Behaviormetrika
Vol.41, No.1, 2014, 41–64

Invited paper

LEARNING CAUSAL STRUCTURE FROM MULTIPLE DATASETS WITH SIMILAR VARIABLE SETS

Robert E. Tillman* and Frederick Eberhardt**

While randomized controlled experiments are often considered the gold standard for predicting causal relationships between variables, they are expensive if one is interested in understanding the complete set of causal relationships governing a large set of variables and it may not be possible to manipulate certain variables due to ethical or practical constraints. To address these scenarios, procedures have been developed which use conditional independence relationships among variables when they are passively observed to predict which variables may or may not be causally related to other variables. Until recently, most of these procedures assumed that the data consisted of a single i.i.d. dataset of observations, but in practice researchers often have access to multiple similar datasets, e.g. from multiple labs studying the same problem, which measure slightly different variable sets and where recording conventions and procedures may vary. This paper discusses recent state of the art approaches for predicting causal relationships using multiple observational and experimental datasets in these contexts.

1. Introduction

In many domains researchers are interested in understanding the causal relationships which govern a set of variables in order to better understand the underlying systems governing those variables and to predict the effects of interventions on those variables. While randomized controlled experiments are often considered the gold standard for predicting causal relationships between variables, these methods are expensive if one is interested in understanding many different potential causal relationships among a large set of variables, and it is often not possible to manipulate certain variables due to ethical or practical constraints. For these scenarios researchers have developed methods which consider the conditional independence relations between variables when they are passively observed, or equivalently, how the joint distribution among those variables can be factored, to establish which variables may or may not be causally related and to predict the results of potential interventions on those variables (Spirtes et al., 2000; Pearl, 2000; Chickering, 2002; Tsamardinos et al., 2006).

Most of these algorithms assume that the data are collected in a single dataset, possibly with some values that are missing at random. While the availability of large amounts of data continues to increase, researchers are not always able to obtain a single dataset measuring every variable of interest; it is often, however, the case that several datasets are available which each measure subsets of these variables. For example, models of the United States and United Kingdom economies share some but not all variables, due to different financial recording conventions; fMRI studies with

Key Words and Phrases: Causal inference, Structure learning, Multiple datasets, Overlapping variables, Missing data

* Blander Technologies, Chicago, IL, United States. E-mail: robert.e.tillman@gmail.com

** California Institute of Technology, Pasadena, CA, United States. E-mail: fde@caltech.edu

similar stimuli may record different variables, since the images vary according to magnet strength, data reduction procedures, etc., and U.S. states report some of the same educational testing variables, but also report state-specific variables (Tillman et al., 2009). Similarly, even if a researcher is able to obtain a single dataset which measures every variable of interest, he or she may wish to combine this dataset with several similar datasets which measure slightly different variable sets to produce a more reliable result. Recently, newer algorithms have been developed for these scenarios which predict causal relationships from multiple datasets which measure similar variable sets when it is assumed that the underlying causal relationships among variables are the same across datasets (Tillman, 2009; Ramsey et al., 2010; Tillman et al., 2009; Triantafillou et al., 2010; Tillman and Spirtes, 2011; Eberhardt et al., 2010; Claassen and Heskes, 2010; Hyttinen et al., 2013b). This paper discusses recent state of the art approaches to these problems and their practical performance and limitations.

The algorithms mentioned above use directed graphs to depict causal relationships among a set of variables. Section 2 first provides the necessary formal background and notation to understand these algorithms and the results they output. Section 3 then describes some of the traditional algorithms for predicting causal relationships from a single observational dataset. In Section 4, we discuss the case where a single dataset measuring every variable of interest exists, but other datasets, which may or may not measure all of these variables also exist and the researcher would like to take advantage of this additional information. In Section 5 we then discuss the harder problem where multiple datasets are available which measure similar variable sets, but no dataset exists which measures the complete set of variables of interest. Section 6 then discusses procedures that can be used when multiple datasets are available where some variables have been experimentally manipulated and others are passively observed. Finally, Section 7 offers conclusions and directions for future research.

2. Background

We now introduce some terminology. A *directed graph* $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a set of nodes $\mathcal{V}$ and a set of directed edges $\mathcal{E}$ connecting distinct nodes. Directed graphs are useful in the context of predicting causal relationships among a set of variables because each variable can be considered a node in a directed graph where a directed edge from X into Y indicates that X is a direct cause of Y (relative to the set of nodes $\mathcal{V}$). If two nodes are connected by an edge then the nodes are *adjacent*. For pairs of nodes $\{X, Y\} \subseteq \mathcal{V}$, X is a parent of Y and Y is a child of X if there is a directed edge from X to Y in $\mathcal{E}$. A *trail* in $\mathcal{G}$ is a sequence of nodes such that each consecutive pair of nodes in the sequence is adjacent in $\mathcal{G}$ and no node appears more than once in the sequence. A trail is a *directed path* if every edge between consecutive pairs of nodes points in the same direction. If there is a directed path from X to Y (all edges point towards Y), then X is an *ancestor* of Y and Y is a *descendant* of X. $\mathbf{Adj}_{\mathcal{G}}^{X}$, $\mathbf{Pa}_{\mathcal{G}}^{X}$, $\mathbf{Ch}_{\mathcal{G}}^{X}$, $\mathbf{An}_{\mathcal{G}}^{X}$, and $\mathbf{De}_{\mathcal{G}}^{X}$ refer to the sets of only those nodes that are adjacencies, parents, children, ancestors, and descendants of X in $\mathcal{G}$, respectively. $\mathcal{G}$ contains a

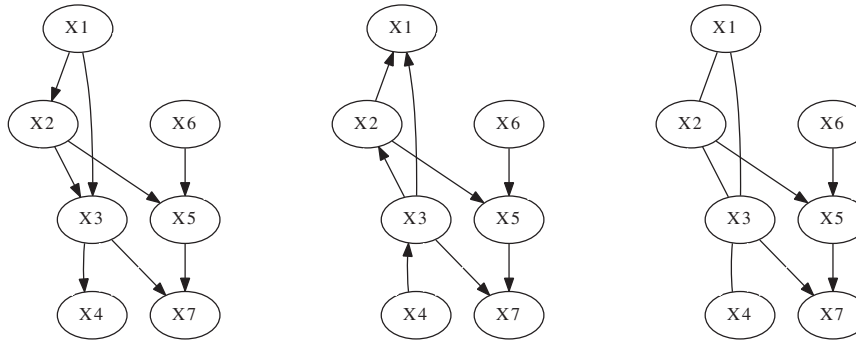

Figure 1: Two Markov equivalent DAGs (a) and (b) and their corresponding PDAG

directed cycle between two nodes X and Y if $X \rightarrow Y$ is an edge in $\mathcal{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}^{X}$. $\mathcal{G}$ is a *directed acyclic graph* (DAG) if it does not contain any directed cycles between any pair of distinct nodes. Figure 1 (a) and (b) show examples of DAGs.

Directed graphs are also useful in the context of predicting causal relationships because we can relate graph theoretic properties to properties of an associated joint probability distribution. A joint probability distribution $\mathcal{P}$ over variables represented as nodes in $\mathcal{V}$ obeys the *local Markov property* with respect to a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ if every variable is conditionally independent of its nondescendents in $\mathcal{G}$ given its parents in $\mathcal{G}$. The local Markov property is equivalent to a *global Markov property* (Lauritzen et al., 1990) which relates more general conditional independence relationships among variables represented as nodes in the graph to a graph theoretic criterion known as *d-separation*, which we define below. A *v-structure* (*collider*) is a triple of nodes $\langle X, Y, Z \rangle$ such that X and Z are parents of Y. If X and Y are not adjacent in a v-structure $\langle X, Y, Z \rangle$, then $\langle X, Y, Z \rangle$ is an *immorality (unshielded collider)*.

Definition 2.1 (Active trail)**.** A trail $\langle V_1, \ldots, V_n \rangle$ in a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is *active* given $\mathbf{Z} \subseteq \mathcal{V} \backslash \{V_1, V_n\}$ if the following hold for $1 < i < n$

(i) if $V_i \in \mathbf{Z}$, then $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, and

(ii) if $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, then either $V_i \in \mathbf{Z}$ or there exists some $X \in \mathbf{De}_{\mathcal{G}}^{V_i}$ such that $X \in \mathbf{Z}$.

Definition 2.2 (D-separation)**.** For disjoint sets of nodes $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we say that $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* given $\mathbf{Z}$, denoted $dsep_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$, if and only if there are no active trails between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given $\mathbf{Z}$. Otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are *d-connected* given $\mathbf{Z}$, denoted $\neg dsep_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$.

The global Markov property states that if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{G}$, then $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{P}$ (denoted as $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$) (Pearl, 1988). The local and global Markov properties are equivalent to the following factorization of the joint distribution $\mathcal{P}$.

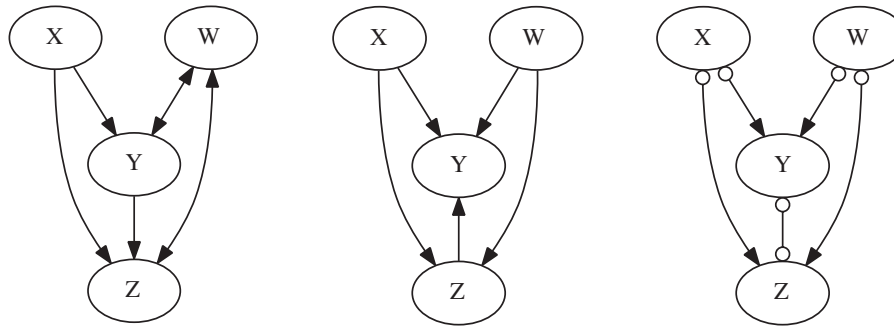

Figure 2: Two Markov equivalent MAGs (a) and (b) and their corresponding PAG. Note that (b) is both a MAG and a DAG.

$$\mathbb{P}(\mathcal{V}) = \prod_{X \in \mathcal{V}} \mathbb{P}\left(X | \mathbf{Pa}_{\mathcal{G}}^{X}\right) \tag{1}$$

A related property is *faithfulness*, which is the converse of the global Markov property. he faithfulness property holds when all conditional independences observed in the joint distribution $\mathcal{P}$ are entailed by the global Markov property, i.e. for disjoint sets of nodes, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, it holds that $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{G}$ whenever $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{P}$ (Spirtes et al., 2000). Meek (1995b) shows that the faithfulness property holds in almost all multinomial distributions which obey the Markov properties, i.e. the set of unfaithful distributions has Lebesgue measure 0. When the Markov and faithfulness properties both hold for $\mathcal{G}$, conditional independence and d-separation are equivalent.

While DAGs provide a simple representation for causal relationships among variables which are observed, there is no straightforward way of using DAGs to represent dependencies among variables which are due to unmeasured confounding variables, i.e. an unmeasured variable U which is a cause of two or more measured variables X and Y. In general it may be wise to consider that such dependencies occur, since important variables may not have been measured, but such dependencies will almost certainly be present when dealing with multiple datasets where only a subset of some variables of interest are measured in each dataset. We thus need a richer type of graphical structure to represent these dependencies for instances where we do not assume the absence of unobserved confounding variables. *Maximal ancestral graphs (MAGs)* are *mixed graphs*, i.e. consisting or directed, undirected, and bidirected edges, which can be used to represent causal relationships among a set of variables and dependencies due to unobserved confounding variables. They are a natural extension of DAGs, which are special cases of MAGs where all edges are directed. Bidirected edges in MAGs indicate that the corresponding nodes have an unobserved common cause. Undirected edges are used to indicate that the corresponding nodes have an association due to the presence of sample selection bias[1)]. Figure 2 (a) and (b) show

[1)] See Zhang (2007) for an example, though we will not consider sample selection bias in detail in this

examples of MAGs. Below, we define MAGs formally. A mixed graph $\mathcal{G}$ contains an *almost directed cycle* between two nodes X and Y if $X \leftrightarrow Y$ is in $\mathcal{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}^{X}$. X and Y are *spouses* in $\mathcal{G}$ if they are connected by a bidirected edge.

Definition 2.3 (Inducing path)**.** A trail $\langle V_1, \ldots, V_n \rangle \subseteq \mathcal{V}$ in a MAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is an *inducing path* between $\mathcal{V}_1$ and $\mathcal{V}_n$ relative to $\mathbf{Z} \subseteq \mathcal{V}$ if the following hold for $1 < i < n$

(i) if $V_i \notin \mathbf{Z}$, then $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, and

(ii) if $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, then $V_i \in \mathbf{An}_{\mathcal{G}}^{V_1} \cup \mathbf{An}_{\mathcal{G}}^{V_n}$.

Definition 2.4 (Maximal ancestral graph (MAG))**.** A mixed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a *maximal ancestral graph (MAG)* if the following hold

(i) $\mathcal{G}$ does not contain any directed cycles or almost directed cycles,

(ii) for any undirected edge $X - Y$ in $\mathcal{G}$, X and Y have no parents or spouses, and

(iii) for distinct pairs of nodes $\{X, Y\} \subseteq \mathcal{V}$, if X and Y are not adjacent in $\mathcal{G}$, then $\mathcal{G}$ contains no inducing paths between X and Y with respect to $\varnothing$.

The first two conditions in the above definition simply extend the acyclicity property of DAGs to MAGs. The third condition ensures that MAGs have a criterion similar to d-separation connecting their topology to individual conditional independence relations, which we define below.

Definition 2.5 (Active trail)**.** A trail $\langle V_1, \ldots, V_n \rangle$ in a MAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is *active* given $\mathbf{Z} \subseteq \mathcal{V} \backslash \{V_1, V_n\}$ if the following hold for $1 < i < n$

(i) if $V_i \in \mathbf{Z}$, then $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, and

(ii) if $\langle V_{i-1}, V_i, V_{i+1} \rangle$ is a v-structure, then either $V_i \in \mathbf{Z}$ or there exists some $X \in \mathbf{De}_{\mathcal{G}}^{V_i}$ such that $X \in \mathbf{Z}$.

Definition 2.6 (M-separation)**.** For disjoint sets of nodes $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we say that $\mathbf{X}$ and $\mathbf{Y}$ are *m-separated* given $\mathbf{Z}$, denoted $msep_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$, if and only if there are no active trails between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given $\mathbf{Z}$. Otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are *m-connected* given $\mathbf{Z}$, denoted $\neg msep_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$.

If a MAG consists of only directed edges (it is a DAG), then m-separation reduces to d-separation. A MAG essentially represents the collection of DAGs over the observed and unobserved variables represented by the MAG which have the same d-separation and ancestral relations among the observed variables (Zhang, 2008).

3. Traditional algorithms for learning causal structure from a single observational dataset

Since d-separation/m-separation and conditional independence are equivalent when

paper.

the Markov and faithfulness properties hold, many algorithms for predicting causal relationships attempt to search for a graph where either the d-separation/m-separation relationships are consistent with the conditional independences observed in the data or the factorization of the joint probability distribution according to the graph topology given in Section 2 maximizes a score function based on the penalized data log-likelihood. Algorithms exist which are asymptotically guaranteed to return a graph which is consistent with the true conditional independences and dependences under different assumptions. Such graphs, however, are not, in general, unique. Two graphs which entail exactly the same sets of conditional independences are said to be *Markov equivalent*. The goal of algorithms which employ these types of methods to predict causal relationships is thus to learn the set of graphs that are Markov equivalent for the observed conditional independences, i.e. the *Markov equivalence class*, and identify features that are common across all of these graphs to predict causal relationships.

One of the first and most straightforward of these algorithms is the SGS (Spirtes et al., 2000) or IC (Pearl, 2000) algorithm, which will provably (in the large sample limit) return a set of DAGs which contains the true causal structure whenever (i) the Markov and faithfulness properties hold, (ii) there are no unobserved variables which are common causes of observed variables, and (iii) there are no directed cycles in the true causal structure, i.e. it is a DAG. The returned set is the smallest set that can be distinguished using conditional independence information alone. In practice, rather than returning a set of DAGs, it is possible to use a more compact representation of this set, referred to as a *PDAG*, since all DAGs in the same Markov equivalence class have the same adjacencies and immoralities (Verma and Pearl, 1990). A PDAG is a mixed graph consisting of directed and undirected edges where a directed edge $X \rightarrow Y$ indicates that X directly causes Y and an undirected edge $X - Y$ indicates that either X causes Y or Y causes X (not to be confused with an undirected edge of a MAG!). Figure 1 shows two Markov equivalent DAGs and the corresponding PDAG.

The SGS/IC algorithm learns a PDAG from data by performing all possible conditional independence tests among all observed variables. Algorithm 1 formally describes the SGS/IC algorithm. The algorithm begins with the undirected complete graph, where all variables are adjacent to each other via an undirected edge. The results of conditional independence tests are then used to discover the correct set of adjacencies and immoralities (which defines the Markov equivalence class) and a set of correct and complete orientation rules described in Meek (1995a) are applied to make some additional edge orientations, e.g. to ensure no additional immoralities can be formed by orienting undirected edges in the graph. This algorithm is not practical, however, because it is both computationally intractable, since the number of conditional independence tests that must be performed is exponential in the number of variables, and not statistically robust, since conditional independence tests become increasingly unreliable as the size of the conditioning set grows. Thus, the current state of the art structure learning algorithms based on conditional independence tests try to limit the number of conditional independence tests that are performed and specif-

ically avoid tests with large conditioning sets. In the PC algorithm (Spirtes et al., 2000), instead of searching all subsets of $\mathbf{V}\backslash\{V_i, V_j\}$ for an $\mathbf{S}$ such that $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$, the algorithm initially sets $\mathbf{S} = \varnothing$ for all $\{V_i, V_j\}$ pairs, and checks to see if any edges can be removed based on the results of conditional independence tests with these $\mathbf{S}$ sets. It then iteratively increases the cardinality of sets $\mathbf{S}$ considered in the checks for edge-removal until there are no nodes that are adjacent to a greater number of nodes than the current cardinality of $\mathbf{S}$ sets considered. Additionally, the only $\mathbf{S}$ sets that are considered are those consisting of subsets of nodes adjacent to V_i or V_j at the current iteration of the algorithm. The PC algorithm learns the correct PDAG in the large sample limit when the Markov and faithfulness conditions hold, there are no unobserved common causes, and there are no directed cycles in the true causal structure. Algorithms which use conditional independence tests to learn graphs whose structure predicts causal relationships, such as SGS/IC and PC, are referred to as *constraint-based* causal structure learning algorithms.

Input : Observed data for variables in $\mathbf{V}$
Output: PDAG $\mathcal{G}$ over nodes $\mathbf{V}$

1 $\mathcal{G} \leftarrow$ the complete undirected graph over the variables in $\mathbf{V}$
2 For $\{V_i, V_j\} \subseteq \mathbf{V}$, if $\exists \mathbf{S} \subseteq \mathbf{V}\backslash\{V_i, V_j\}$, such that $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$, remove the edge between V_i and V_j
3 For $\{V_i, V_j, V_k\} \subseteq \mathbf{V}$ such that V_i and V_j are adjacent and V_j and V_k are adjacent, but V_i and V_k are not adjacent, if $\nexists \mathbf{S} \subseteq \mathbf{V}\backslash\{V_i, V_j, V_k\}$, such that $V_i \perp\!\!\!\perp V_k \mid \{\mathbf{S} \cup V_j\}$, orient the edges between V_i and V_j and between V_j and V_k to make $\langle V_i, V_j, V_k\rangle$ an immorality
4 Make any additional edge orientation to prevent additional immoralities and directed cycles from being formed when orienting undirected edges using the rules given in Meek (1995a)

Algorithm 1: SGS/IC algorithm

The FCI algorithm (Spirtes et al., 1995) is a constraint-based algorithm similar to the PC algorithm, but does not require assuming that there are no unobserved common causes. The FCI algorithm returns an equivalence class of MAGs instead of an equivalence class of DAGs. The true causal structure, represented as a MAG, will provably be included in the output set of FCI (in the large sample limit) whenever the Markov and faithfulness properties hold and there are no directed cycles in the true causal structure. Due to a result from Zhang (2007) characterizing the structural features of Markov equivalence classes of MAGs, a compact structure referred to as a *partial ancestral graph (PAG)* (Richardson and Spirtes, 2002; Zhang, 2007) can be used to represent the graphs output by FCI. PAGs are mixed graphs with a third type of edge endpoint, $\circ$. Whenever an edge has a $\circ$-marked endpoint this indicates that there is at least one MAG in the Markov equivalence class that has an arrowhead at that endpoint and at least one such MAG that has a tail at that endpoint (Zhang, 2007). Figure 2 shows two Markov equivalent MAGs and the corresponding PAG.

Score-based algorithms take a different approach to learning causal structure.

Score-based algorithms consider the factorization of the joint probability distribution equivalent to the Markov property given in Section 2 and search for a graph which maximizes some score function (usually a penalized data log-likelihood based score) when the score is decomposed according to the graph structure using this factorization. The Greedy Equivalence Search (GES) algorithm (Chickering, 2002) is a score based algorithm for learning causal structure that searches over the space of PDAGs. GES learns the correct PDAG in the large sample limit when the Markov and faithfulness properties hold, there are no unobserved common causes, the true causal structure does not contain directed cycles, and a score function is used, which has a property referred to as *local consistency* which we define below.

Definition 3.1 (Locally consistent score)**.** Let $S(\mathcal{G}, \mathcal{P})$ be the score assigned to the factorization of the joint probability distribution $\mathcal{P}$ entailed by some DAG $\mathcal{G}$ and let $\mathcal{H}$ be a DAG which does not entail the correct factorization of the joint probability distribution. S is *locally consistent* if it has the following two properties:

(i) Let $\mathcal{H}'$ be the result of adding an edge to $\mathcal{H}$ which eliminates an independence constraint entailed by $\mathcal{H}$ that is not true in $\mathcal{P}$. Then $S(\mathcal{H}', \mathcal{P}) > S(\mathcal{H}, \mathcal{P})$.

(ii) Let $\mathcal{H}'$ be the result of adding an edge to $\mathcal{H}$ which eliminates an independence constraint entailed by $\mathcal{H}$ that is true in $\mathcal{P}$. Then $S(\mathcal{H}', \mathcal{P}) < S(\mathcal{H}, \mathcal{P})$.

The BDeu and BIC scores are two popular locally consistent scores for discrete multinomial and multivariate Gaussian distributed data, respectively (Heckerman et al., 1995). The GES algorithm is a two stage greedy search procedure. In the first stage, edges are iteratively added to the empty graph and certain types of edges are reversed to increase the current score until doing so no longer increases the score. In the second stage, edges are then iteratively removed and certain edges are reversed to increase the current score until doing so no longer increases the score. In the large sample limit this greedy approach is guaranteed to return the highest scoring PDAG. In general, score-based structure learning is an NP-complete problem (Chickering, 1995), but for most problems with up to 50 variables, this does not seem to be an issue for GES in practice.

Instead of a greedy approach, more recent score based procedures based on the dynamic programming approach described in Koivisto and Sood (2004) are exact: They return a DAG that maximizes the posterior probability, from which the corresponding PDAG can be determined easily. Their worst-case runtime grows "only" exponentially in the number of variables (Koivisto and Sood, 2004; Silander and Myllymaki, 2006; Parviainen and Koivisto, 2009). Unlike GES, however, these methods are not "anytime" algorithms, i.e. they only produce an optimal structure once the search has been completed, while GES can be stopped at any point to obtain the highest scoring structure found so far at that point in the search. de Campos et al. (2009) proposes a branch and bound based score-based learning procedure that is both an anytime algorithm and has the same optimality guarantees as Koivisto and Sood (2004) and

the related dynamic programming approaches. Jaakkola et al. (2010) proposes a similar approach with both of these properties which searches for the optimal structure by solving a series of linear programs.

A significant disadvantage of score-based causal inference procedures is that they require the assumption that there are no unobserved confounding variables; there is no score-based equivalent of the constraint-based FCI algorithm. While it is possible to score MAGs when data have multivariate Gaussian distributions (Silva and Ghahramani, 2009), there are a number of issues which make score-based searches for PAGs more difficult than for PDAGs (Richardson et al., 1999; Claassen and Heskes, 2012). We know of no general consistent score-based algorithms for learning PAGs.

The Max-Min Hill-Climbing (MMHC) algorithm (Tsamardinos et al., 2006) is a recent state of the art algorithm for learning causal structure which uses both constraint-based and score-based methods. First, constraint-based methods are used to find candidate sets of possible parents and children for each variable. Then, score-based methods are used to combine these candidate sets and find a graph which maximizes a score function. MMHC learns the correct PDAG in the large sample limit when the Markov and faithfulness properties hold, there are no unobserved common causes, the true causal structure does not contain directed cycles, and a locally consistent score function is used.

4. Learning from multiple datasets which measure the same variables

It is often the case that researchers have access to multiple observational datasets which record the same variables, e.g. datasets from different labs studying the same problem. When learning causal structure from such data, it can be advantageous to use all of the data from each dataset since this may produce a more reliable result than a single dataset due to the increased total sample size. In the simplest case, where every dataset measures exactly the same variable set, researchers may be tempted to simply concatenate these datasets into a single dataset, possibly after applying some type of standardization procedure to each dataset, and then apply one of the causal structure learning algorithms described in Section 3 to this combined dataset. Even when it can be assumed that the causal relationships governing the recorded variables in each dataset are the same, the use of different types of recording instruments and procedures as well as other factors can lead to slight differences in the joint distributions each dataset is sampled from. When this is the case, statistical problems, such as spurious associations, frequently result when the data are concatenated in this manner. Tillman (2009) found that in these contexts, using the PC algorithm with a concatenated dataset leads to poor results, in some cases worse than simply using one of the individual datasets used to produce the concatenated dataset.

Tillman (2009) proposes a different constraint-based procedure, which incorporates methods used in meta-analyses, for learning causal structure from multiple datasets when each dataset measures the same variable set. Meta-analyses use statistical procedures to combine results from multiple independent experiments and obtain a global

assessment of whether certain effects may be present in the population. One of the most common and convenient ways of doing this involves combining the experimenters' reported p-values from a group of independent studies into a new test statistic which is then used to decide whether the null hypothesis each experimenter is investigating should be rejected (Sutton et al., 2000). Since all that is being combined are the p-values associated with different statistical tests on different datasets, these methods are not affected by differences in the distributions across different experiments, provided the same variables are manipulated in each experiment.

A naive method for combining p-values, which is sometimes used in the literature, is to simply average the individual p-values. This average can then be compared to the original α threshold for rejecting the null hypothesis in a single hypothesis test with a specified type I error rate. This averaging procedure, however, fails to take advantage of the fact that each p-value constitutes an independent source of information; unlikely evidence from multiple sources is more compelling than unlikely evidence from a single source and should thus lower the overall combined p-value (Tillman, 2009). For example, assume we set $\alpha = .05$ and we observe $p = .06$ in two separate experiments. The average p-value is thus .06 which fails to reject the null hypothesis even though it is much less likely that we would observe p-values this extreme and this close to the α threshold in two independent experiments than in a single experiment. A more statistically appropriate way of combining these results would take this into account and likely result in the rejection of the null hypothesis.

Fisher (1950) introduced a theoretically sound method for combining p-values from multiple experiments into a new test statistic. This method, which we will refer to as *Fisher's method* combines p-values into the following test statistic:

$$T_F = -2 \sum_{i=1}^{k} \log(p_i). \tag{2}$$

T_F has a χ^2 distribution with $2k$ degrees of freedom under the null hypothesis, where k is the number of p-values combined, which follows from the fact that p-values have a standard Uniform distribution under the null hypothesis. We can thus compare T_F to the χ^2_{2k} quantile function to make the rejection decision. When two low p-values are combined, Fisher's method results in a much lower combined p-value, unlike taking the average. Simulation studies have found that Fisher's method, in general, performs better than similar competing methods for combining p-values that have been investigated in the literature, and satisfies an optimality criterion known as *Bahadur efficiency* (Bahadur, 1971), which is related to the effective use of data as the number of samples increases (Lazar et al., 2002).

The method proposed in Tillman (2009) shows how any constraint-based causal structure learning algorithm can be adapted to learn from multiple datasets which measure the same variable set by replacing each conditional independence test that would be performed by the algorithm with a method for combining p-values, such as Fisher's method, with the p-values that are associated with performing conditional

independence tests using each individual dataset. Tillman (2009) found that this method, when used to modify the PC algorithm, results in a significant increase in performance when compared to applying the PC algorithm to a single dataset or to a concatenated dataset. Tillman (2009) found that Fisher's method outperforms other competing methods for combining p-values, which is consistent with the results described in Lazar et al. (2002) when these methods are used in other contexts.

While the method proposed in Tillman (2009) is designed for cases where each dataset measures the same variable set, it can be adapted to cases where the variable sets are not the same, provided at least one dataset measures every variable that is measured in any of the datasets (or some variables can be excluded to ensure that this condition is satisfied.) This condition is required since it is not possible to perform certain conditional independence tests, which may be required by a causal structure learning algorithm like PC, when some variables are never jointly measured in a single dataset. Section 5 discusses alternative algorithms specifically designed for cases when this condition is not satisfied. Tillman and Spirtes (2011) provide a generalization of the conditional independence testing procedure using combined p-values with Fisher's method given in Tillman (2009), which we give below as Algorithm 2. This generalization permits conditional independence testing using as many datasets as possible for a particular conditional independence query whenever the variables involved in a particular test are jointly measured in at least one dataset. $F^{-1}_{\chi^2_k}$ is used to represent the χ^2 quantile function with k degrees of freedom.

Input : Set of datasets $\mathbf{D}$, variables $X, Y, \mathbf{Z}$ to test $X \perp\!\!\!\perp Y|\mathbf{Z}$, significance level α
Output: `reject` or `fail to reject` the conditional independence hypothesis

$k \leftarrow 0$
foreach $D_i \in \mathbf{D}$ **do**
  if *X, Y and $\mathbf{Z}$ are all measured in D_i* **then**
    $p_i \leftarrow$ the p-value associated with the test $X \perp\!\!\!\perp Y|\mathbf{Z}$ using D_i
    $k \leftarrow k + 1$
  else
    $p_i \leftarrow 1$
  end
end
if $-2\sum_{i=1}^{k} \log(p_i) < F^{-1}_{\chi^2_{2k}}(1-\alpha)$ **then**
  return `fail to reject`
else
  return `reject`
end

Algorithm 2: Test Conditional Independence

Ramsey et al. (2010) proposes a score based procedure for learning causal structure using multiple datasets which measure the same variable set. Ramsey et al. (2010) shows that the BIC score can be modified so that the log-likelihood component can be replaced with the average of several log-likelihood scores from different datasets

without compromising the local consistency property of the BIC score. Thus, GES can be used with this modified BIC score to learn causal structure using multiple datasets. Ramsey et al. (2010) uses this modified GES procedure to learn cascades of activity between different brain regions from a collection of datasets which each correspond to multiple fMRI BOLD readings for a different individual. While the procedure they describe assumes that each dataset measures exactly the same variable set and contains the same number of samples, the procedure works in practice with slight further modifications to the BIC score when sample sizes are not constant or some variables are missing from some datasets, but there is at least one dataset which measures every variable.

5. Learning when no dataset measures the full set of variables

When a researcher is interested in learning the causal structure for a set of variables which span across multiple datasets, but no single dataset exists which measures all of these variables, it may not be possible to test every conditional independence fact that is needed by a constraint-based causal structure learning algorithm designed for a single dataset, e.g. we cannot test whether $X \perp\!\!\!\perp Y$ if there are no datasets where X and Y are jointly recorded. We thus cannot simply modify existing constraint-based procedures to permit causal structure learning from multiple datasets as described in Section 4 in this context. Furthermore, the causal structure we can learn from such data will be more underdetermined than in the case of a single dataset since the PDAG or PAG structures produced by the algorithms described in Section 3 represent a Markov equivalence class, which describes all conditional independence and dependence relationships among a set of variables, and some conditional independence facts will be untestable in this context. While it is not possible, in general, to learn a unique PDAG or PAG in these contexts, the conditional independence information that is available can be used to learn a set of possible PDAGs or PAGs, one of which will be the correct PDAG or PAG in the large sample limit when the Markov and faithfulness conditions hold.

Tillman et al. (2009) proposed the first procedure for learning from multiple datasets in this context, the *Integration of Overlapping Networks (ION)* algorithm. ION does not learn directly from datasets, but rather accepts as input a set of PAGs which have *overlapping variables* (variables in common) and do not entail contradictory conditional independence information among common variables. Figure 3 (a) and (b) show two possible input PAGs to ION, where two variables from the true causal structure (c) were unmeasured in each input source used to generate the PAGs. ION returns a set of PAGs over the total set of variables that are included in any of the input PAGs, each of which represents a possible Markov equivalence class over this variable set that is consistent with all of the Markov equivalence classes represented by the input PAGs, i.e. for any MAG $\mathcal{G}$ represented by an input PAG, every output PAG represents a MAG $\mathcal{H}$ over all of the variables in any input PAG such that marginalizing out the variables in $\mathcal{H}$ which are not contained in $\mathcal{G}$ produces the MAG

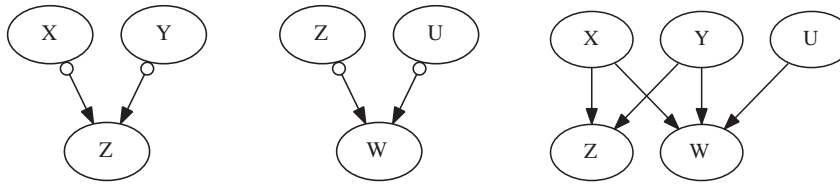

Figure 3: Two input PAGs to ION (a) and (b) corresponding to the true causal structure (c).

$\mathcal{G}$. To learn from a set of datasets which have some variables in common but where no dataset measures all of the variables contained in each dataset using ION, one must first learn a set of PAGs by applying an algorithm for learning a PAG from a single dataset, such as FCI, to each of the datasets and then input these resulting PAGs to ION. Since the output of ION is a set of PAGs over the combined set of input variables, ION is appropriate even in the case where there are unobserved variables in addition to the complete set of variables observed in each dataset.

The ION algorithm begins by transferring nonadjacencies and endpoint orientations to a fully connected PAG over all of the variables where each edge has $\circ$-marked endpoints. The next two stages identify edges to remove and orient as v-structures to produce a candidate set of PAGs which may be consistent with all of the input PAGs. The candidate PAGs are then checked to determine whether they entail all of the m-separations and m-connections that are true in the input PAGs. If this condition is satisfied, then the candidate PAG is added to the output set.

Tillman et al. (2009) shows that the ION algorithm is sound and complete under the Markov and faithfulness assumptions, but it is not scalable in practice to very many variables. One of the most significant issues which limits the practical performance of ION is the fact that input PAGs, when learned from datasets using an algorithm like FCI, often contain inconsistent conditional independences and dependences due to statistical errors. ION attempts to resolve inconsistent conditional independence and dependence information whenever it is detected by treating the conditional independence and dependence relationships which conflict as if they are unknown. However, since sets of conditional independence facts entail other conditional independences, it is not always possible to detect inconsistent information. Furthermore, treating inconsistent conditional independence and dependence facts as unknowns increases the underdetermination of the final output and significantly increases the algorithm's runtime since a larger number of possible structures must be searched over throughout the course of the algorithm. Another significant limitation of ION is that each candidate structure in the final stage of the algorithm is explicitly checked to ensure that it entails the same m-separation and m-connection facts as each of the input PAGs, rather than relying on a simpler condition that can be checked more efficiently.

Tillman and Spirtes (2011) proposes the *Integration of Overlapping Datasets (IOD)* algorithm, which overcomes many of the practical limitations of the ION algorithm. IOD accepts a set of datasets, which have variables in common, but where it is as-

sumed that no dataset measures all of the variables contained in each dataset, as input and returns a set of possible PAGs over all of the variables measured in each of the input datasets as output.

IOD is a constraint-based algorithm and thus relies on the results of conditional independence tests, but assumes that it is not possible, in general, to perform all of the conditional independence tests necessary to distinguish a unique PAG that is consistent with each dataset. Since IOD uses a set of datasets as input, it is able to avoid the problem of inconsistent conditional independence information which ION suffers from by using Algorithm 2, described above, to obtain any necessary conditional independence facts. Since Algorithm 2 only requires the p-values associated with the conditional independence tests used with each dataset to return a result, IOD, like ION, does not require that the datasets each be sampled according to the same distribution; IOD, like ION, only assumes that the same underlying causal relationships govern the variables sampled in each dataset.

The IOD algorithm is based on the theoretical result given below which enables one to determine whether a possible candidate PAG should be included in the algorithm's output without checking whether all of the m-separations and m-connections equivalent to the conditional independence and dependence facts consistent with the datasets are entailed by the candidate PAG. The result says that if we want to check whether a MAG over a set of variables is Markov equivalent to some other MAG over a subset of those variables, we need only check to ensure that the MAG over the larger variable set has the same m-separations as the MAG over the smaller variable set and the MAG over the larger variable set has certain inducing paths for each adjacency in the MAG over the smaller variable set.

Theorem 5.1 (Tillman and Spirtes (2011))**.** Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ and $\mathcal{H} = \langle \mathcal{W}, \mathcal{F} \rangle$ be MAGs where $\mathcal{W} \subset \mathcal{V}$ and $\mathcal{G}'$ be the MAG which results when the variables $\mathcal{V} \backslash \mathcal{W}$ are marginalized from $\mathcal{G}$. $\mathcal{H}$ is Markov equivalent to $\mathcal{G}'$ if the following hold for all pairs $\{X, Y\} \subseteq \mathcal{W}$

(i) for every $\mathbf{Z} \subseteq \mathcal{W} \backslash \{X, Y\}$, if $msep_{\mathcal{H}}(X, Y|\mathbf{Z})$, then $msep_{\mathcal{G}}(X, Y|\mathbf{Z})$, and

(ii) if X and Y are adjacent in $\mathcal{H}$, then $\mathcal{G}$ has an inducing path between X and Y with respect to $\mathcal{V} \backslash \mathcal{W}$.

While the first condition above appears difficult to check, a result from Spirtes et al. (1995) shows that if we check a smaller set of m-separations, referred to as a *sepset*, then this is sufficient to ensure that all of the necessary m-separations required by the condition are entailed. The first stage of the IOD algorithm finds these sepsets and the necessary inducing paths that must be checked when considering a candidate PAG to include in the output set and produces an initial graph with only some m-separation information encoded. The second stage then uses this initial graph to generate candidate PAGs and check whether they entail the necessary m-separations and include the necessary inducing paths. Tillman and Spirtes (2011) shows that this procedure is correct and complete under the Markov and faithfulness assumptions.

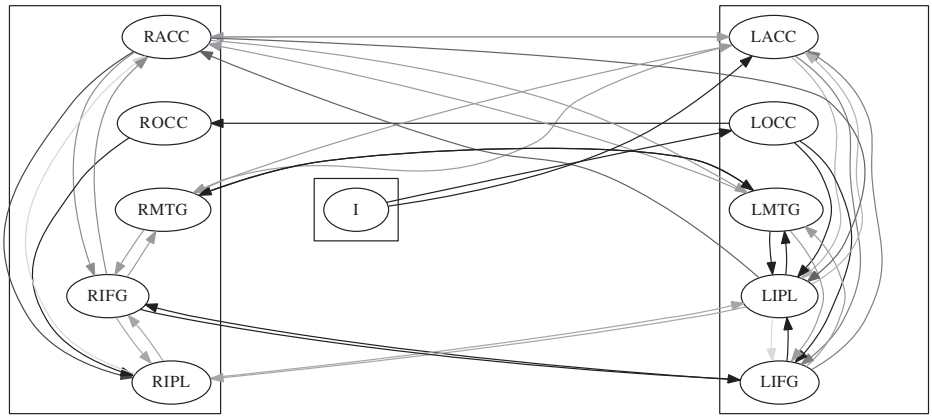

Figure 4: Edges in fMRI datasets equivalence class

Tillman and Spirtes (2011) shows that the IOD algorithm is much faster and requires less memory than the ION algorithm in practice and overcomes the problems associated with inconsistent conditional independence information which ION suffers from through simulations. They also applied the algorithm to datasets resulting from an fMRI experiment analyzed in Ramsey et al. (2010) where experimenters presented individuals with pairs of words and asked the subjects whether the words rhymed. BOLD activity in the following brain regions was recorded in the experiment and included in each dataset: left occipital cortex (LOCC), left middle temporal gyrus (LMTG), left anterior cingulate (LACC), left inferior frontal gyrus (LIFG), left inferior parietal (LIPL), right occipital cortex (ROCC), right middle temporal gyrus (RMTG), right anterior cingulate (RACC), right inferior frontal gyrus (RIFG), right inferior parietal (RIPL). Additionally, an input variable (I) was included in each dataset which indicates the presentation of rhyming or non-rhyming words. Each dataset consists of the recordings for a particular individual. Due to individual differences and differences in fMRI machine calibration and standardization procedures, accurate recordings for every brain region were not available for every individual. Tillman and Spirtes (2011) applied the IOD algorithm to 13 of these datasets where each dataset contained recordings for 6 to 9 different brain regions. Figure 4 summarizes the IOD output equivalence class by showing every edge that is present in some MAG in the equivalence and where the darkness of the edges indicates the number of MAGs that contain the edge (darker edges are included in more MAGs). While the ground truth is unknown, these results are consistent with domain knowledge indicating a cascade of interactions on the left side of the brain after the presentation of the stimulus eventually leading to a cascade on the right side of the brain.

Triantafillou et al. (2010) proposes a different approach to learning causal structure from multiple datasets or existing PAGs with common variables where no dataset or PAG measures all of the variables contained in each input dataset or PAG. They show how the problem of finding a single MAG that entails all of the m-separations and

m-connections observed in an input set of PAGs can be transformed into a satisfiability (SAT) problem. An existing SAT solver (Gomes et al., 2007) can then simply be used to obtain the solution. Their approach can also be used to determine whether a MAG exists that is consistent with the set of input PAGs but also has certain specified edges and endpoint orientations, and if so, produce such a MAG. Thus, one can start with an initial structure that contains some necessary adjacencies and endpoint orientations and then iteratively use this procedure to find all PAGs over the complete variable set that are consistent with the set of input PAGs by checking whether a MAG exists for every possible assignment of unknown edges and endpoint orientations. Triantafillou et al. (2010) shows that this procedure is more scalable than ION, in terms of runtime and memory. This approach, however, assumes that there are no inconsistencies in conditional independences and dependences entailed by the input PAGs due to erroneous conditional independence test results, which would prevent a SAT solver from finding any MAGs that are consistent with the input PAGs.

6. Integrating experimental datasets

In many cases, in addition to observational studies, researchers will have performed small interventional studies, manipulating some subset of the measured variables. If there are multiple groups of researchers, there may be multiple experimental datasets in which different sets of variables have been manipulated. Given the causal inference methods for multiple datasets described in the previous sections, the insights gained from experiments should not go unused. Experimental results can be enormously useful to determine the orientation of a causal relationships and, if the intervention amounts to a randomization of the intervened variable, it can also be used to break confounding due to unobserved variables. In this section we discuss some extensions and variations of causal inference procedures for multiple datasets that can integrate datasets where some variables have been manipulated experimentally.

We define an *experiment* $E_i = (\mathbf{J}_i, \mathbf{U}_i)$ on a set of variables $\mathbf{V}_i$ as a partition of $\mathbf{V}_i$ into two mutually exclusive and jointly exhaustive sets $\mathbf{J}_i$ and $\mathbf{U}_i$, where $\mathbf{U}_i$ consists of variables that are passively observed and $\mathbf{J}_i$ contains the variables that are subject to a randomization simultaneously and independently. An experiment thus gives rise to an experimental dataset in which the intervened variables have been (repeatedly) randomized independently to different values and the remaining variables are measured. Experiments of this type (resembling randomized controlled trials) are often referred to as *surgical interventions* because the randomization breaks the causal influence between each intervened variable and its causes (see Figure 5). The passive observational setting is trivially included when $\mathbf{J}_i = \varnothing$. One can consider other forms of experiments as well, e.g. *fat-hand* interventions, where the randomization of multiple variables are correlated, and experiments in which the intervened variable is not fully controlled but only "nudged." An example of the latter would be an intervention on *income* where instead of determining the value of *income* entirely by a randomizing distribution, one instead only adds, say, $5,000 to the income of the participants in the

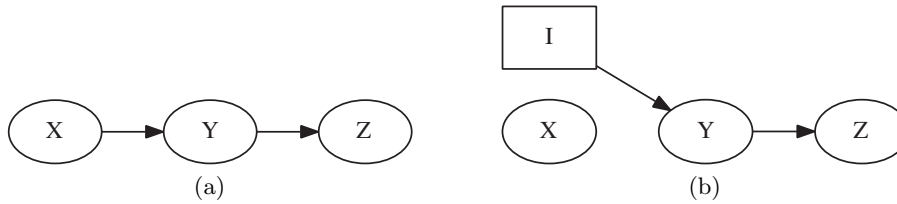

Figure 5: True causal structure (a) and the resulting manipulated structure (b) when Y has been subject to an intervention.

experiment. In that case the variable *income* is still influenced by its normal causes, such as *education* and not fully controlled by the intervention. We will leave these types of experimental interventions aside, but the interested reader can pursue the implications for causal inference in Nyberg and Korb (2006); Eberhardt and Scheines (2007); Eaton and Murphy (2007) and Claassen and Heskes (2010).

Given a set of datasets obtained from a set of experiments (and observations) $\{E_1, \ldots, E_k\}$ over a set of variables $\mathbf{V} = \bigcup_{i=1}^{k} \mathbf{V}_i$, the causal inference task is to determine as much as possible about the causal structure $\mathcal{G}$ over the joint set of variables $\mathbf{V}$ that gave rise to the measured data in the different experiments (and observations). As in the observational case, discovery methods can be divided into those that presuppose that all the (experimental) datasets contain the same variables and only differ in their intervention sets, i.e. the $\mathbf{V}_i$ are identical for all i, but the $\mathbf{J}_i$ differ, and methods for the more general case where datasets do not share the same set of variables, i.e. the $\mathbf{V}_i$ only overlap and the $\mathbf{J}_i$ still differ.

The integration of experimental datasets poses two challenges that distinguish it from the observational case discussed in the previous sections. First, a randomized experiment manipulates the system under investigation. It fully controls the intervened variables and makes them independent of their non-effects (and of each other). In the causal structure learning framework this is reflected in the manipulated distribution and the manipulated causal structure. For example, if the true causal structure is the graph in Figure 5 (a), then a randomization of Y will break the causal influence of X on Y, so the manipulated causal structure now is the graph in Figure 5 (b). Correspondingly, the passive observational distribution and the manipulated distributions are given by

$$\begin{aligned} \mathcal{P}_{\text{obs}}(X, Y, Z) &= \mathbb{P}(X)\mathbb{P}(Y|X)\mathbb{P}(Z|Y) \\ \mathcal{P}_{\text{manip}}(X, Y, Z) &= \mathbb{P}(X)\mathbb{P}^*(Y)\mathbb{P}(Z|Y) \end{aligned}$$

where $\mathbb{P}^*(Y)$ is the randomizing distribution on Y as determined by the intervention I. Unless the causal structure is known, one cannot simply pool the data (or the p-values) across different experimental (or observational) datasets, as was done, for example, in the IOD algorithm in Section 5. Any integration of the experimental datasets has to keep track of the experimental set-ups, i.e. the partitions into $\mathbf{J}_i$ and $\mathbf{U}_i$ that gave rise to the datasets.

A second difference from the purely observational case is that, given enough exper-

iments, one can in principle determine uniquely the true underlying causal structure, as opposed to only an equivalence class. The representations of the equivalence classes used for observational inference methods (PDAGs and PAGs) are no longer appropriate when experimental datasets are considered. In the literature, two different approaches have been used to handle this situation. The first is to develop a new representation for the interventional Markov equivalence classes. This works well for constrained search spaces where one can ensure that the causal structures within one equivalence class do not differ in too many respects. For more general situations, however, this is not feasible as the causal structures that are equivalent given the input datasets, may differ in the orientations of the edges and in their adjacency and ancestral relations. The alternative is to reduce the detail of the information given in the output and simply indicate for each possible edge in the graph $\mathcal{G}$, whether it is present in all graphs consistent with the input dataset, whether it is absent in all graphs consistent with the input dataset, or unknown, i.e. present in some and absent in others. Advantages and disadvantages to both approaches are discussed below.

Under the assumptions that (i) the set of variables $\mathbf{V}_i$ coincide for all i ($\mathbf{V}_i = \mathbf{V}$ for all i), (ii) there are no unmeasured common causes, and (iii) the true causal structure does not contain any directed cycles, Hauser and Bühlmann (2012) have extended the score-based GES algorithm, described in Section 3, to experimental datasets, calling it the Greedy Interventional Equivalence Search algorithm (GIES). GIES requires that the set of experiments is *conservative*, which means that for all variables in $X \in \mathbf{V}$ there is an experiment E_i, such that $X \in \mathbf{U}_i$. For example, one passive observation of $\mathbf{V}$ with $\mathbf{J} = \emptyset$, is sufficient to ensure that the set of experiments is *conservative*, but a set of experiments in which X is always subject to intervention is not *conservative*. The assumption implies that the output of GIES, an interventional Markov equivalence class, is a refinement of the observational Markov equivalence class output by GES. That is, the interventional Markov equivalence classes that result from conservative sets of experiments only contain DAGs that share the same skeleton and the same immoralities, like those output by GES. But in general, GIES will orient additional edges of the PDAG. Nevertheless, much of the theory can be transferred: The score can be determined locally for each variable given its parents, and since all members of the interventional equivalence class have the same skeleton and the same immoralities, an approach analogous to the one used by the observational GES algorithm can be applied, where the score can be evaluated on the basis of any member in the equivalence class. GIES builds on the greedy search over equivalence classes of GES, but adds a third "turning" stage at the end that resolves additional orientations where possible. Thus, using the arguably rather minor restriction to a *conservative* set of experiments, Hauser and Bühlmann (2012) have developed a score-based causal inference procedure that can integrate different experimental datasets and return an interventional Markov equivalence class of causal structures (under the assumption that there are no latent common causes or directed cycles).

When the $\mathbf{V}_i$ are not identical but only overlapping (analogous to Section 5), perhaps the most general approach to learning causal structure is described in Hyttinen

et al. (2013b). Their procedure starts with the independence and dependence relations in the various experimental datasets, encodes the corresponding d-separation and d-connection relations[2)] as satisfiability constraints on the possible pathways that must be present (or absent) in the true causal structure over $\mathbf{V}$, and uses a SAT solver to determine which parts of the underlying graph can be identified. It does not assume that there are no unobserved common causes of the variables in $\mathbf{V}$ or that the underlying causal graph $\mathcal{G}$ is acyclic[3)]. Their approach is very versatile: Any d-separation relation that has been determined in one of the experimental datasets is encoded as a constraint on the paths that must be absent in the true graph $\mathcal{G}$ over the joint set of variables $\mathbf{V}$. The encoding keeps track of the fact that particular paths may have been broken due to the interventions that were performed in the experiment that the d-separation relation was inferred from. For example, suppose that there are three variables $\mathbf{V} = \{X, Y, Z\}$, but we are considering an experiment E_i in which $\mathbf{V}_i = \{X, Z\}$ and $E_i = (\mathbf{J}_i, \mathbf{U}_i) = (\{X\}, \{Z\})$, i.e. X was subject to intervention, Z was observed, and Y was unobserved. Suppose, for purposes of illustration, it was found that in this experiment, X was d-separated from Z, which we denote $dsep_{\mathcal{G},E_i}(X, Z|\varnothing)$. The implication of this d-separation for the underlying graph $\mathcal{G}$ over $\mathbf{V}$ is that there cannot be a direct edge $X \to Z$ in $\mathcal{G}$, nor can there be a path via Y, i.e. $X \to Y \to Z$. These constraints can easily be represented in propositional form where each term $\langle . \to .x \rangle$ in the propositional formula represents the presence of an edge in $\mathcal{G}$ (not a logical implication):

$$dsep_{\mathcal{G},E_i}(X, Z|\varnothing) \implies \neg\langle X \to Z\rangle \wedge \neg[\langle X \to Y\rangle \wedge \langle Y \to Z\rangle]$$

The propositional constraint obviously still allows the possibility that, among other things, there is an unmeasured confounder between X and Z, i.e. $X \leftrightarrow Z$ in $\mathcal{G}$, that Y is a common cause of X and Z, so $X \leftarrow Y \to Z$, or that there is a (direct or indirect) path from Z to X, so $X \leftarrow Z$ or $X \leftarrow Y \leftarrow Z$. All these connections between X and Z would have been broken by the intervention on X in the experiment, and thus the detected d-separation between X and Z in experiment E_i does not disallow them in $\mathcal{G}$. If instead of the d-separation, we had found a d-connection between X and Z in experiment E_i, then we would have a constraint on $\mathcal{G}$ that required that $X \to Z$ in $\mathcal{G}$ or $X \to Y \to Z$ in $\mathcal{G}$ (or both), i.e.

$$\neg dsep_{\mathcal{G},E_i}(X, Z|\varnothing) \implies \langle X \to Z\rangle \vee [\langle X \to Y\rangle \wedge \langle Y \to Z\rangle].$$

Hyttinen et al. (2013b) thus encode any detected d-separation and d-connection as a constraint on the paths that must or must not be in $\mathcal{G}$. The constraints are propositionally encoded and a SAT-solver is used to check for each possible edge in $\mathcal{G}$ whether

[2)] The definition of d-separation used by Hyttinen et al. (2013b) is slightly different than Definition 2.2 used here. It is extended naturally to handle bi-directed edges (for the latent variables) and experimental set-ups. See also the next footnote.

[3)] One should note, though, that in the cyclic case the correspondence between independence and d-separation relations is only known to hold for linear Gaussian models (Spirtes, 1995).

the constraints imply that the edge is either (i) absent in all causal structures consistent with the constraints, (ii) present in all causal structures consistent with the constraints, or (iii) unknown, i.e. it is present in some and absent in other causal structures that are consistent with the constraints. The resulting algorithm is rather simple: It iterates over the datasets and searches for d-separation and d-connection relations in the order of increasing conditioning set size. The d-separation/d-connection relations are encoded as propositional constraints on each pass, and the SAT-solver is used to determine the status of each edge before new constraints are added. Some simple heuristics are used to avoid performing all possible independence tests.

Since the encoding is done in terms of constraints on the presence or absence of particular paths in $\mathcal{G}$, it is very easy to include a wide variety of background knowledge in the search procedure. For example, it would be trivial to add the background knowledge that there is a path from X to Y via Z, without having to specify which other variables (assuming there are more) are on the path or in what order.

In principle, one could develop a representation of the equivalence class of causal structures in the output. Hyttinen et al. (2013b) do not pursue this route because in the case of overlapping datasets, the equivalence classes very quickly become very large (see the discussion in this regard by Triantafillou et al. (2010) on the size of the equivalence classes of the ION algorithm in a more restricted search space). Instead, the SAT-based approach lends itself to answering specific queries: The default output is the status of each edge in the graph, as described above, but if one were interested in some other graphical feature, such as an ancestral relation or the presence of a common effect, then this can also be encoded as a query and determined by the SAT-solver in light of the constraints that it has obtained.

Like any constraint based procedure, this procedure also suffers when statistical errors give rise to conflicting d-separation constraints (recall the discussion of the ION vs. IOD algorithms in Section 5). There are a variety of standard ways in which such conflicts can be handled, e.g. retract the conflicting constraints and solve without them, use different thresholds to infer d-separation constraints from d-connection constraints, etc., but perhaps the more interesting direction to explore is to use weighted maxSAT techniques, where the SAT-solver finds solutions that maximize the number of (weighted) constraints that are satisfied. The details of such an approach are still under active research.

The approach just described is nonparametric in the sense that it is based only on d-separation and d-connection relations. This enables a wide applicability of the search procedure but may undermine how much one can actually infer about the underlying causal structure. For example, Eberhardt et al. (2010) provide an example where even if one can intervene on each variable individually, the underlying causal structure cannot be uniquely determined. If one has reasons that support parametric assumptions, more can be done. In Hyttinen et al. (2012) the authors show how results from different experiments can be integrated when the causal relations are assumed to be linear. This approach is extended in Hyttinen et al. (2011) to discrete models with a so-called *noisy-or* parameterization. In all these cases one can char-

acterize the conditions that a set of experiments must satisfy in order for the true causal structure to be uniquely identified. The approaches thus lend themselves not only to the integration of overlapping experimental datasets, but also to the selection of what experiment would be the best to do next (see Hyttinen et al. (2013a) for further discussion).

7. Conclusion

The above sections provide an overview, along with the necessary background information, of state of the art approaches for predicting causal relationships among sets of variables when multiple observational and experimental datasets are available which each measure some or all of these variables. As described in Section 4, existing constraint-based causal structure learning algorithms can be easily adapted for use in this scenario if all of the datasets are observational and measure the same variable set or there is at least one dataset which measures all of the variables contained in each dataset. Score-based causal structure learning algorithms can also be easily adapted when all datasets measure the same variable set and have the same number of samples. These procedures cannot, however, be easily adapted when no dataset measures the complete variable set and when some variables are subject to experimental manipulations. As described in Section 5, when there are some variables which are never jointly measured, there is conditional independence information which cannot be tested, so specialized procedures are necessary to learn as much as can be known about the Markov equivalence class corresponding to the causal structure which generated the data. While such algorithms tend to be computationally expensive, in general, more recent approaches significantly outperform the first algorithm designed for learning causal structure in this scenario and allow for efficiently querying whether a specific causal relationship may be consistent with observed conditional independence information. Section 6 shows how these procedures can be further generalized when some variables have been manipulated, but as indicated, the handling of conflicting test results is still not well understood, as the pooling techniques described for the observational case cannot be easily transferred to this setting. Given the generality of this search space, there will also be a need for more search constraints, maybe of a parametric form, to avoid massive underdetermination.

There are a number of open problems and possible directions for future research in the area of learning from multiple datasets when no dataset measures the complete variable set. Developing score based-procedures as general as ION or IOD is complicated by the fact that it is difficult to search over the space of and score PAGs, as mentioned in Section 3. However, if one is willing to assume that there are no confounding variables with respect to the complete variable set, it may be possible to develop a dynamic programming approach to learn a set of PDAGs that have optimal substructures with respect to the information available that can be scored. Another possibility is to combine the related methods described in Tsamardinos et al. (2012) for predicting whether two variables measured in different datasets are conditionally

and unconditionally dependent in certain contexts with existing search procedures like IOD to further refine the equivalence classes these procedures produce. Another area for further research involves developing theoretically sound procedures for relaxing constraints when using SAT based search procedures, such as the procedure described in Triantafillou et al. (2010) to learn all possible PAGs, to make these procedures easier to apply to real data. All of these current procedures face the problem that there is no known simple characterization of the equivalence class of causal structures that they output when there is heavy underdetermination. Developing compact informative representations for the information contained in these equivalence classes is an area of research that all of these methods could benefit from. Finally, there is a need for research in further analyzing the robustness and scalability of these procedures as well as applying them to many more application areas.

References

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

Chickering, D. M. (1995). Learning Bayesian networks is NP-complete. In *Proceedings of the 5th International Conference on Artificial Intelligence and Statistics*.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Claassen, T. and Heskes, T. (2010). Causal discovery discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems 22*.

Claassen, T. and Heskes, T. (2012). A Bayesian approach to constraint-based causal inference. In *Proceedings of 28th Conference on Uncertainty in Artificial Intelligence*.

de Campos, C., Zeng, Z., and Ji, Q. (2009). Structure learning of Bayesian networks using constraints. In *Proceedings of the 26th International Conference on Machine Learning*.

Eaton, D. and Murphy, K. (2007). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics*.

Eberhardt, F., Hoyer, P. O., and Scheines, R. (2010). Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74(5):981–995.

Fisher, R. A. (1950). *Statistical Methods for Research Workers*. Oliver and Boyd, London, 11th edition.

Gomes, C., Kautz, H., Sabharwal, A., and Selman, B. (2007). Satisfiability solvers. In Hendler, J., Kitano, H., and Nebel, B., editors, *Handbook of Knowledge Representation*. Elsevier B.V., Oxford.

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

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2011). Noisy-OR models with latent confounding. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2013a). Experiment selection for causal discovery.

Journal of Machine Learning Research. In press.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Järvisalo, M. (2013b). Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*.

Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010). Learning Bayesian network structure using LP relaxations. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Koivisto, M. and Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks*, 20:491–505.

Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, 16:538–550.

Meek, C. (1995a). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*.

Meek, C. (1995b). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*.

Nyberg, E. and Korb, K. (2006). Informative interventions. In Russo, F. and Williamson, J., editors, *Causality and Probability in the Sciences*. College Publications, London.

Parviainen, P. and Koivisto, M. (2009). Exact structure discovery in Bayesian networks with less space. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffmann Publishers.

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

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage*, 49(2):1545–1558.

Richardson, T., Bailer, H., and Banerjee, M. (1999). Specification searches using MAG models. In *Proceedings of the 52nd ISI World Statistics Congress*.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030.

Silander, S. and Myllymaki, M. (2006). A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22th Conference on Uncertainty in Artificial Intelligence*.

Silva, R. and Ghahramani, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 491–498.

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

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons, New York.

Tillman, R. E. (2009). Structure learning with independent non-identically distributed data. In *Proceedings of the 26th International Conference on Machine Learning*.

Tillman, R. E., Danks, D., and Glymour, C. (2009). Integerating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems 21*.

Tillman, R. E. and Spirtes, P. (2011). Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of*

the 14th International Conference on Artificial Intelligence and Statistics.

Triantafillou, S., Tsamardinos, I., and Tollis, I. G. (2010). Learning causal structure from overlapping variable sets. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

Tsamardinos, I., Triantafillou, S., and Lagani, V. (2012). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 13:1097–1157.

Verma, T. S. and Pearl, J. (1990). On equivalence of causal models. Technical report, Cognitive Systems Laboratory, University of California, Los Angeles.

Zhang, J. (2007). A characterization of Markov equivalence classes for causal models with latent variables. In *Proceedings of Uncertainty in Artificial Intelligence*.

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

(Received August 8 2013, Revised October 1 2013)