Lauritzen, S. L. (1996) *Graphical models*. Oxford University Press.

Lin, S., Uhler, C., Sturmfels, B. and Bühlmann, P. (2012) Hypersurfaces and their singularities in partial correlation testing. *arXiv preprint arXiv:1209.0285*.

Maathuis, M. H. and Colombo, D. (2013) A generalized backdoor criterion. *arXiv preprint arXiv:1307.5636*.

Maathuis, M. H., Colombo, D., Kalisch, M. and Bühlmann, P. (2010) Predicting causal effects in large-scale systems from observational data. *Nature Methods*, **7**, 247–248.

Maathuis, M. H., Kalisch, M. and Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, **37**, 3133–3164.

Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika*, **82**, 669–688.

Pearl, J. (2009a) Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.

Pearl, J. (2009b) *Causality: models, reasoning and inference*. Cambridge Univ Press, 2nd edn.

Peters, J. and Bühlmann, P. (2013a) Identifiability of Gaussian structural equation models with equal error variances. *To appear in Biometrika.*

Peters, J. and Bühlmann, P. (2013b) Structural intervention distance (sid) for evaluating causal graphs. *arXiv preprint arXiv:1306.1043.*

Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2013) Causal discovery with continuous additive noise models. *arXiv preprint arXiv:1309.6779.*

R Core Team (2013) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Richardson, T. and Spirtes, P. (2002) Ancestral graph Markov models. *The Annals of Statistics*, **30**, 962–1030.

Richardson, T. S., Robins, J. M. and Shpitser, I. (2012) Nested Markov properties for acyclic directed mixed graphs. In *Proceedings of the 28th conference on Uncertainty in artificial intelligence.*

Robins, J. M., Scheines, R., Spirtes, P. and Wasserman, L. (2003) Uniform consistency in causal inference. *Biometrika*, **90**, 491–515.

Robinson, R. W. (1977) Counting unlabeled acyclic digraphs. In *Combinatorial mathematics V*, 28–43. Springer.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

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

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O. and Bollen, K. (2011) Directlingam: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, **12**, 1225–1248.

Shpitser, I. and Pearl, J. (2008a) Complete identification methods for the causal hierarchy. *The Journal of Machine Learning Research*, **9**, 1941–1979.

Shpitser, I. and Pearl, J. (2008b) Dormant independence. In *Proceedings of the Twenty-third AAAI Conference on Artificial Intelligence*, 1081–1087.

Shpitser, I., Richardson, T. S. and Robins, J. M. (2011) An efficient algorithm for computing interventional distributions in latent variable causal

models. In *Proceedings of the 27th conference on Uncertainty in artificial intelligence*.

Shpitser, I., Richardson, T. S., Robins, J. M. and Evans, R. (2012) Parameter and structure learning in nested Markov models. *arXiv preprint arXiv:1207.5058*.

Silander, T. and Myllymaki, P. (2006) A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the 22nd conference on Uncertainty in artificial intelligence*, 445–452.

Spirtes, P. (2001) An anytime algorithm for causal inference. In *Proc. of International Workshop on Artificial Intelligence and Statistics*.

Spirtes, P. (2010) Introduction to causal inference. *The Journal of Machine Learning Research*, **99**, 1643–1662.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, prediction, and search*, vol. 81. The MIT Press.

Spirtes, P., Meek, C. and Richardson, T. (1995) Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 499–506.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H. and Bühlmann, P. (2012) Causal stability ranking. *Bioinformatics*, **28**, 2819–2823.

Tian, J. and Pearl, J. (2001) Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 512–521.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006) The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, **65**, 31–78.

Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2013) Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, **41**, 436–463.

van de Geer, S. and Bühlmann, P. (2013) ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, **41**, 536–567.

Verma, T. and Pearl, J. (1990) On the equivalence of causal models. In *Proceedings of the Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, 220–227.

Verma, T. and Pearl, J. (1991) Equivalence and synthesis of causal models (technical report). *Cognitive Systems Laboratory, University of California at Los Angeles.*

Wright, S. (1921) Correlation and causation. *Journal of agricultural research*, **20**, 557–585.

Zhang, J. (2008a) Causal reasoning with ancestral graphs. *The Journal of Machine Learning Research*, **9**, 1437–1474.

Zhang, J. (2008b) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, **172**, 1873–1896.