Journal of Machine Learning Research 11 (2010) 1643-1662 Submitted 2/10; Published 5/10

Introduction to Causal Inference

Peter Spirtes PS7Z@ANDREW.CMU.EDU
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Lawrence Saul

Abstract

The goal of many sciences is to understand the mechanisms by which variables came to take on the values they have (that is, to find a generative model), and to predict what the values of those variables would be if the naturally occurring mechanisms were subject to outside manipulations. The past 30 years has seen a number of conceptual developments that are partial solutions to the problem of causal inference from observational sample data or a mixture of observational sample and experimental data, particularly in the area of graphical causal modeling. However, in many domains, problems such as the large numbers of variables, small samples sizes, and possible presence of unmeasured causes, remain serious impediments to practical applications of these developments. The articles in the Special Topic on Causality address these and other problems in applying graphical causal modeling algorithms. This introduction to the Special Topic on Causality provides a brief introduction to graphical causal modeling, places the articles in a broader context, and describes the differences between causal inference and ordinary machine learning classification and prediction problems.

Keywords: Bayesian networks, causation, causal inference

1. Introduction

The goal of many sciences is to understand the mechanisms by which variables came to take on the values they have (that is, to find a generative model), and to predict what the values of those variables would be if the naturally occurring mechanisms were subject to outside manipulations. For example, a randomized experiment is one kind of manipulation that substitutes the outcome of a randomizing device to set the value of a variable (for example, whether or not a particular new medication is given to a patient who has agreed to participate in a drug trial) in place of the naturally occurring mechanism that determines the variable's value. In non-experimental settings, biologists gather data about the gene activation levels in normally functioning systems in order to understand which genes affect the activation levels of which other genes, and to predict what the effects of manipulating the system to turn some genes on or off would be. Epidemiologists gather data about dietary habits and life expectancy in the general population and seek to find what dietary factors affect life expectancy and predict the effects of advising people to change their diets. Finding answers to questions about the mechanisms by which variables come to take on values, or predicting the value of a variable after some other variable has been manipulated, is characteristic of causal inference. If only non-experimental data are available, predicting the effects of manipulations typically involves drawing samples from one probability density (in the unmanipulated population) and making inferences about the values of a variable in a population that has a different probability density (in the manipulated population).

The rapid spread of interest in the last three decades in principled methods of search or estimation of causal relations has been driven in part by technological developments, especially the changing nature of modern data collection and storage techniques, and the increases in the processing power and storage capacities of computers. Statistics books from 30 years ago often presented examples with fewer than 10 variables, in domains where some background knowledge was plausible. In contrast, in new domains such as climate research (where satellite data now provide daily quantities of data unthinkable a few decades ago), fMRI brain imaging, and microarray measurements of gene expression, the number of variables can range into the tens of thousands, and there is often limited background knowledge to reduce the space of alternative causal hypotheses. Even when experimental interventions are possible, performing the many thousands of experiments that would be required to discover causal relationships between thousands or tens of thousands of variables is often not practical. In such domains, non-automated causal discovery techniques from sample data, or sample data together with a limited number of experiments, appears to be hopeless, while the availability of computers with increased processing power and storage capacity allow for the practical implementation of computationally intensive automated search algorithms over large search spaces.

The past 30 years has also seen a number of conceptual developments that are partial solutions to these causal inference problems, particularly in the area of graphical causal modeling. Sections 3 and 4 of this paper describe some of these developments: a variety of well defined mathematical objects to represent causal relations (for example, directed acyclic graphs); well defined connections between aspects of these objects and sample data (for example, the Causal Markov and Causal Faithfulness Assumptions); ways to compute those connections (for example, d-separation); and a theory of representation and calculation of the effects of manipulations (for example, by breaking edges in a graph); and search algorithms (for example, the PC algorithm). However, in many domains, problems such as the large numbers of variables, small samples sizes, and possible presence of unmeasured causes, remain serious impediments to practical applications of these developments.

The articles in the Special Topic on Causality (containing articles from 2007 to 2009) address these and other problems in making causal inferences. Although there are some superficial similarities between traditional supervised machine learning problems and causal inference (for example, both employ model search and feature selection, the kinds of models employed overlap, some model scores can be used for both purposes), these similarities can mask some very important differences between the two kinds of problems. This introduction to the Special Topic on Causality provides a brief introduction to graphical causal modeling, places the articles in a broader context, and describes the differences between causal inference and ordinary machine learning classification or prediction problems; it is not intended to provide a broad overview or a tutorial surveying all methods of causal inference.

Section 2 describes the problem of causal inference in more detail, and differentiates it from the typical machine learning supervised classification or prediction problem; Section 3 describes several different kinds of causal models; Section 4 describes some problems associated with search for causal models, and why algorithms appropriate for the discovery of good classification or prediction models in machine learning are not always appropriate for the discovery of good causal models; and Section 5 describes some major open problems in the field. The various articles in the Special Topic on Causality are described throughout this article, depending upon which topic they address.

2. Manipulating Versus Conditioning

This section will describe three different kinds of problems (one typical machine learning or statistical problem, and two kinds of causal problems), and three different kinds of probability densities (conditional, manipulated, and counterfactual) that are useful for solving the problems.

2.1 Conditional Probabilities

Suppose that there is a population of individuals with the following random variables at time t: rw_t is the average number of glasses of red wine consumed per day in the 5 years prior to t, bmi_t is the body mass index of a person at time t, sex_t is the person's sex (0 = male, 1 = female) at time t, and ha_t is whether or not an individual had a heart attack in the 5 years prior to t. Since sex_t is rarely time-dependent, it will be replaced simply by *sex*.

Suppose an insurance company at time t wants to determine what rates to charge an individual for health insurance who has $rw_t = 1$, $bmi_t = 25$, and $sex = 0$, and that this rate is partly based on the probability of the individual having a heart attack in the next 5 years. This can be estimated by using the rate of heart attacks among the subpopulation matching the subject, that is $rw_t = 1$, $bmi_t = 25$, $sex = 0$. It is impossible to measure the values of ha_{t+5} at time t, because they haven't occurred yet, but if the probability density is stable across time, the density of ha_{t+5} among the subset of the population with $rw_t = 1$, $bmi_t = 25$, and $sex = 0$ will be the same as the density of ha_t among the subpopulation for which $rw_{t-5} = 1$, $bmi_{t-5} = 25$, and $sex = 0$. The density in a subpopulation is a conditional density, in this case $P(ha_t \mid rw_{t-5} = 1, bmi_{t-5} = 25, sex = 0)$.

Conditioning maps a given joint density, and a given subpopulation (typically specified by a set of values for random variables) into a new density. The conditional density is a function of the joint density over the random variables, and a set of values for a set of random variables.[1] The estimation of a conditional probability is often non-trivial because the number of people with $rw_{t-5} = 1$, $bmi_{t-5} = 25$, $sex = 0$ might be small. A large part of statistics and machine learning is devoted to estimating conditional probabilities from realistic sample sizes under a variety of assumptions.

If the insurance company is not attempting to change anyone's behavior then the question of whether drinking the right amount of red wine *prevents* heart attacks is irrelevant to their concerns; the only relevant question is whether the amount of red wine that someone drinks *predicts* heart attack rates. It is possible that people who drink an average of between 1 and 2 glasses of red wine per day for 5 years have lowered rates of heart attacks because of socio-economic factors that both cause average daily consumption of red wine and other life-style factors that prevent heart attacks. But even if moderate red wine consumption does not prevent heart attacks, the insurance company can still use the conditional probability to help determine the rates to charge.

If $\mathbf{X}$ is a set of measured variables, the conditional probability density $P(\mathbf{Y}| \mathbf{X})$ is not only useful for predicting future values of $\mathbf{Y}$, it is also useful for predicting current unmeasured values of $\mathbf{Y}$, and for classifying individuals in cases where $\mathbf{Y}$ is categorical.

Problem 1: Predictive Modeling

Input: Samples from a density $P(\mathbf{O})$ (where $\mathbf{O}$ is a set of observed random variables), and two sets of variables $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$.

Output: A consistent, efficient estimate of $P(\mathbf{Y} \mid \mathbf{X})$.

[1] In order to avoid technicalities, I will assume that the set of values conditioned on do not have measure 0.

2.2 Manipulated Probabilities

In contrast to the previous case, suppose that an epidemiologist is deciding whether or not to recommend providing very strong incentives for adults to drink an average of 1 to 2 glasses of red wine per day in order to prevent heart attacks. Suppose further that if adopted the incentives will be very widely effective. The density of heart attacks observationally *conditional* on drinking an average of 1 to 2 glasses of red wine per day is not the density relevant to answering this question, and the question of whether drinking red wine prevents heart attacks is crucial. Suppose drinking red wine does not prevent heart attacks, but the heart attack rate is lower among moderate red wine drinkers because some socio-economic variable causes both moderate red wine drinking and other healthy life-styles choices that prevent heart attacks. In that case, after the incentives to drink red wine are in place, the density of socioeconomic status among red wine drinkers will be different than prior to the incentives, and the conditional density of heart attacks among moderate red wine drinkers will not be the same after the incentives were adopted as prior to their adoption. Thus, using observational conditional densities to predict heart attacks after the incentives are in place will lead to incorrect predictions.

The density that is relevant to determining whether or not to recommend drinking a moderate amount of red wine is not the density of heart attacks among people who have chosen to drink red wine (choice being the mechanism for determining red wine consumption in the unmanipulated population), but the density of heart attacks among people who would drink red wine after the incentives are in place. If the incentives are very effective, the density of heart attacks among people who would drink red wine after the incentives are in place is approximately equal to the density of heart attacks among people who are assigned to drink moderate amounts of red wine in an experimental study.

The density of heart attacks among people who have been *assigned* to drink red wine (as opposed to those who have *chosen* to drink red wine, as is currently the case) is a *manipulated* density, that results from taking action on a given population - it may or may not be equal to any observational conditional density, depending upon what the causal relations between variables are. Manipulated probability densities are the appropriate probability densities to use when making predictions about the effects of taking actions ("manipulating" or "doing") on a given population (for example, assigning red wine drinking), rather than observing ("seeing") the values of given variables. Manipulated probabilities are the probabilities that are implicitly used in decision theory, where the different actions under consideration are manipulations.[2]

A simple form of manipulation specifies what new density P' is assigned to some variable in a population at a given time. For example, forcing everyone in an (adult) population to drink an average of 1 glass of red wine daily from t–10 to t–5, assigns $P'(rw_{t-5} = 1) = 1$. (Since rw_{t-5} measures red wine drinking for the past 5 years, an intervention on rw_{t-5} begins at t–10.) After this density has been assigned, there is a resulting joint density for the random variables at time t, denoted by $P(sex, bmi_{t\text{-}5}, ha_{t\text{-}5}, rw_{t\text{-}5}, bmi_t, ha_t, rw_t \,||\, P'(rw_{t-5} = 1) = 1)$, where the double bar indicates the density that has been assigned to rw_{t-5}, in this case that everyone has been assigned the value $rw_{t-5} = 1$.[3] This is in contrast to the conditional density $P(sex, bmi_{t\text{-}5}, ha_{t\text{-}5}, rw_{t\text{-}5}, bmi_t, ha_t,$

[2] The use of manipulated probability densities in decision theory is often not explicit. The assumption that the density of states of nature are independent of the actions taken (act-state independence) is one way to ensure that the manipulated densities that are needed are equal to observed conditional densities that can be measured.

[3] There is no completely standard notation for denoting a manipulated density. This notation is adapted from Lauritzen (1999).

$rw_t \mid rw_{t-5} = 1)$, which is the density of the variables in the subpopulation where $rw_{t-5} = 1$ because people have been observed to drink that amount of red wine, as in the unmanipulated population.

$P(sex, bmi_{t\text{-}5}, ha_{t\text{-}5}, rw_{t\text{-}5}, bmi_t, ha_t, rw_t \mid\mid P'(rw_{t-5} = 1) = 1)$ is a density, so it is possible to form marginal and conditional probability densities from it. For example, $P(ha_t \mid bmi_{t-5} = 25 \mid\mid P'(rw_{t-5} = 1) = 1)$ is the probability of having had a heart attack between t–5 and t among people who have a *bmi* of 25 at t–5, everyone having been assigned to drink an average of 1 glass of red wine daily between t–10 and t–5. In this paper, in order to simplify the exposition, it will be assumed that all attempted manipulations are successful; that is, if $P'(rw_{t-5} = 1) = x$ then $P(rw_{t-5} = 1 \mid\mid P'(rw_{t-5} = 1) = x) = x$ (that is, if rw_{t-5} is manipulated to have value 1 with probability x, then in the manipulated population, rw_{t-5} has value 1 with probability x.) For example, if it is assumed that $P'(rw_{t-5} = 1) = 1$ then $P(rw_{t-5} = 1 \mid\mid P'(rw_{t-5} = 1) = 1) = 1$, that is if everyone has been assigned to drink an average of 1 glass of red wine per day for 5 years (denoted $P'(rw_{t-5} = 1) = 1$), that everyone has done so.

In a randomized trial, a manipulation could set $P'(rw_{t-5} = 1) = 0.5$ and $P'(rw_{t-5} = 0) = 0.5$, in which case the resulting density is $P(sex, bmi_{t\text{-}5}, ha_{t\text{-}5}, rw_{t\text{-}5}, bmi_t, ha_t, rw_t \mid\mid \{P'(rw_{t-5} = 1) = 0.5, P'(rw_{t-5} = 0) = 0.5\})$.

In more complex manipulations, different probabilities can be assigned to different subpopulations. For example, the amount of red wine someone is assigned to drink could be based on *sex*: $P'(rw_{t-5} = 0 \mid sex = 0) = 0.25$, $P'(rw_{t-5} = 1 \mid sex = 0) = 0.75$, $P'(rw_{t-5} = 0 \mid sex = 1) = 0.\ 5$, $P'(rw_{t\text{-}5} = 2 \mid sex = 1) = 0.5$. The resulting density is $P(sex, bmi_{t\text{-}5}, ha_{t\text{-}5}, rw_{t\text{-}5}, bmi_t, ha_t, rw_t \mid\mid \{P'(rw_{t-5} = 0 \mid sex = 0) = 0.25, P'(rw_{t-5} = 1 \mid sex = 0) = 0.75, P'(rw_{t-5} = 0 \mid sex = 1) = 0.5, P'(rw_{t\text{-}5} = 2 \mid sex = 1) = 0.5\})$. In general, which manipulations are performed on which subpopulations can be a function both of the values of various random variables, and of what other past manipulations have been performed.

In many cases the values of some variables in the pre-manipulation density are stable, and the temporal indices on those variables are omitted. Similarly, if it is assumed that variables in the post-manipulation population eventually stabilize to fixed values, the time indices of those variables are omitted in the post-manipulation density, and the time-independent variables refer to the stable values. Both of these kinds of omissions of time indices are illustrated by the use of *sex* in the example.

In contrast to conditional probabilities, which can be estimated from samples from a population, typically the gold standard for estimating manipulated densities is an experiment, often a randomized trial. However, in many cases experiments are too expensive, too difficult, or not ethical to carry out. This raises the question of what can be determined about manipulated probability densities from samples from a population, possibly in combination with a limited number of randomized trials. The problem is even more difficult because the inference is made from a set of measured random variables **O** from samples that might not contain variables that are causes of multiple variables in **O**.

Problem 2: Causal Predictive Modeling

Input: Samples from a population with density $P(\mathbf{O})$, and a (possibly empty) set of manipulated densities $P(\mathbf{O} \mid\mid M_1), \ldots\ P(\mathbf{O} \mid\mid M_n)$, a manipulation M, and sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$.
Output: A consistent, efficient estimate of $P(\mathbf{Y} \mid \mathbf{X} \mid\mid M)$ if possible, and an output of “not possible” otherwise.

With causal inference, as with statistical inference, it is generally the case that in order to make inference tractable both computationally and statistically, simplifying assumptions are made. One kind of simplifying assumption common to both statistical and causal inference is the assumption that the population distribution lies in some parametric family (for example, Gaussian) or that relationships between variables are exactly linear. An example of a simplifying assumption unique to causal inference is that multiple causal mechanisms relating variables do not exactly cancel (Section 3). So, although the goal of Problem 2 is stated as finding a consistent estimate of a manipulated density, it is more realistic to state the goal as finding a sufficiently good estimate of a manipulated density when the sample size is large enough.

Problem 2 is usually broken into two parts: finding a set of causal models from sample data, some manipulations (experiments) and background assumptions (Sections 3 and 4), and predicting the effects of a manipulation from a set of causal models (Section 3). Here, a "causal model" (Section 3) specifies for each possible manipulation that can be performed on the population (including the manipulation that does nothing to a population) a post-manipulation density over a given set of variables. In some cases, the inferred causal models may contain unmeasured variables as well as measured variables.

Problem 3: Constructing Causal Models from Sample Data

Input: Samples from a population with density $P(\mathbf{O})$, a (possibly empty) set of manipulated densities $P(\mathbf{O}||M_1), \ldots\ P(\mathbf{O}||M_n)$, and background assumptions.

Output: A set of causal models that is as small as possible, and contains a true causal model that contains at least the variables in **O**.

Problem 4: Predicting the Effects of Manipulations from Causal Models

Input: An unmanipulated density $P(\mathbf{O})$, a set **C** of causal models that contain at least the variables in **O**, a manipulation M, and sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$.

Output: A function g such that $P(\mathbf{Y} \mid \mathbf{X} \mid\mid M) = g(P(\mathbf{O}), \mathbf{C}, M, \mathbf{X}, \mathbf{Y})$ if one exists, and an output of "no function" otherwise.

In analogy to the goals of statistical modeling, it would be more accurate but much more vague to state that the goal in Problem 3 is to find a useful (for example, sufficiently simple, sufficiently accurate, etc.) causal model, rather than a true causal model.

The reason that the stated goal for the output of Problem 3 is a set of causal models, is that it is generally not possible to reliably find a true causal model given the inputs. Furthermore, in contrast to predictive models, even if a true causal model can be inferred from a sample from the unmanipulated population, it generally cannot be validated on a sample from the unmanipulated population, because a causal model contains predictions about a manipulated population that might not actually exist. This has been a serious impediment to the improvement of algorithms for constructing causal models, because it makes evaluating the performance of such algorithms difficult. It is possible to evaluate causal inference algorithms on simulated data, to employ background knowledge to check the performance of algorithms, and to conduct limited (due to expense, time, and ethical constraints) experiments, but these serve as only partial checks how algorithms perform on real data in a wide variety of domains. For examples, see the Causality Challenge (http://www.causality.inf.ethz.ch/challenge.php).

In the Special Topic on Causality in this journal, Shpitser and Pearl (2008) and Zhang (2008) address Problem 4. Bromberg and Margaritis (2009), Pellet and Elisseeff (2008), He and Geng (2009), and (indirectly) Kang and Tian (2009), Aliferis et al. (2010a), and Aliferis et al. (2010b) address Problem 3. Both the problems and the papers will be described in more detail in subsequent sections.

2.3 Effects of Counterfactual Manipulations

There are cases in ethics, the law, and epidemiology in which there are questions about applying a manipulation to a subpopulation whose membership cannot be measured at the time that the manipulation is applied. For example, epidemiologists sometimes want to know what would the effect on heart attacks have been, if a manipulation such as assigning moderate drinking of red wine from t–10 to t–5, had been applied to the subpopulation which has *not* moderately drunk red wine from t–10 to t–5. When the manipulation under consideration assigns a value to a random variable to a subpopulation with a different actual value of the random variable, the probability in question is a *counterfactual* probability. If the subpopulation that did not moderately drink red wine between t–10 and t–5 differs systematically from the rest of the population with respect to causes of heart attacks, the subpopulations' response to being assigned to drink red wine would be different than the rest of the population.

Questions about counterfactual probabilities arise naturally in assigning blame in ethics or in the law. For example, the question of whether tobacco companies were negligent in the case of someone who smoked and developed lung cancer depends upon the probability that person would not have gotten lung cancer if they had not smoked.

A counterfactual probability cannot be estimated directly from a randomized experiment, because it is impossible to perform a randomized experiment that assigns moderate red wine drinking between t–10 to t–5 to a group of people who already have not been moderate wine drinkers between t–10 and t–5. This raises the question of how counterfactual probabilities can be estimated. One general approach is to assume that the value of red wine drinking between t–10 and t–5 contains information about hidden causes of red wine drinking that are also causes of heart attacks.

Problem 5: Counterfactual predictive modeling

Input: An unmanipulated density $P(\mathbf{O})$, a set $\mathbf{C}$ of causal models that contain at least the variables in $\mathbf{O}$, a counterfactual manipulation M, and sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$.

Output: A function g such that $P(\mathbf{Y} \mid \mathbf{X} \mid\mid M) = g(P(\mathbf{O}), \mathbf{C}, M, \mathbf{X}, \mathbf{Y})$ if one exists, and an output of "no function" otherwise.

In the Special Topic on Causality in this journal, Shpitser and Pearl (2008) describes a solution to Problem 5 in the case where the causal graph is known, but may contain unmeasured common causes.

3. Causal Models

This section describes several different kinds of commonly used causal models, and how to use them to calculate the effects of manipulations. The next section describes search algorithms for discovering causal models.

A (parametric) *statistical model* (with free parameters) is a set of probability densities that can be mapped into a single density by specifying the values of the free parameters (for example, a family of multivariate Gaussian densities).[4] For example, a Hidden Markov Model with a fixed structure but free parameters is a statistical model that represents a certain set of probability densities. A *causal model with free parameters* also specifies a set of probability densities over a given set of variables; however, in addition, for each manipulation that can be performed on the population it also specifies a set of post-manipulation probability densities over a given set of variables. A causal model with free parameters together with the values of the free parameters is a *causal model with fixed parameters*; a causal model with fixed parameters is mapped to a single density given a specification of a manipulation.

Often, a causal model is specified in two parts: a statistical model, and a causal graph that describes the causal relations between variables. The most frequently used causal models belong to two broad families: (1) causal Bayesian networks, (2) structural equation models. Causal Bayesian networks (and related models), specify a density for a variable as a function of the values of its causes. Structural equation models (SEMs) specify the value of a variable as a function of the values of its causes (typically including some unmeasured noise terms.) However, not surprisingly, the two kinds of models are closely linked, as explained in Section 3.2.

The statistical setup for both causal Bayesian networks and structural equation models is a standard one. There is a population of units, where depending upon the problem, the units could be people, cities, cells, genes, etc. It is assumed that there is a density over the population, which assigns probabilities to each measurable subset (event) of the population. Each unit also has a set of properties at a time, where the properties are represented by random variables, which are functions from the units to real numbers. The following sections describe the causal part of the model.

3.1 Causal Bayesian Networks

A *Bayesian network* is a pair $\langle G, P \rangle$, where G is a directed acyclic graph (DAG) whose vertices are random variables, and P is a density such that each variable V in G is independent of variables that are neither descendants nor parents of V in G,[5] conditional on the parents of V in G. In this case P is said to satisfy the *local directed Markov condition* for G.

There are two conditions that are equivalent to the local directed Markov condition described below that are useful in causal inference: the global directed Markov condition, and factorization according to G, both of which are described next.

The conditional independence relations specified by satisfying the local directed Markov condition for DAG G might also entail other conditional independence relations. There is a fast algorithm for determining from G whether a given conditional independence relation is entailed by satisfying the local directed Markov condition for G, that uses the d-separation relation, a relation among the

[4] In the nomenclature of machine learning, what this article calls a “model (with free parameters)” is often called a “model family” or “learning machine” and a “model (with fixed parameter values)” is often called a “model instance” or “model”.

[5] X is a *parent* of Y if the graph contains the edge $X \rightarrow Y$. Y is a *descendant* of X if there is a directed path from X to Y.

vertices of G. A variable B is a *collider* (*v-structure*) *on a path* U if and only if U contains a subpath $A \rightarrow B \leftarrow C$. For disjoint sets of vertices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ in a DAG G, $\mathbf{X}$ is *d-connected* to $\mathbf{Y}$ given $\mathbf{Z}$ if and only if there is an acyclic path U between some member X of $\mathbf{X}$, and some member Y of $\mathbf{Y}$, such that every collider on U is either a member of $\mathbf{Z}$ or an ancestor of a member of $\mathbf{Z}$, and every non-collider on U is not in $\mathbf{Z}$.[6] For disjoint sets of vertices, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is *d-separated* from $\mathbf{Y}$ given $\mathbf{Z}$ if and only if $\mathbf{X}$ is not d-connected to $\mathbf{Y}$ given $\mathbf{Z}$. $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in DAG G if and only if $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in every density that satisfies the local directed Markov condition for G (Pearl, 1988). If $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in P whenever $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in G, then P satisfies the *global directed Markov condition* for G.

For the set of random variables $\mathbf{V}$ in G, a density $P(\mathbf{V})$ *factors according to* DAG G iff

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V|\mathbf{Parents}(V,G))$$

where $\mathbf{Parents}(V,G)$ is the set of parents of V in G.

The local directed Markov condition, the global directed Markov condition, and factorization according to a DAG G are all equivalent under mild regularity assumptions (Lauritzen et al., 1990).

A DAG can also be used to represent causal relations between variables. A is a *direct cause* of B relative to a set of variables $\mathbf{V}$ in a population when there exist two manipulations of $\mathbf{V}\backslash\{B\}$ (that is, all the variables in $\mathbf{V}$, except B, are manipulated to specific values) that differ only in the values assigned to A and that produce different probability densities of B. A *causal DAG G* for a population contains an edge $A \rightarrow B$ iff A is a direct cause of B in the specified population.

In order to use samples from probability densities to make causal inferences, some assumptions relating causal relations to probability densities need to be made. The following Causal Markov Assumption is commonly made, if only implicitly. A set of variables $\mathbf{V}$ is *causally sufficient* iff there is no variable C not in $\mathbf{V}$ that is a direct cause of more than one variable in $\mathbf{V}$ (relative to $\mathbf{V} \cup \{C\}$).

Causal Markov Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population N with density $P(\mathbf{V})$, $P(\mathbf{V})$ satisfies the local directed Markov condition for the causal DAG of N.

Under the Causal Markov Assumption, in a causal Bayesian network a manipulation of X to $P'(X \mid \mathbf{Y})$ (where $\mathbf{Y}$ is assumed to contain only non-descendants of X in a causal DAG G) simply replaces the term $P(X \mid \mathbf{Parents}(X,G))$ in the factorization of the joint density by the manipulated density $P'(X \mid \mathbf{Y})$:

$$P(\mathbf{V}||P'(X|\mathbf{Y})) = P'(X|\mathbf{Y}) \prod_{V \in \mathbf{V}\backslash\{X\}} P(V|\mathbf{Parents}(V,G)).$$

This is called the *manipulation rule*. The importance of the manipulation rule is that if the causal DAG is known, and the unmanipulated density can be estimated from a sample, it allows the prediction of the effect of an unobserved manipulation. Hence the manipulation rule is the solution to Problem 4, in the special case where the observed variables are causally sufficient, and the unique correct causal DAG is known.

[6] For both the d-separation relation and the independence relation, if $\mathbf{X}$ contains a single vertex X, then X will be written instead of $\{X\}$, and similarly for $\mathbf{Y}$ and $\mathbf{Z}$. D-connection can also be defined for cyclic graphs and graphs with double-headed arrows (Spirtes, 1995; Koster, 1999; Cox and Wermuth, 1996).

The solution to Problem 4 is more difficult when the set of observed variables is not causally sufficient. There are sufficient and (almost) necessary rules for determining which manipulated conditional probability densities are invariant under a given manipulation (that is, which densities are the same in the unmanipulated population and the manipulated population) and rules for how to express some non-invariant conditional densities as functions of observed densities (Spirtes et al., 1993). Pearl's do-calculus extended the sufficient and (almost) necessary conditions for determining which conditional densities were invariant from single manipulations to sequences of manipulations, and showed how a broader range of non-invariant manipulated densities could be expressed in terms of observed densities (Pearl, 1995). In the Special Topic on Causality of this journal, Shpitser and Pearl (2008) describe an algorithm that has recently been developed and show that it is a complete solution to Problem 4 in the special case where a unique causal DAG is known (Shpitser and Pearl, 2006a,b; Huang and Valtorta, 2006).

Calculation of the effect of a counterfactual manipulation when causal sufficiency does not hold among the observed variables is a complex operation that requires several copies of the causal graph in order to keep track both of the actual value of the variable being manipulated, and the counterfactual value of the variable being manipulated. In the Special Topic on Causality, Shpitser and Pearl (2008) describe for the first time an algorithm that is a complete solution to Problem 5 in the special case where a unique causal DAG is known, even if the set of observed variables is not causally sufficient.

3.2 Structural Equation Models (SEMs)

Structural equation models are widely used in the social sciences (Bollen, 1989) and in some natural sciences. The set of random variables in a structural equation model (SEM) can be divided into two subsets, the "error variables" or "error terms," and the substantive variables (for which there is no standard terminology in the literature). The substantive variables are the variables of interest, but they are not necessarily all observed. Which variables are substantive, and which variables are error terms can vary with the analysis of the problem. Each substantive variable is a function of other substantive variables and a unique error term. The joint density over the substantive variables is a function of the density over the error terms and of the functions relating each variable to its causes. There is an edge $A \to B$ in the graph ("path diagram") of a SEM when A is a non-trivial argument in the function for B. A manipulation of a variable B to a constant c is represented in a SEM by replacing the equation for B with $B = c$.

In general, the graph of a SEM may have cycles (that is, directed paths from a variable to itself), and may explicitly include error terms with double-headed arrows between them to represent that the error terms are dependent (for example, $\varepsilon_A \leftrightarrow \varepsilon_B$); if no such double-headed edge exists in the graph, the error terms are assumed to be independent of each other. An error term is not explicitly included in the graph unless it is the endpoint of a double-headed arrow; otherwise, an error term occurs in the SEM model, but is not shown in the graph. If the graph has no directed cycles and no double-headed arrows, then the graph is a DAG and the SEM is said to be *recursive*; otherwise it is said to be *non-recursive*.

In a recursive SEM, if the marginal density over the substantive variables is $P(\mathbf{V})$, then $\langle G, P(\mathbf{V})\rangle$ is a Bayesian network (Spirtes et al., 2001; Pearl, 2000); for short, say that a SEM with an associated graph that is a DAG is a Bayesian network (although the SEM contains some extra structure in that it entails that any non-determinism among the substantive variables is only due to the marginalization of the error terms.)

Non-recursive SEMs are of interest because they allow for the representation of feedback (with cycles) or unmeasured common causes (represented by double-headed arrows.) In the case of linear non-recursive SEMs, it is still possible to deduce the conditional independencies (or more generally the zero partial correlations) entailed for all Gaussian SEMs (or more generally linear SEMs) from the graph G of a non-recursive linear SEM using a minor modification of the d-separation relation (Koster, 1999; Spirtes, 1995).

For both theoretical interest and for the purposes of efficient (constraint-based) search of the space of linear non-recursive SEMs without cycles (Section 4.2), it is of interest to find some proper subset of the set of all conditional independence relations entailed by the (modified) d-separation which entail all the rest, that is, a modified form of the local directed Markov condition. (In contrast to the recursive case, where such a subset is given by the independencies entailed by the local directed Markov condition, in the non-recursive case SEMs do not generally satisfy the local directed Markov condition). One such subset of conditional independencies was described by Richardson (2003). In this special issue, the paper by Kang and Tian (2009) describes another such subset, which is often smaller than the one described by Richardson, and hence might be more useful for the purposes of search.

4. Model Search

Traditionally, there have been a number of different approaches to causal discovery. The gold standard of causal discovery has typically been to perform planned or randomized experiments (Fisher, 1971). There are obvious practical and ethical considerations that limit the application of experiments in many instances, particularly on human beings. Moreover, recent data collection techniques and causal inference problems raise several practical difficulties regarding the number of experiments that need to be performed in order to answer all of the outstanding questions.

In the absence of experiments, in practice (particularly in the social sciences) search for causal models is often informal, and based on a combination of background assumptions about causal relations together with statistical tests of the causal models. If a model is rejected by a statistical test, the researcher looks for a modification of the original hypothesized model that will pass a statistical test. The search typically halts when a model that is compatible with background knowledge does not fail a statistical test (Rodgers and Maranto, 1989). Often, the final model is presented, and the search itself is not described. Informal searches of this kind fail to account for multiple testing problems, and can potentially lead to severe overfitting problems. The reliability of such a search depends upon the correctness of the background assumptions, and the extent to which the space of alternatives compatible with the background assumptions was searched. Furthermore, unless the background assumptions are very extensive, or the number of variables is tiny, it is not feasible to estimate and test all of the models compatible with background assumptions. This is further complicated by the fact that, as explained below, for reliable causal inference it is not sufficient to find one model that passes a statistical test; instead it is necessary to find all such models. Recent developments in automated model search have attempted to address these problems with traditional methods of search.

There are several major differences between model search in the case of predicting the unmanipulated value of *Y*, and model search in the case of predicting the post-manipulation value of *Y*, based on the different uses of statistical models and causal models described in the following section.

4.1 Underdetermination of Causal Models by Data

Causal model (with fixed parameter) search is often broken into two parts: search for a causal graph, and estimation of the free parameters from sample data and the causal graph. (In some cases, the prediction of the effects of manipulations does not require estimating all of the free parameters, but does require estimating functionals of the free parameters.) Generally, the estimation of the free parameters employs standard statistical methods. For example, in a linear SEM with a recursive DAG, no unmeasured variables, and Gaussian errors, the maximum likelihood estimate of the edge coefficients is given by regressing each variable on its parents in the DAG. This section concentrates on the search for causal graphs, because the search for causal graphs is significantly different than the search for graphs that are to be used only as statistical models.

In causal model search based on unmanipulated data, if no preference for simpler models over more complex models is made, then the causal models are underdetermined to such an extent that useful causal inference is impossible for many important parametric families (for example, Gaussian or multinomial) or unrestricted probability densities. There are a variety of simplicity assumptions that select simpler models over more complex models that can be made. In the case of search based upon maximizing some model score given sample data (such as the Bayesian Information Criterion), the simplicity assumption is a penalty for complexity built into the score (Chickering, 2002). For search that is not based upon model scores, the following simplicity assumption is often, if implicitly made:

Causal Faithfulness Assumption: For a causally sufficient set of variables **V** in a population *N*, every conditional independence relation true in the density over **V** is entailed by the local directed Markov condition for the causal DAG of *N*.

There are several other versions of the assumption that are considerably weaker than the one stated here (and more intuitively justifiable) but still permit reliable causal inference, at the cost of requiring more complicated algorithms with more complex and somewhat less informative output (Ramsey et al., 2006).

However, even given the Causal Markov and Faithfulness Assumptions and the assumption that the observed variables are causally sufficient, the true causal model is underdetermined by the available evidence and background assumptions, because of the hierarchy of equivalence relations described below.

Two different DAGs *G* and *G’* that have the same set of d-separation relations are said to be *Markov* (*conditional independence*, *d-separation*) *equivalent*.

For each DAG *G*, there is a set **P** of probability densities that satisfy the local directed Markov condition for *G*, denoted **P**(*G*) that are said to be *represented* by *G*. In many cases, some subset of **P** that belongs to a parametric or semi-parametric family **F** is of interest; for example, the Gaussian subset of **P**. Two DAGs *G* and *G’* are *statistically equivalent with respect to* **F** iff **P**(*G*) ∩ **F** = **P**(*G’*) ∩ **F**. Two DAGs that are statistically equivalent with respect to **F** are the same statistical model with respect to **F**.

Two DAGs are *causally equivalent* (with respect to a family of densities **F**) iff they represent the same set of probability densities (in family **F**) for every manipulation (including the null manipulation.) It is easy to see that no pair of DAGs that differ in their structure can be causally equivalent.

As an example, $A \rightarrow B \leftarrow C \leftarrow D$ and $A \rightarrow B \leftarrow C \rightarrow D$ are Markov equivalent, but not causally equivalent. They are statistically equivalent with respect to Gaussian SEMs, but they are not statistically equivalent with respect to linear SEMs with at most one Gaussian error term, and no determinism among the substantive variables (Shimizu et al., 2006).[7]

In the absence of further information (for example, samples from manipulated densities or background domain knowledge) all of the DAGs in a statistical equivalence class fit the data and the background assumptions equally well, and are equally simple. Hence standard scores such as Bayesian Information Criterion, Minimum Description Length, chi-squared statistics, etc. all produce equal scores for the alternative DAGs in a statistical equivalence class for all data sets -- in general, there is no one DAG with the highest score, but rather, there is a set of DAGs with equally high scores. Furthermore, for computational and statistical reasons, it is sometimes easier to search for the Markov equivalence class of DAGs, even if it is known that the statistical equivalence class is a proper subset of the Markov equivalence class.

If the DAG is to be used to estimate observational (not manipulated) conditional densities, this is not a problem, because all of the statistically equivalent models will produce the same estimate. However, if the DAG is to be used to predict the effects of manipulations, then the different models will make different predictions about at least some manipulations. So in the case of causal modeling, unlike observational statistical modeling, it is not enough to simply output one arbitrarily selected DAG from a set of highest scoring DAGs -- it is important to output the entire set, so that all of the different answers given by the different models can be taken into account. Once the set of highest scoring DAGs is found, the problem of dealing with the underdetermination of the effects of manipulations must also be dealt with. These problems are described in more detail in the next two subsections.

If the assumption of causal sufficiency of the observed variables is not made, all three kinds of equivalence classes have corresponding equivalence classes over the set of observed variables, and the problem of causal underdetermination becomes much more severe. For example, for a given set of observed variables **O**, the Markov equivalence class relative to **O** consists of the set of all DAGs (possibly containing variables not in **O**) that have the same set of d-separation relations among the variables in **O**; this might be much larger than the Markov equivalence class that consists of the set of DAGs (containing only variables in **O**) that have the same set of d-separation relations among the variables in **O**.

4.2 Constraint-based Search

First, the problem where only sample data from the unmanipulated population density is available will be considered. The number of DAGs grows super-exponentially with the number of vertices, so even for modest numbers of variables it is not possible to examine each DAG to determine whether it is compatible with the population density given the Causal Markov and Faithfulness Assumptions. Constraint based search algorithms, given as input an oracle that returns answers about conditional independence in the population and optional background knowledge about orientations of edges, return a representation of a Markov equivalence class (or if there is background knowl-

[7] In a linear SEM it is assumed that each variable is a linear function of its causal parents and a unique error term; in a Gaussian SEM it is assumed in addition that the errors term are Gaussian.

edge, a subset of a Markov equivalence class) on the basis of oracle queries. One example of a constraint-based algorithm is the PC algorithm (Spirtes and Glymour, 1991). If the oracle always gives correct answers, and the Causal Markov and Causal Faithfulness Assumptions hold, then the PC algorithm always outputs a Markov equivalence class that contains the true causal model, even though the algorithm does not check each directed acyclic graph. In the worse case, it is exponential in the number of variables, but for sparse graphs it can run on hundreds of variables in an acceptable amount of time (Spirtes and Glymour, 1991; Spirtes et al., 1993; Meek, 1995). Kalisch and Buhlmann (2007) showed that under a strengthened version of the Causal Faithfulness Assumption, the PC algorithm is uniformly consistent for very high-dimensional, sparse DAGs where the number of nodes is allowed to quickly grow with sample size n, as fast as $O(n^a)$ for any $0 < a < \infty$. In practice, the judgments about conditional independence are made by performing (fallible) statistical tests. A number of other variants of constraint-based algorithms have been proposed that improve on either the accuracy or speed of the PC algorithm, or to weaken the assumptions under which it is guaranteed to be correct.

There are both advantages and disadvantages of constraint based searches as compared to either a Bayesian approach to the problem of causal discovery (Heckerman and Geiger, 1995), or an approach based upon assigning a score to each causal model for a given data set (for example, Bayesian information criterion) and searching for the set of causal models that maximize the score (Chickering, 2002).

The disadvantages of constraint-based search include that the output of constraint-based searches give no indication of how much better the best set of output models is compared to the next best set of models; at small sample sizes tests of conditional independence have low power, particularly when many variables are conditioned on; mistakes made early in a constraint based searches can lead to later mistakes; and if the only constraints used are conditional independence constraints, as is often but not always the case, then at best the search outputs a Markov equivalence class, rather than a statistical equivalence class.[8] In addition, constraint-based methods have the problem of multiple testing. If no control is made for multiple testing, the models may overfit the data. However, adjustments to control for overfitting, such as the Bonferroni correction, are often too conservative and as a result the corrected statistical tests are not very powerful. The issue of multiple testing appears in Bayesian approaches to causal discovery as multiple causal model scoring. The issue is handled automatically by Bayesian methods by their use of prior probabilities (Heckerman et al., 1999).

The advantages of constraint-based algorithms are that they are easier to generalize to the case where the observed variables are not causally sufficient, they are generally fast, and given recent developments of non-parametric conditional independence tests, they are applicable without parametric assumptions (Tillman et al., 2009).

In the Special Topic on Causation, Bromberg and Margaritis (2009) models the problem of low power of statistical tests as a knowledge base containing a set of independence facts related through conditional independence axioms that may contain errors due to errors in the tests of conditional independence. The inconsistencies are resolved through the use of a defeasible logic called argumentation that is augmented with a preference function. The logic is used to reason about and possibly correct errors in these tests. Experimental evaluation shows significant improvements in the accuracy of argumentative over purely statistical tests, and improvements on the accuracy of causal

[8] For searches that use non-conditional independence constraints see Silva et al. (2006) and Shpitser et al. (2009).

structure discovery from sampled data from randomly generated causal models and on real-world data sets.

The contributions to the Special Topic on Causality by Aliferis et al. (2010a) and Aliferis et al. (2010b) show that a general framework for localized causal membership structure learning is very accurate even in small sample situations and can thus be used as a first step for efficient global structure learning, as well as accurate prediction and feature selection. It also provides extensive empirical comparisons of state of the art causal learning methods with non-causal methods for the above tasks. In addition, they show that unexpectedly some constraint-based methods are self-correcting with respect to multiple testing, and this may constitute a new methodology for control of multiple statistical testing.

Another problem with constraint-based algorithms is to make them feasible for even higher dimensional data sets. In the Special Topic on Causality, Pellet and Elisseeff (2008) link the causal model search problem to a classic machine learning prediction problem. They show how a generic feature-selection algorithm returning strongly relevant variables can be turned into a causal model search algorithm. Under the Causal Markov and Causal Faithfulness Assumptions, the smallest set of features relevant to predicting a vertex V is the set of parents, children, and parents of children of V. Ideally, the variables returned by a feature-selection algorithm identify those features of the causal graph. Then further processing removes the extra edges (between V and those variables that are parents of children of V but that are neither parents nor children of V) and provides as many orientations as possible. This algorithm is more accurate than PC and other constraint-based algorithms, and has the advantage that it can use arbitrary feature-selection techniques developed for high-dimensional data sets under different assumptions to provide causal model learning algorithms for high-dimensional data under those assumptions.

4.3 Dealing with Underdetermination

One possibility for dealing with the underdetermination of causal models by observational data is to strengthen the available information by sampling from manipulated densities, or in other words, performing experiments.

In the Special Topic on Causality, He and Geng (2009) propose an algorithm for distinguishing between members of a Markov equivalence class by a set of optimally designed experiments. They consider several kinds of experiments, and both a batch-design and a sequential design to minimize the required number of manipulations using both minimax and maximum entropy criteria.

If some members of the Markov equivalence class cannot be eliminated through experimentation, there are several different approaches to using the entire Markov equivalence class to predict the effects of manipulations. (This is Problem 4 in the case where the predictions are made from a set of causal models $\mathbf{C}$ rather than a single causal model, and the set of observed variables may not be causally sufficient.) One possibility is to predict an interval for the potential effects of the manipulated quantity, instead of a point value. Theoretically, an interval could be obtained by calculating the manipulated quantity for each DAG G in the Markov equivalence class, and taking the lower and upper limits. Depending upon how many different SEMs there are in the output, this is sometimes computationally feasible (Maathuis et al., 2009).

A second possibility is to use a Bayesian approach, and perform model averaging. That is, a prior probability is placed over each causal DAG G, and a posterior probability for each G is calculated. Then the manipulated quantity is calculated for each G in the output of the search, and the results are averaged together. This requires putting a prior probability over each graph; in addition, if there are many graphs in the output, then this may not be computationally feasible (Hoeting et al., 1999).

A third alternative is to have an algorithm that determines whether each DAG in the Markov equivalence class predicts the same effect of a given manipulation. For example, if the Markov equivalence class contains $A \rightarrow B \leftarrow C \rightarrow D$ and $A \rightarrow B \leftarrow C \leftarrow D$, then the two causal DAGs disagree about the effect of manipulating D on C, but agree about the effect of manipulating A on B. Even when the observed variables are not causally sufficient there is an algorithm (the Prediction Algorithm) for determining when all of the DAGs in a Markov equivalence class relative to the observed variables agree about the effect of a particular manipulation, and returns the common value of the predicted manipulation when they do all agree (Spirtes et al., 1993). However, this algorithm is known to be correct but incomplete (that is, it sometimes returns “don’t know” even when all models in the equivalence class agree on the effect of a particular manipulation). In this special issue, Zhang (2008) provides a modified version of Pearl’s do-calculus that is more complete than the Prediction algorithm.

5. Open Questions

The following is an overview of important problems that remain in the domain of causal modeling.

1. Matching causal models and search algorithms to causal problems. There are a wide variety of causal models that have been employed in different disciplines. What new models and search algorithms are appropriate for different domains such as feedback or reversible systems (Richardson, 1996)? What search algorithms are appropriate for different combinations of kinds of data, such as experimental and observational data (Eberhardt et al., 2005; Cooper and Yoo, 1999; Yoo and Cooper, 2004; He and Geng, 2009)? What search algorithms are appropriate for different kinds of background knowledge, and different families of probability densities?

2. Model selection, and prior knowledge. What kind of scores can be used to best evaluate causal models from various kinds of data? In a related vein, what are good families of prior densities that capture various kinds of background knowledge?

3. Improve efficiency and efficacy of search algorithms. How can search algorithms be improved to incorporate different kinds of background knowledge, search over different classes of causal models, run faster, handle more variables and larger sample sizes, be more reliable at small sample sizes, and produce output that is as informative as possible?

4. Characterization of search algorithms. For causal search algorithms, what are their semantic and syntactic properties (for example, soundness, consistency, maximum informativeness)? What are their statistical properties (pointwise consistency, uniform consistency, sample efficiency)?[9] What are their computational properties (computational complexity)?

5. Adding or relaxing simplifying assumptions. What plausible alternatives are there to the Causal Markov and Faithfulness Assumptions? Are there other assumptions that might be weaker and hold in more domains and applications without much loss about what can be reliably inferred?

[9] Intuitively, an estimator is pointwise consistent when as the sample size increases without limit, regardless of the true value, with probability 1 the absolute value of the difference between the estimator and the true value approaches zero. An estimator is uniformly consistent if for any given ε and δ, there is a single sample size such that in the worst case, the probability is less than ε that the absolute value of the difference between the estimator and the true value is greater than δ. For precise definitions in the causal context, see Robins et al. (2003).

Are there stronger assumptions that are plausible for some domains that might allow for stronger causal inferences? How often are these assumptions violated, and how much do violations of these assumptions lead to incorrect inferences? Can various statistical assumptions be relaxed? For example, what if the sample selection process is not i.i.d., but may be causally affected by variables of interest?

6. Derivation of consequences from causal graph and unmanipulated densities. Shpitser and Pearl have given complete algorithms for deriving the consequences of various causal models with hidden common causes in terms of the unmanipulated density and the given manipulation (Shpitser and Pearl, 2008). Partial extensions of these results to deriving consequences from sets of causal models have been given (Zhang, 2008); are there further extensions to derivations from sets of causal models?

7. New constraints for structure learning. The Causal Markov and Causal Faithfulness Assumptions, in addition to entailing conditional independence constraints on densities, also entail other constraints on densities. For example, in a linear SEM, if an unobserved variable T causes observed variables X_1, X_2, X_3, X_4, and there are no other causal relations among these variables, then there are no entailed conditional independence relations among just the observed variables X_1, X_2, X_3, X_4. However, the SEM entails $\text{cov}(X_1,X_2)\ \text{cov}(X_3,X_4) = \text{cov}(X_1,X_3)\ \text{cov}(X_2,X_4) = \text{cov}(X_1,X_4)\ \text{cov}(X_2,X_3)$ regardless of the values of the free parameters. This information is useful in finding causal structure with unmeasured variables. In addition, there are sometimes constraints that are not conditional independence constraints on the density of the observed variables that do not depend upon any parametric assumptions (Shpitser et al., 2009). How can these non-parametric constraints be incorporated into search algorithms?

8. Find variable definitions. In many domains, such as fMRI research, there are thousands of variables, but the measured variables do not correspond to functional units of the brain. How is it possible to define new variables that are functions of the measured variables, but more useful for causal inference and more meaningful?

9. Find new applications of causal inference. Applications of causal inference algorithms in many domains (Cooper and Glymour, 1999) help test and improve causal inference algorithms, suggest new problems, and produce domain knowledge.

10. Creating good benchmarks. What are the most appropriate performance measures for causal inference algorithms? What benchmarks can be established? What is the best research design for testing causal inference algorithms?

11. Formal connections between different causal modeling approaches. Many different fields have studied causal discovery including Artificial Intelligence, Econometrics, Operations Research, Control Theory, and Statistics. What are the formal connections between the different models, assumptions, and algorithms used in each of these fields? What can each of these domains learn from the others?

Acknowledgments

I would like to thank Isabelle Guyon, Constantin Aliferis, Greg Cooper, and Rob Tillman for many helpful comments.

References

Constantin Aliferis, Alexander Statnikov, Ionnis Tsamardinos, Subramani Mani, and Xenophon Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification, Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010a.

Constantin Aliferis, Alexander Statnikov, Ionnis Tsamardinos, Subramani Mani, and Xenophon Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification, Part II: Analysis and extensions. *Journal of Machine Learning Research*, 11:235–284, 2010b.

Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley-Interscience, 1989.

Facundo Bromberg and Dimitris Margaritis. Improving the reliability of causal discovery from small data sets using argumentation. *Journal of Machine Learning Research*, 10:301–340, 2009.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Greg Cooper and Clark Glymour. *Computation, Causation, and Discovery*. AAAI Press, 1999.

Gregory Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In Kathryn Laskey and Henri Prade, editors, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 116–125, San Francisco, CA, 1999. Morgan Kauffman.

David Cox and Nanny Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation (Monographs on Statistics and Applied Probability)*. Chapman and Hall, 1996.

Frederick Eberhardt, Richard Scheines, and Clark Glymour. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In Fahiem Bacchus and Tommi Jaakkola, editors, *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 178–184, Arlington, VA, 2005. AUAI Press.

Ronald Fisher. *The Design of Experiments*. Macmillan Pub Co, 1971.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 10:2523–2547, 2009.

David Heckerman and Dan Geiger. Learning Bayesian networks: a unification for discrete and Gaussian domains. In Philippe Besnard and Steve Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 274–282. Morgan Kaufman, 1995.

David Heckerman, Chris Meek, and Gregory Cooper. A Bayesian approach to causal discovery. In Greg Cooper and Clark Glymour, editors, *Computation, Causation, and Discovery*, pages 141–165. MIT Press, Cambridge, MA, 1999.

Jennifer Hoeting, David Madigan, Adrian Raftery, and Chris Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999.

Yimin Huang and Marco Valtorta. Identifiability in causal Bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1149–1154, Edinboro, Scotland, 2006. AAAI Press.

Markus Kalisch and Peter Buhlmann. Estimating high dimensional directed acyclic graphs with the PC algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Changsung Kang and Jin Tian. Markov properties for linear causal models with correlated errors. *Journal of Machine Learning Research*, 10:41–70, 2009.

Jan Koster. On the validity of the Markov interpretation of path diagrams of Gaussian structural equation models with correlated errors. *Scandinavian Journal of Statistics*, pages 413–431, 1999.

Steffen Lauritzen. Causal inference from graphical models. In D. Barnsdorf-Nielsen and C. Kluppenberg, editors, *Complex Stochastic Systems*, pages 141–165. Chapman and Hall, Baton Rouge, LA, 1999.

Steffen Lauritzen, Phil Dawid, B. Larsen, and H. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

Marloes Maathuis, Markus Kalisch, and Peter Buhlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A):3133–3164, 2009.

Chris Meek. Strong completeness and faithfulness in Bayesian networks. In Phillipe Besnard and Steve Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–419, Montreal, Quebec, 1995. Morgan Kaufman.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Jean-Philippe Pellet and Andre Elisseeff. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.

Joseph Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In Rina Dechter and Thomas Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 401–408, Cambridge, MA, 2006. AUAI Press.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In Eric Horvitz and Finn Jensen, editors, *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 454–462, Cambridge, MA, 1996. Morgan Kaufmann.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30:145–157, 2003.

James Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

R. Rodgers and C. Maranto. Causal-models of publishing productivity in psychology. *J Appl Psychol*, 74(4):636–649, 1989.

Shohei Shimizu, Aapo Hyvarinen, Patrick Hoyer, and Yutaku Kano. Finding a causal ordering via independent component analysis. *Comput Stat Data An*, 50(11):3278–3293, 2006.

Ilya Shpitser and Judea Pearl. Identification of conditional intervention distributions. In Rina Dechter and Thomas Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 437–444, Cambridge, MA, 2006a. AUAI Press.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226, Menlo Park, California, 2006b. AAAI Press.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Ilya Shpitser, Thomas Richardson, and James Robins. Testing edges by truncation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1957–1963. AAAI Press, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In Phillipe Besnard and Steve Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 491–499, Montreal, Canada, 1995. Morgan Kaufmann.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):67–72, 1991.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics, 1993.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

Robert Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Proceedings of Advances in Neural Processing Information Systems 22*, pages 1847–1855, Vancouver, BC, 2009. Curran Associates, Inc.

Changwon Yoo and Gregory Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine*, 31(2):169–182, 2004.

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