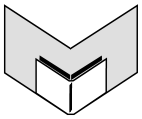
Causality: Objectives and Assessment
Challenges in Machine Learning, Volume 4

Causality: Objectives and Assessment

Challenges in Machine Learning, Volume 4

Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors

Karin Bierig and Nicola Talbot, production editors



Microtome Publishing
Brookline, Massachusetts
www.mtome.com

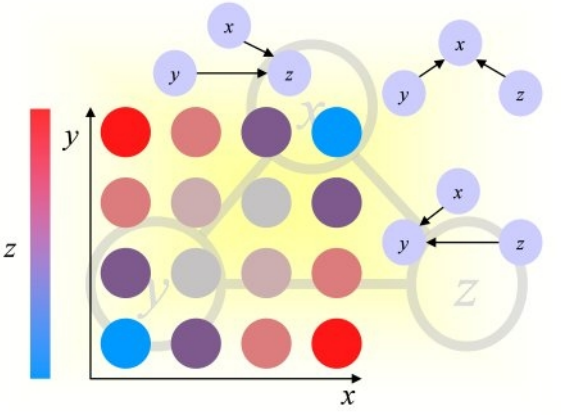
Causality: Objectives and Assessment
Challenges in Machine Learning, Volume 4

Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors

Karin Bierig and Nicola Talbot, production editors

ISBN-13: 978-0-9719777-4-7

Causality Workbench
⟨http://clopinet.com/causality⟩



Foreword

About six months ago, Isabelle Guyon asked me to write a foreword for the first volume on the Causality Challenge (Volume 2 of the Challenges in Machine Learning series), a collection of papers on causal discovery and a collection of "discovery challenges" that involved simulated or real datasets along with particular discovery objectives. I praised the work, and publicly salivated over what the work might mean for the future. Bringing the vast talent of the Machine Learning community to bear on the problems of causal discovery that have been articulated and pioneered by philosophers, statisticians, artificial intelligence researchers and social scientists seemed sure to produce exciting science. Seemingly just minutes after the printer cooled down from that volume, Isabelle sent me the second and present volume, which is all I could have hoped for and more. The collection of excellent papers that follow contain something for everyone and something by everyone. If foundational and representational issues are of interest, Judea Pearl begins the volume with a lucid tutorial on Causal Bayes networks, followed immediately by a philosophical challenge to this framework from Phil Dawid. Later in the volume, Kevin Murphy and colleagues argue that causal discovery can be done without Directed Acyclic Graphs (DAGs), Still further on, Voortman, Dash and Druzdzel consider causal processes represented at equilibrium and those represented as dynamical systems, and argue that only by using an algorithm motivated by none other than Nobel Laureate Herb Simon can we do causal discovery for systems in which the two representations clash.

If general frameworks within which causal discovery can be situated is of interest, then the volume includes a beautiful piece by Frederick Eberhardt casting causal discovery as a game theoretic duel between a "scientist" and "nature," as well as a piece by Lemeire and Steenhaut that looks at causal discovery as an instance of Kolmogorov complexity, as well as a piece on Bayesian algorithms for causal data mining, as well as a piece by Tillman and Spirtes which looks at when causal structure matters, even when the task is strictly predictive, as well as a piece from Uganda on fast causal learning by committee. If causal discovery involving time series is of interest, articles by the Intelligent Data Analysis Group in Germany on multivariate time series and Granger causality and a group from Beijing are rich and new, and the collection also includes fascinating pieces on discovering cyclic causal processes, discovering non-linear networks, as well as pieces on algorithms specialized to more local tasks like discovering the causal direction between a single pair of variables and making predictions about particular manipulations.

The second part of this volume presents a number of datasets and analyses of them that were part of the second Causality Challenge. These range from protein-signaling networks to silicon wafer manufacturing data to simulated data sets for a variety of purposes. They are exactly the sort of community wide scientific challenges that advance a field. In short, Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf are due

enormous praise for causing the existence of another important collection of papers that practically define the state-of-the-art for causal discovery in 2010. I'm going to send this off and go down to the mailbox to look for the next volume.

Richard Scheines
Professor of Philosophy, Machine Learning, and Human-Computer Interaction
Carnegie Mellon University

Preface

This book reprints papers of the Neural Information Processing Systems 2008 (NIPS 2008) workshop "Causality: Objectives and Assessment", December 12, 2008, Whistler, Canada. The papers were initially published on-line in JMLR Workshop and Conference proceedings (JMLR W&CP), Volume 6: http://jmlr.csail.mit.edu/proceedings/papers/v6/.

The project, which led to this book, is an activity of the Causality Workbench http://www.causality.inf.ethz.ch/ supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746 and by the National Science Fundation under Grant N0 ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this book are those of the authors and do not necessarily reflect the views of the sponsors. We are very grateful to all the members of the causality workbench team for their contribution to organizing the pot-luck challenge: Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov.

The Editorial Team:

Isabelle Guyon
Clopinet, California, USA
guyon@clopinet.com

Dominik Janzing
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany
dominik.janzing@tuebingen.mpg.de

Bernhard Schölkopf
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany
bernhard.schoelkopf@tuebingen.mpg.de

Table of Contents

SECOM task (Manufacturing)

Pot Luck Challenge Fact Sheets

JMLR: Workshop and Conference Proceedings 6:1–42, 2010 NIPS 2008 workshop on causality

Causality: Objectives and Assessment

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet, California, USA

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany

Editor: Neil Lawrence

Abstract

The NIPS 2008 workshop on causality provided a forum for researchers from different horizons to share their view on causal modeling and address the difficult question of assessing causal models. There has been a vivid debate on properly separating the notion of causality from particular models such as graphical models, which have been dominating the field in the past few years. Part of the workshop was dedicated to discussing the results of a challenge, which offered a wide variety of applications of causal modeling. We have regrouped in these proceedings the best papers presented. Most lectures were videotaped or recorded. All information regarding the challenge and the lectures are found at http://www.clopinet.com/isabelle/Projects/NIPS2008/. This introduction provides a synthesis of the findings and a gentle introduction to causality topics, which are the object of active research.

Keywords: Causality, Bayesian Networks, Benchmark, Challenge, Competition, resimulated data, probe method

1. Motivations

Machine learning has traditionally been focused on prediction: Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on "data modeling", *i.e.,* on the estimation of a probability law that has generated the data. During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other, however, it is probably fair to say that neither of them has yet fully embraced the field of causal modeling, *i.e.,* the detection of causal structure underlying the data. This has probably different reasons. Many statisticians would still shun away from developing and discussing formal methods for inferring causal structure, other than through experimentation, as they would traditionally think of such questions as being outside statistical science and internal to any science where statistics is applied. Researchers in machine learning, on the other hand, have too long

focused on a limited set of problems, shying away from non i.i.d. data and problems of distribution shifts between training and test set, neglecting the mechanisms underlying the generation of the data, including issues like stochastic dependence, and all too often neglecting statistical tools like hypothesis testing, which are crucial to current methods for causal discovery.

Since the Eighties there has been a community of researchers, mostly from statistics and philosophy, who, in spite of the pertaining views described above, have developed methods aiming at inferring causal relationships from observational data, building on the pioneering work of Glymour, Scheines, Spirtes, Pearl, and others. While this community has remained relatively small, it has recently been complemented by a number of researchers from machine learning. This introduces a new viewpoint to the issues at hand, as well as a new set of tools, including algorithms of causal feature selection, nonlinear methods for testing statistical dependencies using reproducing kernel Hilbert spaces, and methods derived from independent component analysis. Presently, there is a profusion of algorithms being proposed, mostly evaluated on toy problems or in application contexts where models cannot be falsified because of the lack of appropriate data. One of the main challenges in causal learning consists of developing strategies for an objective evaluation. This includes finding methods to acquire large representative data sets of both "observational" and "experimental" data. This, in turn, raises the question to what extent the regularities observed in these data sets provide sufficient evidence on unknown causal structures.

The two themes discussed at the NIPS 2008 workshop on causality reflect these concerns: (1) **Objectives:** Define **causal problems** *i.e.,* generic tasks involving causal modeling illustrated across various application domains. Formalize such tasks mathematically to clearly outline the objectives to be optimized. (2) **Assessment:** Devise reliable protocols of evaluation of solutions to causal problems. To address these objectives, we stated a program of data exchange and benchmarking: the "causality workbench" (Guyon et al., 2010). As part of the effort, we organized for NIPS 2008 a "pot-luck challenge" in which participants were invited to either contribute a solution to one of six proposed tasks or propose a new task.

This introduction is directed to researchers, students, and practitioners with no prior exposure to causality problems, but with some background in machine learning or data mining. It gently guides them through the maze of problems and techniques, without burdening them with mathematical notations and discusses the main outcomes of the workshop. A glossary is appended.

2. Contents overview

In these proceedings, we have gathered the contributions of researchers from a wide variety of horizons. Our collection of papers includes:

- a tutorial paper by one of the founders of the field, Judea Pearl, who revisits the problem of causal modeling with graphical models taking a counterfactual viewpoint,

- a thought provoking paper by Philip Dawid questioning the sanity of the "causal Bayesian network" methodology and proposing a different way of using graphical models for causal modeling, not necessarily interpreting arrows as causal relationships,
- an insightful paper by Lemeire and Steenhaut justifying some of the common model selection choices made in causal discovery using graphical models with the notion of Kolmogorov complexity,
- a novel vision of causal discovery as a game by Frederick Eberhardt,
- a machine learning approach by Haufe and collaborators to learning causal relationships from multivariate time series by enforcing model sparsity,
- two new algorithms by Mani and collaborators for discovering unconfounded causal relationships from observational data without assuming causal sufficiency (which precludes hidden common causes for the observed variables),
- a paper by Tillman and Spirtes analyzing under which conditions models using classical variable or feature selection methods may or may not outperform causal models, shedding light on the results of the causation and prediction challenge (WCCI 2008 (Guyon et al., 2008)).

The proceedings also include selected contributions to the NIPS 2008 "causality pot-luck challenge", proposing innovative solutions to:

- reverse engineering Boolean networks (the SIGNET task),
- finding local causal relationships around a target variable (the LOCANET task),
- finding all possible Markov boundaries, when there is a large number of possible solutions (the TIED task),
- learning a causal network from "heavy handed" manipulations affecting several variables simultaneously (the CYTO task),
- learning causal relationships among pairs of variables isolated from their context – therefore making impossible the use of conditional dependencies to unravel causal direction (the CauseEffectPairs task),
- quantifying the causal effect of promotions on sales (the PROMO task).

Kun Zhang and Aapo Hyvärinen received the best benchmark result award for their contribution to the CauseEffectPairs task (8/8 correct answers). The following authors received mentions: Ernest Mwebaze and John Quinn (for their work on the REGED dataset of the LOCANET task), and You Zhou, Changzhang Wang, Jianxin Yin, Zhi Geng (SIDO dataset, LOCANET task), Mehreen Saeed and the team of Cheng Zheng and Zhi Geng (SIGNET task), and Eugene Tuv (TIED task).

The tasks of the challenge and new proposed tasks contributed by the participants are summarized in Tables 1 and 2. The proceedings include papers describing these tasks, including the new contributions, which will be used in future challenges:

- learning causal relationships using time series when noise is corrupting data in a way that the classical Granger causality method may fail (the NOISE task),

- learning the structure of a fairly complex dynamic system that disobeys the equilibration-manipulation commutability, and predicting the effect of manipulations accurately when a manipulation does not cause an instability (the MIDS task),
- in a manufacturing process (wafer production), identifying measurements on the production line that allow engineers to detect early the pass/fail status at the end of the line (the SECOM task) or identifying faulty manufacturing steps affecting a performance metric (the SEFTY task).

The donor of the dataset NOISE (Guido Nolte) received the best dataset award. The reviewers appreciated that the task includes both real and artificial data and we want to encourage future data donors to move in this direction.

To facilitate the work of practitioners, we have also assembled a collection of "Fact Sheets" containing brief descriptions of the tasks of the challenge and their proposed solutions.

In the rest of this introduction, we develop the main problems addressed in the NIPS 2008 workshop on causality: "objectives" and "assessment". At the risk of missing important aspects, we focus on those concepts most related to machine learning. Section 3 reviews the various settings of causal modeling. Section 4 identifies **objectives** for causal modeling and indicates the role that machine learning may play in pursuing such objectives. Section 5 gives a brief overview of **assessment** methods. Finally, in a discussion section (Section 6) we provide a perspective on challenges being faced, success stories, and open problems.

3. Causal systems *vs.* causal models

A proper definition for causality that regroups all the notions it encompasses in philosophy, psychology, history, law, religion, statistics, physics, and engineering has eluded scientists and philosophers for centuries. However, to avoid accusations of circularity, we give in this section tentative definitions, which, although not universally accepted, are useful to pursue machine learning objectives.

3.1. Causal systems

In the branch of causal studies closest to engineering, the notion of causality is intimately related to the idea that there exist self-contained systems, which have a number of input variables and output variables. Given values of the input variables (set by an external agent), there is a mechanism (a function), which determines the values of the output variables, eventually up to some uncontrollable "stochastic noise". In a certain sense, the values assumed by the input variables cause those of the output variables. There is an intrinsic asymmetry: inversely, if the external agent would force the output variables to assume given values, one would not expect the input variables to be influenced. Take the example of TV remote controllers: you can press a button and turn on or off the TV, but turning on or off the TV does not affect the buttons of the remote controller.

Name (TP; NP; V)	Size	Description	Objective
CEP (Real; 5; 218)	P=8 pairs. N=2 variables.	**Cause Effect Pairs.** Pairs of real variables with known causal relationships.	Find the causal direction in all pairs.
CYTO (Real; 2; 394)	P$\simeq$800 samples per experimental condition $\times$ 9 conditions. N=11 proteins.	**Causal Protein-Signaling Networks in human T cells.** Protein activity monitored by flow cytometry. "Heavy-handed" manipulations are performed using chemical activators or inhibitors.	Learn the architecture of the protein signaling network.
LOCANET (Semi-artificial; 10; 558)	REGED & MARTI: P=500 patients; N=999 genes + target (disease). CINA: P=16033 persons; N=132 attributes + target (earnings). SIDO: P=12678 drugs; N=4932 descriptors + target (activity).	**LOcal CAusal NETwork.** Four datasets: REGED and MARTI (genomics), CINA (marketing), and SIDO (drug discovery). The datasets also include large test sets that were used in the "causation and prediction challenge" (Guyon et al., 2008).	Find the local causal structure around a given target variable (depth 3 network).
SECOM (Real; NA; 59)	P=1567 wafers. N=591 QC measurements + 1 binary target (pass/fail) and 1 date of processing	**Semiconductor manufacturing.** Production entities (wafers) are associated with quality control (QC) measurements on a fabrication line. The labels represent a pass/fail yield in line testing (classification problem).	Predict pass/fail in test data and identify predictive features.
TIED (Artificial; 1; 330)	P=750 training ex. N=1000 variables (including target).	**Target Information Equivalent Dataset.** A Bayesian network with 72 equivalent Markov blankets of the target variable.	Find all Markov blankets.

Table 1: **Atemporal datasets.** "TP" is the data type, "NP" the number of participants who returned results and "V" the number of views as of December 2008. The semi artificial datasets are generally "re-simulated" data, *i.e.,* data obtained from simulators of real tasks, usually trained with real data. Two datasets of LOCANET are made of real data augmented with artificial "probe" variables (SIDO and CINA). N is the number of variables and P is the number of examples (in training data; some datasets have test data too).

Name (TP; NP; V)	Size	Description	Objective
MIDS (Artificial; NA; 65)	T=12 sampled values in time (unevenly spaced); R=10000 simulations. N=9 variables.	**Mixed Dynamic Systems.** Simulated time-series based on linear Gaussian models with no latent common causes, but with multiple dynamic processes.	Use the training data to build a model able to predict the effects of manipulations on the system in test data.
NOISE (Real + artificial; NA; 43)	**Artificial:** T=6000 time points; R=1000 simulations; N=2 variables. **Real:** R=10 subjects. T$\simeq$200000 points sampled at 256Hz. N=19 channels.	**Real and simulated EEG data.** Learning causal relationships using time series when noise is corrupting data causing the classical Granger causality method to fail.	**Artificial task:** find the causal dir. in pairs of var. **Real task:** Find which region of the brain influences which other one.
PROMO (Semi-artificial; 3; 570)	T=365$\times$3 days; R=1 simulation; N=1000 promotions + 100 products.	**Simulated marketing task.** Daily values of 1000 promotions and 100 product sales for three years incorporating seasonal effects.	Predict a 1000$\times$100 boolean influence matrix, indicating for each (i,j) element whether the i^{th} promotion has a causal influence of the sales of the j^{th} product.
SEFTI (Semi-artificial; NA; 35)	R=4000 manufacturing lots; T=300 asynchronous operations (pair of values {one of N=25 tool IDs, date of processing}) + continuous target (circuit performance for each lot).	**Semiconductor manufacturing.** Each wafer undergoes 300 steps each involving one of 25 tools. A regression problem for quality control of end-of-line circuit performance.	Find the tools that are guilty of performance degradation and eventual interactions and influence of time.
SIGNET (Semi-artif.; 2; 415)	T=21 asynchronous state updates; R=300 pseudodynamic simulations; N=43 rules.	**Abscisic Acid Signaling Network.** Model inspired by a true biological signaling network.	Determine the set of 43 boolean rules that describe the network.

Table 2: **Time dependent datasets.** "TP" is the data type, "NP" the number of participants who returned results and "V" the number of views as of December 2008. The semi-artificial datasets are obtained from simulators of real tasks. N is the number of variables, T is the number of time samples (not necessarily evenly spaced) and R the number of simulations with different initial states or conditions.

A wide variety of physical systems under equilibrium do not fall into that category. For instance, a perfect gas governed by the law $pV = nRT$, which states that the product of pressure p and volume V is proportional to the temperature T, would not constitute a "causal system" in the sense described above since any change in two of the variables $\{p, V, T\}$ results in a change in the third one. The role of the three variables p, V and T seems completely symmetrical. Even though there is much to say about the causal interpretation of particular systems subject to the law of perfect gases, we shy away from such controversial cases and limit ourselves to systems in which there is a consensus on their causal interpretation. For instance, there can hardly be any disagreement that if we record the altitude of given villages and their average yearly temperature, if there is a cause-effect relationship, it ought to be altitude that causes temperature and not the opposite.

In many applications, it is useful to broaden the notion of causality to a set of interrelated variables, not necessarily assuming either a role of input or output variable. It becomes then more difficult to define causality and determine to what extent we can say that a variable "causes" another variable. An "operational criterion of causality" (Glymour and Cooper, 1999) is sometimes adopted: consider a system characterized by a set of interdependent random variables (RV) generated by a "natural" stationary distribution, some of which corresponding to directly actionable variables (their values can be set by means of action or manipulation performed by an agent external to the system rather than drawn from the "natural" distribution). **A random variable C may be called a cause of another RV E, called its effect or consequence, if actions performed on C by an external agent result in changes in the distribution of E.** For instance, the variable C=*smoking* and E=*lung cancer* may have given "natural" distributions in a given population. Banning smoking (at least in some places) is an action that may be taken by an external agent (*e.g.,* the Surgeon General). Changes in lung cancer incidence as a result of this action would indicate a causal link between smoking and lung cancer, according to this criterion. This operational criterion of causality provides a sufficient condition for C to be called a cause of E, but not a necessary condition, hence it cannot serve as a definition: An absence of change in the distribution of E under manipulation of C does not exclude that C is a cause of E. For instance, consider the outcome of tossing two fair coins C_1 and C_2 and the variable E that is positive if both coins fall on the same side and negative otherwise. Performing the action of forcing C_1 to be constantly on the "face" side does not change the distribution of E even though C_1 is a cause of E (in the sense that, in the unmanipulated system, E is determined both by C_1 and C_2). To broaden the notion of causality, we give a definition of causal relevance of a variable C to a target E, in the context of other variables (Guyon et al., 2007). For other definitions, see also (Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003; Koller and Friedman, 2009).

The notion of causality between RVs allows us to make simple connections to machine learning and to feature selection applications in which data are often represented as random vectors. It implicitly makes the assumption that similar events repeat themselves and statistics can be computed, hence it does not encompass single event causal-

ity (like legal responsibility in a crime). There are alternative ways of thinking of causality as relationships between objects, events or system states, which we do not cover in this introduction.

Our everyday-life concept of causality is very much linked to time dependencies (the causes precede their effects). However, many machine learning problem are concerned with "cross-sectional studies", which are studies where many samples are drawn at a given point in time. Thus, sometimes the reference to time is replaced by the notion of "causal ordering". Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t + \Delta t$, where Δt can be made as small as we want. But, we will also consider applications in which time dependencies are critical (for instance to continuously monitor treatment in a changing environment) corresponding to problems encountered in so-called "longitudinal studies".

From the point of view described in this section, a "causal system" is characterized by a set of variables, including at least some observable and some directly actionable variables, and a set of permitted actions or manipulations, which may be performed by an external agent to evidence causal relationships between these variables. With some abuse of language we refer to such variables as "random variables" to indicate that they are governed by a "natural" probability distribution when the system is left to evolve according to its own dynamics, and that causal conclusions will be drawn from samples and have only a statistical validity (like "price" influences "sales" or "age" influences "health"). Throughout this introduction, we often use a population of patients under the care of a physician as an example of a causal system. Variables of interest include socio-economic factors, environmental factors, clinical variables, etc. and the physician plays the role of an external agent administering treatments (thought of as actions or manipulations). We put forward this setting for concreteness, but acknowledge that requiring a separation between an inside and an outside of the system and the notion of external agent and manipulations is the object of much debate. In particular, causality is sometimes defined in terms of **counterfactuals** (see glossary): "C causes E" means that "had C not occurred, E would not have taken place". However, because we cannot rewind history and replay events after making small controlled changes, causation can only be inferred, never exactly known. In that sense, it can be understood that the role of "external agents" performing scientific experiments and of statisticians analyzing observations is to approximate as well as possible counterfactuals.

3.2. Causal models

A long time debate in machine learning has been whether predictive models should or not model the data's generative process. Years of research and the results of recent benchmarks (Clopinet, 2009) seemed to have settled the question: there is no need to be concerned with the data's generative process; "agnostic" predictive models, in the vein of neural networks, decision trees and kernel methods, perform as well or better than generative models, at least for data-mining style tasks for which data are i.i.d. But one should be careful not to jump too quickly to conclusions: might the situation change

when we switch from making predictions in a stationary environment (the i.i.d. case) to predicting the consequences of actions?

Assume that we have a system of only two random variables X and Y. In a stationary i.i.d. setting, all that is needed to make predictions is the joint distribution $P(X,Y)$, which does not inform us on whether X was generated from Y or vice versa. However, if actions are being performed, it is useful to know how data were generated. Assume that X is generated first according to $P(X)$ (say X is the atmospheric temperature) and then Y according to $P(Y|X)$ (say Y is the position of the needle of a thermometer). Then, if we force X to assume a given value (by a manipulation like by making a big bonfire), we expect a certain change in Y. Conversely, if we force the thermometer needle position, we do not expect this should have an impact on temperature. The effect of interventions on the joint distribution cannot be predicted by the Bayes formula $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$. In particular, borrowing Pearl's notations (Pearl, 2000), $P(X|do(Y = y)))$ may be different from $P(X|Y = y)$, where $do(Y = y)$ means that Y has been forced to take the value y (by an external agent), while $Y = y$ means that Y has been observed to have the value y. In the case of the temperature example, we have $P(Y|do(X = x)) = P(Y|X = x)$ (observing a given temperature or forcing it artificially to attain the same value results in the same thermometer reading), but we have $P(X|do(Y = y)) \neq P(X|Y = y)$. In fact, $P(X|Y = y)$ obeys the Bayes formula $P(X|Y = y) = P(Y = y|X)P(X)/P(Y = y)$, but $P(X|do(Y = y))$ does not: $P(X|do(Y = y)) = P(X)$ (temperature does not change as a result of forcing the needle position).

From the above considerations, we can conclude that **some knowledge of how the data were generated should be useful** to build predictive models, **if predictions of the consequences of actions are to be made**. In our example, it is useful to choose between two alternative generative models: X generated first according to $P(X)$, then Y generated according to $P(Y|X)$; or, Y generated first according to $P(Y)$, then X generated according to $P(X|Y)$. Importantly, $P(X|Y)P(Y)$ is not the same as $P(Y|X)P(X)$, if the "do" operator is inserted. However, this does not mean that the data's generative process should be modeled faithfully to obtain best prediction performances. As always in machine learning, overfitting must be avoided when modeling data and **the best predictive model does not necessarily belong to the class of systems that generated the data**, owing to the celebrated bias-variance tradeoff (Geman et al., 1992). Therefore, what may appear at first sight to be over-simplifying assumptions (some of which are discussed in Section 6) may turn out to reduce the variance of the model class so effectively that, even though some bias is introduced, good performance is attained.

Before moving forward, we want for concreteness to give some examples of causal models. The use of graphical models in causality has a long history that can be traced back to "path analysis" (Wright, 1921), "structural equations" (Haavelmo, 1943), and modern graphical models that can have a causal interpretation (Spiegelhalter et al., 1993; Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003; Koller and Friedman, 2009). Many other types of models have been used to model causal relationships, including artificial neural networks, Boolean networks, and vari-

ous types of Markov models, including hidden Markov models (HMM), partially observable Markov decision processes (POMDP). The type of causal relationships under consideration have often been modeled as **Bayesian causal networks** or **structural equation models** (SEM) (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). In the graphical representation of such models, an arrow between two variables $A \to B$ indicates the direction of a causal relationship: A causes B. A node of the graph, labeled with a particular variable X, represents a mechanism to generate the value of X given the parent node variable values. For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|\texttt{Parents}(X))$ while for structural equation models it is carried out by a function of the parent variables, eventually distorted by stochastic noise (often but not necessarily additive noise). Learning a causal graph can be thought of as a model selection problem: Alternative graph architectures are considered and a selection is performed, either by ranking the architectures with a global score (*e.g.,* a marginal likelihood, or a penalty-based cost function), or by retaining only graphs that fulfill a number of constraints, such as dependencies or independencies between subsets of variables. Such graphical models usually make at least two simplifying assumption: the **causal Markov condition** (CMC) and the **causal faithfulness condition** (CFC), both of which are discussed in more details is Section 6.

The task of training and selecting causal models is significantly harder than that of training and selecting regular predictive models (classical machine learning from i.i.d. data). The main hurdle in classical machine learning is generally the lack of training data: In most practical applications, with a sufficient amount of training data, the true data distribution may be approached with arbitrary precision, then the problem is "solved". In the jargon of causal modeling, the data commonly used in machine learning are called **observational data**; those are data collected from systems, which are let to evolve according to their own dynamics, without external intervention. Cross-validation is highly effective to perform model selection in this setting.

In contrast, causal models can often not be effectively trained with only "observational data" and cross-validation is ineffective to perform causal model selection, because many models with entirely different causal architectures may perform equally well in an observational setting. It is still debated what the most effective causal model selection strategy should be, but many penalty-based cost functions privileging simple models or stable models have been proposed (Koller and Friedman, 2009). In addition, training and selecting causal models often require data collected after **external interventions** (also referred to as actions, manipulations, or experiments). Such **experimental data** can better distinguish between mere statistical dependence (due for instance to an unknown common cause, referred to as **confounding variable** or **confounder**) and true causation. A widely recognized methodology of unraveling causal relationships or validating causal assumptions is **randomized controlled trial** (RCT). RCTs are most often used for conducting planned experiments in healthcare, but are also employed in other areas of application including judicial, educational, and social research. RCTs involve the random allocation of different interventions (treatments or conditions) to subjects. As long as numbers of subjects are sufficient, this ensures that

both known and unknown confounding factors are evenly distributed between treatment groups. Methods for learning cause-effect links without experimentation (learning from observational data) are attractive because observational data is often available in abundance and experimentation may be costly, unethical, impractical, or even plain impossible (London and Kadane, 2002). Still, many causal relationships cannot be ascertained without the recourse of experimentation and the use of a mix of observational and experimental data might be the most cost effective.

4. Objectives of causal modeling

One of the central topics of the NIPS 2008 workshop was to define objectives for causal modeling. In the previous section, we have tentatively defined causal systems and introduced causal models, not as data generative models, but as tools to predict the consequences of action. Predicting the consequences of actions is often considered to be the main charter of causal modeling. We now review a number of other related causal problems worth pursuing and then put them in the context of applications.

4.1. Causal problems

We collectively call "causal problems" problems requiring the notion of causality. We contrast such problems with machine learning applications using i.i.d. training and test data. In the i.i.d. setting, variables predictive of the target, regardless of causal relationships, may be useful. For instance, in medical diagnosis, the abundance of a protein in serum may be used as a predictor of disease. It is not relevant to know whether the protein is a cause of the disease (*e.g.,* resulting from a gene mutation), or a consequence (*e.g.,* an antibody responding to inflammation). If one is interested in a diagnosis, the abundance of this protein is enough and means disease. We differentiate the problem of making predictions in a stationary environment (diagnosis) with two other types of predictions: the prediction of the consequences of actions performed deliberately by and external agent and counterfactual predictions:

Prediction of the consequences of actions: More and more applications require the assessment of the results of given actions (also referred to as "manipulations" or "experiments"), performed by agents external to the system, thus disturbing the natural functioning of the system. Such assessment is essential in many domains, including epidemiology, medicine, ecology, economy, sociology and business, to assist the development of new treatments and new policies. Assessing the consequences of actions is radically different from making predictions in a stationary environment when the system is subject to its own dynamics. For instance, one might observe that both smoking and coughing are predictive of respiratory disease in a general population and use either predictor for diagnosis. One is a cause (smoking) and the other a symptom (coughing). Acting on the cause can change the disease state, but not acting on the symptom. Therefore if we are interested in treatment rather than in diagnosis it is extremely important to distinguish between causes and symptoms to predict the consequences of actions.

Counterfactual prediction: Another landmark of causal reasoning is **counterfactual** prediction. In fact, some philosophers and practitioners like defining causality via counterfactuals. A typical counterfactual question is: considering that a given patient who took a nicotine substitute stopped smoking, what would have happened if he had not take the medicine? Would he have stopped smoking anyway? More generally, considering a self-contained system of interdependent RVs, what would have been the values assumed by certain variables had some other variables taken values different from the ones observed?

We see that there are subtle differences between predicting the consequence of actions and counterfactuals. First, counterfactuals have to do with hypothetical events that could have taken place in the past whereas predicting the consequences of actions projects events into the future. Second, counterfactual predictions are usually pointwise predictions. For instance, we want to predict what would have happened to one particular patient. In contrast, we might want to optimize the consequences of future actions on a population of patients.

There are many other causal questions. Here we mention a few, which were raised at the NIPS 2006 workshop on causality:

- Determine what manipulations are needed to reach a desired system state with maximum probability (e.g., select variables and propose values to achieve a certain value of a response/target variable, with perhaps a cost per variable).
- Find a causal explanation for a certain observed state y of a target variable Y, *i.e.,* a set of variables having assumed given values, which lead with high probability to the given observation $Y = y$.
- Propose system queries to acquire training data, *i.e.,* design experiments, with perhaps an associated cost per variable and per sample and perhaps with constraints on variables, which cannot be controllable.
- Determine a local causal region around a response/target variable (causal adjacency).
- Determine the source cause(s) for a response/target variable.
- Predict the existence of unmeasured variables (not part of the set of variables provided in the data), which are potential confounders (are common causes of an observed variable and the target).
- Predict which variables called "relevant" by feature selection algorithms are potentially causally irrelevant because their statistical dependency to the target is the result of an experimental artifact (e.g. sampling bias or systematic error).
- Determine causal direction in time series data in which one variable is causing the other.

Defining causal problems supersedes the need for defining causal systems, if we think of causality as a means to an end (solving problems, attaining objectives). We do not need to ascertain that data are generated by a causal system to address a causal problem or answer a causal question. Let us go back to our example of the perfect gas

for which the system of variables $\{p, V, T\}$ did not seem to be in any particular causal relationship. If we use a bicycle pump, the action of pumping has predictable consequences linking the reduction of volume of the gas in the pump to the increase in pressure. Hence, via action/manipulation/experimentation we can evidence a cause-effect relationship for this particular setup, without the prerequisite of solving the problem of whether the set of RVs involved form a "causal" system (this question might not even make sense). What matters to us, in this case, is that we can predict the consequence of actions, *i.e.,* we can use the bicycle pump for a purpose.

4.2. The role of machine learning

The main charter of Machine Learning is learning from data the structure and parameters of an optimal predictive model. We refer to this task as **model inference**. Once a model structure and its parameters are computed, another kind of inference can take place: the inference of variable statistics (point estimations, estimation of expectations, or distribution calculation) given the values of other variables. We refer to this other problem as **variable inference** to distinguish it from the first one. When authors refer to causal inference, they may either refer to variable inference, model inference, or both. We briefly review both aspects to contrast them.

VARIABLE INFERENCE

A lot of effort has been put into solving the problem of variable inference in Bayesian networks (BN) and Structural Equation Models (SEMs), independently of solving the problem of model inference. In many applications, the structure of a causal models is derived from prior knowledge. For instance, in the PROMO task of the challenge, the model structure is given by expect knowledge ("promotions" influence "sales"); only the parameters need to be estimated from data. In some applications, the parameters themselves cannot be subject to learning because of lack of training data, but they can be derived from expert knowledge. For example, the methodology of the noisy-or model, which has been widely deployed for medical diagnosis (Russell and Norvig, 2003) and fault diagnosis (Yongli et al., 2006), allows mapping expert knowledge to parameters. It makes simple independence assumptions between direct causes X_i, $i = 1,...n$ of a target Y. The influence of the X_i on Y is parameterized by only n parameters p_i, easy and intuitive to evaluate for experts. Using n intermediary influence variables Y_i such that Y is the simple logical OR of the Y_i, the parameters p_i represent the probabilities of successful influence: $P(Y_i = 1|X_i = 1) = p_i$ and $P(Y_i = 1|X_i = 0) = 0$. The models thus constructed are used for variable inference.

Variable inference makes use of the model to predict values of certain variables in various situations, including when values of some other variables are missing or imposed (manipulations or counterfactuals). Here is a typical example of variable inference in a simplified "alarm network" (Pearl, 1988): $Burglary \rightarrow Alarm \leftarrow Earthquake$. Assume that the alarm goes off and alerts the police by telephone. The question is: has there been a burglary. If there has just been an earthquake, the probability of

a burglary goes down and it may be unnecessary to send a police officer. Calculation of conditional probabilities (such as $P(Burglary|Alarm, Earthquake)$) can be facilitated by causal networks in complex cases involving a large number of variables. Such uses of causal networks inherit directly from expert systems in artificial intelligence, adding the additional "uncertainty" dimension to the logical constructs. Bayesian networks or SEMs with designated architecture and parameters can be thought of as **motors of calculation of conditional probabilities**. Going one step beyond, in some cases, it is possible to predict conditional probabilities in a **post-manipulation distribution** given the pre-manipulation distribution (the so-called "natural" distribution) and some causal assumptions. For instance, one might want to compute $P(Burglary|do(Alarm), Earthquake)$, where $do(Alarm)$ means that the alarm is triggered by an external agent. The action of the agent disconnects the $Alarm$ variable from its original causes $Burglary$ and $Earthquake$, hence $P(Burglary|do(Alarm), Earthquake) = P(Burglary)$. A complete methodology to carry out such variable inference problems using causal networks (implemented with BNs or SEMs) has been developed by Pearl and his collaborators under the name of "do-calculus" (Pearl, 2000).

MODEL INFERENCE

While variable inference is an important aspect of causal inference with a well developed set of algorithms, model inference has recently become the focus of interest. In that realm, machine learning has various important roles to play:

- **The finite sample case.** Traditionally in the causal discovery community, algorithms for learning causal network structure have been developed with the assumption that there exists an "oracle" having perfect knowledge of the data distribution, and which is capable of answering without mistake questions about **conditional independence** between subsets of variables. This implicitly make the assumption that an infinite amount of training data are available. This raises the questions of developing robust and powerful statistical tests of conditional independence (Margaritis and Thrun, 2001). Kernel methods have moved in this direction (Gretton et al., 2005).
- **Feature and model selection.** Another tradition of the causal discovery community is to dismiss cross-validation for model selection and focus on penalty-based cost functions (most often using Bayesian priors) for reasons alluded to in Section 3.2. Yet, as demonstrated in the "causation and prediction" challenge, regular feature selection methods and cross-validation can take you very far to prune feature space (Guyon et al., 2008). Also, purely frequentist penalty-based model selection methods based on regularization, which have been developed in machine learning, may provide effective means of causal model selection (Pellet and Elisseeff, 2008; Lozano et al., 2009). In the problem of cause-effect pairs for instance, where constraint-based methods using conditional independence tests are not applicable, such methods have proved to be effective (see the papers of Mooij and Janzing and that of Zhang and Hyvärinen in these proceedings).

- **Learning algorithms.** There is a wealth of algorithms developed in machine learning, which can find applications in learning causal models. We saw recently the application of the ICA algorithm to learning SEMs with non-Gaussian noise for linear models (Shimizu et al., 2006), extended in these proceedings to non-linear models by Zhang and Hyvärinen. Recent methods also use non-linear regression techniques to distinguish between cause and effect (Hoyer et al., 2008). Novel methods for identifying latent confounders use a combination of nonlinear dimensionality reduction and kernel dependence measures (Janzing et al., 2009).

4.3. Examples of applications

Recently, there has been a surge of interest in causal models in data mining, prompted by the need of assisting policy making and the availability of massive amounts of "observational data". Examples of applications of causal models include: biology, medicine and pharmacology (Oniśko et al., 1997; Herskovits and Dagher, 1997; Friedman et al., 2000; Kononenko, 2009), epidemiology (Aickin, 2002), climatology (Chu and Glymour, 2008), social and economic sciences (Kaplan, 2000; Demiralp and Hoover, 2003; Moneta, 2005), marketing (CFMDCY, 2006), neuroscience (Ding et al., 2006; Neves et al., 2008), psychology, law enforcement and crime prevention (Young, 2008), manufacturing,, quality control, and fault or security diagnosis (Qin and Lee, 2003; Kraaijeveld and Druzdzel, 2005). Among the most prominent applications, which have taken off in the past decade, uncovering regulatory networks of chemicals in living organisms and connecting those networks to disease, has been the object of much research. For a rather extensive bibliography, see (Markowetz, 2007). Epidemiology has long been one of the main areas of application of causal modeling (Rubin, 1974; Herskovits and Dagher, 1997; J.M. Robins, 2000). Epidemiologists have also embraced the new tools of genomics and proteomics to investigate gene-environment interactions (Vinei and Kriebel, 2006; Jenab et al., 2009).

5. Assessment of causal solutions

A second objective of the NIPS 2008 workshop was to find means of assessing the performances of solutions proposed to causal problems. We present in this section assessment methods, which have been used in our challenges, and point to other methods of interest.

5.1. Experimental verifications

The most established way of assessing causal theories is to carry out randomized controlled experiments to test hypothetical causal relationships. Fisher's book "The Design of Experiments" in 1935 laid the mathematical foundations for experimental design. The central idea is the systematic use of randomization to avoid confounding.

For example, in the medical domain, a causal relationships $C \rightarrow E$ between a treatment C and an effect E may be tested in a **Randomized Controlled Trial** (RCT). Variable C may be the choice of one of two available treatments for a patient with lung

cancer and E may represent 5-year survival. If we randomly assign a large number of patients to the two treatments by flipping a fair coin and observe that the probability distribution for 5-year survival differs between the two treatment groups, it may be concluded that the choice of treatment causally determines survival in patients with lung cancer. The double blind placebo-controlled Randomized Controlled Trial, where allocations are randomized and neither patient not doctor knows which treatment has been assigned, is now standard in clinical trials. In agriculture, complex experiments in which many factors are controlled simultaneously are commonly performed. Unfortunately, experimenting is a long and costly process, and, in many domains it is impractical or infeasible.

An ideal benchmark of causal discovery methods (uncovering causal relationships from observational data) would compare predictions obtained by applying algorithms to large observational databases with the outcome of well designed experimental studies. Because of the rarity of adequate observational data sets paired with appropriate randomized experiments, to our knowledge no such comparisons have been made.

The Causality Workbench project has started a program of benchmarks in which realistic simulated systems will be used for generating observational data and performing virtual experiments (Guyon et al., 2010). In the "causation and prediction challenge" (Guyon et al., 2008), we used matched sets of artificially generated data for various tasks: a training dataset drawn from a "natural" **unmanipulated distribution** and several test sets drawn from various types of **post-manipulation distributions**.

We present alternative evaluation methods in the following sections.

5.2. Established ground truth

Second best to pairing observational studies and the outcome of designed experiments is to compare causal relationships inferred from observational data to **ground truth** established from human expertise (see glossary). This method has been used for instance by Cooper and Spirtes, 1998 (Spirtes et al., 2000, page 369) to compare cause-effect relationships inferred from a database on hospitalized pneumonia patients to expert medical judgement. Here are a few examples of cause-effect pairs tested in this study: *Coronary artery disease* $\rightarrow$ *Myocardial infection*, *Employment status* $\rightarrow$ *Illegal drug abuse*, *Nausea* $\rightarrow$ *Vomiting*, and *Number of comorbid conditions* $\rightarrow$ *Dire outcome*. In the pot-luck challenge organized for NIPS 2008, one dataset used human judgement as ground truth: the CauseEffectPairs dataset. Examples include the pairs *Altitude* $\rightarrow$ *Temperature* and *Longitude* $\rightarrow$ *Precipitation* in German cities and *Age* $\rightarrow$ *Length* for the snail Abalone.

In biology, regulatory pathways obtained by curating thousands of peer reviewed papers constitute reference human knowledge for discovery studies performed with genomic and proteomic observational data (Kanehisa et al., 2008). In the pot-luck challenge, the CYTO dataset is a good example using this type of ground truth. Note, however that due to many inconsistencies in the biological literature there is a lot of uncertainty in the reference regulatory pathways.

Using artificially generated data is another way of having access to an established ground truth (*i.e.,* the structure of the data generative model). In the NIPS 2008 challenge, several datasets resorted to this means of assessment. The dataset TIED is purely artificial and was designed to illustrate a particular technical difficulty. The datasets REGED and MARTI were build from a simulator of a gene regulatory network influencing lung cancer, trained with real data. The dataset SIGNET was simulated from a set of Boolean rules representing knowledge of a plant regulatory pathway gathered from several published papers.

5.3. Statistical tests

We regroup in this section a variety of techniques making solely use of observational data to *validate causal structures* using some statistical argument. We think of such methods as the weakest way of validating causal relationships, yet they are much useful because there are often no better alternatives.

1. **Validation of theoretical models by hypothesis testing.** Statistical hypothesis testing is used as "confirmatory analysis" (not for structure discovery via tests of conditional independence) in social sciences, psychology, and econometrics to validate theoretical models proposed by experts. The parameters of a causal model (typically a SEM) whose structure is determined from domain knowledge, are fitted to data. In ordinary least square regression (with several input features that represent alleged causes and a single target variable), the residuals of the model are compared to the residuals of a null model (*e.g.,* the expected value of the target, another previously proposed model, or, for time series, an autoregressive model). Statistical tests used to perform such comparisons include the Chi-square test. The tested model is invalidated if its predictions cannot be found statistically significantly better than those of the null model. Individual parameters of the model can also be examined within the estimated model in order to see how well the proposed model fits the driving theory.

   For structural equation models (SEMs) assuming Gaussian noise models, the parameter calculations are based on the covariance matrix of the variables. Goodness-of-fit is based on comparing the observed covariance matrix with the covariance matrix estimated by the model. In the early literature on SEMs, analysts tested simply the null hypothesis that the specified model leads to an exact reproduction of the observed covariance matrix with a chi-square test, but this was later replaced by a comparison with the predictions of a null model (*e.g.,* a baseline model assuming that all variables are uncorrelated) (Bollen and Long, 1992). Recently, methods for testing structural parts of a model rather than the whole model have been proposed, providing a more detailed and insightful validation (Tsamardinos and Brown, 2008).

   Another type of test investigates whether the explanatory variables and the error terms are statistically independent, as recently used in (Shimizu et al., 2006;

Hoyer et al., 2008), and by Kun Zhang and Aapo Hyvärinen in these proceedings. Since these dependencies are typically non-linear, tests must be able detect higher-order dependencies, not just simple correlations, . Kernel-based methods like HSIC (Gretton et al., 2005) seem to be useful for this task.

It is important to remember that if such methods are to be used for structure validation, the structure of the tested model should not be obtained from the data used for testing (otherwise it is like testing on training data).[1] Also, a model passing such a test is not confirmed, but rather it is not rejected, because the evidence obtained from observational data is usually insufficient to confirm a causal model. The tested model should have falsifiable implications, which can be tested against the data.

2. **Instrumental variables.** In econometrics, epidemiology and related disciplines, the method of instrumental variables is used to estimate causal relationships when controlled experiments are not feasible. In attempting to estimate the causal effect of some variable C on another E, an instrument is a third variable I which affects E only through I's effect on C: $I \to C \to E$. The method can be thought of as a "natural" experiment in which the instrument variables play the role of the "external agent". The success of the method hinges on the selection of suitable instruments. For instance, Cooper and Spirtes, 1998 (Spirtes et al., 2000, page 372) used *race*, *age*, and *gender* as an instruments in the determination of cause-effect pairs in the example of pneumonia covariates mentioned in the previous section. In Section 6.2, we give examples of Mendelian randomization in which naturally occurring *gene mutations* are used as instruments to manipulate the level of certain proteins in blood.

   Other natural and quasi-natural experiments of various types are commonly exploited, for example (Miguel et al., 2004) use weather shocks to identify the effect of civil conflict on economic growth. Jared Diamond (Diamond, 1997) defends the thesis of the influence of climate and natural resources on societal development (including food production *vs.* hunting and gathering) using a natural controlled experiment: the scattering of populations of homogeneous ancestry over a relatively short period of time in the widely diverse Polynesian islands.

3. **Re-simulation and model architecture stability.** The consistency of the findings obtained by causal discovery algorithms on real data may also be tested by "re-simulation". The re-simulation method consists in: (1) Training a data generative model with real observational data; (2) Generating simulated datasets with the model under various noise conditions; (3) Training new models for every the

1. This section focusses on model assessment or "validation" (testing), not on model selection, which we consider part of training. Statistical tests are also used sometimes for model selection. For instance, nested models with increasing numbers of variables may be created and pvalues may be computed. This can be understood as testing a model not only against a single model, but against all simpler models. Pvalues must be adjusted correctly to take into account the multiple testing problem.

simulated dataset; (4) Studying the model stability with respect to its architectures and its predictions made under manipulation. This methodology was used by Statnikov and collaborators (Aliferis et al., 2006) on the problem of lung cancer. The REGED dataset used in our challenges emerged from this study, but re-simulation was not used as an assessment method is the challenge.

Re-simulation is a variant of an assessment methods often used for clustering algorithms in which the stability of the model under various perturbations of the data is studied (Ben-Hur et al., 2002). Perturbations may include resampling the training dataset or adding noise to the input variables. Clustering and other unsupervised learning methods including principal component analysis and factor analysis can be thought of as latent causal constructs (the latent variables or cluster centers being alleged hidden causes).

4. **Probe method.** Yet another type of method of assessment, very popular in the field of variable or feature selection, is to introduce in real data a number of artificial "distracter" variables called "contrasts" (Tuv et al., 2006) or "probes" (Stoppiglia et al., 2003; Guyon and Dreyfus, 2006), which are, by construction, not predictive of a target variable of interest. In the first causality challenge (Guyon et al., 2008; Guyon et al., 2008), we extended this method to the assessment of causal discovery algorithms.

   The use of probes is relatively straightforward for "regular" feature selection from i.i.d. data, with the goal of selecting predictive variables of a given target variable, regardless of causal relationships. In statistics, for algorithms providing a ranking of variables in order of relevance, it is standard to compare the index of ranked variables to the index of hypothetical variables (called probes) drawn from a null distribution representing irrelevant variables (Guyon and Dreyfus, 2006). In this way, one can test the null hypothesis that variables are irrelevant. For instance, assume that our target variable is binary (*e.g.,* the patient health status "cancer" or "healthy") and that we want to determine whether a given predictor variable of mean μ is individually predictive of the target (univariate association). A possible null hypothesis may be that variables are drawn from a Gaussian distribution of mean μ and the alternative hypothesis may be that it is drawn from a mixture model of two Gaussians with different means (but same variance). The t-test may then be used to test the hypothesis of *equality of the means of the two classes* and determining *whether the predictor variable of interest significantly separates the two classes*. Choosing the right ranking criterion and a good null distribution has been the object of a lot of study and there is no one-size-fit all solution (see Guyon and Dreyfus, 2006, for a review). A completely non-parametric solution to the problem is to select a well suited ranking criterion, not corresponding to any known tabulated statistic (*e.g.*, the Relief criterion Kira and Rendell, 1992), then to generate random "probes" by permuting the values of randomly chosen real variables. In this way, the marginal distribution of the probes mimics that of the real variables, but the randomization

of the order of the values make them independent of the target variable. This method bears resemblance with permutation tests (Pitman, 1937). It is widely applied in genomics.

Extending the idea of probes for the problem of "causal" feature selection is not as simple as it may seem. We move from the relatively simple question of separating "relevant" from "irrelevant" features to a multi-class problem including "causes" of the target, "effects" of the target, "confounded" variables and "unrelated" variables. Suppose for simplicity that we only want to determine *whether an algorithm correctly uncovers causes of a target variable*. "Irrelevant" variables include "unrelated" variables, "effects" and "confounded" variables. So, to test the efficacy of an algorithm to uncover causes of the target, we must introduce artificial distracter variables (probes) of several kinds. Specifically, we need to construct variables with a "null mechanism" (*e.g.,* a function plus some noise or a posterior distribution), taking as input subsets of the available real variables (including eventually the target) and previously constructed probes. This ensures that no probe will be a cause of the target, but that some will be predictive and some not.

One way of assessing the validity of a proposed set of causes of the target is to compute the fraction of probes (all non-causes of the target) in that subset. Large fractions of probes shed doubt to the validity of the proposed causes. The probability of getting a number of probes smaller than a certain threshold can serve as a basis for a statistical test.

In the causality challenges that we organized, we assessed "causal relevance" using the probe method. Algorithms were required to return an ordered list of variables, with, for instance, all causes coming first in order of preference or confidence. If the truth values of the causal relationships had been known, this ranking could simply have been evaluated with the Area Under the ROC curve (AUC, the area under the curve plotting the fraction of correctly detected causes *vs.* the fraction of false alarms, when a threshold on the number of top ranking causes is varied). Instead, we used the probe AUC (called *PAUC*) as a proxy (*correctly detecting causes* being replaced by *correctly excluding probes*). In (Guyon et al., 2008), we prove that, if the null distribution used to generate the probes is correct, in the limit of an infinite number of probes, we have $PAUC = (n_+/n_r)AUC + 0.5n_-/n_r$, where AUC is the true AUC (which cannot be computed) and n_+ and n_- are the unknown numbers of positive examples (causes) and negative examples (non-causes) for the $n_r = n_+ + n_-$ real variables. Hence, asymptotically *PAUC* is monotonically related to the real AUC and therefore it can be used as a proxy to assess the relative performance of models.

The introduction of probes among the real variables induces a perturbation, which may distort the causal discovery problem (*e.g.,* by creating spurious conditional dependencies between the target and real variables). These perturbations may alter the real cause-effect relationships in unsuspected ways. Hence, for

discovery, we recommend to re-run the algorithm on real data only, without the addition of probes.

6. Discussion: Failure breeds success

The old timers of machine learning and artificial neural networks will remember that the field has long been traumatized by the XOR problem. In the 1960's, Frank Rosenblatt, Bernard Widrow and others introduced various training algorithms for one layer neural networks. In 1969, Minsky and Papert in their book on Perceptrons (Minsky and Papert, 1969), inventoried problems, which were "non linearly separable", *i.e.,* could not be solved with one layer neural networks. The archetype or such problems is the XOR problem: the Boolean function XOR is not linearly separable. The book had a great impact and put the field of artificial neural networks in dormancy for nearly 20 years. During its revival in the 1980's when algorithms to train multi-layer Perceptrons emerged, no paper on artificial neural networks failed to address the XOR problem. It is worth noting though that linear discriminant functions are tremendously useful and failing to solve the XOR problem is not an indication that a learning machine is useless. For example, in the 1990's, the non-linear Support Vector Machine was invented (Boser et al., 1992), which brought attention to its linear version dating back from the 1960's. The linear SVM is now a very widely used method in text processing and bioinformatics.

The field of causal discovery has many problems similar to the XOR problem. However, neither solving them nor failing to solve them is necessarily an indication that the methods will not perform well in real world applications. While such problems should be used as tools to improve our methodology and we also should constantly remind ourselves that "failure breeds success" and that stumbling on any of these problems does not mean that unraveling causal relationships is a hopeless task and much less that causality is a useless concept. In this section, we first play devil's advocate and give 10 reasons why causal discovery might be a hopeless enterprize. Then, we tell 10 success stories proving the pessimists wrong. Finally, we list 10 open problems on which researchers are still stumbling.

6.1. Ten challenging problems

Several papers in these proceedings present cases in which common assumptions made are violated or cases in which common causal models either find spurious causal relationships or fail to uncover existing ones. Most of these problems are discussed thoroughly in causality textbooks (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). We present briefly ten of them.

1. **No formal definition of causality.** There is so far no formal mathematical definition of causality. Two approaches attempt to fill this vacuum: (1) Operational tests of causality (Glymour and Cooper, 1999) allow us to detect causality experimentally using controlled experiments, but they provide only sufficient criteria for causality, not necessary conditions (see Section 3.1). (2) Data generative

models propose ways in which variables values may be generated from each other using defined mechanisms. Algorithms, which can reconstruct the architecture of a model using data generated by that model are called "causal discovery algorithms".

2. **Statistically dependent is not the same as correlated.** Two random variables X and Y are called independent if $P(X,Y) = P(X)P(Y)$ which is a stronger condition than absence of correlation, *i.e.,* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. This distinction is often overlooked by causal discovery algorithms, which use correlation as a symptom of causation instead of statistical dependency. Non-linear mechanisms can generate dependencies without any correlation (although, in typical cases, dependent variables are at least weakly correlated). Reciprocally, (partial) correlation can arise in the absence of (conditional) statistical dependence: Partial correlations are given by the correlations of the residuals after *linear* regression. If X and Y are non-linear functions of Z (up to noise terms independent of each other and of Z), only non-linear regression would render them independent and uncorrelated. Hence X and Y can remain partially correlated, given Z, even though they are conditionally independent, given Z. Therefore, neglecting the possibility of non-linear mechanism and using statistical tests based only on correlation can lead both to false negative and false positive dependencies.

3. **Statistical dependence does not imply causation**[2]. According to the **principle of common cause** (PCC), every statistical dependency between two random variables X and Y has a causal explanation. Reichenbach (Reichenbach, 1956) formulated the following three (not necessarily exclusive) cases: (1) X causes Y, (2) Y causes X, or (3) there is a third variable Z (common cause or **confounder**) causing both X and Y. In this last case, conditioning on Z renders X and Y independent, if cases (1) and (2) do not hold. For instance, assume that "chocolate intake" (variable X) is found to positively correlate with "life expectancy" (variable Y). This does not necessarily imply that eating more chocolate will improve your chances of living longer. It is possible that in fact "gender" (variable Z) affects both "life expectancy" (females live longer) and "chocolate intake" (females eat more chocolate), but that in each "gender" sub-population (male or female) there is no dependence between "chocolate intake" and "life expectancy" (Simpson's paradox).

   The problem that confounders are often unobserved, unobservable or even unknown and that Z can be a high-dimensional vector of relevant factors, is one of the main obstacles of causal inference from **observational data**. Often it is even hard to quantify latent factors such as subject's personality and physical condition in a medical study. In other words, there is no way for reliably deciding

2. In light of the previous item, we avoid using the terser motto "correlation does not mean causation".

whether the set of observed variables is **causally sufficient** (*i.e.,* does not exclude any common cause of any pair of variables)[3].

For causally sufficient sets of variables, the postulate of the **causal Markov condition** (CMC) provides a practical principle for selecting candidate causal structures from observational data, by providing conditions under which statistical dependency may be linked to causality. Several equivalent versions of the CMC exist. The most commonly used version postulates **conditional independence between every variable and its non-effects, given its direct causes**. Pearl justified the CMC by a model of causality where every variable is a function of its direct causes and a noise variable that renders the causal mechanism probabilistic (**structural equation model** or SEM). Then the CMC follows, assuming joint statistical independence of the noise terms[4]. The most common violations of the CMC arise from violations of causal sufficiency or existence of correlated noise. In deterministic systems, violations of the CMC may result from the existence of constraints (such as conservation of mass, energy, or momentum); a classical example is that of the trajectories of two billiard balls hit by a third one (see the paper of Lemeire and Steenhaut in these proceedings).

4. **"Faithfulness" is not always justifiable by "stability".** This sentence is a shorthand to bag together a variety of related hypotheses commonly referred to as Causal Faithfulness Condition (CFC). While the CMC essentially states that dependency implies the existence of a causal arrow, the CFC states the opposite, namely that independence implies no causal arrow. The CFC is more controversial than the CMC and it is the XOR problem of causality. Imagine two identical fair coins tossed simultaneously and let us call X_1 and X_2 the binary random variables corresponding to the outcome (heads or tail). Consider a outcome Y, which is whether or not both coins fell on different sides. Note that the logical relation $Y = X_1 \text{XOR} X_2$ is fulfilled. Since both X_1 and X_2 are individually independent of Y (and independent of each other), according to the CFC one should not draw any causal arrow. However, clearly, there is a joint dependency between X_1, X_2 and Y. In the causal network framework one could represent the dependency with the unfaithful graph $X_1 \rightarrow Y \leftarrow X_2$, but the representation $[X_1, X_2] \rightarrow Y$ might be more suitable since the two variables jointly cause Y. Other classical examples of faithfulness violation for non-binary variables include cases in which two causal paths exactly cancel each other with a particular choice of parameters. In either case (XOR or canceled causal path) the "stability" argument in favor of the CFC

3. See section 6.2 in (Pearl, 2000), called "Why there is no statistical test for confounding, why many think there is, and why they are almost right".

4. There is a tight relation between CMC and PCC: The conditional independence of two effects, given their common cause, is just a special case of CMC with three variables. It can also be argued that the independence of noise terms in Pearl's model corresponds to an absence of common noise-generating mechanism, which follows from the PCC. Spirtes et al. (2000) proposed a weak causal Markov assumption, similar to the converse of the PCC, stating that if X and Y have no common cause (including each other), they are probabilistically independent. This weaker assumptions implies the CMC for SEMs

is that if there is the smallest defect in the generative process (a coin not exactly fair or parameters not exactly tuned to cancel the causal paths), then the symmetry is broken and faithfulness is re-established. Critics of the CFC point out that, in practice, small asymmetries are difficult to detect from empirical data and that there are many systems in which there is an equilibrium leading to canceled causal paths (see for instance the paper of Voortman, Dash, and Druzdzel in these proceedings). Hence, many technical systems like systems of logical gates easily violate faithfulness.

5. **Markov equivalences.** Many causal graphs may generate identical probability distributions or at least entail the same set of conditional independencies between variables (Markov equivalent graphs). For instance $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \to Z$ all have the same unique Markov property that X and Z are independent given Y. Most structure learning algorithms (from observational data) rely on the existence of so-called unshielded colliders of the form $X \to Y \leftarrow Z$, which do not have any other Markov equivalent graph. Such methods can unravel causal relationships in systems of at least three variables, up to Markov equivalent graphs. Hence, they are not applicable to the problem of cause-effect pairs. Recent methods have addressed this problem, such as the solutions proposed in these proceedings to the CauseEffetPairs task.

6. **Model selection.** When learning from observational data, classical cross-validation is not very useful to perform model selection since predictions are to be made on data from a different, post-manipulation, distribution. Hence, penalty-based methods like AIC (Akaike, 1973) or BIC (Schwarz, 1978) are sometimes used to drive model choices toward fewer parameters or minimal architectures. Yet, obviously, minimal models are not always the best. For an analysis, see the paper of Lemeire and Steenhaut in these proceedings.

7. **Measurement errors, quantization, and aggregation distort dependencies.** In his presentation at the NIPS 2008 workshop, Richard Scheines gave several examples in which measurement errors or data quantization limit causal discovery. For instance, a causal system of three variables X, Y, and Z may have the Markov property that X is independent of Z given Y (*i.e.,* one of these three graphs is valid: $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, $X \leftarrow Y \to Z$), and yet, this Markov property may go undetected if Y is observed through a noisy or quantized version Y' (technically, Y' is a consequence of Y and therefore it does not d-separate X and Z). Similarly, variables X, Y, and Z may be the result of averaging over populations $X = \sum_i X_i$, $Y = \sum_i Y_i$, and $Z = \sum_i Z_i$. So, even though X_i might be independent of Z_i given Y_i for every i, it is possible that the property does not hold for the average.

8. **Sample bias and attrition bias plague experimental design.** The validity of randomized experiments relies on the quality of randomization. Spurious relationships may be found because of sampling. For instance, it may be found that there is a correlation between pregnancy and flu. If the patients were sampled

only from an emergency room, this may simply indicate that patients with acute nausea or vomiting symptoms arising from multiple conditions are more likely to show up in the emergency room, not that the two conditions are causally related or have a common cause. Sample bias plagues retrospective studies, which analyze observational data collected without any particular design. Prospective longitudinal studies following patients over a period of time are usually less prone to sample bias because they are more carefully designed, but they are prone to attrition bias (some patients quit the study before the end, for instance when a treatment has undesirable side effects.)

9. **Markovian causal graphs do not represent suitably all data's generative processes.** Directed Acyclic Graphs (DAGs) cannot represent cyclic systems, by definition. This can be remedied by unfolding cycles in time, which, for discrete time systems amounts to using a classical Markov model. But, symmetric relationships (such as gravitational or electrical forces) or constraints (such as energy, mass and momentum conservation) are not suitably represented by arrows (which are usually interpreted as directional relationships). Accordingly, a given event Y may simultaneously generate multiple related consequences (a classical example is that of the billiard ball hitting two balls simultaneously). The notation $X \leftarrow Y \rightarrow Z$ suggests that X and Z are generated by Y from two independent mechanisms, rather that a single mechanism with underlying constraints. A new notation such as $Y \rightarrow [X,Z]$ may be more suitable. See the paper of Lemeire and Steenhaut in these proceedings for a discussion of this issue.

10. **Causality in time series is not necessarily an easier problem.** Causality is commonly thought of as a time-related concept (causes precede their effects). So how can causality in time series be harder to investigate than causality in time independent data? On one hand, the problem is indeed simpler because events that took place in the future may be pruned from the set of candidate causes of an event. Thus temporal causal models use only past values of variables to predict future values. On the other hand, modeling can be harder (i) if the time series are non-stationary (spurious correlations are easily found), (ii) if the variables are measured in presence of noise (see the NOISE dataset in these proceedings), (iii) if data are scarce (overfitting problems can be severe since the data points are not independent, therefore more data points are required than for i.i.d. data), (iv) if experiments are not properly designed (in particular, the non-commutativity of equilibration and manipulation might complicate matters, see the paper of Voortman, Dash, and Druzdzel in these proceedings).

This list is pretty scary, although non-exhaustive. On top of that, the availability of (quality) data, particularly experimental data, is usually limited. Causal models are perhaps even more prone to overfitting than regular predictive models, because in addition to estimating dependencies, one must estimate the direction of the causal relationships. When there is enough data, causal models suffer from a high computational complexity.

Hence for a large number of variables, sub-problems must usually be solved (*e.g.,* focusing on the local neighborhood of a variable). And yet, there are success stories!

6.2. How causal conclusions changed our life - ten stories

Researchers working on causal inference are often confronted with three kind of objections:

(1) Philosophical concerns about **whether causality is a well-defined scientific concept**. In 1913, Bertrand Russel stated "the law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (Russell, 1913). On one hand, this perspective seems to be supported by the way many physical laws are formulated, *e.g.,* as time-inversion symmetric differential equations in space time in Einstein's theory of special relativity, just discovered at Russel's time. On the other hand, physics can also be seen as predicting how outcomes of experiments depend on the experimental setup, which is an inherently causal formulation.

(2) Scepticism about **whether causal conclusions can be drawn from non-interventional observations**. The most radical version of this concern would be the belief that only randomized controlled studies yield valuable causal conclusions. A more moderate version, which probably many statisticians would agree to, states that causal inference from non-randomized studies relies essentially on background knowledge of the domain of the data, which makes it part of the respective field rather than being part of statistics.

(3) Scepticism about **whether causal discovery from observational data can be mathematically formalized** up to a degree that admits the implementation of reliable inference algorithms. There is, however, no clear boundary between (2) and (3) because the way the input of an algorithm is specified can contain an arbitrary amount of prior knowledge. For instance, how to formalize observations in terms of random variables already involves human judgements about which representation is natural for the respective problem – a decision that also occurs in other machine learning tasks.

The practical relevance of concern (1) is questionable since an essential part of scientific and technological progress consists in deriving *causal* statements as opposed to purely predictive ones because they are the only results that provide criteria for human actions. The examples of this section illustrate how causal insights from different scientific disciplines already influenced both private decisions and those in public health, economy and politics. Historical examples pre-dating the computer age involved more human reasoning than computerized data analysis, but we include them because of their exemplary nature. They also show that progress has been made by exploiting non-interventional data and not only by randomized control studies, which responds to concern (2). The more recent examples, including the practical impact of Granger causality, the use of instrumental variables in genetic studies (via Mendelian random-

ization) and successes of Bayesian network in biology and SEMs in social sciences respond to concern (3).

In selecting our success stories, we have applied very stringent criteria, which prevented us from including many promising on-going efforts mentioned in Section 4.3 (either because the conclusions have not yet been sufficiently validated or because their socio-economic impact has not been evaluated). Consequently, many recent algorithms are not yet illustrated in these success stories. However, in response to concern (3), it should be emphasized that causality research must not be reduced to developing algorithms (even though this is an important part). Human causal reasoning requires rationales to rely on. The goal to develop automatic causal discovery has already created a conceptual clarity (Pearl, 2000; Spirtes et al., 2000) that many previous discussions were lacking. Human causal inference also requires reliable criteria that state-of-the-art statistics do not provide. Pioneering work from (Janzing and Schölkopf, 2008) present a formal basis for causal inference that also work with single observations rather than relying on statistical ensembles. In deriving further principles, one should be encouraged by the following successes of causal thinking in science.

1. **Vitamin C and scurvy: A historical RCT.** Observational epidemiology and controlled experiments have revolutionized our understanding of causal risk factors predisposing to a variety of common diseases. While at sea in May 1747, a ship surgeon of the British Royal Navy, James Lind, provided some crew members affected by scurvy with two oranges and one lemon per day, in addition to normal rations, while others continued on their regular diet. In the history of science, this is considered to be the first occurrence of a controlled experiment comparing results of two populations where one factor is applied to one group only with all other factors the same. Following this discovery, in 1795 the Royal Navy provided a daily ration of fresh lime or lemon juice to the sailors and successfully fought scurvy. It is now established that citrus fruits contain Vitamin C, which is necessary for the treatment and prevention of scurvy. However, there is continuing debate within the scientific community over the best dose schedule of vitamin C for maintaining optimal health in humans and whether overdose may have adverse effects.

2. **Hygiene and infectious diseases: Can you believe what you can't see?** It is hard to believe that the use of basic hygiene precautions was at some point fought by the medical establishment. Yet, when the Hungarian physician Ignaz Philipp Semmelweis discovered in the 1840's that cases of puerperal fever (childbed fever) could be cut drastically if doctors washed their hands in a chlorine solution before gynaecological examinations, he was ridiculed and harassed. The validation of the germ theory by Pasteur's experiments in the 1860's was necessary before the cause-effect relationship between hygiene and infectious diseases was accepted. He exposed freshly boiled broth to air in vessels either directly exposed to air or protected by a filter stopping all particles. Nothing grew in the protected broths, therefore the living organisms that grew in unprotected

broths came from outside (as spores on dust) rather than being generated within the broth. This initial work stimulated the development of techniques to kill germs in beverages (Pasteurization), protocols of antiseptic surgery, and immunization methods (vaccination). With the advent of more powerful microscopes and the progresses made in microbiology, a large body of work now supports that the underlying mechanisms of infectious diseases involve germs, which can be killed with anti-bacterial agents, thus providing an explanation for the causal link between hygiene and infectious diseases.

3. **Crop yield optimization in agriculture: Mathematical foundations of experimental design.** The first statistician to consider a formal mathematical methodology for designing experiments was Fisher, in his book "The Design of Experiments" (1935). He developed his methodology while working at the Rothamsted Experimental Station (England), one of the oldest agricultural research institutions, founded in 1843. Partly through these methods, researchers at Rothamsted have made significant contributions to agricultural science, including the discovery and development of systemic herbicides and pyrethroid insecticides, as well as pioneering contributions to the fields of virology, nematology, soil science and pesticide resistance. During World War II, aiming to increase crop yields for a nation at war, a team under the leadership of Judah Hirsch Quastel developed 2,4-D, still the most widely used weed-killer in the world. In medicine, the double blind Randomized Controlled Trial (RCT), where allocations are randomized and neither patient not doctor knows which treatment has been assigned, is now a standard experimental design in clinical trials.

4. **The smoking ban and lung cancer: Better err on the safe side.** Prior to World War I, lung cancer was considered to be a rare disease, which most physicians would never see during their career. With the postwar rise in popularity of cigarette smoking, however, came an epidemic of lung cancer. In 1950, Richard Doll undertook with Austin Bradford Hill a study of lung cancer patients in 20 London hospitals, at first under the belief that it was due to the new material tarmac, or motor car fumes, but rapidly discovering that tobacco smoking was the only factor they had in common. Sir Ronald A. Fisher and other statisticians opposed the conclusions of Doll and Hill that smoking caused lung cancer on the ground that correlation does not imply causation. For instance, there may be an unknown genetic factor, which causes both lung cancer and craving for tobacco. Many studies followed (see Spirtes et al., 2000, page 239 for a detailed account), eventually leading to tobacco smoking bans in public places in several countries. Interestingly, the results of controlled studies on the effect of smoking on lung cancer are mixed, but there is a large consensus that the smoking ban reduced heart disease (Sources include, the US National Cancer Institute and the American Lung Association).

5. **NSAIDs, drug efficacy and drug toxicity.** Non-steroidal anti-inflammatory drugs (NSAIDs) include some of the most commercially successful drugs like

Aspirin or Tylenol. They are used to treat pain, fever and inflammation. Most NSAIDs act as non-selective inhibitors of the enzyme cyclooxygenase, which catalyzes the formation of prostaglandins, messenger molecules in the process of inflammation causing pain and fever. This mechanism of action was elucidated by John Vane, who later received a Nobel Prize for his work in 1982. Medicines containing derivatives of salicylic acid, structurally similar to aspirin, have been in medical use since ancient times. A French chemist, Charles Frederic Gerhardt, was the first to prepare acetylsalicylic acid in 1853. In 1899, Bayer patented it for its use as a drug under the name Aspirin. Aspirin's popularity grew over the first half of the twentieth century, spurred by its effectiveness in the wake of the Spanish flu pandemic of 1918, and aspirin's profitability led to fierce competition and the proliferation of aspirin brands and products, especially after the American patent held by Bayer expired in 1917. Aspirin is no longer used in children and adolescents due to the risk of Reye's syndrome; paracetamol (the international non-proprietary name for the drug Tylenol) is now often used instead. In 1887 the clinical pharmacologist Joseph von Mering first tried paracetamol on patients. In 1893 he published his results comparing paracetamol with phenacetin, another aniline derivative, claiming that, unlike phenacetin, paracetamol had a slight tendency to produce methemoglobinemia (abnormal oxidation of hemoglobin to methemoglobin, reducing the oxygen transport capabilities of red blood cells). The toxicity of paracetamol was not challenged until the late 1940's when it was shown that phenacetin metabolizes to paracetamol (von Mering's results may have been due to some impurity). Paracetamol was first marketed in the United States in 1953 by Sterling-Winthrop Co., which promoted it as preferable to aspirin since it was safe to take for children. More recently, the commercially successful NSAID Vioxx, approved by the FDA in 1999, was voluntarily withdrawn from the market by Merck in 2004 because of concerns about increased risk of heart attack and stroke. This example illustrates the intricacy of determining positive and negative effects via a combination of observational, controlled studies, and understanding of mechanisms.

6. **Genetic epidemiology: Towards personalized medicine.** Genetic epidemiology is concerned with understanding heritable aspects of disease risk, individual susceptibility to disease, and ultimately with contributing to a comprehensive molecular understanding of pathogenesis and a medicine tailored to the individuals. It is also an area of intensive causal studies. According to Kraft and Hunter (Kraft and Hunter, 2009): "A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase or decrease the risk of complex diseases. The completion of the International HapMap Project and the development of new methods for genotyping individual DNA samples at 500,000 or more loci have led to a wave of discoveries through genome-wide association studies. These analyses have identified common genetic variants that are associated with the risk of more than 40 diseases and human phenotypes. Several companies have begun offering direct-to-

consumer testing that uses the same single-nucleotide polymorphism chips that are used in genomewide studies." And, according to Goldstein (Goldstein, 2009): "More than 100 genomewide association studies have been conducted for scores of human diseases, identifying hundreds of polymorphisms that are widely seen to influence disease risk. After many years in which the study of complex human traits was mired in false claims and methodological inconsistencies, genomics has brought not only comprehensive representation of common variation but also welcome rigor in the interpretation of statistical evidence."

7. **Reverse causation and confounding resolved by Mendelian randomization.** Mendelian randomization makes a bridge between observational epidemiology studying environmental factors and genetic epidemiology. The problem of "reverse causality" occurs when the direction of a cause-effect relationship is inverted because the onset of the cause was not detectable. The problem was studied by Martijn Katan in 1986 (Katan, 2004; Keavney, 2004) for the association between low serum cholesterol levels and cancer. In this case, a pre-existing occult tumor might cause lower cholesterol levels, rather than lower cholesterol levels causing cancer (Garcia-Palmier et al., 1981). The association might also be explained by confounding factors (such as cigarette smoking) related both to future cancer risk and to lower circulating cholesterol (McMichael et al., 1984). Katan proposed a method using genetics to emulate a RCT without performing actual manipulations. His method was never tested but it was then generalized by Gray and Wheatley in 1991 (Gray and Wheatley, 1991; Wheatley and Gray, 2004; Smith, 2007) in a method called "Mendelian Randomization". The idea is to use a naturally occurring genetic polymorphism, with a well understood regulatory effect, as an instrument to manipulate a variable of interest (*e.g.,* raising blood cholesterol). Importantly, the genotype must only affect the disease status indirectly via its effect on the variable of interest (*e.g.,* blood cholesterol). Because genotypes are assigned randomly when passed from parents to offspring, the statistical dependence between the population genotype and the cancer cannot be confounded (as opposed to cholesterol, where confounding by social, behavioral or physiological factors is possible). The biggest success so far of Mendelian randomization studies were obtained using a mutation of *methylene tetrahydofolate reductase* as randomization instrument in studies of the implication of folate in coronary heart disease, fetus neural tube defects, and cancer (see Smith, 2007).

8. **System biology: Reverse engineering the cell.** One branch of system biology, which is an active area of causal studies, aims at modeling a whole cell. As part of that effort, Nir Friedman and his collaborators wrote several of the key papers using Bayes Networks for gene expression analysis and pathway modeling. This approach generalized the method of Boolean networks for pathway modeling traditionally used by chemical engineers to abstract metabolic and biochemical networks by modeling uncertainty and introducing hidden variables. For a re-

view of Friedman's work see (Friedman, 2004). For the most part, published papers in this area propose networks based on analyzing empirical data and then compare the results with the existing literature. Few papers are followed by an experimental validation of new findings. Still, these results, which incorporate global simultaneous measurements, are a good complement to results coming from other sources investigating in more details the interactions of few chemical species, including *e.g.,* via gene knockout experiments.

9. **College dropouts: Assisting policy-making in social sciences.** Using causal discovery algorithms to learn the structure of Structural Equation Models, Spirtes and collaborators have worked out a large number of problems previously published in the literature and found structures matching or closely resembling those built with expert knowledge (Spirtes et al., 2000). The examples include finding the causes of publishing probability, finding the influence of parent education on children education, and finding what influences abortion opinions. For illustration, we give an end-to-end story, which actually led to a change in policy: (Druzdzel and Glymour, 1999) performed a study at the request of the provost of Carnegie Mellon University (CMU) to investigate policies for lowering dropout rates. Using the US News and World Report database on American college and universities, they found that all variables in the database to be independent of college dropout given the results of test scores of the entering class (SAT test scores). Subsequent higher selection of students based on the SAT test results at CMU correlated with lower dropout rates (but may have been affected by other factors).

10. **Granger causality: Causality in time series.** Clive Granger and his collaborators published in 1970's and 1980's methods for determining whether some time series are useful in forecasting others. A time series $x(t)$ "Granger causes" another $y(t)$ if the bivariate model (using past values of x and y to predict y) is more predictive than the auto-regressive model (using only past values of y to predict y). This conclusion, however, is only correct if there are no instantaneous causal influences between $x(t)$ and $y(t)$ and if there is no common cause influencing both.

    Granger received the 2003 Nobel prize in economics for his work on co-integration and modeling of non-stationary time series. If both $x(t)$ and $y(t)$ are non-stationary, but some linear combination $ax(t)+by(t)$ is stationary, then $x(t)$ and $y(t)$ are said to be co-integrated. Granger proved that co-integrated time series must be in a Granger causal relationship. In spite of its limitations, Granger causality is a big leap forward as it eliminates many spurious correlation or spurious regression found by fitting models making stationarity assumptions using ordinary least squares. Granger's work has transformed the way economists deal with time-series data. Today, tests of stationarity and co-integration are carried out routinely as a stepping-stone to the specification of dynamic econometric models relating exchange rates and price levels, consumption and wealth, divi-

dends and stock prices, and interest rates of different maturities (source: Nobel web site (Granger, 2003)).

6.3. Ten open problems

Much remains to be done in the domain of causal modeling. While successful causal studies have focused primarily on systems of just a few variables, more ambitious recent endeavors have ventured to unravel causal relationships in systems of thousands of variables, facing new challenges. We give ten research directions, which we think deserve attention.

1. **Optimizing directly defined objectives.** One of the two themes of the NIPS 2008 workshop was to define objectives for causal modeling. Assuming that we made a step in the right direction, the next step will be to develop methods to optimize such objectives. In pattern recognition, the old paradigm which consisted in developing separately the building blocks of recognition systems (preprocessing, classifier, and post-processing) has made way to approaches, which globally optimize simultaneously all the parameters of the processing chain with respect to a global objective. Similarly, we anticipate that in causal modeling searching directly for optimal modes of action (policies) to attain given objectives may be easier and yield better solutions than attempting to faithfully unravel the data's generative process. The causal model would then just be a means to an end, not an end in itself. Such approaches may bridge between causal modeling, operations research and identification and control.

2. **Improving and comparing assessment methods.** The second theme of the workshop was the development and study of methods of assessment of causal models. As we pointed out in the course of the paper, the problems of model selection, model performance prediction, and model assessment are more difficult for causal models than for regular statistical models because data are not i.i.d. We briefly reviewed some assessment methods in Section 5. The next step will be to study and compare such methods (and others), eventually leading to best practice recommendations for data analysts.

3. **Understanding and modifying regularly made assumptions.** Assumptions like the CMC, causal sufficiency, the CFC, Gaussianity of the noise, linearity of the relationships, are often made out of convenience rather than out of an understanding of the data's generative process and of the possible consequences on the solution. Collecting pedagogical examples violating such assumptions should facilitate the work of data analysts and, in turn, inspire theoreticians to modify the assumptions. For instance, unfaithful distributions can arise from deterministic relations (Lemeire, 2007), which are not uncommon in nature. Finding appropriate meta-principles which imply faithfulness under specific conditions would be an option for future foundations of causal inference (see the paper of Lemeire

and Steenhaut in these proceedings). In a Bayesian setting, this task would corresponds to finding good priors on the parameter space of a Bayesian network. One could ask for abstract properties such priors should have, following earlier work of (Meek, 1995).

4. **Developing versatile regularized models.** Bayesian networks based on directed acyclic graphs (DAGs) are praised for their simplicity, but have the limitations that we mentioned in Section 6.1. Many other models have been proposed to generalize them and/or address their limitations, including partial ancestral graphs (which model uncertainties about arrow directions), Markov random fields (for bi-directional connections), cyclic and dynamic models. Linear structural equation models (SEMs) with Gaussian noise variables have been generalized to non-linear and non-Gaussian noise models. Practitioners are at a loss to determine without domain knowledge which model may be best suited and avoid either underfitting or overfitting data. It may facilitate their work to move towards general-purpose versatile causal models, and use regularization methods to bias the search for optimal structures and parameters towards simpler solutions. Efforts in this direction have stated to emerge (see Lozano et al., 2009, and the paper of Zhang and Hyvärinen in these proceedings)

5. **Developing efficient and effective algorithms.** Much progress has been made recently towards scaling up algorithms to large numbers of variables and large numbers of examples. One approach has been to make use of regular feature selection methods developed in machine learning to prune the search for causes and effects (Aliferis et al., 2003). This and other efforts in the same direction need to be pursued.

6. **Developing a methodology for feature construction.** Variable definition and coding is not innocuous in causal modeling. We have seen in Section 6.1 that variable aggregation can occlude some conditional independencies. Coding a categorical variable into several (dependent) variables using a complete disjunctive coding may result in similar problems. Hence a methodology for defining, constructing, and coding variables must be developed to guide practitioners. Steps in this direction have recently be made (Spirtes, 2008).

7. **Addressing imperfections in data.** Imperfection in data such as measurement errors, data quantization, missing values, sampling bias, attrition bias, and correlated noise may be responsible for modeling errors. While classical statistical models may degrade gracefully with such data imperfections, structural errors in causal models may yield entirely wrong conclusions as to which actions are susceptible to influence a desired outcome. Although it may not be possible to inventory all possible adverse situation, it is important to raise awareness among practitioners, find methods for diagnosing a number of classical problems, and eventually find remedies.

8. **Integrating heterogeneous information.** Merging data from a variety of sources is going to be one of the major challenge in some domains. In genomics and proteomics, for instance, understanding the role of specific genes and proteins in disease requires multidisciplinary approach. Relevant data come from sources as diverse as high-throughput tools (like DNA microarrays and mass-spectrometry), gene knock-out/knock-down techniques, protein characterization, metabolic profiling, high-content screening, phenotype, and clinical data. In medicine, it is generally admitted that the strongest evidence for therapeutic interventions is provided by systematic review of multiple Randomized Controlled Trials. The Cochrane Collaboration is a group of over 15,000 volunteers in more than 90 countries who review the effects of health care interventions tested in biomedical randomized controlled trials. There may be value in developing methods to integrate information from various sources, identify possible contradictions, and track them back to confounding factors or experimental errors.

9. **Designing studies combining observational and experimental data.** Observational studies and expert opinions are usually not considered reliable evidence, compared to controlled experiments. However, experiments being costly, time consuming and sometimes unethical or impractical, it seems that it could make sense to design studies in which both observational and experimental data would be collected, in an effort to maximize information for a given budget.

10. **Quantifying uncertainty.** Learning from a finite amount of observational and/or experimental data yields models and predictions tainted with uncertainty. Most causal discovery algorithms are justified in the infinite sample size limit. There is a need to quantify uncertainty *e.g.,* with bounds on the prediction error involving model complexity, data quality, and data quantity.

7. Conclusion

There is an intense activity in a nucleus of machine learning researchers interested in causality. We hope that this activity will result in improving techniques to unravel cause-effect relationships and expand the domain of application in areas where the number of features and variables is much larger than those usually considered in the past. At this stage, there seems to be an abundance of algorithms looking for good applications. Hence the most urgent questions are: How to get good problems? How to get good data? How to get conclusive results? For that reason, we are continuing our effort of data exchange and benchmark through the Causality Workbench project.

While we hope that our effort will lead to an improvement in methodology, we would like to borrow the wisdom of Petitti (Petitti, 2004), who makes the following four recommendations: (1) Do not turn a blind eye to contradiction. Do not ignore contradictory evidence but try to understand the reasons behind the contradictions. (2) Do not be seduced by mechanism. Even where a plausible mechanism exists, do not assume that we know everything about that mechanism and how it might interact with

other factors. (3) Suspend belief. Do not be seduced by your desire to prove your case. (4) Maintain scepticism. Question whether the factors under investigation can really be that important; consider what other differences might characterize the case and control groups. Do not extrapolate results beyond the limits of reasonable certainty.

Acknowledgments

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are very grateful to all the members of the causality workbench team for their contribution to organizing the pot-luck challenge: Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. We thank Hans Bitter, Jean-Philippe Pellet, Alexander Statnikov, and Ioannis Tsamardinos for commenting on the manuscript.

References

Mikel Aickin. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation*. Chapman and Hall/CRC, 2002.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.

C. F. Aliferis, I. Tsamardinos, A. Statnikov, and L.E. Brown. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, Las Vegas, Nevada, USA, June 23-26 2003. CSREA Press.

C. F. Aliferis, A. Statnikov, and P. P. Massion. Pathway induction and high-fidelity simulation for molecular signature and biomarker discovery in lung cancer using microarray gene expression data. In *APS Conference: Physiological Genomics and Proteomics of Lung Disease*, 2006.

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

K. A. Bollen and J. S. Long. Tests for Structural Equation Models: Introduction. *Sociological Methods Research*, 21(2):123–131, 1992. doi: 10.1177/0049124192021002001. URL http://smr.sagepub.com.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

CFMDCY. Committee on Food Marketing and the Diets of Children and Youth. *Food Marketing to Children and Youth: Threat or Opportunity*. The National Academies Press, Washington, D.C., 2006. URL http://www.nap.edu/catalog.php?record_id=11514#orgs.

Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *JMLR*, 9:967–991, 2008. ISSN 1533-7928.

Clopinet. Challenges in machine learning, 2009. URL http://clopinet.cm/challenges.

Selva Demiralp and Kevin D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65(s1):745–767, December 2003. URL http://ideas.repec.org/a/bla/obuest/v65y2003is1p745-767.html.

J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W.W. Norton and Company, 1997.

Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. Granger causality: Basic theory and application to neuroscience. *WILEY-VCH VERLAGE*, 2006:451, 2006. URL http://www.citebase.org/abstract?id=oai:arXiv.org:q-bio/0608035.

Marek J. Druzdzel and Clark Glymour. Causal inferences from databases: Why universities lose students. In Clark Glymour and Gregory F. Cooper, editors, *Computation, Causation, and Discovery*, pages 521–539, Menlo Park, CA, 1999. AAAI Press.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *RECOMB*, pages 127–135, 2000. URL citeseer.ist.psu.edu/friedman99using.html.

Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, February 2004. ISSN 1095-9203. doi: 10.1126/science.1094068. URL http://dx.doi.org/10.1126/science.1094068.

M. R. Garcia-Palmier, P. D. Sorlie, Jr. Costas, R., and R. J. Havlik. An apparent inverse relationship between serum cholesterol and cancer mortality in Puerto Rico. *Am. J. Epidemiol.*, 114(1):29–40, 1981. URL http://aje.oxfordjournals.org/cgi/content/abstract/114/1/29.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58, 1992. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1992.4.1.1.

C. Glymour and G.F. Cooper, editors. *Computation, Causation, and Discovery*. AAAI Press/The MIT Press, Menlo Park, California, Cambridge, Massachusetts, London, England, 1999.

David B. Goldstein. Common Genetic Variation and Human Traits. *N Engl J Med*, 360 (17):1696–1698, 2009. doi: 10.1056/NEJMp0806284. URL http://content.nejm.org.

C.W.J. Granger. Statistical methods for economic time series, 2003. URL http://nobelprize.org/nobel_prizes/economics/laureates/2003/public.html.

Richard Gray and Keith Wheatley. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant*, 7(Suppl. 3):9–12, 1991.

A. Gretton, O. Bousquet, A. Smola, and B. Schoelkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT 2005*, pages 63–78, 10/08/ 2005.

I. Guyon and G. Dreyfus. Chapter 2: Assessment methods. In I. Guyon etal Eds., editor, *Feature Extraction, Foundations and Applications*, Series Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer, 2006.

I. Guyon, C. Aliferis, and A. Elisseeff. Causal feature selection. In Huan Liu and Hiroshi Motoda, editors, *Computational Methods of Feature Selection*, pages 63–82. Chapman and Hall/CRC Press. Longer TR: http://clopinet.com/isabelle/Papers/causalFS.pdf, 2007.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*, volume 3, pages 1–33, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008. URL http://jmlr.csail.mit.edu/papers/topic/causality.html.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Causality workbench. In P. McKay Illaria, F. Russo, and J. Williamson, editors, *Causality in the Sciences*. Oxford University Press, (to appear), 2010.

I. Guyon et al. Datasets of the causation and prediction challenge. Technical Report, 2008. URL http://clopinet.com/isabelle/Projects/WCCI2008/Datasets.pdf.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11, 1943.

Edward H. Herskovits and Azar P. Dagher. Application of Bayesian networks to health care, 1997.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696. MIT Press, 2008. URL http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. http://arxiv.org/abs/0804.3678, 2008.

D. Janzing, J. Peters, J. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *25th Conference on Uncertainty in Artificial Intelligence*, pages 1–9, Corvallis, OR, USA, 06 2009. AUAI Press. URL http://www.cs.mcgill.ca/~uai2009/.

M. Jenab, N. Slimani, M. Bictash, P. Ferrari, and S. Bingham. Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human Genetics*, June 2009. URL http://dx.doi.org/10.1007/s00439-009-0662-5.

B. Brumback J.M. Robins, M.A. Hernan. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:D480–D484, 2008.

D. Kaplan. *Structural Equation Modeling: Foundations and Extensions*, volume 10 of *Advanced Quantitative Techniques in the Social Sciences*. SAGE, 2000. ISBN 0-7619-1407-2.

Martjin B Katan. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Int. J. Epidemiol.*, 33(1):9–, 2004. doi: 10.1093/ije/dyh312. URL http://ije.oxfordjournals.org.

Bernard Keavney. Commentary: Katan's remarkable foresight: genes and causality 18 years on. *Int. J. Epidemiol.*, 33(1):11–14, 2004. doi: 10.1093/ije/dyh056. URL http://ije.oxfordjournals.org.

Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *ML92: Proceedings of the ninth international workshop on Machine learning*, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. ISBN 15586247X. URL http://portal.acm.org/citation.cfm?id=142034.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2009.

Pieter C. Kraaijeveld and Marek J. Druzdzel. Genierate: An interactive generator of diagnostic Bayesian network models. In *16th International Workshop on Principles of Diagnosis*, Monterey, California, USA, 2005. URL www.kbs.twi.tudelft.nl/Publications/MSc/2005-Kraaijeveld-MSc.html.

Peter Kraft and David J. Hunter. Genetic Risk Prediction – Are We There Yet? *N Engl J Med*, 360(17):1701–1703, 2009. doi: 10.1056/NEJMp0810107. URL http://content.nejm.org.

J. Lemeire. Learning causal models of multivariate systems. PhD thesis, Brussels, 2007.

Alex John London and Joseph B. Kadane. Placebos that harm: sham surgery controls in clinical trials. *Statistical Methods in Medical Research*, 11:413–427, October 2002.

Aurelie C. Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical Granger modeling methods for temporal causal modeling. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586, Paris, France, 2009.

D. Margaritis and S. Thrun. A Bayesian multiresolution independence test for continuous variables. In *17th Conference on Uncertainty in Artificial Intelligence (UAI)*, Seattle, Washington, August 2001.

Florian Markowetz. A bibliography on learning causal networks of gene interactions, March 2007. URL http://genomics.princeton.edu/~florian/docs/network-bib.pdf.

A. J. McMichael, O. M. Jensen, D. M. Parkin, and D. G. Zaridze. Dietary and endogenous cholesterol and human cancer. *Epidemiol Rev*, 6(1):192–216, 1984. URL http://epirev.oxfordjournals.org.

C. Meek. Strong completeness and faithfulness in Bayesian networks. Proceedings of *11th Uncertainty in Artificial Intelligence (UAI), Montreal, Canada,* Morgan Kaufmann, pages 411–418, 1995.

E. Miguel, S. Satyanath, and E. Sergenti. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753, 2004. doi: 10.1086/421174. URL http://www.journals.uchicago.edu/doi/abs/10.1086/421174.

M. L. Minsky and S. A. Papert. *Perceptrons*. MIT Press, Cambridge, MA, 1969.

Alessio Moneta. Causality in macroeconometrics: some considerations about reductionism and realism. *Journal of Economic Methodology*, 12(3):433–453, September 2005. URL http://ideas.repec.org/a/taf/jecmet/v12y2005i3p433-453.html.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

G. Neves, S. F. Cooke, and T. V. P. Bliss. Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. *Nature Reviews Neuroscience*, 9:65–75, 2008. URL http://www.nature.com/nrn/journal/v9/n1/abs/nrn2303.html.

Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Application of Bayesian belief networks to diagnosis of liver disorders. In *Proceedings of the Third Conference on Neural Networks and Their Applications*, pages 730–736, Kule, Poland, 14–18 October 1997.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, California, 1988.

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

Jean-Philippe Pellet and André Elisseeff. Using Markov blankets for causal structure learning. *JMLR*, 9:1295–1342, 2008. ISSN 1533-7928.

Diana Petitti. Commentary: Hormone replacement therapy and coronary heart disease: four lessons. *Int. J. Epidemiol.*, 33(3):461–463, 2004. doi: 10.1093/ije/dyh192. URL http://ije.oxfordjournals.org.

E. J. G. Pitman. Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, 4, 1937.

Xinzhou Qin and Wenke Lee. Statistical causality analysis of INFOSEC alert data. In *RAID*, pages 73–93, 2003.

H. Reichenbach. *The Direction of Time*. University of Los Angeles Press, Berkeley, 1956.

Donald Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

B. Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1913.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (Second Edition)*. Prentice Hall, 2003.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. doi: 10.2307/2958889. URL http://dx.doi.org/10.2307/2958889.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006. ISSN 1533-7928.

George Davey Smith. Capitalizing on Mendelian randomization to assess the effects of treatments. *J R Soc Med*, 100(9):432–435, 2007. doi: 10.1258/jrsm.100.9.432. URL http://jrsm.rsmjournals.com.

D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–247, August 1993.

P. Spirtes. Variable definition and causal inference. In *13th International Congress of Logic Methodology and Philosophy of Science*, 2008.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, London, England, 2000.

Hervé Stoppiglia, Gérard Dreyfus, Rémi Dubois, and Yacine Oussar. Ranking a random feature for variable and feature selection. *JMLR*, 3:1399–1414, 2003. URL http://www.jmlr.org/papers/v3/stoppiglia03a.html.

Ioannis Tsamardinos and Laura E. Brown. Bounding the false discovery rate in local Bayesian network learning. In *AAAI*, pages 1100–1105, 2008.

Eugene Tuv, Alexander Borisov, and Kari Torkkola. Feature selection using ensemble based ranking against artificial contrasts. In *IJCNN*, pages 2181–2186, 2006.

P. Vinei and D. Kriebel. Causal models in epidemiology: past inheritance and genetic future. *Environ Health*, 5(21), 2006.

Keith Wheatley and Richard Gray. Commentary: Mendelian randomization–an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *Int. J. Epidemiol.*, 33(1):15–17, 2004. doi: 10.1093/ije/dyg313. URL http://ije.oxfordjournals.org.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Zhu Yongli, H. Limin, and Lu Jinling. Bayesian networks-based approach for power systems fault diagnosis. *IEEE Transactions on Power Delivery*, 21(2):634–639, April 2006.

Gerald Young. Causality and causation in law, medicine, psychiatry, and psychology: Progression or regression? *Psychological Injury and Law*, 1(3):161–181, 2008.

Glossary

Action: An intervention performed by an external agent to disrupt the normal functioning of a system, which would otherwise be left to evolve according to its own dynamics.

Causal Bayesian Network: A model frequently used in causal discovery, using a directed acyclic graph (DAG) to model causal relationships between random variables. Using the network, it is possible to infer the probability distribution of some variable given measured values of others.

Causal Faithfulness Condition (CFC): The CFC is the faithfulness condition applied to a causal model (see "faithfulness"). The CFC essentially states that independence implies absence of a causal arrow. Complying with the CFC excludes modeling the XOR problem and cases in which multiple paths compensate each other.

Causal Markov Condition (CMC): The CMC is the Markov condition applied to a causal model (see "Markov property" or "Markov condition"). The CMC essentially states that statistical dependency implies the existence of a causal arrow. See also "principle of common cause". Systems with hidden confounders violate the CMC. See also "causal sufficiency".

Causal sufficiency: Causal sufficiency essentially states that there are no hidden variable that is a common cause of two variables considered, i.e., no hidden confounder. This commonly made assumption is very difficult to verify and the presence of a hidden confounder may invalidate completely a study. See "confounder".

Cause (as system state or event): Informally, a cause can be defined as a state C of a system of interest consistently followed by another state E (its effect) whenever the system is (actually or hypothetically) forced to assume the state C. The eventual existence of unobservable state variables makes it possible that correlated events succeeding each other are not in a causal relationship: both may be the consequence of an earlier common cause. For instance, lightning may trigger both thunder, followed by a fire alarm. "Thunder" and "fire alarm" are the consequence of the common cause "lightning", but are not causally related, even though "thunder" might happen consistently before "fire alarm". This ambiguity could be resolved if an external agent could perform an experiment and force "thunder" to happen with or without "lightning". See also manipulation or action.

Cause (as random variable): If a random variable is an indicator of presence/absence of an event, causal relationships between random variables are simple extensions of causal relationships between events. More generally, causal relationships between random variables can be defined via manipulations. For instance, given

two random variables C and E and a manipulation $do(C)$, a univariate causal relationship between C (cause) and E (effect) is found if $P(E|do(C)) \neq P(E)$. For instance, in a randomized clinical trial, C can be the amount of medicine taken and E the health status of the patient. If the health status of patients having taken the medicine differs from that of patients in the control group, a causal effect is detected.

Conditional independence (CI): Two random variables X and Y are conditionally independent of a third one *iff* $P(X,Y|Z) = P(X|Z)P(Y|Z)$. This may be extended to subsets of variables. Regular statistical independence is equivalent to conditioning on the empty set.

Confounded variable: An alleged cause of a target variable whose dependency with the target can be explained by the presence of a confounder (see "confounder").

Confounding factor or confounder: A variable that shows statistical dependencies to a target variable and its alleged cause and that may be a common cause to both, hence potentially making us confuse statistical dependence and causation.

Consequence, effect: The effect can be defined as the manifestation of the cause, see cause.

Counterfactual: An event contrary to the fact. Causality and counterfactuals are intimately tied together. Some authors argue that all causal statements can be phrased in terms of counterfactuals: "the throw of the stone caused the window to break" may be replaced by "had the stone not been thrown, the window would not have broken". Causal models allow us to predict what would have happened under a situation that did not occur (*e.g.,* "would the patient have died had he not taken the treatment").

Cross-validation (CV): A method frequently used in machine learning to select models, *e.g.,* with different architectures or hyper-parameters. One selects the model with the best CV performance, obtained by splitting repeatedly the available (observational) training data into training and validation set and averaging the prediction results on the validation sets. If observational data are used, CV is not a good method for selecting among alternative causal architectures.

Do-calculus: A method for calculating conditional probabilities of certain variables in a post-manipulation distribution given only conditional probabilities from the pre-manipulation distribution and some causal assumptions. The method was originally developed by Judea Pearl.

D-separation (D-connection): A set $\boldsymbol{C}$ is said to d-separate A from B if $\boldsymbol{C}$ blocks every path between A and B. If A and B are not d-separated, then they are d-connected. A path Π between two variables A and B is blocked by a set of nodes $\boldsymbol{C}$ if (1) Π contains a chain $I \rightarrow C \rightarrow J$ or a fork $I \leftarrow C \rightarrow J$ such that C is

in $\boldsymbol{C}$, or (2) Π does not contain a collider $I \to C \leftarrow J$ such that C or any of its descendants are in $\boldsymbol{C}$. D-separation is an algorithm to compute all the conditional independence relations entailed by a Bayesian network or a SEM.

Endogenous variable: A variable having explicit causes within a particular causal model. The characterization depends on the set of variables under consideration and the chosen causal model. Complementary concept: exogenous variable.

Exogenous variable: A variable having no explicit causes within a particular causal model. The characterization depends on the set of variables under consideration and the chosen causal model. Complementary concept: endogenous variable.

Experiment: Planned manipulations designed to determine causal relationships (see also "randomized controlled trial").

Experimental data: Data collected as a result of an experiment (see also "observational data").

Faithfulness: In the Bayesian network framework, a graph is faithful to a distribution if all the conditional independencies entailed by the distribution are reflected by Markov properties that can be read from the graph (see "Markov property"). A distribution is faithful if there exists a faithful graph representing it.

Features: Variables potentially predictive of the target variable, also called *covariates*, *explanatory variables* , or *predictor variables* in statistics.

Ground truth: In the pattern recognition jargon, "ground truth" refers to verified information obtained by scouting the terrain *on the ground* as opposed to information collected from far away observations, like satellite images.

I.i.d: Independent and identically distributed. A common assumption about the data distribution in machine learning, which assumes a stationary data generating process. This assumption is violated when external agents perform manipulations on the system.

Inference: There are two types of inference: model inference and variable inference (see the corresponding definitions).

Instrumental variable: A variable I used to test an alleged causal relationship $C \to E$ by performing a "natural manipulation" of C. It must be known that I is exogenous and cannot influence E in any other way than though C.

Latent variable: An unobserved (hidden) variable, possibly unknown.

Manipulation: A set of actions performed by an external agent on a system under study to disrupt the normal functioning of a system. A manipulation of a random variable C denoted as $do(C)$ consists in making C assume values according to a distribution decided by the agent, distinct from the "natural" distribution of C conditioned on the other variables of the system.

Markov blanket and Markov boundary (MB): A Markov blanket of a target variable (called MB) is a sufficient set of variables such that all other variables are independent of the target, given MB. A minimal Markov blanket is called a Markov boundary. Under some conditions, the Markov boundary is unique. Under the faithfulness assumption (see "faithfulness") it coincides with the set of parents, children, and spouses of the target. Many people include the minimality restriction in the definition of Markov blankets, therefore identifying the Markov blanket and the Markov boundary.

Markov property and Markov condition: A stochastic process of random variables has the Markov property if its future states are independent of far away past states given the present and a finite number of near past states (*i.e.,* it is memoryless). All Markov processes have an equivalent first order Markov process in which future states are independent of past states given the present state. By extension, atemporal Bayesian networks and SEMs are (first order) Markov models in the sense that each node is independent of its non-descendants given its parents. This is also called the "Markov condition". For these models, a "Markov property" is a conditional independence property between a subset of variables. "Markov properties" read from the graph (see "d-separation") are all valid conditional independence properties.

Model inference, model fitting, training: In a learning problem, inference refers to choosing the model, its structure, hyper-parameters and parameters.

Model over-fitting: Training a model to make excellent predictions for training examples, but obtaining poor prediction performance on test examples.

Natural distribution: Synonym of "observational distribution" or "pre-manipulation" distribution.

Non-interventional observations: See "observational data".

Observational data: Data collected from the observation of a system let to evolve according to its own dynamics, without controlled intervention (see also "experimental data").

Observational distribution: The joint distribution of the variables of a system in the absence or any external perturbation. Also called "pre-manipulation distribution".

Pre-manipulation distribution: Same as "observational distribution".

Principle of Common Cause (PCC): The PCC states that if two variables are correlated but neither is the cause of the other, then there should be at least one common cause influencing both variables.

Post-manipulation distribution: The joint distribution of the variables of a system after an action was performed by an external agent.

Predictive model, predictor: A mathematical construct $y = f(x;a)$ parameterized by a parameter vector a, allowing to make predictions of an outcome y given an input datum x.

Randomized Controlled Trial (RCT): Planned experiments involving a random allocation of different interventions (treatments or conditions) to subjects. As long as the numbers of subjects are sufficient, this ensures that both known and unknown confounding factors are evenly distributed between treatment groups. There are many variants of RCTs including various blinding and randomizing techniques (see also "experiment").

Structural Equation Model (SEM): A model to represent causal relationships as a directed acyclic graph (DAG), similar to a Bayesian network, but in which variables are interconnected by functional relationships (eventually altered by stochastic noise) rather than conditional distributions. Noise variables are called "exogenous"; other (dependent) variables are called "endogenous". (See "exogenous variables" and "endogenous variables").

Target variable (or target): The outcome under study.

Variable inference: A trained model (e.g. a Bayesian network) can then be used to infer variable probability distributions from the partial knowledge of other variables.

JMLR: Workshop and Conference Proceedings 6:39–58, 2010 NIPS 2008 workshop on causality

Causal Inference

Judea Pearl JUDEA@CS.UCLA.EDU
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA

Editors: Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf

Abstract

This paper reviews a theory of causal inference based on the Structural Causal Model (SCM) described in (Pearl, 2000a). The theory unifies the graphical, potential-outcome (Neyman-Rubin), decision analytical, and structural equation approaches to causation, and provides both a mathematical foundation and a friendly calculus for the analysis of causes and counterfactuals. In particular, the paper establishes a methodology for inferring (from a combination of data and assumptions) the answers to three types of causal queries: (1) queries about the effect of potential interventions, (2) queries about counterfactuals, and (3) queries about the direct (or indirect) effect of one event on another.

Keywords: Structural equation models, confounding, graphical methods, counterfactuals, causal effects, potential-outcome.

1. Introduction

The research questions that motivate most quantitative studies in the health, social and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone.

Remarkably, although much of the conceptual framework and algorithmic tools needed for tackling such problems are now well established, they are hardly known to researchers in the field who could put them into practical use. Why?

Solving causal problems mathematically requires certain extensions in the standard mathematical language of statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, large segments of the research community find it hard to appreciate and benefit from the many results that causal analysis has produced in the past two decades. These results rest on advances in three areas:

1. Nonparametric structural equations

2. Graphical models

3. Symbiosis between counterfactual and graphical methods.

This paper aims at making these advances more accessible to the general research community by, first, contrasting causal analysis with standard statistical analysis, second, comparing and unifying existing approaches to causal analysis, and finally, providing a friendly formalism for counterfactual analysis, within which most (if not all) causal questions can be formulated, analyzed and resolved.

We will see that, although full description of the data generating process cannot be inferred from data alone, many useful features of the process can be estimated from a combination of (1) data, (2) prior qualitative knowledge, and/or (3) experiments. Thus, the challenge of causal inference is to answer causal queries of practical interest with minimum number of assumptions and with minimal experimentation. Following an introductory section which defines the demarcation line between associational and causal analysis, the rest of the paper will deal with the estimation of three types of causal queries: (1) queries about the effect of potential interventions, (2) queries about counterfactuals (e.g., whether event x would occur had event y been different), and (3) queries about the direct and indirect effects.

2. From Associational to Causal Analysis: Distinctions and Barriers

2.1. The Basic Distinction: Coping With Change

The aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or external interventions.

This distinction implies that causal and associational concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions which identify relationships that remain invariant when external conditions change.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: one cannot substantiate causal claims from

associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.[1]

2.2. Formulating the Basic Distinction

A useful demarcation line that makes the distinction between associational and causal concepts crisp and easy to apply, can be formulated as follows. An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odd ratio, marginalization, conditionalization, "controlling for," and so on. Examples of causal concepts are: randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution, and so on. The former can, while the latter cannot be defined in term of distribution functions.

This demarcation line is extremely useful in causal analysis for it helps investigators to trace the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must rely on some premises that invoke such concepts; it cannot be inferred from, or even defined in terms statistical associations alone.

2.3. Ramifications of the Basic Distinction

This principle has far reaching consequences that are not generally recognized in the standard statistical literature. Many researchers, for example, are still convinced that confounding is solidly founded in standard, frequentist statistics, and that it can be given an associational definition saying (roughly): "U is a potential confounder for examining the effect of treatment X on outcome Y when both U and X and U and Y are not independent." That this definition and all its many variants must fail (Pearl 2000a, Section 6.2)[2] is obvious from the demarcation line above; if confounding were definable in terms of statistical associations, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders and obtain unbiased estimates of causal effects. This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies. Hence the definition must be false. Therefore, to the bitter disappointment of generations of epidemiologist and social science researchers, confounding bias cannot be detected or corrected by statistical methods alone; one must make some judgmental assumptions regarding causal relationships in the problem before an adjustment (e.g., by stratification) can safely correct for confounding bias.

1. The methodology of "causal discovery" (Spirtes, et al. 2000; Pearl 2000a, chapter 2) is likewise based on the causal assumption of "faithfullness" or "stability."
2. Any intermediate variable U on a causal path from X to Y satisfies this definition, without confounding the effect of X on Y.

Another ramification of the sharp distinction between associational and causal concepts is that any mathematical approach to causal analysis must acquire new notation for expressing causal relations – probability calculus is insufficient. To illustrate, the syntax of probability calculus does not permit us to express the simple fact that "symptoms do not cause diseases", let alone draw mathematical conclusions from such facts. All we can say is that two events are dependent—meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease|symptom)$ from causal dependence, for which we have no expression in standard probability calculus. Scientists seeking to express causal relationships must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease."

2.4. Two Mental Barriers: Untested Assumptions and New Notation

The preceding two requirements: (1) to commence causal analysis with untested,[3] theoretically or judgmentally based assumptions, and (2) to extend the syntax of probability calculus, constitute the two main obstacles to the acceptance of causal analysis among statisticians and among professionals with traditional training in statistics.

Associational assumptions, even untested, are testable in principle, given sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to prior causal assumptions, say that treatment does not change gender, remains substantial regardless of sample size.

This makes it doubly important that the notation we use for expressing causal assumptions be meaningful and unambiguous so that one can clearly judge the plausibility or inevitability of the assumptions articulated. Statisticians can no longer ignore the mental representation in which scientists store experiential knowledge, since it is this representation, and the language used to access it that determine the reliability of the judgments upon which the analysis so crucially depends.

How does one recognize causal expressions in the statistical literature? Those versed in the potential-outcome notation (Neyman, 1923; Rubin, 1974; Holland, 1988), can recognize such expressions through the subscripts that are attached to counterfactual events and variables, e.g. $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions, e.g. $Y(0)$, $Y(1)$, $Y(x,u)$ or $Z(x,y)$.) The expression $Y_x(u)$, for example, stands for the value that outcome Y would take in individual u, had treatment X been at level x. If u is chosen at random, Y_x is a random variable, and one can talk about the probability that Y_x would attain a value y in the population, written $P(Y_x = y)$. Alternatively, Pearl (1995)

3. By "untested" I mean untested using frequency data in nonexperimental studies.

used expressions of the form $P(Y = y|set(X = x))$ or $P(Y = y|do(X = x))$ to denote the probability (or frequency) that event $(Y = y)$ would occur if treatment condition $X = x$ were enforced uniformly over the population.[4] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality.[5]

However, few have taken seriously the textbook requirement that any introduction of new notation must entail a systematic definition of the syntax and semantics that governs the notation. Moreover, in the bulk of the statistical literature before 2000, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations, and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate not be affected by a treatment, a necessary assumption for the control of confounding (Cox, 1958, p. 48), is expressed in plain English, not in a mathematical expression.

Remarkably, though the necessity of explicit causal notation is now recognized by most leaders in the field, the use of such notation has remained enigmatic to most rank and file researchers, and its potentials still lay grossly underutilized in the statistics based sciences. The reason for this, can be traced to the unfriendly and ad-hoc way in which causal analysis has been presented to the research community, resting primarily on the restricted paradigm of controlled randomized trials advanced by Rubin (1974).

The next section provides a conceptualization that overcomes these mental barriers; it offers both a friendly mathematical machinery for cause-effect analysis and a formal foundation for counterfactual analysis.

3. Structural Causal Models (SCM) and The Language of Diagrams

3.1. Semantics: Causal Effects and Counterfactuals

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920's by the geneticist Sewall Wright (1921), who used a combination of equations and graphs. For example, if X stands for a disease variable and Y stands for a certain symptom of the disease, Wright would write a linear equation:

$$y = \beta x + u \tag{1}$$

where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u stands for all factors, other than the disease in question, that could possibly affect Y. In interpreting this equation one should think of a physical process whereby Nature *examines* the values of x and u and, accordingly, *assigns* variable

4. Clearly, $P(Y = y|do(X = x))$ is equivalent to $P(Y_x = y)$, This is what we normally assess in a controlled experiment, with X randomized, in which the distribution of Y is estimated for each level x of X.
5. These notational clues should be useful for detecting inadequate definitions of causal concepts; any definition of confounding, randomization or instrumental variables that is cast in standard probability expressions, void of graphs, counterfactual subscripts or $do(*)$ operators, can safely be discarded as inadequate.

Y the value $y = \beta x + u$. Similarly, to "explain" the occurrence of disease X, one could write $x = v$, where V stand for all factors affecting X.

To express the directionality inherent in this process, Wright augmented the equation with a diagram, later called "path diagram," in which arrows are drawn from (perceived) causes to their (perceived) effects and, more importantly, the absence of an arrow makes the empirical claim that the value Nature assigns to one variable is not determined by the value taken by another. In Figure 1, for example, the absence of arrow from Y to X represent the claim that symptom Y is not among the factors V which affect disease X.

The variables V and U are called "exogenous" ; they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model.

If correlation is judged possible between two exogenous variables, U and V, it is customary to connect them by a dashed double arrow, as shown in Figure 1(b).

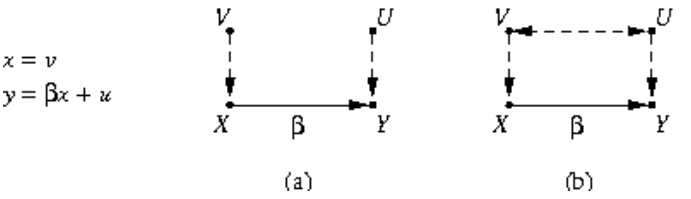


Figure 1: A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

To summarize, path diagrams encode causal assumptions via missing arrows, representing claims of zero influence, and missing double arrows (e.g., between V and U), representing the (causal) assumption $Cov(U,V)$=0.

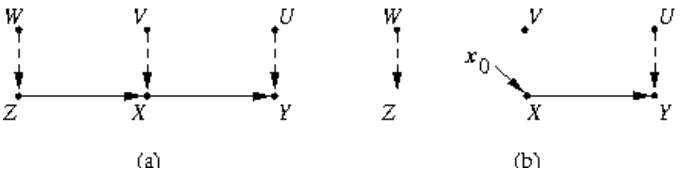


Figure 2: (a) The diagram associated with the structural model of equation (2). (b) The diagram associated with the modified model, M_{x_0}, of equation (3), representing the intervention $do(X = x_0)$.

The generalization to nonlinear systems of equations is straightforward. For example, the non-parametric interpretation of the diagram of Figure 2(a) corresponds to a set

of three functions, each corresponding to one of the observed variables:

$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z,v) \\ y &= f_Y(x,u) \end{aligned} \tag{2}$$

where W, V and U are assumed to be jointly independent but, otherwise, arbitrarily distributed.

Remarkably, unknown to most economists and pre-2000 philosophers,[6] structural equation models provide a formal interpretation and symbolic machinery for analyzing counterfactual relationships of the type: "Y would be y had X been x in situation $\boldsymbol{U}=\boldsymbol{u}$," denoted $Y_x(u) = y$. Here $\boldsymbol{U}$ represents the vector of all exogenous variables.[7]

The key idea is to interpret the phrase "had X been x_0" as an instruction to modify the original model and replace the equation for X by a constant x_0, yielding the sub-model.

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x,u) \end{aligned} \tag{3}$$

the graphical description of which is shown in Figure 2(b).

This replacement permits the constant x_0 to differ from the actual value of X (namely $f_X(z,v)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multi-stage models, where the dependent variable in one equation may be an independent variable in another (Balke and Pearl, 1994ab; Pearl, 2000b). For example, to compute $E(Y_{x_0})$, the expected effect of *setting* X to x_0, (also called the average causal effect of X on Y, denoted $E(Y|do(x_0))$ or, generically, $E(Y|do(x))$), we solve equation (3) for Y in terms of the exogenous variables, yielding $Y_{x_0} = f_Y(x_0,u)$, and average over U and V. It is easy to show that in this simple system, the answer can be obtained without knowing the form of the function $f_Y(x,u)$ or the distribution $P(u)$. The answer is given by:

$$E(Y_{x_0}) = E(Y|do(X = x_0) = E(Y|x_0)$$

which is estimable from the observed distribution $P(x,y,z)$. This result hinges on the assumption that W, V, and U are mutually independent and on the topology of the graph (e.g., that there is no direct arrow from Z to Y.)

6. Connections between structural equations and a restricted class of counterfactuals were recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995) who used modified models to permit counterfactual conditioning on dependent variables.
7. Because $U = u$ may contain detailed information about a situation or an individual, $Y_x(u)$ is related to what philosophers called "token causation," while $P(Y_x = y|Z = z)$ characterizes "Type causation," that is, the tendency of X to influence Y in a sub-population characterized by $Z = z$.

In general, it can be shown (Pearl 2000a, Chapter 3) that, whenever the graph is Markovian (i.e., acyclic with independent exogenous variables) the post-interventional distribution $P(Y = y|do(X = x))$ is given by the following expression:

$$P(Y = y|do(X = x)) = \sum_t P(y|t,x)P(t) \tag{4}$$

where T is the set of direct causes of X (also called "parents") in the graph. Again, we see that all factors on the right hand side are estimable from the distribution P of observed variables and, hence, the counterfactual probability $P(Y_x = y)$ is estimable with mere partial knowledge of the generating process – the topology of the graph and independence of the exogenous variables is all that is needed.

When some variables in the graph (e.g., the parents of X) are unobserved, we may not be able to learn (or "identify" as it is called) the post-intervention distribution $P(y|do(x))$ by simple conditioning, and more sophisticated methods would be required. Likewise, when the query of interest involves several hypothetical worlds simultaneously, e.g., $P(Y_x = y, Y_{x'} = y')$[8], the Markovian assumption may not suffice for identification and additional assumptions, touching on the form of the data-generating functions (e.g., monotonicity) may need to be invoked. These issues will be discussed in Sections 3.2 and 5.

This interpretation of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation models, used in economics and social science and the Neyman-Rubin potential-outcome framework to be discussed in Section 4. But first we discuss two long-standing problems that have been completely resolved in purely graphical terms, without delving into algebraic techniques.

3.2. Confounding and Causal Effect Estimation

The central target of most studies in the social and health sciences is the elucidation of cause-effect relationships among variables of interests, for example, treatments, policies, preconditions and outcomes. While good statisticians have always known that the elucidation of causal relationships from observational studies must be shaped by assumptions about how the data were generated, the relative roles of assumptions and data, and ways of using those assumptions to eliminate confounding bias have been a subject of much controversy. The structural framework of Section 3.1 puts these controversies to rest.

COVARIATE SELECTION: THE BACK-DOOR CRITERION

Consider an observational study where we wish to find the effect of X on Y, for example, treatment on response, and assume that the factors deemed relevant to the problem are structured as in Figure 3; some are affecting the response, some are affecting the treatment and some are affecting both treatment and response. Some of these factors

8. Read: The probability that Y would be y if X were x and y' if X were x'.

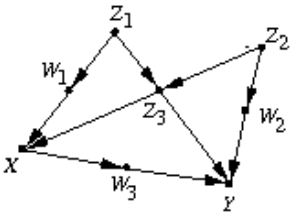


Figure 3: Graphical model illustrating the back-door criterion. Error terms are not shown explicitly.

may be unmeasurable, such as genetic trait or life style, others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set" or a set "appropriate for adjustment". The problem of defining a sufficient set, let alone finding one, has baffled epidemiologists and social science for decades (see Greenland et al., 1999; Pearl, 1998, 2003 for review).

The following criterion, named "back-door" in Pearl (1993a), settles this this problem by providing a graphical method of selecting a sufficient set of factors for adjustment. It states that a set S is appropriate for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.[9]

Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained above, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The implication of finding a sufficient set S is that, stratifying on S is guaranteed to remove all confounding bias relative the causal effect of X on Y. In other words, it renders the causal effect of X on Y estimable, via

$$\begin{aligned} & P(Y = y|do(X = x)) \\ & \quad = \sum_s P(Y = y|X = x, S = s)P(S = s) \end{aligned} \tag{5}$$

Since all factors on the right hand side of the equation are estimable (e.g., by regression) from the pre-interventional data, the causal effect can likewise be estimated from such data without bias.

9. A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. See (Pearl, 2000a, pp. 16-7). If S blocks *all* paths from X to Y it is said to "d-separate X and Y."

The back-door criterion allows us to write equation (5) directly, after selecting a sufficient set S from the diagram, without resorting to any algebraic manipulation. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether "X is conditionally ignorable given S," a formidable mental task required in the potential-outcome framework (Rosenbaum and Rubin, 1983). The criterion also enables the analyst to search for an optimal set of covariate—namely, a set S that minimizes measurement cost or sampling variability (Tian et al., 1998).

GENERAL CONTROL OF CONFOUNDING

Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. A much more general identification criterion is provided by the following theorem:

Theorem 1 (*Tian and Pearl, 2002*)
A sufficient condition for identifying the causal effect $P(y|do(x))$ *is that every path between* X *and any of its children traces at least one arrow emanating from a measured variable.*[10]

For example, if W_3 is the only observed covariate in the model of Figure 3, then there exists no sufficient set for adjustment (because no set of observed covariates can block the paths from X to Y through Z_3), yet $P(y|do(x))$ can nevertheless be estimated since every path from X to W_3 (the only child of X) traces either the arrow $X \rightarrow W_3$, or the arrow $W_3 \rightarrow Y$, both emanating from a measured variable (W_3). In this example, the variable W_3 acts as a "mediating instrumental variable" (Pearl, 1993b; Chalak and White, 2006) and yields the estimand:

$$\begin{aligned} P(Y = y|do(X = x)) &= \sum_{w_3} P(W_3 = w_3|do(X = x))P(Y = y|do(W_3 = w_3)) \\ &= \sum_{w_3} P(w_3|x) \sum_{x'} P(y|w_3, x')P(x') \end{aligned} \tag{6}$$

More recent results extend this theorem by (1) presenting a necessary and sufficient condition for identification (Shpitser and Pearl, 2006), and (2) extending the condition from causal effects to any counterfactual expression (Shpitser and Pearl, 2007). The corresponding unbiased estimands for these causal quantities are readable directly from the diagram.

The mathematical derivation of causal effect estimands, like equations (5) and (6) is merely a first step toward computing quantitative estimates of those effects from finite samples, using the rich traditions of statistical estimation and machine learning. Although the estimands derived in (5) and (6) are non-parametric, this does not mean that

10. Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.

one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (6) can be converted to the product $E(Y|do(x)) = r_{W_3X} r_{YW_3 \cdot X} x$, where $r_{YZ \cdot X}$ is the (standardized) coefficient of Z in the regression of Y on Z and X. More sophisticated estimation techniques can be found in Rosenbaum and Rubin (1983), and Robins (1999). For example, the "propensity score" method of Rosenbaum and Rubin (1983) was found to be quite useful when the dimensionality of the adjusted covariates is high and the data is sparse (See Pearl 2000a, 2nd edition, 2009a, pp. 348–52).

It should be emphasized, however, that contrary to conventional wisdom (e.g., Rubin (2009)), propensity score methods are merely efficient estimators of the right hand side of (5); they cannot be expected to reduce bias in case the set S does not satisfy the back-door criterion (Pearl 2009abc).

3.3. Counterfactual Analysis in Structural Models

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, in much the same way that not all causal questions can be answered from experimental studies. For example, questions of attribution (e.g., I took an aspirin and my headache is gone, was it *due* to the aspirin?) or of susceptibility (e.g., I am a healthy non-smoker, would I be as healthy had I been a smoker?) cannot be answered from experimental studies, and naturally, this kind of questions cannot be expressed in $P(y|do(x))$ notation.[11] To answer such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "Y would be y had X been x in situation $\boldsymbol{U}$=$\boldsymbol{u}$," denoted $Y_x(\boldsymbol{u}) = y$.

As noted in Section 3.1, the structural definition of counterfactuals involves modified models, like M_{x_0} of equation (3), formed by the intervention $do(X = x_0)$ (Figure 2(b)). Call the solution of Y in model M_x the *potential response* of Y to x, and denote it by the symbol $Y_x(\boldsymbol{u})$. In general, then, the formal definition of the counterfactual $Y_x(\boldsymbol{u})$ in SCM is given by (Pearl 2000a, p. 98):

$$Y_x(\boldsymbol{u}) = Y_{M_x}(\boldsymbol{u}).$$

The quantity $Y_x(\boldsymbol{u})$ can be given experimental interpretation; it stands for the way an individual with characteristics $(\boldsymbol{u})$ would respond, had the treatment been x, rather than the treatment $x = f_X(\boldsymbol{u})$ actually received by that individual. In our example, since Y does not depend on v and w, we can write:

$$Y_{x_0}(u, v, w) = Y_{x_0}(u) = f_Y(x_0, u).$$

11. The reason for this fundamental limitation is that no death case can be tested twice, with and without treatment. For example, if we measure equal proportions of deaths in the treatment and control groups, we cannot tell how many death cases are actually attributable to the treatment itself; it is quite possible that many of those who died under treatment would be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

Clearly, the distribution $P(u,v,w)$ induces a well defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as '$Y_{x_0} = y$ AND $Y_{x_1} = y'$,' which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming linear equations (as in Figure 1),

$$x = v, \qquad y = \beta x + u,$$

the conditions $Y = y_0$ and $X = x_0$ yield $v = x_0$ and $u = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{x_1} must take on the value: $Y_{x_1} = \beta x_1 + u = \beta(x_1 - x_0) + y_0$. In other words, if X were x_1 instead of x_0, Y would increase by β times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of U and, for that reason, attributional queries are generally not identifiable in nonparametric models (Pearl, 2000a, Chapter 9).

In general, if x and x' are incompatible then Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$."[12] Such concerns have been a source of objections to treating counterfactuals as jointly distributed random variables (Dawid, 2000). The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels neutralizes these objections (Pearl, 2000b), since the contradictory joint statement is mapped into an ordinary event, one where the background variables satisfy both statements simultaneously, each in its own distinct submodel; such events have well defined probabilities.

The structural interpretation of counterfactuals also provides the conceptual and formal basis for the Neyman-Rubin potential-outcome framework, an approach to causation that takes a controlled randomized trial (CRT) as its starting paradigm, assuming that nothing is known to the experimenter about the science behind the data. This "black-box" approach, which has thus far been denied the benefits of graphical or structural analyses, was developed by statisticians who found it difficult to cross the two mental barriers discussed in Section 2.4. The next section establishes the precise relationship between the structural and potential-outcome paradigms, and outlines how the latter can benefit from the richer representational power of the former.

4. The Language of Potential Outcomes and Counterfactuals

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that outcome Y would obtain in experimental unit u, had treatment X been x" (Neyman, 1923; Rubin, 1974). Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In Section 3.3 we saw that this counterfactual entity has the natural interpretation as representing the solution for Y in a modified system of equations, where *unit* is interpreted

12. For example, "The probability is 80% that Joe belongs to the class of patients who will be cured if they take the drug and will die otherwise."

a vector $\boldsymbol{u}$ of background factors that characterize an experimental unit. Each structural equation model thus carries a collection of assumptions about the behavior of hypothetical units, and these assumptions permit us to derive the counterfactual quantities of interest. In the potential-outcome framework, however, no equations are available for guidance and $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined; not a quantity that can be derived from some model. In this sense the structural interpretation of $Y_x(u)$ provides the formal basis for the potential-outcome approach; the formation of the submodel M_x explicates mathematically how the hypothetical condition "had X been x" could be realized, and what the logical consequence are of such a condition.

4.1. The "Black-Box" or "Missing-data" Paradigm

The distinct characteristic of the potential-outcome approach is that, although investigators must think and communicate in terms of undefined, hypothetical quantities such as $Y_x(u)$, the analysis itself is conducted almost entirely within the axiomatic framework of probability theory. This is accomplished, by postulating a "super" probability function on both hypothetical and real events. If U is treated as a random variable then the value of the counterfactual $Y_x(u)$ becomes a random variable as well, denoted as Y_x. The potential-outcome analysis proceeds by treating the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects (written $P(y|do(x))$ in the structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y_x = y)$. The new hypothetical entities Y_x are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence.

Naturally, these hypothetical entities are not entirely whimsy. They are assumed to be connected to observed variables via consistency constraints (Robins, 1986) such as

$$X = x \implies Y_x = Y, \tag{7}$$

which states that, for every u, if the actual value of X turns out to be x, then the value that Y would take on if 'X *were* x' is equal to the actual value of Y. For example, a person who chose treatment x and recovered, would also have recovered if given treatment x by design. Whether additional constraints should tie the observables to the unobservables is not a question that can be answered in the potential-outcome framework, which lacks an underlying model.

The main conceptual difference between the two approaches is that, whereas the structural approach views the intervention $do(x)$ as an operation that changes the distribution but keeps the variables the same, the potential-outcome approach views the variable Y under $do(x)$ to be a different variable, Y_x, loosely connected to Y through relations such as (7), but remaining unobserved whenever $X \neq x$. The problem of inferring probabilistic properties of Y_x, then becomes one of "missing-data" for which estimation techniques have been developed in the statistical literature.

Pearl (2000a, Chapter 7) shows, using the structural interpretation of $Y_x(u)$, that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (7) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{8}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{9}$$

Equation (8) ensures that the interventions $do(Y = y)$ results in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that may be applied to variables other than Y. Equation (9) generalizes (7) to cases where Z is held fixed, at z.

4.2. Problem Formulation and the Demystification of "Ignorability"

The main drawback of this black-box approach surfaces in problem formulation, namely, the phase where a researcher begins to articulate the "science" or "causal assumptions" behind the problem at hand. Such knowledge, as we have seen in Section 1, must be articulated at the onset of every problem in causal analysis – causal conclusions are only as valid as the causal assumptions upon which they rest.

To communicate scientific knowledge, the potential-outcome analyst must express assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For instance, in our example of Figure 2(a), to communicate the understanding that the (Z) is randomized (hence independent of V and U), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$.[13] To further formulate the understanding that Z does not affect Y directly, except through X, the analyst would write a, so called, "exclusion restriction": $Y_{xz} = Y_x$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest; in other cases, only bounds on the solution can be obtained. For example, if one can plausibly assume that a set Z of covariates satisfies the conditional independence

$$Y_x \perp\!\!\!\perp X | Z \tag{10}$$

(an assumption that was termed "conditional ignorability" by Rosenbaum and Rubin, 1983, then the causal effect $P^*(Y_x = y)$ can readily be evaluated to yield

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x,z)P(z) \quad \text{(using (10))} \\ &= \sum_z P^*(Y = y|x,z)P(z) \quad \text{(using (7))} \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{11}$$

13. The notation $Y \perp\!\!\!\perp X|Z$ stands for the conditional independence relationship $P(Y = y, X = x|Z = z) = P(Y = y|Z = z)P(X = x|Z = z)$ (Dawid, 1979).

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the standard covariate-adjustment formula of equation (5).

We see that the assumption of conditional ignorability (10) qualifies Z as a sufficient covariate for adjustment; it is entailed indeed by the "back-door" criterion of Section 3.2, which qualifies such covariates by tracing paths in the causal diagram.

The derivation above may explain why the potential-outcome approach appeals to mathematical statisticians; instead of constructing new vocabulary (e.g., arrows), new operators ($do(x)$) and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of probability calculus. Save for an occasional application of rule (9) or (7)), the analyst may forget that Y_x stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

However, this mathematical orthodoxy exacts a very high cost: all background knowledge pertaining to a given problem must first be translated into the language of counterfactuals (e.g., ignorability conditions) before analysis can commence. This translation may in fact be the hardest part of the problem. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (10), the key to the derivation of (11), holds in any familiar situation, say in the experimental setup of Figure 2(a). This assumption reads: "the value that Y would obtain had X been x, is independent of X, given Z". Even the most experienced potential-outcome expert would be unable to discern whether any subset Z of covariates in Figure 3 would satisfy this conditional independence condition.[14] Likewise, to derive equation (6) in the language of potential-outcome (see Pearl 2000a, page 233), one would need to convey the structure of the chain $X \rightarrow W_3 \rightarrow Y$ using the cryptic expression: $W_{3_x} \perp\!\!\!\perp \{Y_{w_3}, X\}$, read: "the value that W_3 would obtain had X been x is independent of the value that Y would obtain had W_3 been w_3 jointly with the value of X". Such assumptions are cast in a language so far removed from ordinary understanding of scientific theories that, for all practical purposes, they cannot be comprehended or ascertained by ordinary mortals. As a result, researchers in the graph-less potential-outcome camp rarely use "conditional ignorability" (10) to guide the choice of covariates; they view this condition as a hoped-for miracle of nature rather than a target to be achieved by reasoned design.[15]

Replacing "ignorability" with a simple condition (i.e., back-door) in a graphical model permits researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are

14. Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z|Y$ holds in Figure 2(a).

15. The opaqueness of counterfactual independencies explains why many researchers within the potential-outcome camp are unaware of the fact that adding a covariate to the analysis (e.g., Z_3 in Figure 3) may actually *increase* confounding bias. Paul Rosenbaum, for example, writes: "there is no reason to avoid adjustment for a variable describing subjects before treatment" Rosenbaum (2002), p. 76. Don Rubin (2009) goes as far as stating that refraining from conditioning on an available measurement is "nonscientific ad hockery" for it goes against the tenets of Bayesian philosophy (see Pearl 2009bc for a discussion of this fallacy).

selected, and what experiments we can do to test, at least partially, if we have the knowledge needed for covariate selection.

Aside from offering no guidance in covariate selection, formulating a problem in the potential-outcome language encounters three additional hurdles. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are *redundant*, or whether those judgments are *self-consistent.* The need to express, defend, and manage formidable counterfactual relationships of this type explain the slow acceptance of causal analysis among health scientists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential-outcome alternatives advocated in Angrist et al. (1996); Holland (1988); Sobel (1998).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist et al., 1996), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian and Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2000a). Pearl (2000a, p. 232) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into counterfactual notation, performing the mathematics in the algebraic language of counterfactuals (using (7), (8), and (9)) and, finally, interpreting the result in plain causal language. The next section illustrates such symbiosis.

5. Mediation: Direct and Indirect Effects

5.1. Direct versus Total Effects:

The causal effect we have analyzed so far, $P(y|do(x))$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

Another example concerns the identification of neural pathways in the brain or the structural features of protein-signaling networks in molecular biology (Brent and Lok, 2005). Here, the decomposition of effects into their direct and indirect components

carries theoretical scientific importance, for it predicts behavior under a rich variety of hypothetical interventions.

In all such examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between gender (X) and hiring (Y) for a given level of qualification Z, because, by conditioning on the mediator Z, we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \rightarrow Z \leftarrow U \rightarrow Y$, where X has no direct effect on Y. Physically holding Z constant would permit no association between X and Y, as can be seen by deleting all arrows entering Z. But if we were to condition on Z, a spurious association would be created through U (unobserved) that might be construed as a direct effect of X on Y.

Using the $do(x)$ notation, and focusing on expectations, this leads to a simple definition of *controlled direct effect*:

$$CDE \stackrel{\Delta}{=} E(Y|do(x),do(z)) - E(Y|do(x'),do(z))$$

or, equivalently, using counterfactual notation:

$$CDE \stackrel{\Delta}{=} E(Y_{xz}) - E(Y_{x'z})$$

where Z is any set of mediating variables that intercept all indirect paths between X and Y. Graphical identification conditions for expressions of the type $E(Y|do(x),do(z_1),do(z_2),\ldots,do(z_k))$ were derived by Pearl and Robins (1995) (see Pearl 2000a, Chapter 4) and invoke sequential application of the back-door conditions discussed in Section 3.2.

5.2. Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y; therefore, the direct effect is independent of the values at which we hold Z. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z).

When the direct effect is sensitive to the levels at which we hold Z, it is often meaningful to average the direct effect over those levels. Conceptually, we can define the average direct effect $DE_{x,x'}(Y)$ as the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they *would have obtained* under $do(x)$. This hypothetical change, which Robins and Greenland (1991) called "pure" and Pearl (2001) called "natural," mirrors what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.)

and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)).

Extending the subscript notation to express nested counterfactuals Pearl (2001) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{12}$$

Here, Y_{x',Z_x} represents the value that Y would attain under the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. We see that $DE_{x,x'}(Y)$, the natural direct effect of the transition from x to x', involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified, even with the help of ideal, controlled experiments (see footnote 11 for intuitive explanation). Pearl (2001) has nevertheless shown that, if certain assumptions of "no confounding" are deemed valid,[16] the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y|do(x',z)) - E(Y|do(x,z))]P(z|do(x)). \tag{13}$$

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z|do(x))$ as a weighing function.

In particular, expression (13) is both valid and identifiable in Markovian models, where each term on the right can be reduced to a "*do*-free" expression using equation (4).

5.3. Natural Indirect Effects

Remarkably, the definition of the natural direct effect (12) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the $do(x)$ operator, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The natural indirect effect, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads (Pearl, 2001):

$$IE_{x,x'}(Y) \stackrel{\Delta}{=} E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{14}$$

which is almost identical to the direct effect (equation (12)) save for exchanging x and x'.

Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \stackrel{\Delta}{=} E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{15}$$

16. One sufficient condition is that $Z_x \perp\!\!\!\perp Y_{x',z} | W$ holds for some set W of measured covariates. See details and graphical criteria in Pearl (2001, 2005) and in Petersen et al. (2006).

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{16}$$

Since each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.

Note that, although it cannot be expressed in *do*-notation, the indirect effect has clear policy-making implications. For example: in the hiring discrimination context, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of X on Y through a selected set of paths (Avin et al., 2005).

Note that in all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001) has suggested therefore that *signal sensing* is more fundamental to the notion of causation than *manipulation*; the latter being but a crude way of stimulating the former in experimental setup. The mantra "No causation without manipulation" must be rejected. (See Pearl 2000a, Section 11.4.5, 2nd Ed.)

It is remarkable that counterfactual quantities like DE and ID that could not be expressed in terms of $do(x)$ operators, and appear therefore void of empirical content, can, under certain conditions be estimated from empirical studies. A general characterization of those conditions is given in Shpitser and Pearl (2007).

Additional examples of this "marvel of formal analysis" are given in (Pearl, 2000a, Chapters 7, 9, 11). It constitutes an unassailable argument in defense of counterfactual analysis, as expressed in Pearl (2000b) against the stance of Dawid (2000) and Geneletti (2007).

6. Conclusions

Statistics is strong in devising ways of describing data and inferring distributional parameters from sample. Causal inference require two addition ingredients: a science-friendly language for articulating causal knowledge, and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomenon. This paper introduces nonparametric structural causal models (SCM) as a formal and meaningful language for formulating causal knowledge and for explicating causal concepts used in scientific discourse. These include:

randomization, intervention, direct and indirect effects, confounding, counterfactuals, and attribution. The algebraic component of the structural language coincides with the potential-outcome framework, and its graphical component embraces Wright's method of path diagrams (in its nonparametric version). When unified and synthesized, the two components offer empirical investigators a powerful methodology for causal inference which resolves long-standing problems in the empirical sciences. These include the control of confounding, the evaluation of policies, the analysis of mediation and the algorithmization of counterfactuals.

Acknowledgments

Portions of this paper are based on my book *Causality* (Pearl, 2000, 2nd edition forthcoming 2009a). This research was supported in parts by grants from NSF #IIS-0535223 and ONR #N000-14-09-1-0665.

References

J.D. Angrist, G.W. Imbens, and D.B Rubin. Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, 91(434):444–472, June 1996.

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, pages 357–363, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994a.

A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, pages 230–237. MIT Press, Menlo Park, CA, 1994b.

A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

R. Brent and L. Lok. A fishing buddy for hypothesis generators. *Science*, 308(5721): 523–529, 2005.

K. Chalak and H. White. An extended class of instrumental variables for the estimation of causal effects. Technical Report Discussion Paper, UCSD, Department of Economics, July 2006.

D.R. Cox. *The Planning of Experiments*. John Wiley and Sons, NY, 1958.

A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41(1):1–31, 1979.

A.P. Dawid. Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association*, 95(450):407–448, June 2000.

S. Geneletti. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)*, 69(2):199–215, 2007.

S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington, D.C., 1988.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

J. Pearl. Comment: Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–269, 1993a.

J. Pearl. Mediating instrumental variables. Technical Report Technical Report R-210, Computer Science Department, UCLA, 1993b. http://ftp.cs.ucla.edu/pub/stat_ser/r210.pdf.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, December 1995.

J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284, 1998.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000a. Second Edition forthcoming 2009.

J. Pearl. Comment on A.P. Dawid's, causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):428–431, June 2000b.

J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

J. Pearl. Statistics and causal inference: A review. *Test Journal*, 12(2):281–345, December 2003.

J. Pearl. Direct and indirect effects. In *Proceedings of the American Statistical Association, Joint Statistical Meetings*, pages 1572–1581. MIRA Digital Publishing, Minn., MN, 2005.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition, 2009a. Forthcoming.

J. Pearl. Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine*, 28:1420–1423, 2009b. http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf.

J. Pearl. Myth, confusion, and science in causal analysis. Technical Report R-348, University of California, Los Angeles, CA, 2009c. http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf.

J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

J.M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic with structural nested models. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 349–405. AAAI Press/The MIT Press, Menlo Park, CA, 1999.

J.M. Robins and S. Greenland. Estimability and estimation of expected years of life lost due to a hazardous exposure. *Statistics in Medicine*, 10:79–93, 1991.

P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P.R. Rosenbaum. *Observational Studies*. Springer-Verlag, New York, second edition, 2002.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D.B. Rubin. Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28:1420–1423, 2009.

I. Shpitser and J Pearl. Identification of conditional interventional distributions. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, Corvallis, OR, 2006.

I. Shpitser and J Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359. AUAI Press, Vancouver, BC, Canada, 2007.

H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33: 323–340, 1966.

M.E. Sobel. Causal inference in statistical models of the process of socioeconomic achievement. *Sociological Methods & Research*, 27(2):318–348, November 1998.

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

J. Tian, A. Paz, and J. Pearl. Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA, 1998.

J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 2000.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

JMLR: Workshop and Conference Proceedings 6:59–86, 2010 NIPS 2008 workshop on causality

Beware of the DAG!

A. Philip Dawid APD@STATSLAB.CAM.AC.UK
Statistical Laboratory
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB, UK

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Directed acyclic graph (DAG) models are popular tools for describing causal relationships and for guiding attempts to learn them from data. They appear to supply a means of extracting causal conclusions from probabilistic conditional independence properties inferred from purely observational data. I take a critical look at this enterprise, and suggest that it is in need of more, and more explicit, methodological and philosophical justification than it typically receives. In particular, I argue for the value of a clean separation between formal causal language and intuitive causal assumptions.

Keywords: Directed acyclic graph, conditional independence, probabilistic causality, statistical causality, causal DAG, augmented DAG, Pearlian DAG, causal discovery, causal Markov condition, reification fallacy, instrumental variable

1. Introduction

This article is based on a talk given at the 2008 NIPS Workshop *Causality: Objectives and Assessment*, where I was commissioned to play "devil's advocate" in relation to the enterprise of *causal discovery*, which — as evidenced by the other contributions to the workshop — has become an important and vibrant strand of modern machine learning. Like Cartwright (2007, Chapter II), I take a sceptical attitude to the widespread view that we can learn about causal processes by constructing DAG models of observational data.

In taking on this sceptical rôle I would not wish to be thought entirely negative and destructive: on the contrary, I am impressed by the overall quality of work in this area, be it fundamental methodology, algorithmic development, or scientific application. Moreover, many of the cautions I shall raise have been clearly identified, appreciated and enunciated by the major players in the field, and will be at the back, if not the forefront, of the minds of many of those who use the techniques and algorithms. I do feel, however, that there is still a useful contribution to be made by reiterating and to some extent reframing these cautions. A companion paper (Dawid, 2010) makes similar points, with emphasis on the variety of concepts of causality rolled up in the statistical methodology.

My principal concern here is to clarify and emphasise the strong assumptions that have to be made in order to make progress with causal modelling and causal discovery, and to argue that these should never be accepted glibly or automatically, but deserve careful attention and context-specific discussion and justification whenever the methods are applied. And, in a more positive vein, I describe a formal language that can assist in expressing such assumptions in an unambiguous way, thereby facilitating this process of discussion and justification.

1.1. DAG models

A unifying feature of the discussion is the use of directed acyclic graph (DAG) representations. These can interpreted and applied in a number of very different ways, which I attempt to elucidate and contrast. Here I give a very brief preview of these different interpretations.

Consider for example the simple DAG

(a) $X \leftarrow Z \rightarrow Y$.

One interpretation of this is as a *probabilistic DAG*, which is just a graphical way of describing the probabilistic conditional independence (CI) property $X \perp\!\!\!\perp Y \mid Z$ — and is thus interchangeable with the entirely equivalent descriptions of this CI property by means of the DAGs

(b) $X \rightarrow Z \rightarrow Y$; or

(c) $X \leftarrow Z \leftarrow Y$.

A totally different interpretation of **(a)** is as a *causal DAG*, saying that Z is (in some sense) a "common cause" of both X and Y, which are otherwise causally unrelated. Under this causal interpretation the DAGs **(a)**, **(b)** and **(c)** are *not* interchangeable.

Although these interpretations (probabilistic and causal) have absolutely nothing in common, it is often assumed that a single DAG can fulfil both these interpretative functions simultaneously. When this is so, it follows that any variable will be independent of its non-effects, given its direct causes — the *causal Markov* property that forms the basis of "causal discovery" algorithms that attempt to infer causal relationships from observationally discovered probabilistic conditional independencies. Unfortunately there is no clear way of deciding when (if ever) it is appropriate to endow a DAG with this dual interpretation.

Pearlian DAGs aim to clarify this connexion, using interventions to define causal relationships, and making strong assumptions to relate the non-interventional probabilistic regime with various interventional causal regimes. For example, this interpretation of DAG **(a)** would require that the observational joint conditional distribution of (X,Y) given $Z = z$ (under which X and Y are in fact conditionally independent) is the same as the joint distribution of (X,Y) that would ensue when we intervene on Z to set its value to z. Such Pearlian assumptions, which are testable (at least in principle), support a rich causal calculus. There are also valuable variations on this approach that

require fewer assumptions (*e.g.*, we envisage intervention at some, but not all, of the nodes in the DAG).

This abundance of different interpretations of the same DAG is rich in possibilities, but at the same time a potential source of confusion.

1.2. Outline

In § 2 I recall the importance of distinguishing between passive observation ("seeing") and intervention ("doing"). Section 3 introduces the algebraic theory of conditional independence (CI), relevant to the "seeing" context, while graphical representations of CI are described and discussed in § 4. In § 5 I switch to considering causal models and their graphical representations, as relevant to the "doing" context, while § 6 discusses possible relationships that might be assumed to hold between graphical representations in the two contexts. Section 7 treats the more specific assumptions underlying Judea Pearl's use of DAGs to represent and manipulate causality. In § 8, I comment on the strong assumptions that are implicit in these causal models. Section 9 then presents an approach to modelling causality that does not require any such assumptions (though these can be represented when desired), and this is further illustrated, and contrasted with other approaches, in §§ 10 and 11. The need for (and possibilities for) contextual justification of causal assumptions is highlighted in § 12, while § 13 summarises the arguments presented and considers what might be an appropriate rôle for causal discovery.

2. Seeing and doing

Spirtes et al. (2000) and Pearl (2009), among others, have stressed the fundamental importance of distinguishing between the activities of *Seeing* and *Doing*. *Seeing* involves passive observation of a system in its natural state. *Doing*, on the other hand, relates to the behaviour of the system in a disturbed state, typically brought about by some external intervention. For statistical applications a strong case can be made (Dawid, 2000, 2002b) for regarding the philosophically problematic concept of *causation* as simply describing how the system responds to external intervention — a stripped-down "agency" or "manipulationist" interpretation of causality (Hausman, 1998; Woodward, 2003). *Causal inference* then refers to the problem of drawing conclusions, from available data, about such responses to interventions.

The cleanest case is when the data were collected under the very interventional regime in which we are interested. "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)" (Box, 1966). This is the credo underlying the whole discipline of experimental design and inference, as exemplified by the most important medical advance of the 20th century: the controlled clinical trial.

Often, however, for reasons of cost, practicality, ethics, *etc.*, we can not experiment, but are confined to passive observation of the undisturbed system. Now it is a logically trivial but fundamentally important point that there is no necessary connexion between

the different regimes of seeing and doing: a system may very well behave entirely differently when it is kicked than when it is left alone. So any understanding one might achieve by observation of the system's undisturbed behaviour is at best indirectly relevant to its disturbed behaviour, and thus to causal inference. We might attempt to proceed by *assuming* connexions between the different regimes, which — if valid — would allow us to transfer knowledge gained from *seeing* to inferences about the effects of *doing*. But it is important to be entirely explicit about such assumptions; to attempt, so far as is possible, to justify them; and to be fully aware of the sensitivity of any conclusions drawn to their validity.

In recent years there has grown up a body of methodology, broadly described as *causal discovery*, that purports to extract causal (doing) conclusions from observational (seeing) data in fairly automatic fashion (Spirtes et al., 2000; Glymour and Cooper, 1999; Neapolitan, 2003). This approach largely revolves around directed acyclic graph (DAG) models, which have interpretations in both the seeing and the doing contexts, so that a DAG model identified from observational (seeing) data can be imbued with causal (doing) content. However, these two interpretations of DAGs, while related, are logically distinct, and have no necessary connexion. Hence it is important to clearly identify, understand, and provide contextual justification for, the assumptions that are needed to support replacement of one interpretation by another. There can be nothing fully automatic about causal discovery.

I will survey various different interpretations of DAG models, and their relationships with conditional independence.

3. Seeing: Conditional independence

We start by concentrating on the behaviour, under a single stable regime, of a collection of variables of interest. We assume that this behaviour will be modelled by means of a fixed joint probability distribution P.[1] If we can obtain and record repeated observations under the same regime, we might hope to estimate P. Here we largely ignore problems of inference, and restrict attention to purely probabilistic properties.

One of the most important of such properties is that of *conditional independence*, CI (Dawid, 1979a, 1980). We write $X \perp\!\!\!\perp Y \mid Z\ [P]$ to denote that, under the distribution P, variables X and Y are probabilistically independent given $Z = z$, for any observable value z of Z. When P can be understood we write simply $X \perp\!\!\!\perp Y \mid Z$. This can be interpreted in various equivalent ways, but for our purposes the most useful is the

1. There are of course many interpretations of probability (Galavotti, 2005). For present purposes a naïve frequentist view, which can also be given a subjective Bayesian intepretation in terms of exchangeability (de Finetti, 1975), will suffice. Williamson (2005) argues for an "objective Bayesian" interpretation as most appropriate for causal inference. The formal mathematical framework is the same in all cases.

following:[2]

$$P(X = x \mid Y = y, Z = z) \text{ depends only on } z, \text{ and not further on } y. \tag{1}$$

Universal[3] qualitative properties of probabilistic CI include (Dawid, 1979a; Spohn, 1980; Pearl and Paz, 1986):

$$\begin{array}{lll} X \perp\!\!\!\perp Y \mid X & & \\ X \perp\!\!\!\perp Y \mid Z & \Rightarrow & Y \perp\!\!\!\perp X \mid Z \\ X \perp\!\!\!\perp Y \mid Z, \quad W \leq Y & \Rightarrow & X \perp\!\!\!\perp W \mid Z \\ X \perp\!\!\!\perp Y \mid Z, \quad W \leq Y & \Rightarrow & X \perp\!\!\!\perp Y \mid (W,Z) \\ \left.\begin{array}{l} X \perp\!\!\!\perp Y \mid Z \\ \text{and} \\ X \perp\!\!\!\perp W \mid (Y,Z) \end{array}\right\} & \Rightarrow & X \perp\!\!\!\perp (Y,W) \mid Z \end{array} \tag{2}$$

(where $W \leq Y$ denotes that W is a function of Y).

There is another useful property, which is however valid not universally, but only under additional conditions (Dawid, 1979b, 1980):[4]

$$X \perp\!\!\!\perp Y \mid (Z,W) \text{ and } X \perp\!\!\!\perp Z \mid (Y,W) \Rightarrow X \perp\!\!\!\perp (Y,Z) \mid W. \tag{3}$$

While (2) (and, where appropriate, (3)) do not exhaust all the general properties of probabilistic CI (Studený, 1992), they are adequate for most statistical purposes.

4. Graphical representation

It can be helpful to use mathematical constructions of various kinds to represent and manipulate CI (Dawid, 2001a). This involves making formal analogies between properties of probabilistic CI and non-probabilistic properties of the representations we use. The representations themselves can look very different from probability distributions, and we need to be very clear as to how we are to interpret properties of such a representation as "saying something about" properties of CI. As with any use of representations to assist understanding and construct arguments, the *semantics* (or *meaning*) of a representation — describing exactly just how it is to be taken as relating to the external "reality" it is intended to represent — is at least as important as its *syntax* — describing its internal grammar.

One of the most popular and useful of such representations is the *directed acyclic graph* (DAG). A DAG † has a set $\mathcal{V}$ of nodes, and arrows joining them, with no loops or directed cycles. A full description and analysis of the formal semantics of the relationship between DAGs and the collections of CI properties they represent, together with

2. Purely for simplicity, we may here suppose the variables are discrete, and all combinations of logically possible values have positive probability. For a rigorous definition in the general case, see Dawid (1980).
3. *i.e.* holding for any distribution P and any variables $X, Y, \ldots$
4. For example, when the sample space is discrete and each elementary outcome has positive probability.

the associated notation and terminology, can be found in Cowell et al. (2007). Although this theory will be familiar to many readers, I repeat here the specific features I wish to emphasise — more to clarify what is *not* being said than what is.

4.1. d-separation

Given node-sets $S, T, U \subseteq \mathcal{V}$, we say U *d-separates* S *from* T *in* $\dagger$, and write $S \perp_d T \mid U\ [\dagger]$, if the following somewhat complex geometric property[5] is satisfied. First we delete all nodes that are not "ancestors" of some node in $S \cup T \cup U$, as well as all their incoming arrows; then we add undirected edges between any two nodes that are "parents" of a common "child" node, if they are not already joined by an arrow; next we delete all arrowheads, so obtaining an undirected graph, the relevant *moralized ancestral graph*. Finally, in this graph we look for paths joining S and T that do not intersect U. If there are none such, then S and T are d-separated by U in $\dagger$.

It turns out (Lauritzen et al., 1990) that this graph-theoretic separation property also obeys the formal rules (2) (with $\leq$ interpreted as $\subseteq$), and is thus potentially able to represent some collections of probabilistic conditional independence properties. Specifically, when the nodes of $\dagger$ represent random variables, we say that $\dagger$ *represents* a collection $\mathcal{C}$ of conditional independence relations between sets of variables if the graph-theoretic property $S \perp_d T \mid U\ [\dagger]$ holds exactly when the CI relation $S \perp\!\!\!\perp T \mid U$ either belongs to $\mathcal{C}$, or can be logically deduced from $\mathcal{C}$ by application of the rules in (2). For a probability distribution P over $\mathcal{V}$, we say $\dagger$ *represents* P if (the *Markov condition)*:

$$S \perp_d T \mid U\ [\dagger] \Rightarrow S \perp\!\!\!\perp T \mid U\ [P]. \tag{4}$$

This will be so if and only if, under P, for each $V \in \mathcal{V}$, V is conditionally independent of its parents in $\mathcal{V}$, $\mathrm{pa}(V)$, given its non-descendants in $\mathcal{V}$, $\mathrm{nd}(V)$. Such a representation is termed (probabilistically) *faithful* when the converse implication to (4) also holds, *i.e.* the *only* conditional independence properties holding in P between the variables in $\mathcal{V}$ are those represented by $\dagger$. These relationships between the d-separation properties of a DAG and a collection of CI properties, or a joint distribution P, constitute the *semantic interpretation* of the DAG.

As a simple example, Figure 1 shows the unique DAG over four variables (Z, U, X, Y) that represents the following pair of CI properties:

$$U \perp\!\!\!\perp Z \tag{5}$$

$$Y \perp\!\!\!\perp Z \mid (X, U). \tag{6}$$

It is important to note that, for given variable set $\mathcal{V}$, the collections of CI properties $\mathcal{C}$ that can be represented by a DAG are very special.[6] Thus with $\mathcal{V} = \{X, Y, Z\}$, the pair

5. We here describe the "moralisation" version of this property (Lauritzen et al., 1990). This is logically equivalent to the d-separation property as described by Pearl (1986); Verma and Pearl (1990).

6. They are exactly those that are logically equivalent (using (2)) to a collection of the form $V_i \perp\!\!\!\perp \{V_1, \ldots, V_{i-1}\} \mid S_i$ for $i = 1, \ldots, N$, where $V_1, \ldots, V_N$ is an ordering of $\mathcal{V}$, and $S_i \subseteq \{V_1, \ldots, V_{i-1}\}$. In this case the associated DAG has an arrow into each V_i from each node in S_i.

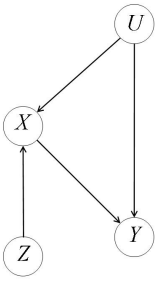


Figure 1: Simple DAG

of properties $\{X \perp\!\!\!\perp Y,\ X \perp\!\!\!\perp Y \mid Z\}$ has no DAG representation, and this is indeed the typical state of affairs. Conversely, when a DAG representation is available, it need not be unique. Distinct DAGs on $\mathcal{V}$ are termed *Markov equivalent* when they represent the same collection of CI relations: this will be so if and only if they have the same *skeleton* (undirected version) and *immoralities* (configurations of the form $A \to C \leftarrow B$ with no arrow between A and B) (Frydenberg, 1990; Verma and Pearl, 1991). Thus the three DAGs **(a)**, **(b)** and **(c)** of § 1.1 are Markov equivalent, all representing the same single CI property $X \perp\!\!\!\perp Y \mid Z$, and are all equally valid for this purpose. This representational flexibility is extended further when we allow the set $\mathcal{V}$ of variables considered to vary: thus both DAGs of Figure 2 represent the single CI property $A \perp\!\!\!\perp B$.

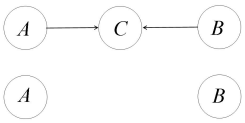


Figure 2: Two DAGs representing $A \perp\!\!\!\perp B$

4.2. What do the arrows mean?

According to the theory presented above, the purpose of a DAG representation is to mirror, *via* the d-separation semantics described in § 4, the probabilistic relationship of conditional independence — a relationship that, it is worth emphasising, is entirely symmetrical, as captured by the second line of (2). However, it is in the very nature, and indeed name, of a directed acyclic graph that it contains *directed* arrows between variables, so that this particular graphical representation embodies a non-symmetrical relationship between nodes. But this is a pure artifact: thus Figure 1, although composed of directed arrows, is nothing but an alternative way of representing the symmetrical CI relationships (5) and (6). The rôle of an arrow in a DAG model is much like that of a construction line in an architect's drawing: although it plays an important rôle in the

formal syntax of the model, it has no direct counterpart in the world, and contributes only indirectly to the semantic interpretation of the model.

4.3. Reification

Nevertheless, having built a DAG representation of a probability distribution, it is hard to resist the temptation to interpret an arrow from node X to node Y in the DAG as representing something meaningful in the real-world system that the DAG is modelling — for example, as embodying some conception of the non-symmetrical relation of *cause and effect*: that X is, in some sense, a "direct cause" of Y. Likewise, we might be tempted to read off from the Figure 1 such intuitive properties as "X lies on the causal pathway between Z and Y". But no such inferences are justified from the formal semantics relating DAG representations to conditional independence. Such interpretation of an incidental formal attribute of a mathematical representation of the world as corresponding to something real in the external (physical or mental) world[7] may be termed "reification". While reification can often be indicative and fruitful, it is important to be very clear as to when we are reaching beyond the formal semantics by which the representation has been supposed to encode real-world properties, and in that case to consider very carefully whether, when and how this might be justifiable.

5. Causal DAGs

An entirely different use of a DAG representation is to model causal relations directly. Unlike conditional independence, which is a clearly defined property of a probability distribution, causality is a slippery and ambiguous concept. In dealing with causal relations, we can either regard them as fundamental undefined primitives in themselves, or as defined in terms of still more basic ingredients, such as the effect of interventions. In either case the important thing, if a representation is to be used for meaningful communication, is that all parties have the same (explicit or implicit) understanding of the things it is supposed to be representing, and of the nature and mechanics of the representation.

A common causal interpretation of a DAG is along the following lines, quoted from Hernán and Robins (2006) (their Figure 2 is redrawn here as our Figure 3), in discussion of a certain problem relating to the use of "instrumental variables" (see § 10 below):

> "A `causal DAG` is a DAG in which:
>
> (i). the lack of an arrow from V_j to V_m can be interpreted as the absence of a `direct causal effect` of V_j on V_m (`relative to` the other variables on the graph)[8]
>
> (ii). all `common causes`, even if unmeasured, of any pair of variables on the graph are themselves on the graph. In Figure 2... the inclusion

7. Bourdieu (1977, *p.* 29) speaks of "sliding from the model of reality to the reality of the model"
8. A stronger and potentially more useful requirement is that an arrow be present *if and only if* there is such a direct causal effect.

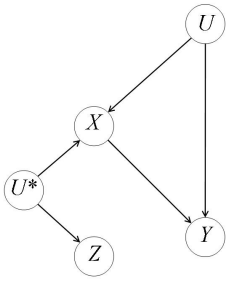


Figure 3: Figure 2 of Hernán and Robins (2006)

> of the measured variables (Z, X, Y) implies that the causal DAG must also include their unmeasured common causes (U, U^*)."

Here we have used `teletype font` (not in the original) to highlight non-mathematical causal concepts.[9] Only when we have pre-existing understanding and interpretation of these concepts will it be possible to say whether or not a given DAG is indeed a *causal* DAG for the problem it is intended to represent. Such a question can never be addressed solely in formal terms, by reference only to the DAG. In particular, we can not define concepts such as "direct causal effect" or "common cause" by reference to a putative DAG model unless that model has previously been justified as "causal" by other, necessarily non-graphical, considerations not involving these terms.

However we may choose to understand the causal terms involved, it is clear that the semantics whereby a DAG represents causal properties are qualitatively totally different from those whereby it represents conditional independence properties. In the former case, the arrows are supposed to have a direct interpretation in terms of cause and effect; whereas, as emphasised in §4.2, for conditional independence the arrows are nothing but incidental construction features supporting the d-separation semantics. A related distinction is that conditional independence is an externally determined all-or-nothing affair, whose validity is unaffected by which other variables and properties we may choose to represent in our DAG: this means, in particular, we can not interpret the presence or absence of an arrow from X to Y in some probabilistic DAG representation as representing a fundamental CI property, since such arrows can come and go as we vary the set of variables represented. In contrast, the very meaning of the properties (such as "direct effect") represented by a causal DAG may be dependent on the specification of the variable set $\mathcal{V}$, and may change (with corresponding changes in the relevant representation) as we vary $\mathcal{V}$. Correspondingly an arrow in a causal DAG *can* be considered as having independent meaning (*e.g.* as representing a "direct effect") — albeit only in terms of causal concepts defined *relative* to the specific variables represented.

Contrasting conditional independence and causality, we see that both in their subject matter and in their graphical representation they differ markedly. We could simply

9. Note that (i) involves a causal concept that is not regarded as absolute, but rather as relative to a specific collection of variables under consideration.

keep them in entirely different pockets, with nothing whatsoever to do with each other. However, there is a long-standing tradition of attempting to forge connexions between the two. Indeed, some thinkers (Shafer, 1996; Spohn, 2001) regard the very concept of causality as entirely supervenient on that of conditional independence (for an appropriate collection of variables).

6. Probabilistic causality

One approach, going back at least to Reichenbach (1956) and pursued by Suppes (1970) among others, essentially proceeds by relating the *probabilistic conditional independence* of two variables X and Y given some third variable Z to the *causal independence* of X and Y, in the sense that neither of these variables causally affects the other. However it is not easy to make this precise. In one direction, it apparently implies that, if X and Y are completely independent probabilistically (so that we can take Z to be vacuous), then neither can causally affect the other. However, this degree of implication is not usually claimed, since such independence could be an accidental result of numerical cancellation between two or more non-null causal probabilistic relationships involving these and other variables.[10] Likewise, if we find that X and Y are not independent given any variable Z currently under consideration, we can not immediately deduce causal dependence between X and Y, since we can not rule out the possibility that we have simply not examined enough Zs.

In the converse direction, it might be claimed (the "weak causal Markov assumption", Scheines and Spirtes, 2008) that, if X and Y are "causally disconnected", in the sense that neither X nor Y causally affects the other and they have no other common cause Z, then they should be probabilistically independent.

6.1. Causal Markov condition

A still more thoroughgoing approach is based on the "Causal Markov condition", CMC (Spohn, 1980; Spirtes et al., 2000). Essentially, this supposes that, when we do have a causal DAG representation[11] of a system, the identical DAG will also represent its CI properties. Equivalently,[12] CMC requires that any variable be probabilistically independent of its non-effects,[13] conditional on its direct causes — all understood relative to the set of variables in the causal DAG. When valid, CMC allows us to infer conditional independence properties from causal assumptions (so long as these can be represented by a causal DAG).

10. Such a state of affairs is sometimes dismissed as being due to a "non-faithful" DAG representation of the problem. But at this level of generality we do not have a DAG.
11. *e.g.*, as described in § 5.
12. At any rate, with the stronger interpretation of footnote 8.
13. The "effect" relation here is the transitive closure of the "direct effect" relation.

6.2. Other interpretations of probabilistic causality

Recently other ideas have been suggested for relating causal relationships between variables to properties of their joint probability distribution. For example, Janzing and Schölkopf (2008b,a) distinguish between the two decompositions of a joint distribution, $p(x,y) = p(x)\,p(y|x)$ and $p(x,y) = p(y)\,p(x|y)$, in terms of their algorithmic complexity: if, say, the former is simpler by this criterion, one might regard this as indicating a causal effect of X on Y. Similarly (Zhang and Hyvärinen, 2009), if one can reasonably describe $p(y|x)$, but not $p(x|y)$, in terms of an implicit additive error structure, one might again interpret that as implying that X is a cause of Y. It is clear that the assumptions underlying such claims are very different from those of "probabilistic causality" above. The extent to which they might be appropriate, and indeed whether they even relate to the same conception of causality, deserves deeper attention.

6.3. Causal discovery

"Causal discovery" aims to deduce causal properties of a system from its CI properties, themselves typically inferred (with a consequent degree of uncertainty) from an analysis of data generated by the system. There are many variations and algorithms, but all share the same basic philosophy. The fundamental assumptions[14] needed to validate this enterprise in any particular application are:

Assumption 6.1 (Causal representation) *There exists* some *DAG* † *that is a causal DAG representation of the system.*

Assumption 6.2 (Causal Markov condition) *The identical DAG* † *also represents (by means of the Markov condition (4)) the probabilistic conditional independence properties of the system.*

The more sophisticated causal discovery methods appreciate that (especially in the light of (ii) of § 5) it would not generally be reasonable to expect the causal DAG † to involve only the variables that happen to have been measured, and so will allow for the inclusion of additional unobserved variables.

Some putative causal DAG representations might be eliminated directly on *a priori* grounds, *e.g.* taking into account temporal order. Under Assumption 6.2, any remaining putative causal DAG representation will have implications for the probability distribution over the observed variables — either directly in terms of conditional independencies when there are no unobserved variables in the causal DAG, or more subtle consequences of "latent conditional independence" when there are. Consequently, if those implications are not supported by the data, then the hypothesised causal DAG may be eliminated on empirical grounds. In this way, and under the assumptions made, we can gain partial knowledge of causal structure from a combination of *a priori* reasoning and observational data.

To make further progress, it is common to strengthen Assumption 6.2 as follows:

14. There are also variations using different graphical representations of causal and CI properties, such as partial ancestral graphs (Richardson and Spirtes, 2002; Zhang, 2008).

Assumption 6.3 (Causal faithfulness) *The causal DAG † is a probabilistically faithful representation of the system.*

In this case, the *only* conditional independence properties enjoyed by the variables in † will be those represented by the causal DAG †. Under Assumption 6.1 and Assumption 6.3, knowledge of which (patent or latent) conditional independencies do or do not hold between the observed variables allows us to eliminate still more putative causal DAG representations of the problem. However, even if we knew which variables were to be included in the causal DAG †, and all the conditional independence properties they possess, we might not be able to identify † uniquely, since we could never distinguish observationally[15] between distinct DAGs that are Markov (but not causal) equivalent. Even with this remaining ambiguity, it may be possible to make some causal inferences: thus if every uneliminated causal DAG description of the problem involves the same variable set $\mathcal{V}$, and all contain an arrow from X to Y, we can infer that X is a direct cause of Y relative to $\mathcal{V}$.

Zhang and Spirtes (2008) point out that certain implications of the combination of Assumptions 6.1 and 6.3 can be tested empirically. For it follows from Assumption 6.3 that the conditional independence properties of the observational distribution P are faithfully represented by *some* DAG, and this property has testable consequences. But this does not make much progress towards *causal* inference without the additional strong Assumption 6.1.

7. Pearlian DAGs

Judea Pearl, through his book (Pearl, 2009) and many other works, has popularised a particular use of DAGs to represent both CI and causal properties simultaneously — the latter understood as describing the effects of interventions. We shall refer to a DAG imbued with such an interpretation as a *Pearlian DAG*.[16]

Such a representation applies to a collection of variables measured on some system, such that we can intervene (or at least can conceive of the possibility of intervening) on any one variable or collection of variables, so as to "set" the value(s) of the associated variable(s) in a way that is determined entirely externally. This gives rise to a wide variety of interventional regimes, while the observational regime arises as the special case that no variables are set. A DAG † is then a Pearlian representation of the system when the following properties hold:

15. although we may be able to do so on *a priori* grounds

16. We in fact shall deal only with Pearl's initial, fully stochastic, theory. More recently (see the second-half of Pearl (2009), starting with Chapter 7), he has moved to an interpretation of DAG models based on deterministic functional relationships, with stochasticity deriving solely from unobserved exogenous variables. That interpretation does however imply all the properties of the stochastic theory, and can be regarded as a alternative description of it. (This is however not so when we move from DAG models to more general representations, when such deterministic models have restricted generality: see Example 11.2 below.)

Property 7.1 (Locality) *Under regimes in which all variables other than V are set, at arbitrary values, the associated distribution of V depends only on the settings of its parents,* $\mathrm{pa}(V)$*, in* †.

This can be interpreted as requiring that only the DAG parents of V have a direct effect on V, relative to the other variables in the DAG.

Property 7.2 (CMC) *Under any regime,* † *represents (by means of the Markov condition (4)) the probabilistic conditional independence properties of the associated joint distribution.*

Under any interventional regime that sets the value of $V \in \mathcal{V}$, there trivially can be no dependence of the (one-point) distribution of V on $\mathrm{pa}(V)$: the arrows into V could thus be removed while retaining a DAG representation of this regime.

Under Property 7.1, † can plausibly be interpreted as a causal DAG representation of the problem: property (i) of § 5 is incorporated in Property 7.1, while property (ii), though not directly interpreted or represented, might be regarded as implicit in Property 7.2. With this interpretation, Property 7.2, applied to the observational regime, implies the causal Markov condition.

However, a Pearlian DAG representation must also satisfy an additional "modularity" (or "invariance") condition:

Property 7.3 (Modularity) *For any node $V \in \mathcal{V}$, its conditional distribution, given its DAG parents* $\mathrm{pa}(V)$*, is the same, no matter which variables in the system (other than V itself) are intervened on.*

(Note that Property 7.1 follows from Property 7.2 combined with Property 7.3).

Property 7.3 extends CMC: not only can we relate the *qualitative* conditional independence properties and causal properties represented by † (as embodied in CMC), but we can further relate the various *quantitative* distributional behaviours of the system when subjected to different interventions. In particular, from purely observational data on $\mathcal{V}$ we could estimate the modular parent-child distributions, and piece these together to deduce the joint distribution for the system under any set of interventions: a fully quantitative solution to the problem of inferring causality from observational data.

We see that a Pearlian DAG representation embodies CMC (for its particular interpretation of causality), but much more besides. When we can assume that a system is represented by some Pearlian DAG, we can attempt "quantitative causal discovery", in which we attempt to learn the quantitative as well as the qualitative causal structure of the problem, as embodied in the underlying Pearlian DAG.

8. How do we get started?

As is brilliantly attested by the work of Pearl, an extensive and fruitful theory of causality can be erected upon the foundation of a Pearlian DAG. So, when we can assume that a certain DAG is indeed a Pearlian DAG representation of a system, we can apply

that theory to further our causal understanding of the system. But this leaves entirely untouched the vital questions: when is a Pearlian DAG representation of a system appropriate at all?; and, when it is, when can a specific DAG † be regarded as filling this rôle? As we have seen, Pearlian representability requires many strong relationships to hold between the behaviours of the system under various kinds of interventions.

Causal discovery algorithms, as described in § 6.3, similarly rely on strong assumptions, such as Assumption 6.1 and Assumption 6.2, about the behaviour of the system. The need for such assumptions chimes with Cartwright's maxim "No causes in, no causes out" (Cartwright, 1994, Chapter 2), and goes to refute the apparently widespread belief that we are in possession of a soundly-based technology for drawing causal conclusions from purely observational data, without further assumptions.[17] This belief perhaps arises because every DAG model can be given both a probabilistic and a causal interpretation, so it is easy to conclude that, once we have derived a DAG model to describe observational conditional independencies, it must necessarily also be interpretable according to more sophisticated causal semantics (*e.g.*, as a Pearlian DAG). While this is evidently untrue (in particular, distinct but Markov equivalent DAG models, representing identical observational CI properties, will always have different implications when interpreted causally), such reification of a DAG CI representation can be very tempting.

In my view, the strong assumptions needed even to get started with causal interpretation of a DAG are far from self-evident as a matter of course,[18] and whenever such an interpretation is proposed in a real-world context these assumptions should be carefully considered and justified. Without such justification, why should we have any faith at all in, say, the application of Pearl's causal theory, or in the output of causal discovery algorithms?

But what would count as justification? We return to this important question in § 12. For the moment we merely remark that it cannot be conducted entirely within a model, but must, as a matter of logic, involve consideration of the interpretation of the terms in the model in the real world.

9. A formal language for causality

Another difference between a DAG representation of CI and a DAG representation of causality is that the former is always available, while the latter is not. In particular, a complete DAG over a collection of variables is totally non-committal as to their CI properties, and so (vacuously) correct. However, interpreted causally, even a complete DAG makes strong assertions. If we do not wish to make any such assertions, we can not even begin to consider using a causal DAG representation.

17. See Geneletti (2005) for further discussion of the hidden assumptions made in this enterprise.
18. It is commonly recognised (Scheines and Spirtes, 2008) that there are cases where such assumptions should *not* be expected to hold, such as in the presence of measurement error or coarsening (which might however be rehabilitated by including the original variables in the DAG), and, more fundamentally, when dealing with dynamic processes in equilibrium (Dash, 2005).

A less restrictive approach to causal modelling (Didelez and Sheehan, 2007a) is to develop a formal framework, with clear semantics relating mathematical properties of a putative representation to causal properties of the external system it is intended to represent, but without any commitment as to what properties the system should have: such properties should be expressible within the system, but not imposed by it. In particular, no rigid assumptions about how causality relates to probability need be made. Rather, the aim is to present a completely general language, in terms of which we can clearly express and manipulate whatever tentative causal assumptions we may wish to entertain in a specific context (in particular, it should be possible to make no such assumptions whatsoever). In these respects the rôle of such a theory would be similar to that of the theory of probabilistic conditional independence, as described in Sections 3 and 4.

One way of proceeding involves extending that same CI theory into the causal domain, using a manipulationist conception of causality (similar to that underlying the approach of Pearl). The basic ingredients are of two kinds, intended to represent, respectively, the variables ("domain variables") in the system, and the "regimes" under which those variables are generated. For application to modelling a particular external system, we must fully understand what real-world variables are supposed represented by the domain variables in the model, and what real-world regimes by the regime variables in the model. To accommodate our manipulationist stance, at least one of the regimes modelled should result from an external intervention.

The kind of causal property that will be expressible in this theory will concern relationships between the probabilistic behaviours of the domain variables, across the various regimes. Specifically, we are able (but are not obliged!) to postulate the identity, across two or more regimes, of the *conditional distribution* for one set of domain variables given another set of domain variables. When this holds we can regard that conditional distribution as a stable "modular component",[19] transferable across regimes.

This invariance or (stochastic) "stability" concept, in addition to being fundamental to my interpretation of causality, has other useful applications, arguably outside "causal inference", which can be modelled and analysed in essentially the same way. Thus we might consider the differing probabilistic behaviours of some collection of random variables in various different hospitals. We could then introduce a non-random regime indicator (but now without an interventional interpretation) to index which hospital we are looking at: this would allow us to express an assumption that a certain conditional distribution is the same in all hospitals. Or (see Example 9.1 below), we could express the property that a certain imperfect diagnostic test has the same error probabilities, no matter who it is used on. Such "reusable invariant modules" can be conveniently

19. Modularity — though more typically conceived in terms of transferable *deterministic* relationships between variables — has often been taken as an essential or defining property of causality, though this view has been challenged (Cartwright, 2007, Chapter II-3). While I make no metaphysical commitment to modularity as essential to the understanding of causality, nor even to the expression of modularity solely in terms of invariant conditional distributions, I consider that this particular approach covers a very great deal of ground, and is able to handle most aspects of "statistical" causality. A similar approach, regarding causality as residing in the "structural stability" of random variation, is taken by Russo (2008).

implemented in "object-oriented" software such as HUGIN 6[20](Dawid et al., 2007), and have been found useful in generic schemes for handling and interpreting evidence (Hepler et al., 2007).

We observe that Property 7.3 of a Pearlian DAG representation is of just this modular form. The essential difference between Pearl's approach and that described here is that, in a Pearlian DAG model, Property 7.3 requires many modularity properties to hold — for each $V \in \mathcal{V}$, under many different observational-interventional regimes — in a way that is fully determined by the form of the DAG. In contrast, we do not seek to impose any particular modularity requirements, nor do we require that the problem be representable by a DAG. We simply provide a language for expressing and manipulating any modularity properties that we might think it appropriate, on the basis of subject matter understanding, to impose or hyothesise. As we shall see in §9.2, in some (special) cases such more limited assumptions can themselves be usefully represented by DAG-type models, but these will be non-prescriptive, and will make explicit exactly what modularity assumptions it has been considered appropriate to incorporate.

9.1. Extended conditional independence

Suppose then that there is a collection of domain variables that together describe relevant aspects of the behaviour of a system under each regime of interest. Under any one of these regimes, these variables will have a joint distribution. Any conditional independence properties that distribution may have could be expressed algebraically as in §3 or — where appropriate — graphically as in §4. We now indicate how to extend such mathematical representations to incorporate any relationships, as described above in terms of invariant conditional distributions, that might be assumed to hold between the various different regimes.

Example 9.1 As a simple example, let X be a patient's actual systolic blood pressure, and Y the value of this as recorded on a certain sphygnomanometer. The same sphygnomanometer might be used on different patients at different times, but it might be reasonable to assume that the distribution of Y given X is stable, irrespective of the circumstances of use. We could introduce a regime indicator F, whose values specify the conditions, environment, kind of patient, *etc.*. Note that whereas the domain variables (X,Y) are random, F is not: rather, it has the status of a *statistical parameter*, indexing the probabilistic regime under consideration. In particular, any probability or independence statements must, explicitly or implicitly, be conditioned on the value of F.

The stability assumption is just that the conditional density $p(y \mid F = f, X = x)$ for Y, given $X = x$, in regime $F = f$, is in fact the same for all values of f. In the light of (1), we see that this can be expressed in the form of a conditional independence property:

$$Y \perp\!\!\!\perp F \mid X. \tag{7}$$

□

20. http://www.hugin.com/

It is important to note that expression (7) makes sense, even though F is not a random variable. In general, for the expression $X \perp\!\!\!\perp Y \mid Z$ in (1) to be meaningful, while X must be random there is no requirement that the conditioning variables Y and Z be random: either or both could be a parameter variable or regime indicator (see Dawid (1979a, 2002b) for further details). This language of *extended conditional independence* (ECI) thus provides a natural way of expressing stability across regimes of modular conditional distributions. In particular, ECI supplies an appropriate formal language (syntax and semantics) for describing and handling causality in our modular manipulationist understanding of the term.

Example 9.2 Consider a system involving domain variables Z, U, X, Y. We wish to model the effect of an intervention that sets the value of X. To this end we introduce an *intervention variable*, F_X, a special regime indicator with values corresponding to the different regimes that arise on intervening to set the value of X in various ways (Spohn, 1976; Spirtes et al., 2000; Pearl, 2009). If X is binary, then F_X might have values $\emptyset$, 0 and 1, the interpretation being that, when $F_X = \emptyset$ (the *idle* regime), the domain variables arise from the undisturbed system; whereas when $F_X = 0$ [resp., 1] they arise from the system disturbed by an external intervention that forces X to take the value 0 [resp., 1].

In general, the joint distributions of (Z, U, X, Y) under the three different regimes (*i.e.*, given $F_X = \emptyset$, 0 or 1) could be entirely arbitrary,[21] and unrelated to each other. But *should* we wish to specify or describe connexions between them, we can usefully do so using ECI. This programme can be effected, in great generality, in entirely algebraic fashion: we can use the general properties (2) and (with due care) (3) to manipulate ECI properties, almost exactly as for probabilistic CI. We just have to ensure that no non-random variable occurs as the first term in any ECI relation in either our assumptions or our conclusions.

Again, these manipulations are most conveniently described and conducted in graphical terms — though we once again warn that by no means every problem that can be manipulated algebraically can be modelled graphically. □

9.2. Augmented DAGs

Just as for regular CI it is sometimes possible, and then is helpful, to represent a collection of ECI properties by means of a DAG[22] — but now extended to include nodes to represent non-random regime variables (generally drawn as square), in addition to nodes representing domain variables (generally drawn as round). Indeed, this can be done with essentially the identical constructions and interpretations as for regular DAGs. Such a DAG is termed an *influence diagram* (ID) (Dawid, 2002b).

Many of the IDs considered in a causal context have a specific form, as "*augmented DAGs*"

21. Except that, to express our intended interpretation of F_X, under $F_X = 0$ [resp., 1] we should require $X = 0$ [resp., 1] with probability 1. There is however no immediate implication for the distribution of any other variables.
22. Other kinds of graphical CI representations can be similarly extended to include intervention variables (Dawid, 2002a; Zhang, 2008; Eichler and Didelez, 2009).

(Pearl, 1993). Figure 4 shows an augmented DAG, a variation on the simple, purely probabilistic, DAG of Figure 1, that also incorporates, in a particular way, an *intervention node* F_X, interpreted as in Example 9.2.

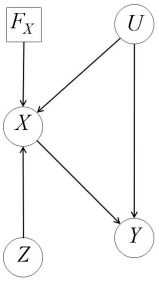


Figure 4: Augmented DAG

What does it mean to say that a particular system is modelled by this augmented DAG? To address this question, we apply the "d-separation semantics" described in § 4 — but now ignoring the distinction between domain and regime variables. The DAG thus represents the following (algebraically expressed) conditional independence properties:

$$(U,Z) \perp\!\!\!\perp F_X \tag{8}$$
$$U \perp\!\!\!\perp Z \mid F_X \tag{9}$$
$$Y \perp\!\!\!\perp F_X \mid (X,U) \tag{10}$$
$$Y \perp\!\!\!\perp Z \mid (X,U;F_X). \tag{11}$$

Using (1), property (8) is to be interpreted as saying that the joint distribution of (U,Z) is independent of the regime F_X: *i.e.*, it is the same in all three regimes. In particular, it is unaffected by whether, and if so how, we intervene to set the value of X. The identity of this joint distribution across the two interventional regimes $F_X = 0$ and $F_X = 1$ could be interpreted as expressing a causal property: manipulating X has no (probabilistic) effect on the pair of variables (U,Z). Furthermore, since this common joint distribution is also supposed the same in the idle regime, $F_X = \emptyset$, we could in principle use observational data to estimate it — thus opening up the possibility of causal inference.

Property (9) asserts that, in their (common) joint distribution in any regime, U and Z are independent: this however is a purely probabilistic, not a causal, property.

Property (10) says that the conditional distribution of Y given (X,U) is the same in both interventional regimes, as well as in the observational regime, and can thus be considered as a modular component, fully transferable between the three regimes — again, I regard this as expressing a causal property.

Finally, property (11) asserts that this common conditional distribution is unaffected by further conditioning on Z (not in itself a causal property).

Just as for regular CI, it is possible for a collection of ECI properties to have more than one representation as an augmented DAG. This is the case for Figure 5, where the direction of the arrow between U and V is not determined.

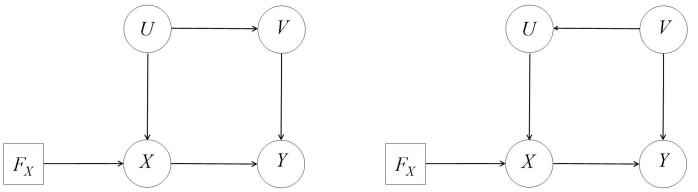


Figure 5: Two Markov-equivalent augmented DAGs

We see that the ingredients required for at least some causal assertions and inferences — namely that certain marginal or conditional distributions be unaffected by whether or how certain interventions are made — are readily expressible using the familiar language of conditional independence (specifically, they arise when the second argument of an ECI relation is a regime variable). They are just as readily manipulated by means of the rules embodied in (2). And in those special cases that it is possible to express all the causal and probabilistic assumptions made in a form that can be represented by an augmented DAG, we can use the d-separation semantics of §4 as a "theorem-proving machine" to discover their logical implications.

9.3. Pearlian DAGs as augmented DAGs

A Pearlian DAG is readily represented as a special kind of augmented DAG. To do this, we elaborate the given DAG by including, for *every* domain variable V in it, an intervention node F_V, and an arrow pointing from F_V to V. Properties 7.1, 7.2 and 7.3 are then explicitly represented by the d-separation semantics. Correspondingly all the implications of a Pearlian representation can be deduced from this augmented DAG and d-separation.

In Pearl's earlier work (Pearl, 1993, 1995) he moved backwards and forwards between explicit and implicit representation of the intervention variables in the DAG. More recently he, and most of those following him, have been using only the implicit version, in which the intervention variables F_V are not explicitly included in the diagram, but (to comply with the Pearlian interpretation) the DAG is nevertheless to be interpreted as if they were. I regard this demotion of the intervention indicators as a retrograde move, since the resulting graphical representation, while imbued with Pearlian causal semantics, is visually indistinguishable from a DAG used to describe purely probabilistic CI. Consequently, great care is needed to be clear just what a given DAG is intended to represent, and to avoid slipping unthinkingly from one interpretation to

another. Explicit representation of intervention nodes helps to guard against such confusion, as well as simplifying interpretation and manipulation.[23]

10. Instrumental variables

To clarify the similarities and differences between augmented DAG representations and other causal DAG representations, we revisit the example of § 5. Hernán and Robins (2006) present the causal DAG of Figure 3 as a counterexample to the supposition (Martens et al., 2006) that the following conditions are necessary for a variable Z to qualify as an "instrumental variable" for estimating the causal effect of an "exposure" X on a "response" Y, in the presence of an additional, unmeasured, variable U (a confounder), that affects both X and Y, when we can not directly manipulate X:

(i). Z has a causal effect on X

(ii). Z affects the outcome Y only through X (*i.e.*, no direct effect of Z on Y)

(iii). Z does not share common causes with the outcome Y (*i.e.*, no confounding for the effect of Z on Y).

The causal DAG presented by Hernán and Robins (2006) as embodying these assumptions is essentially the same as our Figure 1. This is contrasted with the causal DAG of Figure 3, which is not regarded as embodying condition (i), since Z has no direct causal effect on X, but is merely associated with it through sharing a common cause U^*.

Note that the descriptions of both problems employ intuitive causal terms, and that these are associated with the presence and directionality of the arrows in the causal DAG representations.

DAG representations of the ECI versions of these stories are presented in Figure 4 and Figure 6. In each case an intervention node F_X associated with X has been added, describing three regimes of interest: the idle regime $F_X = \emptyset$ corresponding to pure observation, and the two interventional regimes $F_X = 0$ and 1, corresponding to an intervention in which X is externally manipulated to take values 0 and 1, respectively. While data can be gathered only under the idle regime, which is thus all that can be directly estimated, our interest is nevertheless in estimating (if possible), and, especially, comparing, the distributions of the response Y under the interventional regimes, $F_X = 0$ and $F_X = 1$.

Now in the story represented by Figure 3 or Figure 6, the variable U^*, while apparently required for a full causal specification of the structure of the problem, plays no rôle in the analysis of Z as an instrumental variable. So we can restrict attention to the joint

23. For example, Pearl (1995) derives his "do-calculus" rules using an explicit augmented DAG representation, but then re-expresses them in terms of the unaugmented graph — when they become considerably more complex. It is not clear what is gained to compensate for this loss of transparencqy.

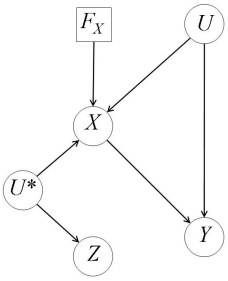


Figure 6: Augmented DAG corresponding to Figure 3

distribution, under the various regimes, of the variables (U,X,Y,Z), and their independence properties. We then find that the augmented DAG of Figure 6 embodies the identical conditional independence properties (8)–(11) as the alternative augmented DAG of Figure 4, in which U^* does not figure at all. Consequently, for our purposes this is just as good an augmented DAG representation of the problem as Figure 6. This equivalence should be contrasted with the apparent attitude of Hernán and Robins (2006), that if Figure 3 is a "true" causal DAG representation of the problem, then Figure 1 is not. We would likewise have to distinguish these two DAG representations under a Pearlian interpretation, which would be equivalent to attaching intervention nodes to *every* domain variable. But this would involve making many additional and possibly questionable assumptions, none of which is needed for the analysis.

An important interpretive difference between causal DAGs as described in § 5 and as represented by augmented DAGs is that, in the former, causal meaning is understood as carried by the arrows, whereas, in the latter, it is entirely carried by extended conditional independence properties, involving intervention variables, which are represented only indirectly in the DAG, *via* d-separation. In particular, in Figure 4 (and in contrast to the causal interpetation of Figure 1) the arrow from Z to X is *not* to be construed as representing a relationship of cause and effect between Z and X (see Didelez and Sheehan (2007b) for more on this in the context of Mendelian randomization).

The ECI properties (8)–(9) are "core conditions" for a variable Z to be an *instrument* for the effect of X on Y.[24] Once so characterised, these properties can be manipulated algebraically using the rules of (2) (together with properties such as $F_X = 0 \Rightarrow X = 0$), without reference to any graphical representation: the "theorem-proving" properties of DAG representations, while immensely useful, are logically inessential. But if we do want to use graphical representations to help us, there is no point in arguing whether it is Figure 4 or Figure 6 that is "correct" — since each of them embodies (8)–(9) equally well.

24. There is one more core condition, expressible in terms of ECI though not graphically representable: $X \not\perp\!\!\!\perp Z \mid F_X = \emptyset$. In addition to these core conditions, precise identification of a causal effect by means of an instrumental variable requires further modelling assumptions, such as linear regressions (Didelez and Sheehan, 2007b).

11. Non-DAG modularity

Bertrand Russell (Russell, 1913) famously opined "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm". Many scientific laws are symmetric, and thus inappropriate for representation in terms of directional causal properties. Nevertheless, they can still be described by means of modularity properties, and these can frequently be expressed using ECI.[25]

Example 11.1 Ideal gas

Consider a system involving a fixed number N of molecules of a monatomic ideal gas in an impermeable container, whose volume can be adjusted by manipulating a piston. The container is immersed in a heat bath of constant absolute temperature t^*. Let V, P denote, respectively, the volume and pressure of the gas.

Suppose first that the manipulations of the piston are *isothermal*, *i.e.* slow enough that, by heat transfer through the walls of the container, the gas always remains at the external temperature t^*. Then P and V are functionally related by Boyle's law:

$$PV = c \tag{12}$$

where the constant c is kNt^*, k being Boltzmann's constant.

Let the regime indicator F describe various isothermal manipulations of the piston, in either fixed or random ways. In all cases the relation (12) will hold. In particular, given either V or P, the other is determined, irrespective of the regime that brought the situation about. We shall thus have, simultaneously, the ECI properties

$$P \perp\!\!\!\perp F \mid V \tag{13}$$

$$V \perp\!\!\!\perp F \mid P \tag{14}$$

where, for example, the modular conditional distribution of P given $V = v$ associated with (13) is a 1-point distribution at c/v. We note that there is no DAG representation of the pair of ECI properties (13) and (14).

We could alternatively consider *adiabatic* manipulations, which proceed sufficiently fast that no heat can transfer through the walls of the container (but not so fast as to add energy to the gas). Then (12) is replaced by

$$PV^{\frac{5}{3}} = \text{ constant.} \tag{15}$$

Again (13) and (14) will hold, but now with different specifications for the modular conditional distributions.

25. Alternative modular descriptions can also be used in this more general context, for example, based on non-recursive systems of simultaneous structural equations, such as in Pearl (2009, Chapter 7) (see Example 11.2 below). An immediate advantage of the ECI description over representations in terms of equations is that, because of its relationship with conditional distributions, each ECI property, of the form $X \perp\!\!\!\perp Y \mid Z$, automatically comes with an associated directionality: from its conditioning variables (Y, Z) to its response variables X.

Finally, consider arbitrary manipulations of the piston. There are now no modular relationships holding between P and V, but modularity can be restored by introducing the additional variable T, the absolute temperature of the gas. The (symmetrical) invariant relationship is

$$PV = kNT. \tag{16}$$

In terms of ECI we have

$$P \perp\!\!\!\perp F \mid (V,T) \tag{17}$$
$$V \perp\!\!\!\perp F \mid (P,T) \tag{18}$$
$$T \perp\!\!\!\perp F \mid (P,V) \tag{19}$$

where, for example, the modular conditional distribution of P given $V = v, T = t$ associated with (17) is a 1-point distribution at kNt/v. Again, there is no DAG representation of this collection of ECI properties. □

Example 11.2 Price and demand

A simple econometric model relates price, P, and quantity demanded, Q, for some good. It is supposed possible to manipulate either of these to any given value. There are additional unobserved explanatory variables U_P, U_Q, which are supposed unaffected by such manipulations, having a given joint distribution. Let F_P, F_Q denote the indicators for interventions at P, Q respectively. We suppose we have specified the interventional conditional distribution of Q, given $(P = p, U_P, U_Q; F_P = p, F_Q = \emptyset)$, and that this does not in fact depend on U_P; and similarly we have specified the conditional distribution for P, given $(Q = q, U_P, U_Q; F_P = \emptyset, F_Q = q)$, which is independent of U_Q.

The idle regime, when $F_P = F_Q = \emptyset$, is taken as referring to the joint distribution "in equilibrium". On the basis of economic theory it is supposed — constituting our "modular assumptions" — that all the above specified marginal and conditional distributions continue to apply, simultaneously, in this equilibrium regime. (Note that consistency conditions then constrain the possible specifications of the conditional distributions for Q and P).

The modular assumptions made are encapsulated in the following ECI properties:

$$(U_P, U_Q) \perp\!\!\!\perp (F_P, F_Q) \tag{20}$$
$$Q \perp\!\!\!\perp (F_P, U_P) \mid (P, F_Q, U_Q) \tag{21}$$
$$P \perp\!\!\!\perp (F_Q, U_Q) \mid (Q, F_P, U_P). \tag{22}$$

There is no DAG representation of this collection of properties, but they can be represented using the more general graphical semantics of *chain-graphs* (Cowell et al., 2007; Dawid, 2002a), involving undirected as well as directed links: the relevant diagram is shown in Figure 7.[26]

26. In reality we can not vary F_P and F_Q independently: at least one of them must be idle. This "variation non-independence" (Dawid, 2001a,b) could be represented in Figure 7 by a further undirected link between F_P and F_Q; however this is of no real consequence here.

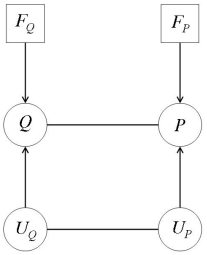


Figure 7: Chain-graph for price-demand relationship

When each of the modular distributions of Q given (P,U_Q), and of P given (Q,U_P), is concentrated on a single point, so that they represent deterministic functional relationships, and moreover those relationships are linear and (U_P,U_Q) are bivariate normal, our model is isomorphic to the structural equation model considered by Pearl (2009, §7.2.1).[27] However — and in sharp contrast to the analogous case for DAG models — even if we consider only the case that U_P and U_Q are independent, this model is *not* equivalent, in terms of the properties of the observables (Q,P) under the different regimes, to a case where the latent variables (U_P,U_Q) are absent (equivalently, taken as trivial), but we allow genuine stochasticity in the above conditional distributions. In particular, the appropriate generalisation of d-separation (Frydenberg, 1990) applied to the chain-graph of Figure 7 (even without the link U_Q — U_P) does not yield $Q \perp\!\!\!\perp F_P \mid (P,F_Q)$, although (by (21) with the U's absent) this does hold in the stochastic model. A confirmation of this non-equivalence is that, in the stochastic model without U's, the observational and interventional distributions of Q given P are identical, as follows immediately from the above relation; whereas Pearl's own analysis shows that this is typically not the case for the deterministic model incorporating U's.

We can entertain more general models, in which the U's are non-trivial and the specified distributions are genuinely stochastic. Again, the observational and interventional and distributions of Q given P will differ, but the relationship between them will be different from both the cases considered above.

Our general ECI model (20)–(22) thus incorporates Pearl's deterministic structural model, but allows other cases too — which, it could be argued, are no less obviously appropriate as descriptions of the problem. A moral of this analysis is that we should not be cavalier in setting out the ingredients (variables and modular conditional distributions) of such a model, but need to think very carefully about them in the context of the problem we are modelling and any relevant theory. And in order for us to be able to approach this task in a meaningful way, we must be able to identify the unobserved explanatory variables U_P and U_Q as real-world quantities. We can not just treat them as convenient mathematical fictions ("error terms"), for then how are we to decide whether

27. We have omitted Pearl's observable explanatory variables I, W for simplicity.

our model should be deterministic or stochastic? — a choice that will make a difference to our analysis and conclusions.

□

12. Justifying assumptions

Perhaps the most important characteristic of my suggested approach to causality, using extended conditional independence and (where appropriate) augmented DAGs or other graphical representations, is that it is descriptive, not prescriptive. It makes no assumptions as to how causality ought to behave or be represented; rather, it supplies a language by which we are able clearly to express and manipulate any such assumptions we might wish to make in any given context. In this respect it differs from other theories of "probabilistic causality"[28] in much the same way as Kolmogorov's purely formal theory of probability differs from other theories such as the "classical theory" based on the assumption that intuitively "equally possible" outcomes should be assigned equal probabilities, or von Mises's theory of collectives, which sought to represent assumed empirical properties of probability, such as the existence and stability of limiting relative frequencies, directly within the formal theory. This strict separation of the formal general-purpose language from any special assumptions that might be made in specific contexts allows for much greater clarity and flexibility. It also protects against the ever-present danger of unthinking reification of incidental formal properties of our representations. In particular, it does not in itself support causal interpretation of a probabilistic DAG. If we wish to represent this, we have very explicitly to introduce (using ECI) whatever additional assertions we are making about effects of interventions. ECI is a purely mechanical tool for manipulating causal properties, not a philosophical foundation for defining them.

This purely formal approach does, of necessity, leave entirely untouched such essential questions as "Where do we get our causal assumptions from?" and "How can they be justified?" It is at this point, entirely removed from representational issues, that we might find a place for more informal arguments, based on intuitive understandings of cause and effect.

In principle, the meaning of ECI assumptions such as (8)–(11) is straightforward; and they could indeed all be tested empirically if we had access to data collected on (U,Z,X,Y) under the various regimes. In practice, however, we will usually not have such data (and it may not even be clear which unobserved external variable or variables are represented by the symbol U). Then the appropriateness of the assumptions made requires and deserves further, necessarily context-dependent, argument.

28. By this term I do not mean to include general theories of " statistical causality," such as that of Rubin (1978), which likewise make no prescriptive assumptions. See Dawid (2000, 2002b) for comparisons and contrasts between my own approach and other approaches to statistical causality. The general points I have made could have been developed from the viewpoint of those other theories, though these mostly do not focus, as I do, on modularity at the level of conditional distributions, which supplies a natural point of contact with the intuitive concepts of "probabilistic causality".

For example, physical *randomization* of a treatment T in the "idle" regime is generally agreed to provide a convincing reason for believing that the observational distribution of a response Y, given $T = t$, is the same as its distribution would be under an intervention to set T to t (formally: $Y \perp\!\!\!\perp F_T \mid T$), thus justifying causal interpretation of these conditional distributions. Although this property of randomization is usually taken as intuitively obvious, I am not aware of any argument for it based on deeper principles. One such argument could be based on the assumed existence of some *sufficient covariate* U, such that (a) $U \perp\!\!\!\perp F_T$ and (b) $Y \perp\!\!\!\perp F_T \mid (T,U)$ (Dawid, 2002b). Here, (a) says that the distribution of U is unaffected by which regime is operating — typically believable if U is a "pre-treatment" variable; while (b) says that, conditional on U and *which* treatment T is applied, the response Y of the system is unaffected by *how* (*i.e.*, in which regime) it is applied. While it may not be easy to identify a specific pre-treatment variable U with this property, one might be willing to accept that some such variable does exist. Randomization, and the pretreatment status of U, now gives good cause to accept $T \perp\!\!\!\perp U \mid F_T = \emptyset$, whence (since T is in any case non-random in any interventional regime) (c) $T \perp\!\!\!\perp U \mid F_T$. Using the rules of (2), it is straightforward to deduce, from the three CI properties (a), (c), (b), the desired conclusion $Y \perp\!\!\!\perp F_T \mid T$. Alternatively, these CI properties can be represented by the augmented DAG of Figure 8, from which we can readily read off $Y \perp\!\!\!\perp F_T \mid T$. Similar arguments can be made to

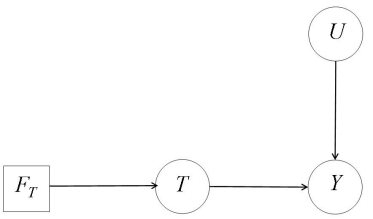


Figure 8: Augmented DAG for randomization

justify suitably expressed causal interpretations of data generated under more complex randomization schemes. But the appropriateness of any such argument needs to be carefully considered, not just taken for granted.[29]

When physical randomization is not possible, it will be necessary to attempt to justify causal CI assumptions on other grounds. For example, in the instrumental variable

29. Indeed, even in a randomized double-blind clinical trial — the "gold standard" of evidence-based medicine — one could argue that the very artificiality of the trial negates assumption (b) above: we would not expect the same response process to operate for future treated patients as for those in the trial. To make progress we might make weaker assumptions, such as transferability from the clinical trial into general practice of the "specific causal effect": $\mathrm{E}(Y \mid T = 1, U = u, F_T = \emptyset) - \mathrm{E}(Y \mid T = 0, U = u, F_T = \emptyset) = \mathrm{E}(Y \mid F_T = 1, U = u) - \mathrm{E}(Y \mid F_T = 0, U = u)$. While not expressible in terms of ECI, such an assumption still relates to the invariance of probabilistic properties across different regimes. Again, it should be made explicit, and justified (ideally empirically).

problem of § 10, we need to argue for the appropriateness of the assumptions (8)–(11). (Once again, it is enough that there exist *some* variable U, which we need not however specify in detail, for which the conditions can be assumed to hold.)

Property (8) essentially requires that both U and Z be pre-treatment variables, and then (10) implies that U must be a sufficient covariate.

Properties (9) and (11) are more problematic. Property (9) could be plausible if Z is itself determined by randomization: a scenario in which this occurs is that of "incomplete compliance" (Dawid, 2003), where patients are randomized to treatment, with randomization indicator Z, but the treatment X actually taken might not be the same as that assigned. Alternatively, in "Mendelian randomization" (Didelez and Sheehan, 2007b), Z might be a gene that naturally affects X: property (9) might then be justified, for suitable U, on the basis of the random assortment of genes under Mendelian genetics. As described by Didelez and Sheehan (2002): "If we think of U as some behavioural pattern or life style, this independence condition can be justified as long as we are reasonably certain that any possible genetic factors influencing the behavioural pattern are unrelated to this particular gene".

Finally, (11) requires that the distribution of Y given (X,U) (which has been assumed the same in all regimes) is unaffected, in any regime, by further conditioning on Z — intuitively expressed as "no direct effect of Z on Y". This might be plausible in the imperfect compliance context, where we could believe that behaviour of the response Y could depend on the treatment X actually taken and further pre-existing individual characteristics U, but not further on the treatment Z that the individual was supposed to take. In the context of Mendelian randomization, we require that "there is no association between the genotype and the disease status given the intermediate phenotype and the life style" (Didelez and Sheehan, 2002). (However, core conditions (9) and (11) can be violated in the presence of various complications, such as linkage disequilibrium, pleiotropy, genetic heterogeneity or population stratification (Didelez and Sheehan, 2007b).)

When attempting to justify the core conditions in a specific context, it is plausible that thinking about the problem in terms of further unobserved variables, such as U^* in Figure 6, can play a valuable rôle in the process. However, once these conditions have been settled on as the assumptions we wish to introduce, there is no need to make irrelevant distinctions between alternative, equally valid representations of them, such as Figure 4 and Figure 6.

In the ECI framework, attention is clearly drawn to any assumptions we may choose to make, since these have to be clearly expressed as explicit ingredients added to our model, and justified in the context of the real-world application under consideration. In other approaches the assumptions are often hidden, and it is easy to be misled into believing that they are not in need of justification. For example, the weak causal Markov assumption (§ 6) rules out certain ECI representations purely on the basis of ordinary CI properties in the observational regime; but there is no logical reason why this should be so, and its validity should be carefully considered in every intended application.

13. Conclusion

We have contrasted various approaches to the interpretation of graphical models of probabilistic causal processes. Each of these purports to relate properties of the mathematical model and properties of the process.

The most common approach, "probabilistic causality" (see § 6), works with intuitive understandings of causal terms, which are often taken as undefined and self-evident primitives, although they can also be regarded as deriving from an underlying manipulationist conception. Its most important feature is that it assumes links (*via e.g.* the "Causal Markov Condition") between such causal concepts and certain probabilistic conditional independence properties — links that, however, there is no reason to believe hold in complete generality.

In contrast the approach described in § 9, based on the algebraic theory of extended conditional independence and its graphical representations, is based on a clearly defined internal mathematical structure (syntax), and clearly described rules of interpretation (semantics). In these respects it is similar to Pearl's approach. However, unlike both that approach and that of probabilistic causality, it does not suppose any special relationship between causality and conditional independence. It merely supplies a formal language by means of which we can express and explore interesting causal conjectures, phrased as the identity of certain conditional distributions across a variety of different regimes (typically encompassing both intervention and pure observation). This surgical separation of the formal language from *ad hoc* causal assumptions enforces clear and unambiguous articulation of those assumptions, allows us to develop the logical implications of our assumptions, and clarifies exactly what needs to be justified in any particular context. That justification is itself, however, an entirely separate task, that can not rely on formal representations of any kind but must relate to the real-world context of the problem. Perhaps the most important contribution of modelling "causality" in terms of ECI is to highlight the vital need for such external justification.

13.1. What rôle for "causal discovery"?

The enterprise of "causal discovery" aims to extract causal conclusions from observationally inferred conditional independencies. However it can not do so without making (explicitly or, more often, implicitly) strong causal assumptions — which may rest unjustified, so invalidating the process. Such methods can nevertheless be useful in suggesting interesting causal conjectures for further investigation. Ideally we should then gather data from appropriate interventional studies, to investigate — and if necessary revise — the validity of conjectures, made purely on the basis of observational data, about the effects of interventions. Williamson (2005), among others, has argued for such a "hybrid hypothetico-deductive/inductive" approach.

Alternatively, when we can collect data under a variety of regimes, including interventional studies, we could directly apply variations of causal discovery techniques, to uncover genuinely causal properties. Thus, if we had data on variables (U,Z,X,Y) under all three regimes $F_X = \emptyset$, $F_X = 0$, $F_X = 1$, we could empirically test the ECI

properties (8) and (10), by (for example) simple χ^2-tests (which are equally valid for testing homogeneity of conditional distributions as they are for testing conditional independence); alternatively, Bayesian techniques could be used (Cooper and Yoo, 1999). Only with such experimental data could we hope to obtain genuine empirical evidence in favour of a causal DAG representation such as Figure 4.

Acknowledgments

My thanks to the Editors for their encouragement to prepare both the NIPS talk and this paper very loosely based on it. I am grateful to Nancy Cartwright, Vanessa Didelez, Sara Geneletti, Paul Rosenbaum, Federica Russo and Jon Williamson, the referees, and many contributors to the "Causality and Machine Learning Reading Group" http://www.afia-france.org/tiki-index.php?page=Groupe+de+lecture, for valuable feedback on an earlier draft.

References

Pierre Bourdieu. *Outline of a Theory of Practice*. Cambridge University Press, 1977.

George E. P. Box. Use and abuse of regression. *Technometrics*, 8:625–629, 1966.

Nancy Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1994.

Nancy Cartwright. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, 2007.

Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 116–125, San Francisco, CA, 1999. Morgan Kaufmann.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, New York, 2007.

Denver Dash. Restructuring dynamic causal systems in equilibrium. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005. URL http://www.gatsby.ucl.ac.uk/aistats/fullpapers/264.pdf.

A. Philip Dawid. Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979a.

A. Philip Dawid. In discussion of Lauritzen and Richardson (2002). *Journal of the Royal Statistical Society, Series B*, 64:348–351, 2002a.

A. Philip Dawid. Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, 95:407–448, 2000.

A. Philip Dawid. Conditional independence for statistical operations. *Annals of Statistics*, 8:598–617, 1980.

A. Philip Dawid. Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In Peter J. Green, Nils L. Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*, pages 45–81. Oxford University Press, 2003.

A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002b. Corrigenda, *ibid*., 437.

A. Philip Dawid. Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B*, 41:249–52, 1979b.

A. Philip Dawid. Seeing and doing: The Pearlian synthesis. In Rina Dechter, Hector Geffner, and Joseph Y. Halpern, editors, *Festschrift for Judea Pearl*. College Publications, London, 2010. To appear.

A. Philip Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32:335–372, 2001a.

A. Philip Dawid. Some variations on variation independence. In Tommi Jaakkola and Thomas S. Richardson, editors, *Artificial Intelligence and Statistics 2001*, pages 187–191, San Francisco, California, 2001b. Morgan Kaufmann Publishers.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies. Research Report 262, Department of Statistical Science, University College London, 2005. URL http://www.ucl.ac.uk/Stats/research/reports/abs05.html#262.

A. Philip Dawid, Julia Mortera, and Paola Vicard. Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, 169: 195–205, 2007.

Bruno de Finetti. *Theory of Probability (Volumes 1 and 2)*. John Wiley and Sons, New York, 1975. (Italian original Einaudi, 1970).

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation: Why epidemiology needs a formal language for causality. In Federica Russo and Jon Williamson, editors, *Causality and Probability in the Sciences*, volume 5 of *Texts In Philosophy Series*, pages 263–292. College Publications, London, 2007a.

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16: 309–330, 2007b.

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation and instrumental variables: What can and what can't be done. Technical Report 05-02, University of Leicester Department of Health Sciences, 2002.

Michael Eichler and Vanessa Didelez. On Granger-causality and the effect of interventions in time series. Research Report 09:01, Statistics Group, University of Bristol, 2009.

Morten Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353, 1990.

Maria Carla Galavotti. *Philosophical Introduction to Probability*. CSLI Publications, Stanford, 2005.

Sara G. Geneletti. *Aspects of Causal Inference in a Non-Counterfactual Framework*. PhD thesis, Department of Statistical Science, University College London, 2005.

Clark Glymour and Gregory F. Cooper, editors. *Computation, Causation and Discovery*. AAAI Press, Menlo Park, CA, 1999.

Daniel Hausman. *Causal Asymmetries*. Cambridge University Press, Cambridge, 1998.

Amanda B. Hepler, A. Philip Dawid, and Valentina Leucari. Object-oriented graphical representations of complex patterns of evidence. *Law, Probability & Risk*, 6:275–293, 2007. doi: 10.1093/lpr/mgm005.

Miguel A. Hernán and James M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17:360–372, 2006.

Dominik Janzing and Bernhard Schölkopf. Distinguishing between cause and effect via the algorithmic Markov condition. Paper presented at NIPS 2008 Workshop "Causality: Objectives and Assessment", Whistler, Canada, 2008a.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition, 2008b. URL http://arxiv.org/abs/0804.3678.

Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations (with Discussion). *Journal of the Royal Statistical Society, Series B*, 64:321–361, 2002.

Steffen L. Lauritzen, A. Philip Dawid, Birgitte N. Larsen, and Hanns-Georg Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

Edwin P. Martens, Wiebe R. Pestman, Anthonius de Boer, Svetlana V. Belitser, and Olaf H. Klungel. Instrumental variables: Applications and limitations. *Epidemiology*, 17:260–267, 2006.

Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, New Jersey, 2003.

Judea Pearl. Causal diagrams for empirical research (with Discussion). *Biometrika*, 82: 669–710, 1995.

Judea Pearl. A constraint–propagation approach to probabilistic reasoning. In Laveen N. Kanal and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 357–370, Amsterdam, 1986. North-Holland.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, second edition, 2009.

Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8:266–269, 1993.

Judea Pearl and Azaria Paz. Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *ECAI*, pages 357–363, 1986.

Hans Reichenbach. *The Direction of Time*. University of Los Angeles Press, Berkeley, 1956.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

Donald B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6:34–68, 1978.

Bertrand Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13: 1–26, 1913.

Federica Russo. *Causality and Causal Modelling in the Social Sciences: Measuring Variations*, volume 5 of *Methodos Series*. Springer, 2008.

Richard Scheines and Peter Spirtes. Causal structure search: Philosophical foundations and future problems. Paper presented at NIPS 2008 Workshop “Causality: Objectives and Assessment", Whistler, Canada, 2008.

Glenn Shafer. *The Art of Causal Conjecture*. MIT Press, Cambridge, Mass, 1996.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. Springer-Verlag, New York, Second edition, 2000.

Wolfgang Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9:73–99, 1980.

Wolfgang Spohn. Bayesian nets are all there is to causal dependence. In Maria Carla Galavotti, Patrick Suppes, and Domenico Costantini, editors, *Stochastic Dependence and Causality*, chapter 9, pages 157–172. University of Chicago Press, Chicago, 2001.

Wolfgang Spohn. *Grundlagen der Entscheidungstheorie*. PhD thesis, University of Munich, 1976. (Published: Kronberg/Ts.: Scriptor, 1978).

Milan Studený. Conditional independence relations have no finite complete characterization. In *Transactions of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, volume B, pages 377–396, Prague, 1992. Academia.

Patrick Suppes. *A Probabilistic Theory of Causality*. North Holland, Amsterdam, 1970.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In Ross D. Shachter, Tod S. Levitt, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence 4*, pages 69–76, Amsterdam, 1990. North-Holland.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In Piero P. Bonissone, Max Henrion, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 255–268. North-Holland, Amsterdam, 1991.

Jon Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press, Oxford, 2005.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, 2003.

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

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. *Journal of Machine Learning Research*, 6:157–164, 2010.

JMLR: Workshop and Conference Proceedings 6:87–96, 2010 NIPS 2008 workshop on causality

Causal Discovery as a Game

Frederick Eberhardt EBERHARDT@WUSTL.EDU
Department of Philosophy
Washington University in St. Louis
St. Louis, MO 63130, USA

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

This paper presents a game theoretic approach to causal discovery. The problem of causal discovery is framed as a game of the Scientist against Nature, in which Nature attempts to hide its secrets for as long as possible, and the Scientist makes her best effort at discovery while minimizing cost. This approach provides a very general framework for the assessment of different search procedures and a principled way of modeling the effect of choices between different experiments.

Keywords: causal discovery, interventions, search strategy, game theory, worst and expected case analysis

1. Introduction

In machine learning much of the literature on causal discovery has focused on discovery in passive observational data. The analysis of experimental data has been left to the field of experimental design, but there the focus has been on the optimal allocation of samples to a pre-determined set of treatment variables, and the subsequent analysis of the data. Very little work has been done on the selection of experiments. The specification of the best sequence of experiments to discover particular causal relations has largely been left to the "good judgment of the scientist." Only recently have first steps been taken to automate this process: Tong and Koller (2001); Murphy (2001); Yoo and Cooper (2003); Meganck et al. (2005) and He and Geng (2008) have presented approaches to select the next best experiment based on information theoretic measures or expected utility, and Eberhardt (2007) provided worst case bounds for such search strategies under different assumptions. In this paper a game theoretic analysis of sequences of experiments is proposed that identifies appropriate guidelines for the choice and comparison of different experimental strategies.

Randomized controlled trials (RCTs) are perhaps the most widely accepted standard to determine cause and effect. If, as intended by the randomization, the intervention makes the intervened variable independent of its normal causes, then it breaks any confounding of the causal effect of the intervened variable on the outcome variable by

measured or unmeasured common causes.[1] Given a set of, say, three variables X, Y and Z, a scientist has many choices of which variable(s) to randomize. She could intervene on any one and measure the other two. She could randomize any two, independently or not, and measure the third, etc. Whichever choice she makes, one experiment will in general not guarantee – even in the large sample limit – the discovery of the *true* causal structure among the 25 *possible* (directed acyclic) causal structures over the three variables. *Sequences* of different experiments are often necessary to determine all the causal relations between variables. But what is the best sequence, and in what sense of "best"?

2. Worst Case Analysis

One way to compare different search strategies is to consider their worst case performance. In Eberhardt et al. (2005) we gave worst case analyses of different search procedures for causal discovery involving different types of interventions under a variety of different assumptions. The quality of different search procedures was measured in terms of the number of experiments *sufficient and in the worst case necessary* to discover the true causal structure among N variables. The worst case was characterized by the causal structure that required the longest sequence of experiments that could not be avoided (by more appropriate choices of experiments given the available knowledge at the choice point). The following table summarizes the results for sequences of experiments with single or multiple simultaneous RCT-type interventions per experiment on a set of N causal variables, with and without latent variables:[2]

Interventions per Experiment	Latent Variables Present	Number of Experiments
Single	No	2 if $N = 2$ & $N-1$ if $N > 2$
Single	Yes	impossible
Multiple	No	$\lfloor \log_2(N) + 1 \rfloor$
Multiple	Yes	N

A worst case analysis provides an upper bound, but in practice the worst case may be very rare whereas a "typical" search problem might be resolved much faster. Consequently, the expected performance is often considered. The computation of an expectation depends on a distribution over the possible hypotheses. In the case of three variables, it would require a distribution over the 25 possible (acyclic) causal structures. In many cases, the uniform distribution is used, but without more specific knowledge of the domain under consideration, it is not clear why the uniform distribution is more appropriate than any other. Often sparsity assumptions play a crucial role in restricting the hypothesis space and it is not clear that a uniform distribution over hypotheses is

1. With regard to causal discovery, weaker forms of interventions can also provide insights but we will leave that issue aside here.
2. The second row indicates that no sequence of experiments, in which only a single variable is subject to an RCT-type intervention, is sufficient to discover the causal structure *in the worst case* if there are latent variables. Under different assumptions, such as linearity, discovery is possible (see Eberhardt (2007)).

"uninformative" when a sequence of experiments is used for discovery. What, then, can be said about an expected case performance without commitment to a particular distribution?

3. Expectation and Optimization

One approach supported by a game-theoretic interpretation of the discovery problem is the *worst case expected* performance, i.e. the upper bound on the expected length of sequences of experiments sufficient and in the worst case necessary to discover the causal structure, *no matter what the probability distribution over the set of directed acyclic graphs is*. That is, for each distribution $P(\mathcal{G})$ over the set $\mathcal{G}$ of directed acyclic graphs, take the expectation $E_P(.)$ of the number of experiments $\#ex(.)$ sufficient and in the worst case necessary to uniquely discover the true causal graph G, whatever G is. Then take the upper bound – the supremum – of those expectations. Or formally:

$$\sup_{E} E_P(\#ex(G)) \quad \text{over all} \quad P(\mathcal{G}). \tag{1}$$

The key to determining this quantity is the specification of $\#ex(G)$ for some true underlying causal structure G. To specify this quantity we need to specify how experiments are chosen. But how and which experiments are chosen affects which causal structures are difficult to learn, so the supremum is affected by both the underlying distribution over causal structures and the sequence of experiments that is used to identify which one is true.[3]

Let $\mathcal{S}$ be a strategy that specifies a sequence of experiments, in which the next experiment is determined with probability 1 contingent on the evidence revealed in all previous experiments. Given a set $\mathcal{G}' \subseteq \mathcal{G}$ of possible causal structures (determined by non-zero probability in the probability distribution $P(.)$ over causal structures), let $\#ex_{\mathcal{S}}(G)$ be the number of experiments according to strategy $\mathcal{S}$ that is necessary and sufficient to uniquely identify a particular causal structure $G \in \mathcal{G}'$. Since we are interested in an optimal number of experiments, we can now define $\#ex(G)$ as the number of experiments necessary and sufficient to uniquely identify $G \in \mathcal{G}'$ using a strategy $\mathcal{S}^+$ where

$$\forall \mathcal{S} \neq \mathcal{S}^+ \qquad E_P(\#ex_{\mathcal{S}}(G)) \geq E_P(\#ex_{\mathcal{S}^+}(G)). \tag{2}$$

That is, for a given set of possible causal structures $\mathcal{G}'$, $\#ex(G)$ specifies the expected number of experiments necessary and sufficient to uniquely identify the causal structure $G \in \mathcal{G}'$ using a most efficient strategy $\mathcal{S}^+$. However, since we are interested in the

3. For example: If one always intervenes on X first, then causal structures in which X is an effect (but not a cause!) of the other variables, are more difficult to discover because any incoming causal influence on X is destroyed by the intervention, and so the structure cannot be distinguished from one in which X is causally independent of the other variables. Consequently, a distribution that puts more weight on those graphs will be a candidate for the maximum expectation. But such a distribution results in a much lower expectation if the first intervention always intervenes on one of the causes of X (say Y), since the $Y \rightarrow X$ edge is discovered immediately.

supremum of the expectations, the distribution $P(.)$ that specifies the set of possible graphs, must be such that it is implies the largest expected number of experiments for a given strategy, i.e. given a strategy S, $P^+(.)$ is chosen such that for any

$$P(.) \neq P^+(.) \qquad E_{P^+}(\#ex_{\mathcal{S}}(G)) \geq E_P(\#ex_{\mathcal{S}}(G)). \tag{3}$$

Definitions (2) & (3) make the interdependence between a search strategy and the distribution over hypotheses explicit: Given a hypothesis space one can specify the optimal search strategy. Given a search strategy one can specify the hypothesis space that will make search most difficult. We can thus rephrase the supremum in (1) above as the following optimization:

$$E_{P^*}(\#ex_{\mathcal{S}^*}(G)) \tag{4}$$

where it is simultaneously the case that for any

$$P(.) \neq P^*(.) \qquad E_{P^*}(\#ex_{\mathcal{S}^*}(G)) \geq E_P(\#ex_{\mathcal{S}^*}(G)) \tag{5}$$

and given the set of possible causal structures implied by $P^*(.)$,

$$\forall \mathcal{S} \neq \mathcal{S}^* \qquad E_{P^*}(\#ex_{\mathcal{S}}(G)) \geq E_{P^*}(\#ex_{\mathcal{S}^*}(G)). \tag{6}$$

For fixed sequences of experiments that specify a particular experiment for a given history of evidence there is no solution to the above double optimization. That is, for any specific strategy $\mathcal{S}$ there is a probability distribution $P(.)$ that maximizes the number of experiments with respect to $\mathcal{S}$ (in fact, the expectation can always be forced to the absolute worst case bound). However, an alternative strategy $\mathcal{S}'$ would do better on $P(.)$, but then there is another probability distribution $P'(.)$ that would trouble strategy $\mathcal{S}'$.

The main problem is that knowledge of the proposed sequence of experiments permits a choice of distribution over hypotheses that is specifically geared towards making discovery hard for that sequence of experiments. A natural solution to this problem is to consider search strategies that do not commit to a particular experiment in light of a particular history of evidence, but rather to a distribution over possible experiments, i.e. a mixture of search strategies. This suggests a game-theoretic analysis.

4. Discovery as a Game

We can recast the above analysis of search strategies as a two person zero-sum game between Nature and the Scientist. The Scientist attempts to discover the true causal structure as efficiently as possible and Nature tries to make discovery as difficult as possible – in our case (for now) in terms of the number of experiments.

Nature initially gets to decide what the truth is – the underlying causal structure – but then has to stick with it, while the Scientist performs her experiments. Nature's pure strategies are all the directed acyclic causal structures over N variables. After each experiment by the Scientist the independence relations true in the underlying causal

structure (manipulated by the intervention) are returned, i.e. the equivalence class of directed acyclic graphs that contains the true graph and is consistent with the sequence of experiments so far, is revealed. We refer to this – as is standard in game theory – as an information set. The pure strategies for the Scientist are all possible sequences of experiments. The Scientist may end the game after any sequence of experiments by declaring one of the graphs remaining in her information set as true. If the Scientist is correct, the payoff is the negative number of experiments that were performed (negative, since the Scientist wants to perform as few experiments as possible). If the Scientist is incorrect, the payoff is $-\infty$. Payoffs of $-\infty$ ensure that in order to avoid infinite loss the Scientist must be able to prove that her response is uniquely correct given the evidence.[4]

In game theory a strategy that specifies for each choice point a determinate choice (of experiment) corresponds to a *pure strategy*. A *mixed strategy* permits non-trivial distributions over the choices of experiments. Sometimes a mixed strategy can outperform any pure strategy. In our context the case for mixed strategies for Nature (i.e. distributions over graphs) is obvious – it would not be an interesting search problem if Nature were restricted to selecting one particular causal structure with probability 1. In the case of the Scientist we consider mixed strategies for two reasons. First, we already indicated at the end of Section 3 that there is no solution to the optimization problem when the Scientist is restricted to pure strategies. Second, and perhaps more intuitively, there are many circumstances in which a restriction to a specific experiment in light of the available evidence is artificial. For example, suppose there are two variables X and Y and it is known that either $X \to Y$ or $Y \to X$, each with probability 0.5. In that case a commitment to always intervene on X is artificial. Flipping a fair coin to either intervene in X or Y seems more appropriate.

For simplicity of exposition (and computation), we assume that every variable can be manipulated, that there are no latent variables and we only consider sequences of experiments in which one (or no) variable is subject to an intervention per experiment.[5] Given that the worst case bound on the number of experiments under these circumstances is $N-1$ for $N > 2$ variables, we do not need to consider search strategies for the Scientist that are longer than $N-1$ experiments, i.e. the table of results in Section 2 gives upper bounds on the worst case loss for the Scientist. Consider a simple example.

4.1. Example: Two Variables

Suppose there are just two variables. There are three possible causal structures among two variables X and Y, call them

$$\mathbf{Sa} := X \quad Y, \quad \mathbf{Sb} := X \to Y \quad \text{and} \quad \mathbf{Sc} := X \leftarrow Y.$$

Two experiments involving single interventions are sufficient and in the worst case necessary to discover the causal structure uniquely. The full game of Nature against

4. Of course, one could integrate into the payoff structure some account of *how* wrong a Scientist is, but we leave this for future consideration.
5. See Section 6 and Eberhardt (2007) for more on multiple simultaneous interventions.

Scientist is given in Figure 1. Nature can select among the three structures (grey boxes) **Sa**, **Sb** and **Sc**. The Scientist does not know which structure is selected, so **Sa**, **Sb** and **Sc** form an information set. The Scientist makes the next move and can end the game by guessing one of the structures without collecting any data (represented by the three arrows leaving each grey box upwards with Sa, Sb or Sc and the respective payoffs to Nature of 0 when the choice was correct and ∞ when incorrect). Alternatively, the Scientist can perform a passive observation (**N**), an intervention on the first variable (**X**), or an intervention on the second variable (**Y**). Depending on the choice and the true underlying graph, the game is either resolved because the graph can be uniquely identified (payoffs are indicated), or one of three new information sets – represented in the figure as a box containing the two causal structures that cannot be distinguished given the experiments so far – is returned. Again, the Scientist can end the game at this

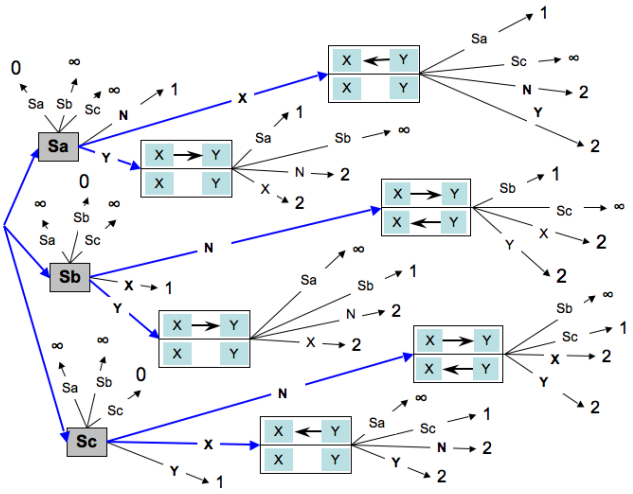


Figure 1: *Discovery of Causal Structure as a game of Nature against the Scientist, here for two causally sufficient variables.*

point with a guess, or can continue with a further experiment. Guesses and experiments that do not make sense in light of the evidence obtained so far, are not included in the game, since the Scientist is assumed to be rational. Given the worst case bound of two, there is no need to consider strategies of more than two experiments.

The game now permits an analysis of the optimal (mixed) search strategy against the most difficult probability distribution over causal structures. Since the game was constructed as a zero-sum game, the Nash equilibrium of the game corresponds to the mini-max solution, i.e. the Nash equilibrium specifies the desired upper bound on the expectation of the number of experiments sufficient and in the worst case necessary to discover the causal structure. The strategies implied by the Nash equilibrium for

Nature and the Scientist, respectively, characterize a state, in which a unilateral change in strategy by Nature or by the Scientist does not improve their individual score.

An analysis of the game shows that the Nash equilibrium is given by a mixed strategy that is uniform over the three possible structures **Sa**, **Sb** and **Sc** for Nature, and a mixed strategy for the Scientist that is uniform over passive observation, an intervention on X and an intervention on Y for the first experiment, and indifferent between possible (relevant) experiments for the second experiment, if a second experiment is necessary. That is, if Nature "selects" the true causal structure among the two variables uniformly, then Nature is making the discovery task maximally difficult for the Scientist. On the other side, by choosing uniformly whether to intervene on X, intervene on Y or just passively observe in the first experiment, the Scientist is doing the best she can to discover Nature's secrets efficiently, given that Nature is an adversarial player. Any other strategy, even mixed, will do no better and may well be worse (or will allow Nature to adapt accordingly to make things worse).

One could take this result to be a justification for the consideration of the uniform distribution over the hypothesis space in the assessment of an expected case performance for algorithms, but we will show below that this argument does not extend beyond the case of two variables.

The value of the Nash equilibrium represents the expected payoff to Nature (and loss to the Scientist) when playing the mixed strategy that is Nash. For this two variable game it is $5/3$ experiments, so the worst case expected performance is slightly better than the absolute worst case bound of 2 experiments. The Scientist's strategy is in this case not an equalizer, since some graphs are resolved in one experiment and others in two. The Nash equilibrium is, if we ignore the indifference for the second experiment, unique. As already indicated in the discussion of the optimization in the previous section, there is no Nash-equilibrium over pure strategies for either side, i.e. there is no Nash equilibrium if Nature selects one particular causal structure with probability 1, and there is no Nash equilibrium if the Scientist picks a experiments with probability 1. In both cases the opponent can adjust to do better. Further, returning with a guess of the true causal structure at any point is (obviously, given the infinities in the payoff structure) not Nash, so the solution that the Scientist returns is guaranteed to be justifiable given the evidence. Guessing (ending the game early) only becomes a viable option, when Nature is restricted to playing a subset of the possible structures.

The mixed strategy for the Scientist that is Nash is a Bayes solution, since it is a best response to the uniform distribution over structures. No two-experiment strategy (using single interventions) is a best response to any pure or mixed strategy by Nature. Interestingly, this last point does not apply in the case of three variable graphs. In the case of three variables, the game is substantially more complicated. There are 25 pure strategies for Nature (all DAGs over three variables) and 67 pure strategies for the Scientist (including all the early stops by guessing). We computed a Nash equilibrium, which determined 2 as the solution for the game: The worst case expected number of experiments sufficient and in the worst case necessary to determine the causal graph over three variables is two. That is, in the case of three variables, Nature can force

the Scientist to the absolute worst case bound $(N-1=2)$ even *in expectation*. To do so, Nature must select the true causal structure using a uniform distribution over the following set of 10 different graphs over three variables: the empty graph, three graphs consisting of a common effect only and the six possible complete graphs. Due to the edge-breaking nature of RCT-type interventions, at least two graphs of the 10 remain indistinguishable after any single intervention experiment or passive observation; hence a second experiment is necessary.

We know from the table in Section 2 that two experiments are sufficient for three variables. Consequently, the uniform distribution over the 10 graphs implies that any sequence of two different experiments is a best response, and obviously an equalizer (same payoff of two experiments, no matter which graph is true). No pure or mixed strategy will fare any better against the above distribution, which is not to say that there are no mixed strategies that do equally well.[6]

5. General Results

For single interventions per experiment, the three variable game is unique: For no other number of variables can Nature force the Scientist to the worst case bound in expectation. The general result for mixed strategies using single interventions per experiment is given by the following theorem.

Theorem *Given a set of $N > 3$ causally sufficient variables, the supremum of the expected number of experiments sufficient and in the worst case necessary to discover the causal structure is $\frac{2}{3}N - \frac{1}{3}$ experiments if only one (or no) variable can be subject to a RCT-type intervention per experiment.* ■

This bound is the value of a Nash equilibrium of the game: Nature plays a mixed strategy that is uniform over the *complete*(!) graphs over N variables only. For any N there are $N!$ such structures. From Nature's perspective, there is no advantage in considering incomplete causal structures, since for $N > 3$ variables, two single interventions have to be performed anyway, and in those two experiments any missing edge would be detected. This implies that a uniform distribution over all possible hypotheses (graphs) is not Nash for Nature, and an analysis based on such a distribution would underestimate the worst case expectation. For the Scientist the following strategy is Nash:

Strategy *Given N causally sufficient variables $X_1, \ldots, X_n$, let each experiment $\mathcal{E}_i$ in the sequence intervene on $\mathbf{I}_i = \{X_j\}$, where X_j is selected uniformly from the variables that have not yet been subject to an intervention so far in the sequence.*

Since the game is symmetric with regard to the ordering of the variables (any variable can occur in any position in the graph), there are no order constraints on the Scientist's strategy. Of course, there may exist for some circumstances a particular order

6. For example, if the passive observation is included as a possible first experiment, then if Nature chooses uniformly, it is a best response, but not an equalizer: $2/5$ of the time it finds the graph in one experiment and $3/5$ of the time it requires $8/3$ experiments (i.e. two experiments on average overall).

of experiments that minimizes the length of the sequence; but the Scientist cannot tell in advance.

For multiple simultaneous interventions per experiment the case is far more complicated. A discussion can be found in Eberhardt (2007); simulations suggest that the absolute worst case bound for multiple simultaneous interventions ($\lfloor \log_2(N) \rfloor + 1$ experiments) is fairly close to the worst case expectation, which would imply that the computationally simple pure strategies for the absolute worst case are fairly efficient even compared to the optimal mixed strategy, which is very hard to compute.

6. Conclusion

We framed the search for causal structure as a game in which Nature gets the first move to determine the graph after which the Scientist has free reign. This follows the approach developed for statistical hypothesis testing by Wald (1950), and generalizes it to sequences of experiments. Needless to say, this is only a first step presented with a very simple example. But the possibilities for generalization should now be obvious: (i) The effect of additional assumptions on the search procedure can be represented in terms of additional or reduced underdetermination in the information sets at any decision point. (ii) Cost other than the number of experiments can be considered. One may consider cost functions in terms of sample size, number of variables subject to intervention, or actual cost of experimentation – ethical or monetary. These cost functions need not be uniform across variables. (iii) Constraints or background knowledge on possible causal structures can be represented by limiting the possible pure strategies for Nature, while constraints on the set of experiments – e.g. it might not be possible to subject all variables to an experiment – limit the pure strategies for the Scientist. (iv) The robustness of search strategies can be analyzed in terms of changes in the optimal strategy with regard to off-equilibrium play by Nature – after all, Nature need not be adversarial; and the sensitivity of the optimal search strategy can be investigated by considering off-equilibrium play by the Scientist.

The game-theoretic approach to the discovery problem provides a general framework in which search strategies can be analyzed for their efficiency using a well-defined terminology and highly developed machinery. General guidelines for search procedures can be discovered and assessed on the basis of the explicit trade-off between discovery and its cost. Addressing these issues in the appropriate generality will require the integration of some of the most sophisticated game-theoretic techniques.

Acknowledgments

This research was supported by a fellowship from the James S. McDonnell Foundation. I am very grateful to Teddy Seidenfeld for making me think about this issue in the first place. Four anonymous reviewers (and one in particular, who was asked to review again) provided very useful comments on earlier drafts.

Appendix: Proofs[7]

Lemma 1 *For $N \geq 4$ the supremum of the expected number of experiments sufficient and in the worst case necessary to uniquely determine the causal graph is greater than 2 if only single interventions are permitted per experiment.*

Lemma 2 *The uniform distribution over complete graphs of N variables maximizes the expected number of experiments sufficient and in the worst case necessary to discover the true graph uniquely when only single intervention experiments are permitted.*

Theorem *Given a set of $N > 3$ causally sufficient variables, the supremum of the expected number of experiments sufficient and in the worst case necessary to discover the causal structure is $\frac{2}{3}N - \frac{1}{3}$ experiments if only one (or no) variable can be subject to a RCT-type intervention per experiment.* ■

Proof By Lemma 2, the uniform distribution over complete graphs is a worst case distribution. Suppose without loss of generality that the true complete graph over the variables $X_1, \ldots, X_N$ is such that for all $i < j$, $X_i \to X_j$. Under these circumstances an intervention on X_i is (1) uninformative with respect to edge-orientation about all pairs of variables X_j, X_k with $j,k < i$; (2) uninformative with respect to edge-orientation about all pairs of variables X_j, X_k with $j,k > i$; and (3) informative for the remaining edges: It resolves (i) edges between variables X_j, X_k with $j > i > k$, (ii) outgoing edges from X_i and, (iii) since it is known that the graph is complete, edges broken by the intervention can be identified, and so all edges incident on X_i are resolved. In other words, an intervention on X_i splits the discovery problem into two subproblems, one with $N - i$ variables and the other with $i - 1$ variables. About these subproblems, the intervention on X_i is uninformative.

Given the uniform distribution over complete graphs, the problem is entirely symmetric in the sense that each node is equally likely to be at any of the possible positions in a complete graph. Similarly, a uniform distribution selecting among the unintervened variables, implies that each variable is equally likely to be subject to an intervention in the first experiment. Consequently, we can give the expected number of experiments for this worst case distribution in terms of the numbers required for the subproblems the intervention creates:

$$E(\#\mathcal{E}(N)) = \frac{1}{N}\sum_{i=1}^{N}(E(\#\mathcal{E}(i-1)) + E(\#\mathcal{E}(N-i)) + 1) = 1 + \frac{2}{N}\sum_{i=1}^{N}E(\#\mathcal{E}(i-1))$$

where $E(\#\mathcal{E}(N))$ is the expected number of experiments required to discover the true graph if the graph is sampled from a Uniform over complete graphs of N variables. So the expected number of experiments for N variables is one plus the average of the sum of the number of experiments that it takes to resolve the two subproblems of size $N - i$ and $i - 1$, respectively. For complete graphs with two and three variables, one can check by hand that $E(\#\mathcal{E}(2)) = 1$ and $E(\#\mathcal{E}(3)) = 5/3$. So finally we prove by induction that

$$E(\#\mathcal{E}(N)) = \frac{2}{3}N - \frac{1}{3} \qquad \text{for } N \geq 2.$$

7. For more detailed proofs see Eberhardt (2007).

It is true for $N = 2$. Suppose it is true for all integers up to some $N-1$. Then

$$E(\#\mathcal{E}(N)) = 1 + \frac{2}{N}\sum_{i=1}^{N} E(\#\mathcal{E}(i-1)) = 1 + \frac{2}{N}\sum_{i=1}^{N}(\frac{2}{3}(i-1) - \frac{1}{3}) = -\frac{1}{3} + \frac{2}{3}N$$

■

References

F. Eberhardt. *Causation and Intervention*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 2007.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the 21st Conference on Uncertainty and Artificial Intelligence*, pages 178–184, 2005.

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

S. Meganck, B. Manderick, and P. Leray. A decision theoretic approach to learning Bayesian networks. Technical report, Vrije Universiteit Brussels, 2005.

K. P. Murphy. Active learning of causal Bayes net structure. Technical report, Department of Computer Science, U.C. Berkeley, 2001.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 863–869. Morgan Kaufmann, 2001.

A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

C. Yoo and G. Cooper. A computer-based microarray experiment design-system for gene-regulation pathway discovery. In *AMIA 2003 Symposium Proceedings*, pages 733–737, 2003.

JMLR: Workshop and Conference Proceedings 6:97–106, 2010 NIPS 2008 workshop on causality

Sparse Causal Discovery in Multivariate Time Series

Stefan Haufe HAUFE@CS.TU-BERLIN.DE
Klaus-Robert Müller KRM@CS.TU-BERLIN.DE
Machine Learning Group, TU Berlin
Franklinstr. 28/29, 10587 Berlin, Germany

Guido Nolte GUIDO.NOLTE@FIRST.FRAUNHOFER.DE
Intelligent Data Analysis Group, Fraunhofer FIRST
Kekuléstr. 7, 12489 Berlin, Germany

Nicole Krämer NKRAEMER@CS.TU-BERLIN.DE
Machine Learning Group, TU Berlin
Franklinstr. 28/29, 10587 Berlin, Germany

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Our goal is to estimate causal interactions in multivariate time series. Using vector autoregressive (VAR) models, these can be defined based on non-vanishing coefficients belonging to respective time-lagged instances. As in most cases a parsimonious causality structure is assumed, a promising approach to causal discovery consists in fitting VAR models with an additional sparsity-promoting regularization. Along this line we here propose that sparsity should be enforced for the subgroups of coefficients that belong to each pair of time series, as the absence of a causal relation requires the coefficients for all time-lags to become jointly zero. Such behavior can be achieved by means of $\ell_{1,2}$-norm regularized regression, for which an efficient active set solver has been proposed recently. Our method is shown to outperform standard methods in recovering simulated causality graphs. The results are on par with a second novel approach which uses multiple statistical testing.

Keywords: Vector Autoregressive Model, Granger Causality, Group Lasso, Multiple Testing

1. Introduction

Causality is commonly defined based on the widely accepted assumption that an effect is always preceded by its cause. Granger (1969) postulates a measure of causal influence between two time series (*Granger Causality*). In a nutshell, a time series z_i Granger-causes time series z_j if knowledge of past values of z_i improves the prediction of z_j (compared to only using past values of z_j). The improvement is assessed by means of the *Granger score*, which is defined as the logarithm of the ratio of the residuals of the two models (1) including only z_j and (2) including both z_i and z_j.

In the case of a set $F = \{z_1, \ldots, z_M\}$ of time series, the pairwise analysis may lead to spurious detection of a causal relation. For this reason it is advisable to additionally

include the set $F \setminus \{z_i, z_j\}$ of all other observable time series in both models. This approach, to which we refer as *complete* (or conditional) Granger Causality, resolves the problem of spurious causality due to common hidden factors z_* if $z_* \in F$. If the z_* are not observable, Granger causality fails and we refer to Nolte et al. (2008) for a detailed discussion and a remedy.

Just to illustrate the problem, consider that a hidden driving factor is equally pronounced in two variables $z_{i'}$ and $z_{i''}$. If both variables contain roughly the same amount of noise, all of the sets F, $F \setminus \{z_i\prime\}$ and $F \setminus \{z_i\prime\prime\}$ provide equal information about z_j, for which reason complete Granger causality will neither identify $z_i\prime$ nor $z_i\prime\prime$ as a driver. This type of mistake can only be avoided if each set $F \setminus \{z_i\prime\}$ is tested against all sets not including $z_i\prime$, which leads to exponential complexity.

An elegant alternative to the pairwise comparisons of (complete) Granger causality is to handle all potential causal relations between all time series at once. Assuming a linear dynamics of the system under study, this leads us to the vector autoregressive (VAR) model. Interestingly, the parameters of the VAR model induce a natural alternative definition of causal influence, which is compliant with Granger's considerations.

In many applications the true causality graph is assumed to be sparse, i.e. only a few causal interactions between time series are expected. Ordinary Least Squares (OLS) and Ridge Regression, which are usually used for fitting VAR models, however, are known for producing dense coefficients. Only recently Valdes-Sosa et al. (2005) have proposed to enforce estimation of sparse AR coefficients using ℓ_1-norm regularized models such as the Lasso (Tibshirani, 1996).

In this paper we propose a novel sparse approach which – unlike Lasso – accounts for the fact that the absence of a causal relation between z_i and z_j requires all AR coefficients belonging to that certain pair of time series to be jointly zero. Furthermore, we consider Ridge Regression in combination with the multiple statistical testing procedure provided by Hothorn et al. (2008). More details on the methodology are given in section 3. These methods are evaluated and compared to standard approaches in extensive simulations.

2. Background

In this section, we briefly summarize related approaches to estimate sparse vector autoregressive models in the context of causal discovery. We roughly distinguish between sparse estimation methods and testing strategies.

Given a multivariate time series $\mathbf{z}(t) \in \mathbb{R}^M$ a linear vector autoregressive process of order P is defined as

$$\mathbf{z}(t) \quad = \quad \sum_{p=1}^{P} A^{(p)} \mathbf{z}(t-p) + \varepsilon(t)\,, \tag{1}$$

where $A^{(p)} \in \mathbb{R}^{M \times M}$, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ and $t \in \mathbb{Z}$ indicates time. Hence, the signal at time t is modeled as a linear combination of its P past values and Gaussian measurement

noise. Inspired by the initial assumption that the cause should always precede the effect, we suggest the following definition of causality. We say that time series z_i has a causal influence on time series z_j if for at least one $p \in \{1,\dots,P\}$, the coefficient $A_{ji}^{(p)}$ corresponding to the interaction between z_j and z_i at the pth time-lag is nonzero.

Thus, causal inference may be conducted by estimating the matrices $A^{(p)}$ from a sample $Z = (\mathbf{z}(1),\dots,\mathbf{z}(T))$. Let us introduce the following shortcuts. We denote by $A = \left(A^{(1)},\dots,A^{(P)}\right)^\top$ the matrix of all VAR coefficients and set $X = (Z_1,\dots,Z_P)$, $Y = Z_0$, $Z_p = (\mathbf{z}(P+1-p),\dots,\mathbf{z}(T-p))^\top$. Here $\mathrm{vec}(\cdot)$ denotes the vectorization operation.

2.1. Sparsity

Probably the most straightforward way to estimate a sparse VAR is to use ℓ_1-regularization on the set of coefficients,

$$\widehat{A}^{\text{lasso}} = \arg\min_A \|\mathrm{vec}(XA - Y)\|_2^2 + \lambda \|\mathrm{vec}(A)\|_1 \ , \ \lambda \geq 0 \, .$$

Recently, Valdes-Sosa et al. (2005) proposed a combination of VAR-estimation and the Lasso (Tibshirani, 1996). While Valdes-Sosa et al. (2005) only consider a VAR model of order 1, there have been extensions to higher orders (e.g. Arnold et al., 2007). However, we note in the latter case, Lasso is not used on the VAR coefficients directly, but that the problem is transformed into the task of estimating partial correlation coefficients between time-lagged copies of the time series (see also Opgen-Rhein and Strimmer, 2007).

2.2. Testing

Just as in the case of sparse methods, it is often suggested to transform the regression task into the estimation of the matrix of partial correlation coefficients between time-lagged copies of the time series. While Drton and Perlman (2008) estimate the correlation matrix in an unregularized way, Opgen-Rhein and Strimmer (2007) propose a shrinkage estimator, which is superior in the case of high-dimensional data (Schäfer and Strimmer, 2005). Afterwards, significant partial correlations are detected by controlling false discovery rates. While the latter approach is only tested for $P = 1$, it is straightforward to extend it to higher order VAR's.

3. Our Approach

In the following, we provide the details regarding the groupwise sparsity and the alternative testing strategy respectively.

3.1. Ridge Regression and Multiple Testing

Under the assumption of Gaussian white noise it is natural to estimate the AR coefficients using regularized least squares, and probably the most straightforward way to do

so is to use Ridge Regression,

$$\widehat{A}^{\text{ridge}} = \arg\min_A \|\text{vec}(XA-Y)\|_2^2 + \lambda \|\text{vec}(A)\|_2^2 = (X^\top X + \lambda I)^{-1} X^\top Y \,, \lambda \geq 0 \,. \quad (2)$$

Thanks to the Ridge penalty, Eq. 2 delivers solutions with small coefficients, which, however, are in general never exactly zero. In the strict sense of Granger, this corresponds to a fully-connected dependency graph, rendering Ridge Regression an improper candidate for sparse causal recovery. On the other side, many of the estimated coefficients are expected to be non-significant. Hence, we propose a sparsification by means of statistical testing, where our approach is, in contrast to e.g. bootstrapping, to explicitly derive p-values.

From Eq. 2 it is apparent that the estimation can be done independently for each column of A, and so does the testing. Let therefore $\boldsymbol{\alpha}_k$ denote the kth column of A and let $\mathbf{y}_k = (z_k(P+1), \ldots, z_k(T))^\top$. Neglecting the dependency of X and Y , the Ridge coefficients depend linearly on Y, we can conclude that under the null-hypothesis $H_0 : \boldsymbol{\alpha}_k = 0$, we have $\widehat{\boldsymbol{\alpha}}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \Sigma)$ with $\Sigma = \left(X^\top X + \lambda I\right)^{-1} X^\top X \left(X^\top X + \lambda I\right)^{-1}$. Furthermore, setting $H = X\left(X^\top X + \lambda I\right)^{-1} X^\top$ an estimate of the model variance σ_k^2 is given by

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{y}_k - H\mathbf{y}_k\|^2}{\text{trace}\left((I-H)(I-H^\top)\right)} \,. \quad (3)$$

Using Eq. 3 we can now construct normalized test statistics $\widetilde{\alpha}_{ik} = \widehat{\alpha}_{ik} / \sqrt{\sigma_k^2 \Sigma_{ii}}$ which are jointly normally distributed with $\widetilde{\boldsymbol{\alpha}} \sim \mathcal{N}(\mathbf{0}, R)$ and $R_{ij} := \Sigma_{ij} / \sqrt{\Sigma_{ii}\Sigma_{jj}}$. Suppose we want to test all individual hypotheses $H_{0,i} : \alpha_{ik} = 0$ simultaneously, then, according to Hothorn et al. (2008), the adjusted p-values are $p_i = 1 - g\left(R, |\widetilde{\alpha}_{ik}|\right)$. We reject a hypothesis, if the p-value is below the predefined significance level γ. Here,

$$g(R,t) = P\left(\max_i |\widetilde{\alpha}_{ik}| \leq t\right) = \int_{-t}^{t} \ldots \int_{-t}^{t} \phi(\alpha_1, \ldots, \alpha_{MP}) \mathrm{d}\alpha_1 \cdots \mathrm{d}\alpha_{MP} \quad (4)$$

and $\phi(\boldsymbol{\alpha})$ is the density function of the multivariate normal distribution $\mathcal{N}(\mathbf{0}, R)$.

3.2. Group Lasso

Sparse causal discovery using Ridge Regression is a two-step procedure and may possibly suffer from the aggregation of assumptions that enter in each step. Direct estimation of sparse VAR coefficients (e.g. via Lasso) is therefore desirable, as this would allow omission of the multiple significance testing step. However, for higher order models, this approach is prone to selecting a different set of causal interactions for each of the P time lags. We here suggest that this behavior can be overcome by enforcing *joint sparsity* of the coefficient vectors that belong to a certain pair of time series. This corresponds to incorporating the prior belief that causal influences between time series are not restricted to only one particular time lag into the estimation. The positive effect of such modeling can be verified in Figure 1 (see Section 4 for more details).

The idea of imposing groupwise sparse coefficients leads to $\ell_{1,2}$-norm regularized regression also known as the *Group Lasso* (Yuan and Lin, 2006), which has also applications in Multiple Kernel Learning (Bach et al., 2004; Sonnenburg et al., 2006) and the EEG/MEG inverse problem (e.g. Haufe et al., 2008). The term $\ell_{1,2}$-norm stands here for an ℓ_1-norm of a vector of ℓ_2-norms. Our proposed objective is given by

$$\widehat{A}^{\text{glasso}} = \arg\min_{A} \|\text{vec}(XA - Y)\|_2^2 \tag{5}$$

$$\text{s.t.} \quad \left\| \left(A_{11}^{(1)}, \ldots, A_{MM}^{(P)} \right) \right\|_2 + \sum_{i \neq j} \left\| \left(A_{ij}^{(1)}, \ldots, A_{ij}^{(P)} \right) \right\|_2 \leq \kappa \,, \tag{6}$$

This penalty leads to a groupwise variable selection, i.e. a whole block of coefficients is jointly zero. Note that the first term in Eq. 6 penalizes all MP coefficients describing univariate relations. In this way, those coefficients are shrunk and hence, overfitting is avoided. Furthermore, we remark that it is also conceivable to to split the the whole estimation of A into M subproblems (as suggested in Subsection 3.1), which is desirable in large-scale scenarios.

Eqs. 5 and 6 define a non-differentiable but convex optimization problem which can be solved in polynomial time by means of Second-order Cone Programming (SOCP). For problems with sparse expected structure, however, the optimization can be carried out much more efficiently using the results of Roth and Fischer (2008). By keeping a set of active coefficient groups, their algorithm needs to call the SOCP solver only for problem sizes far smaller than the original problem – leading to a considerable reduction of memory usage and computation time. In the experiments, we employ the active-set algorithm of Roth and Fischer (2008) in combination with a freely available SOCP solver (Sturm, 1999).

4. Simulations

We conduct a series of experiments in which the causal structure of simulated data has to be recovered. We include the proposed groupwise sparse approach, standard Lasso, Ridge Regression with multiple testing and complete Granger Causality based on AR models in the comparison. All four approaches are applied both with and without knowledge of the true model order. In the latter case $P = 10$ is chosen for the reconstruction. For all methods considered, it is also possible to estimate the model order P, e.g., via cross-validation.

4.1. Setup

Each simulated data set consists of a multivariate time series with parameters $M = 7$ and $T = 1000$ that is generated by a random VAR process of order $P = 5$ according to 1. The distribution of the noise component $\varepsilon(t)$ is chosen to be the standard normal distribution. The VAR coefficients for all but 10 randomly chosen pairs of time series are set to zero, yielding exactly 10 causal interactions. The non-zero coefficients are

drawn randomly from $\mathcal{N}(0, 0.04I)$. Each set of VAR coefficients is tested for the stability of its induced dynamical system by looking at the eigenvalues of the corresponding transition matrix. Only coefficients leading to stable systems (i.e those with transition matrices with eigenvalues of at most 1) are accepted. We consider the following three types of problems, for each of which we created 10 instances: 1) no noise is added to the data generated by the VAR model 2) the data is superimposed by Gaussian noise of approximately the same strength, which is uncorrelated (white) both across time and sensors 3) the data is superimposed by mixed noise of approximately the same strength, which is generated as a random instantaneous mixture of M univariate AR processes of order 20. Note that in none of these cases the noise itself possesses a causal structure which would superimpose the true structure.

For measuring performance we consider Receiver Operating Characteristics (ROC) curves, which allow objective assessment of the performance in different regimes (e.g. very few false positives). As an additional measure of absolute performance we also calculate the Area Under Curve (AUC). ROC curves and AUC values are averaged across the 10 problem instances and standard errors are computed for AUC.

Complete Granger Causality is calculated using the Levinson-Wiggens-Robinson algorithm for fitting AR models (Marple, 1987), which is available in the open Biosig toolbox (Schlögl, 2003). For each pair of variables, the Granger score is calculated. The Granger score is standardized by dividing it by it's standard deviation as estimated by the jackknife. To obtain a ROC-curve, the standardized scores are threshold at different values, ranging from completely sparse to completely dense solutions.

The regularization parameter of Ridge Regression λ is chosen via 10-fold cross-validation (with respect to time-series prediction accuracy). For this value of λ, we derive the test statistics defined in Subsection 3.1. The multidimensional integrals in Eq. 4 are computed using Monte Carlo sampling according to Genz (1992). ROC-curves are constructed by varying the significance level γ.

For Lasso and Group Lasso, solutions ranging from completely sparse to completely dense are obtained through variation of the regularizing constant λ and κ respectively.

4.2. Results and Discussion

First, we illustrate the different behavior of the investigated methods in Figure 1. This example corresponds to the situation without noise and with known model order $P = 5$. The leftmost part of the Figure shows the true underlying causal structure. In the top we show the strength of the generating AR coefficients belonging to each pair of variables. Following Granger, this defines the binary causal influence matrix in the bottom, where black boxes indicate causal interactions.

The reconstructions for the different methods are here based on a point estimate of the VAR coefficients, rather than the whole ROC curve. For Granger causality, this estimate is obtained by thresholding the standardized Granger score. A causal influence is defined to be significant, if the standardized score exceeds a threshold of 0.5. The regularizing constant of Ridge Regression, Lasso and Group Lasso is fixed using

10-fold cross-validation. Note that for the Lasso variants, this already determines the sparse causality structure. For Ridge Regression, we perform subsequent sparsification using a significance level of $\gamma = 0.05$.

We display the estimated binary influence matrices in the bottom row of Figure 1. In the top row, we also show for the sake of comprehensibility the quantities these matrices are derived from by means of thresholding. In cases of Lasso and Group Lasso these quantities are simply the estimated AR coefficients and the threshold is zero (the machine precision). For Ridge Regression we depict the negative logarithmic p-values derived from the AR coefficients, while for complete Granger causality the standardized Granger score is shown.

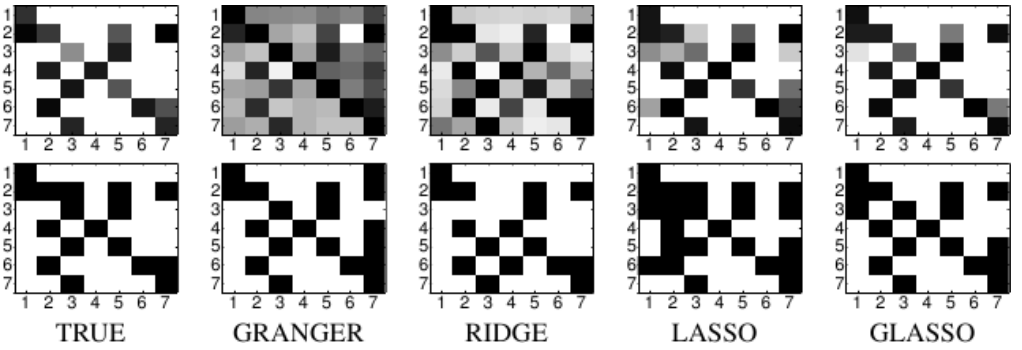


Figure 1: Simulated causal influence matrix and estimates according to Granger Causality, Ridge Regression, Lasso and Group Lasso. In the top row the generating AR coefficients and their Lasso/Group Lasso estimates are shown, as well as the p-values derived from Ridge Regression and the (complete) Granger-score. The bottom row depicts the binarized causal influence matrices.

Table 1 summarizes the AUC scores obtained in the experiments described above. The complementing ROC curves are shown in Figure 2. In short it can be stated that Group Lasso and Ridge Regression outperform their competitors in all scenarios, although not always significantly. While Ridge Regression performs slightly better than Group Lasso in the noiseless condition, Group Lasso has a clearly visible yet insignificant advantage over all methods in the white noise setting. Under the influence of mixed noise Ridge Regression and Group Lasso are on par. Note furthermore that the ROC curve for Lasso is below the ROC curve of Group Lasso, which shows that Lasso tends to be too dense. Interestingly, knowledge of the true model order hardly provided any significant advantage in our simulations.

5. Conclusion

We presented a novel approach for causal discovery in multivariate time series which is based on the Group Lasso. As an alternative we also discussed Ridge Regression

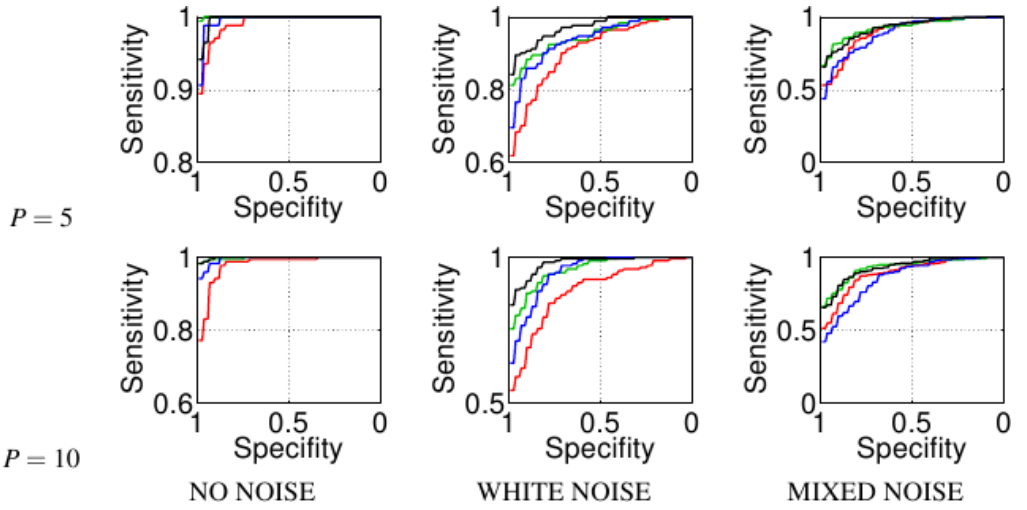


Figure 2: Average ROC curves of Granger Causality (red), Ridge Regression (green), Lasso (blue) and Group Lasso (black) in three different noise conditions and for two different model orders.

		GRANGER	RIDGE	LASSO	GLASSO
$P = 5$	NO NOISE	0.991 ± 0.004	$\mathbf{1.000 \pm 0.000}$	0.996 ± 0.002	0.997 ± 0.002
	WHITE NOISE	0.910 ± 0.023	0.948 ± 0.020	0.941 ± 0.021	0.971 ± 0.016
	MIXED NOISE	0.896 ± 0.012	$\mathbf{0.928 \pm 0.010}$	0.889 ± 0.011	$\mathbf{0.926 \pm 0.012}$
$P = 10$	NO NOISE	0.980 ± 0.005	$\mathbf{0.998 \pm 0.002}$	$\mathbf{0.996 \pm 0.002}$	$\mathbf{0.999 \pm 0.001}$
	WHITE NOISE	0.885 ± 0.019	0.958 ± 0.012	0.948 ± 0.013	$\mathbf{0.979 \pm 0.005}$
	MIXED NOISE	0.893 ± 0.013	$\mathbf{0.931 \pm 0.015}$	0.861 ± 0.014	$\mathbf{0.931 \pm 0.007}$

Table 1: Average AUC scores and standard errors of Granger Causality, Ridge Regression, Lasso and Group Lasso in three different noise conditions and for two different model orders. Entries with significant superior score are highlighted.

with subsequent multiple testing according to Hothorn et al. (2008) which is also novel in the context of VAR modeling. Both approaches were shown to outperform standard methods in simulated scenarios. Future research will aim at applying our techniques to real-world problems. Given that the sparsity assumption is correct, our Group Lasso approach should be able to handle much larger problems than the ones that were considered here by 1) splitting the problem into M independent subproblems and 2) using the active set solver of Roth and Fischer (2008) in combination with strong regularization that ensures staying in the sparse regime. We expect that this will allow large-scale applications such as the estimation of cerebral information flow from functional Mag-

netic Resonance Tomography (fMRI) recordings to benefit from the improved accuracy of our approach.

Acknowledgments

This work was supported in part by the German BMBF (FKZ 01GQ0850, 01-IS07007A and 16SV2234) and the FP7-ICT Programme of the European Community under the PASCAL2 Network of Excellence, ICT-216886. We thank Thorsten Dickhaus for discussions.

References

A. Arnold, Y. Liu, and N. Abe. Temporal Causal Modeling with Graphical Granger Methods. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 66–75, 2007.

F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality and the SMO algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

M. Drton and M.D. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.

C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

S. Haufe, V.V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2): 726–738, 2008.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous Inference in General Parametric Models. *Biometrical Journal*, 3:346–363, 2008.

S.L. Marple. *Digital Spectral Analysis with Applications*. Prentice Hall, Englewood Cliffs, NJ, 1987.

G. Nolte, A. Ziehe, V.V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.R. Müller. Robustly Estimating the Flow Direction of Information in Complex Physical Systems. *Physical Review Letters*, 100(23):234101, 2008.

R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 9, 2007.

V. Roth and B. Fischer. The Group Lasso for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, pages 848 –855, 2008.

J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32, 2005.

A. Schlögl. BIOSIG - an open source software library for biomedical signal processing, http://BIOSIG.SF.NET, 2003.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

P.A. Valdes-Sosa, J.M. Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garcia, and E. Canales-Rodriguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B*, 360:969–981, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

JMLR: Workshop and Conference Proceedings 6:107–120, 2010 NIPS 2008 workshop on causality

Inference of Graphical Causal Models: Representing the Meaningful Information of Probability Distributions

Jan Lemeire JAN.LEMEIRE@VUB.AC.BE

Kris Steenhaut KRIS.STEENHAUT@VUB.AC.BE

Dept. of Electronics and Informatics (ETRO),
Vrije Universiteit Brussel (VUB) - Interdisciplinary Institute for Broadband Technology (IBBT),
Pleinlaan 2, 1050 Brussels, Belgium

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

This paper studies the feasibility and interpretation of learning the causal structure from observational data with the principles behind the Kolmogorov Minimal Sufficient Statistic (KMSS). The KMSS provides a generic solution to inductive inference. It states that we should seek for the minimal model that captures all regularities of the data. The conditional independencies following from the system's causal structure are the regularities incorporated in a graphical causal model. The meaningful information provided by a Bayesian network corresponds to the decomposition of the description of the system into Conditional Probability Distributions (CPDs). The decomposition is described by the Directed Acyclic Graph (DAG). For a causal interpretation of the DAG, the decomposition should imply modularity of the CPDs. The CPDs should match up with independent parts of reality that can be changed independently. We argue that if the shortest description of the joint distribution is given by separate descriptions of the conditional distributions for each variable given its effects, the decomposition given by the DAG should be considered as the top-ranked causal hypothesis. Even when the causal interpretation is faulty, it serves as a reference model. Modularity becomes, however, implausible if the concatenation of the description of some CPDs is compressible. Then there might be a kind of meta-mechanism governing some of the mechanisms or either a single mechanism responsible for setting the state of multiple variables.

1. Introduction

Causal inference is an ambitious research field, as it tries to learn how the world is put together from observations only. The algorithms for causal inference are based on the conditional independencies implied by the causal structure of the system. The theory of graphical causal models, as developed by Pearl et al., gives a probabilistic view on causation and is based on the theory of Bayesian networks. The Directed Acyclic Graph (DAG) of a Bayesian network can be regarded as a representation of the conditional independencies of a probability distribution. A causal model gives a

causal interpretation to the edges of a Bayesian network. The causal interpretation is based on manipulability; the model exhibits the structure of the system such that it is able to predict changes to the system. Hausman and Woodward (1999) show that this interventionist interpretation of causality is tightly linked to modularity. They also defend the equivalence of modularity and the causal Markov condition (Hausman and Woodward, 1999, p. 554).

The causal interpretation of a Bayesian network is often criticized (Freedman and Humphreys, 1999; Cartwright, 2001; Williamson, 2005; Hausman and Woodward, 1999). This paper would like to contribute to the discussion by analyzing causal inference through the concept of the *Kolmogorov Minimal Sufficient Statistic* (KMSS). The idea is that patterns or regularities in the observed data do not happen by accident. They teach us the important properties of the system. We say that the regularities constitute the *meaningful information* of the data. The KMSS allows a formal separation of meaningful and random information, based on the Kolmogorov complexity of objects. The application of Kolmogorov complexity to inductive inference has given rise to different methods, such as Minimum Message Length (MML) (Wallace and Boulton, 1968) and Minimum Description Length (MDL) (Rissanen, 1978). These methods are used for selecting the best model from a given set of models. The choice of model class, however, determines the regularities under consideration. During our discussion, we will not stick to an a priori chosen set of regularities, but search for the relevant regularities. Regularities will show up to be of key importance for testing the validity of causal inference.

This paper puts forward that the meaningful information of a Bayesian network is the decomposition of the system's description into separate components, the Conditional Probability Distributions (CPDs). The correctness of the causal interpretation of this decomposition relies on whether the CPDs correspond to independent mechanisms. We will analyze the correctness by looking at the regularities not incorporated by the Bayesian network.

In Section 2, the concept of KMSS is introduced. In Section 3, we will give a survey of graphical causal model theory and the learning algorithms. The link between a Bayesian network and the KMSS of a probability distribution is discussed in Section 4. In Section 5 we will argue that causal inference is plausible if the Bayesian network gives the KMSS. Section 6 discusses the cases in which the minimal Bayesian network does not provide the minimal description.

2. Meaningful Information

Kolmogorov Complexity provides an objective measure of simplicity so that Occam's razor can be applied. The *Kolmogorov Complexity* of a string x is defined to be the length of the shortest computer program that prints the string and then halts (Li and Vitányi, 1997):

$$K(x) = \min_{p:\mathcal{U}(p)=x} l(p) \tag{1}$$

with $\mathcal{U}$ a universal Turing machine and $l(p)$ the size in bits of program p. Patterns in the string allow for its compression, i.e. to describe the data using fewer symbols than the number of symbols to describe the data literally. The string "00010001000100010001000100010001000100010001" can be described shorter by program `REPEAT 11 TIMES "0001"`. But not all bits of this program can be regarded as containing *meaningful information*. We consider meaningful information as the properties of the string that allow for its compression (Vitányi, 2002). Such properties are called patterns or *regularities*. The regularity of the string is the repetition. The number of repetitions (11) or the substring `"0001"` is random information. A random string, which is incompressible, has no meaningful information at all.

For inductive inference, we will look for a minimal description in 2 parts, one containing the regularities of the data, which we call the model, and one part containing the remaining random noise. Such a description is called a *two-part code*. This results in generic approaches for inductive inference, such as *Minimum Description Length* (MDL). According to MDL we have to pick the model M_{mdl} from model class $\mathcal{M}$ where M_{mdl} is the model which minimizes the sum of the description length of M and of the data D encoded with the help of M (Grünwald, 1998):

$$M_{mdl} = \arg\min_{M \in \mathcal{M}} \{L(M) + L(D \mid M)\} \tag{2}$$

with $L(.)$ the description length.

The MDL approach relies on the a priori chosen model class. It does not tell us how to make sure the models capture all and nothing more than the regularities of the data. The KMSS provides a formal separation of meaningful and meaningless information. We limit the introduction of KMSS to models that can be related to a finite set of objects, called the *model set*. In the context of learning, we are interested in a model set S that contains string x and the objects that share x's regularities. With $|S|$ the size of set S, all elements of a set S can be enumerated with a binary index of length $\log_2 |S|$. We say that x is *typical* for S if

$$K(x \mid p_S) \geq \log_2 |S| - \beta \tag{3}$$

with p_S the shortest program that describes S and β an agreed upon constant. The index is constructed by index enumerating all elements of the set, its length is thus $\log_2 |S|$. Atypical elements have regularities that are not shared by most of the set's members and can therefore be described by a shorter description. Most elements of S are typical, since, by counting arguments, only a small portion of it can be described shorter than $\log_2 |S|$.

The *Kolmogorov Minimal Sufficient Statistic* (KMSS) of x is defined as the shortest program p^* which describes the smallest set S^* such that x is a typical element of S^* and the two-stage description of x is as good as the minimal single-stage description of x (Gács et al., 2001):

$$p^* = \arg\min_p \{l(p) \mid \mathcal{U}(p) = S,\ x \in S,\ K(S) + \log_2 |S| \leq K(x)\} \tag{4}$$

Program p^* minimally describes the meaningful information present in x and nothing else. This can be understood as follows. By the inequality, the two-part description is at least as short as the Kolmogorov complexity of x. Since we seek for the simplest S (minimal p), we will only describe regularities by p. Regularities compress the description and greatly reduce the size of S. Putting random information in S would also reduce $\log_2 |S|$, but would increase $K(S)$ equally.

$K(x)$ depends on the chosen Turing machine $\mathcal{U}$, or, in practice, on the chosen description language. Minimality and compressibility are thus partly dependent on the choice of language. Another problem for directly applying the definitions is the intractability of $K(x)$. It falls out of the scope of this paper to address these problems, consult (Li and Vitányi, 1997) for an in-depth analysis. We will apply the *principles* behind Kolmogorov complexity, MDL and KMSS in order to better understand causal inference, its interpretation and feasibility.

3. Graphical Causal Models

This chapter will introduce graphical causal models and the learning algorithms (Pearl, 2000; Spirtes et al., 1993; Tian and Pearl, 2002).

3.1. Representation of Causal Relations

Graphical causal models intend to describe with a Directed Acyclic Graph (DAG) the structure of the underlying physical mechanisms governing a system under study. The state of each variable, represented by a node in the graph, is generated by a stochastic process that is determined by the values of its parent variables in the graph. All variables that influence the outcome of the process are called *causes* of the outcome variable. An *indirect cause* produces the state of the effect indirectly, through another variable. If there is no intermediate variable among the known variables, the cause is said to be a *direct cause*.

Each process represents a physical mechanism. In it most general form it can be described by a conditional probability distribution (CPD) $P(X \mid Pa(X))$, where $Pa(X)$ is the set of parent nodes of X in the graph and constitute the direct causes of the variable. The combination of the CPDs results in the system's joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \tag{5}$$

The right hand side is called a *factorization* of the joint probability distribution.

3.2. The Effect of Changes to the System

We attribute a causal interpretation to the edges of the graph, but what does this 'causal interpretation' signify? The approach of Pearl and many others is to draw a connection between causation and manipulability (Hausman and Woodward, 1999). The causal interpretation is defined by the model's capacity to predict the effect of changes to the

system. Changes are defined by Pearl as *interventions*. An intervention is defined as an atomic operation that fixates a set of variables to some given states and eliminates the corresponding factors (CPDs) from the factorization (Eq. 5) (Pearl, 2000). Applied on a causal graph, an intervention on variable X sets the value of X and breaks all of the edges in the graph directed into X and preserves all other edges in the graph, including all edges directed out of X. This is called the Manipulation Theorem by Spirtes et al. (1993, p. 51). Intervening on a variable only affects its effects. Causes have to be regarded as if they were levers which can be used to manipulate their effects.

This approach does not directly define causality, but defines the implications of having a thorough knowledge of the mechanisms that make up a system. Manipulability puts a constraint of independentness on the mechanisms. The accuracy of the mutilated model relies on autonomy or modularity; a mechanism can be replaced by another without affecting the rest of the system. It is defined by Hausman and Woodward (1999, p. 545) as follows. Note that they relate each CPD to a structural equation.

Definition 1 *(Modularity) For all subsets* $\mathbf{Z}$ *of the variable set* $\mathbf{V}$*, there is some non-empty range* $\mathbf{R}$ *of values of members of* $\mathbf{Z}$ *such that if one intervenes and sets the value of the members of* $\mathbf{Z}$ *within* $\mathbf{R}$*, then all equations except the equations with a member of* $\mathbf{Z}$ *as a dependent variable (if there is one) remain invariant.*

3.3. Representation of Independencies

The evidence for causal inference is the conditional independencies entailed by the system's causal structure. For a causal model, the *causal Markov condition* gives us the independencies that follow from the causal structure: each variable is probabilistically independent of its non-effects conditional on its direct causes (Spirtes et al., 1993). These independencies are irrespective of the nature of the mechanisms, of the exact parameterization of the conditional probability distributions $P(X_i \mid Pa(X_i))$. All independencies following from the causal Markov condition can be retrieved from the causal graph by the d-separation criterion. A causal graph is called *faithful* if all conditional independencies from the distribution follow from the causal Markov condition.

3.4. Correspondence with Bayesian Networks

Graphical causal models provide a probabilistic account of causality (Spohn, 2001). This resulted in a close correspondence with Bayesian networks. In contrast to causal models, Bayesian networks are only concerned with offering a dense and manageable representation of joint distributions. A joint distribution over n variables can be *factorized* relative to a chosen variable ordering $(X_1, \ldots, X_n)$ as follows:

$$P(X_1, \ldots, X_n) = \prod_i^n P(X_i \mid X_1, \ldots, X_{i-1}) \tag{6}$$

Variable X_j can be removed from the conditioning set of variable X_i if it becomes conditionally independent from X_i by conditioning on the rest of the set:

$$X_j \perp\!\!\!\perp X_i \mid X_1 \ldots X_{j-1}, X_{j+1} \ldots X_{i-1} \tag{7}$$

Such conditional independencies reduce the complexity of the factors in the factorization. The conditioning sets of the factors can be described by a Directed Acyclic Graph (DAG), in which each node represents a variable and has incoming edges from all variables of the conditioning set of its factor. The joint distribution is then described by the DAG and the conditional probability distributions (CPDs) of the variables conditional on their parents: $P(X_i \mid Pa(X_i))$. A *Bayesian network* is a factorization that is edge-minimal, in the sense that no edge can be deleted without destroying the correctness of the factorization.

Causal models attribute a causal interpretation to the edges of the graph of a Bayesian network and are therefore called *causally interpreted Bayesian networks*. Bayesian networks are just dense descriptions of probability distributions and offer an explicit representation of dependencies and independencies. The link is that the causal Markov condition follows from the correctness of the factorization (Hausman and Woodward, 1999, p. 532).

Although edge-minimality of a Bayesian network, the graph depends on the chosen variable ordering. Some orderings lead to the same networks, while others result in different topologies. All networks represent the probabilities just as well, except that some are more complex than others. We call the *minimal Bayesian networks* the Bayesian networks which have the least number of edges in their DAGs.

3.5. Causal Inference

The goal of causal inference is to learn the causal structure of a system based on observational data. Causal structure learning algorithms fall apart in two categories: scoring-based and constraint-based algorithms. Scoring-based algorithms are based on an optimized search through the set of all possible models, which tries to find the minimal model that best describes the data. Each model is given a score that is a trade-off between model complexity and goodness-of-fit. Different scoring criteria have been applied in these algorithms, such as a Bayesian scoring method (Cooper and Herskovits, 1992), an entropy based method (Herskovits, 1991) and one based on the Minimum Description Length (Suzuki, 1996). Irrespective of the exact definition of the scoring criteria, we can say that the algorithms are looking for the minimal Bayesian network.

Constraint-based learning algorithms rely on the conditional independencies detected that follow from the system's causal structure. It is a kind of evidence-based construction, the decisions to include an edge and on the edge's orientation are based on the presence or absence of certain independencies. The algorithms assume minimality, faithfulness and the causal Markov condition (Spirtes et al., 1993). We are also searching for the minimal Bayesian network, since a faithful Bayesian network is minimal (Lemeire et al., 2009).

4. Bayesian Networks as Minimal Descriptions of Distributions

In this section we will draw the connection between Bayesian networks and the KMSS of probability distributions. Let us apply the principle of KMSS to inductive inference

of multivariate data that are independently and identically distributed (i.i.d.). Following the principle, the inferred model should capture the regularities of the data. The type of regularity we have to consider is a dependency between variables; knowing one variable gives information about the state of another variable. The knowledge about the state of a single stochastic variable is captured by a probability distribution over it. Dependency information is captured by the joint probability distribution defined over the variables of interest. The KMSS of the distribution should be a minimal description of the distribution's regularities. In this section we will consider the description given by a Bayesian network, other regularities will be considered in Section 6.

From the theory of Bayesian networks, we know that a joint distribution can be described shorter by a factorization (relative to a certain variable ordering) that is reduced by conditional independencies (given by Eq. 7). This leads to the description of the joint distribution by a factorization: $P(X_1,\ldots,X_n) = \prod CPD_i$, with CPD_i the CPD of variable X_i, defining $P(X_i \mid Pa(X_i))$, a distribution over X_i conditional on a subset of some other variables. A two-part description of a joint distribution is then:

$$descr(P(X_1 \ldots X_n)) = descr(\{Pa(X_1),\ldots,Pa(X_n)\}) + descr(CPD_1) + \cdots + descr(CPD_n) \tag{8}$$

With $descr()$ denoting a description. The parents' lists can be described very compact by a DAG. The descriptive size of the CPDs is determined by the number of variables in the conditioning sets, the number of free parameters for describing the distributions and the chosen accuracy. Eq. 8 corresponds to the description of a Bayesian network. If this results in an incompressible description, the DAG gives the meaningful information and the KMSS. This is proven by the following theorem.

Theorem 2 *Given a set of probability distributions $\mathcal{P}$ defined over a set of n variables $X_1,\ldots,X_n$. Consider a probability distribution $P \in \mathcal{P}$ which can be decomposed by a factorization based on parents' lists $Pa(X_1),\ldots,Pa(X_n)$. Consider that the factorization is described (see Eq. 8) by a minimal code which is able to describe all $P' \in \mathcal{P}$ that can be described by a factorization based on the same parents' lists. If such a description results in an incompressible string, then the first part (the description of the parents' lists) is the Kolmogorov minimal sufficient statistic of P.*

Proof The parents' lists describe a subset $\mathcal{P}' \subset \mathcal{P}$ which includes all elements that can be decomposed by the same factorization. We assume that this description is incompressible, therefore its length corresponds to $K(\mathcal{P}')$ up to a constant [1]. The code used to describe the CPDs allows a description of all elements of $\mathcal{P}'$. It is a minimal code, so its length equals to $\log_2 |\mathcal{P}'|$. The total description is incompressible, so its length equals, up to a constant, to $K(P)$. We have found a set $\mathcal{P}'$, for which $K(\mathcal{P}') + \log_2 |\mathcal{P}'| \leq K(P)$.

Next, we have to prove that there is no other set, $\mathcal{P}''$, which has a shorter description and for which the inequality holds (see Eq. 4). Assume that such a $\mathcal{P}''$ exists. If

1. Two minimal descriptions of x based on different codes or turing machines are equal up to a constant that is independent of x. Since the descriptions are minimal, they are incompressible.

$\mathcal{P}'' \subset \mathcal{P}'$, the description of the CPDs would be compressible. $\mathcal{P}''$ is a smaller set, indicating that there are regularities in P which are not described by $\mathcal{P}'$. If $\mathcal{P}' \subset \mathcal{P}''$, then, similarly, the description of $K(\mathcal{P}'') + \log_2 |\mathcal{P}''|$ would be compressible. It follows that if such a $\mathcal{P}''$ exists, both sets contain exclusive elements that do not belong to the other set. This implies that the descriptions of both sets exploit regularities of P that are not exploited by the other one. Therefore neither of the descriptions is minimal and incompressible. This proves that such a $\mathcal{P}''$ does not exist.

■

The DAG thus minimally describes the dependencies among the variables, the model's complexity is reduced by conditional independencies.

It must be noted that if there exists a faithful and minimal Bayesian network, it is not necessarily unique. Multiple minimal models can exist for a distribution. These models represent the same set of independencies and are therefore statistically indistinguishable. They define a *Markov-equivalence class*. It is proved that they share the same skeleton and v-structures. They only differ in the orientation of some edges (Pearl, 2000). This set can be represented by a partially-directed acyclic graph in which some of the edges are not oriented. The corresponding factorizations have the same number of conditioning variables. Thus, all models of a Markov-equivalence class have the same complexity.

5. Correspondence of Decomposition to Independent Mechanisms

Causal inference from observations is based on finding the minimal Bayesian network (Sec. 3.5) and attaching a causal interpretation to it (Sec. 3.2). In this section we will discuss the case in which, for a given set of observations, there is exactly one minimal Bayesian network which is also the minimal description of the data. This means that there are no other regularities than the conditional independencies the model represents. The DAG is then the KMSS of the data and minimally represents all regularities. We argue that description minimality can be linked to causality.

Note that for not overloading the discussion we will assume causal sufficiency: there are no unknown variables that affect more than one known variable.

5.1. Correspondence of Bayesian Networks to a Decomposition

Let us first analyze what the meaningful information of a Bayesian network exactly represents. A Bayesian network describes a joint probability distribution by DAG G and a list of CPDs as given by Equation 8. The description is decomposed into individual CPDs. The decomposition is described by the DAG, it tells us which CPDs we have to consider. The DAG G contains the meaningful information. It describes a model set of distributions $\mathcal{P}_G$, all sharing the conditional independencies following from the Markov condition. The distributions that only have these independencies are the typical elements of $\mathcal{P}_G$. G is faithful to them. Distributions having other independencies, are atypical, their description based on G is compressible (Lemeire et al., 2009, Theorem

6). We will consider them in the next section. We will first assume that the description is unique and minimal.

A Bayesian network thus describes a decomposition and matches with a reductionist view, according to which the world can be studied in parts. Indeed, if the system cannot be decomposed, if there are no conditional independencies that simplify a factorization, then the DAG does not contain meaningful information (Theorem 2). We end up with a Holist system in which everything depends on everything.

It must be noted that the assumption of a unique minimal Bayesian network is not essential. If the minimal Bayesian network is not unique, the Markov-equivalence class indicates exactly which parts are undecided, namely the orientation of some edges. So, we know exactly for which parts of the model we have not enough information to decide upon the decomposition. For the remainder of the model, the decomposition is known. When we speak of the minimal Bayesian network, we actually mean the class of closely-related minimal Bayesian networks.

5.2. Correspondence of the Decomposition to Mechanisms

Let us now investigate whether the CPDs of the minimal Bayesian network of a probability distribution can be matched up with the mechanisms of the underlying system. The minimal Bayesian network provides modularity in the descriptive sense: the description consists of components. But does this also imply modularity in the causal sense: do the descriptive components correspond to independent parts of reality? In other words, does the model learned by observations reveal the underlying system?

We have found the simplest model. Following Occam's razor this is the model we should 'select'. But does Occam's razor also guarantees that this model tells us something about the real system? Yet, we did not only find the minimal Bayesian networks in the set of all Bayesian networks, we also found the minimal model in the general sense. The model is the KMSS and has extracted all regularities from the data. Moreover, the model is unique. From these facts we argue that:

In absence of background knowledge, experiments with interventions or other information, given that a minimal Bayesian network is the KMSS of the data, the top-ranked hypothesis is that each CPD represents an independent part of reality.

Before explaining what we exactly mean by 'top-ranked' hypothesis, let us consider two counter examples.

5.3. Counter Examples

Consider people living at different latitudes and the amount of vitamin D creation. Melanin is a pigment that protects us against harmful UV radiation. On the other hand, we need a limited amount of UV radiation to produce a necessary amount of vitamin D. To ensure this, evolution has given humans a different amount melanin, which is reflected by skin color, relative to the amount of sun they are exposed to. The latter is mainly affected by the latitude. This results in a nearly constant amount of vitamin D creation independent from the latitude we live at. Figure 1 shows the real causal model

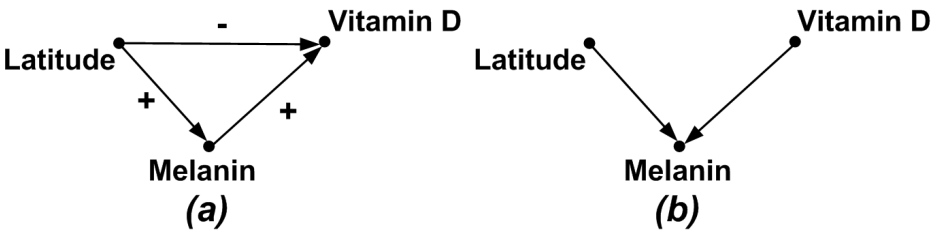


Figure 1: The relation between vitamin D creation and latitude: true causal model (a) and minimal Bayesian network (b).

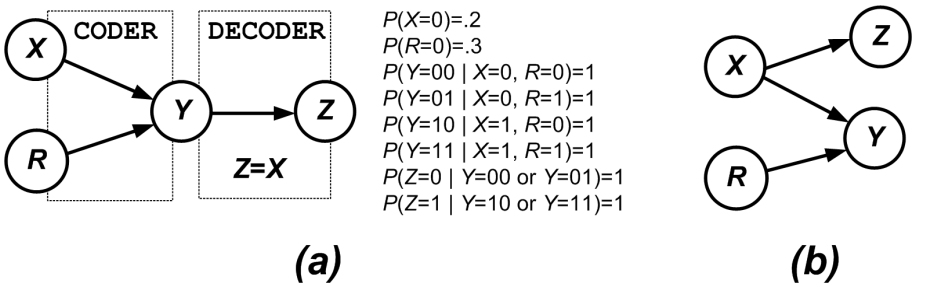


Figure 2: Coder-decoder system, taken from Spirtes et al. (1993, Figure 3.23), in which Z equals to X. Description of the system (a) and minimal model (b).

(a) and the minimal Bayesian network (b). Evolution has controlled the *Latitude* $\rightarrow$ *Melanin* relation such that the parameters were calibrated until the influences from *Latitude* on *Vitamin D* neutralized. This is a counter example of Occam's razor; the simplest model does not give us the true model. There is a *meta-mechanism*, namely evolution, controlling the mechanisms.

Next, consider the coder-decoder example, taken from Spirtes et al. (1993, Figure 3.23), shown by Figure 2(a). Variable Y encodes the values of both R and X, and Z decodes Y to match the value of X. This is possible because the first bit of Y corresponds to the value of X. The coder-decoder system is designed to exhibit the specific behavior that Z equals to X. The model describing such a system, shown in Figure 2(a), is clearly not minimal. In the minimal description of the system given by Figure 2(b), Z is directly related to X. Occam's razor is violated. The CPD components are part of a greater mechanism and are engineered to match each other in such a way that the desired functionality is realized. The causal interpretation of the minimal model is incorrect. If we intervene on Y by manually setting the value of Y to a certain value which is not controlled by X, then Z becomes independent of X.

5.4. Conclusions about the Causal Interpretation of the Decomposition

The examples illustrate that the real system can be more complex than suggested by the complexity of the observations. Does this invalidate our claim about the minimal Bayesian network? It shows that Occam's razor cannot always be trusted. The minimal model may be incorrect, in the sense that the causal interpretation should be considered with care. But even when faulty, the minimal Bayesian network tells us two things about the system.

First, the DAG of the minimal Bayesian network describes the qualitative behavior of the system. This is true for both counter examples. From Figure 1(b) we see that the amount of vitamin D creation is independent from the latitude we live in. From Figure 2(b) we immediately see that Z only depends on X and is totally independent from R and Y. The structure of the real causal models does not reveal these independencies. We have to carefully study the parameterization to understand these independencies. Moreover, these are not accidental independencies. For the first example, it is the result of an evolution triggered by the evolutionary fitness. For the second example, it was the deliberate intention of the engineer to give the system this specific behavior. This corresponds to the rationale behind Occam's razor: regularities are most likely not accidental, but indications of a kind of mechanism. Only the occurrence of accidental (in)dependencies in the data, due to a limited sample size for example, makes the minimal model not correctly describing the system's behavior.

Secondly, we argue that the minimal Bayesian network at least serves as a reference model. The correspondence of the CPDs to real mechanisms might be untrue due to a meta-mechanism controlling the configuration of the system. The likelihood of the occurrence of such a mechanism must be estimated from background knowledge. In that case, experiments with interventions will have to be performed in order to reveal the true causal model (Korb and Nyberg, 2006). But even then will the minimal Bayesian network show its value. It can be used as a *reference model*, which will be compared with the model learned after the application of the interventions. This comparison would reveal the meta-mechanism.

6. Compressibility of the Minimal Bayesian Network

In this section we will dig deeper. We will investigate the consequences for cases in which the minimal Bayesian network does not describe the KMSS. Then, the DAG is not the only meaningful information. There exists a simpler description of the distribution than given by the minimal Bayesian network. First we will consider the compressibility of an individual CPD and then we consider the compressibility of several CPDs taken together.

6.1. Compressibility of a Single CPD

Compressibility of individual CPDs is called *local structure* (Friedman and Goldszmidt, 1996). In this terminology, the DAG describes the global structure, the CPDs the local structure. On top of the independencies following from the causal structure the

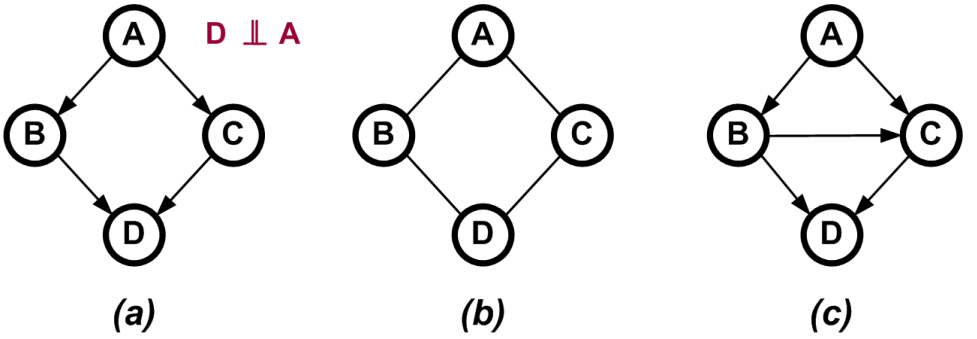


Figure 3: O-structure in which A is independent from D (a). A Markov network (b) and one of the minimal Bayesian networks describing the same system (c).

individual CPDs exhibits additional regularities. For discrete models for example, the conditional probability tables can be described shorter by decision trees when so-called context-specific independencies appear (Boutilier et al., 1996).

A specific type of local structure is the decomposition of a CPD into independent components. In general, a CPD describes the mechanism by which all direct causes together produce the state of a single variable. Various authors report on independent cause-effect relations. They study representations in which the causal influences of the direct causes of a variable are independent, for example by a factorized representation of a CPD (Madsen and D'Ambrosio, 2000). Hausman and Woodward (1999, p. 547) call it disjunctive causes. On top of the decomposition given by Equation 8, the parts of the decomposition can be further decomposed.

In these cases, the decomposition according to Equation 8 is still valid. Modularity is still valid, in the descriptive and causal sense. The same conclusions as those of the previous section apply.

6.2. Compressibility of the Concatenation of CPDs

When the description of some CPDs together can be compressed, the regularity indicates that the CPDs are in some way related. The following counter examples show that this often invalidates the causal interpretation.

6.2.1. O-STRUCTURE

The most-known counter example of causal inference is when in the model of Figure 3(a), A and D appear to be independent (Spirtes et al., 1993). This happens when the influences along the paths $A \to B \to D$ and $A \to C \to D$ exactly balance, so that they cancel each other out and the net effect results in an independence. This cancellation is similar to the vitamin D example of Section 5.3. Except that in this case the DAG of the minimal Bayesian network corresponds to the structure of the model. The CPDs and the mechanisms are, however, not independent. Pearl considers the exact cancellation of

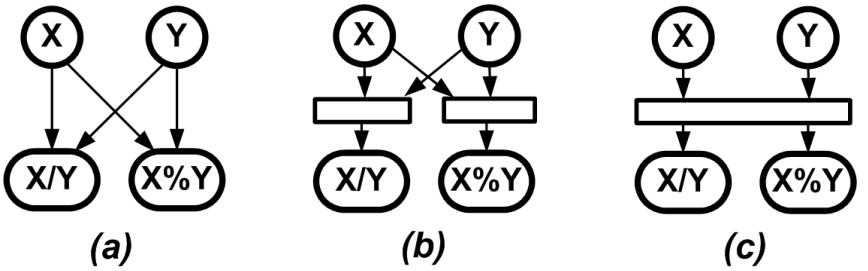


Figure 4: Bayesian network (a) of a system that calculates the quotient and remainder of two integers. The decomposition it represents (b) and a single mechanism calculating both outputs together (c).

the parameters as a measure zero event, since the probability of such a coincidence can therefore be regarded as nearly zero (Pearl, 2000, p. 48). This is true as long as there not a kind of *meta-mechanism* controlling the mechanisms such that the parameters are calibrated until they neutralize. This confirms our conclusions of the previous section that the existence of such mechanisms must be taken into consideration. Then, the likelihood of a cancellation is not zero.

Modularity becomes invalid when the meta-mechanism acts instantly. In the vitamin D case, evolution works slowly. On short term, the mechanisms are independent. It is only on the long term that the calibration will be reestablished.

6.2.2. MARKOV NETWORK

Consider a system that is minimally described by a Markov network, as shown in Figure 3 (b). Variables which are connected by a path in the network are dependent, unless each path is blocked by one of the conditioning variables. So is $B \not\perp\!\!\!\perp C \mid A$, but $B \perp\!\!\!\perp C \mid \{A, D\}$. For describing the same network with a DAG, we have to orient the edges of the network. For acyclicity, we have to create at least one v-structure. We can choose for example $B - D - C$. But then, for keeping the same dependencies, we have to add an edge, as shown in Figure 3 (c). Without $B \to C$ we would have $B \perp\!\!\!\perp C \mid A$. Clearly, this Bayesian network is not minimal; the description is longer than that of a Markov network. The parameterizations of the CPDs contain redundancies. In the model of 3 (c), the parameterizations must ensure that $B \perp\!\!\!\perp C \mid \{A, D\}$, an independency which is not captured by the DAG. When such a distribution is observed, the causal interpretation of the CPDs of the minimal Bayesian networks is incorrect.

6.2.3. MULTIPLE-OUTPUT FUNCTIONS

Consider a system that calculates the quotient and remainder of two integers. Figure 4(a) shows the minimal Bayesian network of the system. The model describes the system as two different mechanisms, one for calculating the quotient and one for the remainder, shown in Figure 4(b). Both mechanisms, however, are related; there is a

lot of overlap in calculating the quotient and the remainder. A model describing the system by one component which calculates both outputs together, as shown in Figure 4(c), is more compact than a Bayesian network which only allows components with single outputs. In that case, the CPDs of the minimal Bayesian network cannot be considered as independent. If the output variables are clearly separated quantities, a mechanism setting the values of multiple variables should be taken into consideration.

6.2.4. OBJECT-ORIENTED NETS

Another regularity is the repetition of similar mechanisms in a system. This results in a causal model in which identical CPDs appear. The model is therefore compressible. The compressibility does not necessarily result in a dependence of the CPDs in terms of manipulability. It depends on the meta-mechanism responsible for the regularities in the system. The system could, for example, be designed by an engineer, such as a digital circuit. Then, modularity holds; one mechanism can be replaced by another without affecting the rest of the model. *Object-Oriented nets* provide a representation format that explicitly capture similarities of mechanisms (Koller and Pfeffer, 1997).

7. Conclusions

We showed that the meaningful information described by a Bayesian network about a probability distribution is the decomposition of the distribution into CPDs. We argue that if the shortest description of the joint distribution is given by separate descriptions of conditional distributions, it is the *top-ranked causal hypothesis*:

(1) The Bayesian network gives a correct description of the behavior of system. The qualitative properties, namely the conditional independencies, are incorporated in the DAG. This only becomes invalid by accidental (in)dependencies due to for instance a limited sample size.

(2) The likelihood that the CPDs correspond to independent mechanisms of the system (modularity) depends on the likelihood of a kind of meta-mechanism. A meta-mechanism could result in a system which is more complex than suggested by its behavior. In that case, the behavior of the system when subjected to an external intervention will not be correctly predicted by the minimal model. Nonetheless, the minimal Bayesian network can serve as a reference model that has to be compared with the behavior of the system after applying the interventions.

By applying the principle of KMSS, we did not only look for the minimal Bayesian network in the set of all Bayesian networks. We also took into consideration the presence of other regularities than the conditional independencies following from the system's causal structure. These regularities might invalidate the above conclusions. If such regularities appear in individual CPDs, the mapping of CPDs onto independent mechanisms still holds (in the above sense). The DAG of the Bayesian network does not have to be the KMSS. The decomposition should be correct. It becomes incorrect if the concatenation of the description of the CPDs is compressible. Then the CPDs are no longer independent and modularity might become invalid. The dependence might be

caused by a meta-mechanism that governs the dependent mechanisms, or a mechanism affecting the state of multiple variables.

References

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

Nancy Cartwright. What is wrong with Bayes nets? *The Monist*, pages 242–264, 2001.

G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

D.A. Freedman and P. Humphreys. Are there algorithms that discover causal structure? *Synthese*, 121:2954, 1999.

N. Friedman and M. Goldszmidt. Learning bayesian networks with local structure. In *In Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence*, 1996.

Péter Gács, J. Tromp, and Paul M. B. Vitányi. Algorithmic statistics. *IEEE Trans. Inform. Theory*, 47(6):2443–2463, 2001.

P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty, ILLC Dissertation series 1998-03*. PhD thesis, University of Amsterdam, 1998.

Daniel M. Hausman and James Woodward. Independence, invariance and the causal Markov condition. *Bristish Journal For the Philosophy Of Science*, 50(4):521–583, 1999.

E.H. Herskovits. *Computer-Based Probabilistic Network Construction*. PhD thesis, Medical information sciences, Stanford University, CA, 1991.

Daphne Koller and Avi Pfeffer. Object-oriented Bayesian networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 302–313, 1997.

Kevin B. Korb and Erik Nyberg. The power of intervention. *Minds and Machines*, 16 (3):289–302, 2006.

Jan Lemeire, Kris Steenhaut, and Abdellah Touhafi. When are graphical causal models not good models? In *Causality in the sciences, J. Williamson, F. Russo and P. McKay, editors*, 2009.

Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

Anders L. Madsen and Bruce D'Ambrosio. A factorized representation of independence of causal influence and lazy propagation. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 8(2):151–165, 2000.

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

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2nd edition, 1993.

Wolfgang Spohn. Bayesian nets are all there is to causal dependence. In *In Stochastic Causality, Maria Carla Galaviotti, Eds*. CSLI Lecture Notes, 2001.

J. Suzuki. Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. In *Procs of the International Conf. on Machine Learning*, Bally, Italy, 1996.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *AAAI/IAAI*, pages 567–573, 2002.

Paul M. B. Vitányi. Meaningful information. In Prosenjit Bose and Pat Morin, editors, *ISAAC*, volume 2518 of *Lecture Notes in Computer Science*, pages 588–599. Springer, 2002.

Chris S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

Jon Williamson. *Bayesian Nets And Causality: Philosophical And Computational Foundations*. Oxford University Press, 2005.

JMLR: Workshop and Conference Proceedings 6:121–136, 2010 NIPS 2008 workshop on causality

Bayesian Algorithms for Causal Data Mining

Subramani Mani SUBRAMANI.MANI@VANDERBILT.EDU
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN, 37232-8340, USA

Constantin F. Aliferis CONSTANTIN.ALIFERIS@NYUMC.ORG
Center for Health Informatics and Bioinformatics
New York University
New York, NY, 10016, USA

Alexander Statnikov ALEXANDER.STATNIKOV@MED.NYU.EDU
Center for Health Informatics and Bioinformatics
New York University
New York, NY, 10016, USA

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We present two Bayesian algorithms CD-B and CD-H for discovering unconfounded cause and effect relationships from observational data without assuming causal sufficiency which precludes hidden common causes for the observed variables. The CD-B algorithm first estimates the Markov blanket of a node X using a Bayesian greedy search method and then applies Bayesian scoring methods to discriminate the parents and children of X. Using the set of parents and set of children CD-B constructs a global Bayesian network and outputs the causal effects of a node X based on the identification of Y arcs. Recall that if a node X has two parent nodes A, B and a child node C such that there is no arc between A, B and A, B are not parents of C, then the arc from X to C is called a Y arc. The CD-H algorithm uses the MMPC algorithm to estimate the union of parents and children of a target node X. The subsequent steps are similar to those of CD-B. We evaluated the CD-B and CD-H algorithms empirically based on simulated data from four different Bayesian networks. We also present comparative results based on the identification of Y structures and Y arcs from the output of the PC, MMHC and FCI algorithms. The results appear promising for mining causal relationships that are unconfounded by hidden variables from observational data.

Keywords: Causal data mining, Markov blanket, Y structures

1. Introduction and Background

Causal knowledge enables us to plan interventions leading to predictable, measurable and desirable outcomes. Experimental data is typically generated for ascertaining cause and effect relationships. However, experimental studies may not be feasible in many situations due to ethical, logistical, cost, technical or other reasons. This study introduces two new Bayesian algorithms CD-B and CD-H for ascertaining causality from

observational data. There are many algorithms available for learning the underlying causal structure from data such as GS (Margaritis and Thrun, 2000), PC (Spirtes et al., 2000, page 84–85), HITON (Aliferis et al., 2003a), OR (Moore and Wong, 2003) and FCI (Spirtes et al., 2000). However, all of these algorithms except FCI make an assumption of *causal sufficiency* which maintains that there are no unobserved common causes for any two or more of the observed variables. Even though FCI does not make such an assumption its usefulness is limited in practical settings due to its scalability limitation.

The Bayesian algorithms proposed in this paper are based on a computationally feasible score-based search to identify some causal effects in the large sample limit while allowing for the possibility of unobserved common causes, and without making any assumptions about the true causal structure (other than acyclicity). There is also no need to assign scores explicitly to causal structures with unobserved common causes in this framework.

We now define some terms that are needed for our causal datamining framework. Our framework for causal discovery is based on causal Bayesian networks (CBNs). A CBN is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network (Pearl, 1991). We proceed to introduce the concept of a Y structure and a Y

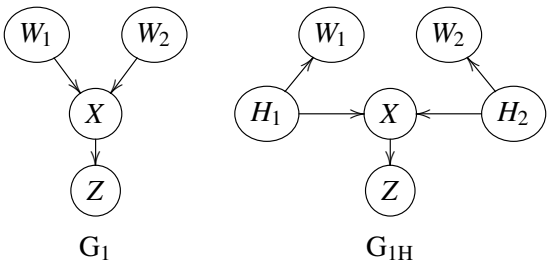


Figure 1: A Y structure G_1 and a Y equivalent structure G_{1H} (H_1 and H_2 denote hidden variables).

arc in a Bayesian network. Let $W_1 \rightarrow X \leftarrow W_2$ be a V structure (there is no arc between W_1 and W_2). If there is a node Z such that there is an arc from X to Z, but no arc from W_1 to Z and no arc from W_2 to Z, then the nodes W_1, W_2, X and Z form a Y structure (see Figure 1, G_1). If such a Y structure over four measured variables $\mathbf{V}$ is learned from an observational dataset D, the arc from X to Z in the Y structure represents an unconfounded causal relationship (Mani et al., 2006). Since G_1 also has the same set of independence/dependence relationships over the observed variables (I-map) as G_{1H} (see Figure 1), the arcs $W_1 \rightarrow X$ and $W_2 \rightarrow X$ in G_1 cannot be interpreted as necessarily representing causal relationships. The arc from X to Z in a Y structure is referred to as a Y arc (YA).

We now define the concept of a *Markov blanket* which is needed for an understanding of the CD-B and CD-H algorithms. The Markov blanket (MB) of a node X in a

causal Bayesian network G is the union of the set of parents of X, the children of X, and the parents of the children of X.

2. Algorithms

In this section we introduce the algorithms in this study for discovering cause and effect relationships from observational data. We first introduce the Bayesian algorithms CD-B and CD-H that learn global CBN models and output the set of Y arcs which represent cause and effect relationships unconfounded by hidden variables. We then provide short descriptions of the PC, FCI and MMHC algorithms and the post-processing procedure that we use to identify the unconfounded causal arcs from the output of PC and MMHC.

2.1. CD-B algorithm

The CD-B algorithm first induces the Markov blanket (MB) of each node $X \in \mathbf{V}$ (where $\mathbf{V}$ is the set of domain variables) using the Bayesian Markov blanket induction (MBI) procedure (Mani, 2005). The MBI procedure finds the Markov blanket of a node X under the assumptions of Markov, faithfulness and large sample size. It uses a greedy forward and backward search in seeking the Markov blanket of X, which we denote MB(X). The set MB(X) is the estimated Markov blanket of X in a data generating network. From the MB of each node X the spouse nodes (parents of children of a node X are referred to as the spouse nodes of X) are excluded by a Bayesian dependence heuristic (Cooper, 1997) to obtain the set of parents and children of X (denoted as PC(X)). Using PC(X) we generate all possible DAGs such that the only arcs are from each parent to X and from X to each child. We refer to these DAGs as PC DAGs.

The key insight is that there are exactly 2^k such PC DAGs where $k = |\mathrm{PC}(X)|$. The highest scoring DAG from each node set $\mathrm{PC}(X) \cup X$ is used to get the $\mathbf{P}(X)$ and the $\mathbf{C}(X)$, that is, the parents and children of X respectively. Using $\mathbf{P}(X)$ from the highest scoring PC DAGs with two or more parents a global directed graph is constructed. Note that a PC DAG G with $\mathbf{P}(X)$ as set of parents and $\mathbf{C}(X)$ as set of children of the node X is unique (the only member of its Markov equivalence class) if $|\mathbf{P}(X)| \geq 2$. Since all the PC DAGs are scored, this step is exponential in the size of the set of parents and children. To the best of our knowledge the PC DAG method introduced here is the only Bayesian method to partition a set of parents and children ($\mathbf{P}(X) \cup \mathbf{C}(X)$) into the set of parents $\mathbf{P}(X)$ and the set of children $\mathbf{C}(X)$.

The global directed graph created using the set of parents may contain directed cycles. A directed cycle is a directed path starting from a node A and ends in node A after traversing two or more nodes. The cycles in the graph are broken iteratively by removing the "weakest" arc using a greedy search heuristic till all cycles are eliminated. The $\mathbf{C}(X)$ edges from the highest scoring PC DAGs with two or more parents are inserted based on a set of constraints (rules). The union of the edges of the highest scoring PC DAGs with less than two parents are inserted using a different set of constraints. As already mentioned, when the PC DAG has two or more parents it is unique, that is, it is the only member of its Markov equivalence class. On the other hand the PC DAGs

with less than two parents are not unique (there is at least one additional member in its Markov equivalence class). Hence the arcs belonging to the two categories of PC DAGs are inserted into the global DAG using different sets of constraints. The resulting DAG is used to identify all the Y arcs. The pseudocode for the CD-B algorithm is provided in Appendix A.1.

2.2. CD-H algorithm

The CD-H algorithm replaces the initial steps of the CD-B algorithm for finding the PC(X) with the MMPC algorithm (Tsamardinos et al., 2003, 2006). The MMPC uses a two-phase search procedure based on tests of independence/dependence. In the first phase of search a candidate set of parents and children called CPC is estimated which is a superset of the parents and children (PC) set. The second phase of the search procedure prunes the CPC set yielding the PC set. A proof of correctness and empirical results showing the validity of the MMPC algorithm are provided in (Tsamardinos et al., 2003). The subsequent steps of the CD-H algorithm are similar to CD-B.

2.3. PC algorithm

The PC algorithm takes as input a dataset D over a set of observed random variables **V**, a conditional independence test, and an α level of significance threshold for a test of statistical independence and then outputs an essential graph. PC also makes an assumption of *causal sufficiency*. This means that all the variables of the causal network are measured and there is no attempt to discover latent (hidden) variables. Hence PC is not designed to discover hidden variables that are common causes of any pair of observed variables. In the worst case, PC is exponential in the largest degree (size of the set of parents and children of a node) in the data generating DAG. See (Spirtes et al., 2000, page 84–85) for more details on the PC algorithm. The PC algorithm outputs both directed and undirected edges. A post-processing step (procedure YA) that we add is performed on the set of arcs to identify the Y structures. The pseudocode for procedure YA is given in Appendix A.1.3.

2.4. FCI algorithm

The FCI algorithm takes as input a dataset D over a set of random variables **V** and outputs a graphical model consisting of edges between variables that have a cause and effect interpretation. While the PC algorithm outputs only directed and undirected edges, the FCI algorithm outputs a richer set of edges to denote the presence of hidden (unmeasured) confounding variables and various levels of uncertainty in the orientation of the edges (Spirtes et al., 2000). The FCI algorithm can handle hidden variables and sample selection bias that are likely to be present in real-world datasets. It is possible to obtain causal relationships that are unconfounded by hidden variables from the partial ancestral graph (PAG) output of the FCI algorithm. The edges oriented as $A \rightarrow B$ in the FCI output can be interpreted as an unconfounded causal arc similar to a Y arc.

2.5. MMHC algorithm

The max-min hill-climbing (MMHC) Bayesian network structure learning algorithm is a hybrid algorithm that combines ideas from constraint-based and score-based methods (Tsamardinos et al., 2006). MMHC has been extensively evaluated on a variety of structure learning tasks from different datasets and outperformed PC, FCI, the Sparse Candidate, Optimal Reinsertion and the Greedy Equivalence Search algorithms. The MMHC algorithm estimates the set of parents and children of a node X denoted by PC(X) using the MMPC algorithm (Tsamardinos et al., 2003) to first obtain an undirected skeleton of the output graph. MMHC then uses greedy steepest-ascent TABU search and the Bayesian scoring measure BDeu (Heckerman et al., 1995) to orient the edges.

3. Experimental methods

In this section we describe the experimental methods used to evaluate our causal discovery approach. We used expert-defined CBNs to (1) generate data from those models, (2) apply the causal discovery algorithm to the data, and (3) evaluate the causal relationships output by the algorithm relative to the data generating CBNs that serve as gold standards. The output of the algorithm was compared with the data generating structure and scored as explained below. CD-B and CD-H algorithms were implemented in Matlab. The PC and FCI algorithms implemented in Tetrad IV (`http://www.phil.cmu.edu/projects/tetrad`) were used. The MMHC implementation in the Causal Explorer package (Aliferis et al., 2003b) was used. For PC, MMHC, CD-B and CD-H algorithms the Y arcs output by the algorithms were compared with the Y arcs of the data generating networks and for FCI the fully oriented arcs were used. Recall that a post-processing step was required for PC and MMHC algorithms to obtain the Y arcs. Precision, recall and F-measure were computed for the algorithms as follows:

Precision: (# of Y arcs correctly identified) / (# of total Y arcs output).

Recall: (# of Y arcs correctly identified) / (# of total Y arcs present in the data generating network).

F measure: (2 * recall * precision) / (recall + precision).

Four Bayesian networks built by domain experts in such varied fields as medicine, atmospheric sciences and agriculture were identified. These networks are Alarm (Beinlich et al., 1990), Hailfinder (Abramson et al., 1996), Barley (Kristensen and Rasmussen, 2002), and Munin (Andreassen et al., 1987). For causal discovery, we generated simulated training instances by stochastic sampling (Henrion, 1986). Varying sample sizes in the range of 1,000 to 20,000 instances were used in our causal discovery experiments. Table 1 gives the distribution of the nodes, arcs and Y structures for the various networks used in our study. Typically default parameters were used to run the algorithms with some adjustments made for uniformity. The PC algorithm was

Table 1: Nodes, arcs and Y structures in the Alarm, Hailfinder, Barley, and Munin networks

Category	Alarm	Hailfinder	Barley	Munin
Nodes	37	56	48	189
Arcs	46	66	84	282
Y structures	13	20	44	147

run with default parameters (significance level 0.05). CD-B was also run with default parameters (maximum MB size 12, dependency threshold for spouse elimination 0.9). CD-H was run with the following parameters: maximum size of conditioning set 10 and significance threshold 0.05. All the algorithms were run on each of the sample sizes using the ACCRE (Linux) cluster in Vanderbilt University consisting of x86 processors with 3.8 GB memory. Each job was assigned to a single processor with a time limit of 48 hours.

4. Results

The results presented below are based on sample sizes of 1K, 2K, 5K, 10K and 20K instances for each of the four domain datasets that were generated. We present a summary performance of all the four algorithms based on Y arcs present in all the data generating networks using precision, recall and F-measure as explained below. The aggregate results are presented based on the following two methods.

1. The various data generating networks are given equal weight in the analysis irrespective of the number of Y arcs.

2. The data generating networks are weighted by the number of Y arcs present in each network.

Table 2: Averages without FCI table weighted by # of Y arcs.

F-measure

Sample	CD-B	CD-H	PC	MMHC
1k	0.34	0.27	0.15	0.32
2k	0.41	0.30	0.22	0.33
5k	0.47	0.34	0.29	0.34
10k	0.47	0.38	0.30	0.52
20k	0.52	0.44	0.34	0.44

Precision

Sample	CD-B	CD-H	PC	MMHC
1k	0.79	0.28	0.90	0.43
2k	0.84	0.31	0.93	0.40
5k	0.82	0.31	0.95	0.40
10k	0.83	0.39	0.79	0.57
20k	0.81	0.39	0.65	0.45

Recall

Sample	CD-B	CD-H	PC	MMHC
1k	0.21	0.27	0.08	0.25
2k	0.27	0.29	0.13	0.29
5k	0.33	0.38	0.17	0.29
10k	0.33	0.38	0.18	0.48
20k	0.38	0.50	0.23	0.44

Altogether there were 224 YA in the four domain CBNs. The results presented are based on averages over all the four networks unless specified otherwise (see Tables 2, 3 and Figures 2, 3). The highest precision of 0.97 (equal weight) and 0.95 (weighted by #

Table 3: Averages without FCI table weighted equally.

F-measure

Sample	CD-B	CD-H	PC	MMHC
1k	0.32	0.41	0.20	0.40
2k	0.39	0.50	0.35	0.46
5k	0.49	0.51	0.45	0.38
10k	0.48	0.52	0.38	0.60
20k	0.54	0.55	0.45	0.58

Precision

Sample	CD-B	CD-H	PC	MMHC
1k	0.76	0.44	0.93	0.61
2k	0.78	0.52	0.93	0.59
5k	0.76	0.51	0.97	0.51
10k	0.76	0.54	0.78	0.69
20k	0.80	0.59	0.73	0.60

Recall

Sample	CD-B	CD-H	PC	MMHC
1k	0.29	0.39	0.14	0.32
2k	0.32	0.49	0.26	0.40
5k	0.39	0.57	0.36	0.32
10k	0.39	0.55	0.33	0.56
20k	0.44	0.57	0.43	0.58

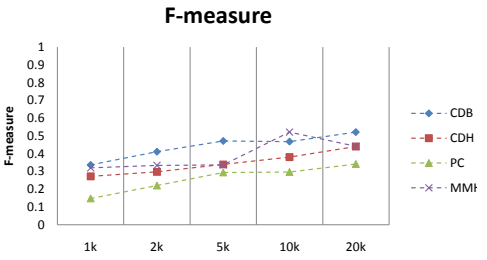


Figure 2: Averages without FCI graph weighted by # of Y arcs.

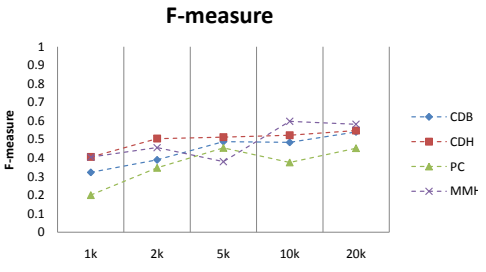


Figure 3: Averages without FCI graph weighted equally for all networks.

of Y arcs) were achieved at the sample size of 5,000 with PC. In general PC and CD-B had higher precision (≥ 0.65) across all the sample sizes tested. The best recall was obtained by MMHC (0.58 equally weighted) and CD-H (0.50 when weighted by # of Y arcs) with a sample size of 20,000. The best F-measure (when equally weighted) of 0.60 was achieved by MMHC (sample size 10K) followed by 0.55 for CD-H (sample size 20K). The best F-measure (weighted by # of Y arcs) of 0.52 was achieved by MMHC (sample size 10K) and CD-B (sample size 20K).

FCI could be run without going out of memory or exceeding the time limit of 48 hours on all sample sizes only for the Alarm dataset. Out of 20 experiments (5 sample sizes x 4 datasets), FCI ran out of memory in 9 cases (for all sample sizes in Barley network and for sample sizes 1K, 5K, 10K, and 20K in Munin network), ran out of time in 1 case (sample size 20K for Hailfinder network), and completed with results in the remaining 10 experiments (see Tables 4, 5 and Figures 4, 5). Based on F measure FCI performance is generally lower when compared with the other algorithms across all the sample sizes. Figure 6 shows total run time for all algorithms for the latter 10 experiments. As can be seen, FCI is the second slowest algorithm after CD-H. However, CD-H was able to complete with results in all 20 experiments, thus it is more useful for practitioners despite being one of the slowest algorithms in the comparison. CD-H runs slow primarily because it includes false positives in the estimated PC sets which makes PC DAG search much more computationally expensive. Table 6 provides the runtimes

of the various algorithms for the Alarm dataset. FCI runtime is an order of magnitude higher compared to the other algorithms on the Alarm dataset.

Table 4: Averages with FCI table weighted by # of Y arcs.

F-measure

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.48	0.39	0.30	0.50	0.22
2k	0.46	0.26	0.23	0.35	0.31
5k	0.53	0.63	0.68	0.51	0.49
10k	0.58	0.63	0.48	0.72	0.46
20k	0.77	0.81	0.88	0.93	0.75

Precision

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.71	0.41	0.86	0.80	0.42
2k	0.84	0.27	0.92	0.39	0.79
5k	0.70	0.53	1.00	0.64	0.65
10k	0.73	0.56	0.62	0.84	0.57
20k	0.77	0.79	0.92	0.87	0.63

Recall

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.36	0.36	0.18	0.36	0.15
2k	0.32	0.25	0.13	0.32	0.19
5k	0.42	0.79	0.52	0.42	0.39
10k	0.48	0.73	0.39	0.64	0.39
20k	0.77	0.85	0.85	1.00	0.92

Table 5: Averages with FCI table weighted equally.

F-measure

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.44	0.46	0.29	0.51	0.22
2k	0.47	0.52	0.46	0.54	0.42
5k	0.53	0.75	0.69	0.51	0.50
10k	0.59	0.72	0.49	0.73	0.49
20k	0.77	0.81	0.88	0.93	0.75

Precision

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.62	0.54	0.90	0.75	0.42
2k	0.77	0.52	0.90	0.62	0.77
5k	0.67	0.66	1.00	0.68	0.65
10k	0.72	0.65	0.62	0.83	0.65
20k	0.77	0.79	0.92	0.87	0.63

Recall

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.45	0.41	0.18	0.42	0.15
2k	0.39	0.51	0.32	0.49	0.32
5k	0.48	0.88	0.53	0.43	0.41
10k	0.53	0.83	0.42	0.70	0.41
20k	0.77	0.85	0.85	1.00	0.92

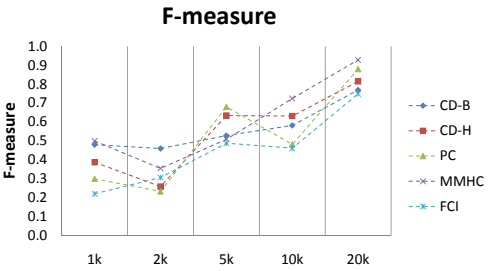


Figure 4: Averages with FCI graph weighted by # of Y arcs.

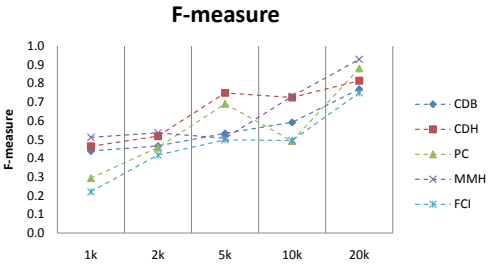


Figure 5: Averages with FCI graph weighted equally for all networks.

We also present results of causal discovery in the presence of hidden variables based on randomly assigning "hidden" status to 25% of the variables for the Alarm dataset (see Tables 7 and 8). The results presented in Table 8 are averaged over 5 such random Alarm networks with with 25% of the nodes hidden. The results show that there is a degradation in performance for all the algorithms when a subset of the variables are unobserved (see Table 9). CD-B and FCI appear more robust in the presence of hidden variables when compared to CD-H, PC and MMHC based on the magnitude of reduction in F measure when hidden variables are introduced. Additional evaluation is needed to understand the effect of hidden variables for causal discovery from observational data.

Table 6: Alarm original network runtimes in minutes for all the algorithms.

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.20	0.30	0.10	0.10	5.00
2k	0.30	0.30	0.10	0.10	1.00
5k	0.50	0.50	0.10	0.20	5.00
10k	0.70	0.70	0.10	0.20	5.00
20k	1.10	0.90	0.20	0.40	58.00

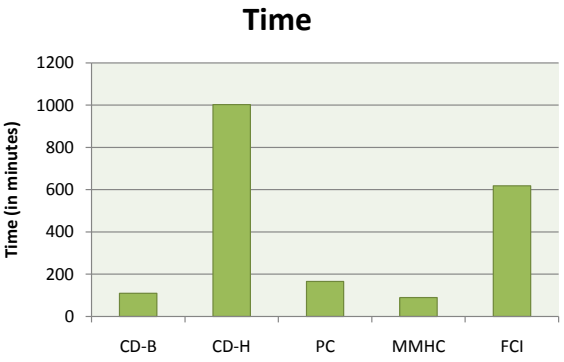


Figure 6: Runtimes for all the algorithms over all the four datasets based on FCI completion.

Table 7: Alarm original.

F-measure

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.79	0.73	0.27	0.78	0.21
2k	0.77	0.92	0.70	0.85	0.60
5k	0.77	0.93	0.76	0.46	0.55
10k	0.77	0.90	0.58	0.93	0.60
20k	0.77	0.81	0.88	0.93	0.75

Precision

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.73	0.89	1.00	0.90	0.33
2k	0.77	0.92	1.00	0.85	0.86
5k	0.77	0.88	1.00	0.46	0.67
10k	0.77	0.81	0.64	0.87	0.86
20k	0.77	0.79	0.92	0.87	0.63

Recall

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.85	0.62	0.15	0.69	0.15
2k	0.77	0.92	0.54	0.85	0.46
5k	0.77	1.00	0.62	0.46	0.46
10k	0.77	1.00	0.54	1.00	0.46
20k	0.77	0.85	0.85	1.00	0.92

Table 8: Alarm 75 percent observed.

F-measure

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.60	0.38	0.22	0.49	0.19
2k	0.60	0.50	0.32	0.43	0.40
5k	0.51	0.65	0.49	0.27	0.43
10k	0.57	0.54	0.54	0.47	0.58
20k	0.58	0.61	0.56	0.57	0.66

Precision

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.75	0.53	0.70	0.75	0.28
2k	0.73	0.57	0.80	0.52	0.41
5k	0.67	0.65	1.00	0.38	0.51
10k	0.64	0.58	0.69	0.53	0.51
20k	0.65	0.60	0.80	0.66	0.58

Recall

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	0.52	0.31	0.13	0.38	0.15
2k	0.52	0.46	0.21	0.38	0.39
5k	0.42	0.67	0.35	0.22	0.38
10k	0.52	0.52	0.45	0.44	0.67
20k	0.54	0.62	0.47	0.54	0.79

Table 9: Alarm 75 performance degradation.

Sample	CD-B	CD-H	PC	MMHC	FCI
1k	-0.18	-0.35	-0.04	-0.30	-0.02
2k	-0.17	-0.43	-0.38	-0.42	-0.20
5k	-0.26	-0.28	-0.27	-0.19	-0.11
10k	-0.20	-0.36	-0.05	-0.45	-0.02
20k	-0.18	-0.21	-0.32	-0.36	-0.09

5. Discussion

In this section we discuss the results and present the implications of our research for discovering causal relationships from observational data. This research has highlighted the role of Y structures for causal discovery from observational data and introduced two new algorithms CD-B and CD-H based on identification of Y structures.

Precision varied within a narrow range of 0.76 to 0.84 for CD-B and between 0.65 and 0.97 for PC (see Tables 2 and 3). The relatively narrow precision range for the different sample sizes combined with a monotonic increase in recall throughout the sample range shows that the performance of CD-B and PC is robust across a wide range of sample sizes. In general precision values are higher compared to recall values for all the sample sizes except for the CD-H algorithm. Note that a higher precision translates to lower number of false positives even though some causal relationships may not be reported. A desirable goal in causal discovery is to keep the proportion of false positives low even if it entails a trade-off in terms of recall.

FCI and CD-H had longer runtimes when compared with PC, MMHC and CD-B. It is possible to use symmetry correction in the MMPC step of the CD-H algorithm to reduce the number of false positives in the PC set and decrease runtime.

The causal discovery framework that we presented for identifying direct causal relationships is dependent on the presence of Y structures in the data generating process. The two medical (Alarm, Munin) and two non-medical (Hailfinder, Barley) networks that were used to generate data had varying numbers of Y structures. These networks were created by domain experts capturing the probabilistic dependencies and independencies in the domain. Hence it seems plausible that Y structures occur in the data generating process of many real-world domains. Presence of Y structures have also been shown in a real world infant birth and death dataset (Mani and Cooper, 2004).

CD-B and CD-H are unique in differentiating the set of parents and the set of children of a node X from the union of the set of parents and children of X. Identification of the parents and children of a node will give us the candidate set of direct causes and the candidate set of direct effects of a node. Due to the presence of hidden variables all the parents cannot be interpreted as direct causes and all the children cannot be interpreted as direct effects. However, the candidate set of parents and children can be used to rule out hypothesized causes or effects. Also, when experimental studies are feasible the candidate sets can act as the first filter and provide the experimenter with a preliminary

set of potential causes and effects. Moreover, the set of parents or the set of children of a node completely specify a directed acyclic graph which can be used to approximate the data generating model.

5.1. Related work

The most related algorithm to the CD-B algorithm is the BLCD (Mani and Cooper, 2004; Mani, 2005). BLCD estimates the Markov blanket of a variable and uses it for the identification of Y structures from sets of four variables. BLCD does not specifically identify the sets of parents and children from the Markov blanket.

Aliferis et al. have introduced HITON, an algorithm to determine the MB of an outcome variable (Aliferis et al., 2003a). Tsamardinos et al. have described an algorithm called MMMB and they discuss that since the MB contains direct causes and direct effects of a variable X, the MB has causal interpretability (Tsamardinos et al., 2003). Note that both HITON and MMMB do not specifically distinguish between causes and effects of a node; however, they do output the variables that have direct edges during the operation of the algorithm. Additional processing (or experimentation) of HITON and MMMB output is required to determine causal directionality.

5.2. Limitations and future work

There are two main types of limitations of this work. The first set of limitations results from the framework and assumptions we have chosen for causal discovery. The second set of limitations is due to the specifics of the algorithm and the experimental methods that were used.

The CBN framework imposes a directed acyclic graph structure on all causal phenomena. Discovering causal mechanisms that incorporate feedback cycles can be problematic unless time is represented explicitly and cycles are "unfolded" to provide a DAG structure (Cooper, 1999). The causal discovery approach we have taken is not complete in the sense that we can discover only causal relationships represented in nature as Y structures. The algorithms also currently requires that the modeled variables be discrete.

The evaluation measures of precision, recall and F measure that were used are structural. Hence the evaluation of the purported causal relationships were structural, leaving out the parametric components. That is, we evaluated how well the algorithm can discover the presence of a causal influence, but leave to future work the characterization of how well the algorithm captures the functional relationships among the causes and effects.

We plan to apply the CD-B and CD-H algorithms to real-world datasets as part of our future work.

Acknowledgments

We thank professor Greg Cooper for helpful discussions. We thank Yerbolat Dosbayev for implementing CD-B and CD-H and Yukun Chen for running the experiments and

generating the results presented in the paper. We also thank the anonymous reviewers for their critical comments and suggestions for improving the paper.

References

Bruce Abramson, John Brown, Ward Edwards, Allan Murphy, and Robert L. Winkler. Hailfinder: A Bayesian System for Forecasting Severe Weather. *International Journal of Forecasting*, 12:57–71, 1996.

Constantin F. Aliferis, Ioannis Tsamardinos, and Alexander Stanikov. HITON, A novel markov blanket algorithm for optimal variable selection. In *Proceedings of the AMIA Fall Symposium*, 2003a.

Constantin F. Aliferis, Ioannis Tsamardinos, Alexander Stanikov, and Laura E. Brown. Causal Explorer: A causal probabilistic network learning toolkit for biomedical discovery. In *Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, 2003b.

Steen Andreassen, Marianne Woldbye, Bjorn Falck, and Stig K. Andersen. MUNIN — A causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 366–372, San Mateo, CA, 1987. Morgan Kaufmann.

Ingo A. Beinlich, H.J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, pages 247–256, London, 1990. Chapman and Hall.

Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

Gregory F. Cooper. An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks. In Clark Glymour and Gregory F. Cooper, editors, *Computation, Causation, and Discovery*, pages 3–62. MIT Press, Cambridge, MA, 1999.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20 (3):197–243, 1995.

Max Henrion. Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Proceedings of the 2nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-86)*, New York, NY, 1986. Elsevier Science Publishing Company, Inc.

K. Kristensen and I.A. Rasmussen. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33:197–217, 2002.

Subramani Mani. *A Bayesian Local Causal Discovery Framework*. PhD thesis, University of Pittsburgh, 2005.

Subramani Mani and Gregory F. Cooper. Causal discovery using a Bayesian local causal discovery algorithm. In M. Fieschi et al. editor, *Proceedings of MedInfo*, pages 731–735. IOS Press, 2004.

Subramani Mani, Peter Spirtes, and Gregory F. Cooper. A theoretical study of Y structures for causal discovery. In Rina Dechter and Thomas S. Richardson, editors, *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 314–323, Corvallis, OR, 2006. AUAI Press.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In S.A.Solla, T.K.Leen, and K.R.Muller, editors, *Advances in neural information processing systems*, volume 12, pages 505–511, Cambridge, MA, 2000. MIT Press.

Andrew Moore and Weng-Keen Wong. Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In T. Fawcett and N. Mishra, editors, *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, pages 552–559, Menlo Park, California, August 2003. AAAI Press.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 2nd edition, 1991.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Stanikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the 9th CAN SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678, 2003.

Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65: 31–78, 2006.

Appendix A. CD-B Pseudocode

In this section we provide the pseudocode for the CD-B algorithm and the details of the various procedures called by CD-B, specifically the Markov blanket induction (MBI) procedure and the Y arc (YA) finding procedure.

A.1. CD-B algorithm

/* Note: When the PC DAG has 0 or 1 parent G_{max} is not unique. We pick any G_{max} PC DAG from its equivalence class. This implies that in the data generating DAG the edges of such PC DAGs can have either $A \rightarrow B$ or $A \leftarrow B$ orientation. */

Input : Dataset D and the set of variables **X**.

Output : Pairwise causal influences of the form $A \rightarrow B$ representing Y arcs.

The following are the steps of the algorithm:

1. For each variable $X \in \mathbf{X}$ estimate MB(X) using the Bayesian MB induction (MBI) procedure.
2. For each variable $X \in \mathbf{X}$ DO
   (a) Update MB(X). If A is in the MB of B, but B is not in the MB of A, we add B to the MB of A.
   (b) Remove the spouse nodes from MB(X) to obtain PC(X). Any node independent of X is excluded from MB(X). Let **B** denote PC(X).
   (c) From $\mathbf{B} \cup X$ generate all possible DAGs such that the only arcs are from each parent to X and from X to each child. Let this set of DAGs be **G**.
   (d) From the set of DAGs **G** identify the maximally scoring DAG G using the BDeu scoring measure (Heckerman et al., 1995). Let this DAG be G_{max}. If there is a tie for G_{max}, it is broken randomly.
   (e) If the G_{max} has 2 or more parents mark the Pa(X) and Ch(X) as oriented (Pa^o(X) and Ch^o(X)).
   (f) If the G_{max} has less than 2 parents mark the Pa(X) and Ch(X) as unoriented (Pa^u(X) and Ch^u(X)).
   (g) OD
3. Using Pa^o(X) of all the nodes with 2 or more parents construct a global directed graph G'. G' may contain cycles.
4. Construct DAG G from G' using Procedure RC.
5. Let **E** be the union of all the arcs from Ch^o(X) of all the nodes with 2 or more parents.
6. While **E** not $\emptyset$, insert the edge $e \in \mathbf{E}$ in G iff it satisfies the following conditions (a), (b) and (c).
   (a) Not already present in G.
   (b) No cycle is introduced in G.

(c) Insert the e that maximizes the score for G.

(d) Remove e from **E**.

(e) OD

7. Let **E** be the union of all the edges ignoring direction from $\text{Pa}^u(X)$ and $\text{Ch}^u(X)$ of all the nodes with less than 2 parents.

8. While **E** not $\emptyset$, insert the edge $e \in \mathbf{E}$ ($A \rightarrow B$ or $A \leftarrow B$) in G iff it satisfies the following conditions (a), (b) , (c) and (d).

   (a) Not already present in G ignoring direction.

   (b) No new V structure is introduced in G.

   (c) No cycle is introduced in G.

   (d) Insert the e that maximizes the score for G.

   (e) Remove e from **E**.

   (f) OD

9. Remove cycles from G using Procedure RC.

10. Identify all the Y arcs (YA) in G using Procedure YA and output the YA.

A.1.1. PROCEDURE MBI

We derive an estimate of the Market blanket (MB) of a node (designated as **H**) using a greedy forward and backward heuristic search which we refer to as the *Procedure MBI*.

Input: Dataset D over observed random variables **X** and a variable $X \in \mathbf{X}$.

Output: Markov blanket of X in a data generating network, which we denote MB(X) i.e. the union of estimated parents, children and spouses (parents of children) of node X, under the assumption the data is being generated by a faithful Bayesian network on measured variables **X**.

The following are the steps of the MBI procedure:

- Identify the set **H'** $\subseteq \mathbf{X} \setminus X$ that maximizes the BDeu score for the structure **H'** $\rightarrow X$ based on a one-step forward greedy search.

- Perform a one step backward greedy search that prunes **H'** to yield set **H** $\subseteq$ **H'** that maximizes the score for the structure **H** $\rightarrow X$.

- Output **H** which represents MB(X).

A.1.2. PROCEDURE RC

This procedure removes the cycles from a directed graph. The "weakest" arc is removed iteratively till all cycles are eliminated.

Input: A directed graph G'.

Output: A directed acyclic graph G.

The following are the steps of the procedure:

1. Check for cycle(s) in G'. If no cycle assign G' to G and return G.
2. Identify all the arcs forming cycle(s). Let these set of arcs be **E**.
3. Identify the weakest arc E $\in$ **E** by iteratively removing each arc from **E** and scoring the graph using the BDeu scoring measure. The arc causing the least reduction in the BDeu score is determined to be the weakest.
4. Remove E from G'. Let the resulting graph be G'. GOTO Step 1.

A.1.3. PROCEDURE YA

We identify all the unique Y arcs (YA) in a DAG G using this procedure. The procedure looks for all the embedded Y structures (EYS) in G. We say that G contains an *embedded* Y structure involving the variables W_1, W_2, X and Z, iff all and only the following adjacencies hold among the variables W_1, W_2, X and Z ($A \square B$ means that there is no arc between A and B):

- $W_1 \square W_2$; $W_1 \square Z$; $W_2 \square Z$
- $W_1 \rightarrow X$; $W_2 \rightarrow X$; $X \rightarrow Z$

Input: A DAG G and a set of nodes **X** in G.

Output: A set of Y arcs denoted as **Y**.

Initialize set of YA as $\mathbf{Y} := \{\}$.
For each $X \in \mathbf{X}$
DO
    Determine Pa(X) for X.
    If $|\text{Pa}(X)| \leq 1$
        Continue /* Next iteration */
    Determine Ch(X) for X.
    If $|\text{Ch}(X)| < 1$
        Continue /* Next iteration */

    /* Look for Y structure */

```
    For each pair of parents W1, W2 of X
    DO
        If W1 and W2 are adjacent then Continue
        For each child Z ∈ Ch(X)
        DO
            If (W1, Z) or (W2, Z) adjacent then Continue
            If (X → Z) ∉ Y
                Y := Y ∪{X → Z}
        OD
    OD
OD
Return Y
```

JMLR: Workshop and Conference Proceedings 6:137–146, 2010 NIPS 2008 workshop on causality

When causality matters for prediction: investigating the practical tradeoffs

Robert E. Tillman RTILLMAN@ANDREW.CMU.EDU
Department of Philosophy and Machine Learning Department, School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, United States

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

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Recent evaluations have indicated that in practice, general methods for prediction which do not account for changes in the conditional distribution of a target variable given feature values in some cases outperform causal discovery based methods for prediction which *can* account for such changes. We investigate some possibilities which may explain these findings. We give theoretical conditions, which are confirmed experimentally, for when particular manipulations of variables should not affect predictions for a target. We then consider the tradeoff between errors related to causality, i.e. not accounting for changes in a distribution after variables are manipulated, and errors resulting from sample bias, overfitting, and assuming specific parametric forms that do not fit the data, which most existing causal discovery based methods are particularly prone to making.

Keywords: causal discovery, prediction, interventions

1. Introduction

Most methods in machine learning are intended primarily for *prediction*. Given *training data* for a target variable T to be predicted and a set of associated *predictor* variables $\mathbf{X}$, the goal is to use the training data to learn a prediction function $T = f(\mathbf{X})$ that can be used to predict values for the target given values for the predictor variables, assuming the conditional distribution $P(T|\mathbf{X})$ does not change after the training data is collected. In general, we are not concerned with whether the prediction function actually depicts true causal relationships between variables in the underlying data generating mechanism; we care only whether it makes accurate predictions.

The advantage of *causal discovery* methods is that they can be be used to learn models that depict true data generating mechanisms. We can discover particular causal relationships between variables to determine which variables should be manipulated when setting policies to achieve a desired effect. We can use the resulting *causal mo-*

dels to predict the effects of such manipulations or make predictions for a target variable when we have data for the predictor variables even if some variables have been manipulated since the training data was collected.

Evaluations of causal discovery methods have focused primarily on how closely the resulting causal models resemble true data generating mechanisms obtained either through simulations or data from controlled experiments. There has been little focus on how accurate predictions made using causal discovery methods after variables are manipulated are relative to known values for a predicted variable, i.e. using test data from the manipulated population, which is the primary means for evaluating most other methods in machine learning. Recent results from a causality challenge[1] have raised questions as to whether existing causal discovery methods are useful for making such predictions. In the challenge, some participants used prediction methods which ignored causality to predict a target variable after predictor variables were manipulated, i.e.. applying support vector machines trained using the unmanipulated data to make predictions for the manipulated data without any adjustments to account for the manipulations, and in some cases achieved results that were better than any of the participants who used causal discovery based methods for prediction.

One possible explanation for these results is that there is a noticeable tradeoff when using causal discovery methods to make such predictions: while causal discovery methods may make the proper adjustments to account for the change in a distribution after variables are manipulated, prediction with causal discovery based methods may result in significant errors due to overfitting and sampling bias as well as parametric assumptions, i.e. linearity, Gaussianity, which do not hold. There is nothing inherent in causal models or causal inference that requires parametric assumptions that are more restrictive than other machine learning methods; however, most[2] existing causal discovery algorithms do require such assumptions. Thus, most causal discovery methods for prediction may result in considerably more of this second type of error than many other nonparametric methods in machine learning, such as support vector machines. Furthermore, in many cases, manipulating a particular variable in a causal system will have no effect on the predicted value of a particular target, e.g. if the manipulated variable is conditionally independent of the target variable given the set of predictor variables. Thus, if certain parametric assumptions made by causal discovery algorithms do not hold for some data, then we should expect nonparametric methods for predictions and methods which make less strict parametric assumptions to outperform causal discovery based methods for predictions even for some cases where variables are manipulated after the training data is collected.

In this paper, we begin to investigate this tradeoff. We review the relevant terminology in section 2. In section 3, we present theoretical conditions which distinguish manipulations which do affect predictions for a target from those which do not and demonstrate how causal discovery methods used for prediction can account for the

1. See http://www.causality.inf.ethz.ch/challenge.php for details.
2. There have been several recent proposals which require less restrictive parametric assumptions, i.e. Shimizu et al. (2006), Hoyer et al. (2008), Hoyer et al. (2009).

change in distribution. In section 4, we then experimentally test these conditions using synthetic data to confirm that they at least hold in the cases favorable for casual discovery methods. Conclusions are offered in section 5.

2. Formal preliminaries

We first introduce some terminology. A *directed graph* $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a set of nodes $\mathcal{V}$, which represent variables, and a set of directed edges $\mathcal{E}$ connecting distinct nodes. For a node $V \in \mathcal{V}$, $\mathbf{Pa}_V^{\mathcal{G}}$ refers to the set of nodes that are parents of V (nodes with an edge directed into V), $\mathbf{Ch}_V^{\mathcal{G}}$ refers to the children of V (nodes with an edge directed out of V), and $\mathbf{Co}_V^{\mathcal{G}}$ refers to the coparents (spouses) of V (parents of children of V other than V). A *trail* in $\mathcal{G}$ is a sequence of nodes such that each adjacent pair in the sequence is connected by an edge (ignoring directions), and no node appears more than once in the sequence. A trail is a *directed path* if every edge points in the same direction. $\mathcal{G}$ is a *directed acyclic graph* (DAG) if for every pair $\{X, Y\} \subseteq \mathcal{V}$, there are not directed paths from both X to Y and Y to X (no directed cycles). X is an *ancestor* (*descendant*) of Y if there is a directed path from X to Y (Y to X). A *v-structure* (*collider*) is a triple of nodes $\langle X, Y, Z \rangle$ such that X and Z are parents of Y.[3] A trail is *active* given a conditioning set $\mathbf{C} \subseteq \mathcal{V}$ if (i) for every v-structure $\langle X, Y, Z \rangle$ in the trail either $Y \in \mathbf{C}$ or some descendant of Y is in $\mathbf{C}$ and (ii) no other node in the trail is in $\mathbf{C}$. For disjoint sets of nodes, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ if and only if there are no active trails between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

A *Bayesian network* $\mathcal{B}$ is a pair $\langle \mathcal{G}, \mathcal{P} \rangle$, where $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a DAG and $\mathcal{P}$ is a joint probability distribution over the variables represented by the nodes in $\mathcal{V}$ such that $\mathcal{P}$ can be factored as follows:

$$\mathcal{P}(\mathcal{V}) = \prod_{V \in \mathcal{V}} P(V | \mathbf{Pa}_V^{\mathcal{G}})$$

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 $\mathcal{P}$ (Pearl, 1988). For disjoint sets of nodes, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ in $\mathcal{V}$, $\mathcal{P}$ is *faithful* to $\mathcal{G}$ if $\mathbf{X}$ 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). $\mathcal{B}$ is said to be a *causal* Bayesian network if an edge from X to Y indicates that X is a direct cause of Y relative to $\mathcal{V}$. When performing causal inference, it is generally assumed that the distribution over the observed variables $\mathcal{P}$ factors according to a DAG $\mathcal{G}$ in a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ and $\mathcal{P}$ is faithful to $\mathcal{G}$. In this paper, we assume that there are no unmeasured common causes of variables in $\mathcal{V}$.

For a Bayesian network $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$, where $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, a *Markov blanket* for some node $V \in \mathcal{V}$ in $\mathcal{G}$, $\mathbf{MB}_V^{\mathcal{G}}$, is a minimal set of variables in $\mathcal{V}/\{V\}$ such that V is conditionally independent of $\mathcal{V}/\{\mathbf{MB}_V^{\mathcal{G}} \cup \{V\}\}$ given $\mathbf{MB}_V^{\mathcal{G}}$. If $\mathcal{P}$ is faithful to $\mathcal{G}$, then $\mathbf{MB}_V^{\mathcal{G}} = \mathbf{Pa}_V^{\mathcal{G}} \cup \mathbf{Ch}_V^{\mathcal{G}} \cup \mathbf{Co}_V^{\mathcal{G}}$, for any $V \in \mathcal{V}$ (Pearl, 1988).

3. We are using the definition given in Koller and Friedman (2008). Other sources use v-structure to refer to only such triples where X and Z are not adjacent (an *immorality* or *unshielded collider*).

We represent manipulations of variables $\mathbf{Z} \subseteq \mathcal{V}$ in a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ where $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, by forming the new DAG $\mathcal{G}(Policy(\mathbf{Z}))$, where we introduce a new exogenous node (node without parents) $Policy(Z)$ to $\mathcal{G}$ that is a parent of only Z, for each $Z \in \mathbf{Z}$. For disjoint sets of non-policy nodes $\mathbf{X}$ and $\mathbf{Y}$, a conditional distribution $P(\mathbf{Y}|\mathbf{X})$ is *invariant* under the manipulation of $\mathbf{Z}$ if $P(\mathbf{Y}|\mathbf{X})$ is the same when the variables in $\mathbf{Z}$ are manipulated and the variables in $\mathbf{Z}$ are unmanipulated. If **Policy(Z)**, the set of all policy nodes, is d-separated from $\mathbf{Y}$ given $\mathbf{X}$ in $\mathcal{G}(Policy(\mathbf{Z}))$, then $P(\mathbf{Y}|\mathbf{X})$ is invariant under manipulation of $\mathbf{Z}$ (Spirtes et al., 2000). $\mathcal{P}_M$, the distribution resulting from manipulating the variables in $\mathbf{Z}$, factors according to the DAG $\mathcal{G}_M$, which is $\mathcal{G}$ changed by removing every edge that is directed into some $Z \in \mathbf{Z}$ (Spirtes et al., 2000).

3. Invariance of predictions under manipulations

In some cases, it is obvious that manipulations of certain variables will not affect predictions for other variables. Consider the simple case of two variables X and Y, where X causes Y and there are no common causes of X and Y. If X is manipulated, this does not change the distribution of $P(Y|X)$, which produces the Bayes optimal prediction for Y. Thus, a classifier or regression method trained using data from a population where X is not manipulated should correctly make predictions for Y using test data from a population where X is manipulated without any adjustments to account for the manipulation. The following theorem distinguishes the more complicated cases where manipulations do not affect predictions for a target variable from those where manipulations do affect predictions.

Theorem 1 *Let $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ be a Bayesian network over variables $\mathcal{V}$, $T \in \mathcal{V}$ a target predicted variable, $\mathbf{X} \subseteq \mathcal{V}$ a set of predictor variables, and $\mathbf{Z} \subseteq \mathcal{V}$ a set of variables that are manipulated. If $\forall Y \in \mathbf{Y}$, $Y \neq T$ and $Y \notin \boldsymbol{Ch}_T^{\mathcal{G}}$, then $P(T|\mathbf{X})$ is invariant under the manipulation.*

Proof Let P be a trail between T and $Policy(Y)$ for some $Y \in \mathbf{Y}$. If P is into T, e.g. some node Z in P is a parent of T, then Z and T do not form a v-structure with some parent or child of Z in P and $Z \in \mathbf{X}$ since $Z \in \mathbf{Pa}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. If P is out of T, e.g. some node Z in P is a child of T, and some child U of Z is in P, then $\langle U, Z, T \rangle$ is not a v-structure and $Z \in \mathbf{X}$ since $Z \in \mathbf{Ch}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. If P is out of T and some parent U of Z other than T is in P, then U and Z do not form a v-structure with some parent or child of U and $Z \in \mathbf{X}$ and $U \in \mathbf{X}$ since $Z \in \mathbf{Pa}_T^{\mathcal{G}}$ and $U \in \mathbf{Co}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. This exhausts all cases so $Policy(Y)$ and T are d-separated given $\mathbf{X}$. ■

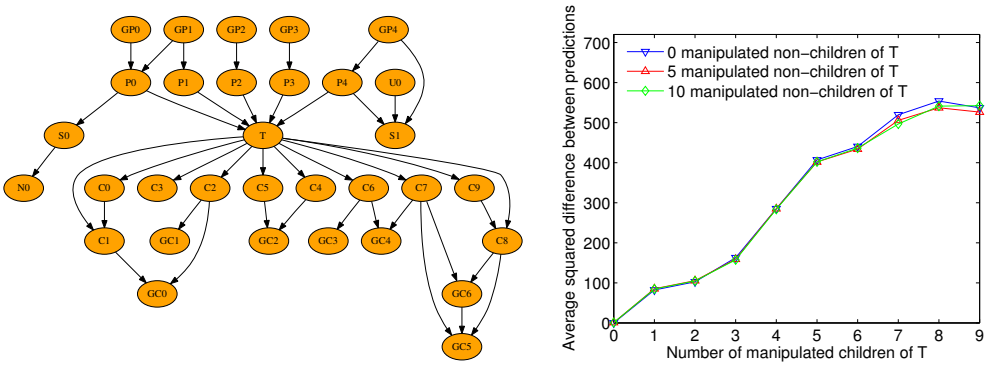
Thus, as long as a set of predictors includes the Markov blanket for a target node, prediction will be unaffected by any manipulation which does not change the value of a child of the target, even if the manipulation changes the value of another variable in the Markov blanket. In such cases, causal knowledge will not improve the accuracy of predicted values in any way (though we will not know whether prediction is affected by a manipulation unless we know the causal relationships in the underlying data generating

mechanism). While this is a straightforward result, it is important in practice. Most methods for prediction which do an explicit or implicit feature selection will likely assign high weight to (at least most of) the features in the Markov blanket. Since in most cases, the children of a particular target will consist of only a small percentage of the nodes in a graph, it is unlikely that the target will have many children that are manipulated to unlikely values, which would lead to high error in prediction. In many cases, errors resulting from manipulated children of a target that are used as features may be negligible and canceled out by other gains made by a prediction method that combats overfitting well or does not make restrictive parametric assumptions. In other cases where we may expect many children to be manipulated to unlikely values, we can use causal knowledge to select the correct set of predictors for the manipulated distribution and avoid errors resulting from manipulated children of a target that are used as features.

Theorem 2 *Let $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ be a Bayesian network over variables $\mathcal{V}$, $T \in \mathcal{V}$ a target predicted variable, $\boldsymbol{X} \subseteq \mathcal{V}$ a set of predictor variables, and $\mathbf{Z} \subseteq \mathcal{V}$ a set of variables that are manipulated. If $\boldsymbol{X} = \boldsymbol{MB}_T^{\mathcal{G}_M}$, then $P(T|\boldsymbol{MB}_T^{\mathcal{G}_M})$ is invariant under the manipulation of $\mathbf{Z}$ if $\forall Z \in \{\mathbf{Z} \cap \boldsymbol{Ch}_T^{\mathcal{G}}\}$, Z is not an ancestor of some $X \in \boldsymbol{Ch}_T^{\mathcal{G}}$ such that $X \notin \mathbf{Z}$.*

Proof If $P(T|\mathbf{MB}_T^{\mathcal{G}_M})$ is not invariant under the manipulation of $\mathbf{Z}$, then there is an active trail R between T and $Policy(Z)$ for some $Z \in \mathbf{Z}$ given $\mathbf{MB}_T^{\mathcal{G}_M}$ in $\mathcal{G}(Policy(\mathbf{Z}))$. Let X be the node in R connected to T, and U the node connected to X in R other than T. If X is a parent of T, then $X \in \mathbf{MB}_T^{\mathcal{G}_M}$ and $\langle U, X, T \rangle$ is not a v-structure so R is not active. Thus, X is a child of T. We have the following 2 cases. Case 1: $\langle U, X, T \rangle$ is not a v-structure. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so $X \notin \mathbf{MB}_T^{\mathcal{G}_M}$. $X \in \mathbf{Ch}_T^{\mathcal{G}}$ and $X \notin \mathbf{MB}_T^{\mathcal{G}_M}$ so $X \in \mathbf{Z}$. $Policy(Z)$ and T are both parents in R, so X is an ancestor of the middle node of a v-structure in R. R is active so the middle node of this v-structure either is contained in or has a descendant in $\mathbf{MB}_T^{\mathcal{G}_M}$. Thus, X is an ancestor of some $W \in \mathbf{MB}_T^{\mathcal{G}_M}$. Case 2: $\langle U, X, T \rangle$ is a v-structure in R. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so $U \notin \mathbf{MB}_T^{\mathcal{G}_M}$. $U \in \mathbf{Co}_T^{\mathcal{G}}$ and $U \notin \mathbf{MB}_T^{\mathcal{G}_M}$ so $X \in \mathbf{Z}$. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so X is either contained in or has a descendant in $\mathbf{MB}_T^{\mathcal{G}_M}$. Thus, for cases 1 and 2, $X \in \mathbf{Z}$ and either X is an ancestor of some $W \in \mathbf{MB}_T^{\mathcal{G}_M}$ or $X \in \mathbf{MB}_T^{\mathcal{G}_M}$. If $W \in \mathbf{Pa}_T^{\mathcal{G}}$ ($X \in \mathbf{Pa}_T^{\mathcal{G}}$), then there is a directed path from T to W (X) and a directed path from W (X) to T. $\mathcal{G}$ is acyclic so W (X) $\in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and either W (X) $\notin \mathbf{Z}$ or W (X) is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$. But since $X \in \mathbf{Z}$, there are only three cases: (i) $X \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and X is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$, (ii) $W \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and $W \notin \mathbf{Z}$ and (iii) $W \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and W is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$. In all three cases X is an ancestor of some $S \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $S \notin \mathbf{Z}$. ■

As long as we are using the the Markov blanket for the manipulated structure, e.g. after policy nodes are added, as our set of predictors, which requires causal knowledge, predictions will not be affected by manipulated children unless there is some manipulated child of the target that is an ancestor of an unmanipulated child of the target. When it is the case that a manipulated child of the target is an ancestor of some unmanipulated child, we can still make predictions using the Markov blanket for the manipulated



(a) (b)

Figure 1: (a) causal structure used in the simulations, (b) averaged squared differences in predictions using the ground truth regression equations for the manipulated and unmanipulated datasets

structure, but a correction[4] needs to be made to subtract out the influence from the manipulated variable. In practice, however, failure to make this correction usually has little effect on predictions. The importance of this theorem is that it allows us to select the correct set of predictors to account for changes in a distribution resulting from manipulations, regardless of what variables are manipulated. Thus, we should expect causal discovery based methods for prediction to perform increasingly better than methods for prediction which ignore causality as we increase the number of children of a target variable that are manipulated if the parametric assumptions made by the causal discovery methods are reasonable for some given data and error due to overfitting and sampling bias are reasonably low, even if such instances are not representative of the majority of cases. This hypothesis is evaluated in the next section.

4. Experimental results

We first constructed the graph of the causal environment around the target node T shown in figure 1a to use in the following experiments. The graph was constructed to be similar to the causal environment we learned for the target variable in one of the challenge datasets. We chose random linear Gaussian parameters for the variables in this structure and used forward sampling to generate a synthetic training dataset of size $N = 1000$. We then generated test datasets of sizes $N = 1000$ after various manipulations were made to variables in the structure. We first manipulated 0, 5, and 10 random

4. To compute this correction, we (i) replace the original equations for predicting the manipulated variables with the new manipulated equations (e.g. if Z is manipulated to 3, then the new equation is Z = 3), (ii) calculate the implied covariance matrix for the manipulated set of equations, and (iii) use the implied covariance matrix for the manipulated set of equations to calculate the equation for predicting the target variable in the usual way.

non-children of T and then for each we also manipulated from 0 to 9 children of T. In our simulations, we manipulated variables by setting each manipulated variable's dependency with all parents to 0, its mean to an unlikely value, and its variance to a small value. Each simulation described below was repeated 100 times and the results averaged.

Before considering a realistic prediction scenario, we first show the isolated causal component of prediction errors to confirm that the distributions are changing as variables are manipulated in our simulations according to the theorems from section 3. Using the chosen parameters for the models, we calculated ground truth regression equations for the target variable in the unmanipulated model and each manipulated model. We then calculated predicted values for T using the values of all predictor variables in each test dataset with these ground truth equations. Figure 1b shows the average squared difference between the predicted values using the equations for the unmanipulated model and each manipulated model. When the number of manipulated variables in $\mathbf{Ch}_T^{\mathcal{G}}$ is 0, there is no difference between the predictions, even when 5 or 10 non-children of T are manipulated. However, as we increase the number of variables in $\mathbf{Ch}_T^{\mathcal{G}}$ that are manipulated, the difference between the predicted values increases at approximately the same rate regardless of the number of non-children of T that are manipulated. This confirms theorem 1. To confirm theorem 2, we repeated this procedure using only the variables in $\mathbf{MB}_T^{\mathcal{G}_M}$ as predictors for T and made corrections for manipulated children of T that are ancestors of unmanipulated children of T in a given simulation. In each of these cases, there was no difference in the predictions for T when using the ground truth regression equations for either the unmanipulated or manipulated models, indicating that the change in the distribution was correctly accounted for using the causal information.

We now consider the scenario from the causality challenge where we have only training data from the unmanipulated population and test data from some manipulated population and we know the variables that were manipulated. We used two simple causal discovery based methods for prediction: LR-MB/C and LR-MB/C*. For both of these methods we used the training data to calculate parameters for the model, then made the appropriate changes to the model and parameters to account for the manipulations, and finally used the parameters to calculate a regression equation for T using only the variables in $\mathbf{MB}_T^{\mathcal{G}_M}$, which was used to obtain a predicted value for T. For LR-MB/C* we added the additional step of correcting for manipulated nodes that are ancestors of unmanipulated nodes, as described in section 3. We used six other methods for prediction where causality was ignored: LR-ALL, LR-MB, LASSO, SVR-RBF, and RVR-RBF. In each case, a prediction function for T was learned using the training data and then applied to the manipulated test data without accounting for the manipulated variables in any way. LR-ALL and LR-MB are simply linear regression using all of the variables other than T as predictors and only the variables in $\mathbf{MB}_T^{\mathcal{G}}$ as predictors, respectively. LASSO is the "least absolute shrinkage and selection operator", which uses the L_1 penalty to obtain a sparse linear regression model (Tibshirani, 1996). SVR-RBF is support vector regression with a Gaussian RBF kernel (Smola and Schölkopf, 1998), RVR-RBF is relevance vector regression with a Gaussian RBF kernel (Tipping, 2001).

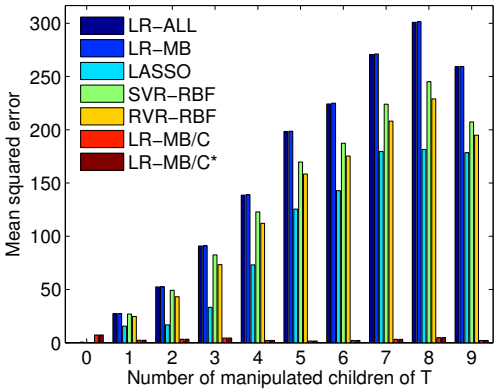


Figure 2: Mean squared error when 0 non-children of T are manipulated

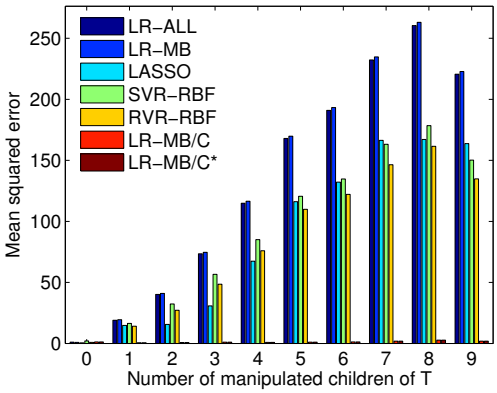


Figure 3: Mean squared error when 5 non-children of T are manipulated

Figures 2, 3, and 4 show the mean squared errors for the predicted values for T for each method as the number of manipulated children increases from 0 to 9, when 0, 5, and 10 non-children of T are manipulated, respectively.

As expected, the methods which take advantage of the causal structure perform no better than the methods that ignore causality when we manipulate 0, 5, or 10 non-children of T as long as no children of T are manipulated. If fact, when no variables are manipulated, the causal methods show the highest error. However, as we manipulate children of T, the accuracy of the causal methods does not change, but the non-causal methods begin to perform progressively worse. The trend as the number of manipulated children increases appears relatively constant for the 0, 5, and 10 manipulated non-children cases. We also note that the difference between LR-MB/C and LR-MB/C* are not noticeable in any case, indicating that the correction applied with the LR-MB/C* method does not make a considerable difference in practice. We attempted the same

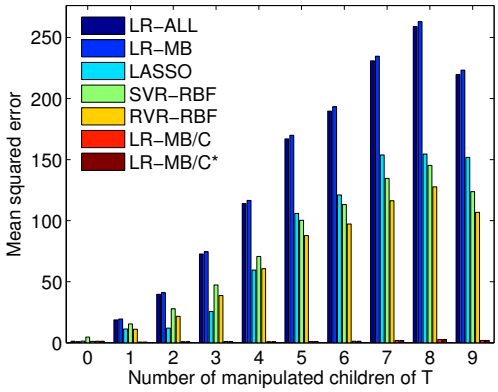


Figure 4: Mean squared error when 10 non-children of T are manipulated

simulations after adding nonlinear dependencies between the variables to test a case where the parametric assumptions made by the causal methods do not hold, but the results were not very informative. We simply note that the nonparametric SVR-RBF and RVR-RBF methods performed best, as we might expect, but made slightly more errors when children of T were manipulated.

5. Conclusions

The conditions from section 3, confirmed experimentally in section 4, may help to explain the surprising results from the recent causality challenge. There is a tradeoff between gains resulting from using the correct causal set of predictors and losses resulting from overfitting and sample bias as well as when parametric assumptions made by causal discovery algorithms do not hold. While the results given in section 4 indicate that there are cases where we should expect causal discovery based methods for prediction to strongly outperform prediction methods which do not account for causality, these cases, where many variables of a target node are manipulated, may not arise frequently in practice, and the exact parametric forms used when generating the data in the experiments may not be reflective of real world data. Thus, in practice, we may see greater performance when nonparametric methods and methods which make less restrictive assumptions about parametric forms and combat overfitting well that ignore causality are used, since even though these methods may make errors related to causality, i.e. not accounting for changes in a distribution after variables are manipulated, these errors may be small when compared to errors resulting from overfitting and sampling bias and when parametric assumptions do not reflect the data when causal discovery algorithms are used.

One possibility for achieving accurate results while still accounting for causality is to use methods which perform well in the prediction scenario with only the causally correct set of variables for each particular setting where certain variables are manip-

ulated, e.g. retrain support vector machines using different sets of "causal" features for each prediction setting where manipulations vary. While using more sophisticated methods for prediction with such causal features may certainly produce models that are less likely to make errors related to sampling bias and overfitting, this still may not overcome problems resulting from assuming a parametric form which does not fit the data. In the experiments in section 4, we assumed that we were able to learn the correct Bayesian network for the data using a causal discovery algorithm, since the variables were linear Gaussians. However, if the data are very nonlinear then the DAG learned may be far from the truth. Thus, it would make more sense to use the features selected by a nonparametric method which does not account for causality, since the causally relevant set of variables for a particular setting where variables are manipulated that is indicated by the DAG may remove important variables and include problematic variables due to errors made by the causal discovery algorithm when a parametric form which does not fit the data is assumed. Fortunately, there has been much recent work in developing causal discovery algorithms which make less restrictive assumptions about the parametric forms, i.e. Shimizu et al. (2006), Hoyer et al. (2008), Hoyer et al. (2009). This may indeed become a possibility for obtaining accurate predictions that are sensitive to changes in a distribution when variables are manipulated in the future.

There are also many other factors which can affect prediction in these contexts. We merely highlighted a few factors relevant for the causality challenge. In particular, we considered only structural or perfect manipulations. In practice, manipulations may not completely break edges into a manipulated node and instead only change the conditional distribution of the node, and may affect other variables as well. We also have not considered the effects of unobserved variables which are causes of more than one of the observed variables on predictions or how well the children of a target variable predict the target compared to other variables in the Markov blanket. A more thorough investigation which considers some of these factors and uses more realistic data for testing may provide a more complete understanding of when causality is useful for making predictions.

Acknowledgments

We thank several anonymous reviewers for helpful comments and suggestions. R.E.T. was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative.

References

P. O. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2008.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. MIT Press, 2009.

D. Koller and N. Friedman. *Structured Probabilistic Models: Principles and Techniques*. Draft Textbook, 2008.

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

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

A. Smola and B. Schölkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, NeuroCOLT, 1998.

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

R. Tibshirani. Regression shrinkage and selction via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

JMLR: Workshop and Conference Proceedings 6:147–156, 2010 NIPS 2008 workshop on causality

Distinguishing between cause and effect

Joris Mooij ORIS.MOOIJ@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics,
72076 Tübingen, Germany

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We describe eight data sets that together formed the `CauseEffectPairs` task in the *Causality Challenge #2: Pot-Luck* competition. Each set consists of a sample of a pair of statistically dependent random variables. One variable is known to cause the other one, but this information was hidden from the participants; the task was to identify which of the two variables was the cause and which one the effect, based upon the observed sample. The data sets were chosen such that we expect common agreement on the ground truth. Even though part of the statistical dependences may also be due to hidden common causes, common sense tells us that there is a significant cause-effect relation between the two variables in each pair. We also present baseline results using three different causal inference methods.

Keywords: causal inference, benchmarks

1. Introduction

Arguably, the most elementary problem in causal inference is to decide whether statistical dependences between two random variables X, Y are due to (a) a causal influence from X to Y, (b) an influence from Y to X, or (c) a possibly unobserved common cause Z influencing X and Y. Most of the state-of-the-art causal inference algorithms address this problem only if X and Y are part of a larger set of random variables influencing each other. In that case, conditional statistical dependences rule out some causal directed acyclic graphs (DAGs) and prefer others (Spirtes et al., 1993; Pearl, 2000).

Recent work (Kano and Shimizu, 2003; Sun et al., 2006; Shimizu et al., 2006; Sun et al., 2008; Hoyer et al., 2009; Janzing and Schölkopf, 2008) suggests that the shape of the joint distribution shows asymmetries between cause and effect, which often indicates the causal direction with some reliability, i.e., one can distinguish between cases (a) and (b).

To enable more objective evaluations of these and other (future) proposals for identifying cause and effect, we have tried to select real-world data sets with pairs of variables where the causal direction is known. The best way to obtain the ground truth of the causal relationships in the systems that generated the data would be by performing

Data set	Number of samples	Variable 1	Variable 2	Causal relationship
pairs01	349	Altitude	Temperature	$1 \to 2$
pairs02	349	Altitude	Precipitation	$1 \to 2$
pairs03	349	Longitude	Temperature	$1 \to 2$
pairs04	349	Sunshine hours	Altitude	$1 \leftarrow 2$
pairs05	4177	Length	Age	$1 \leftarrow 2$
pairs06	4177	Age	Shell weight	$1 \to 2$
pairs07	4177	Diameter	Age	$1 \leftarrow 2$
pairs08	5000	Age	Wage per hour	$1 \to 2$

Table 1: Data sets in the CauseEffectPairs task.

interventions on one of the variables and observing whether the intervention changes the distribution of the other variable. Unfortunately, these interventions cannot be made in practice for many of the existing data sets because the original data-generating system is no longer available, or because of other practical reasons. Therefore, we have selected some data sets in which the causal direction should be clear by common sense.

In selecting the data sets for the CauseEffectPairs task, we applied the following selection criteria:

- the minimum number of data points should be a few hundred;
- the variables should have continuous values;
- there should be a significant cause–effect relationship between the two variables;
- the direction of the causal relationship should be known or obvious from the meaning of the variables;

We collected eight data sets satisfying these criteria, which we refer to as pairs01, ..., pairs08. They can be downloaded from Mooij et al. (2008). Some properties of the data sets are given in Table 1.

In this article, we describe the various data sets in the task and provide our "common sense" interpretation of the causal relationships present in the variables. We also present baseline results of all previously existing applicable causal inference methods that we know of.

2. Climate data

The first four pairs were obtained from climate data provided by the *Deutscher Wetterdienst* (DWD) and are available online at Deutscher Wetterdienst (2008). We merged several of the original data sets to obtain data for 349 weather stations in Germany, selecting only those weather stations with no missing data. After merging the data sets, we selected the following six variables: altitude, latitude, longitude, and annual mean

values (over the years 1961–1990) of sunshine duration, temperature and precipitation. We converted the latitude and longitude variables from sexagesimal to decimal notation. Out of these six variables, we selected four different pairs with "obvious" causal relationships: altitude–temperature, altitude–precipitation, longitude–temperature and sunshine–altitude. We will now discuss each pair in more detail.

2.1. Altitude and temperature

As an elementary fact of meteorology, places with higher altitude tend to be colder than those that are closer to sea level (roughly 1 centigrade per 100 meter). There is no doubt that altitude is the cause and temperature the effect: one could easily think of an intervention where the thermometer is lifted by a balloon to measure the temperature at a higher point of the same longitude and latitude. On the other hand, heating or cooling a location does not change its altitude.

The altitudes in the DWD data set range from 0 m to 2960 m, which is sufficiently large to detect significant statistical dependences. The data is plotted in Figure 1(a).

One potential confounder is latitude, since all mountains are in the south and far from the sea, which is also an important factor for the local climate. The places with the highest average temperatures are therefore those with low altitude but lying far in the south (Upper Rhine Valley). Hence this confounder should induce positive correlations between altitude and temperature as opposed to the negative correlation between altitude and temperature which is already evident from the scatter plot. This suggests that the direct causal relation between altitude and temperature dominates over the confounder.

2.2. Altitude and precipitation

Altitude and precipitation form the second pair of variables that we selected from the DWD data; their relation is plotted in Figure 1(b).

It is known that altitude is also an important factor for precipitation since rain often occurs when air is forced to rise over a mountain range and the air becomes oversaturated with water due to the lower temperature (orographic rainfall). This effect defines an indirect causal influence of altitude on precipitation via temperature. These causal relations are, however, less simple than the causal influence from altitude to temperature because gradients of the altitude with respect to the main direction of the wind are more relevant than the altitude itself. The hypothetical intervention that defines a causal relation could be to build artificial mountains and observe orographic rainfall.

2.3. Longitude and temperature

For the dependence between longitude and temperature, shown in Figure 1(c), a hypothetical intervention could be to move a thermometer between west and east. Even if one could adjust for altitude and latitude, it is unlikely that temperature would remain the same since the climate in the west is more oceanic and less continental than in the east of Germany. Therefore, longitude causes temperature.

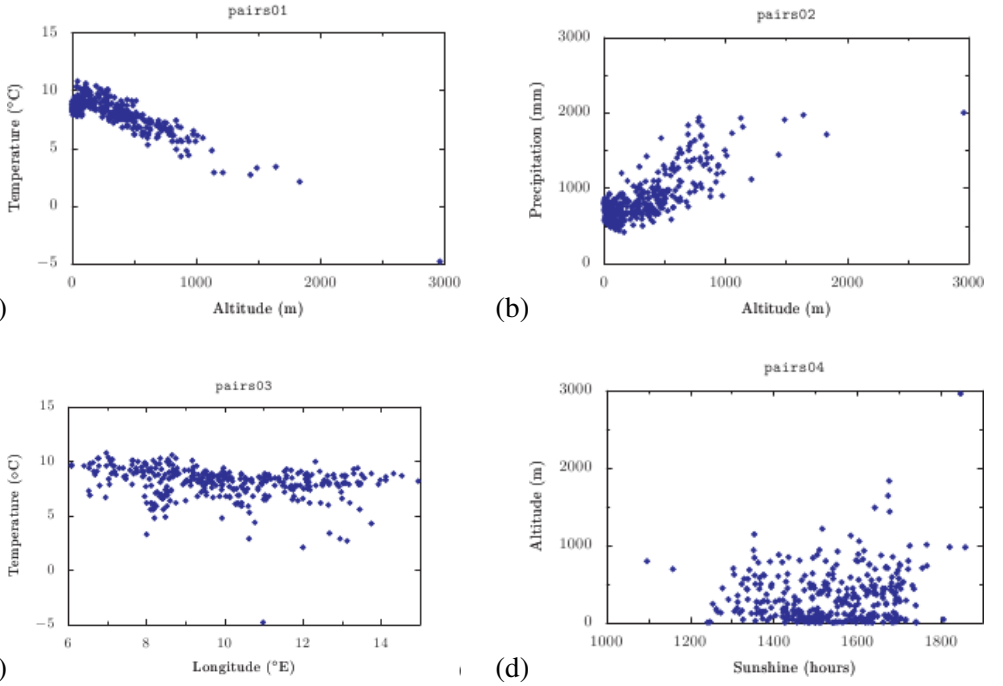


Figure 1: Scatter plots of the German climate data: (a) altitude–temperature, (b) altitude–precipiation, (c) longitude–temperature, (d) altitude–sunshine hours.

2.4. Sunshine hours and altitude

The fourth and final pair of DWD variables are sunshine duration and altitude, shown in Figure 1(d). Linear regression between both quantities shows a slight increase of sunshine duration with altitude. Possible explanations are that higher cities are sometimes above low-hanging clouds. Cities in valleys, especially if they are close to rivers or lakes, typically have more misty days. Moving a sunshine sensor above the clouds clearly increases the sunshine duration whereas installing an artificial sun would not change the altitude. The causal influence from altitude to sunshine duration can be confounded, for instance, by the fact that there is a simple statistical dependence between altitude and longitude in Germany as explained in Subsection 2.1.

3. Abalone data

Another three pairs of variables were selected from the `Abalone` data set (Nash et al., 1994) in the UCI Machine Learning Repository (Asuncion and Newman, 2007). The data set contains 4177 measurements of several variables concerning the sea snail *Abalone*. The original data set contains the nine variables sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and number of rings. The number of rings in the shell is directly related to the age of the snail: adding 1.5 to the number of rings gives the age in years. Of these variables, we selected three pairs with obvious cause-effect relationships, which we now discuss in more detail.

3.1. Length and age

The data for the first Abalone pair, length and age, is plotted in Figure 2(a). For the variable "age" it is not obvious what a reasonable intervention would be since there is no possibility to change the time. However, waiting and observing how the length changes or how it changed from the past to the present can be considered as equivalent to the hypothetical intervention (provided that the relevant background conditions do not change too much). Clearly, this "intervention" would change the probability distributions of the length, whereas changing the length of snails (by a complicated surgery) would not change the distribution of age. Regardless of the difficulties of defining interventions, we expect common agreement on the ground truth (age causes length).

3.2. Age and shell weight

The data are plotted in Figure 2(b). Similar considerations as in Subsection 3.1 hold for the ground truth: age causes shell weight but not vice versa.

3.3. Diameter and age

For the final pair, shell diameter and age, the data are plotted in Figure 2(c). Again, age causes diameter and not the other way around.

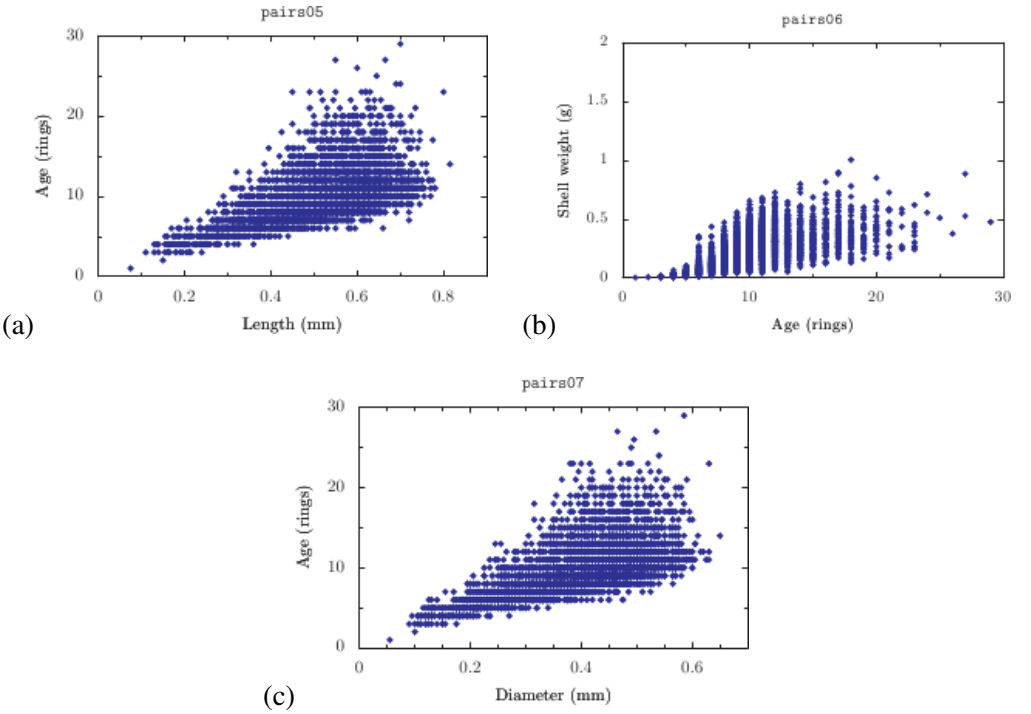


Figure 2: Scatter plots of the Abalone data: (a) length–age, (b) age–shell weight, (c) diameter–age.

4. Age and wage per hour of employees in the USA

Our final data source was the `Census Income` data set (Kohavi, 1996) in the UCI Machine Learning Repository (Asuncion and Newman, 2007). We have selected the following variables: `1 AAGE` (age), and `7 AHRSPAY` (wage per hour) and selected the first 5000 instances for which wage per hour was not equal to zero. The scatter plot for this pair is shown in Figure 3. It clearly shows an increase of wage up to about 45 and decrease for higher age.

As already argued in the Abalone case, interventions on the variable "age" are difficult to define. Compared to the discussion in the context of the Abalone data set, it seems more problematic to consider waiting as a reasonable "intervention" since the relevant (economical) background conditions change rapidly compared to the length of the human life: If someone's salary is higher than the salary of a 20 year younger colleague *because* of his/her longer job experience, we cannot conclude that the younger colleague 20 years later will earn the same money as the colleague earns now. Possibly, the factory or even the branch of industry he/she was working in does not exist any more and his/her job experience is no longer appreciated. However, we know that employees sometimes indeed do get a higher income because of their longer job experience. Pretending longer job experience by a fake certificate of employment would be a

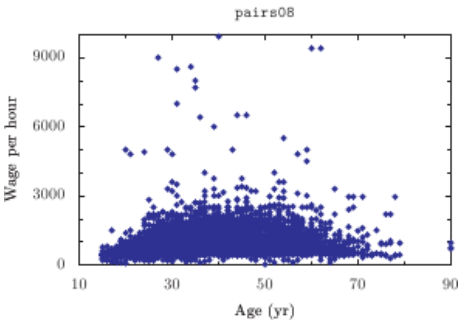


Figure 3: Scatter plot of Census data: age–wage per hour.

possible intervention. On the other hand, changing the wage per hour is an intervention that is easy to imagine (though difficult for us to perform) and this would certainly not change the age.

5. Baseline results

At the time the challenge was held, only three methods existed for deciding upon the causal direction between two real-valued variables, to the best of our knowledge: the method proposed by Friedman and Nachman (2000), LiNGAM (Shimizu et al., 2006) and the causal inference method of Hoyer et al. (2009). In this Section, we report the results of applying these three methods to the data sets of the challenge task. These results may serve as baseline results for future evaluations.

5.1. Comparing marginal likelihood of Gaussian Process regression fits

The basic idea behind the method of Friedman and Nachman (2000) (when applied to the special case of only two variables X and Y) is fitting a Gaussian Process (Rasmussen and Williams, 2006) to the data twice: once with X as input and Y as output, and once with the roles of X and Y reversed. If the former fit has a larger marginal likelihood, this indicates that X causes Y, and otherwise, one concludes that Y causes X. We adopted a squared exponential covariance function and used the GPML code (Rasmussen and Williams, 2007).

The results are shown in Table 2. Only three out of eight causal direction inferences are correct.

5.2. LiNGAM

The causal inference method LiNGAM (an acronym for Linear, Non-Gaussian, Acyclic causal Models) assumes that effects are linear functions of their causes, plus independent additive noise. Shimizu et al. (2006) showed that if all (or all except one of the) noise distributions are non-Gaussian, the correct causal (data-generating) structure can be identified asymptotically using Independent Component Analysis. We have applied

Dataset	$S_{1\to 2}$	$S_{1\leftarrow 2}$	Decision	Ground truth	Correct?
pairs01	2.183×10^{02}	2.171×10^{02}	$1 \to 2$	$1 \to 2$	+
pairs02	3.355×10^{02}	3.385×10^{02}	$1 \leftarrow 2$	$1 \to 2$	-
pairs03	4.858×10^{02}	4.603×10^{02}	$1 \to 2$	$1 \to 2$	+
pairs04	4.821×10^{02}	4.889×10^{02}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs05	5.141×10^{03}	4.291×10^{03}	$1 \to 2$	$1 \leftarrow 2$	-
pairs06	4.568×10^{03}	4.801×10^{03}	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	5.086×10^{03}	4.243×10^{03}	$1 \to 2$	$1 \leftarrow 2$	-
pairs08	6.842×10^{03}	6.869×10^{03}	$1 \leftarrow 2$	$1 \to 2$	-

Table 2: Baseline results for distinguishing the cause from the effect, using the method of Friedman and Nachman (2000); S denotes the logarithm of the marginal likelihood of the Gaussian Process fit.

Dataset	Diagnostic	Decision	Ground truth	Correct?
pairs01	OK	$1 \leftarrow 2$	$1 \to 2$	-
pairs02	Not really triangular at all	$1 \leftarrow 2$	$1 \to 2$	-
pairs03	Not really triangular at all	$1 \to 2$	$1 \to 2$	+
pairs04	Only somewhat triangular	$1 \to 2$	$1 \leftarrow 2$	-
pairs05	OK	$1 \to 2$	$1 \leftarrow 2$	-
pairs06	OK	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	OK	$1 \to 2$	$1 \leftarrow 2$	-
pairs08	OK	$1 \to 2$	$1 \to 2$	+

Table 3: Baseline results for distinguishing the cause from the effect, using LiNGAM Shimizu et al. (2006).

the implementation provided by the authors at (Hoyer et al., 2006) on the data sets of our challenge task.

The results are shown in Table 3. Only two out of eight causal direction inferences are correct.

5.3. Additive noise models

The basic idea of the recent method by Hoyer et al. (2009) is to assume that the effect can be written as some (not necessarily linear) function of the cause, plus additive noise, which is independent of the cause. In practice, one tests the causal model "X causes Y" as follows:

- perform regression of Y on X in order to estimate the function $f : \mathbb{R} \to \mathbb{R}$ that best approximates the functional relationship between X and Y, i.e., such that $Y \approx f(X)$,
- calculate the residuals $Y - f(X)$ for all data points,
- check whether these residuals are independent of X, i.e., whether $(Y - f(X)) \perp\!\!\!\perp X$.

For the regression, we used standard Gaussian Process Regression (Rasmussen and Williams, 2006) using the GPML code (Rasmussen and Williams, 2007), with a squared exponential covariance function. For the independence test, we used the independence test based on the Hilbert Schmidt Independence Criterion (also known as HSIC) (Gretton et al., 2005), using the gamma approximation and Gaussian kernels with heuristically chosen kernel widths. The statistical test assumes independence as a null hypothesis and calculates corresponding p-values. Now in order to decide whether "X causes Y" or, alternatively, "Y causes X", one simply takes the model with the highest p-value for independence between residuals and regressor.

We report the results in Table 4. By using this method, we correctly classify six out of eight data sets. The small p-values may indicate that the assumption of additive noise is violated in these data sets, even in the correct causal direction. Still, by comparing the p-values in both directions, the correct decision is made in most cases.[1]

6. Discussion and remarks on submitted solutions

Finding data sets satisfying the criteria mentioned in Section 1 turned out to be challenging, which explains why the number of data sets in our task is relatively small (another reason is that we only decided to submit a task to the challenge just shortly before the deadline). For future evaluations, the number of data sets should be increased in order to obtain more significant conclusions when used as benchmarks for comparing causal inference algorithms.

1. Meanwhile, we have improved the method by replacing the regression step by a dependence minimization procedure, which yields similar qualitative results, but with more plausible p-values (Mooij et al., 2009).

We received 6 submissions as suggested solutions of this task. The number of correctly identified pairs were $2, 8, 5, 3, 5, 7$, while the submission with 7 correct solutions was (unfortunately) later changed to 5 correct ones. The winner team (Zhang and Hyvärinen) correctly identified 8 out of 8 causal directions. Their method will be described in the paper *Distinguishing causes from effects using nonlinear acyclic causal models*, published elsewhere in this workshop proceedings. One group (not the winning group) used the fact that the pairs contained common variables and used conventional methods in addition to a new method. Since the goal of our task was to consider only pairs of variables at a time, it was a weakness of our task to allow for such a solution strategy (the submission was accepted nevertheless, of course).

An additional desideratum for data sets used in similar future challenges would therefore be that all variable pairs should be disjoint. On the other hand, the constraint that the variables should have continuous values could be removed, which would make the task more challenging for the participants (and would also make it easier to find suitable data).

Acknowledgments

We would like to thank Bernhard Schölkopf for suggesting the DWD climate data. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Deutscher Wetterdienst. Website of the German weather service, 2008. URL http://www.dwd.de/.

N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 2000.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference (ALT 2005)*, pages 63–78, August 2005.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS*2008)*. MIT Press, 2009.

Patrik O. Hoyer, Antti Kerminen, and Shohei Shimizu. LiNGAM v. 1.4.2, December 2006. URL http://www.cs.helsinki.fi/group/neuroinf/lingam/.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition, 2008. URL http://arxiv.org/abs/0804.3678.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Distinguishing between cause and effect, 2008. URL http://www.causality.inf.ethz.ch/repository.php?id=14.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. To appear at the 26th International Conference on Machine Learning (ICML 2009), 2009.

W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288), 1994.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.

C. E. Rasmussen and C. Williams. GPML code, 2007. URL http://www.gaussianprocess.org/gpml/code.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

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

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer, New York, 1993.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71:1248–1256, 2008.

Dataset	$p_{1\to 2}$	$p_{1\leftarrow 2}$	Decision	Ground truth	Correct?
pairs01	1.64×10^{-02}	9.43×10^{-15}	$1 \to 2$	$1 \to 2$	+
pairs02	1.50×10^{-13}	2.88×10^{-16}	$1 \to 2$	$1 \to 2$	+
pairs03	7.89×10^{-03}	7.02×10^{-04}	$1 \to 2$	$1 \to 2$	+
pairs04	5.50×10^{-05}	1.08×10^{-02}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs05	1.13×10^{-70}	7.79×10^{-23}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs06	1.56×10^{-210}	1.98×10^{-113}	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	2.66×10^{-82}	5.85×10^{-26}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs08	$0.00 \times 10^{+00}$	1.60×10^{-80}	$1 \leftarrow 2$	$1 \to 2$	-

Table 4: Baseline results for distinguishing the cause from the effect, using the method of Hoyer et al. (2009).

JMLR: Workshop and Conference Proceedings 6:157–164, 2010 NIPS 2008 workshop on causality

Distinguishing Causes from Effects using Nonlinear Acyclic Causal Models

Kun Zhang KUN.ZHANG@CS.HELSINKI.FI
Dept of Computer Science and HIIT
University of Helsinki
00014 Helsinki, Finland

Aapo Hyvärinen AAPO.HYVARINEN@CS.HELSINKI.FI
Dept of Computer Science, HIIT, and Dept of Mathematics and Statistics
University of Helsinki
00014 Helsinki, Finland

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Distinguishing causes from effects is an important problem in many areas. In this paper, we propose a very general but well defined nonlinear acyclic causal model, namely, post-nonlinear acyclic causal model with inner additive noise, to tackle this problem. In this model, each observed variable is generated by a nonlinear function of its parents, with additive noise, followed by a nonlinear distortion. The nonlinearity in the second stage takes into account the effect of sensor distortions, which are usually encountered in practice. In the two-variable case, if all the nonlinearities involved in the model are invertible, by relating the proposed model to the post-nonlinear independent component analysis (ICA) problem, we give the conditions under which the causal relation can be uniquely found. We present a two-step method, which is constrained nonlinear ICA followed by statistical independence tests, to distinguish the cause from the effect in the two-variable case. We apply this method to solve the problem "CauseEffectPairs" in the Pot-luck challenge, and successfully identify causes from effects.

Keywords: causal discovery, sensor distortion, additive noise, nonlinear independent component analysis, independence tests

1. Introduction

Given some observable variables, people often wish to know the underlying mechanism generating them, and in particular, how they are influenced by others. Causal discovery has attracted much interest in various areas, such as philosophy, psychology, machine learning, etc. There are some well-known algorithms for causal discovery. For example, conditional independence tests can be exploited to remove unnecessary connections among the observed variables and to produce a set of acyclic causal models which are in the d-separation equivalence class (Pearl, 2000; Spirtes et al., 2000).

Recently, some methods have been proposed for model-based causal discovery of continuous variables (see, e.g., Shimizu et al., 2006; Granger, 1980). Model-based causal discovery assumes a generative model to explain the data generating process. If the assumed model is close to the true one, such methods could not only detect the causal relations, but also discover the form in which each variable is influenced by others. For example, Granger causality assumes that effects must follow causes and that the causal effects are linear (Granger, 1980). If the data are generated by a linear acyclic causal model and at most one of the disturbances is Gaussian, independent component analysis (ICA) (Hyvärinen et al., 2001) can be exploited to discover the causal relations in a convenient way (Shimizu et al., 2006).

However, the above causal models seem too restrictive for real-life problems. If the assumed model is wrong, model-based causal discovery may give misleading results. Therefore, when the prior knowledge about the data model is not available, the assumed model should be general enough such that it could be adapted to approximate the true data generating process. On the other hand, the model should be identifiable such that it could distinguish causes from effects. In a large class of real-life problems, the following three effects usually exist. 1. The effect of the causes is usually nonlinear. 2. The final effect received by the target variable from all its causes contains some noise which is independent from the causes. 3. Sensors or measurements may introduce nonlinear distortions into the observed values of the variables. To address these issues, we propose a very realistic model, called post-nonlinear acyclic causal model with inner additive noise. In the two-variable case, we show the identifiability of this model under the assumption that the involved nonlinearities are invertible. We conjecture that this model is identifiable in very general situations, as illustrated by the experimental results.

2. Proposed Causal Model

Let us use a directed acyclic graph (DAG) to describe the generating process of the observed variables. We assume that each observed continuous variable x_i, corresponding to the ith node in the DAG, is generated by two stages. The first stage is a nonlinear transformation of its parents pa_i, denoted by $f_{i,1}(pa_i)$, plus some noise (or disturbance) e_i (which is independent from pa_i). In the second stage, a nonlinear distortion $f_{i,2}$ is applied to the output of the first stage to produce x_i. Mathematically, the generating process of x_i is

$$x_i = f_{i,2}(f_{i,1}(pa_i) + e_i). \tag{1}$$

In this model, we assume that the nonlinearities $f_{i,2}$ are continuous and invertible. $f_{i,1}$ are not necessarily invertible. This model is very general, since it accounts for the nonlinear effect of the causes pa_i (by using $f_{i,1}$), the noise effect in the transmission process from pa_i to x_i (using e_i), and the nonlinear distortion caused by the sensor or measurement (using $f_{i,2}$). In particular, in this paper we focus on the two-variable case. Suppose that x_2 is caused by x_1. The relationship between x_1 and x_2 is then assumed to be

$$x_2 = f_{2,2}(f_{2,1}(x_1) + e_2), \tag{2}$$

where e_2 is independent from x_1.

3. Identifiability[1]

3.1. Relation to post-nonlinear mixing ICA

We first consider the case where the nonlinear function $f_{2,1}$ is also invertible. Let $s_1 \triangleq f_{2,1}(x_1)$ and $s_2 \triangleq e_2$. As e_2 is independent from x_1, obviously s_1 is independent from s_2. The generating process of (x_1, x_2), given by Eq. 2, can be re-written as

$$\begin{cases} x_1 = f_{2,1}^{-1}(s_1), \\ x_2 = f_{2,2}(s_1 + s_2). \end{cases} \tag{3}$$

We can see that clearly x_1 and x_2 are post-nonlinear (PNL) mixtures of independent sources s_1 and s_2 (Taleb and Jutten, 1999). The PNL mixing model is a nice special case of the general nonlinear ICA model.

ICA is a statistical technique aiming to recover independent sources from their observed mixtures, without knowing the mixing procedure or any specific knowledge of the sources (Hyvärinen et al., 2001). The basic ICA model is linear ICA, in which the observed mixtures, as components of the vector $\mathbf{x} = (x_1, x_2 \cdots, x_n)^T$, are assumed to be generated from the independent sources $s_1, s_2 \cdots, s_n$, with a linear transformation $\mathbf{A}$. Mathematically, we have $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s} = (s_1, s_2 \cdots, s_n)^T$. Under weak conditions on the source distribution and the mixing matrix, ICA can recover the original independent sources up to the permutation and scaling indeterminacies with another transformation $\mathbf{W}$, by making the outputs as independent as possible. That is, the outputs of ICA, as components of $\mathbf{y} = \mathbf{W}\mathbf{x}$, produce an estimate of the original sources s_i. In the general nonlinear ICA problem, $\mathbf{x}$ is assumed to be generated from independent sources s_i with an invertible nonlinear mapping $\mathcal{F}$, i.e., $\mathbf{x} = \mathcal{F}(\mathbf{s})$, and the separation system is $\mathbf{y} = \mathcal{G}(\mathbf{x})$, where $\mathcal{G}$ is another invertible nonlinear mapping. Generally speaking, nonlinear ICA is ill-posed: its solutions always exist but they are highly non-unique (Hyvärinen and Pajunen, 1999). To make the solution to nonlinear ICA meaningful, one usually needs to constrain the mixing mapping to have some specific forms (Jutten and Taleb, 2000).

The PNL mixing ICA model plays a nice trade-off of linear ICA and general nonlinear ICA. It is described as a linear transformation of the independent sources $s_1, s_2, ..., s_n$ with the transformation matrix $\mathbf{A}$, followed by a component-wise invertible nonlinear transformation $\mathbf{f} = (f_1, f_2, ..., f_n)^T$. Mathematically,

$$x_i = f_i\Big(\sum_{k=1}^{n} \mathbf{A}_{ik} s_k\Big).$$

1. When this paper was finalized, a systematic investigation of the identifiability of the proposed causal model was already reported in Zhang and Hyvärinen (2009), which contains some different results from this paper. Please refer to Zhang and Hyvärinen (2009) for more rigorous results on the identifiability.

In matrix form, it is denoted as $\mathbf{x} = \mathbf{f}(\mathbf{A}\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ and $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$. In particular, from Eq. 3, one can see that for the causal model Eq. 2, the mixing matrix is $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, and the post-nonlinearity is $\mathbf{f} = (f_{2,1}^{-1}, f_{2,2})^T$.

3.2. Identifiability of the Causal Model

The identifiability of the causal model Eq. 2 is then related to the separability of the PNL mixing ICA model. The PNL mixing model $(\mathbf{A}, \mathbf{f})$ is said to be separable if the independent sources s_i could be recovered only up to some trivial indeterminacies (which includes the permutation, scaling, and mean indeterminacies) with a separation system $(\mathbf{g}, \mathbf{W})$, The output of the separation system is $\mathbf{y} = \mathbf{W} \cdot \mathbf{g}(\mathbf{x})$, where $\mathbf{g}$ is a componentwise continuous and invertible nonlinear transformation. The separability of the PNL mixing model has been discussed in several contributions. As Achard and Jutten (2005) proved the separability under very general conditions, their result is briefly reviewed below.

Theorem 1 (Separability of the PNL mixing model, by Achard & Jutten) *Let $(\mathbf{A}, \mathbf{f})$ be a PNL mixing system and $(\mathbf{g}, \mathbf{W})$ the separation system. Let $h_i \triangleq g_i \circ f_i$. Assume the following conditions hold.*

- *Each source s_i appears mixed at least once in the observations.*
- *$h_1, h_2, ..., h_n$ are diffenrentiable and invertible (same conditions as $f_1, f_2, ..., f_n$).*
- *There exists at most one Gaussian source.*
- *The joint density function of the sources s_i is differentiable, and its derivative is continuous on its support.*

Then the output of the separation system $(\mathbf{g}, \mathbf{W})$ has mutually independent components if and only if each h_i is linear and $\mathbf{WA}$ is a generalized permutation matrix.

The above theorem states that under the conditions stated above, by making the outputs of the separation system ($\mathbf{g}$, $\mathbf{W}$) mutually independent, the original sources s_i and the mixing matrix $\mathbf{A}$ could be uniquely estimated (up to some trivial indeterminacies). If $f_{2,1}$ is invertible, the causal model Eq. 2, as a special case of the PNL mixing model, can then be identified. Thus, the theorem above implies the following proposition.

Proposition 1 (Identifiability of the causal model with invertible nonlinearities) *Suppose that x_1 and x_2 are generated according to the causal model Eq. 2 with both $f_{2,2}$ and $f_{2,1}$ differentiable and invertible. Further assume that at most one of $f_{2,1}(x_1)$ and e_2 is Gaussian, and that their joint density is differentiable, with the derivative continuous on its support. Then the causal relation between x_1 and x_2 can be uniquely identified.*

In the discussions above, we have constrained the nonlinearity $f_{2,1}$ to be invertible. Otherwise, $f_{2,1}^{-1}$ does not exist, and the causal model Eq. 2 is no longer a PNL mixing one. A rigorous proof of the identifiability of the causal model in this situation is under investigation. But it seems that it is identifiable under very general conditions, as

verified by various experiments. It should be noted that when all the nonlinear functions $f_{i,2}$ are constrained to be identity mappings, the proposed causal model is reduced to the nonlinear causal model with additive noise which was recently investigated by Hoyer et al. (2009). Interestingly, for this model, it was shown that in the two-variable case, the identifiability actually does not depend on the invertibility of the nonlinear function $f_{2,1}$.

4. Method for Identification

Given two variables x_1 and x_2, we identify their causal relation by finding which one of the possible relations ($x_1 \to x_2$ and $x_2 \to x_1$) satisfies the assumed causal model. If the causal relation is $x_1 \to x_2$ (i.e., x_1 and x_2 satisfy the model Eq. 2), we can invert the data generating process Eq. 2 to recover the disturbance e_2, which is expected to be independent from x_1. One can then examine if a possible causal model is preferred in two steps: the first step is actually a constrained nonlinear ICA problem which aims to retrieve the disturbance corresponding to the assume causal relation; in the second step we verify if the estimated disturbance is independent from the assume cause using statistical tests.

4.1. A two-step method

Suppose the causal relation under examination is $x_1 \to x_2$. According to Eq. 2, if this causal relation holds, there exist nonlinear functions $f_{2,2}^{-1}$ and $f_{2,1}$ such that $e_2 = f_{2,2}^{-1}(x_2) - f_{2,1}(x_1)$ is independent from x_1. Thus, we first perform nonlinear ICA using the structure in Figure 1. The outputs of this system are $y_1 = x_1$, and $y_2 = g_2(x_2) - g_1(x_1)$. In our experiments, we use multi-layer perceptrons (MLP's) to model the nonlinearities g_1 and g_2. Parameters in g_1 and g_2 are learned by making y_1 and y_2 as independent as possible, which is achieved by minimizing the mutual information between y_1 and y_2. The joint density of $\mathbf{y} = (y_1, y_2)^T$ is $p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})/|\mathbf{J}|$, where $\mathbf{J}$ is the Jacobian matrix of the transformation from (x_1, x_2) to (y_1, y_2), i.e., $\mathbf{J} = \left[\partial(y_1, y_2)/\partial(x_1, x_2)\right]$. Clearly $|\mathbf{J}| = |g_2'|$. The joint entropy of $\mathbf{y}$ is then

$$H(\mathbf{y}) = -E\{\log p_{\mathbf{y}}(\mathbf{y})\} = -E\{\log p_{\mathbf{x}}(\mathbf{x}) - \log|\mathbf{J}|\} = H(\mathbf{x}) + E\{\log|\mathbf{J}|\}.$$

Finally, the mutual information between y_1 and y_2 is

$$\begin{aligned} I(y_1, y_2) &= H(y_1) + H(y_2) - H(\mathbf{y}) \\ &= H(y_1) + H(y_2) - E\{\log|\mathbf{J}|\} - H(\mathbf{x}) \\ &= -E\{p_{y_1}(y_1)\} - E\{p_{y_2}(y_2)\} - E\{\log|g_2'|\} - H(\mathbf{x}), \end{aligned}$$

where $H(\mathbf{x})$ does not depend on the parameters in g_1 and g_2 and can be considered as constant. One can easily find the gradient of $I(y_1, y_2)$ w.r.t. the parameters in g_1 and g_2, and minimize $I(y_1, y_2)$ using gradient-descent methods. Details of the algorithm are skipped.

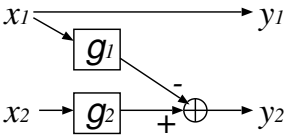


Figure 1: The constrained nonlinear ICA system used to verify if the causal relation $x_1 \to x_2$ holds.

y_1 and y_2 produced by the first step are the assumed cause and the estimated corresponding disturbance, respectively. In the second step, one needs to verify if they are independent, using statistical independence tests. We adopt the kernel-based statistical test (Gretton et al., 2008), with the significance level $\alpha = 0.01$. If y_1 and y_2 are not independent, indicating that $x_1 \to x_2$ does not hold, we repeat the above procedure (with x_1 and x_2 exchanged) to verify if $x_2 \to x_1$ holds. If y_1 and y_2 are independent, usually we can conclude that x_1 causes x_2, and that g_1 and g_2 provide an estimate of $f_{2,1}$ and $f_{2,2}^{-1}$, respectively. However, it is possible that both $x_1 \to x_2$ and $x_2 \to x_1$ hold, although the chance is very small. Therefore, for the sake of reliability, in this situation we also test if $x_2 \to x_1$ holds. Finally, we can find the relationship between x_1 and x_2 among all four possible scenarios: 1. $x_1 \to x_2$, 2. $x_2 \to x_1$, 3. both causal relations are possible, and 4. there is no causal relation between x_1 and x_2 which follows our model.

4.2. Practical considerations

The first issue that needs considering in practical implementation of our method is the model complexity, which is controlled by the number of hidden units in the MLP's modelling g_1 and g_2 in Figure 1. The system should have enough flexibility, and at the same time, to avoid overfitting, it should be as simple as possible. To this end, two ways are used. One is 10-fold cross-validation. The other is heuristic: we try different numbers of hidden units in a reasonable range (say, between 4 and 10); if the resulting causal relation does not change, we conclude that the result is feasible.

The second issue is the initialization of the nonlinearities g_1 and g_2 in Figure 1. If the nonlinear distortions $f_{2,2}$ and $f_{2,1}$ are very strong, it may take a long time for the nonlinear ICA algorithm in the first step to converge, and it is also possible that the algorithm converges to a local optimum. This can be avoided by using reasonable initializations for g_1 and g_2. Two schemes are used in our experiments. One is motivated by visual inspection of the data distribution: we simply use a logarithm-like function to initialize g_1 and g_2 to make the transformed data more regular. The other is by making use of Gaussianization (Zhang and Chan, 2005). Roughly speaking, the central limit theorem states that sums of independent variables tend to be Gaussian. Since $f_{2,2}^{-1}(x_2)$ in the causal model Eq. 2 is the sum of two independent variables, it is expected to be not very far from Gaussian. Therefore, for each variable which is very far from Gaussian, its associated nonlinearity (g_1 or g_2 in Figure 1) is initialized by the strictly increasing function transforming this variable to standard Gaussian. In all experiments, these two schemes give the same final results.

Data Set	#1	#2	#3	#4	#5	#6	#7	#8
Result	$x_1 \to x_2$	$x_1 \to x_2$	$x_1 \to x_2$	$x_1 \leftarrow^{\ddagger} x_2$	$x_1 \leftarrow x_2$	$x_1 \to x_2$	$x_1 \leftarrow x_2$	$x_1 \to x_2$

Table 1: Causal directions obtained. (‡ indicates that the causal relation is not significant.)

5. Results

The proposed nonlinear causal discovery method has been applied to the "CauseEffectPairs" task proposed by Mooij et al. (2008) in the Pot-luck challenge. In this task, eight data sets are given; each of them contains the observed values of two variables x_1 and x_2. The goal is to distinguish the cause from the effect for each data set. Figure 2 gives the scatterplots of x_1 and x_2 in all the eight data sets. Table 1 summaries our results. In particular, below we take data sets 1 and 8 as examples to illustrate the performance of our method.

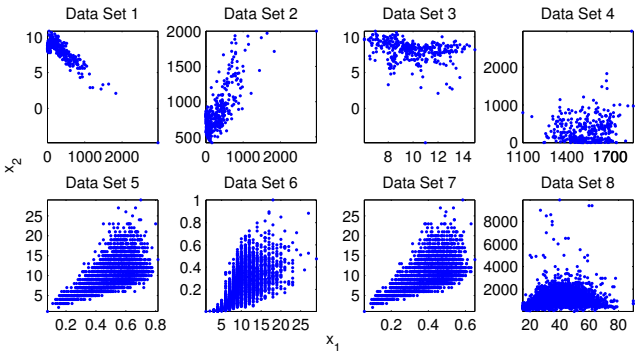


Figure 2: Scatterplot of x_1 and x_2 in each data set of the "CauseEffectPairs" task (Mooij et al., 2008).

The variable x_1 in Data set 1 is non-negative and extremely non-Gaussian. We initialized the nonlinearity g_1 with the transformation $\log(2+x_1)$ (Gaussianization was also tested and it finally produced the same causal relation). The scatterplot of y_1 and y_2 (as outputs of the constrained nonlinear ICA system in Figure 1) under each hypothesis ($x_1 \to x_2$ or $x_2 \to x_1$) is given in Figure 3(a,b). Clearly y_1 and y_2 are much more independent under hypothesis $x_1 \to x_2$. This is verified by the independence test results in the third row of Table 2. Note that a large test statistic tends to reject the null hypothesis (the independence between y_1 and y_2). Figure 4 shows the result on Data set 8. In this case, we applied the transformation $\log(x_2+50)$ for initialization. By comparing (a) and (b) in Figure 4, also by inspecting the independence test results in the fourth row of Table 2, one can see clearly that $x_1 \to x_2$.

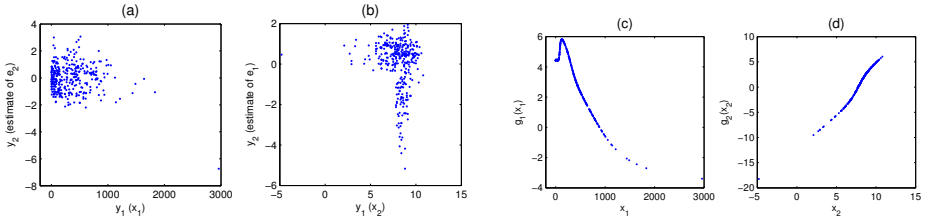


Figure 3: Result on Data set 1. (a) y_1 vs. y_2 under hypothesis $x_1 \to x_2$. (b) that under $x_2 \to x_1$. (c & d) x_1 vs. $g_1(x_1)$ and x_2 vs. $g_2(x_2)$ under hypothesis $x_1 \to x_2$.

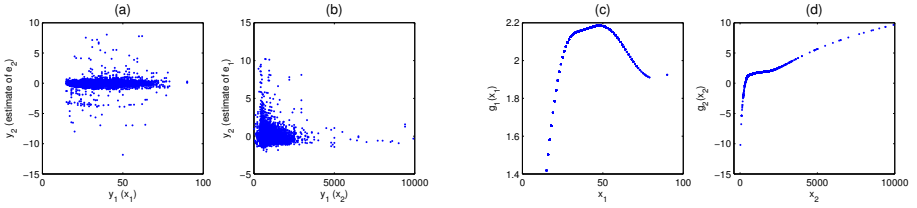


Figure 4: Result on Data set 8. For captions of the sub-figures, please refer to Figure 3.

Data Set	$x_1 \to x_2$ assumed		$x_2 \to x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}
#8	1.2×10^{-4}	1.2×10^{-4}	1.1×10^{-4}	7.4×10^{-4}

Table 2: Result of independence test on y_1 and y_2 for Data sets 1 and 8 under different assumed causal directions. For both data sets, the independence hypothesis is accepted in the scenario $x_1 \to x_2$, and rejected in the other scenario, with the significance level $\alpha = 0.01$.

6. Conclusion

We proposed a very general nonlinear causal model for model-based causal discovery. This model takes into account the nonlinear effect of the causes, inner noise effect, and the sensor distortion, and is capable of approximating the data generating process of some real-life problems. We presented the identifiability of this model under the assumption that the involved nonlinearities are invertible. Experimental results illustrated that based on this model, one could successfully distinguish the cause from the effect, even if the nonlinear function of the cause is not invertible. An on-going work is to investigate the identifiability of this model under more general conditions.

References

S. Achard and C. Jutten. Identifiability of post-nonlinear mixtures. *IEEE Signal Processing Letters*, 12:423–426, 2005.

C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 1980.

A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A.J. Smola. A kernel statistical test of independence. In *NIPS 20*, pages 585–592, Cambridge, MA, 2008.

P.O. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS 21*, Vancouver, B.C., Canada, 2009. To appear.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

C. Jutten and A. Taleb. Source separation: From dusk till dawn. In *2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, pages 15–26, Helsinki, Finland, 2000.

J. Mooij, D. Janzing, and B. Schölkopf. Distinguishing between cause and effect, Oct. 2008. URL http://www.kyb.tuebingen.mpg.de/bs/people/jorism/causality-data/.

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

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

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

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

K. Zhang and L. W. Chan. Extended Gaussianization method for blind separation of post-nonlinear mixtures. *Neural Computation*, 17(2):425–452, 2005.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Canada, June 2009.

JMLR: Workshop and Conference Proceedings 6:165–176, 2010 NIPS 2008 workshop on causality

Structure Learning in Causal Cyclic Networks

Sleiman Itani* SSOLOMON@MIT.EDU
77 Massachusetts Ave,
32-D760
Cambridge, MA, 02139, USA

Mesrob Ohannessian* MESROB@MIT.EDU
77 Massachusetts Ave,
32-D740
Cambridge, MA, 02139, USA

Karen Sachs* KAREN.SACHS@STANFORD.EDU
Stanford University School of Medicine,
269 Campus Drive
Stanford, CA, 94305, USA

Garry P. Nolan GNOLAN@STANFORD.EDU
Stanford University School of Medicine,
269 Campus Drive
Stanford, CA, 94305, USA

Munther A. Dahleh DAHLEH@MIT.EDU
77 Massachusetts Ave,
32-D734
Cambridge, MA, 02139, USA
* *These authors contributed equally to this work.*

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Cyclic graphical models are unnecessary for accurate representation of joint probability distributions, but are often indispensable when a causal representation of variable relationships is desired. For variables with a cyclic causal dependence structure, DAGs are guaranteed not to recover the correct causal structure, and therefore may yield false predictions about the outcomes of perturbations (and even inference.) In this paper, we introduce an approach to generalize Bayesian Network structure learning to structures with cyclic dependence. We introduce a structure learning algorithm, prove its performance given reasonable assumptions, and use simulated data to compare its results to the results of standard Bayesian network structure learning. We then propose a modified, heuristic algorithm with more modest data requirements, and test its performance on a real-life dataset from molecular biology, containing causal, cyclic dependencies.

1. Introduction

Bayesian network models encode probabilistic relationships among random variables, providing a framework for tasks such as inference and decision making. In some settings, it is useful for model edges to represent probabilistic dependence resulting from causal mechanisms. This is the case when the goal is structure recovery for the sake of revealing causal interactions for prediction of perturbation effects in some domain, for instance, when learning the structure of molecular pathways from biological measurements.

Causal Bayesian network models have been described Pearl (2000), relying on the *framework of causation*, which enables causal interpretation under proper assumptions Spirtes et al. (1993). These models may be learned from *observational data*, i.e. passive observations of the domain. However, such methods yield entire equivalence classes, leaving the causal direction of many edges unknown. A solution to this problem is offered by the *framework of intervention*, where interventions effectively override variables, and halt the influence of the network on them, enabling the use of *interventional* or *experimental* data Pearl (1995) and Pearl (2000). In this framework, it is possible to ask: "how can the graphical structure of the causal model be recovered from observational and experimental data?"

Research in Bayesian networks has predominantly focused on directed acyclic graphs (DAGs), even when the acyclicity assumption is knowingly violated Friedman et al. (2000). Within that context, solutions to this question abound, e.g. Cooper and Yoo (1999). In cyclic domains, DAGs represent an inaccurate causal structure, consequently, prediction of perturbation effects will fail, as in Figure 1. To avoid these inaccuracies, a representation which encompasses cycles must be employed.

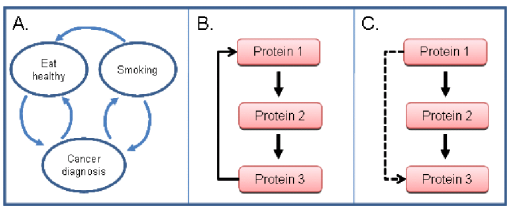


Figure 1: **Cyclic causal networks**. **A.** Risk assessment network for predicting the effect of behavior interventions. Smoking positively influences diagnosis of lung cancer, while eating healthy does so negatively. A cancer diagnosis may influence eating and smoking choices, though cessation of smoking can deteriorate eating habits. **B.** In protein networks, feedback loops are ubiquitous modes of positive and negative regulation of biological processes. **C.** A DAG representation of the cyclic structure in B. The dotted line indicates an incorrectly oriented edge: perturbing protein 3 would inaccurately be assessed as having no effect on proteins 1 and 2.

BN models with directed cyclic graphs (DCGs), though inherently possible, Pearl (1988), had unclear interpretation and applicability Spirtes et al. (1993). Cycles also occur in an alternative modeling paradigm called structural equation models (SEMs),

which model functional dependence directly. Key developments Spirtes (1995) and Koster (1996), endowed DCGs with some of the properties of their DAG counterparts, and it was also shown that some SEMs are amenable to the same analysis Spirtes (1995); Pearl and Dechter (1996). Based on these, Richardson (1996) established an algorithm for discovering a partial structure on DCGs. Recently, Lacerda et al. (2008) provided an alternative algorithm. Both procedures lie in the framework of causation, and use solely observational data to output equivalence classes, rather than a single DCG.

In this paper, we are interested in modeling cycles, yet tapping into the power of experimental data. At the extreme of exhaustive interventions, the problem appears trivial. However, discovering structure by such brute force is a daunting task, and in truth one is constrained by the number and type of interventions at hand. We address the problem through the following contributions:

- In Section 2, we give a novel formalization of cyclic networks by characterizing them locally with stochastic kernels, which bridge the SEM context with that of BNs by replacing deterministic equations with exogenous variables by a direct probabilistic description. We call the resulting models generalized Bayesian networks (GBNs). The framework of intervention extends directly to such a description, resulting in causal GBNs (CGBNs).

- In Section 3, we prove that interventions allow us to discover descendants and children. Such discovery is robust, in that in general it does not result in false discovery and, given natural properties, it always succeeds as the size of the data grows to infinity. Interventions can affect either the *abundance* or the *activity* of variables (corresponding to ingoing or outgoing edges, respectively). However, in this work, we assume the *activity* of a perturbed variable is affected. We elaborate in Section 2.3.

- In Section 4, we cast these results into an algorithm for structure learning. Rather than searching over all causal interactions by brute force, we first discover cycle breakers. Upon intervention on these quantities, we reduce the task into an acyclic problem which can be learned generically. Finally we close cycles to recover the cyclic structure. We illustrate these results on synthetic data with 14 nodes, 2 cycles and 3 interventions.

- In Section 5, we develop a modified heuristic algorithm for the structure learning from more limited data, containing only one perturbation per sample. This algorithm is inspired by our previous one, and is motivated by limitations on experiment technologies. We illustrate the usefulness of this algorithm by studying a biological dataset of 11 variables from the MAPK/AKT pathway (CYTO) Sachs et al. (2005).

Finally, a related research area is that concerned with structure learning with time course data. In this case, alternative representations exist in the form of dynamic Bayesian networks (DBNs) Friedman et al. (1999) and continuous-time Bayesian networks

(CTBNs) Nodelman et al. (2002, 2003). These models represent cycles by 'unrolling' them in time. As with other efforts to learn static representations of underlying dynamic systems Friedman et al. (2000); Sachs et al. (2005), what we propose here can be interpreted as learning a DBN or CTBN in the absence of time-course data, or from single time-point data (constituting a snapshot of a dynamic system).

2. Problem formulation

2.1. Generalized Bayesian networks

Definition 1 (Generalized Bayesian network) *We define a* generalized Bayesian network *(GBN) as a pair (G,F), where G is a directed graph $G=(V,E)$ and F is a set of stochastic kernels (conditional probability tables) $f_i : \mathcal{X} \times \mathcal{X}^{|\pi_i|} \to \mathbf{R}_+$ indexed by all nodes $i \in V$, for a finite set $\mathcal{X}$. Here, π_i is the set of parents of i in G. With each node i of the GBN we associate a random variable X_i. In this paper, we restrict ourselves to discrete random variables taking values in a common alphabet $\mathcal{X}$.* [1] *The GBN then* induces *a joint distribution on $X_1, \cdots, X_N$ satisfying the following* characterizations*:*

1. *Local characterization:*

$$\mathbb{P}(X_i = x_i, X_{\pi_i} = x_{\pi_i}) = \mathbb{P}(X_{\pi_i} = x_{\pi_i}) f_i(x_i; x_{\pi_i}), \qquad \forall i \in V. \tag{1}$$

2. *Independence under d-separation: Given any two nodes i and j in G, if i and j are d-separated Pearl (1988) by a set $Z \subset V$, then X_i and X_j given $\{X_k, k \in Z\}$.*

We make the following assumption:

Assumption *[Existence and Uniqueness] For every F that we consider, there exists a unique induced (global) joint distribution that satisfies all the local characterizations in Equation (1).*

Since *GBN*'s are generalizations of BN's to the cyclic case, the previous assumption doesn't hold for any graph G and stochastic kernels $\{f_i\}$. This is just like the fact that a dynamic system with feedback (cycle) is not necessarily causal even if all of the subsystems are causal. Of course, it is expected that in the applications of interest, the variables measured *do* come from a unique underlying joint distribution.. Another view of this assumption is that it is the same as the one in the case of the Gibb's sampler: for the sampling to guarantee convergence, a unique joint that is compatible with the given conditionals must exist.

When the graph of a GBN is acyclic, the product of all the stochastic kernels gives a valid joint distribution satisfying (1). Thus, by uniqueness, an acyclic GBN reduces to a BN:

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i \in V} f_i(x_i; x_{\pi_i}). \tag{2}$$

1. Although we restrict ourselves to discrete variables, this is in general not restrictive since any continuous variable can approximated arbitrarily well by a discrete variable.

2.2. Causal generalized Bayesian networks

Let an *intervention* (I,ξ) be a pair, where $I \subset V$ is a subset of the nodes of a graph G, and $\xi \in \mathcal{X}^{|I|}$ is a tuple of values in an alphabet $\mathcal{X}$.

Definition 2 (Causal generalized Bayesian network) *We define a* causal generalized Bayesian network *(CGBN) as a GBN with which we associate a collection of joint distributions* $\mathbb{P}_{(I,\xi)}$ *indexed by all interventions* (I,ξ)*, for each of which it satisfies:*

$$\mathbb{P}_{(I,\xi)}(X_i = x_i, X_{\pi_i \setminus I} = x_{\pi_i \setminus I}) = \mathbb{P}_{(I,\xi)}(X_{\pi_i \setminus I} = x_{\pi_i \setminus I}) f_i(x_i; x_{\pi_i \setminus I}, \xi_{\pi_i \cap I}), \qquad \forall i \in V. \quad (3)$$

When ξ is implicit we only use I as subscript, and when $I = \varnothing$ we drop the subscript altogether. Below, we provide more intuition about this definition. Meanwhile, we extend the assumption of existence and uniqueness to CGBNs by taking it to hold for every intervention (I,ξ). With this, an acyclic CGBN reduces to a causal BN, in the sense of interventions Pearl (2000):

$$\mathbb{P}_{(I,\xi)}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i \in V} f_i(x_i; x_{\pi_i \setminus I}, \xi_{\pi_i \cap I}). \quad (4)$$

2.3. σ-μ characterization

By carefully examining Equation (3), we can see how interventions effectively decouple nodes into *seen* and *measured* values. Just as in the do-calculus of Pearl, the intervention value supersedes the node variable itself as far as its influence on the network goes, and can thus be interpreted as what is (internally) *seen* by all descendants. The value of the *seen* variable is determined solely by the intervention. However, and this is in contrast to traditional intervention models, we (externally) *measure* or *observe* the value (i.e. abundance) of the intervention variables. These can be thought of as shadow copies, which are still influenced by the network *but no longer influence it*, because its activity is externally set by the intervention. This formulation is motivated by some inhibition models in molecular biology, where the inhibitors do not change the amount of a given protein but rather halt its activity. Thus the correct modeling of this situation is to separate the inhibited node from its children. The σ-μ characterization simply does that while staying in the framework of probability theory. All of our results extend to the case when measured values are lost, by eliminating the variables intervened at.

We can capture this decoupling via an explicit characterization which reduces a CGBN with an intervention to a GBN. In particular, given a CGBN (G,F) describing N variables and an intervention (I,ξ), one can construct a GBN (G',F') which describes $N + |I|$ variables, such that the restriction to the first N of the variables has a joint distribution evaluating to $\mathbb{P}_{(I,\xi)}$. We call this construction the σ-μ characterization of a CGBN. We do not elaborate on this further, and leave its illustration to the second example below.

2.4. Examples

2.4.1. CYCLE WITH 2 NODES

Consider the GBN with binary-valued variables X_1 and X_2 described in Figure 2. The local characterizations of the joint distribution $\mathbb{P}$ induced by the GBN are as follows: $\mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_2 = x_2) f_1(x_1; x_2)$, and $\mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_1 = x_1) f_2(x_2; x_1)$, for all binary configurations of x_1 and x_2. Under the proper choice of f_1 and f_2, these yield linearly independent equations, in which case a distribution satisfying the local characterizations exists and is unique.

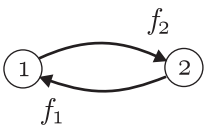


Figure 2: A GBN with two nodes.

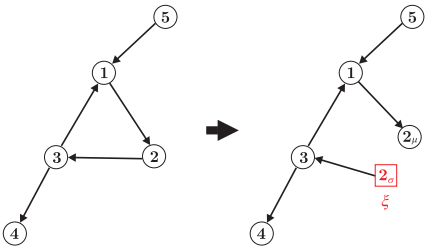


Figure 3: A CGBN with an intervention at node 2.

2.4.2. BREAKING CYCLES

Consider the CGBN described in Figure 3a. In Figure 3b, we illustrate what happens when node 2 is intervened at. We use the σ-μ characterization, and represent the seen node with a σ subscript and the measured node with a a μ subscript. Note how node 2_μ is effectively a leaf under intervention. As such, the resulting graph is a DAG. It follows that, the product of all the f_i's is a valid characterization, and by uniqueness it is the distribution induced by the CGBN under the intervention. The resulting network is thus exactly equivalent to a BN. We say that the cycle has been *broken*. This notion, in more generality, will be used throughout our algorithm (Section 4).

3. Interventions and Descendent Detection

We now introduce analytical results which we subsequently use to justify the correctness of our algorithm for structure learning. For conciseness, we state and prove only the forward direction of the results. The converses hold under some natural properties of the network and interventions. Please see supporting materials for additional proofs and associated assumptions.

Theorem 3 *Consider a CGBN, and let the existence and uniqueness assumption hold. Intervene at a single node i, that is let $(I, \xi) = (i, \xi_i)$ and consider a node j. If j is not a descendant of i then $\mathbb{P}(X_j = x_j) = \mathbb{P}_{(i,\xi_i)}(X_j = x_j)$ for all $x_j \in \mathcal{X}$.*

Proof Partition V into two: V_i (nodes that are descendants of i, including i) and $\overline{V_i}$ (nodes that are not descendants of i). Consider the network restricted to $\overline{V_i}$, by restricting the graph. Since there are no incoming edges from V_i to $\overline{V_i}$, we can also restrict F to contain only f_j, $j \in \overline{V_i}$. Since none of the local characterizations of the distribution induced by the restricted network depend on the intervention, and by the uniqueness of the solution, the restricted distribution is unchanged. Thus the marginal distributions of all $j \in \overline{V_i}$ is unchanged. ■

In other words, Theorem 3 states that if a node j experiences a change in marginal distribution when i is intervened at, then it is a descendant of i. As mentioned, the converse also holds under proper assumptions, detailed in the supporting materials. One of these assumptions states that a child variable must be sensitive to perturbations imposed upon its parent variables, an assumption which may in general be violated, particularly if the network compensates in the face of perturbations. Such *insensitive* descendants may still be detectable with the use of multiple perturbations.

Theorem 4 *Consider a CGBN, an intervention (I^1, ξ^1), and an incremental intervention (I^2, ξ^2) by a single node i, as in $I^2 \setminus I^1 = \{i\}$. Let the existence and uniqueness assumption hold. Define $\mathbb{P} := \mathbb{P}_{(I^1,\xi^1)}$ and $\mathbb{Q} := \mathbb{P}_{(I^2,\xi^2)}$. Consider a node j and let $\tilde{\pi}_j = \pi_j \setminus I^2$. If j is not a child of i then $\mathbb{P}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) = \mathbb{Q}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})$ for all $x_j \in \mathcal{X}$ and $x_{\tilde{\pi}_j} \in \mathcal{X}^{|\tilde{\pi}_j|}$.*

Proof We shall split the parents of j into three groups: i itself if it is a parent, the never-intervened-at parents $\tilde{\pi}_j$, and the always-intervened-at parents $\hat{\pi}_j$. When j is not a child of i the inclusion pattern for the parents of j in the local characterization is unchanged. Hence:

$$\begin{aligned}
\mathbb{P}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) &= \frac{\mathbb{P}(X_j = x_j, X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})}{\mathbb{P}(X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})} = f_j(x_j; x_{\tilde{\pi}_j}, \xi^1_{\hat{\pi}_j}), \\
\mathbb{Q}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) &= \frac{\mathbb{Q}(X_j = x_j, X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})}{\mathbb{Q}(X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})} = f_j(x_j; x_{\tilde{\pi}_j}, \xi^2_{\hat{\pi}_j}).
\end{aligned}$$

But since ξ^1 and ξ^2 agree on $\hat{\pi}_j$, the claim follows. ■

In other words, Theorem 4 states that if a node j experiences a change in marginal conditional distribution given the never-intervened-at parents $\tilde{\pi}_j$ when i is intervened at, then it is a child of i. Again, the converse also holds under proper assumptions.

4. Algorithm for structure learning

Consider a CGBN from which we can sample both observational and experimental data, from an intervention set I and its subsets. Assume that I is 'rich', in the sense that it has at least one representative node from every cycle in the underlying graph. The

following algorithm effectively guides the experimental procedure (or uses previously collected data) and recovers the CGBN's structure. In what follows, we elaborate the subroutines that are used, and show correctness.

Algorithm: Learn CGBN structure

0: Start with a CGBN and an intervention set I.
1: [Probing experiments] Collect sets of i.i.d. samples under no-intervention and single-intervention data, i.e. when node i is intervened at, for each i in I.
2: Call subroutine 'detect descendants' to recover descendant information for all nodes in I.
3: Identify the minimal subset of nodes in I which are sufficient to break all cycles, and denote it by $I_{\mathcal{C}}$.
4: [Cycle-breaking experiment] Collect i.i.d. samples when all nodes in $I_{\mathcal{C}}$ are intervened at.
5: Recover an embedded DAG.
6: [Leave-one-out experiments] Collect sets of i.i.d. samples when nodes in $I_{\mathcal{C}} \setminus \{i\}$ are intervened at, for each $i \in I_{\mathcal{C}}$.
7: Call subroutine 'detect children' to recover child information for all nodes in $I_{\mathcal{C}}$.
8: Recover all missing edges in the DAG, and complete the DCG structure of the CGBN.

The following is the subroutine that obtains descendant information based on no-intervention and single-intervention i.i.d. data. The correctness of the subroutine follows from Theorem 3 and the convergence of empirical distributions, since non-descendants will exhibit no change of marginal, whereas descendants will. The choice of distance is not critical, and thresholding can be automated.

Subroutine: Detect descendants

0: Start with sets of n i.i.d. samples generated by a CGBN, under no interventions as well as single-interventions at each i in I. Initialize a binary $|V| \times |I|$ descendant information matrix.
1: For each $j \in V$:
2: Compute $\hat{\mathbb{P}}^n(X_j)$, the empirical marginal of X_j under no interventions.
3: For each $i \in I$:
4: Compute $\hat{\mathbb{P}}^n_i(X_j)$, the empirical marginal of X_j under the single-intervention i.
5: Evaluate some distance between $\hat{\mathbb{P}}^n(X_j)$ and $\hat{\mathbb{P}}^n_i(X_j)$.
6: If the distance exceeds a threshold, mark j as a descendant of i.
7: Next i.
8: Next j.
9: Compute the transitive closure of the descendant information matrix, and return it.

$I_{\mathcal{C}}$ can then be identified as the set of all self-descendants. Since the intervention set I has at least one node from each cycle in the underlying graph, $I_{\mathcal{C}}$ constitutes a cycle-breaking intervention set, meaning that if all nodes in $I_{\mathcal{C}}$ are intervened at, the CGBN behaves like a BN. Thus with i.i.d. data obtained as such, we can recover the corresponding embedded DAG using generic BN structure learning, which we do not elaborate further on. Note that I itself is a cycle-breaking intervention set, the merit here being that $I_{\mathcal{C}}$ can be much smaller.

Note that the only edges that are in the underlying graph but are missing from the embedded DAG are those from cycle breakers to their children. The following

subroutine obtains a child information matrix, based on $I_{\mathcal{C}}$-intervention and leave-one-out from $I_{\mathcal{C}}$ intervention i.i.d. data. Once this information is obtained, all cycles can be closed in a straightforward fashion, recovering the underlying structure. Once again, the correctness of the subroutine follows from Theorem 4 and the convergence of empirical distributions, since only children will exhibit a change in marginal conditional.

Subroutine: Detect children

0: Start with the recovered DAG, and sets of n i.i.d. samples generated by the CGBN, under $I_{\mathcal{C}}$-intervention as well as leave-one-out interventions, i.e. on $I_{\mathcal{C}} \setminus \{i\}$ for each i in $I_{\mathcal{C}}$. Initialize a binary $|V| \times |I_{\mathcal{C}}|$ child information matrix. Denote by $\tilde{\pi}_j$ the parents of node j according to the recovered DAG.
1: For each $j \in V$:
2: For each $\alpha \in \mathcal{X}^{|\tilde{\pi}_j|}$:
3: Compute the empirical marginal conditional $\hat{\mathbb{P}}^n_{I_{\mathcal{C}}}(X_j | X_{\tilde{\pi}_j} = \alpha)$, call it $\mathbb{Q}_1$.
4: For each $i \in I_{\mathcal{C}}$:
5: Compute the empirical marginal conditional $\hat{\mathbb{P}}^n_{I_{\mathcal{C}} \setminus \{i\}}(X_j | X_{\tilde{\pi}_j} = \alpha)$, call it $\mathbb{Q}_2$.
6: Evaluate some distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$.
7: If the distance exceeds a threshold, mark j as a child of i.
8: Next i.
9: Next α.
10: Next j.
11: Return the completed child information matrix.

To illustrate the algorithm, we simulated a GBN that has fourteen variables,shown in Figure 4, each with three states $\mathcal{X} = \{0,1,2\}$, two cycles $5 \to 6 \to 7 \to 5$ and $8 \to 9 \to 10 \to 11 \to 8$, and nodes 7, 8 and 10 available for intervention. The stochastic kernels were sampled continuously from the 3-simplex. The simulation was performed using Gibbs-like sampling Chou et al. (1991); Sharma et al. (1989), and up to 4000 data points were sampled for every required intervention.

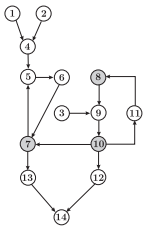


Figure 4: Test network, recovered exactly by GBN learning algorithm

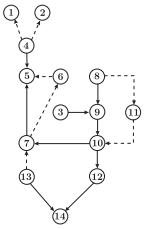


Figure 5: Best network recovered by BN structure learning

GBN algorithm

Data	Correct	Inverted	Added
1000	14	0	0
2000	15	0	0
4000	16	0	0

BN structure learning

Data	Correct	Inverted	Added
1000	9	3	0
2000	9	7	0
4000	12	4	2

Figure 6: Performance tables

In the tables of Figure 6, we compare the performance of our algorithm to a plain BN structure learning algorithm for the various data sizes. In particular, the tables document the number of true edges that the algorithms uncover, the number of reversed edges that they give, and the number of edges that they add but which are absent in the original graph. Observe that the GBN algorithm recovers the network exactly with 4000 data points. The comparison is inherently unfair, because BN structure learning does not handle cycles, but the emphasis here is on illustrating the type of pitfalls in using BNs to capture data that is generated by a GBN. Using the best recovered DAG in Figure 5, for instance, will mistakenly predict that an intervention at node 9 will not affect node 8.

5. Single Perturbations

In this section we introduce an algorithm that is inspired by our previous one but doesn't require data with multiple simultaneous perturbations. Due to practical considerations, sometimes multiple inhibition data-sets are not available. This is why we are interested in an algorithm that can recover the causal structure (even when it's cyclic) without the need for multiple simultaneous perturbations. We assume that the interventions that are available are activity interventions, and so the amount of the variable x can be measured when x is intervened at. The algorithm we have when such perturbations are available is as follows:

Algorithm: Learn CGBN structure without multiple simultaneous perturbations

0: Start with a CGBN and an intervention set I.

1: [Probing experiments] Collect sets of i.i.d. samples under no-intervention and single-intervention data, i.e. when node i is intervened at, for each i in I.

2: Call subroutine 'detect descendants' to recover descendant information for all nodes in I.

3: Identify the subset of all nodes in I which are in cycles, and denote it by $I_{\mathcal{C}}$.

4: Use a regular CBN learning algorithm to recover an approximation of the structure of the causal relations. This is done with the standard structure learning algorithm using the complete dataset, as in Sachs et al. (2005).

5: For every variable i in $I_{\mathcal{C}}$:

   a- Recover the paths from i's descendants in the cycle back to it using BN learning on the data where i was perturbed. As in the original algorithm, this recovers the linearized structure with the perturbed node as a leaf.

   b- Overwrite the paths from i's descendants in the BN approximate graph. This step may alter the parent set of i as well as the direction of edges among i's ancestors. Because the approximate graph is expected to have incorrect edge directionality imposed by the cycles, the graph under perturbations is considered more accurate.

6: Call subroutine 'detect children' to recover child information for all nodes in $I_{\mathcal{C}}$. Use the data with no perturbations and the data with i inhibited for all $i \in I_{\mathcal{C}}$.

7: Recover all missing edges in the DAG, and complete the DCG structure of the CGBN. This proceeds as in the original algorithm, using only the observational data to detect direct edges and indirect paths from each variable in $I_{\mathcal{C}}$ to its descendants.

This algorithm is a heuristic, although it inherits some of the intuition and reasoning of our previous algorithm: It recovers the structure of every cycle by first breaking it and finding its partial structure. To illustrate the performance of this algorithm, we applied it to a real data set from the MAPK/AKT pathway Sachs et al. (2005).

6. Results from the CYTO dataset

The heuristic algorithm from Section 5 was applied to the CYTO dataset Sachs et al. (2005), a real-life dataset of eleven protein measurements, which employs single perturbations (per sample), including three activity inhibitors and one abundance inhibitor. Model results (figure 6) show the edges from regular BN structure learning in blue (solid lines), novel edges resulting from the GBN approach in purple (broken lines). To assess this model's accuracy in representing the true underlying causal structure, as compared to the original model, we turned to the biological literature. There are seven edges unique to the GBN model, of which three represent canonical, well established causal connections that were completely missed by standard BN structure learning efforts. One of these, the connection between PIP2 and Akt, our model represents somewhat inaccurately, shifting the canonical edge (PIP3 $\rightarrow$ Akt). PIP2 and PIP3 are precursors of each other, so this edge incorrectly assigns the parent of Akt as the precursor of the actual parent, perhaps due to confounding effects of the dynamics of the system (i.e. PIP2 abundance may more accurately represent the quantity of PIP3 that influenced the current level of Akt, see Itani et al. (2009). An additional perturbation, or a more idealized one, may have helped resolved this inaccuracy. It can be argued that the GBN model with the shifted edge comes closer to representing the true structure than the BN model that fails to represent this interaction all together. Another canonical edge present only in the GBN model is PKC $\rightarrow$ Plcγ, known in the classic literature, but in the reverse orientation. While this may be an inaccuracy in direction of the edge, the data clearly support this connection (with the Plcγ distribution strongly affected by PKC perturbation), and it has been reported by previous studies Xu et al. (2001); Quinlan et al. (2003), leading us to believe it is a correct edge. Like its BN counterpart, the GBN model misses the edge in the Plcγ $\rightarrow$ PKC direction, but unlike the BN model, it successfully represents the dependence between these two proteins. Finally, the canonical edge (PIP2 $\rightarrow$ PKC) is missed by the BN model but correctly represented in the GBN model. For these canonical edges, the GBN model is somewhat imperfect but nevertheless strongly outperforms the BN model.

Of the remaining four edges, both p38 $\rightarrow$ PKC and PKA $\rightarrow$ PIP3 are supported by previous literature findings Shimizu et al. (1999); Deming et al. (2008). We did not find specific evidence for the edges from Erk to PKC and PKA, though several studies report feedback on PKA and PKC, with potential roles for Erk Geritsa et al. (2008). Although confirmation of all model results requires experimental validation, comparison to literature studies indicates a clear improvement in accuracy for the GBN model. Additionally, the GBN model improves on the BN result by *accurately representing all causal connections and conditional independencies found in the data*, something the standard BN model is unable to achieve.

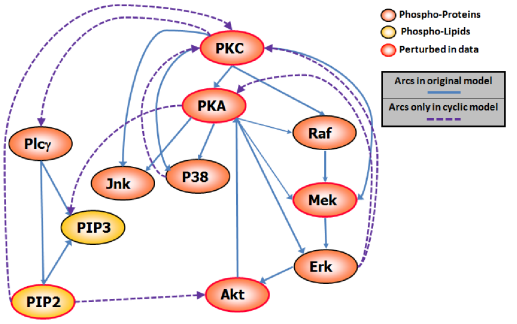


Figure 7: **Application of the heuristic algorithm to real-life protein dataset.** Application of the heuristic structure learning algorithm to this dataset from Sachs et al. (2005) yields this cyclic structure. Edges found in the original graph, resulting from standard Bayesian network structure learning, are in blue (solid lines), edges unique to the GBN result are in purple (broken lines). Several of the cycles in this result structure are supported by literature findings (see text).

7. Conclusion and future work

In this paper we reviewed previous work in incorporating both causality and cyclic structure within the context of Bayesian networks. We then presented the formalism of generalized BNs, which preserves only the local characterizations with stochastic kernels, applying it equally well to the cyclic case, under an existence and uniqueness assumption for the joint distribution. In the acyclic case, this reduces to BNs. The framework of interventions easily extends to this formalism, resulting in causal GBNs. We present an algorithm that uses no-intervention and single- intervention data to detect cycle breakers, then uses multiple simultaneous interventions to learn an embedded DAG, close cycles, and recover the underlying DCG. This algorithm relies on a minimal set of perturbations. We illustrate the procedure via a numerical example. Finally, we present a modified algorithm with more modest, one-intervention-at-a-time data requirements and demonstrate its performance on a real-life biological dataset, successfully recovering many known connections, and strongly outperforming standard structure learning with respect to recovery of the known causal structure. This work can be extended in several directions. We are currently expanding its application to biological data by extending the algorithm to one which explicitly handles the imperfect specificity and efficacy of biological inhibitors. A more theoretical direction is that of relating snapshot structure embodied in GBNs to that of underlying time-dynamics. For that, one needs to start with a dynamic hypothesis of data generation, e.g. CTBNs, stochastic differential equations, etc. Conditions under which the static and dynamic structures coincide would further motivate the current paradigm.

References

C. Chou, Bentler, P.M. Satorra, A. Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, 44, 347-357.(1991)

G. Cooper and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 116–125.

P. B. Deming, S. L. Campbell, L. C. Baldor, and A. K. Howe (2008). Protein Kinase A Regulates 3-Phosphatidylinositide Dynamics during Platelet-derived Growth Factor-induced Membrane Ruffling and Chemotaxis. 10.1074 *J. Biol Chem*

N. Friedman, K. Murphy, and S. Russell (1999). Learning the structure of dynamic probabilistic networks. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 139–147.

N. Friedman, N. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, vol. 7, pp 601–620.

N. Geritsa, S. Kostenkoa, A. Shiryaeva, M. Johannessena and U. Moens (2008). Relations between the mitogen-activated protein kinase and the cAMP-dependent protein kinase pathways: Comradeship and hostility. 20(9):1592-607. *Cellular Signaling*

S. Itani, K. Sachs, J. Fitzgerald, L. Wille, B. Schoeberl, G. Nolan and M. Dahleh (2009). Single timepoint models of dynamic systems. *In preparation*

J. T. A. Koster (1996). Markov properties of nonrecursive causal models. *Annals of Statistics*, vol. 24, no. 5, pp. 2148–2177.

G. Lacerda, P. Spirtes, J. Ramsey, P. O. Hoyer (2008). Discovering cyclic causal models by independent components analysis. *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence*.

U. Nodelman, C. Shelton, and D. Koller (2002). Continuous time Bayesian networks. *Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 378–387.

U. Nodelman, C. Shelton, and D. Koller (2003). Learning continuous time Bayesian networks. *Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 451–458.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffman.

J. Pearl (1995). Causal diagrams for empirical research. *Biometrika*, vol. 82, no. 4, pp. 669–710.

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

J. Pearl and R. Dechter (1996). Identifying independence in causal graphs with feedback. *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 420–426.

L. Quinlan, S. Faherty, and M. Kane (2003). Phospholipase C and protein kinase C involvement in mouse embryonic stem-cell proliferation and apoptosis. 126(1):121-31. *Reproduction*

T. S. Richardson (1996). A discovery algorithm for directed cyclic graphs. *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 454–461.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan (2005). Causal protein-signaling networks derived from multiparameter single-cell data. 308(5721):523 - 529.*Science*.

S. Sharma, S., Durvasula S., Dillan, W. R.. Some results on the behavior of alternate covariance structure estimation in the presence of non-normal data. *Journal of Marketing research* 26, 214-221. (1989)

T. Shimizu, T. Kato, Jr. , A. Tachibana and M. S. Sasaki (1999). Coordinated Regulation of Radioadaptive Response by Protein Kinase C and p38 Mitogen-Activated Protein Kinase. 251(2):424-32. *Experimental Cell Research*

P. Spirtes, C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*, Lecture Notes in Statistics, vol. 81. Springer-Verlag.

P. Spirtes (1995). Directed cyclic graphical representations of feedback models. *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 491–498.

A. Xu, Y. Wang, L. Yi Xu, and R. Stewart Gilmour (2001). Protein Kinase C-mediated Negative Feedback Regulation Is Responsible for the Termination of Insulin-like Growth Factor I-induced Activation of Nuclear Phospholipase C 1 in Swiss 3T3 Cells. 276(18):14980-6. *J. Biol. Chem*

JMLR: Workshop and Conference Proceedings 6:177–190, 2010 NIPS 2008 workshop on causality

Causal learning without DAGs

David Duvenaud DUVENAUD@CS.UBC.CA
Daniel Eaton DEATON@CS.UBC.CA
Kevin Murphy MURPHYK@CS.UBC.CA
Mark Schmidt SCHMITDM@CS.UBC.CA
University of British Columbia
Department of Computer Science
2366 Main Mall
Vancouver, BC V6T 1Z4
Canada

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Causal learning methods are often evaluated in terms of their ability to discover a true underlying directed acyclic graph (DAG) structure. However, in general the true structure is unknown and may not be a DAG structure. We therefore consider evaluating causal learning methods in terms of predicting the effects of interventions on unseen test data. Given this task, we show that there exist a variety of approaches to modeling causality, generalizing DAG-based methods. Our experiments on synthetic and biological data indicate that some non-DAG models perform as well or better than DAG-based methods at causal prediction tasks.

Keywords: Bayesian Networks, Graphical models, Structure Learning, Causality, Interventions, Cell signalling networks, Bioinformatics.

1. Introduction

It is common to make causal models using directed acyclic graphs (DAGs). However, one problem with this approach is that it is very hard to assess whether the graph structure is correct or not. Even if we could observe "nature's graph", it probably would not be a DAG, and would contain many more variables than the ones we happened to have measured. Realistic mechanistic (causal) models of scientific phenomena are usually much more complex, involving coupled systems of stochastic partial differential equations, feedback, time-varying dynamics, and other complicating factors.

In this paper, we adopt a "black box" view of causal models. That is, we define causality in *functional* terms, rather than by committing to a particular representation. Our framework is as follows. Suppose we can measure d random variables, X_i, for $i = 1{:}d$. For example, these might represent the phosphorylation levels of different proteins. Also, suppose we can perform k different actions (interventions), A_j, for $j = 1{:}k$. For example, these might represent the application of different chemicals to the system. For simplicity, we will think of the actions as binary, $A_j \in \{0,1\}$, where a

value of 1 indicates that we performed action A_j. We *define* a causal model as one that can predict the effects of actions on the system, i.e., a conditional density model of the form $p(\mathbf{x}|\mathbf{a})$. These actions may or may not have been seen before, a point we discuss in more detail below. Note that our definition of causal model is even more general than the one given in Dawid (2009), who defines a causal model as (roughly speaking) any model that makes conditional independence statements about the X and A variables; as Dawid points out, such assumptions may or may not be representable by a DAG.

To see that our definition is reasonable, note that it includes the standard approach to causality (at least of the non-counterfactual variety) as a special case. In the standard approach (see e.g., (Spirtes et al., 2000; Pearl, 2000; Lauritzen, 2000; Dawid, 2002)), we assume that there is one action variable for every measured variable. We further assume that $p(\mathbf{x}|\mathbf{a})$ can be modeled by a DAG, as follows:

$$p(X_1,\ldots,X_d|A_1=0,\ldots,A_d=0,G,f)=\prod_{j=1}^{d} f_j(X_j,X_{\pi_j}) \tag{1}$$

where G is the DAG structure, π_j are the parents of j in G, and $f_j(X_j,X_{\pi_j})=p(X_j|X_{\pi_j},A_j=0)$ is the conditional probability distribution (CPD) for node j, assuming that node j is not being intervened on (and hence $A_j=0$). If node j *is* being intervened on, we modify the above equation to

$$p(X_1,\ldots,X_d|A_j=1,A_{-j}=0,G,f,g)=g_j(X_j,X_{\pi_j})\prod_{k\neq j} f_k(X_k,X_{\pi_k}) \tag{2}$$

where $g_j(X_j,X_{\pi_j})=p(X_j|X_{\pi_j},A_j=1)$ is the CPD for node j given that node j is being intervened on. In the standard model, we assume that the intervention sets the variable to a specific state, i.e., $g_j(X_j,X_{\pi_j})=I(X_j=S_j)$, for some chosen target state S_j. This essentially cuts off the influence of the parents on the intervened-upon node. We call this the *perfect* intervention assumption. A real-world example of this might be a gene knockout, where we force X_j to turn off (so $S_j=0$). The crucial assumption is that actions have local effects, and that the other f_j terms are unaffected.

If we do not know which variables an action affects, we can learn this; we call this the *uncertain* intervention model (Eaton and Murphy, 2007). In particular, this allows us to handle actions which affect multiple nodes. These are sometimes called "fat hand" actions; the term arises from thinking of an intervention as someone "sticking their hand" into the system, and trying to change one component, but accidently causing side effects. Of course, the notion of "fat hands" goes against the idea of local interventions. In the limiting case in which an action affects all the nodes, it is completely global. This could be used to model the effects of a lethal chemical that killed a cell, and hence turned all genes "off".

If we model $p(\mathbf{x}|\mathbf{a})$ by a DAG, and make the perfect intervention assumption, then we can make predictions about the effects of actions we have never seen before. To see this, suppose we have collected N samples from the non interventional regime, $\mathcal{D}=\{\mathbf{x}_n\}_{n=1}^N$, where $\mathbf{x}_n\sim p(\mathbf{x}|\mathbf{a}=\mathbf{0})$ (this is called observational data). We can use this data to learn the non-interventional CPDs f_j. Then we make a prediction about what

would happen if we perform a novel action, say turning A_j on, by simply replacing f_j with g_j, which we assume is a delta function, $I(X_j = S_j)$. Of course, if the data is only observational, we will not, in general, be able to uniquely infer the DAG, due to problems with Markov equivalence. However, if some of the data is sampled under perfect interventions, then we can uniquely recover the DAG (Eberhardt et al., 2005, 2006).

The key question is: is the assumption of DAGs and perfect interventions *justified* in any given problem? What other models might we use? It seems that the only way to choose between methods in an objective way, without reference to the underlying mathematical representation, is to collect some real-world data from a system which we have perturbed in various ways, partition the data into a training and test set, and then evaluate each model on its ability to predict the effects of interventions. This is what we do in this paper.

An important issue arises when we adopt this functional view of causality, which has to do with generalizing across actions. In the simplest case, we sample training data from regimes $p(\mathbf{x}|\mathbf{a}_1), \ldots, p(\mathbf{x}|\mathbf{a}_r)$, for r different action combinations, and then sample test data *from the same regimes*. We will see an example of this in Section 3.1, where we discuss the intracellular flow cytometry dataset analyzed in Sachs et al. (2005). In this setup, we sample data from the system when applying one chemical at a time, and then ask the model to predict the protein phosphorylation levels when the same chemical is applied.

A more interesting task is to assume that the test data is drawn from a different sampling regime than the training data. This clearly requires that one make assumptions about how the actions affect the variables. We will see an example of this in Section 3.2, where we discuss another flow cytometry dataset, used in the Dream 2008 competition. In this setup, we sample data from the system when applying one inhibitory chemical and one excitatory chemical at a time, but then ask the model to predict the protein phosphorylation levels when a novel pair of chemicals is applied. For example, we train on data sampled from $p(\mathbf{x}|a_1 = 1, a_2 = 1, a_3 = 0)$ and $p(\mathbf{x}|a_1 = 0, a_2 = 1, a_3 = 1)$, and test on data sampled from $p(\mathbf{x}|a_1 = 1, a_2 = 0, a_3 = 1)$. That is, we have seen A_1 and A_2 in combination, and A_2 and A_3 in combination, and now want to predict the effects of the A_1, A_3 combination. Another variation would be to train on data from $p(\mathbf{x}|a_1 = 1, a_2 = 0)$ and $p(\mathbf{x}|a_1 = 0, a_2 = 1)$, and test on data sampled from $p(\mathbf{x}|a_1 = 1, a_2 = 1)$. This is similar to predicting the effects of a double gene knockout given data on single knockouts.

The most challenging task is when the testing regime contains actions that were never tried before in the training regime, neither alone nor in combination with other actions. For example, suppose we train on data sampled from $p(\mathbf{x}|a_1 = 1, a_2 = 0)$ and test on data sampled from $p(\mathbf{x}|a_1 = 0, a_2 = 1)$. In general, these distributions may have nothing to do with each other. Generalizing to a new regime is like predicting the label of a novel word in a statistical language model. In general, this is impossible, unless we break the word down into its component pieces and/or describe it in terms of features (e.g., does it end in "ing", does it begin with a capital letter, what is the

language of origin, what is the context that it was used in, etc). If we represent actions as "atomic", all we can do is either make the DAG plus perfect intervention assumption, or assume that the action has no affect, and "back-off" to the observational regime. We will compare these approaches below.

2. Methods

In this section, we discuss some methods for learning conditional density models to represent $p(\mathbf{x}|\mathbf{a})$, some based on graphs, others not. We will compare these methods experimentally in the next section. Code for reproducing these experiments is available at www.cs.ubc.ca/~murphyk/causality.

2.1. Approaches to Modeling Interventions

We consider several classes of methods for creating models of the form $p(\mathbf{x}|\mathbf{a})$:

1. **Ignore**: In this case, we simply ignore A and build a generative model of $P(X)$. This has the advantage that we gain statistical strength by pooling data across the actions, but has the disadvantage that we make the same prediction for all actions.

2. **Independent**: In this case, we fit a separate model $P(X|A)$ for each unique joint configuration of A. This is advantageous over the *ignore* model in that it makes different predictions for different actions, but the disadvantage of this model is that it does not leverage information gained between different action combinations, and can not make a prediction for an unseen configuration of A.

3. **Conditional**: In this case, we build a model of $P(X|A)$, where we use some parametric model relating the A's and X's. We give the details below. This will allow us to borrow strength across action regimes, and to handle novel actions.

2.2. Approaches based on DAGs

In the ignore case, we find the exact MAP DAG using the dynamic programming algorithm proposed in (Silander and Myllmaki, 2006) applied to all the data pooled together. We can use the same algorithm to fit independent DAGs for each action, by partitioning the data. In the conditional case, there are two ways to proceed. In the first case, which we call **perfect**, we assume that the interventions are perfect, and that the targets of intervention are known. In this case, it is simple to modify the standard BDeu score to handle the interventional data, as described in Cooper and Yoo (1999). These modified scores can then be used inside the same dynamic programming algorithm. In the second case, which we call **uncertain**, we learn the structure of an augmented DAG containing A and X nodes, subject to the constraint that there are no $A \rightarrow A$ edges or $X \rightarrow A$ edges. It is simple to modify the DP algorithm to handle this; see (Eaton and Murphy, 2007) for details.

2.3. Approaches based on undirected graphs

DAG structure learning is computationally expensive due to the need to search in a discrete space of graphs. In particular, the exact dynamic programming algorithm mentioned above takes time which is exponential in the number of nodes. Recently, computationally efficient methods for learning undirected graphical model (UGM) structures, based on L1 regularization and convex optimization, have become popular, both for Gaussian graphical models (Meinshausen and Buhlmann, 2006; Friedman et al., 2007; Banerjee et al., 2008), and for Ising models (Wainwright et al., 2006; Lee et al., 2006). In the case of general discrete-state models, such as the ternary T-cell data, it is necessary to use a *group* L1 penalty, to ensure that all the parameters associated with each edge get "knocked out" together. Although still convex, this objective is much harder to optimize (see e.g., (Schmidt et al., 2008) and (Duchi et al., 2008) for some suitable algorithms). However, for the small problems considered in this paper, we found that using L2 regularization on a fully connected graph did just as well as L1 regularization, and was much faster. The strength of the L2 regularizer is chosen by cross validation.

To apply this technique in the *ignore* scenario, we construct a Markov random field, where we create factors for each X_i node and each $X_i - X_j$ edge. For the *independent* scenario, one such Markov random field is learned for each action combination in the training set. In the *interventional* scenario, we construct a conditional random field, in which we additionally create factors for each $X_i - A_j$ edge, and for each X_i, X_j, A_k triple (this is similar to a chain graph; see (Lauritzen and Richardson, 2002) for a discussion.) Since it does not contain directed edges, it is harder to interpret from a causal perspective. Nevertheless, in Section 3.1, we show that the resulting model performs very well at the task of predicting the effects of interventions.

2.4. Other methods

There are of course many other methods for (conditional) density estimation. As a simple example of a non graph based approach, we considered mixtures of K multinomials. In the ignore case, we pool the data and fit a single model. In the independent case, we fit a separate model for each action combination. In the conditional case, we fit a mixture of independent logistic regressions:

$$p(\mathbf{x}|\mathbf{a}) = \sum_k p(z=k) \prod_{j=1}^{d} p(x_j|z=k,\mathbf{a}) \tag{3}$$

where $p(z=k)$ is a multinomial, and $p(x_k|\mathbf{a}, z=k)$ is multinomial logistic regression. This is similar to a mixture of experts model (Jordan and Jacobs, 1994).

2.5. Summary of methods

In summary, we have discussed 10 methods, as follows: 3 models (Mixture Model, UGM or DAG), times 3 types (ignore, independent, conditional), plus perfect intervention DAGs. We did not try independently trained DAGs, because it was substantially

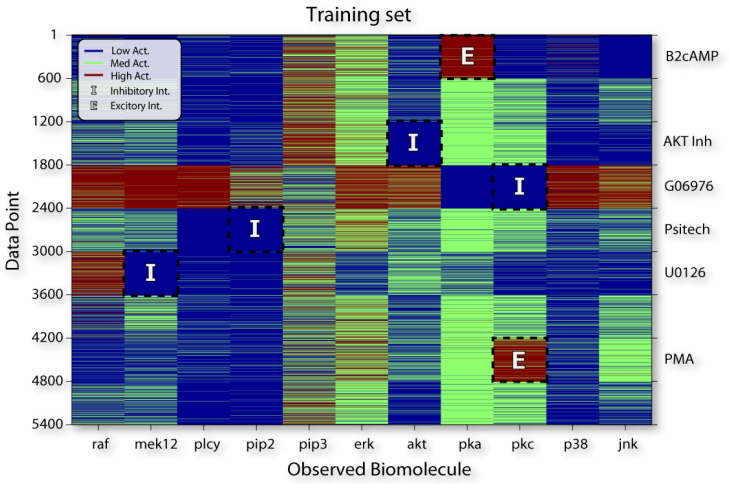


Figure 1: **T-cell data.** 3-state training data from (Sachs et al., 2005). Columns are the 11 measured proteins, rows are the 9 experimental conditions, 3 of which are "general stimulation" rather than specific interventions. The name of the chemical that was added in each case is shown on the right. The intended primary target is indicated by an E (for excitation) or I (for inhibition). There are 600 measurements per condition. This figure is best viewed in colour.

slower than other methods (using exact structure learning), so we only consider 9 methods in total.

3. Experimental results

In the introduction, we argued that, in the absence of a ground truth graph structure (which in general will never be available), the only way to assess the accuracy of a causal model is to see how well it can predict the effects of interventions on unseen test data. In particular, we assume we are given a training set of $(\mathbf{a}, \mathbf{x})$ pairs, we fit some kind of conditional density model $p(\mathbf{x}|\mathbf{a})$, and then assess its predictive performance on a different test set of $(\mathbf{a}, \mathbf{x})$ pairs.

3.1. T-cell data

Flow cytometry is a method for measuring the "status" of a large number of proteins (or other molecules) in a high throughput way. In an influential paper in Science in 2005, Sachs et al. used flow cytometry to collect a dataset of 5400 samples of 11 proteins which participate in a particular pathway in T-cells. They measured the protein phosphorylation levels under various experimental conditions. Specifically, they applied 6 different chemicals separately, and measured the status of the proteins; these chemicals

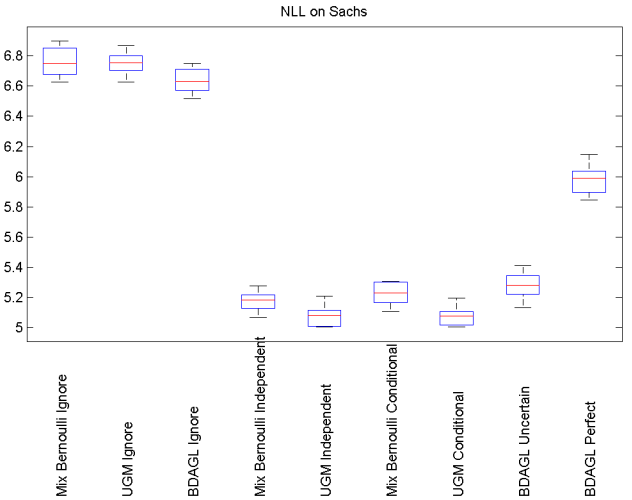


Figure 2: 10-fold cross-validated negative log likelihood on the T-cell data (lower is better). The methods are divided based on their approach to modeling interventions (*Ignore* the interventions, fit *Independent* models for each intervention, fit a *Conditional* model that conditions on the interventions, or assume *Perfect* interventions). Within each group, we sub-divide the methods into MM (mixture of multinomials), UGM (undirected graphical model), and DAG (directed acyclic graphical model).

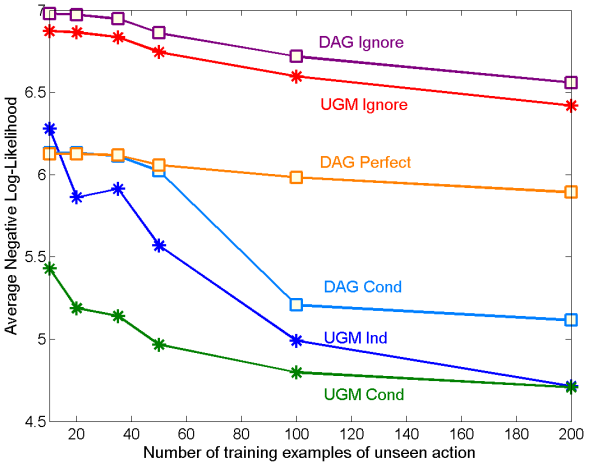


Figure 3: Average (per-case) negative log-likelihood on the T-cell test data as a function of the amount of training data for one particular action regime, given the data from all other action regimes. Results when choosing other actions for the "sparse training regime" are similar. "DAG Cond" is a DAG with uncertain interventions. "UGM Ind" is a UGM fit independently for each action.

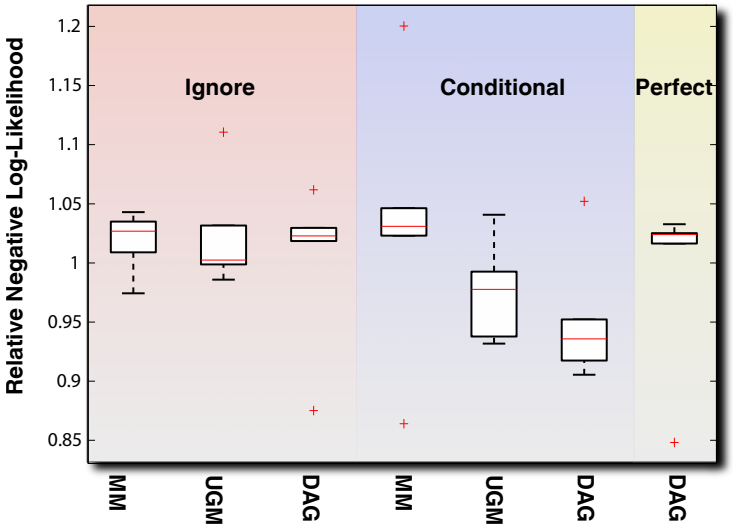


Figure 4: Negative log-likelihood on the T-cell data for different methods when predicting a novel action, using data from all the other actions as training. The boxplot shows the variation when different actions are chosen as the prediction targets. We plot performance relative to the mean over all methods for each chosen action, since some actions are easier to predict than others.

were chosen because they target the state of individual proteins. They also measured the status in the unperturbed state (no added chemicals).[1] Sachs et al. then discretized the data into 3 states, representing low, medium and high activation (see Figure 1), and learned a DAG model using simulated annealing and the scoring function described in (Cooper and Yoo, 1999). The resulting DAG was quite accurate, in that it contained most of the known edges in the biological network, and few false positives. However, it is known that the "true" graph structure contains feedback loops, which cannot be modeled by a DAG. In addition, there are many variables in the "true" model that are not measured in the data. Hence assessing performance by looking at the graph structure is not ideal. Instead, we will measure predictive accuracy of the learned models.

We used the same *discretized* version of the data as in the original Sachs paper. There are 600 samples in each interventional regime, and 1800 samples in the observational regime, for a total of 5400 samples. There is no pre-specified train/test split in the T-cell data, so we have to make our own. A natural approach is to use cross validation, but a subtlety arises: the issue is whether the test set folds contain novel action combinations or not. If the test data contains an action setting that has never been seen before, in general we cannot hope to predict the outcome, since, for example, the distribution $p(\mathbf{x}|a_1 = 0, a_2 = 1)$ need have nothing in common with $p(\mathbf{x}|a_1 = 1, a_2 = 0)$.

Initially we sidestep this problem and follow the approach taken by Ellis and Wong (2008), whereby we assess predictive performance using 10-fold cross validation, where the folds are chosen such that each action occurs in the training and test set. Hence each training set has 540 samples and each validation set has 60 samples.

The results of evaluating various models in this way are shown in Figure 2. We see that the methods which ignore the actions, and pool the data into a single model, do poorly. This is not surprising in view of Figure 1, which indicates that the actions do have a substantial affect on the values of the measured variables. We also see that the approach that learns the targets of intervention (the *conditional DAG*) is significantly better than learning a DAG assuming that the interventions are perfect (see last two columns of Figure 2). Indeed, as discussed in Eaton and Murphy (2007), the structure learned by the uncertain DAG model indicates that each intervention affects not only its suspected target, but several of its neighbors as well. The better prediction performance of this model indicates that the perfect intervention assumption may not be appropriate for this data set. However, we also see that *all* the independent and conditional models not based on DAGs do as well or better than the DAG methods.

It was somewhat surprising how well the independent models did. This is presumably because we have so much data in each action regime, that it is easy to learn separate models. To investigate this, we considered a variant of the above problem in which we trained on all 600 samples for all but one of the actions, and for this remaining action we trained on a smaller number of samples (and tested only on this remaining action). This allows us to assess how well we can borrow statistical strength from the data-rich regimes to a data-poor regime. Figure 3 shows the results for several of the

1. This original version of the data is available as part of the 2008 Causality Challenge. See the CYTO dataset at http://www.causality.inf.ethz.ch/repository.php.

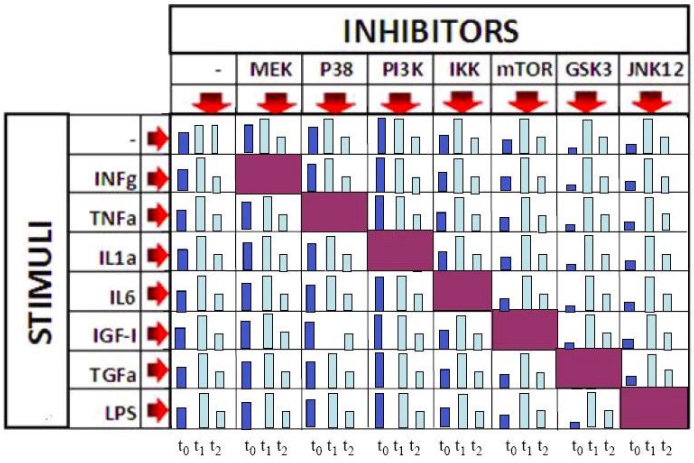


Figure 5: Dream 3 phosphoprotein data. See text for details.

models on one of the actions (the others yielded largely similar results). We see that the conditional models need much less training data when faced with a novel action regime than independent models, because they can borrow statistical strength from the other regimes. Independent models need much more data to perform well. Note that even with a large number of samples, the perfect DAG model is not much better than fitting a separate model to each regime.

The logical extreme of the above experiment is when we get no training samples from the novel regime. That is, we have 600 training samples from each of the following: $p(\mathbf{x}|1,0,0,0,0,0)$, $p(\mathbf{x}|0,1,0,0,0,0)$, ... $p(\mathbf{x}|0,0,0,0,1,0)$, and we test on 600 samples from $p(\mathbf{x}|0,0,0,0,0,1)$, where the bit vector on the right hand side of the conditioning bar specifies the state of the 6 A_j action variables. We can then repeat this using leave-one-action out. The results are shown in Figure 4. (We do not show results for the independently trained models, since their predictions on novel regimes will be based solely on their prior, which is essentially arbitrary.) We see that all methods do about the same in terms of predictive accuracy. In particular, the perfect DAG model, which is designed to predict the effects of novel actions, is actually slightly worse than conditional DAGs and conditional UGMs in terms of its median performance.

3.2. DREAM data

One weakness of the CYTO dataset discussed above is that the actions are only performed one at a time. A more recent dataset has been collected which measures the status of proteins under different action combinations.[2] This data is part of the DREAM 3 competition, which took place in November 2008. (DREAM stands for "Dialogue for Reverse Engineering and Assessment of Methods".) The data consists of measure-

2. This data is available from http://wiki.c2b2.columbia.edu/dream/index.php/The_Signaling-Response_Prediction_Challenge._Description.

Stimulus							Inhibitor								
INFg	TNFa	IL1a	IL6	IGF1	TGFa	LPS	MEK	P38	P13K	IKK	mTOR	GSK3	JNK12	X1	X17
0	0	0	0	0	0	0	0	0	0	0	0	0	0	5578	275
0	0	0	0	0	0	0	1	0	0	0	0	0	0	454	89
0	0	0	0	0	0	0	0	1	0	0	0	0	0	1001	99
						⋮								⋮	
0	0	0	0	0	0	1	0	0	0	0	0	1	0	22	33

Figure 6: The dream 3 training data represented as a design matrix. We treat each cell type and time point separately, and show the response of the 17 phosphoproteins to 58 different action combinations (58 is 8×8 minus the 6 test conditions shown in Figure 5.) Each 14-dimensional action vector has 0, 1 or 2 bits turned on at once. For example, the last row corresponds to stimulus=LPS, inhibitor = GSK3.

Team	MSE
PMF	1483
Linear regression	1828
Team 102	3101
Team 106	3309
Team 302	11329

Figure 7: Mean squared error on the DREAM 3 dataset, using the training/test set supplied with the challenge. Also listed is the performance of the three other teams who competed in the challenge.

ments (again obtained by flow cytometry) of 17 phosphoproteins and 20 cytokines at 3 time points in 2 cell types under various combinations of chemicals (7 stimuli and 7 inhibitors). In the challenge, the response of the proteins under various stimulus/ inhibitor pairs is made available, and the task is to predict the response to novel stimulus/ inhibitor combinations. In this paper, we focus on the phosphoprotein data. The data is illustrated in Figure 5. Another way to view this data is shown in Figure 6.

The DREAM competition defines a train/test split, and evaluates methods in terms of their mean squared error for predicting the responses of each variable separately to 6 novel action combinations. In Table 7, we show the scores obtained by the 3 entrants to the competition in November 2008. The method used by these teams has not yet been disclosed, although the organizer of the Dream competition (Gustavo Stolovitzky) told us in a personal communication that they are not based on graphical models. We also show two approaches we tried. The first uses simple linear regression applied to the 14-dimensional binary action vector $\mathbf{a}$ to predict each response X_j (since the methods are evaluated in terms of mean squared-error, this is equivalent to using a conditional DAG model with linear-Gaussian CPDs) We see that this beats all the submitted entries by a

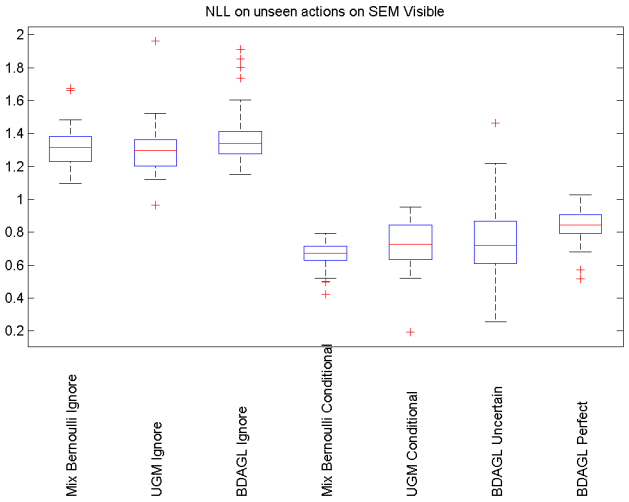


Figure 8: Negative log-likelihood for novel action combinations on synthetic data generated from a fully visible SEM. We plot NLL relative to the mean performance over all methods on each action.

large margin. However, the significance of this result is hard to assess, because there is only a single train/test split. We also tried probabilistic matrix factorization, using $K = 3$ latent dimensions. This is similar to SVD/PCA but can handle missing data (see Salakhutdinov and Mnih (2008) for details). This choice was inspired by the fact that the data matrix in Figure 5 looks similar to a collaborative filtering type problem, where the goal is to "fill in" holes in a matrix. We see that PMF does even better than linear regression, but again it is hard to assess the significance of this result. Hence in the next section, we will discuss a synthetic dataset inspired by the design of the DREAM competition.

3.3. Synthetic Data

Since the DREAM data uses population averaging rather than individual samples, it does not contain enough information to learn a model of the underlying system. Thus, we sought to validate some of the approaches discussed here on a synthetic data set. To this end, we generated synthetic data sets that simulate the DREAM training/testing regime (i.e., where we train on pairs of actions and test on novel pairs).

We sought to generate a data set that has a clearly defined notion of intervention, but that is not a DAG. To do this we simulated data from a discrete structural equation model (SEM) (see Pearl (2000)). In particular, we generated a data set where each

variable X_j is updated based on

$$p(X_j = 1|\mathbf{x}_{\pi_j}, \boldsymbol{\theta}_j) = \sigma(w_{0j} + \mathbf{w}_j^T \mathbf{x}_{\pi_j}) \quad (4)$$

$$p(X_j = -1|\mathbf{x}_{\pi_j}, \boldsymbol{\theta}_j) = 1 - p(X_j = 1|\mathbf{x}_{\pi_j}, \boldsymbol{\theta}_j) \quad (5)$$

where $\sigma()$ is the sigmoid function $\sigma(x) \triangleq 1/(1+\exp(-x))$, and $\boldsymbol{\theta}_j = (w_{0j}, \mathbf{w}_j)$ are the paramters for each node; here w_{0j} is the bias term and $\mathbf{w}_j$ are the regression weights. We generated each w_0 from a standard Normal distribution, and to introduce strong dependencies between nodes we set each element of each $\mathbf{w}$ vector to $U_1 + 5sgn(U_2)$, where U_1 and U_2 were generated from a standard Normal distribution. For each node j, we included each other node in its parent set π_j with probability 0.25. To generate samples that approximate the equilibrium distribution of the model, we started by sampling each node's value based on its bias w_0 alone, then we performed 1000 updates, where in each update we updated all nodes whose parents were updated in the previous iteration. We assume perfect interventions, which force a variable into a given state. In the special case where the dependency structure between the nodes is acyclic, this sampling procedure is exactly equivalent to ancestral sampling in a DAG model (and the update distributions are the corresponding conditional distributions), and these interventions are equivalent to perfect interventions in the DAG. However, we do not enforce acyclicity, so the distribution may have feedback cycles (which are common in biological networks).

We considered 2 variants of this data, one where all variables are visible, and one with hidden variables (as is common in most real problems). In the *visible* SEM data set, we generated from an 8-node SEM model under all 28 pairs of action combinations. In our experiments, we trained on 27 of the action pairs and tested on the remaining action pair, for all 28 pairs. In the *hidden* SEM data set, we generated from a 16-node SEM model under the 28 pairs of actions combinations for the first 8 nodes, but we treat the odd-numbered half of the nodes as hidden (so half of the actions affect a visible node in the model, and half of the actions affect a hidden node). We think that this is a slightly more realistic synthetic data set than a fully visible DAG with perfect interventions, due to the presence of hidden nodes and feedback cycles, as well as interventions that affect both visible and hidden nodes. When the data is visualized, it looks qualitatively similar to the T-cell data in Figure 1 (results not shown).

The results on the visible data are shown in Figure 8. Since we are only testing on new action combinations, independent models cannot be applied. As expected, conditional models do better than ignore models. However, amongst the conditional models there does not appear to be a clear winner. In particular, DAG models, even perfect DAGs which are told the target of intervention, do no better than non-DAG models.

The results on the hidden data are not shown, since they are qualitatively similar to the visible case. Note that in this setting, we cannot use the perfect intervention model, since some of the interventions affected hidden nodes; hence the target of intervention is not well defined. We have obtained qualitatively similar results on other kinds of synthetic data.

4. Conclusions

In this paper, we have argued that it is helpful to think of causal models in functional terms, and to evaluate them in terms of their predictive performance, rather than in terms of graph structures that they learn. In particular, we view causal modeling as equivalent to learning a conditional density model of the form $p(\mathbf{x}|\mathbf{a})$.

A criticism of this work could be that we are not really doing causality because we can't predict the effects of new actions. However, in general, this is impossible unless we know something (or assume something) about the new action, since in general $p(\mathbf{x}|a_1 = 1, a_2 = 0)$ need have nothing to do with $p(\mathbf{x}|a_1 = 0, a_2 = 1)$. Indeed, when we tested the ability of various methods, including causal DAGs, to predict the effects of a novel action in the T-cell data, they all performed poorly — not significantly better than methods which ignore the actions altogether. This is despite the fact that the DAG structure we were using was the globally optimal DAG, which had previously been shown to be close to the "true" structure, and that we knew what the targets of the novel action were.

We think a promising direction for future work is to describe actions, and/or the variables they act on, in terms of feature vectors, rather than treating them as atomic symbols. This transforms the task of predicting the effects of new actions into a standard structured prediction problem, that could be addressed with CRFs, M3Ns, etc. Just like predicting the labeling of a new sentence or image given only its features, if there is some regularity in the action-feature space, then we can predict the effects of a new action given only the features of the action, without ever having to perform it.

Acknowledgments

We would like to thank Guillaume Alain for help with some of the experiments, and Gustavo Lacerda and Oliver Schulte for comments on a draft version of this paper.

References

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. of Machine Learning Research*, 9:485–516, 2008.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *UAI*, 1999.

A. P. Dawid. Influence diagrams for causal modelling and inference. *Intl. Stat. Review*, 70:161–189, 2002. Corrections p437.

A. P. Dawid. Beware of the DAG! *J. of Machine Learning Research*, 2009. To appear.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *UAI*, 2008.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *AI/Statistics*, 2007.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *UAI*, 2005.

F. Eberhardt, C. Glymour, and R. Scheines. N-1 experiments suffice to determine the causal relations among N variables. In *Innovations in Machine Learning*. Springer, 2006.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation the graphical lasso. *Biostatistics*, 2007.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

S. Lauritzen. Causal inference from graphical models. In D. R. Cox O. E. Barndoff-Nielsen and C. Klueppelberg, editors, *Complex stochastic systems*. 2000.

S. Lauritzen and T. Richardson. Chain graph models and their causal interpretations. *J. of the Am. Stat. Assoc.*, 3(64):321–361, 2002.

S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using L1-regularization. In *NIPS*, 2006.

N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge Univ. Press, 2000.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure Learning in Random Fields for Heart Motion Abnormality Detection. In *CVPR*, 2008.

T. Silander and P. Myllmaki. A simple approach for finding the globally optimal Bayesian network structure. In *UAI*, 2006.

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

M. Wainwright, P. Ravikumar, and J. Lafferty. Inferring graphical model structure using ℓ_1-regularized pseudo-likelihood. In *NIPS*, 2006.

JMLR: Workshop and Conference Proceedings 6:191-202, 2010 NIPS 2008 workshop on causality

Discover Local Causal Network around a Target to a Given Depth

You Zhou ZHOUYOU@PKU.EDU.CN
Changzhang Wang CHANGZHANG@PKU.EDU.CN
Jianxin Yin JIANXINYIN@GMAIL.COM
Zhi Geng ZGENG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing 100871, China

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

For a given target node T and a given depth $k \geq 1$, we propose an algorithm for discovering a local causal network around the target T to depth k. In our algorithm, we find parents, children and some descendants (PCD) of nodes stepwise away from the target T until all edges within the depth k local network cannot be oriented further. Our algorithm extends the PCD-by-PCD algorithm for prediction with intervention presented in Yin et al. (2008). Our algorithm can construct a local network to depth k, has a more efficient stop rule and finds PCDs along some but not all paths starting from the target.

Keywords: Causal network, Local structural learning

1. Introduction

In some applications, we may be interested in discovering a local causal network around a target variable rather than the whole network over all variables. For example, we want to predict the target in the cases with external interventions, or we are interested in direct and indirect causes of a disease and further discriminate direct causes from other indirect causes. There are many algorithms for structural learning, but most of them are for constructing a whole network over all variables, such as Pearl (2000); Spirtes et al. (2000); Heckerman (1999); Tsamardinos et al. (2006); Xie et al. (2006) and Xie and Geng (2008). To discover a local causal network, however, it is inefficient to construct the whole network over a large number of variables. In Causation and Prediction Challenge of IEEE WCCI2008, Yin et al. (2008) proposed local structural learning approaches for prediction with external interventions, in which only edges connecting to the target are discovered and oriented. But it cannot be used to discover a larger local structure or more indirect causes of the target.

In this paper, for a given target node T and a given depth $k \geq 1$, we propose an algorithm for discovering a local causal network around the target T to depth k. Our al-

gorithm extends the PCD-by-PCD algorithm for prediction with intervention presented in Yin et al. (2008). First our algorithm can construct a depth k local network, and Yin's PCD-by-PCD algorithm is a special case of the depth 1. Second, our algorithm has a more efficient stop rule than Yin's algorithm. In Yin's algorithm, a main stop condition is 'until all edges connecting the target are oriented', but in our algorithm, we make this condition weaker so that our algorithm can stop earlier than Yin's algorithm without loss of validity. Third, our algorithm continues to find PCDs along only some paths away from the target which are necessary to orient the undirected edges within the depth k local network, while Yin's algorithm continues to find PCDs along all paths starting from the target.

In Section 2, we propose the local structural learning algorithm. In Section 3, we theoretically show the correctness of our algorithm. Section 4 gives definitions of scores to be used for evaluation of algorithm performance. In Section 5, we compare our algorithms with other algorithms via simulation. We discuss the challenge task: LOCANET in Section 6. Discussion is given in Section 7. Proof of theorem is shown in Appendix.

2. Learning a local structure around the target to a given depth

Let U denote the full set of all nodes. For a node u, let $PC(u)$ denote the set of all parents and all children of u, and let $PCD(u)$ denote a set which contains $PC(u)$ and may contain some descendants of u. There are several algorithms which can be used to find $PCD(u)$, such as Min Max Parent and Children (MMPC) algorithm Tsamardinos et al. (2006).

Let T be the target node. Suppose that we are interested in the local network around the target T to a depth k. In our algorithm, we first find parents, children and some descendants (PCD) of the target T to obtain a local skeleton with a radius 1, and then repeatedly find PCDs of nodes in the previous PCDs until the radius of the local skeleton is up to the given depth k. To orient the edges in the local skeleton, we may need to find more PCDs further away from the target T along some but not all paths. We expect to orient all undirected edges within the local network, but some of the undirected edges cannot be oriented essentially from observational data even if we construct a correct global network, which is an equivalence class of causal networks. Thus we propose a stopping rule so that the process of finding PCDs can stop early even if some edges within the local network are unoriented. Our stopping rule is based on the fact that when the unoriented edges are surrounded by directed edges, they cannot be oriented by finding further structures. We theoretically show that our algorithm can correctly obtain the local causal network with the given depth. Our algorithm does not need to construct the global network and thus it can greatly reduce computational complexity of structural learning.

In the following algorithm, we separate the process into two parts. Part 1 is to find edges within length $k-1$ from the target T. Part 2 is to find edges at the outer layer k and to orient undirected edges within the local network with depth k. When $k=1$, we only need to run Part 2 but no need to run Part 1.

Algorithm 1: Local structural learning around the target T to depth k

Part 1: Find edges within length $k-1$ from the target T.

1 **Initialization:** Find the PCD of T, $PCD(T)$.
$V = \{T\}$ (V is a set of variables whose PCDs have been obtained)
$Layer(0) = \{T\}$, $Layer(i) = \emptyset$ for $i = 1, \ldots k$ ($Layer(i)$ is the node set on layer i)
$TotalLay = \{T\}$, (The total set of nodes on all layers)
$depth = 1$, (the counter of depth)
$canU(1) = PCD(T)$, $canU(i) = \emptyset$ for $i = 2, \ldots, k$;
($canU(i)$ is an ordinal waiting list for layer i whose PCD will be found.)
Repeat
2 Take X from the head of list $canU(depth)$ out.
3 If $X \notin V$ (i.e., $PCD(X)$ has not been gotten before) then
Find $PCD(X)$, and set $V = V \cup \{X\}$.
For each $Y \in V$, if [$X \in PCD(Y)$ and $Y \in PCD(X)$],
then create an undirected edge (X, Y).
Find v-structures within V including X:
{Within V, find possible v-structures only for the triple of X and other two variables in V if an intermediate node is not in the separator set of two nonadjacent nodes.}
Orient undirected edges within V:
{Orient other edges between nodes in V if each opposite of them creates either a directed cycle or a new v-structure Meek (1995).}
End if
4 If $X \notin TotalLay$ and $X \notin Layer(depth)$ and X is adjacent to
a node in $Layer(depth-1)$ then
$Layer(depth) = Layer(depth) \cup \{X\}$,
add $PCD(X) \backslash TotalLay$ to the tail of list $canU(depth+1)$
End if.
5 If $canU(depth) = \emptyset$ then
$TotalLay = TotalLay \cup Layer(depth)$ and $depth = depth + 1$
End if
6 **Until** $canU(depth) = \emptyset$ or $depth \geq k$.

Algorithm 1 (continued)

Part 2: Find edges at layer k and orient undirected edges within the local structure.

Initialization. Set the list of nodes at the layer $k-1$ whose PCDs will further be found:

$canV = \{struct('leaf', v, 'length', 1, 'path', u) : u \in Layer(depth = k-1), v \in PCD(u) \backslash TotalLay\}$

Repeat

Take X from the head of the list $canV$ out.

If all edges on path X.path are undirected then

If $X.leaf \notin V$ then

find $PCD(X.leaf)$ and set $V = V \cup \{X.leaf\}$;

for each $Y \in V$, if $X.leaf \in PCD(Y)$ and $Y \in PCD(X.leaf)$, then create an undirected edge (X.leaf,Y);

find v-structures within V including $X.leaf$;

orient undirected edges within V.

End if.

If there is an undirect edge between X.leaf and the last node u of $X.path$ then

add $\{struct('leaf', v, 'length', X.length+1, 'path', [X.path, X.leaf]) : v \in PCD(X.leaf) \backslash X.path \backslash TotalLay\}$ to the tail of $canV$

End if.

End if.

Until $canV = \emptyset$.

Return

Example. Consider the revised ALARM network in Figure 1 where the arrows $16 \to 20$ and $25 \to 20$ in the original ALARM Beinlich et al. (1989) are reversed as $16 \leftarrow 20$ and $25 \leftarrow 20$ respectively. The revision makes more edges unoriented in its Markov equivalence class and thus it becomes more complicated for structural learning. Suppose that we want to discover a local network around node 20 to depth 2, denoted as $G_2(20)$. Applying Part 1 of our algorithm, we obtain a local network with all edges undirected as shown in Figure 2. Applying Part 2, we first obtain the local network with $depth = 2$ in Figure 3. Since there are some edges unoriented, we extend the network along undirected paths to orient these undirected edges. Finally Part 2 returns a local network as shown in Figure 4, which is larger than $G_2(20)$ and has four nodes 9, 13, 19 and 21 outside $G_2(20)$. Nodes 19 and 21 are used to find two v-structures $19 \to 15 \leftarrow 18$ and $21 \to 17 \leftarrow 16$ respectively, and thus they help to orient edges $17 \leftarrow 16$ and $15 \leftarrow 18$ within $G_2(20)$. Nodes 9 and 13 are used to find a v-structure $13 \to 9 \leftarrow 14$ such that all undirected edges within $G_2(20)$ are surrounded by directed edges, and thus the algorithm stops.

3. Theoretical result for algorithm's correctness

We show below the correctness of the algorithm proposed in the previous section.

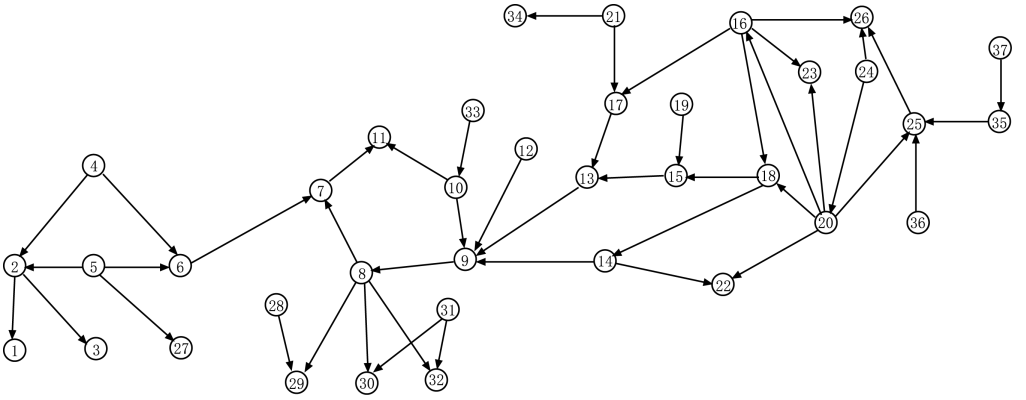


Figure 1: A revised ALARM

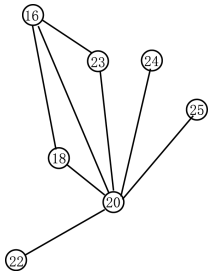


Figure 2: Network by Part 1

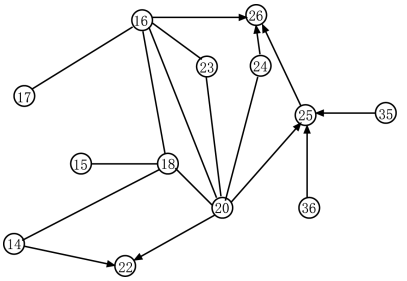


Figure 3: Network to $depth = 2$ by Part 2

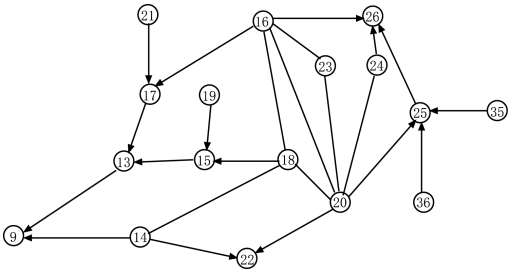


Figure 4: Network returned by Part 2

Theorem 1 *Suppose that a causal network is faithful to a probability distribution and all conditional independencies are correctly checked by using data. Then the algorithm proposed in the previous section can correctly discover the edges within the depth k local causal network around the target variable T. Further it can obtain the same orientations of these edges as a partially directed graph for the Markov equivalence class of the underlying global causal network.*

The proof of this theorem is given in Appendix. Under the suppositions of the faithfulness and correctness of conditional independence tests, the above result ensures that our algorithm can return the correct local network. Notice that some edges in the local network may not be oriented. It is because these edges cannot be oriented by using data from observational studies, but it is not because our algorithm does not finish the learning process of the whole network.

4. The scores for evaluation

In this section, we introduce the two kinds of evaluation methods that are used in the causal challenge to evaluate the performance for discovering a local causal network (Guyon et al., 2008). The first method uses the average edit distance score. In the causal challenge, the task is to construct a depth 3 causal network around a given target variable. Thus the relationship of a variable to the target variable is encoded as a string of up (u) and down (d) arrows from the target:

- Depth 1 relatives: parents (u) and children (d);
- Depth 2 relatives: spouse (du), grand-children (dd), siblings (ud), grand-parents (uu); and
- Depth 3 relatives: great-grand-parents (uuu), uncles/aunts (uud), nices/nephews (udd), parents of siblings (udu), spouses of children (ddu), parents in law (duu), children of spouses (dud), great-grand-children (ddd).

A confusion matrix $C = \{C_{ij}\}$ is defined to record the number of relatives confused for another type of relative among 14 types of relatives in a depth 3 network. A cost matrix $A = \{A_{ij}\}$ is defined to account for the distance between the true and obtained relatives, as shown in Table 1. The edit distance score is defined as

$$S = \sum_{i,j} A_{ij} C_{ij}.$$

The second method uses a score-pair (precision, recall) for each kind of variable subsets: parents, children, Markov blanket, all depth 1 variables, all depth 2 variables, all depth 3 variables. Precision and recall (also called sensitivity) are defined respectively as:

- Precision = # of true positive found/# of found, and

Depth	Desired		1	1	2	2	2	2	3	3	3	3	3	3	3	3	X
Obtained	Relationship		P	C	Sp	GC	Si	GP	GGP	uud	N	PS	SC	IL	CP	GGC	Other
			u	d	du	dd	ud	uu	uuu	uud	udd	udu	ddu	duu	dud	ddd	
1	Parents	u	0	1	1	2	1	1	2	2	2	2	2	2	2	3	4
1	Children	d	1	0	1	1	1	2	3	2	2	2	2	2	2	2	4
2	Spouse	du	1	1	0	1	2	1	2	2	2	1	1	1	1	2	4
2	Gchildren	dd	2	1	1	0	1	2	3	2	1	2	1	2	1	1	4
2	Siblings	ud	1	1	2	1	0	1	2	1	1	1	2	2	1	2	4
2	Gparents	uu	1	2	1	2	1	0	1	1	2	1	2	1	2	3	4
3	Ggparents	uuu	2	3	2	3	2	1	0	1	2	1	2	1	2	3	4
3	Uncles/Aunts	uud	2	2	2	2	1	1	1	0	1	2	3	2	1	2	4
3	Nieces/Nephews	udd	2	2	2	1	1	2	2	1	0	1	2	3	2	1	4
3	ParentsOfSiblings	udu	2	2	1	2	1	1	1	2	1	0	1	2	2	2	4
3	SpousesOfChildren	ddu	2	2	1	1	2	2	2	3	2	1	0	1	2	1	4
3	ParentsInLaw	duu	2	2	1	2	2	1	1	2	3	2	1	0	1	2	4
3	ChildrenOfSpouses	dud	2	2	1	1	1	2	2	1	2	2	2	1	0	1	4
3	GgChildren	ddd	3	2	2	1	2	3	3	2	1	2	1	2	1	0	4
X	Other		4	4	4	4	4	4	4	4	4	4	4	4	4	4	0

Table 1: A cost matrix $A = \{A_{ij}\}$.

- Recall = # of true positive found/ # of true positive.

In the cases with a 0 denominator, a very small number are added to both the numerator and the denominator.

5. Simulation

In this section, we compare the algorithm proposed in this paper with other algorithms via simulations. Consider again the example in Section 2 and the goal is to get the depth 3 network around node 20. In Table 2, we show the simulation results for the revised ALARM network depicted in Figure 1. We compare our algorithm (PCD-path) with the PC algorithm, the MMHC algorithm proposed by Tsamardinos et al. (2006) and the recursive algorithm proposed by Xie and Geng (2008). The 'distscore' is the edit distance score defined for the task LOCANET to measure the difference between the obtained local network and the true local network. We consider several cases with different significance levels and different sample sizes. In the simulation, we do 1000 repetitions and obtain average values for each case of different sample size n and significance level α. For each repetition, we draw a training data set from the distribution whose parameters for the unchanged structures are obtained from the FullBNT code package: `http://www.cs.ubc.ca/~murphyl/Software/BNT/bnt.html`, and parameters for the changed structure are set by chance. All of our computations are performed on a computer with CPU 2.1 GHz×2 and 2 GB RAM. 'CPU time' is the total CPU time of 1000 repetitions for each algorithm. It can be seen from Table 2 that our algorithm takes much less CPU times and it has also less distscores than other three algorithms for every case.

n	α	Algorithm	distscore	CPU time (second)
500	0.05	PCD-path	1.0305	881
		Recursive	1.1635	3,293
		MMHC	1.1162	3,405
		PC	1.2213	11,638
	0.10	PCD-path	1.0993	1,175
		Recursive	1.2057	3,404
		MMHC	1.1692	3,515
		PC	1.3150	11,979
	0.15	PCD-path	1.1621	1,489
		Recursive	1.2498	3,509
		MMHC	1.1919	3,949
		PC	1.4083	12,457
1000	0.05	PCD-path	0.8573	993
		Recursive	1.1205	3,739
		MMHC	1.1133	3,864
		PC	1.0804	8,823
	0.10	PCD-path	0.8836	1,204
		Recursive	1.2958	3,889
		MMHC	1.1431	4,326
		PC	1.1241	9,635
	0.15	PCD-path	0.9082	1,415
		Recursive	1.3702	4,008
		MMHC	1.1724	4,823
		PC	1.1508	10,440

Table 2: Comparison of algorithms for the revised ALARM network.

In Table 3, we give the total (precision, recall) scores in 1000 simulations and the average scores can be obtained by dividing it by 1000. By the (precision, recall) scores, there is no algorithm which always is better than others. The Recursive one seems to be better averagely. In some cases, the PCD-path algorithm seems to be better at 'pa' and 'ch' than the MMHC algorithm, but worse at 'pc' and 'mb'. From Tables 2 and 3, it can be seen that the PCD-path algorithm proposed in this paper runs fastest among the four algorithms without loss of performance. The main advantage of the PCD-path algorithm is to construct a local network around the given target, and this is more important for the cases with a large number of variables.

n	α	Algorithm	precision						recall					
			pa	ch	pc	mb	D2	D3	pa	ch	pc	mb	D2	D3
500	0.05	PCD-path	41	883	904	873	755	669	93	364	713	640	686	402
		Recursive	282	747	971	824	751	692	971	705	780	876	788	573
		MMHC	50	768	944	933	833	777	21	652	928	887	909	656
		PC	41	932	836	822	823	901	222	441	861	839	713	377
	0.10	PCD-path	145	836	894	853	812	780	307	472	836	751	782	502
		Recursive	262	750	967	822	750	692	934	717	789	880	802	597
		MMHC	56	749	943	935	829	738	28	651	945	900	918	688
		PC	74	882	815	818	823	891	426	393	921	876	728	435
	0.15	PCD-path	174	821	879	839	816	774	409	519	882	796	809	542
		Recursive	246	750	962	816	742	676	893	725	794	880	806	605
		MMHC	64	739	941	933	818	707	35	657	955	908	925	703
		PC	80	848	804	808	806	846	485	378	949	893	735	469
1000	0.05	PCD-path	263	968	990	894	847	657	68	652	643	754	814	576
		Recursive	196	782	961	819	862	840	812	866	894	969	883	681
		MMHC	55	740	986	958	909	804	23	574	911	873	932	769
		PC	49	952	990	810	931	822	190	701	767	817	861	534
	0.10	PCD-path	315	911	983	900	895	780	160	722	776	824	905	688
		Recursive	169	765	935	815	845	813	755	872	904	965	863	689
		MMHC	71	748	980	953	904	801	30	617	943	914	956	804
		PC	88	906	980	802	943	882	359	759	847	879	911	627
	0.15	PCD-path	318	893	974	899	898	811	174	745	826	853	930	737
		Recursive	158	753	917	806	822	774	737	877	911	962	850	693
		MMHC	73	751	976	950	892	785	32	639	959	936	965	814
		PC	99	884	970	796	942	907	422	781	890	906	925	670

Table 3: Precision and recall of algorithms for the revised ALARM network.

6. Challenge task LOCANET

We applied the algorithm to three data sets: REGED, CINA and MARTI to find local structures around targets to depth 3. The data set MARTI is preprocessed in the way proposed by Yin et al. (2008), which is available at http://clopinet.com/isabelle/Projects/WCCI2008/MARTI/JY/ We summarize our results for the Potluck challenge task LOCANET in Table 4. 'NoNode' denotes the number of nodes

for a data set, 'NoNodeLN' denotes the number of nodes in the local network around the target to depth 3, 'NoPcds' denotes the number of PCDs found by our algorithm, and we give CPU times for every data set. Our algorithm takes so much longer on the dataset CINA than other datasets. First, the sample size of CINA is much larger than REGED and MARTI. Second, the independence tests of discrete variables for CINA runs much slower than the tests of continuous variables for REGED and MARTI under the assumption of Gaussian distribution. For the data set SIDO, there are 4933 variables, the observed frequencies are very unbalanced, some cells have very small frequencies, and some have very large ones. In this case, the approach for finding parents and children sets is not so efficient as the approach for finding Markov blankets. Thus the recursive algorithm via Markov blankets proposed by Xie and Geng (2008) is used to find a local networks including the target and 400 nodes which are strongly associated with the target.

Data set	NoNodes	NoNodeLN	NoPcds	CPU time
REGED	1000	136	212	10 minutes
CINA	133	108	116	4 hours
MARTI	1025	224	309	10 minutes

Table 4: Results for the challenge task LOCANET.

The (precision, recall) scores and the edit distance scores of our performance on the four datasets of LOCANET are shown in Figure 5. For the (precision, recall) scores, the more the symbols are in the upper right corner, the better the performance is. We have about 7 symbols in the upper right quadrant. Most of the symbols in the lower left corner are for the MARTI dataset which are generated by adding noises to the dataset REGED. The performance for the dataset REGED is quite better, and thus the noises in MARTI may not be filtered throughout in our algorithm. Since there are no known parents in the CINA task, it is not surprising that our result for the parents in the CINA is on the vertical axis (which means we have a recall 1 while precision 0). Thirteen in total 24 symbols are close to the right boundary which presents a high precision, and this means that the results we found are mostly true.

7. Conclusion

We proposed an algorithm for local structural learning of a causal network around a given target node to depth k. Our algorithm finds PCDs stepwise starting from the target node and stops the process when the local structure is obtained, and thus it can reduce the computational complexity. We theoretically show its correctness. The algorithm can be used for prediction with external interventions and for local causal discovery.

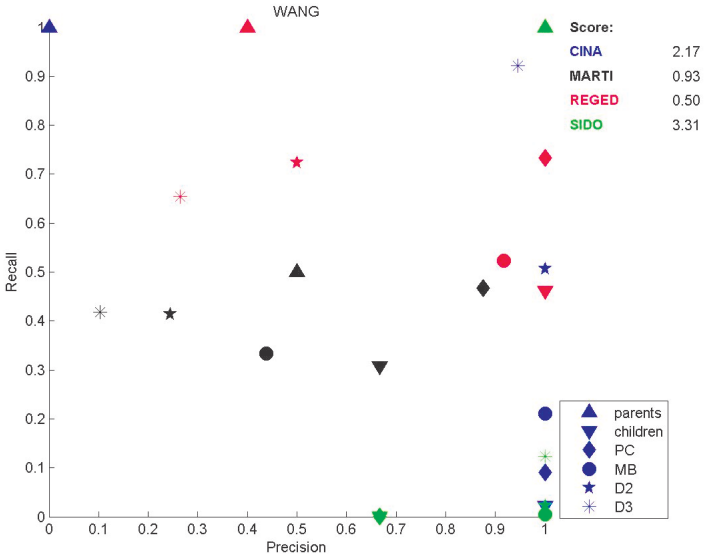


Figure 5: (Precision, Recall) scores and edit-distance scores for four datasets of LOCANET.

Acknowledgments

We would like to thank the four reviewers for their valuable comments and suggestions. We would appreciate I. Guyon and the competition committee for their encouragement and support to our work. This research was supported by NSFC (10771007, 10721403), 863 Project of China (2007AA01Z437), MSRA and MOE-Microsoft Key Laboratory of Statistics and Information Technology of Peking University.

References

I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the 2nd European Conference in Artificial Intelligence in Medicine*, pages 247–256, Germany, 1989. Springer-Verlag.

I. Guyon, A. Statnikov, and C. Aliferis. Pot-luck challenge: FACT SHEET. Technical Report.

D. Heckerman. A tutorial on learning with bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*, Cambridge, MA, 1999. MIT Press.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–41. Morgan Kaufmann, 1995.

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

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

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

X. Xie and Z. Geng. A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research*, 9:459–483, 2008.

X. Xie, Z. Geng, and Q. Zhao. Decomposition of structural learning about directed acyclic graphs. *Artificial Intelligence*, 170(4):422–439, 2006.

J. Yin, Y. Zhou, C. Wang, P. He, C. Zheng, and Z. Geng. Partial orientation and local structural learning of causal networks for prediction. In *JMLR: Workshop and Conference Proceedings*, volume 3, pages 93–104, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

Appendix

In this appendix we prove Theorem 1 presented in Section 3.

Proof We first show the correctness of Part 1. Step 1 is initialization. We take X from list $canU(depth)$ to find PCD(X) at step 2. At Step 3, we obtain an undirect edge $X - Z$ if and only if both $Z \in PCD(X)$ and $X \in PCD(Z)$, and thus all edges can be created correctly if independence tests for finding PCD are correctly performed. After finding a new edge connecting node X newly taken at Step 2, we can determine whether there is a v-structures with X as one node and other two nodes in V, such as $X - Y - Z$ with X and Z nonadjacent and all X, Y and Z in the set V. It is because X, Y and Z are all in the set V, and edges between them have been correctly determined. We can correctly find a v-structure $X \rightarrow Y \leftarrow Z$ if Y is not in the separator X and Z (that is, $X \perp\!\!\!\perp Z|S$ and $Y \notin S$). Note that the separator S has been obtained during finding $PCD(X)$ if $Z \notin PCD(X)$ or during finding $PCD(Z)$ if $X \notin PCD(Z)$. After finding a new v-structure or adding a new undirected edge, we need to orient again undirected edges within V using Meek's rules.

At Step 4, we add X to $Layer(depth)$ because X is adjacent to a node in $Layer(depth-1)$ and not in the previous layers. Thus $Layer(depth)$ can correctly be formed if $Layer(depth-1)$ was correctly formed. Nodes in $PCD(X) \setminus TotalLay$ are added to the list $canU(depth+1)$ as candidate nodes in the next layer. Thus we can make sure all nodes which have length $depth+1$ from T in $canU(depth+1)$ if $Layer(depth)$ are

correct. Inductive, we showed the correctness of $Layer(i)$ for all i since at the initiation step, $Layer(0) = T$ is correctly set.

At step 5, we obtain the final $Layer(depth)$, add it to $TotalLay$ and add 1 to $depth$ after we treated all nodes in $canU(depth)$.

Finally, we stop Part 1 if (1) all nodes having a path to T have a distance shorter than k or (2) the first $k-1$ Layers have been obtained.

Next we show the correctness of Part 2. In Part 2 of the algorithm, we sequentially search nodes outside $TotalLay$ along an undirected path starting from a node in $Layer(k-1)$ through finding PCDs of the terminal node of the path until a directed edge is found. By Part 2, we obtain a network G which covers the local network $G_k(T)$ we want to find. The network G has mixed types of directed and undirected edges and has directed edges as its boundary. Define A as a set which contains all undirected edges and the first $k-1$ layers in the local network finally obtained by the algorithm, that is, $A = \{u \in V : u$ has an undirected path starting from a node in $Layer(k-1)\} \cup TotalLay$. Then any edge (u,v) which connects a node $u \in A$ and a node $v \notin A$ must be a directed edge otherwise v should be contained in A. Define B as a set of nodes which surrounds A, that is, $B = \{v \in PCD(u) \setminus A : u \in A\}$. Define E as a set of edges each of which has at least one node in A, that is, $E = \{(u,v) : u \in A,\ v \in A \cup B\}$. We can have that all undirected edges within E cannot be oriented even if the global network is obtained. It is because any undirected (u,v) in E must have both of its two nodes u and v contained in A and all undirected edges in E must be surrounded by directed edges. Thus, if these undirected edges cannot be oriented by applying Meek's rules to E, then they cannot still be oriented by finding more edges outside E. ■

JMLR: Workshop and Conference Proceedings 6:203–214, 2010 NIPS 2008 Workshop on Causality

Fast Committee-Based Structure Learning

Ernest Mwebaze EMWEBAZE@CIT.MAK.AC.UG
Faculty of Computing & I.T.
Makerere University
Kampala, Uganda

John A. Quinn JQUINN@CIT.MAK.AC.UG
Faculty of Computing & I.T.
Makerere University
Kampala, Uganda

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Current methods for causal structure learning tend to be computationally intensive or intractable for large datasets. Some recent approaches have speeded up the process by first making hard decisions about the set of parents and children for each variable, in order to break large-scale problems into sets of tractable local neighbourhoods. We use this principle in order to apply a structure learning committee for orientating edges between variables. We find that a combination of weak structure learners can be effective in recovering causal dependencies. Though such a formulation would be intractable for large problems at the global level, we show that it can run quickly when processing local neighbourhoods in turn. Experimental results show that this localized, committee-based approach has advantages over standard causal discovery algorithms both in terms of speed and accuracy.

Keywords: Bayesian Network, feature ranking, relevance learning, committee method

1. Introduction

Current methods for causal structure learning tend to be computationally intensive or intractable for large datasets. Most approaches towards causal structure learning can be categorized into two classes: constraint-based approaches that use independence tests and score-based techniques that search for Bayesian networks. The former are slow because independence has to be tested between variables under many different conditioning sets. The latter are slow because of the possible number of Bayesian networks; a naïve scoring with just 10 variables would have to consider around 10^{18} configurations (Robinson, 1977).

Some recent approaches have speeded up the process by finding the network skeleton first and then doing local neighborhood learning to orient the skeleton edges, such as MMHC (Tsamardinos et al., 2006). Building the skeleton of a network is an easier task than orientating the edges as we only look at associations between variables and not the causal relationships between them.

In this paper we propose a method for fast structure learning based on finding the set of parents and children for each variable, and then applying a committee of structure learners to make a joint decision about edge orientation. Some of the structure learning methods we use would be intractible when applied globally to a dataset with many variables, but can run rapidly at neighbourhood level. When the structure learners are based on different principles (e.g. a mixture of constraint-based and score-based) it is significant when they agree with each other, and in particular we find that this strategy gives good worst-case accuracy.

The contributions of this paper are:

- We generalise previous work on restricting the search space to speed up structure learning;
- We present a novel local structure learning algorithm, EPC, specifically intended for analysing a target variable and its immediate neighbourhood;
- We show how different structure learners can be combined in a committee to give results with better consistency.

The rest of the paper is organized as follows. In Section 2 we discuss the initialization step for finding the skeleton of a network of variables. Section 3 discusses the causal discovery committee. We present experimental evidence in Section 4, and summarise our findings in Section 5.

2. Skeleton Discovery

The overall aim of skeleton discovery is to consider each variable in a dataset and find the set of directly neighbouring variables. To find the neighbourhood of one variable, we begin by considering all variables as potential neighbours and then filtering down this set in two phases. We first employ Relevance Learning Vector Quantization (RLVQ), a fast prototype-based classification method, to do an initial feature selection for each variable. The variables found to have low relevance during this stage are removed from the estimated set of neighbours. We then apply the HITON algorithm on the resulting variables to narrow down this set.

LVQ and RLVQ are prototype-based classification methods applied in supervised learning. They employ a distance measure (typically Manhattan distance or quadratic Euclidean distance) that quantifies the similarity of a given feature vector with a prototype (representative) of any particular class. The distance measure (Manhattan distance) for two arbitrary vectors $x, y \in \mathbb{R}^N$ can be defined as:

$$d(x,y) = \sum_{j=1}^{N} \left| x_j - y_j \right| . \tag{1}$$

Because the features have varied meanings and magnitudes in the data, quantifying their similarity by a uniform distance measure tends to be problematic. These differences

are accounted by relevance learning schemes like RLVQ that employ adaptive scaling factors that scale the features based on their relevance for classification. This takes the form

$$d^i_\lambda\left(w^i,\xi\right)=\sum_{j=1}^{N}\lambda^i_j\left|w^i_j-\xi_j\right| \tag{2}$$

where w denotes a prototype or representative vector of a particular class, ξ denotes a data vector and the adaptive relevance factors λ^i_j are restricted to non-negative values and obey the normalization $\sum_{j=1}^{N}\lambda^i_j=1$. The special case $\lambda^i_j=1/N$ for all $j=1,\ldots N$ is analogous to the original LVQ measure. The RLVQ adapts the prototypes and the relevance factors for each training run through the data until the error rate is at a minimum. Further details on LVQ and RLVQ can be obtained from Bojer et al. (2001).

These methods have been used is several applications because on top of being intuitively easy to understand they are easy to implement and their complexity is controlled by the user. For our purposes however we draw from the fact that they are fast and have been shown to give high accuracy in identifying relevant features for classification (Biehl et al., 2007).

HITON is a standard algorithm for feature selection that, assuming the joint data distribution is faithful to a Bayesian Network, carries out statistical tests on the data to determine the Markov boundary and the Markov blanket of a target variable. HITON has been proven to accrue two main advantages over other feature selection algorithms: 1) it reduces the number of variables in the prediction models roughly by three orders of magnitude relative to the original variable set while improving or maintaining accuracy, and 2) it outperforms the baseline algorithms by selecting smaller variable sets than the baselines (Aliferis et al., 2003). Because HITON takes several hours to run for datasets with hundreds or thousands of variables, the RLVQ preprocessing step is useful to speed the process of obtaining Markov boundaries for each variable.

To summarise, for each variable in the local neighbourhood a set of features relevant for its classification are obtained using RLVQ (phase 1). For each of these sets of relevant features, the HITON algorithm is used to further narrow down the set of parents and children of the variable under consideration. Given this skeleton of undirected edges between variables, a committee of structure learning methods is then used to vote on the causes (parents) and effects (children), as described in the next section.

3. Causal Discovery Committees

Once a skeleton of the network is found we apply a structure learning committee for orientating edges between variables. We find that a combination of weak structure learners can be effective in recovering causal dependencies. Though such a formulation would be intractable for large problems at the global level, we show that it can run quickly when processing local neighbourhoods in turn.

The structure learning committee method takes the neighbourhood of each variable and applies different algorithms to determine whether each neighbour of that variable is a cause or an effect. If the majority of the algorithms determine that a given neighbour

is a cause, then we classify it as a cause. Effects are classified in the same way. We do not apply any conflict resolution at the moment; our method might return bi-directional causes. Algorithm 1 shows the committee voting method.

Algorithm 1: Localized causal discovery committee.

input: $\mathbf{c}_1 \ldots \mathbf{c}_N$, data vectors for variables $C_1, \ldots, C_N$
  $PC(i)$, set of parents and children for each variable C_i.
for each variable C_i, $i = 1 \ldots N$ **do**
  for each algorithm $Algo_j$ **do** /* PC, EPC, GES, MWST, LiNGAM, K2 */
    $causes(C_i, j), effects(C_i, j) \leftarrow Algo_j(C_i, PC(i))$
  $\mathbf{C_i} \leftarrow$ majorityVote($causes(C_i, :)$)
  $\mathbf{E_i} \leftarrow$ majorityVote($effects(C_i, :)$)
return: {**C**,**E**}, causes and effects of variables $C_1, \ldots, C_N$.

In the remainder of this section, we first introduce a novel structure learning algorithm, EPC, and then list the other committee members.

3.1. Expected Partial Correlation (EPC) Method

EPC is a simple local neighborhood structure discovery algorithm. Given the set of parents and children of a target variable, it returns a probability of each neighbourhood variable being either a cause or an effect. It is based on partial correlation as a measure of conditional independence, which is true in certain cases such as binary or linear Gaussian networks (Baba et al., 2004). We denote the Pearson correlation coefficient between A and B as ρ_{AB}, and the partial correlation between A and B conditioned on C as $\rho_{AB \cdot C}$.

The algorithm works by considering different three-variable subsets of the target and neighbourhood. We divide 3-variable acyclic connected models into three interesting classes: the collider or V-structure $(A \rightarrow C \leftarrow B)$; the chain $(A \rightarrow C \rightarrow B, A \leftarrow C \leftarrow B)$; and the fork $(A \leftarrow C \rightarrow B)$, as shown in Figure 1(i-iii). The chain and the fork have the same conditional independency $A \perp\!\!\!\perp B \mid C$, while the V-structure has the unique property $A \perp\!\!\!\perp B$ but $A \not\perp\!\!\!\perp B \mid C$. We only consider variables B which are not directly connected to A, as this would imply a cycle, which we cannot make any inferences about.

Given a particular sample size and type of distribution, we can work out what distribution of empirical correlation and partial correlation we expect from each different class. We show histograms of correlation and partial correlation in simulated networks in Figure 2. 10,000 binary models in each class (collider, chain, fork) were randomly created, with conditional probability tables sampled from the uniform distribution. We can see that the V-structure is the only case where conditioning on the variable C increases the scale of the correlation between A and B, from the distribution of $|\rho_{AB \cdot C}| - |\rho_{AB}|$ in Figure 2 (right).

The histogram in Figure 2(right) therefore gives us a probability distribution on the likelihood $P(\delta_{ABC} | class(A,B,C))$, where $\delta_{ABC} = |\rho_{AB \cdot C}| - |\rho_{AB}|$ and $class(A,B,C)$

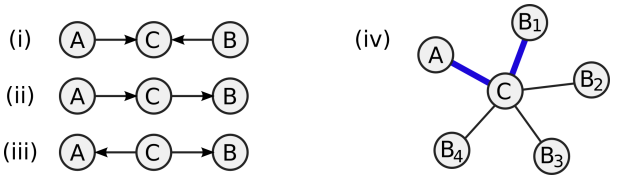


Figure 1: Possible 3-variable structures: (i) collider, (ii) chain, (iii) fork. Panel (iv) shows an example local neighbourhood for a variable C. The EPC algorithm orientates the edge AC by looking at the supporting evidence from each of the B_i's.

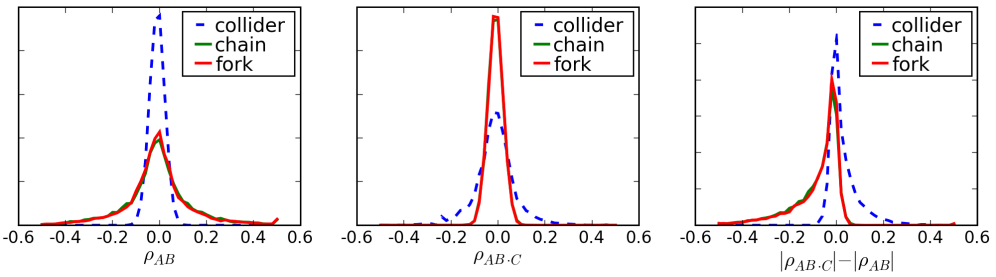


Figure 2: Histograms of correlation and partial correlation from 10,000 simulated 3-variable binary networks of each class, with 1000 samples drawn from each. Chains and forks have indistinguishable correlation distributions.

can be "collider" or "chain/fork" as in Figure 1(i-iii). By specifying priors on $P(class(A,B,C))$ we can then calculate the probability that A is a cause of C, using the assumption that in the "collider" class A is always a cause of C, whereas in the "chain/fork" class, there are 3 possible orientations, in only one of which A is a cause of C. While trying to calculate whether A is a cause or an effect, we incorporate evidence from each of the B_i's in the neighbourhood and obtain $P\left(Cause(A,C)|\delta_{AB_1C},\delta_{AB_2C},\ldots,\delta_{AB_{N-1}C}\right)$ for a neighbourhood of size N and threshold at 0.5 to determine whether A is a cause or an effect. Algorithm 2 shows the steps of this calculation.

The algorithm is limited to certain distributions, such as binary or Gaussian networks, where partial correlation is a measure of conditional independence. The method would fail in non-linear relationships between variables such as an XOR function. We also do not have an analytical form for the likelihood function; we currently have to estimate the distribution through simulations.

However the advantages of the algorithm are as follows. First, it is cheap to run: $O(N^2)$ in the neighbourhood size and $O(M)$ in the sample size. Second, it provides probabilities rather than categorical outputs – most methods based on CI constraints simply accept or reject a causal hypothesis. Third, we have the ability to incorporate prior beliefs about the orientations of edges. Fourth, it is useful as a committee member, as it gives high confidence when there is a V-structure and low confidence otherwise.

The performance of the EPC algorithm in recovering true causes and effects is evaluated in section 4.

Algorithm 2: EPC Algorithm to distinguish between local causes and effects.

input: $\mathbf{c},\mathbf{b_1},\ldots,\mathbf{b_N}$, data vectors for target variable C and set of parents/children $B_1,\ldots,B_N$.
  $P(Cause(B_i,C))$ for all i, priors for each B_i being a cause of C.
for each variable B_i **do**
  for each variable $B_{j\neq i}$ (B_i not a neighbour of B_j) **do**
    $\delta_{ij} \leftarrow |\rho_{B_iB_j\cdot C}| - |\rho_{B_iB_j}|$
    Compute likelihoods $L(\delta_{ij}|class(B_i,B_j,C))$ where $class(B_i,B_j,C) \in$ {"collider', "chain/fork"}
  $causeodds(i) \leftarrow P(Cause(B_i,C))\prod_{i\neq j}(L(\delta_{ij}|\text{collider}) + L(\delta_{ij}|\text{chain/fork}))$
  $effectodds(i) \leftarrow (1-P(Cause(B_i,C)))\prod_{i\neq j} 2L(\delta_{ij}|\text{chain/fork})$
  $P(Cause(B_i,C)|\mathbf{c},\mathbf{b_1},\ldots,\mathbf{b_N}) \leftarrow \frac{causeodds(i)}{causeodds(i)+effectodds(i)}$
return: $P(Cause(B_i,C)|\mathbf{c},\mathbf{b_1},\ldots,\mathbf{b_N})$ for each i, posterior probabilities that each B_i is a cause of C.

3.2. Other Committee Members

Standard algorithms were used in conjunction with EPC to form the structure learning committee. The strength of the committee method is derived from applying each of these methods to the same skeleton obtained from Section 2 for each dataset. These

methods were selected to include methods based on different principles. These methods were as follows[1].

PC : This is a common benchmark constraint-based causal discovery algorithm, introduced by Spirtes et al. (1993). A confidence level of 0.05 was used with this method.

MWST : The MWST algorithm as introduced by Chow and Liu (1968) is based on the maximum weight spanning tree. It essentially associates a weight with each edge obtained according to some similarity criterion (mutual information between variables or BDeu score) and then builds the maximum spanning tree of the obtained graph. For our experiments we used the mutual information between variables as a measure of (conditional) dependence.

GES : The greedy search (GS) algorithm is an implementation of a standard optimization heuristic. Greedy Equivalent Search is an extension of the GS algorithm that optimizes searching the DAG space by searching in the Markov equivalent space. This method initially starts with an empty graph, adds arcs until the score cannot be improved then tries to suppress some irrelevant arcs (Munteanu and Bendou, 2002). For our experiments we used the Bayesian Information Criterion (BIC) as our scoring function with an instantiation cache of 300.

K2 : The K2 algorithm (Cooper and Herskovits, 1992) is a probabilistic algorithm that maximizes structure probability given the data. It defines the Bayesian measure(BIC/BDeu) which is a quality measure of the network given the data. We use it in the committee to vote on whether a feature is an effect of the target variable only and not a cause because it is easier to specify a node order for the former. For our experiments we used the Bayesian Score (BIC) as our scoring function.

LiNGAM : LiNGAM (Shimizu et al., 2006) is a more specific technique that attempts to discover the causal structure in linear non-Gaussian acyclic models. We include it in the committee because it provides a relatively different technique form the rest of the committee members and hence can account for certain distributions on which the other members may produce poor results. For our experiments default settings were used, as provided in the author's implementation.

4. Experiments

We test our methods on several standard datasets with known generating structures. The causal structures found by our methods are evaluated using the same edit distance score based evaluation method used to evaluate the NIPS 2008 causality challenge entries. In this method, a confusion matrix C_{ij} that specifies the number of relatives confused for another type of relative is computed. It evaluates the 14 types of relatives in a depth-3 network. A cost matrix A_{ij} is also computed to account for the edit distance between the relatives. The edit distance specifies the number of substitutions, insertions, or deletions to go from one string to another. A score for a particular structure is then computed as

1. Implementations from four packages were used: BNT (http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html), BNT-SLT (http://bnt.insa-rouen.fr/), LiNGAM (http://www.cs.helsinki.fi/group/neuroinf/lingam/) and Causal Explorer (http://discover.mc.vanderbilt.edu/discover/public/causal_explorer/)

$S = \sum_{ij} A_{ij} C_{ij}$. The results are summarised in Table 1 for the different datasets. In this table, we first show the performance of different standard methods and the EPC algorithm when applied in a local neighbourhood setting. Each of the algorithms in Table 1 are applied on the same skeleton (for each dataset) obtained using feature selection/reduction techniques discussed in Section 2. We then show the performance of all methods when combined in committee, using a majority voting scheme. As a benchmark we then show performance of the PC algorithm when applied as standard to the whole datasets (no localization).

Table 2 (a-f) shows confusion matrices for the committee output of three datasets. The confusion matrices give an idea of the recall and precision rates of the method. An ideal confusion matrix would be a diagonal matrix indicating the true positives and true negatives. The figures that are not on the diagonal represent the numbers of false positives (spurious causes, bottom left) and the numbers of false negatives (spurious independencies, top right).The performance of our committee method in the LOCANET challenge is given in Table 2 (g), compared to other participants. Execution time for HAILFINDER using the standard PC algorithm for example was 4452.4 seconds, while for the committee this time was 190 seconds for obtaining the skeleton and 13.3 seconds for obtaining the local graph from the committee.

Table 3 shows other performance metrics used for datasets where we know the ground truth. We calculate precision (ratio of true causes found to total causes found), recall (proportion of true causes found to actual number of causes), and Fmeasure $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

Method	LUCAS (2000)	LUCAP (2000)	ALARM (5000)	ASIA (2000)	INSURANCE (2000)	HAILFINDER (20000)
PC	1.91	2.14	2.43	2.08	2.81	1.79
EPC	0.91	1.81	0.57	2.94	2.2	2.2
GES	1.86	2.14	1.5	2.96	3.38	2.58
MWST	2.86	2.46	2.21	1.7	2.7	1.68
K2	2.18	1.95	2.1	1.78	2.15	1.79
LiNGAM	1.73	3.08	1.93	1.38	2.81	1.43
Committee (M)	1.65	1.9	1.07	2.86	2.46	2.39
PC‡	2.91	3.38	2.72	3.29	2.81	2.73

Table 1: Evaluation of edit distances for various algorithms with known networks (sample size in brackets). Committee (M) denotes the committee decision with *Majority Voting*. PC‡ represents results obtained on running the standard PC on the whole dataset.

	→	x
→	66	123
x	89	20458

(a) Lucap-EPC

	→	x
→	78	236
x	178	20244

(b) Lucap-Committee

	→	x
→	10	22
x	74	1263

(c) Alarm-EPC

	→	x
→	18	18
x	131	1202

(d) Alarm-Committee

	→	x
→	2	27
x	28	3079

(e) LiNGAM

	→	x
→	7	27
x	45	3057

(f) Committee

Dataset	Score	Others(range)
CINA	2.32	1.70 - 3.31
REGED	0.22/0.439	0.27 - 0.50
SIDO	3.46	3.31 - 3.48

(g) Results on challenge datasets

Table 2: (a-g) Confusion matrices showing the algorithms with the winning edit distance and the corresponding majority committee decisions in terms of true causes found (top left), spurious causes (bottom left), true independencies (bottom right) and spurious independencies (top right) for Lucap (a-b), Alarm (c-d), and Hailfinder (e-f). (g) LOCANET challenge results.

Dataset/Method	PRECISION	RECALL	FMEASURE
Lucap-EPC	0.43	0.54	0.47
Lucap-Committee	0.30	0.33	0.32
Alarm-EPC	0.12	0.45	0.19
Alarm-Committee	0.12	1.00	0.22
Hailfinder-LiNGAM	0.07	0.07	0.07
Hailfinder-Committee	0.13	0.26	0.18

Table 3: Evaluation of Precision, Recall and Fmeasure scores for three of the datasets contrasting different individual algorithms with the committee

5. Discussion

From the challenge results we can see that our method gives performance comparable to other entries , while employing a method designed to give fast inference time. We provide two scores for the REGED dataset; the first one represents the initial submission where our method failed to find causes but due to the bias in the scoring statistic, this did quite well, the second represents a score obtained by the committee with the voting threshold decreased. For the benchmark datasets we find that the quality of the committee decisions is close to the best committee member in each case. Results obtained for applying PC to whole datasets without localization generally indicate a lower accuracy rate than either PC with localization or the causal discovery committee. In principle to increase precision (at the expense of recall) we can increase the voting

threshold upwards towards the unanimous voting level. Conversely it is also possible to increase the recall rate by altering the voting threshold in the opposite fashion. For applications where a causal relationship needs to be established with high precision, a unanimous voting scheme may be used though we have not so far analysed the accuracy of this approach.

The confusion matrices in Table 2 indicate that the committee generally obtains more true positives (higher recall - Table 3) than the corresponding committee member with the best average edit distance score. However the committee also generally obtains more false positives (lower precision) which accounts for the committee score not being as good as that of the best algorithm in each case.

Currently our method does not explicitly handle conflict of orientation so it is possible to have a situation where we find that $P \rightarrow Q$ and also $Q \leftarrow P$. The output is therefore not a DAG. We conjecture that finding bi-directional causes $P \leftrightarrow Q$ may indicate the presence of a hidden variable which influences both P and Q.

We have used our local committee framework with particular structure learning algorithms, but anticipate that other algorithms can be used in future work. Future research will also look at weighting the committee members based on derived properties of the dataset.

Acknowledgments

We would like to thank Michael Biehl for helpful discussions on relevance learning. The work was supported in part by the Dutch NUFFIC NPT project.

References

C. F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection. In *Proc. of the 2003 American Medical Informatics Association (AMIA) Annual Symposium*, pages 21–25, 2003.

K. Baba, R. Shibata, and M. Sibuya. Partial Correlation and Conditional Correlation as Measures of Conditional Independence. *Australian and New Zealand Journal of Statistics*, 46(4):657–664, 2004.

M. Biehl, R. Breitling, and Y. Li. Analysis of Tiling Microarray Data by Learning Vector Quantization and Relevance Learning. In *Proc. of the 2007 IDEAL*, 2007.

T. Bojer, B. Hammer, D. Schunk, and Tluk von Toschanowitz. Relevance determination in learning vector quantization. In Verleysen M, editor, *European Symposium on Artificial Neural Networks*, pages 271–276. d-facto publications, 2001.

C. K. Chow and C. N. Liu. Approximating discrete probability distribution with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

G. F. Cooper and E. H. Herskovits. The induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

P. Munteanu and M. Bendou. The EQ framework for learning equivalence classes of Bayesian networks. In *First IEEE International Conference on Data Mining (IEEE ICDM)*, San Jose, 2002.

R. W. Robinson. Counting unlabeled acyclic digraphs. In C.H.C. Little, editor, *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*. Springer, Berlin, 1977.

S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*, volume 81. Springer Verlag, Berlin, 1993.

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

Appendix A. Pot-luck challenge: FACT SHEET.

(for a task solved)

Title: LOCANET
Ernest Mwebaze & John A. Quinn
Faculty of Computing & IT
Makerere University
(emwebaze, jquinn)@cit.mak.ac.ug
Task(s) solved: Local Structure Discovery

Method:
Our method uses a relevance learning algorithm (RLVQ) and the HITON algorithm to reduce the feature set to parents and children of each feature. A novel partial correlation algorithm in a committee of standard structure learning algorithms then votes on which of the features are parents and which are children for each Markov boundary obtained from the feature reduction step. Because we employ feature reduction initially, the method is fast and because a committee votes on the edge directions, the method yields high accuracy.

- Preprocessing : None
- Causal discovery : Use of novel probabilistic partial correlation algorithm in committee of standard structure learning algorithms ; PC, GES, MWST, K2 and LiNGAM.
- Feature selection : Use of Relevance Learning Vector Quantization and HITON algorithms
- Classification : None
- Model selection/hyperparameter selection : Majority vote of committee of algorithms

Results:

Dataset/Task	Score 1
CINA	2.32
REGED	0.22/0.439
SIDO	3.46

Table 4: Result table.

Advantages:

- Quantitative advantages : Our method employs feature selection techniques to obtain relevant features for classification from which we can obtain relevant features for causality. This reduces the processing time as indicated in our paper.

- Qualitative advantages : We employ a novel method, Expected Partial Correlation (EPC), that offers comparable results when compared with other standard algorithms on known datasets as illustrated in Table 1 in the paper.

Method Implementation:
We implemented our method in matlab on an Intel Duo CPU T7100 laptop computer with 1024 MB or RAM.
The standard algorithms were implemented using standard packages from different individuals/organizations these included :-

- Bayesian Network Toolkit (http://www.ai.mit.edu/m̃urphyk/Software/BNT/bnt.html),

- Structure Learning Toolkit (http://bnt.insa-rouen.fr/),

- LiNGAM package (http://www.cs.helsinki.fi/group/neuroinf/lingam/)

- Causal Explorer (http://discover.mc.vanderbilt.edu/discover/public/causal_explorer/).

The whole application is built up into two modules, one that does the feature reduction and the other that does the structure learning based on the novel algorithm EPC and a host of standard algorithms including PC, MWST, GES, LiNGAM and K2.

JMLR: Workshop and Conference Proceedings 6:215–224, 2010 NIPS 2008 workshop on causality

SIGNET: Boolean Rule Determination for Abscisic Acid Signaling

Jerry Jenkins JWJ@CFDRC.COM
Systems Biology and Bioinformatics Group
CFD Research Corporation
601 Genome Way
Huntsville, AL 35806, USA

Editors: Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf

Abstract

This paper describes the SIGNET dataset generated for the Causality Challenge. Cellular signaling pathways are most elusive types of networks to access experimentally due to the lack of methods for determining the state of a signaling network in an intact living cell. Boolean network models are currently being used for the modeling of signaling networks due to their compact formulation and ability to adequately represent network dynamics without the need for chemical kinetics. The problem posed in the SIGNET challenge is to determine the set of Boolean rules that describe the interactions of nodes within a plant signaling network, given a set of 300 Boolean pseudodynamic simulations of the true rules. The two solution methods that were presented revealed that the problem can be solved to greater than 99% accuracy.

Keywords: Boolean pseudodynamics, plant signaling network

1. Introduction

Development of accurate models to predict cellular response to stimulus must begin with a proper characterization of the interaction between the various cellular processes. It is estimated that each individual gene or protein, on average, interacts with four to eight other genes and is involved in ten biological functions (Arnone and Davidson, 1997; Miklos and Rubin, 1996). A seamless interaction between all cellular processes is essential for a living cell to thrive.

Kinetic models have been successfully applied to the analysis of a wide variety of biological systems, recent examples include neuronal signaling and the role of synaptic plasticity (Ajay and Bhalla, 2006), phase sensitivity in circadian rhythms (Gunawan and F. J. Doyle, 2007), and prediction of IL-2 response from T-cell receptor activation (Kemp et al., 2007). By providing a global view of the underlying system, a kinetic model can be used to interpret new experimental data in the proper biological context (von Dassow et al., 2000), provide mechanistic explanations for counter-intuitive observations (Fallon and Lauffenburger, 2000), and facilitate the formulation of experimentally testable hypotheses (Abouhamad et al., 1998; Endy et al., 2000). Unfortunately, accurate descriptions of underlying chemical kinetics are difficult to determine *in vivo*,

with reliable kinetic coefficient estimation being a non-trivial and frequently impossible challenge due to a lack of identifiability (Yao et al., 2003).

Experimental observations of cellular function indicate that the input-output behavior of signaling networks has a sigmoidal time dependence, and often can be adequately explained using the Heaviside, or step function (Thomas, 1973). This observation suggests that a two state Boolean model could be employed to represent signaling network nodes, with nodal values being determined using an associated logical rule, representing network edges. Recent research has focused on applying rule-based Boolean models to the challenging problem of predicting biological network dynamics (Li et al., 2006). In a Boolean network model, the nodes of the network represent biological entities and the edges represent the interactions between them. The nodes can have a value of 0 or 1, representing an inactive or an active state, respectively. The network dynamics are determined by Boolean rules for each node, that determine the state of the node at the next time-step based on the state of the upstream nodes, and the nodal update strategy. Rule-based Boolean network models have been successfully used to aid in explaining experimentally observed robustness of cellular networks (Albert and Othmer, 2003; Kauffman et al., 2003; Thomas, 1973), and to determine the effects of an alteration in the network components and individual reaction rates (Chaves et al., 2005).

At CFDRC, we have developed an augmented Boolean pseudo-dynamics approach to identify and quantitatively rank the importance of a node using a Boolean description of a cellular interaction network. The approach, known as the Boolean Network Dynamics and Target Identification (BNDTI), combines network topology and dynamic state information to determine the relative importance of a particular node with respect to the overall response of the network (Soni et al., 2008). In order to perform a demonstration of the utility of the newly developed approach, the guard cell signaling network in plant cells was selected (Li et al., 2006). This signaling network has been painstakingly translated into a Boolean network, and centers around abscisic acid (ABA) signal transduction, which for many decades has been known to play a role in ABA induced stomatal closure, regulating the plant water balance and imparting drought resistance. Two major secondary messengers involved in the closure of the stomata via ABA signal transduction are cytosolic calcium and the cytosolic pH. These two messengers are in turn regulated by a variety of other enzymes, secondary messengers, small molecules, and membrane channels. Figure 1 is a rendering of the interaction network, illustrating the complex regulatory interactions between species.

In the remainder of the paper, Section 2 provides a statement of the particular problem posed in the SIGNET challenge, along with some comments on the importance of the problem and how researchers addressed the problem. Section 3 includes a summary of the challenge results along with relevant comments.

2. SIGNET Challenge Problem Description

The problem posed in the SIGNET challenge is to determine the set of Boolean rules that describe the interactions of nodes within a plant signaling network, given a set of

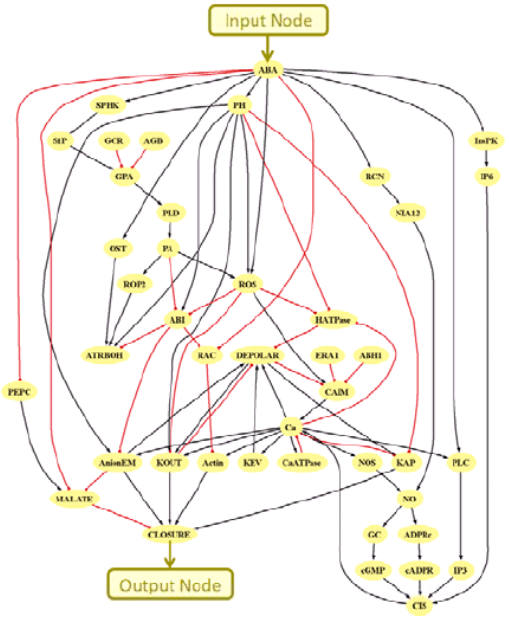


Figure 1: A Schematic of the Guard Cell Signaling Network. Inhibition reactions are shown with red edges and inverse arrowheads, whereas activation interactions are shown as black edges.

300 Boolean pseudodynamic simulations of the true rules. The relevance of this problem arises from the trend in the biological sciences toward the increased availability of large datasets generated using high-throughput, high-content experimental technologies (such as gene expression microarrays). Experimental methods are currently able to probe the interactions of many thousands of cellular components simultaneously. However, cellular signaling pathways are still one of the most challenging and illusive types of networks to access experimentally due to the lack of methods for determining the state of a signaling network in an intact living cell. The SIGNET problem anticipates that experimental techniques for signaling network measurements will continue to progress, and assumes the availability of a large set of high throughput data that will be used to determine the set of Boolean rules describing the signaling network.

The expense and time in signaling measurements necessitate that the majority of signaling network models in the published literature are manually constructed using the relatively sparse literature data. The typical methodology includes a thorough assimilation of all relevant literature, followed by the construction of a table that formalizes the nodes (components) and edges (interactions) of the network. Using the table, a necessary and sufficient network capable of predicting the relevant behavior is generated. Often times the network is manually generated, introducing human bias and utilizing a significant amount of time and resources.

Automated methods for Boolean network inference have focused a significant amount of attention to the problem of identifying gene networks. The REVEAL (REVerse Engineering ALgorithm) was one of the first algorithms designed for this purpose (Liang et al., 1997), which combines information-theory tools with an exhaustive search to generate a network that is consistent with the data. An alternative algorithm is the BOOL-1 algorithm (Akutsu et al., 1999), which consists of examining all possible k-tuples of inputs and testing all Boolean function for each k-tuple until a consistent network is generated. The difficulty has motivated the utilization of heuristic approaches. An example of a heuristic approach is ID3 (Quinlan, 1986), which is a well known algorithm in Machine Learning. ID3 is based on the incremental construction of the input set for each variable using a greedy search. The approach presented in next section is based on the synergistic utilization of evolutionary algorithms and existing heuristics such as ID3. More recent approaches include the p-ary transitive reduction (TR_p) (Albert et al., 2007), have been demonstrated that produce an optimal network given the constraints of minimal false positive inferences. Unfortunately, due to the lack of the necessary quantities of experimental data little effort has been expended for the automated identification of cellular signaling networks. Therefore, the overall goal of the SIGNET challenge was, therefore, to increase awareness of this problem area and stimulate interest in novel methods of solution.

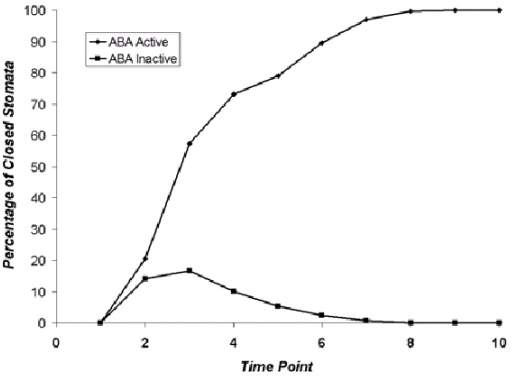


Figure 2: Effect of the presence and absence of abscisic acid on the percentage of closed stomata in the plant guard cell signaling model.

The SIGNET dataset was generated using the procedure that follows. Nodes, edges, and Boolean rules were obtained from the work of Li et al. (2006). The network consists of a total of 43 nodes. Five nodes are input nodes (nodes that have only out-degree), and are the initiators of network action. The state of the input nodes ABA, GCR, ABH1, ERA1 was fixed at a value of '1' throughout the simulation. The state ('0' or '1') of the remaining 38 variable nodes was selected at random at the start of each simulation. 300 Boolean pseudodynamic (BPD) simulations were then generated using the asynchronous update strategy. The choice of BPD update scheme depends on the distribution of kinetic timescales within the network. The two most popular update schemes are synchronous and asynchronous. The synchronous method updates nodes in a fixed order at each time step, the order being determined at the start of the simulation. Synchronous updating assumes that the physical interactions within the network all occur at approximately the same time scale. Though synchronous updating is an efficient simulation method, it is rarely used for realistic systems due to the limiting assumption of similar time scales. In contrast, the asynchronous update method randomly determines the update order at each time step, which is equivalent to the assumption that the kinetic time scales within the network have a Gaussian distribution. Asynchronous updating is known to mimic realistic events in complex networks, and has been shown to effectively capture rare events (Chaves et al., 2005; Li et al., 2006). Figure 2 is a plot of the response of the CLOSURE node averaged over the 300 randomly selected initial conditions with ABA=1 and ABA=0.

The overall objective of the SIGNET challenge was to determine the set of Boolean rules that describe the interactions of the nodes within this plant signaling network. The dataset includes 300 separate Boolean pseudodynamic simulations of the true rules, using an asynchronous update scheme. The results for 300 separate simulations are included in the dataset. Each simulation consists of a matrix of 0's and 1's, with 21 rows and 43 columns. The first row is the randomly generated initial condition for the particular simulation, with the next 20 rows being the output from the Boolean pseudodynamics simulation. Each of the 43 columns represents the transient response of a particular node. The nodal names are identified at the top of the data file.

3. Summary of SIGNET Challenge Results

Solutions to the SIGNET challenge were submitted by Mehreen Saeed of the Department of Computer Science at the National University of Computer and Emerging Sci-

ences (Lahore Campus, Pakistan) (Saeed, 2009), and Cheng Zheng of the School of Mathematical Sciences at Peking University (Beijing, China) (Zheng and Geng, 2009).

3.1. Performance Assessment

Solution methodology performance was assessed using the original SIGNET case, and for a second case generated by Prof. Isabelle Guyon. The organizers of the challenge provided a Matlab code for the evaluation of the algorithm performance. The evaluation code consisted of the generation of a truth table for each true rule and computing a prediction error by comparing output values of the extracted rule to the output from the true rule. The error is computed by averaging over all rules. In addition to calculating the overall error rate of prediction, Prof. Saeed also calculated the training set error of the inferred rules. This was done by applying each rule to the individual Boolean vectors of the simulation data and predicting the output value. The output value was compared with the actual value to obtain the overall training accuracy rate.

3.2. Bernoulli Mixture Model (BMM)

The paper submitted by Prof. Saeed develops a Bernoulli distribution-based probabilistic model for the data, and combines this with the mixture densities to identify the Boolean rules from the SIGNET dataset. Parameters for the underlying Bernoulli distribution are estimated from the raw data using the expectation maximization (EM) algorithm. This methodology is stated to be ideal for estimating the probability distribution of non uni-modal data. Prof. Saeed has considerable experience applying this same methodology to the problem of dimensionality reduction and feature selection.

Optimal values for the number of mixtures as well as the probability thresholding value are given. The number of mixtures determines the underlying Bernoulli distribution, and the complexity of parameter extraction for estimating priors. Probability thresholding values are used to identify high data density areas on the corners of a hypercube. Each corner represents a conjunct of Boolean variables and together the set, of all the corners, forms a disjunction of rules, yielding a disjunctive normal form of a Boolean rule.

The results presented indicate that three mixtures produced the optimal training and evaluation accuracy of 94.55% and 82.98%, respectively, for the original SIGNET set. The dataset generated by Prof. Isabelle Guyon yielded a training accuracy of 95.88% and an evaluation accuracy of 87.61% for the three mixture model. A thresholding value of 0.70 produced good results for the case of a single mixture, but poor results for 2 and 3 mixtures. A thresholding value of 0.80 produced optimal results for 3 mixtures, and showed good accuracy for 1 and 2 mixtures. Thresholding values of 0.90 produced low accuracy results due to the number of results being ignored.

3.3. Minimum Explanatory Set and Maximum Likelihood (MESML)

The paper submitted by C. Zheng uses a method for finding the minimum explanatory set for a particular node (Ideker et al., 2000), and then determines a Boolean func-

tion that generates maximal log likelihood for a particular node. The methodology is specifically modified for the reconstruction of asynchronous Boolean networks, where the nodal update order is selected at random.

Accuracy of the method was assessed in the same manner as with Prof. Saeed's solution. The accuracy of the proposed method was evaluated on the original SIGNET dataset and two other datasets generated by C. Zheng. Accuracy rates as a function of the number of assumed parent nodes are given for evaluation of the method. Interestingly, C. Zheng finds that as the number of parent nodes increases, the accuracy rate also decreases. This result is in agreement with the expectation that the average nodal accuracy should exhibit a maximum around the most probable in-degree, which for this network is 1 (58% of nodes). The averaged accuracy rate for a single parent node is 95%, which is excellent.

3.4. Discussion

The primary strength of the BMM methodology is the straight forward, novel approach of converting a probabilistic model into a rule based model in an intuitive manner. Information concerning the total runtime to expect in practice was not provided in the final manuscript, which would have aided the reader in making an implementation decision. However, the only bottleneck to performance would be the expectation maximization and I would not anticipate that it scales poorly with the number of mixtures.

The MESML method demonstrated by C. Zheng is the most accurate, with an average accuracy of 95%. The major drawback of the methodology is that the computational time scaling is roughly proportional to 10^n (see Table 2 of Zheng and Geng (2009)), where n is the number of parent nodes. This is likely to cause a potential problem for networks that contain a large number of hub nodes, where the most probable in-degree is larger than one.

As experimental techniques become more sophisticated, computational methods will be called upon to provide biologically relevant insight into cellular behavior and interactions. Boolean networks will continue to play an ever increasing role in signaling network modeling due to their simplicity and predictive capability. Based upon the accuracy of the predictions, the results of the SIGNET challenge should provide significant confidence to researchers seeking to unravel the secrets of signaling networks.

Acknowledgements

I would like to thank Constantin Aliferis (NYU Medical Center) for encouraging our submission of the SIGNET dataset. I thank Prof. Saeed and C. Zheng for submitting their solutions. Guidance and encouragement provided by Isabelle Guyon and the organizers of the challenge is also acknowledged. Funding was provided by the U.S. Army MRMC (Program Manager: COL Alan Magill, WRAIR).

References

W. N. Abouhamad, D. Bray, M. Schuster, K. C. Boesch, R. E. Silversmith, and R.B. Bourret. Computer-aided resolution of an experimental paradox in bacterial chemotaxis. *J Bacteriol*, 180:3757–64, 1998.

S. M. Ajay and U. S. Bhalla. Synaptic plasticity in vitro and in silico: insights into an intracellular signaling maze. *Physiology (Bethesda)*, 21:289–96, 2006.

T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, page 29, 1999.

R. Albert and H. Othmer. The topology of the regulatory interactions predicts the expression pattern of the drosophila segment polarity genes. *J. Theor. Biol.*, 223: 1–18, 2003.

R. Albert, B. DasGupta, R. Dondi, S. Kachalo, E. Sontag, A. Zelikovsky, and K. Westbrooks. A novel method for signal transduction network inference from indirect experimental evidence. *J Comput Biol*, 14:927–49, 2007.

M.I. Arnone and E.H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124:1851–64, 1997.

M. Chaves, R. Albert, and E. D. Sontag. Robustness and fragility of boolean models for genetic regulatory networks. *J Theor Biol*, 235:431–49, 2005.

D. Endy, L. You, J. Yin, and I. J. Molineux. Computation, prediction, and experimental tests of fitness for bacteriophage t7 mutants with permuted genomes. In *Proc Natl Acad Sci U S A*, volume 97, pages 5375–80, 2000.

E. M. Fallon and D. A. Lauffenburger. Computational model for effects of ligand/receptor binding properties on interleukin-2 trafficking dynamics and t cell proliferation response. *Biotechnol Prog*, 16:905–16, 2000.

R. Gunawan and 3rd F. J. Doyle. Phase sensitivity analysis of circadian rhythm entrainment. *J Biol Rhythms*, 22:180–94, 2007.

T. Ideker, V. Thorsson, and R. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. In *Pacific Symposium on Biocomputing*, volume 5, pages 302–313, 2000.

S. Kauffman, C. Peterson, B. Samuelson, and C. Troein. Random boolean network models and the yeast transcription network. In *Proc Natl Acad Sci USA*, volume 100, pages 14796–14799, 2003.

M. L. Kemp, L. Wille, C. L. Lewis, L. B. Nicholson, and D. A. Lauffenburger. Quantitative network signal combinations downstream of tcr activation can predict il-2 production response. *J Immunol*, 178:4984–92, 2007.

S. Li, S. Assmann, and R. Albert. Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biol*, 4: e312, 2006.

S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, volume 3, page 29, 1997.

G. L. Miklos and G. M. Rubin. The role of the genome project in determining gene function: insights from model organisms. *Cell*, 86:521–9, 1996.

J. Quinlan. Induction of decision trees. *Machine Learning*, 1:106, 1986.

M. Saeed. The use of bernoulli mixture models for identifying corners of a hypercube and extracting boolean rules from data. *JMLR: Workshop and Conference Proceedings*, 6, 2009.

A. S. Soni, J. W. Jenkins, and S. S. Sundaram. Determination of critical network interactions: an augmented boolean pseudo-dynamics approach. *IET Syst Biol*, 2:55–63, 2008.

R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42:563–85, 1973.

G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406:188–92, 2000.

K. Yao, B. Shaw, B. Kou, K. McAuley, and D. Bacon. Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design. *Polym. React. Eng*, 11:563–588, 2003.

C. Zheng and Z. Geng. Reverse engineering of asynchronous boolean networks via minimum explanatory set and maximum likelihood. *JMLR: Workshop and Conference Proceedings*, 6, 2009.

JMLR: Workshop and Conference Proceedings 6:225–236, 2010 NIPS 2008 workshop on causality

The Use of Bernoulli Mixture Models for Identifying Corners of a Hypercube and Extracting Boolean Rules From Data

Mehreen Saeed MEHREEN.SAEED@NU.EDU.PK
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore Campus, Pakistan

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

This paper describes the use of Bernoulli mixture models for extracting boolean rules from data. Bernoulli mixtures identify high data density areas on the corners of a hypercube. One corner represents a conjunction of literals in a boolean clause and the set of all identified corners, of the hypercube, indicates disjuncts of clauses to form a rule. Further class labels can be used to select features or variables, in the individual conjuncts, that are relevant to the target variable. This method was applied to the SIGNET dataset of the causality workbench challenge. The dataset is derived from a biological signaling network with 21 time steps and 43 random boolean variables. Results indicate that Bernoulli mixtures are quite effective at extracting boolean rules from data.

Keywords: boolean networks, boolean rules, Bernoulli mixture models, feature selection

1. Introduction

Recently, the causality workbench team launched a challenge that involved many tasks related to causal discovery (see http://www.causality.inf.ethz.ch/pot-luck.php). One of the tasks, called 'SIGNET', was to discover boolean rules from raw data. The data was generated by the simulation of boolean rules, which describe the interaction of various variables in a boolean network representing a plant signaling network (Li et al., 2006). This is an interesting problem involving not only the modeling of time series data but also developing feature selection algorithms to explain the causes of a target variable.

The SIGNET data was derived by simulating an asynchronous boolean network. Boolean networks find important applications in many areas, especially for modeling biological systems such as plant signaling networks, genetic regulatory networks, signal transduction pathways, etc. In the past many scientists have modeled boolean networks using various techniques. A well established method was developed by Ideker et al. (2000) for inferring a genetic network from gene expression data. They used a method, called the predictor method, to infer a set of genetic networks. The genetic networks were derived using the concept of minimum set covering. A 'chooser' method, based

upon entropy, was then used to discriminate between various networks, generated by the predictor. The predictor and chooser were used iteratively to derive the final gene network.

Other methods for modeling regulatory networks include graphical models like Bayesian networks (Friedman et al., 2000), which are directed graphs representing joint probabilities of the variables in the network and also captures the conditional independences between them. Neural networks have also been used for gene expression analysis and representing regulatory relationships between genes in a genetic network (Weaver et al., 1999). Such networks have also been constructed using the information theoretic measure, i.e., mutual information criterion between the input and output states (Liang et al., 1998).

This paper describes a novel technique for extracting boolean rules from the 'SIGNET' dataset by converting a probabilistic model of variable relationships into a rule based system. The probabilistic model is governed by a mixture of Bernoulli distributions. Mixture densities determine different groups or clusters, within data, based upon the various observational characteristics of data. They are ideal for estimating the overall probability distribution of data when the data is not uni-modal. The use of Bernoulli mixture models in machine learning and pattern classification is not new. The basic formula for a Bernoulli mixture model was first proposed by Duda and Hart (1973). They have been successfully used for OCR tasks by Juan and Vidal (2004) and Grim et al. (2000) and in supervised text classification tasks (Juan and Vidal, 2002). Bernoulli mixtures have also been used for supervised dimensionality reduction tasks (Sajama and Orlitsky, 2005). Prior to this work we used them for dimensionality reduction and feature selection (Saeed, 2008; Saeed and Babri, 2008) tasks.

Bernoulli mixture models identify areas of high data density in the form of probability vectors. We threshold these probabilities to obtain points that lie on the corners of a hypercube. Each corner of the hypercube represents a clause containing the conjunction of literals in a binary rule. Together the set of different corners specify the disjunction of various clauses to specify a complete rule. For feature selection within these rules we partition the data into 0 and 1 class labels and generate Bernoulli mixtures from these rules separately. We then discard features in the corresponding mixtures, based upon their probability values. The results obtained on the SIGNET dataset show that this method is quite effective in rule extraction from raw data.

The outline of this paper is as follows: Section 2 introduces the reader to the notion of Bernoulli mixture models and how they can be used for boolean rules extraction. Section 3 describes the results of applying this method to the 'SIGNET' task. Finally, the paper concludes via Section 4.

2. Multivariate Bernoulli Mixtures

A probability mixture model represents an overall probability distribution of data via a convex combination of various component probability distributions also called mixtures. Each mixture is a probability distribution over a discrete or continuous variable

and has its own set of parameters. A mixture model with D components is described by a probability function given by $p(\mathbf{x}) = \sum_{d=1}^{D} \pi_d p(\mathbf{x}|d)$. Here, π_d is the prior or the mixing proportion of each mixture so that $\sum_d \pi_d = 1$ and $p(\mathbf{x}|d)$ is its component-conditional probability function.

A multivariate Bernoulli mixture model assumes that each component of the model is an n-dimensional multivariate Bernoulli probability distribution. Suppose we have data given by $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. For a single binary vector $\mathbf{x}_k \in \{0,1\}^n$, the form of this distribution, in the d^{th} mixture is given by (Bishop, 2006):

$$p(\mathbf{x}_k|d) = \prod_{i=1}^{n} (p_{di})^{x_{ki}} (1-p_{di})^{1-x_{ki}} \qquad \forall k, 1 \leq k \leq m, \forall d, 1 \leq d \leq D$$

Here, $p_{di} \in [0,1]$ is the probability of success of the i^{th} component of vector $\mathbf{x}_k$ for the d^{th} mixture, i.e., $p_{di} = p(x_{ki} = 1|d)$. We assume that the n-dimensional vector $\mathbf{x}$ has n independent component attributes. The parameter θ governing this distribution is the probability of success for each attribute of vector $\mathbf{x}$, i.e., $\theta = \mathbf{p}$ where $\mathbf{p} \in [0,1]^n$. We can use any appropriate optimization algorithm to estimate the parameters of these mixtures. For our work we used expectation maximization (EM) algorithm to estimate these parameter values from raw data.

2.1. Bernoulli Mixtures For Identifying Edges of a Hypercube and Extracting Boolean Rules from Data

In essence, one Bernoulli mixture is a vector of probabilities, each component representing the probability/chances of success of an individual feature or attribute. A single Bernoulli distribution over the entire data does not tell us anything about the inter-relationship of variables with each other. However, a mixture of such distributions can be used to determine the covariances and hence the correlations between pairs of attributes. They can also be used to identify high data density areas on the corners of a hypercube by thresholding the probability vector. We extract a vector $\mathbf{v}_d$ from a probability vector $\mathbf{p}_d$ and call it the main vector. The value of an attribute in a main vector can be taken as a 1 (0) if its probability is greater (less) than a certain threshold. The probability values around 0.5 can be taken as don't cares. So mathematically,

$$v_{di} = \begin{cases} 1 & \text{if } p_{di} > \alpha \\ 0 & \text{if } p_{di} < 1-\alpha \\ \phi & \text{otherwise} \end{cases} \tag{1}$$

where ϕ represents a don't care value. As an example consider Figure 1. Here, we have two mixtures extracted from three dimensional data and two corresponding main vectors. The figure shows the corners of the hypercube represented by these vectors and the corresponding boolean rule extracted from the 2 mixtures.

Looking at Figure 1, we can see that it is easy to infer boolean rules from data once the corners of the hypercube are identified. Each corner represents a conjunct of boolean variables and together the set, of all the corners, forms a disjunction of rules

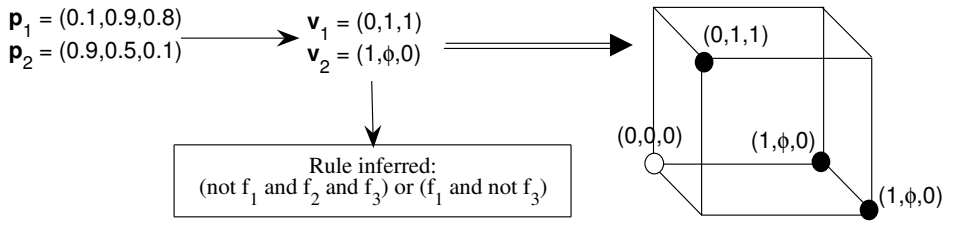


Figure 1: An example with two mixtures

to give us a disjunctive normal form of a boolean rule. The variables with a don't care value do not form a part of the rule.

2.2. Feature Selection in Rules

Once we have inferred the various disjuncts of a boolean rule we need to further identify the features that are irrelevant to our target variable. Suppose our training data is given by a set of m input vectors, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, $\mathbf{x}_k \in \{0,1\}^n$. Each vector $\mathbf{x}_k$ is assigned a boolean label, $y_k \in \{0,1\}$, called the target variable. We partition the data into the two corresponding classes of 0 and 1 labels and generate D and D' Bernoulli mixture models from both these sets separately. Let $M = \{\mathbf{p}_d; \pi_d\}_{d=1}^{D}$ be the mixtures extracted from instances with target variable equal to 1 and $M' = \{\mathbf{p}'_d; \pi'_{d'}\}_{d'=1}^{D'}$ be the mixtures estimated from data with 0 class labels.

Let $r = \{\mathbf{v}_d\}_{d=1}^{D}$ be the main vectors for mixtures in M and $r' = \{\mathbf{v}'_{d'}\}_{d'=1}^{D'}$ be the corresponding main vectors for mixtures in M'. A feature, which has don't care values in all vectors of r (r') is eliminated automatically for the rules for target variable = 1 (0). Also, if a feature's value does not change in both r and r' then it means that its value doesn't affect the target variable and hence this variable is irrelevant for its prediction. Hence, we can discard the literal/feature f from a rule if its value is the same in both the 0 and 1 labels. This variable is, therefore, assigned a don't care value, i.e.,

$$v_{df} = v'_{d'f} = \phi \quad \text{if } v_{df} = v_{d'f} \quad \forall d, \forall d'$$

Figure 2 illustrates the generation of rules for 0 and 1 class labels. Here we can see that the value of feature 2 does not vary in the main vector for both the classes and hence we can assign it a don't care value and eliminate it from the rules.

The above mentioned procedure gives us rules for both 0 and 1 class labels for a target variable. Ideally, one rule should be the negation of the other. However, it is not the case as we treat the data with 0 and 1 labels separately. Out of the two rules, we choose the rule, which gives us better accuracy over the training set. A point to note here is that, our current model does not make any use of mixture priors. The mixture priors in a way represent the proportion of data being generated by a particular distribution. So even if part of the data (say a larger percentage, e.g., 80%) was generated by one mixture and a part of it (say a smaller percentage, e.g., 20%) by another mixture, we should still take both mixtures into account in the DNF of the rule for that feature

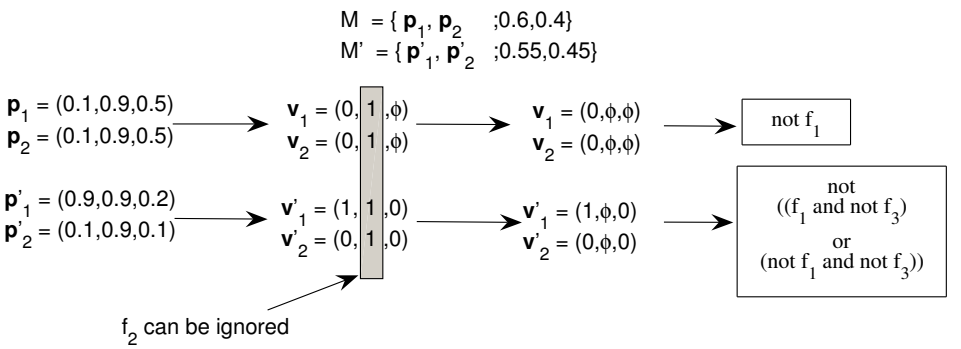


Figure 2: An example with two classes. The second rule is a negation as it is derived from 0 class labels.

as both the mixtures account for the generation of data and constitute the individual clauses in the rule.

Now we briefly outline the main steps of our algorithm:

1. For each variable i in the dataset repeat the following:
   i. Build the training data by using variable i as the target/class and the values of the rest of variables as input data.
   ii. Partition the data into 0 and 1 class labels
   iii. Generate mixtures M' and M for the corresponding 0 and 1 class labels
   iv. Generate the set of main vectors r' and r from both M' and M
   v. Discard features, which have the same values in the main vectors for both class labels
   vi. Generate two boolean rules from 0 and 1 class labels
   vii. Check the training accuracy for both classes and output the rule with maximum training accuracy

3. Simulations

The method described in this paper was applied to the causality workbench's 'SIGNET' data describing the interaction of variables inside a plant signaling network (Li et al., 2006). The dataset includes 300 pseudo dynamic simulations of 43 boolean rules. Each simulation starts with a randomly generated boolean vector representing the initial state and is depicted by a 21x43 sized matrix. The next 20 vectors in a simulation are output by the boolean pseudo dynamic simulation using an asynchronous update scheme. Our task is to extract 43 boolean rules from the simulation data.

We applied the algorithm of Section 2.2 to extract boolean rules from the SIGNET data. Since all simulations result in stable values after a number of time steps, the dataset includes many repetitive training instances. To alleviate this problem, we removed duplicate entries. We emphasize that we are not using only stable states to

generate the mixtures. All states present in the training data are used. Our data preparation only avoids biasing the data distribution in favor of stable states. For evaluation of the extracted rule, the organizers, of the challenge, suggested generating the truth tables for each true rule and computing a prediction error rate by comparing the output values of the extracted rule to the output value of the actual rule. The overall error rate is thus calculated by averaging over all the rules. The corresponding Matlab code for evaluating the system was also provided to us. In addition to calculating the overall error rate of prediction, we also calculated the training set error of the inferred rules. This was done by applying each rule to the individual boolean vectors of the simulation data and predicting the output value. The output value was compared with the actual value of simulation to get the overall training accuracy rate.

The main parameter of our model is the number of mixtures for positive and negative classes and the threshold α of Eq. (1). If the number of data points in a particular set were less than 100 then we generated only one mixture for this data and if the total data points were less than 350 then we generated only 2 mixtures from this set. The table and graph of Figure 3 show the training accuracy and overall accuracy of our method when generating different number of mixtures and using a threshold value of 0.8 on the original provided data. The results obtained after varying the threshold values are shown in Table 1 and discussed later.

Mixtures	**% Training Accuracy**	**% Evaluation Accuracy**
1	95.19	87.80
2	93.89	79.49
3	94.55	82.98
4	91.23	81.03
5	92.72	76.86
6	93.09	75.23
7	93.12	76.29
8	93.23	75.19
10	91.94	76.25

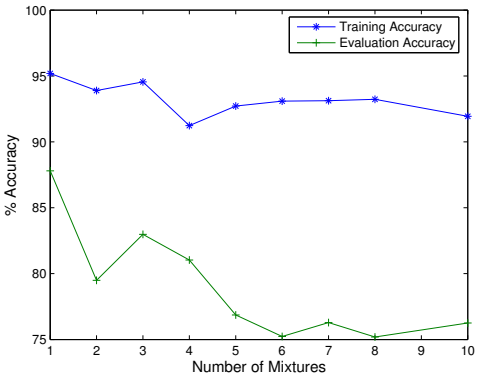


Figure 3: Results with different mixtures using original data (threshold = 0.8)

When generating the rules we noticed that some rules cannot be inferred from data alone. For example, consider the following true rules:

MALATE = PEPC and not ABA and not AnionEM

ABA = 1

Since ABA is 1, hence, MALATE = 0, no matter what the values of the other variables are. However, when evaluating this rule, the system will generate the truth tables for three variables for MALATE and match the output of this rule with ours, resulting in very low accuracy. Hence, we replaced the constants ABA and AGB with 1 in the actual rules. In light of this, the organizers of the challenge generated a new dataset using our new rules. We repeated our experiments on the new dataset and the results obtained are illustrated in Figure 4.

Mixtures	**% Training Accuracy**	**% Evaluation Accuracy**
1	95.41	86.97
2	95.60	85.61
3	95.88	87.61
4	95.37	81.72
5	95.44	83.68
6	94.65	81.04
7	94.63	82.32
8	94.83	79.2
10	94.42	80.87

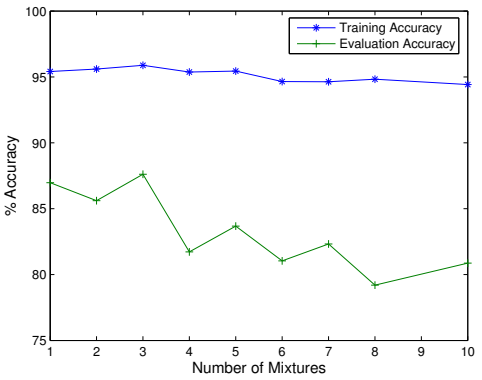


Figure 4: Results with different mixtures using new data (threshold = 0.8)

Figure 3 shows that the best result on the original data was obtained by using only one mixture. This shows that generating different mixtures is not exactly necessary for inferring a rule based systems. However, as pointed out earlier the data is not consistent with the original set of rules. When we repeated our experiments with the new data generated according to the new set of rules we get the accuracies shown in Figure 4. Here, we can see that using 3 mixtures gives us the best results. As we increase the number of mixtures to 10 the accuracy deteriorates considerably owing to the generation of too many clauses for a rule.

To see the effect of threshold α on the quality of the generated rules, we repeated our experiments for different values of α using 1, 2 and 3 mixtures. The results are illustrated in Table 1. Here, we can see that the best overall results are obtained for a threshold value of 0.8. The results for a threshold of 0.7 are good for only a single mixture, but the accuracy for 2 and 3 mixtures is low as compared to the corresponding threshold of 0.8. When α is increased to 0.9 the accuracies deteriorate considerably. This is owing to the fact that too many values are being converted to don't care values for high threshold. In this case it is not possible to capture the actual rules governing the data.

	threshold 0.7		**threshold 0.8**		**threshold 0.9**	
Mix-tures	**% Train Acc**	**% Evaluation Acc**	**% Train Acc**	**% Evaluation Acc**	**% Train Acc**	**% Evaluation Acc**
1	95.84	87.98	95.41	86.97	93.68	76.04
2	95.31	81.47	95.60	85.61	93.68	77.32
3	95.36	81.66	95.88	87.61	94.11	75.82

Table 1: Results with different threshold values on the new dataset

Our method of converting a probabilistic model into a rule based system is quite intuitive and its usefulness is confirmed from experimental results on the SIGNET dataset. We can see that with a threshold value of 0.8 we get a good rule base with three mixtures, showing that most of the attributes can be modeled with rules that have three conjuncts inside them. Also, we get a good result with a single mixture at a threshold of 0.7.

All the results, demonstrated in this section, are, so far, based upon the algorithm described in Section 2.2. According to Step vi and vii of the algorithm, we generate two boolean rules for both the class 0 and class 1 labels. Out of these two rules we select the rule, which has a higher training set accuracy. This may not be the best strategy as the training set data might overestimate the accuracy of a rule. In order to explore this further, we generated new data using the code provided by the organizers of the challenge. We varied the number of simulations in each experiment. For each simulation, we performed rule selection based upon training set accuracy, and then repeated the experiment by performing rule selection based upon validation set accuracy. For the later experiment, we set aside 20% simulations to constitute the validation set and generated the rules using the rest of 80% data. The results of the two methods, as a function of the number of simulations, are shown in Figure 5. Interestingly, for a smaller number of simulations, the method of selection based upon the validation set accuracy outperforms that of selection based upon the training set accuracy. As we increase the number of simulations, the method of selecting a rule based upon training set accuracy gives better results. This is because for smaller datasets, the rule is overfitting the data and hence gives poor performance. In general, we can see that the accuracy of the rule based system does not vary much with the increase in the number of simulations. It is drastically low for 20 simulations (74% using validation set) and then reaches the highest for 400 simulations (85% using validation set). The average accuracy (accuracy averaged over

all the simulations) of rules selected via the training set and the validation set is the same, i.e., around 81%. Hence the validation set does not give us any added advantage over selection with the training set. The important thing to note here is that we can use rule selection based upon training set accuracy for large dataset sizes.

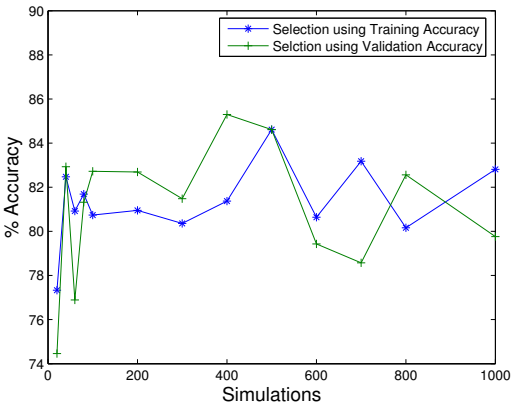


Figure 5: Results with different simulations using 3 mixtures (threshold = 0.8)

4. Discussion and Conclusions

Bernoulli mixtures provide a simple and intuitive method for inferring rules from raw data. The method, described in this paper, is simple and easy to implement. Our approach is novel as it shows how we can convert a probabilistic model into a rule based system. The concept applied here is different from the one adopted by Ideker et al. (2000). They are using only the stable values to infer boolean networks. Here, we are also making use of unstable values to deduce boolean rules. We use mixture models for feature selection and rule deduction. Ideker *et al.* , on the other hand, have used the concept of minimum set covering using greedy searches to determine various perturbations of genetic networks and entropy based measures to select the best network out of these networks.

Zheng and Geng (2008) were the other participants of the causality challenge who presented a solution for the SIGNET task. They have defined a methodology for identifying asynchronous networks. Their method is based upon the concept of finding the minimum explanatory set for a node, similar to Ideker et al. (2000)'s idea of finding minimum set covering. The log likelihood for a particular node is then maximized to determine its corresponding boolean function. Their method depends upon the number of assumed parent nodes and its complexity increases as the number of assumed parent nodes is increased. They achieved an impressive average accuracy of above 99% for a single parent node. However, this method might be too expensive to implement for networks where the in-degree of a node is more than one.

Currently we are converting a probabilistic model into a rule based system. The probabilistic model based on Bernoulli mixtures is a continuous model which is discretized to infer the boolean rules governing data. The priors of mixtures are completely ignored and not used. Another possibility could be to retain the continuous properties of our model and make predictions based on this model. The predictions can then be thresholded to attain truth tables and consequently the DNF of the individual rules.

x_1	x_2	x_1 **XOR** x_2
0	0	0
0	1	1
1	0	1
1	1	0

$\longrightarrow$

Rule from class label 0:
not((not x_1 and not x_2) or (x_1 and x_2))

Rule from class label 1:
((x_1 and not x_2) or (not x_1 and x_2))

Figure 6: The XOR problem and rule extraction

A critical parameter of our model is the actual number of mixtures to generate from data. If we can correctly identify this parameter, then it will increase our chances of inferring rules, from data, that have a higher degree of accuracy. As an example consider the truth table for the XOR problem in Figure 6. The corresponding rules are also shown in the figure. If we know in advance that our data can be modeled by two mixtures for each class then we can come up with accurate rules to represent the data. The rules extracted from both the zero and one class labels are also logically equivalent. However, for this problem one mixture will not be sufficient to infer the rules accurately. We need to come up with a good model selection technique that determines the actual number of mixtures to generate from data. There are several possibilities like the minimum message length (MML) criterion as suggested by Figueiredo and Jain (2002) is based upon concepts from information/coding theory. Minimum description length (MDL) and MML formally coincide with Bayesian inference criterion (BIC) (Schwarz, 1978), which is another possible approach to model selection.

In this paper, we have illustrated how a probabilistic model based upon Bernoulli mixtures can be converted into a rule based system. Two critical parameters of our method are the number of mixtures used to generate the rules and probability threshold value used to determine don't care values in rules. We evaluated the performance of our method by varying both these parameters and achieved good results when generating 3 mixtures with a threshold value of 0.8. We also evaluated the behavior of our method by varying the number of simulations or experiments to generate the raw data from which mixtures are learnt. We found that for very small number of simulations the performance of our rule based system is not very good. However, as we increase the number of simulations, the performance of the system also improves. Here, we have presented some preliminary work and showed how the basic concept of converting a mixture model into a rule based system can be used to give good performance. The method presented here is straightforward and intuitive and can be easily implemented.

Acknowledgments

I would like to thank Isabelle Guyon for her technical support, guidance and prompt replies to all my queries. I also thank Professor Haroon Babri and Mr. Kashif Javed of UET Lahore, for the useful discussions we had on this topic. Another thanks goes to the editors and the anonymous reviewers for their useful comments and feedback which helped me improve the quality of this paper

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

Mario A.T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):381–396, March 2002.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000.

Jiri Grim, Pavel Pudil, and Petr Somol. Multivariate structural Bernoulli mixtures for recognition of handwritten numerals. In *Proceedings of International Conference on Pattern Recognition (ICPR'00)*, 2000.

Trey E. Ideker, Vesteinn Thorsson, and Richard M. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. *Pacific Symposium on Biocomputing*, 5:302–313, 2000.

Alfons Juan and Enrique Vidal. On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705–2710, December 2002.

Alfons Juan and Enrique Vidal. Bernoulli mixture models for binary images. In *Proceedings of 17th International Conference on Pattern Recognition (ICPR'04)*, 2004.

Song Li, Sarah M. Assmann1, and Re ka Albert. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, 4(10):1732–1748, 2006.

Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 3:18–29, 1998.

Mehreen Saeed. *Hands-on pattern recognition challenges in data representation, model selection, and performance prediction*, chapter Hybrid learning using mixture models and artificial neural networks. 2008. To appear, see http://www.clopinet.com/ChallengeBook.html.

Mehreen Saeed and Haroon Babri. Classifiers based on Bernoulli mixture models for text mining and handwriting recognition. In *Proceedings of International Joint Conference on Neural Networks, IEEE WCCI*, 2008.

Sajama and Alon Orlitsky. Supervised dimensionality reduction using mixture models. In *Proceedings of the 22nd international conference on machine learning*, pages 768–775, Bonn, Germany, 2005.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, March 1978.

D.C. Weaver, Christopher T. Workman, and Gary D. Stormo. Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing*, 4:112–123, 1999.

Cheng Zheng and Zhi Geng. Reverse engineering of asynchronous boolean networks via minimum explanatory set and maximum likelihood. In *NIPS 2008 Workshop on Causality*, 2008. see http://clopinet.com/isabelle/Projects/NIPS2008/.

Appendix A. Pot-luck causality challenge: FACT SHEET (SIGNET)

Title: The Use of Bernoulli Mixture Models for Identifying Corners of a Hypercube and Extracting Boolean Rules From Data

Participant name, address, email and website: Mehreen Saeed, FAST National University of Computer and Emerging Sciences, Lahore Campus, Pakistan.

email: mehreen.saeed@nu.edu.pk

Task(s) solved: SIGNET

Reference: Not yet published

Method:

Profile of the method:

- **Preprocessing :** None
- **Causal discovery:** None
- **Feature selection:** Bernoulli mixtures were generated from data and the probabilities in the mixtures were thresholded to identify corners of the hypercube that represent the data. From there, only those features were selected which had varying values in class 1 and class 2.
- **Classification :** The task did not involve classification, only rule generation.

Mixtures	% Training Accuracy	% Evaluation Accuracy
1	95.41	86.97
2	95.60	85.61
3	95.88	87.61
4	95.37	81.72
5	95.44	83.68
6	94.65	81.04
7	94.63	82.32
8	94.83	79.2
10	94.42	80.87

Table 2: Results with different mixtures and threshold of 0.8)

- **Model selection/hyperparameter selection:** None

Results: Shown in Table 2

- **quantitative advantages:** Simplicity, computationally easy, intuitive
- **qualitative advantages:** Converts a probabilistic model into a rule based model which is novel.

Implementation: Code for Bernoulli mixtures was written in C++. The code for boolean rule generation was implemented in Matlab

Keywords:

- **Preprocessing or feature construction:** None
- **Causal discovery:** None
- **Feature selection:** Bernoulli mixtures
- **Classifier:** None
- **Hyper-parameter selection:** None
- **Other:** Boolean rule generation using Bernoulli mixtures

JMLR: Workshop and Conference Proceedings 6:237–248, 2010 NIPS 2008 workshop on causality

Reverse Engineering of Asynchronous Boolean Networks via Minimum Explanatory Set and Maximum Likelihood

Cheng Zheng ZZHENGCCHENG@PKU.EDU.CN
Yuanpei College
Peking University
Beijing, 100871, China

Zhi Geng ZGENG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, 100871, China

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

In this paper, we propose an approach for reconstructing asynchronous Boolean networks from observed data. We find the causal relationships in Boolean networks using an asynchronous evolution approach. In our approach, we first find a minimum explanatory set for a node to reduce complexity of candidate Boolean functions, and then we choose a Boolean function for the node based on the maximum likelihood. This approach is stimulated by the task SIGNET of the causal challenge #2 pot-luck Jenkins (2009). Besides the data set SIGNET, we also applied our approach to two other datasets to evaluate our approach: one is generated by Professor Isabelle Guyon and the other generated ourselves from the signal transduction network of Abscisic acid in guard cell.

Keywords: Boolean network, Reverse engineering, Structural learning

1. Introduction

Boolean network is a useful model in many applications, such as gene regulation networks Thomas (1973) and signal pathway Davidich and Bornholdt (2008). To simulate real biological networks better, various types of Boolean networks are developed, such as probabilistic Boolean networks Shmulevich et al. (2002), and asynchronous Boolean networks Kauffman et al. (2003); Harvey and Bossomaier (1997). Because of the deficiency of knowledge about related chemical reaction speeds, physicists often use the asynchronous evolution approach to study the general dynamic characteristic of these networks Chaves et al. (2005). By using the asynchronous approach, the relative timing of each reaction is chosen randomly for each update round, which is defined as the longest time for a node to respond to a change of its parent nodes. In studies of gene regulation networks, the environmental factors may change during the experiment process, and thus the reaction velocity may be affected by some external variables. Such problems can be treated by using an asynchronous Boolean network. Although many

approaches have been developed for learning Boolean networks from steady states data, which are equivalent to learning the synchronous networks Ideker et al. (2000); Nam et al. (2006), there are a few of approaches for reconstructing asynchronous Boolean network Jenkins (2009); Saeed (2009). In this paper, we propose an approach for reconstructing asynchronous Boolean networks.

In Section 2, we describe the model of asynchronous networks, and we briefly describe the format of the dataset generated from a guard cell signaling network given in Li et al. (2006). In Section 3, we discuss identification of asynchronous Boolean networks. In Section 4, we first give the definition of candidate explanatory sets and propose an algorithm for finding the minimum explanatory set. Then we propose a method for choosing a Boolean function of parent nodes for each node from all possible functions based on the maximum likelihood method. Algorithm complexity and simulation results are given in Section 5. Some possible improvement of our approach and the disadvantage of our approach are discussed in Section 6. The result for the challenge problem is given in Appendix B.

2. Model of Asynchronous Boolean Networks

A Boolean network is represented as a directed graph with N nodes where the orientation of the edges shows the causal relationship among variables. Suppose that the state of each node is determined by a Boolean regulatory function of the states of its parent nodes. For an asynchronous Boolean network, the order to update all nodes in every update round is random. The Boolean updating rules of an asynchronous network can be described as

$$\mathcal{S}_i^t = \mathcal{B}_i(\mathcal{S}_{i_1}^{t-\mathcal{I}(R_{i_1}^{t-1}>R_i^{t-1})}, \mathcal{S}_{i_2}^{t-\mathcal{I}(R_{i_2}^{t-1}>R_i^{t-1})}, \cdots, \mathcal{S}_{i_{n_i}}^{t-\mathcal{I}(R_{i_{n_i}}^{t-1}>R_i^{t-1})}) \tag{1}$$

for node $i = 1, \ldots, N$, where $\mathcal{S}_i^t$ denotes the state of node i in round t, $\mathcal{B}_i(\cdot)$ denotes the Boolean function for node i, R_i^t with the domain $\{1, \cdots, N\}$ denotes the update order of node i in round t, n_i is the number of parent nodes of node i, and $\mathcal{I}(A)$ is defined as

$$\mathcal{I}(A) = \begin{cases} 1, & A = True; \\ 0, & A = False. \end{cases} \tag{2}$$

We write the formula (1) simply as

$$\mathcal{S}_i^* = \mathcal{B}_i(\mathcal{S}_{i_1}, \mathcal{S}_{i_2}, \cdots, \mathcal{S}_{i_{n_i}}). \tag{3}$$

The observed data generated from different initial values are sequentially arranged in a matrix E with N columns and $(T+1)M$ rows, where T denotes the number of update rounds for every initial value and M denotes the number of initial values. The first $T+1$ rows are the first initial value and the data updated consequently in T rounds, the next $T+1$ rows are the data for the second initial value, and finally the last $T+1$ rows are the data for the Mth initial value.

3. Identifiability of Asynchronous Boolean Networks

For an asynchronous update rule, we can obtain more states of a network than a synchronous update rule. Thus more observed states can be used to reconstruct an asynchronous network. However, a disadvantage for an asynchronous network is that we cannot know from the data at which time point the state of a node is affected by the states of its parents. As shown in the following example, there is a network structure which cannot be identified from observed data if the data are generated from an asynchronous network, but which can be identified if the data are generated from a synchronous network.

Example 1 Consider the following two Boolean networks

$$\mathcal{S}_A^* = \mathcal{S}_B,\ \mathcal{S}_B^* = \neg\mathcal{S}_A,\ \mathcal{S}_C^* = (\mathcal{S}_A \wedge \mathcal{S}_B) \vee (\neg\mathcal{S}_A \wedge \neg\mathcal{S}_B),\ \mathcal{S}_D^* = \mathcal{S}_E,\ \mathcal{S}_E^* = \neg\mathcal{S}_D,$$

$$\mathcal{S}_A^* = \mathcal{S}_B,\ \mathcal{S}_B^* = \neg\mathcal{S}_A,\ \mathcal{S}_C^* = (\mathcal{S}_D \wedge \mathcal{S}_E) \vee (\neg\mathcal{S}_D \wedge \neg\mathcal{S}_E),\ \mathcal{S}_D^* = \mathcal{S}_E,\ \mathcal{S}_E^* = \neg\mathcal{S}_D.$$

The two networks cannot be distinguished if they are treated as asynchronous rules since $\mathcal{S}_C^t$ is independent of $(\mathcal{S}_A^t, \mathcal{S}_A^{t-1}, \mathcal{S}_B^t, \mathcal{S}_B^{t-1})$ and $(\mathcal{S}_D^t, \mathcal{S}_D^{t-1}, \mathcal{S}_E^t, \mathcal{S}_E^{t-1})$, which is shown in Appendix A. However, these two networks can easily be distinguished when they are treated as synchronous rules. It is because, for the initial value $(\mathcal{S}_A, \mathcal{S}_B, \mathcal{S}_C, \mathcal{S}_D, \mathcal{S}_E) = (1,1,0,1,0)$, the first network has the updated value $(1,0,1,0,0)$ after one update round, while the second network has the value $(1,0,0,0,0)$.

We guess that if an asynchronous network has a steady state for each initial value, then the network may be identifiable. The identification means that if we obtain enough initial states and enough update processes, we can discover the structure and Boolean rules for the whole network correctly.

4. Reconstruction of Asynchronous Boolean Networks

As shown in formula (1), the state of node i in round t can be affected only by the state of its parents in round t or round $t-1$. For each time point t, we combine the states of N nodes in both rounds t and $t-1$ into one row, and we get a data matrix with $2N$ columns and $T \times M$ rows for M initial values and T rounds. In this section, we propose an algorithm MESML for finding Boolean functions in which for a particular node, first the Minimum Explanatory Set is found, and then a Boolean function that has the Maximum Likelihood is chosen.

4.1. Find the minimum explanatory set

Ideker et al. (2000) proposed an algorithm for reconstructing a synchronous Boolean network. The network contains N nodes: $a_1, \cdots, a_N$, and each node a_i has a Boolean function f_i. Observed data are given in a matrix E', whose each row is for an individual and each column is for a variable. For every node a_i, assume that a Boolean function $a_i = f_i(a_{i_1}, \cdots, a_{i_{n_i}})$ always holds for all rows. Their algorithm has the following three steps. For every node a_i,

1. consider every pair of rows (t_1, t_2) in E' such that the states of a_i differs. For each pair, find the set V_{t_1,t_2} of all other nodes whose states are different for these two rows, i.e., $V_{t_1,t_2} = \{j : \mathcal{S}_j^{t_1} \neq \mathcal{S}_j^{t_2}\}$;

2. find the minimum set of nodes V_{min} which intersects all $V_{j,k}$ obtained at Step 1, i.e., $V_{min} \cap V_{j,k} \neq \emptyset$; V_{min} is used to find a Boolean rule f_i;

3. find a truth table from the data. For a deterministic Boolean network, f_i can be obtained from the observed data of variables in V_{min}. For a combination of levels of variables in V_{min} that does not appear in the data, we cannot get the truth value of node a_i for this combination.

For asynchronous networks, it seems that a difference set V_{t_1,t_2} for node i at a pair (t_1, t_2) of time points could be defined similarly to the above case by comparing a variable at two consecutive time points simultaneously. That is, $V_{t_1,t_2} = \{j : \mathcal{S}_j^{t_1} \neq \mathcal{S}_j^{t_2}$ or $\mathcal{S}_j^{t_1-1} \neq \mathcal{S}_j^{t_2-1}\}$. In the following example, we show that such a difference set is improper.

Example 2 Consider the asynchronous Boolean network

$$\mathcal{S}_A^* = 1,\ \mathcal{S}_B^* = \neg\mathcal{S}_A,\ \mathcal{S}_C^* = \neg\mathcal{S}_B.$$

For the single value $(0,0,0)$ of $(\mathcal{S}_A^t, \mathcal{S}_B^t, \mathcal{S}_C^t)$, we may get the following two different data $(1,1,0)$ or $(1,1,1)$ after one update round. By the above definition of a difference set, we get for variable C that $V_{t_1,t_2} = \emptyset$ for these two data, and thus we cannot get V_{min}, which means that no variable can explain variable C.

To avoid the above mistake, we can revise the definition of the difference set for node i as $V_{t_1,t_2} = \{j : \mathcal{S}_j^{t_1} \neq \mathcal{S}_j^{t_2}$ or $\mathcal{S}_j^{t_1-1} \neq \mathcal{S}_j^{t_2-1}$ or $\mathcal{S}_j^{t_1} \neq \mathcal{S}_j^{t_1-1}\}$. By the revised definition, we get $V_{t_1,t_2} = \{A, B\}$ for Example 2. We say that V is an explanatory set for node i if $V \cap V_{t_1,t_2} \neq \emptyset$ for all pairs (t_1, t_2). In our approach, we find the minimum explanatory set V as a candidate set of parent nodes of node i. This is called the minimum set covering and can be calculated by the branch and bound technique Nemhauser (1988). By the revised difference set, it may be shown that the parent set of node i can be found by using an explanatory set. If a node has n parents, then the algorithm can stop before we search all sets with not more than n elements. Unfortunately, for an explanatory set V, we cannot ensure the existence of a Boolean function that can explain the whole data (see Example 3). Thus unlike the approach for synchronous networks, to decide whether an explanatory set is appropriate, we need to check the existence of a Boolean function which can explain the whole data. If there does not exist such a function, we try the next minimum explanatory set V that satisfies $V \cap V_{t_1,t_2} \neq \emptyset$ for all t_1 and t_2. We repeat this process until finding a Boolean function.

Example 3 This example illustrates that an explanatory set V may not ensure the existence of a Boolean function for explaining all data from an asynchronous network. Consider the following asynchronous network

$$S_A^* = 1,\ S_B^* = S_A,\ S_C^* = \neg S_A.$$

From it, we may get the following data of $(\mathcal{S}_A^t, \mathcal{S}_B^t, \mathcal{S}_C^t; \mathcal{S}_A^{t+1}, \mathcal{S}_B^{t+1}, \mathcal{S}_C^{t+1})$ at three update rounds

$$(1,1,0;1,1,0) \text{ for } t = t_1,\ (0,0,0;1,0,0) \text{ for } t = t_2,\ (0,1,0;1,0,1) \text{ for } t = t_3.$$

It is obvious that $\{B\}$ is an explanatory set of C. From the data at t_1 and t_2, we find that C is a constant 0 for $B = 0$ and 1, and thus we get the function $C = 0$. But for the data at t_3, we have $C_{t_3+1} = 1$. This means that there does not exist a Boolean function $S_C^* = \mathcal{B}_C(S_B)$.

4.2. Find a Boolean Function

To find a Boolean function, we propose the following process. Given the minimum explanatory set $V = \{j_1, j_2, \cdots, j_{n_i}\}$ for node i obtained with the approach presented in the previous subsection, first we extract those rows that satisfy $\mathcal{S}_{j_m}^{t-1} = \mathcal{S}_{j_m}^t$ for all m, and we can directly get a value of the Boolean function from these rows. We extract all different states and calculate a log likelihood for each of possible functions $\mathcal{B}_i(\mathcal{S}_{j_1}, \cdots, \mathcal{S}_{j_{n_i}})$

$$\begin{aligned}
&l[S_i^* = \mathcal{B}_i(\mathcal{S}_{j_1}, \cdots, \mathcal{S}_{j_{n_i}}), data] \\
=&\sum_t \ln P[S_i^t = \mathcal{B}_i(\mathcal{S}_{j_1}^{t-\mathcal{I}(R_{j_1}^{t-1} > R_i^{t-1})}, \cdots, \mathcal{S}_{j_{n_i}}^{t-\mathcal{I}(R_{j_{n_i}}^{t-1} > R_i^{t-1})})] \\
=&\sum_t \ln\{ \sum_{(R_{j_1}^{t-1}, \cdots, R_{j_{n_i}}^{t-1}, R_i^{t-1})} [\mathcal{I}(S_i^t = \mathcal{B}_i(\mathcal{S}_{j_1}^{t-\mathcal{I}(\mathcal{R}_{j_1}^{t-1} > \mathcal{R}_i^{t-1})}, \cdots, \mathcal{S}_{j_{n_i}}^{t-\mathcal{I}(\mathcal{R}_{j_{n_i}}^{t-1} > \mathcal{R}_i^{t-1})})) \\
&\times P(R_{j_1}^{t-1}, \cdots, R_{j_{n_i}}^{t-1}, R_i^{t-1})]\} \\
=&\sum_t \ln\{ \sum_{(a_{j_1}^{t-1}, \cdots, a_{j_{n_i}}^{t-1})} [\mathcal{I}(S_i^t = \mathcal{B}_i(\mathcal{S}_{j_1}^{t-a_{j_1}^{t-1}}, \cdots, \mathcal{S}_{j_{n_i}}^{t-a_{j_{n_i}}^{t-1}})) P(a_{j_1}^{t-1}, \cdots, a_{j_{n_i}}^{t-1})]\},
\end{aligned} \tag{4}$$

where $a_{j_k}^{t-1} = \mathcal{I}(R_{j_k}^{t-1} > R_i^{t-1})$, which has value 0 or 1. The maximum log likelihood equal to $-\infty$ implies that there is not any function $\mathcal{B}_i$ such that $\mathcal{S}_i^* = \mathcal{B}_i(\mathcal{S}_{j_1}, \cdots, \mathcal{S}_{j_{n_i}})$. We find a function $\mathcal{B}_i$ which has the maximum log likelihood. Assume that $(R_1^t, \cdots, R_N^t)$ has a uniform distribution over all permutations of $\{1, 2, \cdots, N\}$ for all t. Then we have

$$P(a_1^t, \cdots, a_{n_i}^t) = \frac{(\sum_{m=1}^{n_i} a_m^t)! \times (n_i - (\sum_{m=1}^{n_i} a_m^t))!}{(n_i+1)!}. \tag{5}$$

4.3. Algorithm MESML for finding Boolean functions

Let $\Omega = \{1, 2, \ldots, (T+1)M\}$ denote a set of row indexes for the data matrix E. Let $\Omega_I = \{1, (T+1)+1, 2(T+1), \ldots, (M-1)(T+1)+1\}$ denote the set of row indexes for M initial values, and let $\Omega_{NI} = \Omega \setminus \Omega_I$ denote the set of row indexes for non-initial data.

For each node $i \in \{1, \ldots, N\}$, we find a Boolean function $\mathcal{B}_i(V)$ as follows:

Step 1 If all non-initial values of variable i are constant, that is, $S_i^t = S_i^{t'}$ for all t and $t' \in \Omega_{NI}$, then return a constant function $\mathcal{B}_i(V) = S_i^t$.

Step 2 For each $t \neq t'$, calculate the difference set $V_{t,t'}$ if $S_i^t \neq S_i^{t'}$.

Step 3 Find a minimum set V which intersects all $V_{t,t'}$ obtained at Step 2, that is, $V \bigcap V_{t,t'} \neq \emptyset$ for all t and $t' \in \Omega_{NI}$.

Step 4 Find a Boolean function $\mathcal{B}_i(V)$ which maximizes the log likelihood $l[\mathcal{B}_i(V), data]$.

Step 5 If $l[\mathcal{B}_i(V), data] = -\infty$, then we cannot obtain a Boolean function of the current set V for the node i and go to Step 3 to try the next minimum set V.

Step 6 Otherwise, output the Boolean function $\mathcal{B}_i(V)$.

Example 4 This example illustrates the algorithm MESML. Consider the following Boolean network with $N = 3$ nodes A, B and C:

$$S_A^* = 1,\ S_B^* = S_A,\ S_C^* = \neg S_A.$$

Suppose that we have $M = 3$ initial vectors of $(\mathcal{S}_A, \mathcal{S}_B, \mathcal{S}_C)$: $(1,1,0)$, $(0,0,0)$ and $(0,1,0)$. After one update round for each initial vector, we get the following data of $(\mathcal{S}_A^t, \mathcal{S}_B^t, \mathcal{S}_C^t; \mathcal{S}_A^{t+1}, \mathcal{S}_B^{t+1}, \mathcal{S}_C^{t+1})$:

1. $(1,1,0;1,1,0)$ for the first initial value labeled as $t = 1$,
2. $(0,0,0;1,0,0)$ for the second initial value labeled as $t = 3$, and
3. $(0,1,0;1,0,1)$ for the third initial value labeled as $t = 5$,

where the first three numbers in each bracket are an initial vector, and the last three numbers are the vector obtained after one update round. To clearly separate the initial values from the updated data, the data matrix E is revised by combining two rows into one as follows

$$\left(\begin{array}{ccc|ccc} \mathcal{S}_A^1 & \mathcal{S}_B^1 & \mathcal{S}_C^1 & \mathcal{S}_A^2 & \mathcal{S}_B^2 & \mathcal{S}_C^2 \\ \mathcal{S}_A^3 & \mathcal{S}_B^3 & \mathcal{S}_C^3 & \mathcal{S}_A^4 & \mathcal{S}_B^4 & \mathcal{S}_C^4 \\ \mathcal{S}_A^5 & \mathcal{S}_B^5 & \mathcal{S}_C^5 & \mathcal{S}_A^6 & \mathcal{S}_B^6 & \mathcal{S}_C^6 \end{array}\right) = \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right),$$

where the left parts in the above matrixes are for the initial values and the right parts are for the data obtained after one update round. First we use the algorithm MESML to find the Boolean function for node A. At Step 1, since $\Omega_{NI} = \{2,4,6\}$ and $\mathcal{S}_A^2 = \mathcal{S}_A^4 = \mathcal{S}_A^6 = 1$, we output $S_A^* = 1$.

Next we try to find a Boolean function for node B. Since $\mathcal{S}_B^2 \neq \mathcal{S}_B^4$, the condition at Step 1 does not hold. At step 2, there are two pairs of (t,t'): (2,4) and (2,6) such that $\mathcal{S}_B^t \neq \mathcal{S}_B^{t'}$. We calculate the difference set for the pair $(2,4)$. Since $\mathcal{S}_A^1 = \mathcal{S}_A^2 = \mathcal{S}_A^4 = 1 \neq 0 = \mathcal{S}_A^3$ and $\mathcal{S}_C^1 = \mathcal{S}_C^2 = \mathcal{S}_C^3 = \mathcal{S}_C^4 = 0$, by definition of difference set, we get $V_{2,4} = \{A\}$. Then we calculate the difference set for the pair $(2,6)$. Since $\mathcal{S}_A^1 = \mathcal{S}_A^2 = \mathcal{S}_A^6 = 1 \neq 0 =$

$\mathcal{S}_A^5$ and $\mathcal{S}_C^1 = \mathcal{S}_C^2 = \mathcal{S}_C^5 = 0 \neq 1 = \mathcal{S}_C^6$, we get $V_{2,6} = \{A, C\}$. At Step 3 we find that for each set that has one element, only $\{A\}$ is an explanatory set. At Step 4, we find the truth table for A and B, and we can obtain the Boolean function $\mathcal{S}_B^* = \mathcal{S}_A$. At Step 6, we output the result.

Finally we want to find a Boolean function for node C. Since $\mathcal{S}_C^2 \neq \mathcal{S}_C^6$, the condition at Step 1 does not hold. At step 2, there are also two pairs of (t,t'): (2,6) and (4,6) such that $\mathcal{S}_C^t \neq \mathcal{S}_C^{t'}$. First we calculate the difference set for the pair $(2,6)$. Since $\mathcal{S}_A^5 = 0 \neq 1 = \mathcal{S}_A^1 = \mathcal{S}_A^2 = \mathcal{S}_A^6$ and $\mathcal{S}_B^1 = \mathcal{S}_B^2 = \mathcal{S}_B^5 = 1 \neq 0 = \mathcal{S}_B^6$, we get $V_{2,6} = \{A, B\}$. Then we calculate the difference set for the pair $(4,6)$. Since $\mathcal{S}_A^4 = \mathcal{S}_A^6 = 1 \neq 0 = \mathcal{S}_A^3 = \mathcal{S}_A^5$ and $\mathcal{S}_B^3 = \mathcal{S}_B^4 = \mathcal{S}_B^6 = 0 \neq 1 = \mathcal{S}_B^5$, we have $V_{4,6} = \{A, B\}$. At Step 3, we find that for each set with a single element, both $\{A\}$ and $\{B\}$ are explanatory sets. At Step 4, we search the Boolean rules for them separately. As shown in Example 3, there does not exist a Boolean rule of $\{B\}$ that can explain the data, which means that $l[\mathcal{B}_C(\mathcal{S}_B), data] = -\infty$. So go back to Step 3 to try the next minimum set. Since there exists a boolean rule of A and C, we get $\mathcal{S}_C^* = \neg\mathcal{S}_A$ such that $l[\mathcal{B}_C(\mathcal{S}_B), data] \neq -\infty$, and then go to Step 6 for output.

5. Simulation and Algorithm Analysis

5.1. Simulation results

From the result obtained by our approach, we find that if the parents of a variable are found correctly, then the Boolean regulatory rule is always obtained correctly. Thus our method based on the maximum likelihood is an efficient way for selecting the Boolean regulatory rule. For our results, the average error rate defined by the challenge organizers is 0.14 for the original SIGNET dataset, 0.06 for our dataset and 0.05 for the dataset generated by Professor Guyon. We also find that the error rate increases as the the number of parent nodes increases. The dataset generated by Professor Guyon is used for further simulation, and the error rate, the number of real parent nodes and CPU time for various cases are shown in Table 1. All of our computations are performed on a computer with CPU 1.73 GHz and 1.00 GB RAM and the algorithm is implemented with R language. The CPU times do not include the data preprocess and the preprocess takes about an hour. In our algorithm, we limit the maximum size of the set of parent nodes up to 4 since the CPU time for more than 5 parent nodes may take more than 30 days. For the last line in Table 1, the number of real parents is 5, but we limit the number of parents to be found up to 4. Thus the CPU time is almost the same as the case with 4 real parents, but the average error rate is much larger than the case with 4 real parents.

5.2. Algorithm Analysis

To find the minimum set, we need to calculate $V_{t,t'}$ for all t and $t' \in \Omega_{NI}$. The complexity is $2N(T \times M)^2$. If a node has K parents, the total computational complexity for finding the minimum explanatory set V is not more than $\sum_{k=1}^{K} C_N^k$ times. Given a minimum set V with K nodes, we need to calculate not more than 2^{2^K} log likelihoods.

Number of real parents	Average error rate	CPU Time
0	0.000	0.1 seconds
1	0.000	10 seconds
2	0.063	10 minutes
3	0.150	12 hours
4	0.110	2.5 days
5	0.281	2.5 days

Table 1: Error rate and CPU time for the different number of parents of a node.

But if we search on the whole space of Boolean functions without using the minimum set, the complexity is about $\sum_{k=1}^{K}(2^{2^k}\times C_N^k)$. So our algorithm can greatly reduce the complexity for a large network with large N and K. However, when K and N are small but M and T are large, it takes much times to find the minimum set, and so we directly try all possible sets V with sizes equal to 1 and 2 without finding the minimum set.

The advantages of our algorithm are that our approach takes full use of the information in the data, especially the information on the dynamic evolution processes. Thus our algorithm can have a higher accuracy for learning structures and Boolean functions, especially for those sparse structures with circles and without steady states.

6. Discussion

In the algorithm MESML, we stop the process as long as we find a single minimum parent set which has a Boolean function that maximizes the likelihood and can explain the whole data set. If we do not consider the computational complexity, we should try all possible minimum parent sets to find a Boolean function which maximizes the likelihood or some other scores over all sets. To compare the models with different numbers of parent nodes, we can use AIC score or other scores for model selection. Pearson's χ^2 test can be used to check whether the Boolean regulatory function fits the observed data. The χ^2 test is described as follow. The hypotheses are

$$H_0 : \mathcal{S}_i^t = \mathcal{B}_i(\mathcal{S}_{j_1}^{t-\mathcal{I}(\mathcal{R}_{j_1}^{t-1}>\mathcal{R}_i^{t-1})},\cdots,\mathcal{S}_{j_{n_i}}^{t-\mathcal{I}(\mathcal{R}_{j_{n_i}}^{t-1}>\mathcal{R}_i^{t-1})}),$$

$$H_1 : \mathcal{S}_i^t \neq \mathcal{B}_i(\mathcal{S}_{j_1}^{t-\mathcal{I}(\mathcal{R}_{j_1}^{t-1}>\mathcal{R}_i^{t-1})},\cdots,\mathcal{S}_{j_{n_i}}^{t-\mathcal{I}(\mathcal{R}_{j_{n_i}}^{t-1}>\mathcal{R}_i^{t-1})}).$$

An asynchronous rule $\mathcal{S}_i^t = \mathcal{B}_i(\cdot)$ may take multiple values depending the orders of parent nodes before or after the order of node i. Under the assumption that the orders have a uniform distribution, we can obtain the distribution of $\mathcal{S}_i^t$ conditional on its parent nodes, which is given in the likelihood (4) (i.e., the argument of ln function). Thus we can test the hypothesis using Pearson's χ^2 statistic:

$$\chi^2 = \sum_{\lambda} T_{i\lambda} = \sum_{\lambda}\sum_{j=0}^{1}\frac{(O_{ij\lambda}-E_{ij\lambda})^2}{E_{ij\lambda}},$$

where λ denotes a state of the conditional set $(S^t_{j_1},\cdots,S^t_{j_{n_i}};\, S^{t-1}_{j_1},\cdots,S^{t-1}_{j_{n_i}})$, $O_{ij\lambda}$ is an observed frequency

$$O_{ij\lambda} = \sum_t \mathcal{I}(S^t_i = j, (S^t_{j_1},\cdots,S^t_{j_{n_i}};S^{t-1}_{j_1},\cdots,S^{t-1}_{j_{n_i}}) = \lambda),$$

$E_{ij\lambda}$ is the expectation

$$E_{ij\lambda} = n_\lambda P(S^t_i = j|(S^t_{j_1},\cdots,S^t_{j_{n_i}};S^{t-1}_{j_1},\cdots,S^{t-1}_{j_{n_i}}) = \lambda;H_0),$$

which can be calculated by (4) and (5), and the total frequency is

$$n_\lambda = \sum_t \mathcal{I}((S^t_{j_1},\cdots,S^t_{j_{n_i}};S^{t-1}_{j_1},\cdots,S^{t-1}_{j_{n_i}}) = \lambda).$$

The statistics $T_{i\lambda}$ for all states of the conditional set are mutually independent. The statistic χ^2 asymptotically has a χ^2 distribution with F degrees of freedom, where F is the number of different states of $(S^t_{j_1},\cdots,S^t_{j_{n_i}};\, S^{t-1}_{j_1},\cdots,S^{t-1}_{j_{n_i}})$.

To evaluate the test, we did a simulation on the above test for checking whether a Boolean rule can be accepted or rejected correctly. Using the dataset of the task SIGNET, we obtain the results shown in Table 2. The significance level α for the test is 0.001. The test results illustrate that Pearson's χ^2 test can be used to improve the accuracy for learning Boolean rules.

Rule	Accept	Reject
True	34	2
False	1	6

Table 2: Simulation results of χ^2 tests for Boolean rules.

The deficiency of our approach is that the accuracy may decrease if the network contains too many hub nodes. We limit the number of parent nodes up to 4 and stop our algorithm after searching all sets with not more than four nodes even if no explanatory set is found. This is a tradeoff between time cost and the error rate.

In our approach, we use a conditional likelihood for a single node i instead of the full likelihood for all nodes to reduce the computational complexity. Thus our approach may not obtain the optimal result. Since we consider a Boolean regulatory rule for each node i separately, the update order we obtained may not be suitable for other Boolean regulatory rules although it maximizes the likelihood of the Boolean rule for node i. Below we give an example to illustrate this.

Example 5 Suppose that the observed data of $(\mathcal{S}^0_A,\mathcal{S}^0_B,\mathcal{S}^1_A,\mathcal{S}^1_B)$ are (1,1,0,0). First consider A only, and the rule $S^*_A = \neg S_B$ can explain the data under the order $(R^1_A = 1, R^1_B = 2)$. Next consider B only, and the rule $S^*_B = \neg S_A$ can also explain the data under another update order $(R^1_A = 2, R^1_B = 1)$. So it seems that the rules $S^*_A = \neg S_B$ and

$S_B^* = \neg S_A$ could explain the data. However, from the rules, we find that possible data are either $(1,1,0,1)$ or $(1,1,1,0)$, which do not contain the observed data $(1,1,0,0)$. Thus we must add one more step after our algorithm to check whether all the Boolean rules we learned can explain the observed data.

Acknowledgments

We would like to thank the four reviewers for their valuable comments. We would appreciate I. Guyon and all the competition organizers for their encouragement and support to our work. This research was supported by NSFC (10771007, 10721403), 863 Project of China (2007AA01Z437), MSRA and MOE-Microsoft Key Laboratory of Statistics and Information Technology of Peking University.

References

M. Chaves, R. Albert, and E. Sontag. Robustness and fragility of boolean models for genetic regulatory networks. *J Theor Biol*, 235:431–449, 2005.

M. Davidich and S. Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One.*, 3:e1672, 2008.

I. Harvey and T. Bossomaier. Time out of joint: Attractors in asynchronous random boolean networks. *Proc. Fourth European Conference on Artificial Life*, pages 67–75, 1997.

T. Ideker, V. Thorsson, and R. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. *Pacific Symp Biocomput*, 5:302–313, 2000.

J. Jenkins. Signet: Boolean rule determination for abscisic acid signaling. *Proc. Machine Learning Research*, 5(In this volume), 2009.

S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Random boolean network models and the yeast transcriptional network. *PNAS.USA*, 100:14796–14799, 2003.

S. Li, S. Assmann, and R. Albert. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLOS Biol*, 4(10),e312:1732–1748, 2006.

D. Nam, S. Seo, and S. Kim. An efficient top-down search algorithm for learning boolean networks of gene expression. *Machine Learning*, 65,1:229–245, 2006.

G. Nemhauser. *Integer and combinatorial optimization*. Wiley, New York, 1988.

M. Saeed. The use of bernoulli mixture models for identifying corners of a hypercube and extracting boolean rules from data. *Proc. Machine Learning Research*, 5(In this volume), 2009.

I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinfo.*, 18:261–274, 2002.

R. Thomas. Boolean formalization of genetic control circuits. *Theor Biol*, 42:563–585, 1973.

Appendix A. Proof of Example 1

Below we show the statement for Example 1: $\mathcal{S}_C^t$ is independent of $(\mathcal{S}_A^t, \mathcal{S}_A^{t-1}, \mathcal{S}_B^t, \mathcal{S}_B^{t-1})$. Since all information on the network is contained by the conditional probabilities $P(S_A^t, S_B^t,\ S_C^t | S_A^{t-1}, S_B^{t-1}, S_C^{t-1})$, we first calculate these probabilities, and then we can calculate the conditional probabilities $P(S_C^t | S_A^{t-1}, S_B^{t-1}, S_A^t, S_B^t)$. It is easy to get that $P(S_C^t = 1 | S_A^{t-1} = i, S_B^{t-1} = j, S_A^t = k, S_B^t = l) = 0.5$ for all i, j, k and l =0 and 1. Thus we showed that $\mathcal{S}_C^t$ is independent of $(\mathcal{S}_A^t, \mathcal{S}_A^{t-1}, \mathcal{S}_B^t, \mathcal{S}_B^{t-1})$. By the symmetry of (A,B) and (D,E), we have that $\mathcal{S}_C^t$ is independent of $(\mathcal{S}_D^t, \mathcal{S}_D^{t-1}, \mathcal{S}_E^t, \mathcal{S}_E^{t-1})$.

Appendix B. Pot-luck challenge: FACT SHEET

Method:

- Preprocessing
  First, we combine the state vectors at two consequent time points into one vector. Then for each variable i, we find a difference set V_{t_1,t_2} of variables for a pair of two combined vectors.

- Possible Explanatory Set finding and Boolean Rule Finding
  Since all initial values for the variable ABA are 1 in the original SIGNET dataset, each node which has a single parent ABA will have a constant value, and then the node is determined incorrectly as a root node. Thus we assume that the five root nodes are known for the dataset. For our dataset and Isabelle's dataset, we have different initial values of root nodes, and thus we need not make this assumption. To reduce the computational complexity, we limit the number of parents of each variable not more than 4. For each variable i that is not a root, we first find a minimum explanatory set V that satisfies $V \cap V_{t,t'} \neq \emptyset$ for all t and t', and then we find the Boolean rule of the parent set V for the variable i based on the maximum likelihood. If there is no such Boolean rules, then we try the next minimum explanatory set and repeat this process until finding a Boolean rule.

- Post Process
  If all minimum sets with not more than 4 nodes cannot explain the data, then we select a set V and a Boolean rule that contradict with the data set at the least.

Results:

Dataset/Task	Score
SIGNET/original dataset generated by Jenkins	0.14
SIGNET/dataset generated by Isabelle	0.05
SIGNET/my dataset	0.06

Table 3: Average error rates for three data sets.

- Quantitative advantages (e.g., compact feature subset, simplicity, computational advantages) The computational complexity is too high to utilize fully observed data from the dynamic process of an asynchronous network. But it is easy to implement our algorithm since our algorithm uses conditional likelihoods and it treats nodes one-by-one. See Section 5.2 for the complexity analysis.

- Given a parent set of a node, its Boolean rule is found based on the maximum likelihood. We use the conditional likelihood for a single node to reduce the computational complexity. When the network is not too complexity (i.e., the number of parent nodes for each node is less than 5), our method can run fast and it may be improved by using the full maximum likelihood of all nodes and using Pearson's χ^2 test or other model-selection scores to check whether a Boolean rule fits observed data well. Besides, our method takes full use of the information in the data, especially the information on the dynamic evolution processes. Using the information, we can discover many structures that cannot be found only from the information on steady states.

Now we briefly explain our implementation. First, we change the names of variables to be numerical style and transform the original data to a matrix. Then for each node, we use R program to process the preprocessed data to find its parent set and its truth table. Finally, we find a Boolean rule from the truth table. Contact us via email to ask for the code.

JMLR: Workshop and Conference Proceedings 6:249–256, 2010 NIPS 2008 workshop on causality

TIED: An Artificially Simulated Dataset with Multiple Markov Boundaries

Alexander Statnikov ALEXANDER.STATNIKOV@VANDERBILT.EDU
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

Constantin F. Aliferis CONSTANTIN.ALIFERIS@NYUMC.ORG
Center for Health Informatics and Bioinformatics
New York University
New York, NY 10016, USA

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We present an artificially simulated dataset (*TIED*) constructed so that there are many minimal sets of variables with maximal predictivity (i.e., Markov boundaries) and likewise many sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. This dataset was used in the *Potluck Causality Challenge* to determine all statistically indistinguishable sets of direct causes and direct effects and all Markov boundaries of the response variable and also to predict the response variable in the independent test data. We also present baseline results of application of several algorithms to this dataset.

Keywords: local causal discovery, Markov boundary induction, variable selection, classification

1. Introduction

The problem of variable/feature selection is of fundamental importance in machine learning and applied statistics, especially when it comes to analysis, modeling, and discovery from high-dimensional data (Guyon and Elisseeff, 2003; Kohavi and John, 1997). In addition to the promise of cost-effectiveness, two major goals of variable selection are to improve the prediction performance of the predictors and to provide a better understanding of the data-generative process (Guyon and Elisseeff, 2003). An emerging class of algorithms proposes a principled solution to the variable selection problem by identification of a Markov blanket of the response variable of interest (Aliferis et al., 2009, 2003; Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003). A *Markov blanket* is a set of variables conditioned on which all the remaining variables excluding the response variable are statistically independent of the response variable. A related concept is a *Markov boundary* (or non-redundant Markov blanket) that is a Markov blanket such that no proper subset of it is a Markov blanket (Pearl, 1988). Un-

der assumptions about the learner and loss function, a Markov boundary is the solution to the variable selection problem (Tsamardinos and Aliferis, 2003).

An important theoretical result states that if the distribution satisfies the *intersection property*[1], then it is guaranteed to have a unique Markov boundary of the response variable (Pearl, 1988). Furthermore, if the distribution satisfies common causal assumptions such as *faithfulness*, *Markov condition*, and *causal sufficiency*, then the Markov boundary is also unique and consists only of direct causes, direct effects, and direct causes of direct effects (also known as "spouses") of the response variable in the underlying causal graph (Tsamardinos and Aliferis, 2003). Even though there are several well-developed algorithms for learning a Markov boundary either in faithful distributions or in distributions where the intersection property holds (Aliferis et al., 2009; Peña et al., 2007; Aliferis et al., 2003; Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003), little research has been done in development of algorithms for learning *multiple* Markov boundaries from the same dataset when the above assumptions do not hold.

We present an artificially simulated dataset (*TIED*) that contains multiple Markov boundaries (and thus violates the intersection and faithfulness properties) and likewise many sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. This dataset was used in the *Potluck Causality Challenge* to determine all statistically undistinguishable sets of direct causes and direct effects and all Markov boundaries of the response variable and also to predict the response variable in the independent test data. We also present baseline results of application of several algorithms to this dataset.

2. Dataset

Using the principles from (Lemeire, 2006), we constructed a discrete Bayesian network *TIED* with 1,000 variables (including a response variable T). Figure 1 shows a fragment of the network structure and specifies which variables contain the same information about T by the color of highlighting. The parameterization of the network fragment shown in Figure 1 is provided in Table 1. The network fragment contains a response variable T, all variables that participate in all Markov boundaries of the response variable T, and some other variables. The full network can be obtained by adding 10 children to each variable from the set $\{X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{18}, X_{19}, X_{20}\}$ (a total of 110 variables) with conditional probability distribution defined in Table 2 and 860 variables that do not have a path to T in the network. If variables X and Y are shown with the same color in Figure 1, then (a) for every combination of values of X and T such that $P(T = t|X = x) = p$, there exists a value y of variable Y such that $P(T = t|Y = y) = p$, and (b) for every combination of values of Y and T such that $P(T = t|Y = y) = p$, there exists a value x of variable X such that $P(T = t|X = x) = p$. Such variables are interchangeable for prediction of T, and therefore if X belongs to a

1. We use notation $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$ to denote that subset of variables $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ in the underlying probability distribution. Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{W}$ be any four disjoint subsets of variables. Then the probability distribution satisfies the intersection property if $\mathbf{X} \perp \mathbf{Y}|(\mathbf{Z} \cup \mathbf{W})$ and $\mathbf{X} \perp \mathbf{W}|(\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp (Y \cup \mathbf{W})|\mathbf{Z}$.

Markov boundary $\boldsymbol{M}_1$ of T, then $\boldsymbol{M}_2 = (\boldsymbol{M}_1 \setminus \{X\}) \cup \{Y\}$ is another Markov boundary of T. The work of Lemeire (2006) specifically describes why such variables violate the intersection property of the probability distribution. In summary, the network contains 72 Markov boundaries of T. Each of these Markov boundaries contains 5 variables: (i) X_9, (ii) X_4 or X_8, (iii) X_{11} or X_{12} or X_{13}, (iv) X_{18} or X_{19} or X_{20}, and (v) X_1 or X_2 or X_3 or X_{10}. Similarly, there are 72 sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of T. These sets of variables coincide with the Markov boundaries of T.

The dataset *TIED* was obtained by sampling 3,750 instances from the above Bayesian network. 750 (20%) instances were used for discovery of multiple Markov boundaries (or sets of variables that are statistically indistinguishable from the set of direct causes and direct effects) of T, and the remaining 3,000 (80%) instances were used for validation of classification performance of T. We also computed the optimal Bayes classification performance of T which is 0.9663 weighted accuracy[2].

Table 1: Parameterization of the *TIED* network for variables shown in Figure 1 $\{T, X_1, X_2, X_3, X_4, \ldots, X_{29}\}$. Only nonzero probabilities are shown in the table.

T:	$P(T=0\|X_{10}=0)=1.0$	X_5:	$P(X_5=0\|X_4=0)=0.6$	X_{10}:	$P(X_{10}=0\|X_3=0)=1.0$
	$P(T=0\|X_{10}=1)=1.0$		$P(X_5=1\|X_4=0)=0.2$		$P(X_{10}=0\|X_3=1)=1.0$
	$P(T=0\|X_{10}=2)=1.0$		$P(X_5=2\|X_4=0)=0.2$		$P(X_{10}=1\|X_3=2)=0.3$
	$P(T=1\|X_{10}=3)=0.3$		$P(X_5=0\|X_4=1)=0.5$		$P(X_{10}=2\|X_3=2)=0.7$
	$P(T=2\|X_{10}=3)=0.3$		$P(X_5=1\|X_4=1)=0.25$		$P(X_{10}=3\|X_3=3)=1.0$
	$P(T=3\|X_{10}=3)=0.4$		$P(X_5=2\|X_4=1)=0.25$		
			$P(X_5=0\|X_4=2)=0.8$		
			$P(X_5=1\|X_4=2)=0.1$		
			$P(X_5=2\|X_4=2)=0.1$		
X_1:	$P(X_1=0)=0.25$	X_6:	$P(X_6=1\|X_4=0)=0.5$	X_{11}:	$P(X_{11}=0\|T=0)=1.0$
	$P(X_1=1)=0.25$		$P(X_6=2\|X_4=0)=0.5$		$P(X_{11}=0\|T=1)=1.0$
	$P(X_1=2)=0.25$		$P(X_6=0\|X_4=1)=0.8$		$P(X_{11}=0\|T=2)=1.0$
	$P(X_1=3)=0.25$		$P(X_6=1\|X_4=1)=0.2$		$P(X_{11}=1\|T=3)=0.5$
			$P(X_6=0\|X_4=2)=0.2$		$P(X_{11}=2\|T=3)=0.5$
			$P(X_6=1\|X_4=2)=0.3$		
			$P(X_6=2\|X_4=2)=0.5$		

(continued on the next pa

2. Weighted accuracy is defined as the average proportion of correct classifications in each category/class of the response variable.

Table 1: continued from previous page

X_2: $P(X_2=0\|X_1=0)=0.8$ $P(X_2=1\|X_1=0)=0.2$ $P(X_2=0\|X_1=1)=0.1$ $P(X_2=1\|X_1=1)=0.9$ $P(X_2=2\|X_1=2)=1.0$ $P(X_2=3\|X_1=3)=1.0$	X_7: $P(X_7=0\|X_4=0)=0.9$ $P(X_7=1\|X_4=0)=0.1$ $P(X_7=0\|X_4=1)=0.7$ $P(X_7=1\|X_4=1)=0.2$ $P(X_7=2\|X_4=1)=0.1$ $P(X_7=0\|X_4=2)=0.6$ $P(X_7=1\|X_4=2)=0.3$ $P(X_7=2\|X_4=2)=0.1$	X_{12}: $P(X_{12}=0\|X_{11}=0)=1.0$ $P(X_{12}=1\|X_{11}=1)=0.5$ $P(X_{12}=2\|X_{11}=1)=0.5$ $P(X_{12}=1\|X_{11}=2)=0.5$ $P(X_{12}=2\|X_{11}=2)=0.5$
X_3: $P(X_3=0\|X_2=0)=0.3$ $P(X_3=1\|X_2=0)=0.7$ $P(X_3=0\|X_2=1)=0.8$ $P(X_3=1\|X_2=1)=0.2$ $P(X_3=2\|X_2=2)=1.0$ $P(X_3=3\|X_2=3)=1.0$	X_8: $P(X_8=1\|X_4=0)=1.0$ $P(X_8=2\|X_4=1)=1.0$ $P(X_8=0\|X_4=2)=1.0$	X_{13}: $P(X_{13}=0\|X_{12}=0)=1.0$ $P(X_{13}=1\|X_{12}=1)=0.5$ $P(X_{13}=2\|X_{12}=1)=0.5$ $P(X_{13}=1\|X_{12}=2)=0.5$ $P(X_{13}=2\|X_{12}=2)=0.5$
X_4: $P(X_4=1\|T=0)=0.9$ $P(X_4=2\|T=0)=0.1$ $P(X_4=0\|T=1)=0.8$ $P(X_4=1\|T=1)=0.1$ $P(X_4=2\|T=1)=0.1$ $P(X_4=0\|T=2)=0.1$ $P(X_4=1\|T=2)=0.8$ $P(X_4=2\|T=2)=0.1$ $P(X_4=0\|T=3)=0.1$ $P(X_4=1\|T=3)=0.1$ $P(X_4=2\|T=3)=0.8$	X_9: $P(X_9=0\|T=0)=0.1$ $P(X_9=1\|T=0)=0.8$ $P(X_9=2\|T=0)=0.1$ $P(X_9=1\|T=1)=0.1$ $P(X_9=2\|T=1)=0.9$ $P(X_9=0\|T=2)=0.1$ $P(X_9=1\|T=2)=0.8$ $P(X_9=2\|T=2)=0.1$ $P(X_9=0\|T=3)=0.2$ $P(X_9=1\|T=3)=0.7$ $P(X_9=2\|T=3)=0.1$	X_{14}: $P(X_{14}=0\|X_1=0)=0.8$ $P(X_{14}=1\|X_1=0)=0.1$ $P(X_{14}=2\|X_1=0)=0.1$ $P(X_{14}=0\|X_1=1)=0.1$ $P(X_{14}=1\|X_1=1)=0.8$ $P(X_{14}=2\|X_1=1)=0.1$ $P(X_{14}=0\|X_1=2)=0.8$ $P(X_{14}=1\|X_1=2)=0.1$ $P(X_{14}=2\|X_1=2)=0.1$ $P(X_{14}=0\|X_1=3)=0.1$ $P(X_{14}=1\|X_1=3)=0.1$ $P(X_{14}=2\|X_1=3)=0.8$
X_{15}: $P(X_{15}=0\|X_{14}=0)=1.0$ $P(X_{15}=0\|X_{14}=1)=1.0$ $P(X_{15}=1\|X_{14}=2)=0.5$ $P(X_{15}=2\|X_{14}=2)=0.5$	X_{20}: $P(X_{20}=0\|X_{19}=0)=1.0$ $P(X_{20}=1\|X_{19}=1)=1.0$ $P(X_{20}=2\|X_{19}=2)=1.0$	X_{25}: $P(X_{25}=0)=0.5$ $P(X_{25}=1)=0.5$

(continued on the next page)

Table 1: continued from previous page

X_{16}: $P(X_{16}=0\|X_1=0)=0.2$ $P(X_{16}=1\|X_1=0)=0.6$ $P(X_{16}=2\|X_1=0)=0.2$ $P(X_{16}=0\|X_1=1)=0.1$ $P(X_{16}=1\|X_1=1)=0.3$ $P(X_{16}=2\|X_1=1)=0.6$ $P(X_{16}=0\|X_1=2)=0.5$ $P(X_{16}=1\|X_1=2)=0.1$ $P(X_{16}=2\|X_1=2)=0.4$ $P(X_{16}=0\|X_1=3)=0.3$ $P(X_{16}=1\|X_1=3)=0.5$ $P(X_{16}=2\|X_1=3)=0.2$	X_{21}: $P(X_{21}=0\|X_5=0)=0.2$ $P(X_{21}=1\|X_5=0)=0.6$ $P(X_{21}=2\|X_5=0)=0.2$ $P(X_{21}=0\|X_5=1)=0.1$ $P(X_{21}=1\|X_5=1)=0.3$ $P(X_{21}=2\|X_5=1)=0.6$ $P(X_{21}=0\|X_5=2)=0.5$ $P(X_{21}=1\|X_5=2)=0.1$ $P(X_{21}=2\|X_5=2)=0.4$	X_{26}: $P(X_{26}=0\|X_{25}=0)=0.1$ $P(X_{26}=1\|X_{25}=0)=0.9$ $P(X_{26}=0\|X_{25}=1)=0.3$ $P(X_{26}=1\|X_{25}=1)=0.7$
X_{17}: $P(X_{17}=0)=0.25$ $P(X_{17}=1)=0.25$ $P(X_{17}=2)=0.25$ $P(X_{17}=3)=0.25$	X_{22}: $P(X_{22}=0\|X_6=0)=0.3$ $P(X_{22}=1\|X_6=0)=0.2$ $P(X_{22}=2\|X_6=0)=0.5$ $P(X_{22}=0\|X_6=1)=0.8$ $P(X_{22}=1\|X_6=1)=0.1$ $P(X_{22}=2\|X_6=1)=0.1$ $P(X_{22}=0\|X_6=2)=0.6$ $P(X_{22}=1\|X_6=2)=0.2$ $P(X_{22}=2\|X_6=2)=0.2$	X_{27}: $P(X_{27}=0\|X_{25}=0)=0.4$ $P(X_{27}=1\|X_{25}=0)=0.6$ $P(X_{27}=0\|X_{25}=1)=0.8$ $P(X_{27}=1\|X_{25}=1)=0.2$
X_{18}: $P(X_{18}=1\|T=0)=0.1$ $P(X_{18}=2\|T=0)=0.9$ $P(X_{18}=0\|T=1)=0.1$ $P(X_{18}=2\|T=1)=0.9$ $P(X_{18}=0\|T=2)=0.8$ $P(X_{18}=1\|T=2)=0.1$ $P(X_{18}=2\|T=2)=0.1$ $P(X_{18}=0\|T=3)=0.1$ $P(X_{18}=1\|T=3)=0.8$ $P(X_{18}=2\|T=3)=0.1$	X_{23}: $P(X_{23}=0\|X_7=0)=0.5$ $P(X_{23}=1\|X_7=0)=0.1$ $P(X_{23}=2\|X_7=0)=0.4$ $P(X_{23}=0\|X_7=1)=0.6$ $P(X_{23}=1\|X_7=1)=0.3$ $P(X_{23}=2\|X_7=1)=0.1$ $P(X_{23}=0\|X_7=2)=0.7$ $P(X_{23}=1\|X_7=2)=0.1$ $P(X_{23}=2\|X_7=2)=0.2$	X_{28}: $P(X_{28}=0)=0.33$ $P(X_{28}=1)=0.33$ $P(X_{28}=2)=0.33$
X_{19}: $P(X_{19}=1\|X_{18}=0)=1.0$ $P(X_{19}=2\|X_{18}=1)=1.0$ $P(X_{19}=0\|X_{18}=2)=1.0$	X_{24}: $P(X_{24}=0\|X_8=0)=0.8$ $P(X_{24}=1\|X_8=0)=0.1$ $P(X_{24}=2\|X_8=0)=0.1$ $P(X_{24}=0\|X_8=1)=0.6$ $P(X_{24}=1\|X_8=1)=0.2$ $P(X_{24}=2\|X_8=1)=0.2$ $P(X_{24}=0\|X_8=2)=0.5$ $P(X_{24}=1\|X_8=2)=0.3$ $P(X_{24}=2\|X_8=2)=0.2$	X_{29}: $P(X_{29}=0\|X_{15}=0)=1.0$ $P(X_{29}=1\|X_{15}=1)=0.5$ $P(X_{29}=2\|X_{15}=1)=0.5$ $P(X_{29}=1\|X_{15}=2)=0.5$ $P(X_{29}=2\|X_{15}=2)=0.5$

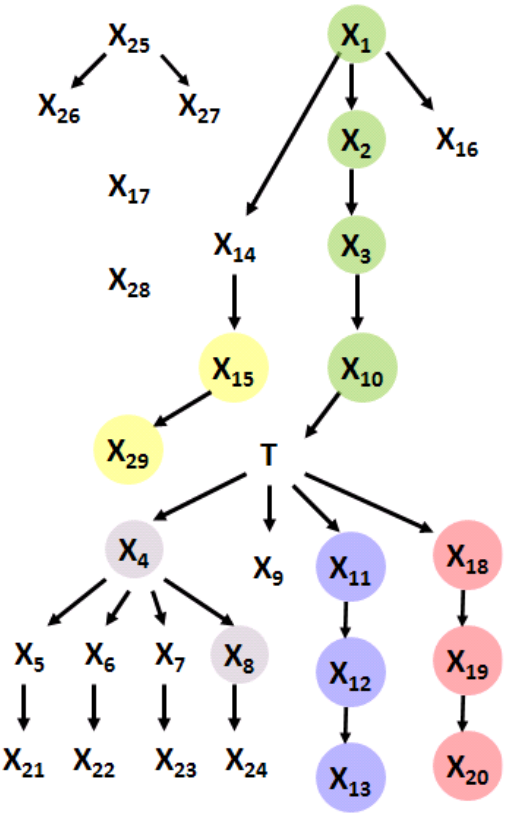


Figure 1: Graphical visualization of the fragment of a discrete Bayesian network *TIED*. Variables that contain exactly the same information about T are highlighted with the same color, e.g. variables X_{11}, X_{12}, and X_{13} provide exactly the same information about T and thus are interchangeable for prediction of T.

Table 2: Conditional probability distribution of each of 110 variables (denoted by Z) mentioned in Section 2 that have a single parent from the set $\{X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{18}, X_{19}, X_{20}\}$ (denoted by X).

$P(Z\|X)$	$X=0$	$X=1$	$X=2$
$Z=0$	0.3	0.4	0.3
$Z=1$	0.3	0.3	0.4
$Z=2$	0.4	0.3	0.3

3. Experiments and Results

The experiments involved running several algorithms for discovery of multiple Markov boundaries:

- Four resampling-based techniques that apply a variable selection algorithm to bootstrap samples from the original dataset: The following variable selection methods were used: (i) SVM-based recursive feature elimination (SVM-RFE) (Guyon et al., 2002); (ii) SVM-RFE with additional application of McNemar's test (Everitt, 1977) to identify the most parsimonious variable set with classification performance statistically indistinguishable from the observed best one; (iii) backward wrapping with linear SVM classifier based on univariate ranking of variables by Kruskal-Wallis non-parametric ANOVA (Hollander and Wolfe, 1999); and (iv) backward wrapping with linear SVM classifier based on Kruskal-Wallis ANOVA with additional statistical comparison step, as in (ii). The above four methods are denoted as *Resampling-SVM-RFE1*, *Resampling-SVM-RFE2*, *Resampling-Univariate1*, *Resampling-Univariate2*, respectively. Since there is no natural termination criterion of these methods, they were run on 5,000 bootstrap samples from the original dataset.

- Three instantiations of KIAMB algorithm (Peña et al., 2007): KIAMB was applied with G^2 test, parameter $K = 0.8$, and three statistical thresholds $\alpha = 0.01$, $\alpha = 0.005$, and $\alpha = 0.001$ (denoted as *KIAMB1*, *KIAMB2*, *KIAMB3*, respectively). The first threshold was used by inventors of the method in the paper that introduced it (Peña et al., 2007). Since there is no natural termination criterion of these methods, they were run 5,000 times.

- *Iterative Removal* method (Natsoulis et al., 2005): This method works as follows: First, it extracts a Markov boundary from the original dataset and estimates its classification performance. Second, it removes all variables from the original dataset that were found to participate in the Markov boundaries, extracts a new tentative Markov boundary from the modified dataset, and estimates its classification performance. Finally third, if the classification performance of the tentative Markov boundary is statistically indistinguishable from the Markov boundary obtained in the first step, then this is also a true Markov boundary and the second and third steps of the algorithm are repeated. The implementation of this method used an algorithm HITON-PC (Aliferis et al., 2009, 2003) to learn a Markov boundary and McNemar's test to compare linear SVM classification performance of resulting variable sets (Everitt, 1977).

All methods were applied to the 750-instance training dataset to indentify Markov boundaries of the response variable T. Once the Markov boundaries were indentified, a linear SVM classifier was trained with these variable sets in the training dataset and it was applied to the 3,000-instance validation dataset. The classification performance was measured by the weighted accuracy metric (Guyon et al., 2006). In independent

tests (not shown here) the choice of a linear SVM versus non-linear one was validated as not compromising classification performance.

The results of experiments are presented in Table 3. The following are observed: (i) *Iterative Removal* identifies only one Markov boundary because all other Markov boundaries have a common variable (X_9) and thus cannot be detected by this method. This is s structural deficiency of that method. (ii) KIAMB fails to identify any true Markov boundaries due to its sample inefficiency (its sample requirements are of exponential order to the number of variables in the Markov boundary), and because of the same reason its output Markov boundaries have poor predictivity; (iii) Resampling-based methods either miss many true Markov boundaries and/or output many false positive variables in the identified Markov boundaries.

Table 3: Results of experiments with artificial dataset *TIED*. All experiments were executed on a cluster with Intel 2.4 GHz Xeon CPU's.

Method	Total number of output Markov boundaries	Number of variables in an average output Markov boundary	Number of true Markov boundaries		Average number of false positive variables in identified true Markov boundaries	Average classification performance in validation data	CPU time in minutes
			identified exactly	identified with false positive variables			
Iterative Removal	3	5.67	0	1	2.00	0.959	0.04
KIAMB1	5000	2.82	0	0	—	0.798	285.42
KIAMB2	5000	2.81	0	0	—	0.796	285.45
KIAMB3	5000	2.80	0	0	—	0.796	285.48
Resampling + Univariate1	5000	11.10	0	72	12.29	0.942	5999.64
Resampling + Univariate2	5000	5.58	0	0	—	0.934	6000.41
Resampling + RFE1	5000	8.70	0	72	6.38	0.952	6235.28
Resampling + RFE2	5000	4.24	0	29	5.76	0.947	6235.93

4. Conclusion

This report introduced an artificially simulated dataset (*TIED*) with multiple Markov boundaries and multiple sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. We also presented baseline results of several algorithms in this dataset. The results demonstrate that *TIED*

is a challenging problem and many methods fail to discover multiple Markov boundaries from this dataset. Therefore, there is a need to create new algorithms to identify multiple Markov boundaries.

References

C. F. Aliferis, I. Tsamardinos, and A Statnikov. HITON: a novel markov blanket algorithm for optimal variable selection. In *AMIA 2003 Annual Symposium Proceedings*, pages 21–25, 2003.

C. F. Aliferis et al. 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*, 2009. (to appear).

B. Everitt. *The analysis of contingency tables*. Chapman and Hall, London, 1977.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

I. Guyon et al. *Feature extraction: foundations and applications*. Springer-Verlag, Berlin, 2006.

M. Hollander and D. Wolfe. *Nonparametric statistical methods*. Wiley, New York, NY, USA, 1999.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

J. Lemeire. The representation and learning of equivalent information in causal models. Technical report, 2006. Technical Report IRIS-TR-0099.

G. Natsoulis et al. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, 15:724–736, 2005.

J. Peña et al. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232, 2007.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, California, 1988.

I. Tsamardinos and C. F. Aliferis. Towards principled feature selection: relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AI & Stats)*, 2003.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 673–678, 2003.

JMLR: Workshop and Conference Proceedings 6:257–266, 2010 NIPS 2008 Workshop on Causality

Learning Causal Models That Make Correct Manipulation Predictions With Time Series Data

Mark Voortman VOORTMAN@SIS.PITT.EDU
Decision Systems Laboratory
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA, 15260, USA

Denver Dash DENVER.H.DASH@INTEL.COM
Intel Research
Pittsburgh, PA, 15213, USA

Marek J. Druzdzel MAREK@SIS.PITT.EDU
Decision Systems Laboratory
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA, 15260, USA

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf.

Abstract

One of the fundamental purposes of causal models is using them to predict the effects of manipulating various components of a system. It has been argued by Dash (2005, 2003) that the *Do* operator will fail when applied to an equilibrium model, unless the underlying dynamic system obeys what he calls *Equilibration-Manipulation Commutability*. Unfortunately, this fact renders most existing causal discovery algorithms unreliable for reasoning about manipulations. Motivated by this caveat, in this paper we present a novel approach to causal discovery of dynamic models from time series. The approach uses a representation of dynamic causal models motivated by Iwasaki and Simon (1994), which asserts that all "causation across time" occurs because a variable's derivative has been affected instantaneously. We present an algorithm that exploits this representation within a constraint-based learning framework by numerically calculating derivatives and learning instantaneous relationships. We argue that due to numerical errors in higher order derivatives, care must be taken when learning causal structure, but we show that the Iwasaki-Simon representation reduces the search space considerably, allowing us to forego calculating many high-order derivatives. In order for our algorithm to discover the dynamic model, it is necessary that the time-scale of the data is much finer than any temporal process of the system. Finally, we show that our approach can correctly recover the structure of a fairly complex dynamic system, and can predict the effect of manipulations accurately when a manipulation does not cause an instability. To our knowledge, this is the first causal discovery algorithm that has demonstrated that it can correctly predict the effects of manipulations for a system that does not obey the EMC condition.

Keywords: Causal discovery, dynamic systems, manipulations.

1. Introduction

One of the fundamental purposes of causal models is the prediction of the effects of manipulating various components of a system. It has been argued by Dash (2005, 2003) that the *Do* operator will fail when applied to an equilibrium model unless the underlying dynamic system obeys what he calls *Equilibration-Manipulation Commutability (EMC)*, a principle which is illustrated by the graph in Figure 1. In this figure, a dynamic system S, represented by a set of differential equations, is depicted on the upper-left. S has one or more equilibrium points such that, under the initial exogenous conditions, the equilibrium model $\tilde{S}$, represented by a set of equilibrium equations, will be obtained after sufficient time has passed. There are thus two approaches for making predictions of manipulations on S on time-scales sufficiently long for the equilibrations to occur. One could start with $\tilde{S}$ and apply the *Do* operator to predict manipulations. This is path A in Figure 1, and is the approach taken whenever a causal model is built from data drawn from a system in equilibrium. Alternatively, in path B the manipulations are performed on the original dynamic system which is then allowed to equilibrate; this is the path that the actual system takes. The EMC property is satisfied if and only if path A and path B lead to the same causal structure.

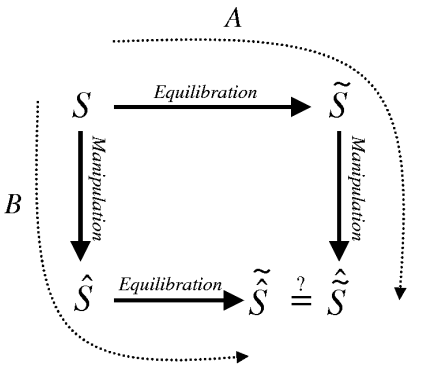


Figure 1: Equilibration-Manipulation Commutability provides sufficient conditions for an equilibrium causal graph to correctly predict the effect of manipulations.

As an example of a system that obeys the EMC condition, consider a body of mass m dangling from a damped spring. The mass will stretch the spring to some equilibrium position $x = mg/k$ where k is the spring constant. As we vary m and allow the system to come to equilibrium, the value of x gets affected according to this relation. The equilibrium causal model $\tilde{S}$ of this system is simply $m \rightarrow x$. If one were to manipulate the spring directly and stretch it to some displacement $x = \hat{x}$, then the mass would be independent of the displacement, and the correct causal model is obtained by applying the *Do* operator to this equilibrium model.

Alternatively, one could have started with the original system S of differential equations of the damped simple-harmonic oscillator by explicitely modeling the acceleration

$a = mg - kx - \alpha v$, where α is the dampening constant, and the velocity v. S can likewise be used to model the manipulation of x by applying the *Do* operator to a, v, and x simultaneously, ultimately giving the same structure as was obtained by starting with the equilibrium model. For examples of systems that do not obey the EMC condition, we refer the reader to Dash (2005, 2003) and the model shown later in this paper.

Unfortunately, requiring a system to obey the EMC condition renders most existing causal discovery algorithms unreliable for reasoning about manipulations, unless the details of the underlying dynamics of the system are explicitly represented in the model. Most classical causal discovery algorithms in AI make use of the class of independence constraints found in the data to infer causality between variables, assuming the faithfulness assumption (e.g., Spirtes et al., 2000; Pearl and Verma, 1991; Cooper and Herskovits, 1992). These methods will not be guaranteed to obey EMC if the observation time-scale of the data is long enough for some process in the underlying dynamic system to go through equilibrium. On the other hand, there have been previous approaches for learning dynamic causal models and Bayesian networks. Friedman et al. (1998) learn the structure of first-order Markov model by using time series data, and it would be straightforward to extend these approaches to higher-order Markovian models. However, the search space rapidly gets very large when searching for arbitrary dependencies across time.

Our approach, by contrast, uses an alternative representation of a dynamic system, explicitely modeling derivatives (or differences) of variables. It is beyond the scope of this paper to perform a quantitative comparison of prediction to these other approaches, however, we argue here that the representation that we learn helps us constrain the search space, and we expect that this reduction in complexity will make our algorithm perform better in practice and be more efficient than methods that try to learn fixed-order Markov structures for all variables.

2. Representation and Assumptions

Our approach uses a representation of dynamic causal models inspired by Iwasaki and Simon (1994), which asserts that all "causation across time" occurs because a variable's derivative has been affected instantaneously. Iwasaki and Simon called these models "mixed causal structures". We use a slightly modified version of them and we call them "differential-based dynamic causal models" (DBD causal models, for short).

We use the notation $X^{(n)}$ to denote the n-th order derivative (or discrete version thereof) of variable X, and we use the convention that $X^{(0)} = X$.

Definition 1 (DBD graphs) *Differential-based dynamic causal graphs over a set of time-dependent variables X are discrete-time directed acyclic causal graphs, in which all "change across time" of a variable X occurs because there exists some n such that* $X^{(n)}$ *is being caused contemporaneously. That is, an edge exists from variable* $Y_t \rightarrow X_{t+1}$ *only if* $Y_t = X_t^{(1)}$ *or* $Y_t = X_t$*, in which case the parent set of* X_{t+1} *is* $\{X_t, X_t^{(1)}\}$.

The reason we constrain the parent set of X_{t+1} to be $\{X_t, X_t^{(1)}\}$ when X_t^1 is determined, is simply that, by definition,

$$X_{t+1} = X_t + X_t^{(1)} dt.$$

DBD models are unique in that they focus on uncovering *contemporaneous* causal relations that impact *derivatives* of some variables. They are motivated by real physical systems based on classical mechanics. For example, systems governed by Newton's 2nd Law are archetypical causal systems: some "force" acts on a body, "causing" it to accelerate. The acceleration of the object, in turn, causes it to change velocity, which can cause the object to change position. The DBD reprentation assumes that all causation can be described in terms of "forces" causing a variable to change by impacting a derivative of some order instantaneously.

We show an example DBD graph in Figure 2. We will also use this example to illustrate the algorithm in one of the next sections. In the graph, two kinds of arcs

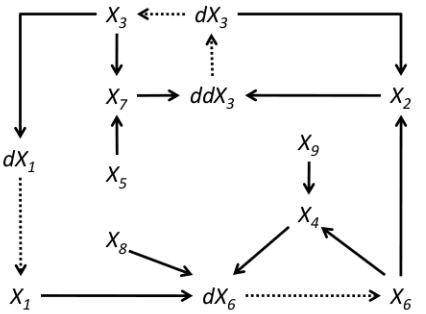


Figure 2: The DBD graph we used to simulate data.

are used: *solid arcs* that denote instantaneous causation, and *dashed arcs* that denote causation across time. The dashed arcs were called *integration links* by Iwasaki and Simon (1994) because they always point from a derivative of order n to a derivative of order $n-1$. The variables that have derivatives in the model are called *dynamic*, as they are solely responsible for the dynamics of the system. Assuming this representation, the learning algorithm has to find out which variables are dynamic and find the instanteneous causal arcs between variables and derivatives. It is important to note that our algorithm only learns contemporaneous causality, all dynamic behavior is then determined by integration over time.

In a dynamic structure, different causal equilibrium models may exist over different time-scales. Which equilibrium models will be obtained over time are determined by the time-scales at which variables equilibrate. The causal structures are derived from the equations by applying the causal ordering algorithm (Iwasaki and Simon, 1994) and by assuming that at fast time-scales, the slower moving variables are relatively constant. In the example of Figure 2, the time-scales could be such that $\tau_6 \ll \tau_3 \ll \tau_1$, where τ_i is the time-scale of variable X_i, in which case, at time $t \sim \tau_6$ it would be safe to assume that X_3 and X_1 are approximately constant. Under these time-scale assumptions, Figure 3 shows the different (approximate) models that exist for the graph in Figure 2.

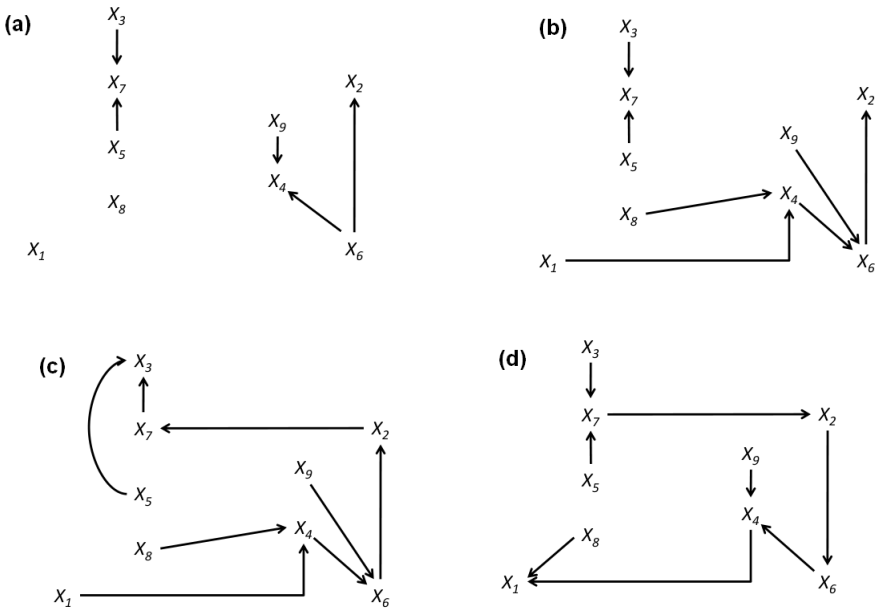


Figure 3: The different equilibrium models that exist in the sytem over time. (a) The independence constraints that hold when $t \sim 0$. (b) The independence constraints when $t \sim \tau_6$. (c) The independence constraints when $t \sim \tau_3$. (d) The independence constraints after all the variables are equilibrated, $t \gtrsim \tau_1$.

One obvious approach to learning the graph of Figure 2 (assuming no derivative variables are present in the data), is to try to learn an arbitrary-order dynamic Bayesian network, for example using the method of Friedman et al. (1998). Figure 4 shows the second order Markov graph that a perfect DBN oracle would produce for this system. The problem with learning an arbitrary Markov model to represent this dynamic system is that there are no constraints as to which variables may affect other variables across time, so in principle, the search space could be unneccessarily large. The DBD representation, on the other hand, implies specific rules for when variables can affect other variables in the future (when they instantaneously effect some derivative of the variable). Given that a derivative $X^{(n)}$ is being instantaneously caused, DBDs also provide constraints on what variables can effect all $X^{(i)}$ for $i \neq n$.

We now state three conjectures concerning DBD models that are useful in explicating these constraints. Conjecture 2 is used to constrain the search space by limiting the number of possible dynamic variables. Conjecture 3 states that only one of the derivatives of a variable, or the variable itself, can have an incoming arc. Conjecture 4 is used to direct additional edges that are not oriented by the regular PC algorithm: If one of the derivatives of a variable has an incoming edge and the variable itself has an undirected

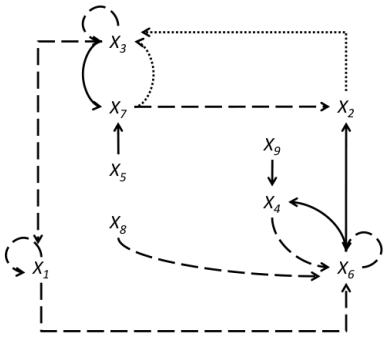


Figure 4: The second order Markov graph of the system. Thick dashed lines represent first-order Markov relations. Thin dotted lines represent second-order relations.

edge, then the edge of the variable must be outgoing. This is necessary, otherwise it would conflict with Conjecture 3.

Conjecture 2 *Every non-exogenous root node that is present in the independence structure at time $t = 0$ is a dynamic variable.*

Conjecture 3 *Let A, B and C be different variables in a DBD model, and $\dot{A}$ any order derivative of A. If the model contains an arc $A \leftarrow B$, then it does not contain the arc $\dot{A} \leftarrow C$.*

We use the standard notation $X - Y$ to indicate that either $X \rightarrow Y$ or $Y \rightarrow X$.

Conjecture 4 *Let A, B and C be different variables in a DBD model, and $\dot{A}$ any order derivative of A. If $A - B$ and $\dot{A} - C$, then the edge $A - B$ must be oriented $A \rightarrow B$.*

Our algorithm is based on the PC algorithm (Spirtes et al., 2000), although we could have used any other causal discovery algorithm as well. Besides the assumptions required for the PC algorithm, we make several additional assumptions. First, we assume that the system is stable, i.e., every dynamic variable must be part of a feedback loop. This implies that the highest order derivative of each dynamic variable must have at least one incoming arc. Second, exogenous variables are held constant over time and thus easily detectable in a data set. Third, in order for our algorithm to discover the dynamic model, it is necessary that the time-scale of the data is much finer than any temporal process of the system. This ensures that we are learning the dynamic model and not an equilibrium model. Finally, we assume that, apart from variable derivatives, the system is *causally sufficient* (i.e., there are no latent common causes).

3. The Algorithm

We present an algorithm that exploits the DBD representation within a constraint-based learning framework. The aim is to learn DBD models like the one given in Figure 2 directly. The input data[1] consisted of multiple time series that were generated first by parametrizing the model of Figure 2 with linear equations with independent Gaussian error terms, then by choosing different initial conditions for exogenous and dynamic variables and simulating 10000 discrete time steps. The integral equations have no noise, because they involve a deterministic relationship.

All derivatives of variables have been omitted from the data; thus, part of the challenge was that our method had to infer from the data which variables were changing due to the presence of derivatives and which were changing due to contemporaneous causation. Since calculating higher order derivatives using differences is sensitive to numerical errors, we opt for an incremental approach that gradually adds derivatives to the data set only when necessary, and exploits constraints given by our conjectures about feasible structures in these DBD graphs.

Our algorithm can be described in a few sentences: First we start with the original variables given in the data set and try to learn the instantaneous independence structure S_0^0 beteween non-derivative variables. This structure (plus our conjectures above) constrain which variables may be affected by derivatives. There may be multiple possible sets $S_0^1, S_1^1, \ldots, S_m^1$ of variables that could be consistent with S_0^0 and our conjectures. We then try to learn additional structure S_j^{i+1} with these new sets of variables assuming all links in S_j^i are correct. We recursively traverse the tree until we reach a set of maximum-order derivative models S_j^n, where n is an input into the algorithm. In instances where the structure from S_j^{i+1} contradicts structure from S_j^i, we assume S_j^i is correct. The output of the algorithm is then the complete set of consistent n-th order graphs.

To illustrate the algorithm, we will use the example model from Figure 2. To find out which variables are dynamic, we run the PC algorithm on a data set containing only non-derivative variables. The resulting structure will be the graph in Figure 3-a. The following four disconnected graphs will be discovered, and using Conjecture 2 we can find which variables are dynamic:

- X_1; this variable has to be dynamic, because it is not exogenous.
- X_8; this variable is exogenous and, therefore, not dynamic.
- $X_5 \to X_7 \leftarrow X_3$; X_5 is exogenous, X_7 is instanteneously caused and not dynamic, X_3 is not exogenous and, therefore, dynamic.
- $X_2 - X_6 \to X_4 \leftarrow X_9$; either X_2 or X_6 is dynamic, X_4 is not, and X_9 is exogenous.

Summarizing, X_1, X_3, and either X_2 or X_6 are dynamic variables so there are only two competing models.

1. Downloadable from http://www.causality.inf.ethz.ch/repository.php?id=16

In the second step, for each of the competing models, the first order derivatives of the dynamic variables are added to the data set and the PC algorithm is executed again. The competing model in which X_2 is a dynamic variable will lead to an inconsistent structure, because there will be a v-structure into X_2, namely $dX_3 \rightarrow X_2 \leftarrow X_6$. This violates Conjecture 3 and so the structure is inconsistent. The other competing model is consistent, although no derivative of X_3 has an incoming edge. Therefore, as the last step, we add the second derivative of X_3 to the data set and run PC again to retrieve the original structure. We used Conjecture 4 as an extra rule to orient edges.

4. Prediction of Manipulations

The following results[2] were obtained by using the data and applying the instructions described in Appendix A. After running our algorithm on the data to obtain a causal structure, we estimated the coefficients in the equations in order to be able to make quantitative predictions. In the next step, we used the model and the values of the first four time steps in the data set to make predictions for time steps $\{5, 50, 100, 500, 1000, 2000, 4000, 10000\}$. We do not attempt to correct our predictions by using the data at times $t > 4$ when predicting later times, although doing this is possible and should improve our results.

The results are shown in Figure 5. Due to space constraints we chose not to present six tables, but instead calculated the average RMSE per time step for each manipulated variable. The graph shows that the error for the first few time steps is relatively small, but for all variables (except X_1) grows large in later times. Three variables in particular (X_2, X_7 and X_4) had astronomical errors in later times. These huge RMS errors are not indicative that our model was poor. In fact, in our case, since we generated the model, we could verify that the structure was exactly correct and the linear Gaussian parameters were very well identified. The reason for the unstable errors is that in the model of Figure 2, manipulating any variable except X_1 will approximately break the feedback loop of a dynamic variable and thus will in general result in an instability (Dash, 2003). Feedback variable X_1 is a relatively slow process, so breaking this feedback loop does not have a large effect on the feedback loops of X_3 and X_6. Thus our absolute rms error is expected to also be unstable all manipulations but X_1, simply because we are predicting such large values.

More important than getting the correct RMS error for these manipulations is the fact that our learned model correctly predicts that an instability will occur when any variable except X_1 is manipulated. In the absence of instability, our method has very low RMS error, as indicated by the curve of variable X_1 in Figure 5. This fact is significant, because the model retrieved from our system when variable X_1 is allowed to come to equilibrium will not obey the EMC condition (Dash, 2005). Thus, to our knowledge, we have presented the first algorithm that has demonstrated that it can correctly predict the effects of manipulations on systems that do not obey this condition.

2. Results can be downloaded from http://pittsburgh.intel-research.net/~dhdash/causalitydata/rmse.zip.

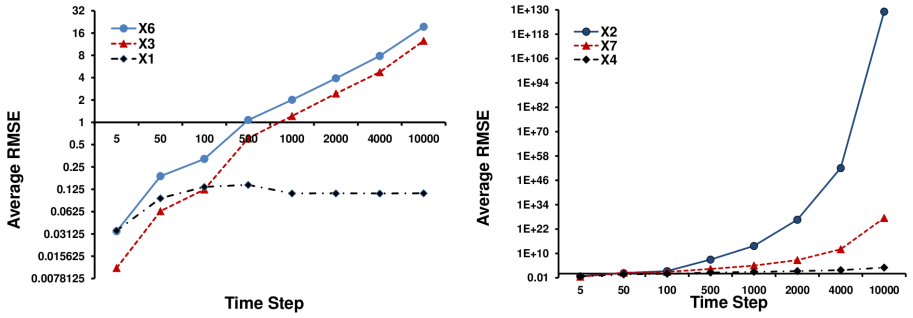


Figure 5: Average RMSE for each manipulated variable.

5. Conclusions

We have described a first effort to construct an algorithm that can predict the effects of manipulations on systems that do not obey the EMC condition. We accomplish this by learning dynamic causal graphs in a representation very similar to that of Iwasaki and Simon. We have proposed a set of conjectures which are effective at constraining the search space for high-order Markovian relationships, which we expect will make this method more reliable and more efficient than other methods for learning temporal models, especially when higher-order relationships are present. We have shown that on a benchmark dataset generated from a fairly sophisticated dynamic system having multiple inter-related processes operating at widely varying time-scales, we were able to correctly learn the structure of the underlying system, and were able to predict that manipulating some variables in that system would result in an instability. Finally, in the absence of instabilities, we were able to predict with high accuracy the results of manipulating a variable, even far into the future. Future work will involve performing quantitative comparisons to other time-series methods. Also, although our conjectures formed useful heuristics for this method, we have been able to construct counter-examples where at least one of them is incorrect, so more work is needed to prove our existing conjectures and finding additional constraints on the search for derivatives.

Acknowledgments

This research was supported by the Air Force Office of Scientific Research grant FA9550–06–1–0243 and by Intel Research. All experimental data have been obtained using SMILE, a Bayesian inference engine developed at the Decision Systems Laboratory and available at http://genie.sis.pitt.edu/.

References

Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

Denver Dash. *Caveats for Causal Reasoning*. PhD thesis, Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, April 2003. http://etd.library.pitt.edu/ETD/available/etd-05072003-102145/.

Denver Dash. Restructuring dynamic causal systems in equilibrium. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AIStats 2005)*, pages 81–88. Society for Artificial Intelligence and Statistics, 2005. (Available electronically at http://www.gatsby.ucl.ac.uk/aistats/).

Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 139–147. Morgan Kaufmann, 1998.

Yumi Iwasaki and Herbert A. Simon. Causality and model abstraction. *Artificial Intelligence*, 67(1):143–194, May 1994.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer Verlag, New York, NY, USA, second edition, 2000.

Appendix A. Pot-luck challenge: FACT SHEET .

Repository URL: http://www.causality.inf.ethz.ch/repository.php?id=16

Title: Mixed Dynamic Systems
Authors: Denver Dash, Mark Voortman, Marek Druzdzel
Contact name, address, email and website: Denver Dash, denver.h.dash@intel.com, http://pittsburgh.intel-research.net/~dhdash

Key facts:
Simulated time series data of 9 variables based on linear Gaussian models with no latent common causes, but with multiple dynamic processes at varying time-scales.
Training Data: 9 Variables, 10000 time series, each time series sampled at 12 distinct times (relative to when exogenous variables were first manipulated).
Testing Data: Manipulation data. Each of the 6 non-exogenous variables is manipulated and held fixed for the duration of the time series. This is repeated 100 times for each of the 6 variables.

Summary: A Mixed Dynamic System is one that consists of multiple dynamic processes operating at widely different time-scales. This data represents a 9 variable (labeled $X1 \ldots X9$) dynamic system with several dynamic processes acting on qualitatively different time scales from one another. The goal is to learn a causal model of the system with the training data, and then correctly predict the effects of various manipulations on the system (using the testing data for a quantitative measure of performance). This dataset was meant to be both simple and extremely challenging. All relations are linear with independent Gaussian error terms. There are no hidden confounders. However, we believe the inter-related dynamic processes will make prediction of manipulations challenging.

Training Data: The training data consists of 9 tab-separated text files (labelled X1.tsv, X2.tsv, etc.) one for each variable, and is arranged so that the rows in each file represent distinct time series for each variable (there are 10000 of these). That time series has been sampled at a few points in time after the exogenous variables of the system have been manipulated (all exogenous variables are held fixed for the duration of the time series). Specifically, the variables have been measured at the following discrete time intervals: t = {1, 2, 3, 4, 5, 50, 100, 500, 1000, 2000, 4000, 10000}, so there are 12 columns in each data file. Variables $X8$, $X5$ and $X9$ are all exogenous as can be verified by looking at X9.tsv, etc.

Test Data: The test data is organized into several (6x9 = 54) data files labeled Xi-manipj.tsv (For example X2-manip3.tsv shows the values of variable $X2$ when $X3$ has been manipulated and held fixed). Each variable in the set of endogenous variables {X1, X2, X3, X4, X6, X7} is manipulated 100 times for the entire 10000 time-step duration of each time series while the remaining variables are measured once at each of the 12 predetermined time-intervals. Thus each Xi-manipj.tsv file has 100 rows and 12 columns, and there are 9 files for each variable manipulated from the set {X1, X2, X3, X4, X6, X7}.

Evaluation: The objective of this problem is to use the first set of data labeled X*.tsv to build a model which is then able to predict the effects of manipulation on the system as given by the X*-manipN.tsv files. When predicting the effect of the manipulations, the goal is to predict the values of non-manipulated variables at times 5–10000 (columns 5 – 12) using the values of the previous times as input. For example, when predicting time 100 (column 7), you could use times 1, 2, 3, 4, 5, 50 (columns 1-6) as input. The output of the evaluation should be one

table for each variable in the set {X1, X2, X3, X4, X6, X7} of manipulated variables. Each table should have 5 rows and 8 columns, one row for each variable in {X1, X2, X3, X4, X6, X7} $\setminus X_j$, (where X_j is the manipulated variable), and one column for each time in the set {5, 50, 100, 500, 1000, 2000, 4000, 10000}. The entry of the table is the RMS error (over the 100 runs) between the predicted value of the variable at that time and the actual value in the test data.

JMLR: Workshop and Conference Proceedings 6:267–276, 2010 NIPS 2008 workshop on causality

Comparison of Granger Causality and Phase Slope Index

Guido Nolte GUIDO.NOLTE@FIRST.FRAUNHOFER.DE
Intelligent Data Analysis Group, Fraunhofer FIRST
Kekuléstr. 7, 12489 Berlin, Germany

Andreas Ziehe ZIEHE@FIRST.FRAUNHOFER.DE
Intelligent Data Analysis Group, Fraunhofer FIRST
Kekuléstr. 7, 12489 Berlin, Germany

Nicole Krämer NKRAEMER@CS.TU-BERLIN.DE
Machine Learning Group, TU Berlin
Franklinstr. 28/29, 10587 Berlin, Germany

Florin Popescu FLORIN.POPESCU@FIRST.FRAUNHOFER.DE
Intelligent Data Analysis Group, Fraunhofer FIRST
Kekuléstr. 7, 12489 Berlin, Germany

Klaus-Robert Müller KRM@CS.TU-BERLIN.DE
Machine Learning Group, TU Berlin
Franklinstr. 28/29, 10587 Berlin, Germany

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We recently proposed a new measure, termed Phase Slope Index (PSI), It estimates the causal direction of interactions robustly with respect to instantaneous mixtures of independent sources with arbitrary spectral content. We compared this method to Granger Causality for linear systems containing spatially and temporarily mixed noise and found that, in contrast to PSI, the latter was not able to properly distinguish truly interacting systems from mixed noise. Here, we extent this analysis with respect to two aspects: a) we analyze Granger causality and PSI also for non-mixed noise, and b) we analyze PSI for nonlinear interactions. We found a) that Granger causality, in contrast to PSI, fails also for non-mixed noise if the memory-time of the sender of information is long compared to the transmission time of the information, and b) that PSI, being a linear method, eventually misses nonlinear interactions but is unlikely to give false positive results.

Keywords: Phase Slope Index, Granger Causality, Noise, Nonlinearity

1. Introduction

To understand the direction of information flow in interacting systems, it is of fundamental importance to distinguish the driver from the recipient. Granger Causality proposed by Granger (1969) is probably the most prominent method to estimate the direction of causal influence in time series analysis.

Apart from Granger Causality, many other methods have been proposed to estimate the direction of information flow both for bivariate and multivariate data. Baccala and Sameshima (1998) suggested to interpret autoregressive matrices in the frequency domain to estimate directionality for bivariate data, which was generalized to multivariate data by Baccala and Sameshima (2001). The approach of Kaminski and Blinowska (1991) is equivalent to the preceding ones for bivariate data, but differs for multivariate data most notably with regard to the question whether estimated information flux is direct or indirect. An information theoretic approach was taken by Schreiber (2000) by analyzing entropies of conditional probabilities (rather than the mean as implicitly done with Granger Causality). A model based method valid for nonlinear and weakly coupled oscillators was proposed by Rosenblum and Pikovsky (2001). With the notable exception of Rosenblum and Pikovsky (2001), all these methods are variations of the highly popular Granger causality, and this will serve as a comparison to our proposed method.

Granger Causality is based on asymmetric prediction accuracies of one time series on the future of another. The difficulty in realistic measurements is that asymmetries can also arise due to other factors, specifically independent background activity having nontrivial spectral properties and eventually being measured in unknown superposition in the channels. In this case the interpretation of the asymmetry as a direction of information flow can lead to significant albeit false results as demonstrated e.g. by Albo et al. (2004). To overcome this difficulty Nolte et al. (2008) recently proposed a method based on a frequency-average of the slope of the phase of coherence defined in such way that it is strictly robust with respect to instantaneous mixtures of independent sources of otherwise arbitrary nature.

In this paper we address two new aspects in more detail. First, in many situations one could argue that, while the measurements are noisy, this noise is not a mixture, and Granger Causality might work for this case. Second, the beneficial properties of PSI might disappear if interactions are nonlinear. We will first shortly recall both methods and then study both mentioned aspects with simulations.

2. Methods

2.1. Granger Causality

The fundamental basis of estimates of causal relations using Granger Causality is the fact that a cause precedes the effect. Probably the simplest way to exploit this idea is to use linear prediction of future values of bivariate data $x_i(t)$ for $i = 1,2$ with AR-modeling:

$$\mathbf{x}(t) = \sum_{p=1}^{P} A(p)\mathbf{x}(t-p) + \boldsymbol{\xi}(t) \tag{1}$$

where $A(p)$ are the AR-matrices up to order P and $\xi(t)$ is white Gaussian noise with estimated covariance matrix Σ.

The diagonal elements of Σ (i.e Σ_{ii} for $i = 1,2$) measure the remaining error when future values of $x_i(t)$ are modeled with both time series', simultaneously. Instead of

one multivariate model one can also model the data by two separate univariate models:

$$x_i(t) = \sum_{p=1}^{P} A_i(p) x_i(t-p) + \xi_i(t) \tag{2}$$

for $i = 1,2$, and where $\xi_i(t)$ has estimated variance Σ_i.

Note that $\Sigma_i \leq \Sigma_{ii}$ because the univariate models do not use information contained in the other time series. The additional information contained in x_j about the future of x_i for $j \neq i$ can be quantified as

$$\Gamma_{j\to i} = \log\left(\frac{\Sigma_i}{\Sigma_{ii}}\right) \tag{3}$$

If $\Gamma_{j\to i} > 0$ one says that channel j 'Granger causes' channel i.

For a unidirectional information flow one has $\Gamma_{1\to 2} = 0$ or $\Gamma_{2\to 1} \neq 0$ or vice versa with obvious direction of information flux. In practice, results are rarely that clear and one can define the effective information flux from the first to the second channel as

$$\tilde{G} = \Gamma_{1\to 2} - \Gamma_{2\to 1} \tag{4}$$

We here normalize $\tilde{G}$ by its standard deviation estimated by the Jackknife procedure. The validity was confirmed in simulations where the same examples were repeated many times. Finally, we define the Granger Causality as

$$G = \frac{\tilde{G}}{std(\tilde{G})} \tag{5}$$

With this normalization we consider any result with absolute value larger than 2 as statistically significant corresponding to a 'pseudo-z-score'. It enables us to compare Granger Causality with Phase Slope Index to defined in the next section.

2.2. Phase Slope Index

In an alternative approach we first divide the whole data set into K segments of duration T (in physical units) and estimate the cross-spectral density as

$$S_{ij}(f) = \frac{1}{K}\sum_k z_i(f,k) z_j^*(f,k) \tag{6}$$

where $z_i(f,k)$ is the Fourier transform of the Hanning-windowed, i.e. multiplied by a raised cosine function, data in channel i and segment k. The 'Phase Slope Index' (PSI) is now defined as (Nolte et al. (2008))

$$\tilde{\Psi}_{ij} = \Im\left(\sum_{f\in F} C_{ij}^*(f) C_{ij}(f+\delta f)\right) \tag{7}$$

where

$$C_{ij}(f) = \frac{S_{ij}(f)}{\sqrt{S_{ii}(f)S_{jj}(f)}} \tag{8}$$

is the complex coherency, $\delta f = 1/T$ is the frequency resolution, and $\Im(\cdot)$ denotes taking the imaginary part. F is the set of frequencies over which the slope is summed. Typically, F contains all frequencies, but it can also be restricted to a specified band for rhythmic activities.

To see that the definition of $\tilde{\Psi}_{ij}$ corresponds to a meaningful estimate of the average slope it is convenient to rewrite it as

$$\tilde{\Psi}_{ij} = \sum_{f \in F} \alpha_{ij}(f)\alpha_{ij}(f+\delta f)\sin(\Phi(f+\delta f) - \Phi(f)) \tag{9}$$

with $C_{ij}(f) = \alpha_{ij}(f)\exp(i\Phi(f))$ and $\alpha_{ij}(f) = |C_{ij}(f)|$ being frequency dependent weights. For smooth phase spectra, $\sin(\Phi(f+\delta f) - \Phi(f)) \approx \Phi(f+\delta f) - \Phi(f)$ and hence $\tilde{\Psi}$ corresponds to a weighted average of the slope.

Let us list the most important qualitative properties of $\tilde{\Psi}$:

1. For an infinite amount of data and for arbitrary instantaneous mixtures of an arbitrary number of independent sources, $\tilde{\Psi}$ is exactly zero, because mixtures of independent sources do not induce an imaginary part of coherencies (Nolte et al. (2004)) which in turn is necessary to generate a non-vanishing $\tilde{\Psi}$. For finite data, $\tilde{\Psi}$ will then fluctuate in this case around zero within error bounds. A special case of this are phase jumps from 0 to $\pm\pi$ which can arise also for mixtures of independent sources.

2. $\tilde{\Psi}$ is expressed in terms of coherencies, only. The standard deviation of a coherency is approximately constant and approximately only depends on the number of averages which is equal for all frequencies. Thus, large but meaningless phase fluctuations in frequency bands containing essentially independent signals are implicitly suppressed.

3. If the phase $\Phi(f)$ is linear in f and provided that the frequency resolution is sufficient (i.e. δf is sufficiently small), the argument in the sum has the same sign across all frequencies and then $\tilde{\Psi}$ will have the same sign as the slope of $\Phi(f)$.

Finally, as for Granger Causality it is convenient to normalize $\tilde{\Psi}$ by an estimate of its standard deviation

$$\Psi = \frac{\tilde{\Psi}}{std(\tilde{\Psi})} \tag{10}$$

with $std(\tilde{\Psi})$ being estimated by the Jackknife method, which was validated in simulations. In the examples below we consider absolute values of each larger than 2 as significant.

3. Simulations and Causality Challenge

3.1. Uncorrelated noise

Granger Causality is based on the assumption that a sender possesses information about the future of the recipient which is not available at the recipient itself, because, roughly speaking, this information has not yet arrived. In contrast, the recipient cannot access any information about the sender other than that already contained in the present and past of the sender because causal interactions are necessarily forward in time.

However, the situation changes when the measurements, especially of the sender, are noisy. In that case the signal of the recipient contains delayed but cleaner information about the sender which is masked/hidden to the sender itself due to the noise. Thus, the slightly outdated information of the receiver may help to predict the future of the sender and yet lead to wrong results in a Granger test. In other words, the disadvantage of the recipient of receiving only old information might have been compensated or even overcompensated by the advantage of being measured in a much cleaner way.

Apparently, the impact of this trade-off depends on the memory time of the sender: If the sender has a long memory and the transmission time is short then the time delay of the interaction is largely irrelevant.

To show this explicitly we simulated clean data of the sources using an $AR(1)$ model with coefficient matrix

$$A(1) = \begin{pmatrix} \alpha & 0 \\ 1 & .5 \end{pmatrix} \tag{11}$$

This system models a unidirectional information flow from channel 1 to channel 2. The memory of the first channel, the sender, is controlled by α: any input decays after n time points as $\alpha^n = \exp(-n\log(1/\alpha))$ and has hence a decay rate of $-1/\log(\alpha)$.

Let us denote the output of this clean system for the $i.th$ source (i.e. true signal of interest) as $x_i(t)$. Then we assume the measurements $y_i(t)$ to be $y_1(t) = x_1(t) + \beta\eta(t)$ and $y_2(t) = x_2(t)$ with $\eta(t)$ being white Gaussian noise and β a free parameter which controls the relative strength of true signal and noise. Results for these systems are shown in Figure 1 for various values of α. We observe that Granger Causality results in significant wrong direction estimates for long memory times of the sender. In contrast, the Phase Slope Index always results in the correct directionality. We note, that with α also the magnitude of the sender changes which also has an impact on the results. However, normalizing the sender leads to essentially identical results provided that influence of the sender on the recipient is at least as large as the innovation process of the recipient, i.e. of $\xi_2(t)$. This leads to the somewhat paradoxical situation that for noisy measurements the larger the causal drive from A to B the more likely Granger Causality estimates a drive from B to A.

In a second example, we simulated the situation based on real EEG data. Results for power and autocorrelation function are shown in Figure 2. The memory time of the system is about 0.5 seconds which is large compared to typical transmission times along neuronal fibers. Neuronal signals in axons in white brain matter, which are relevant for long distance information transfer, travel with a speed of about 1cm/msec and need

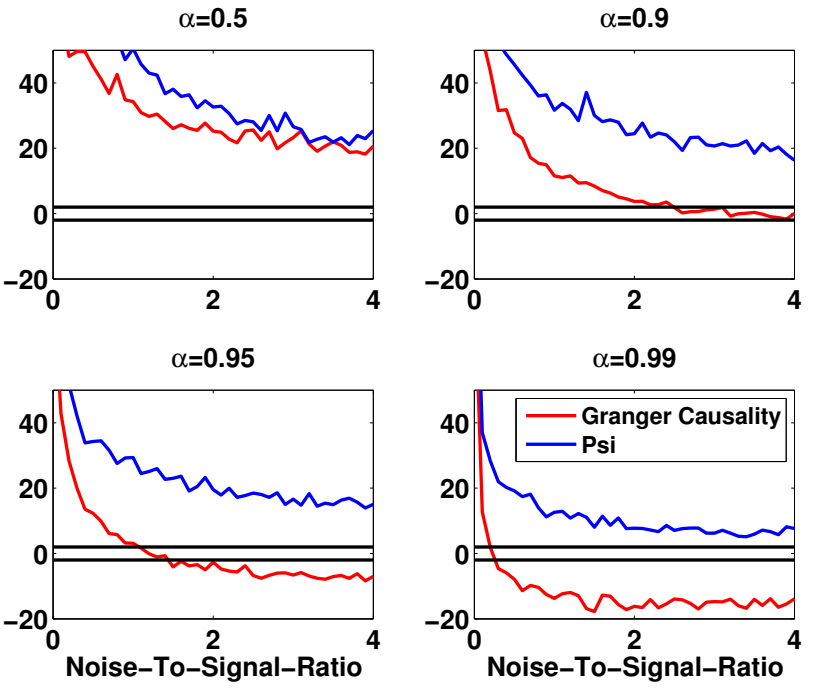


Figure 1: Granger Causality and PSI as a function of noise level for systems with different memory. The memory time in bins is roughly given by $1/(1-\alpha)$. Values outside the narrow horizontal strip are statistically significant.

only a few milliseconds to cross the whole brain. The simulation was identical to the previous one with the exception that the real data $x(t)$, normalized to unit standard deviation, were taken as sender and the recipient was assumed to be $y(t) = 2x(t-3) - .5y(t-1) + \eta(t)$ with $\eta(t)$ being white Gaussian noise with unit standard deviation. Since the sampling rate was 256 Hz, the delay corresponds to a transmission time of about $12ms$. Results again showed that already a fairly small amount of noise put on the measurement of the sender is sufficient to result in significant false direction estimates of Granger Causality while PSI always predicted the correct direction.

3.2. Nonlinear interactions

To test Granger Causality and PSI for bivariate nonlinear systems we included a nonlinearity of specific order into the interaction term and generated 500 examples as randomly as possible. The data $\mathbf{z}(t)$ were generated as

$$\mathbf{z}(t) = (1-\gamma)\frac{\mathbf{x}(t)}{||X||} + \gamma\frac{B\mathbf{y}(t)}{||BY||} \tag{12}$$

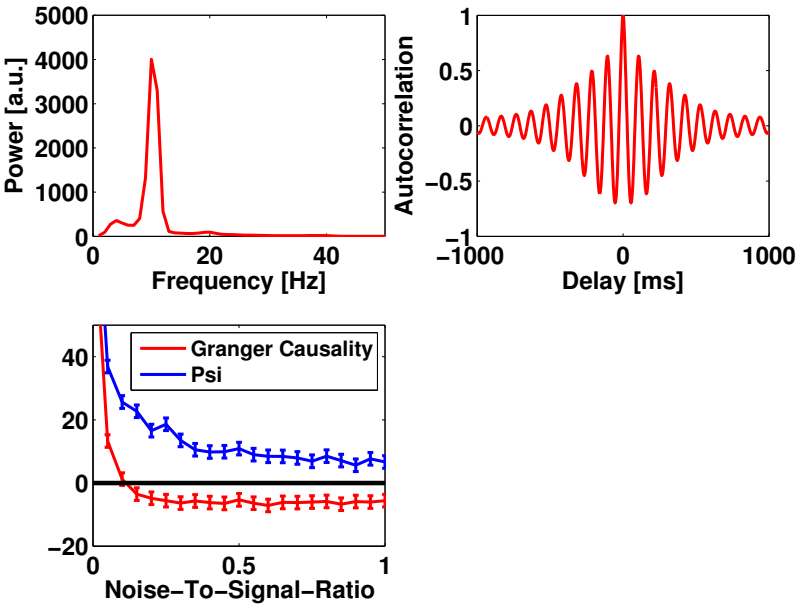


Figure 2: Top panels: Power and autocorrelation of a single channel for a real EEG experiment. Bottom: Granger Causality and PSI as a function of noise level using the real EEG data as driver and adding noise to the sender. The noise level is measured as the power ratio at the peak frequency. The vertical bars correspond to two estimated standard deviations and indicate significance if they do not cross the zero line.

where $\mathbf{x}$ is a unidirectional and in general nonlinear system and $\mathbf{y}$ are two independent noise sources which are mixed into channels by a random matrix B. The parameter γ was set randomly between 0 and 1, $||\cdot||$ denotes Frobenius matrix norm, and X and Y denote the full data as a matrix, e.g. $X = (\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N))$ for N data points. The noise $\mathbf{y}(t)$ was generated with an AR(10)-model with diagonal but otherwise random parameters. The signal $\mathbf{x}(t)$ was generated in the following way. If, e.g., the first channel was the sender then $x_1(t)$ was generated with a random AR-model of order 10, and $x_2(t)$ was generated as

$$x_2(t) = \sum_p A_{22}(p) x_2(t-p) + f(x_1(t-1), ..., x_1(t-P)) \tag{13}$$

where P was set to 10 and f is a in general nonlinear function of specific order chosen in the most general way. E.g., for order 4 the function f was given by

$$f(x_1(t-1), ..., x_1(t-P)) = \sum_{ijkl} a_{ijkl} x_1(t-i) x_1(t-j) x_1(t-k) x_1(t-l) \tag{14}$$

with random parameters a_{ijkl}. The construction for other orders is analogous.

Results for PSI and Granger Causality are shown in Figure 3 and Figure 4, respectively. We observe that PSI, in contrast to Granger Causality, hardly ever results in false significant direction estimates. We also observe that for even order of nonlinearity PSI is also not able to detect any interaction at all. However, this can be explained by the sign symmetry of the interaction and is due to the linear nature of PSI.

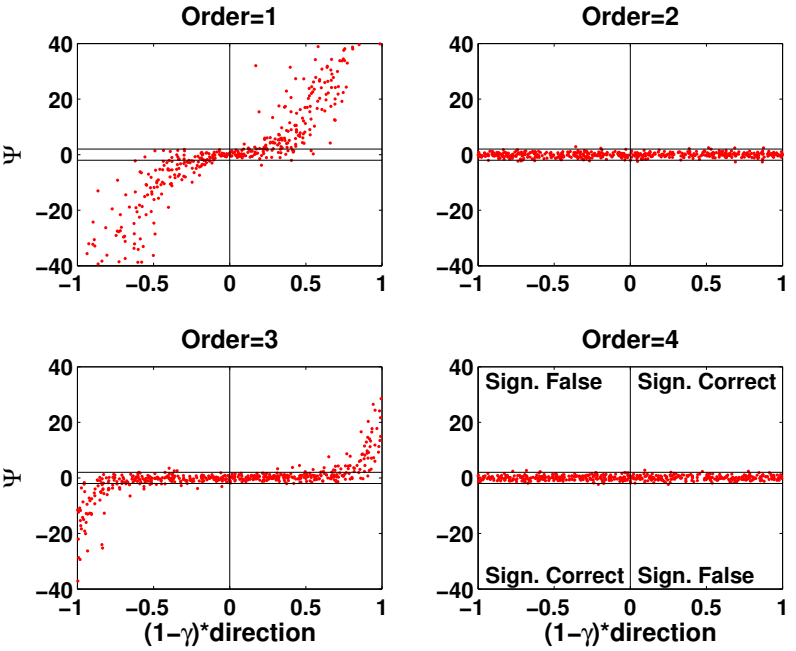


Figure 3: Results for PSI for 500 random systems as a function of noise level times the sign of true direction for different orders of nonlinearity (order=1 is linear). All results in the narrow horizontal strip are insignificant, and the others are as indicated in the lower right panel. For each panel, left and right borders, i.e. $1-\gamma=1$, correspond to zero noise and the center, i.e. $\gamma=1$, corresponds to only noise.

3.3. Causality Challenge

We submitted a dataset to the Causality Challenge[1] which consists of 1000 examples identical to the ones in the previous section for the *order* $=1$ case except for two minor details: for the challenge we chose uniformly distributed innovation processes (i.e. $\boldsymbol{\xi}(t)$ in Eq.(1)) instead of Gaussian distributed input, and we chose three noise sources instead of two.

1. "NOISE", http://www.causality.inf.ethz.ch/repository.php?id=17

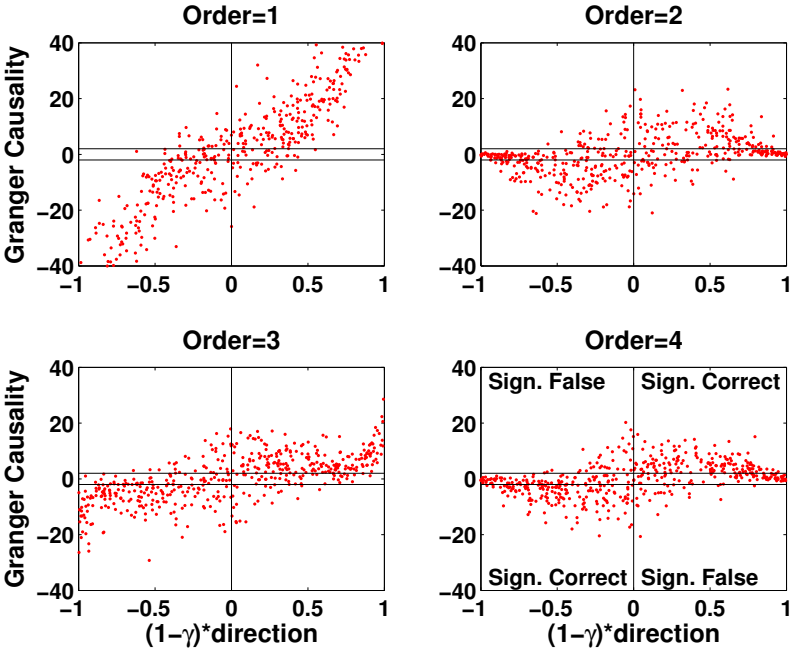


Figure 4: Same as Figure 3 for Granger Causality.

The task is to estimate the causal direction for as many examples as possible. The counting is as follows: +1 point for each correct result, -10 points for each wrong results, and 0 points for each missed example. For the top left panel of Figure 3 this means that one gets +1 point for each dot in the lower left or upper right box, -10 points for each point in the lower right or upper left box, and 0 points for each point in the narrow horizontal stripe.

For the challenge data, Granger causality leads to 736 correct and 100 wrong results scoring a total of -264 points. Note, that 164 insignificant results are not counted. In comparison, PSI[2] leads to 638 correct and 6 wrong results scoring a total of +578 points.

This counting was introduced to address the importance of evidence for scientific claims. A finding which was just guessed right has little value. In many cases conclusions cannot be drawn with the given data measured in a specific situation. Researchers must be able to also recognize these cases and should then not draw conclusions at all.

In a second set of data sets we provided real EEG data for 10 subjects measured at rest in eyes closed condition. A specific feature of this measurement is a strong 10Hz rhythm predominantly in the back part of the brain. Using our methods we found information flow from front to back, i.e. from channels with low signal to ratio to channels with high signal to ratio.

2. The Matlab code can be downloaded at http://ml.cs.tu-berlin.de/causality/

The real data used in section 3.1 were taken from one of these subjects. We showed in that section that Granger Causality has a bias to estimate direction from clean to noisy signals, and a finding using Granger Causality stating that information is flowing from back to front is possibly caused by the different signal to noise ratios rather than by true information flow.

For the challenge we can only put this to discussion since the ground truth is not known. We therefore just presented our own results and provided excellent data sets to let people apply their own methods to this case.

4. Conclusion

The paper presents novel insights on causality measures and carefully evaluates their domain of applicability. In particular, we present simulations that contrast Granger Causality and our new Phase Slope Index.

Interestingly, under noise the classical Granger Causality can fail, even to an extent that a wrong causal direction is inferred with a high significance level and even if noise is uncorrelated.

We could show that the PSI approach does not suffer from such a shortcoming including in simulations modeling random and highly nonlinear interactions. Clearly real-world data are always noisy and many complex technical or biological systems contain nonlinear elements. Therefore inference on causal structure in data is required to be robust, a property that is inherent to our proposed new method.

Appendix. Pot-luck challenge: FACT SHEET .

(for a donated dataset)

Repository URL: http://www.causality.inf.ethz.ch/repository.php?id=17

Dataset name: NOISE

Title: Causal Directions in Noisy Environment
Author: Guido Nolte
Address: Fraunhofer FIRST, Kekulestr. 7, 12489 Berlin, Germany
Email: guido.nolte@first.fraunhofer.de
Homepage: http://ida.first.fraunhofer.de/~nolte/

Key facts:
A: 1000 examples of real valued bivariate data with 6000 time points each. B: Real EEG data of 10 subjects.

Abstract:
This challenge has two parts, a simulation and real data.
Simulation: Data are simulated as superposition of bivariate unidirectional interaction plus additive mixed and non-white noise. The simulations were done with AR-models with uniformly

distributed input. The challenge is to estimate the causal direction. For each out of 1000 examples you get +1 point for the correct answer, -10 points for the wrong answer, and 0 points for no answer.
Real Data: These are high quality EEG data for 10 subjects for 19 channels. The data contain a prominent peak at around 10 Hz predominantly in occipital (back) channels. No ground truth is known. A submission must be a single 19x19 matrix corresponding to a causality estimate between all pairs of channels averaged across subjects. Any submission will be visualized and, with the agreement of the authors, put on the net for an open discussion.

Keywords:
Time series, mixed noise, bivariate, EEG

References

Z. Albo, G.V. Di Prisco, Y. Chen, G. Rangarajan, W. Truccolo, J. Feng, R.P. Vertes, and M. Ding. Is partial coherence a viable technique for identifying generators of neural oscillations? *Biol. Cybern.*, 90:318–326, 2004.

L.A. Baccala and K. Sameshima. Directed coherence: a tool for exploring functional interactions among brain structures. *Methods for Simultaneous Neuronal Ensemble Recordings, CRC Press, Boca Raton*, pages 179–192, 1998.

L.A. Baccala and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biol Cybern.*, 84:463–74, 2001.

C.W.J. Granger. Investigating causal relations by economic models and cross-sprectal methods. *Econometrica*, 37:424–438, 1969.

M. Kaminski and K.J. Blinowska. A new method of the description on information flow. *Biol.Cybern.*, 65:203–210, 1991.

G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett. Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clin. Neurophysiol.*, 115: 2292–2307, 2004.

G. Nolte, A. Ziehe, V.V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Phys Rev Lett*, 100:234101, 2008.

M.G. Rosenblum and A.S. Pikovsky. Detecting direction of coupling in interacting oscillators. *Phys. Rev. E*, 64:045202, 2001.

T. Schreiber. Measuring information transfer. *Phys. Rev Let.*, 85:461–4, 2000.

JMLR: Workshop and Conference Proceedings 6:277–288, 2010 NIPS 2008 workshop on causality

Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control

Michael McCann MCCANN-M15@EMAIL.ULSTER.AC.UK
Yuhua Li Y.LI@ULSTER.AC.UK
Liam Maguire LP.MAGUIRE@ULSTER.AC.UK
Adrian Johnston A.JOHNSTON@EMAIL.ULSTER.AC.UK
Intelligent Systems Research Centre
University of Ulster
Derry
N.Ireland

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

A complex modern manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be used to identify the most predictive signals. Once these signals have been identified causal relevance may then be investigated to try and identify the causal features. The Process Engineers may then use these signals to ensure a small scrap rate further downstream in the process, increase the throughput and reduce the per unit production costs. Working in partnership with industry we aim to address this complex problem as part of their process control engineering in the context of wafer fabrication production and enhance current business improvement techniques with the application of causal feature selection as an intelligent systems technique.

Keywords: Causal discovery, feature selection, semi-conductor manufacturing, industry, business improvement techniques

1. Introduction

In high volume manufacturing close control and monitoring of production processes are required to ensure quality control and efficiency (Jeong and Cho, 2006). Considering the number of process steps in wafer fabrication, typically over 500, and the amount of data recorded during the entire production process, this produces a vast amount of monitoring data. However not all of this data is equally relevant for process control monitoring. Within this environment industry standard business improvement techniques are the tools that are used to try and solve this complex problem. Currently within industry

Six Sigma is one of the main business improvement strategies employed to improve the manufacturing process, although this is a well proven technique throughout industry there are a number of weaknesses inherent within its approach (Johnston, 2007). The application of new computational intelligence techniques is now being introduced in manufacturing environments. The introduction of feature selection techniques is proposed as an intelligent systems approach to solving this issue. These techniques are prevalent in high volume data environments, in this application domain they may be deployed to identify the desired Key Process Input Variables (KPIVs) and assess their causal relevance. Once identified process engineers may then use these KPIVs to significantly reduce the time required to reach mature product target line yield figures for new product integration with an overall impact on bottom line production costs. The aims and objectives are to investigate and understand the nature of the complex process control issues faced on a daily basis by the semi conductor industry particularly in high volume manufacturing, with a view to research and develop a causal feature selection methodology that can be combined in a hybrid approach to business improvement in this domain. This solution will address the impact of KPIVs on production line yield figures failure rates and hence improve efficiency specifically in the area of new product development (NPD).

Section 2 provides an introduction to how intelligent systems are being deployed within industry to enhance their current business improvement strategies. Section 3 gives an overview of feature selection and causal relevance. Section 4 describes the SECOM dataset that has been put forward for the challenge along with some baseline results and Section 5 outlines conclusions and future work proposing how feature selection and causal relevance may be applied to process control engineering within the semi conductor industry.

2. Business Improvement and Intelligent Systems

Within an industrial context there has been a growing requirement for the introduction of intelligent system techniques over the past 10 years to assist process engineers with their decision-making (Johnston, 2007; Peretto, 1999). The advances in hardware automation and control systems have impacted the overall importance of utilizing these new techniques within manufacturing (Harrison and Petty 2002). One of the issues faced by engineers in a modern manufacturing environment is how experiential knowledge is utilized within the decision making process. The use of intelligent systems to aid in this decision-making process helps to overcome this problem, current techniques include Fuzzy Logic (FL), Artificial Neural Networks (ANNs) and Genetic Algorithms (GAs). Cus and Balic (2003) propose the use of GAs for use in metal cutting processes to optimize parameters in machine operation and FL combined with ANNs are proposed for grinding processes by Chen and Kumara (1998) for automation of design. As each of these intelligent techniques have different advantages and disadvantages, see Table 1, hybrid combinations are often used to address complex systems. An example of a hybrid system is proposed by Guh et al. (1999) for use in Statistical Process Control (SPC) combining neural networks and expert systems.

Table 1: Intelligent System Techniques Properties (Johnston, 2007)

	Properties				
	Reasoning	Generalisation	Decision-making	Adaption	Rule Visibility
Neural Networks	✘	✔	✘	✔	✘
Fuzzy Logic	✔	✘	✔	✘	✔
Genetic Algorithms	✘	✘	✔	✔	✘
Expert Systems	✔	✘	✔	✘	✔
Case Based Reasoning	✔	✔	✔	✔	✔

Six Sigma is one of the main business improvement strategies employed in the manufacturing process. The determining factor within Six Sigma is its aim to identify causal KPIVs and therefore ensure that process outputs remain in control (Flott, 2000; Card, 2000; Rao et al., 2000; Schmidt et al., 1998). One of the major issues in applying Six Sigma as an improvement strategy within a high volume production environment, where time to full production for integration products is such a critical milestone, is that due to the nature of Six Sigma projects they tend to be time consuming and project centric (Johnston, 2007). Thus although it is an industry standard technique in certain circumstances it is not always a feasible solution. The Six Sigma process flow for project implementation is shown in Figure 1. Once a project has been defined the initial measure phase is typically conducted by a project team consisting of all parties that have the relevant expertise and a stake hold in the overall project definition.

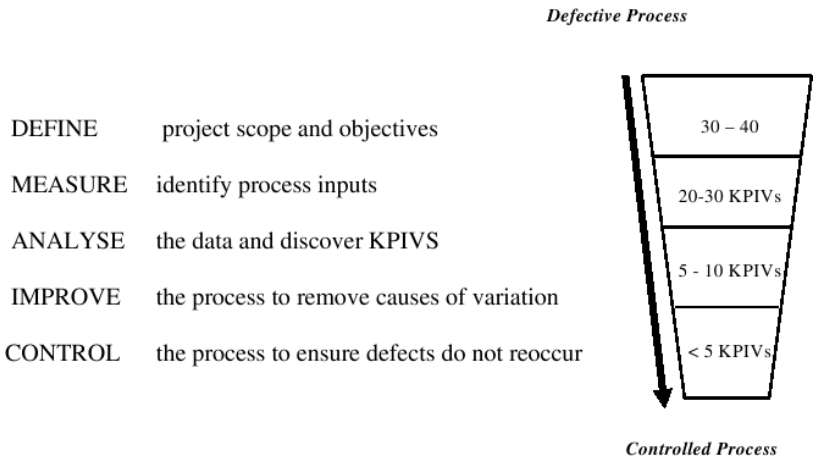


Figure 1: Six Sigma Process Flow (Johnston, 2007)

Therefore this phase and hence the overall success of the project is highly dependent on project team experiential knowledge, which unfortunately can be lost, forgotten or invalid for new projects. The advantage of considering intelligent systems such as causal feature selection methods to solve a similar problem is the fact that it does not

rely on this experiential knowledge as much to narrow down the processes that are under consideration (Patterson et al., 2005). This allows all the data that is relevant to the overall scope of the project definition to be considered when trying to discover the desired KPIVs. This also overcomes another issue known as the "anchoring effect" wherein project teams tend to focus on impact processes that have previously displayed concerns within similar project types. Hence this form of human conditioning can lead to previously undiscovered KPIVs being excluded from investigation. During the analysis phase of the project statistical tools are employed to analyze the data that has been identified from the measure phase. Once again this phase is dependent upon the engineer applying the appropriate statistical analysis techniques for the extraction and interpretation of the data such as hypothesis testing on individual process input variables. This entire phase is extremely time and labour intensive and therefore is not always appropriate for time critical projects (Johnston, 2007). The improvement phase then requires the consideration of implementing the appropriate actions from these findings to be integrated into the current process flow. This may require optimisation with procedures such as design of experiment (DOE) and potentially failure modes and effects analysis (FMEA). Unfortunately this type of procedure is practically unfeasible in a high volume manufacturing environment because of the amount of data and time required to run trials which has an impact on production and scrap rates. It would be much more desirable to introduce an intelligent systems approach that was able to identify causal KPIVs and apply this methodology in tandem with overall business improvement strategies.

3. Feature Selection

In recent years the nature of feature selection has changed in terms of the complexity of the application. For example in 1997 the applications explored in this field seldom contained more than 40 features (Chiang and Pell, 2004; Kohavi and John, 1997), whereas in recent years this has changed as feature selection methods are required for domains with in excess of tens of thousands of features such as in gene selection (Guyon et al., 2002), text categorisation (Liu et al., 2005) and other various engineering applications (Guyon and Elisseeff, 2003). The selection of relevant features, and the elimination of irrelevant ones, is one of the central problems in machine learning (Blum and Langley, 1997). There have been significant advances in feature selection development in recent years and there are a significant number of methods that can be utilised to try and achieve the optimum results. In pattern recognition, the goal of feature selection is to find a feature subset that has the most discriminative information from a given set of a candidate features (Abe and Kudo, 2006).

Data representations tend to be very domain specific (Guyon and Elisseeff, 2003). Once data is available for machine learning it is often required to manipulate this "raw" data into a format that is conducive to the methodology that is to be applied. This is known as feature construction and may involve simple data manipulation or the application of data transformations. This is often achieved through what is known as pre-processing steps some simple examples of which are (Guyon et al., 2006):

Standardisation e.g. measurements that have different scales

Normalisation e.g. pixel intensity values in image processing

Signal enhancement e.g. smoothing or sharpening

Principal component analysis and multidimensional scaling projecting data into a lower dimensional space whilst retaining the information.

Feature selection then is primarily performed to select the most informative features but other motivations include (Guyon et al., 2006):

General data reduction for storage requirements and processing speed

Feature set reduction to save resources

Performance improvement to gain predictive accuracy

Data understanding to gain knowledge of the process that generated the data or visualisation

3.1. Causal Considerations

Although feature selection on its own is mainly concerned with making accurate predictions with as few variables as possible it does not follow that these variables are necessarily causal within a specific domain. The issue faced by semi conductor manufacturing is not a typical predictive or classification one, it has a large causal element to the problem. For high volume manufacturing the key requirement is to determine which of the variables selected prove causal in terms of affecting failure rates on the factory line yield. So the optimum results would involve identifying these KPIVs giving process engineers insight into the hidden causal relationships within individual manufacturing process steps and overall line yield pass/fail rates. Obviously in real life terms validation of results is not always feasible because of the financial impact of experimental alterations on production processes and the associated unknowns on yield excursions. For this reason it is proposed that any intelligent systems approach to process control be sanitised by inclusion in existing business improvement techniques such as Six Sigma.

4. SECOM: SEmiCOnductor Manufacturing dataset

This challenge aims to investigate a range of feature selection techniques and how appropriate they are to identifying the causal effects faced by process control engineering in semi conductor manufacturing. "In the manufacturing process of semiconductor products one deals with a great number of production steps that involve many different machines. Malfunctions can usually not be ruled out or identified in each processing step" (Pfingsten et al., 2007). Operating conditions can change frequently in a process control environment both intentionally and unintentionally identification of the KPIVs allows rapid recovery, optimisation and control (Chiang and Pell, 2004). The goal of

this case study is to develop a causal feature selection approach that applies to this domain, helps to solve process control issues and enhance overall business improvement strategies.

Consider in more detail at the nature of the wafer fabrication production process. In the case of integration products it takes time to tweak the processes to achieve target yield figures. Feature selection techniques may be applied to the production process to provide the process control engineers with the necessary intelligence to decrease this integration time and achieve target yield figures earlier in the product life cycle and hence proceed into full production quicker. As highlighted earlier current strategies depend heavily on experiential knowledge which limits the data under investigation and is time consuming. Figure 2 shows "time to yield" baseline trends for integration products i.e. the time required to get new products up to target yield figures hence improving time to market. Good line yields mean:

- Low cost per product
- Predictable schedule adherence and starts planning
- Can run the factory leaner (fewer starts)
- Better throughput at critical tools
- Better quality downstream
- Better product predictability
- No 'firefighting' – more resource for project work
- Less waste – less use of consumables

By enabling process engineers to identify KPIVs earlier in the production process it should enable them to affect yield figures more accurately and increase productivity using a more efficient strategy and hence achieving target yield figures for integration products.

4.1. Data Structure

The SECOM dataset presented in this paper, (for a summary see Appendix A), represents a selection of process related data taken from a production line. The dataset is presented with features in columns each representing a recorded measurement and product examples in rows. Within the production cycle there are several major check points for in house line testing to ensure product functionality as demonstrated in Figure 3. The labels file then represents a simple pass/fail classification corresponding to each row in the dataset, where -1 corresponds to a pass and 1 corresponds to a fail. A date-time stamp for each pass/fail is also provided in the labels file corresponding to a selected functionality test.

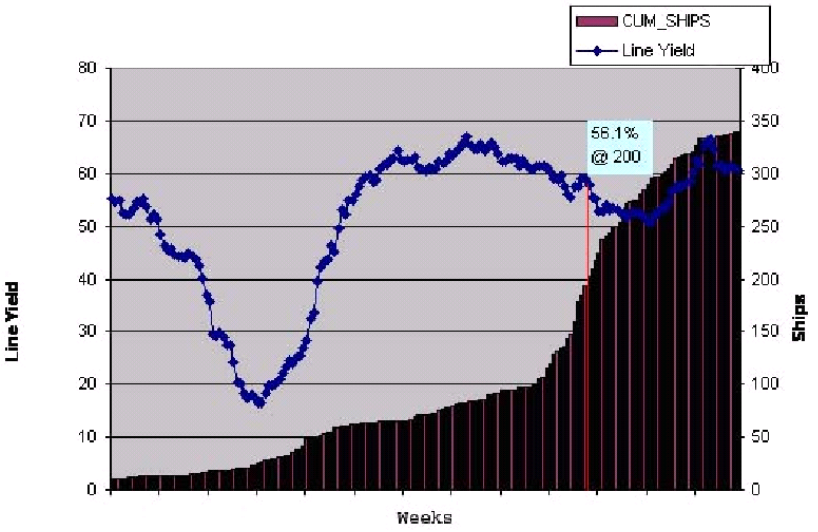


Figure 2: Line Yield Trends

The data consists of 2 files, the dataset file SECOM consisting of 1567 examples each with 591 features, a 1567×591 matrix, and a labels file containing the classifications and date time stamp for each example. As with any real life data situations this data contains null values varying in intensity depending on the individual features corresponding to data-points with no recorded measurement in the original data. This may be taken into consideration when investigating the data either through pre-processing or within the technique applied. Using feature selection techniques it is desired to obtain a sub-set of the most predictive features and then consider the causal relationships within these features and how they impact on the overall pass/fail rates for the product. It is suggested that cross validation be used for generalization performance. Some baseline results are given below.

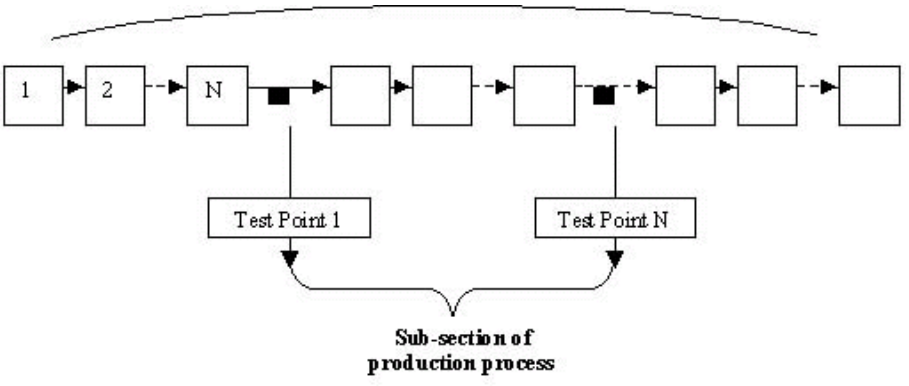


Figure 3: Production Cycle

Baseline Results: Preprocessing objects were applied to the dataset simply to standardize the data and remove the constant features. Then a number of simple statistical feature-ranking techniques were applied with a simple Naïve-Bayes classifier to achieve some initial baseline results. 40 features were selected in each case. 10 fold cross validation was used and the balanced error rate (BER) generated as an initial performance metric to help investigate this dataset. The results are shown in Table 2 below. The desired goals at this stage are to improve upon these error rates for models selecting no more than 40 features and investigate the causal relationships with the target values.

Table 2: SECOM Dataset: 1567 examples 591 features, 104 fails

FSmethod (40 features)	**BER %**	**True + %**	**True − %**
No feature selection	36.9 ±2.4	43.8 ±4.7	82.4 ±1.5
S2N (signal to noise)	34.5 ±2.6	57.8 ±5.3	73.1 ±2.1
Ttest	33.7 ±2.1	59.6 ±4.7	73.0 ±1.8
Relief	40.1 ±2.8	48.3 ±5.9	71.6 ±3.2
Pearson	34.1 ±2.	0 57.4 ±4.3	74.4 ±4.9
Ftest	33.5 ±2.2	59.1 ±4.8	73.8 ±1.8
Gram Schmidt	35.6 ±2.4	51.2 ±11.8	77.5 ±2.3

Initial findings and baseline results suggest it may be desirable to increase the size of the dataset significantly to improve performances and allow for separate final tests sets.

5. Conclusion and Future Work

Introducing intelligent system techniques such as causal feature selection within a high volume manufacturing environment would overcome many of the difficulties that have been outlined. Research by Pfingsten et al suggests the use of feature selection to consider the complete assembly line and detect key processes that affecting yield (Pfingsten et al., 2007). Previously undiscovered KPIVs could then potentially be identified earlier in the product integration life cycle where time is of critical consideration. Although intelligent techniques have seen significant advances in deployment, feature selection has not been seen wide spread use within the semi conductor industry. By investigating how causal feature selection can be deployed within a process control environment, it is proposed that a hybrid approach employing the appropriate feature selection techniques and existing business improvement techniques be designed. This enhanced business improvement strategy may then be deployed to achieve more effective monitoring and process control. This should allow engineers to consider all of the possible KPIVs across the complete production process and overcome some of the disadvantages associated with current methods.

References

N Abe and M. Kudo. Non-parametric classifier-independent feature selection. *Pattern Recognition*, 39(5):737–46, 2006.

A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. 97(1–2):245–71, 1997.

D. N. Card. Sorting out Six Sigma and the CMM. *IEEE Software*, 17(3):1–13, 2000.

Y. T. Chen and S. R. T. Kumara. Fuzzy logic and neural networks for design of process parameters: a grinding process application. *International Journal of Production Research*, 36(2):395–415, 1998.

L. H. Chiang and R. J. Pell. Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control*, 14(2):143–55, 2004.

F. Cus and J. Balic. Optimization of cutting process by GA approach. *Robotics and Computer Integrated Manufacturing*, 19(1–2):113–121, 2003.

L. W. Flott. Six-Sigma controversy. *Metal Finishing 0026-0576*, 98:43–48, 2000.

R. S. Guh, J. D. T. Tannock, and C. O'Brien. IntelliSPC: a hybrid intelligent tool for on-line economical statistical process control. *Expert Systems with Applications*, 17 (3):195–212, 1999.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7):1157–82, 2003.

I. Guyon, J. Weston, and S. Barnhill. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness & Soft Computing)*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, har/cdr edition, July 2006.

B. Jeong and H. Cho. Feature selection techniques and comparative studies for large-scale manufacturing processes. *International Journal of Advanced Manufacturing Technology*, 28(9):1006–11, 2006.

A. Johnston. *Integrating Business Improvement and Intelligent Systems in high volume manufacturing*. PhD thesis, 2007.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324, 1997.

L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. In *Proceedings of the 2005 12th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE '05)*, pages 597–601, 30 October–1 November 2005.

A. Patterson, P. Bonissone, and M. Pavese. Six Sigma applied throughout the lifecycle of an automated decision system. *Quality and Reliability Engineering International*, 21(3):275–292, 2005.

P. F. Peretto. Industrial development, technological change, and long-run growth. *Journal of Development Economics*, 59(2):389–417, 1999.

T. Pfingsten, D. J. L. Herrmann, T. Schnitzler, A. Feustel, and B. Scholkopf. Feature selection for troubleshooting in complex assembly lines. *IEEE Transactions on Automation Science and Engineering*, 4(3):465–9, 2007.

M. Rao, X. Sun, and J. Feng. Intelligent system architecture for process operation support. *Expert Systems with Applications*, 19(4):279–288, 2000.

D. C. Schmidt, J. Haddock, S. Marchandon, G. C. Runger, W. A. Wallace, and R. N. Wright. A methodology for formulating, formalizing, validating, and evaluating a real-time process control advisor. *IIE Transactions*, 30(3):235–245, 1998.

Learning Causal Protein-Signaling Network From Experimental Data

Ping He SUNHP@PKU.EDU.CN
Zhi Geng ZGENG@MATH.PKU.EDU.CN
Wei Yan YANWEI1982@PKU.EDU.CN
Zhihai Liu DEMETRIO@PKU.EDU.CN
School of Mathematical Sciences, Peking University Beijing 100871, China

Task solved: CYTO

Reference:

1. A. J. Hartemink. Principled Computational Methods for the Validation of and Discovery of Genetic Regulatory Networks. Unpublished doctoral thesis, Massachusetts Institute of Technology, 2001.

2. G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 116–125, 1999.

3. K. Sachs, O. Perez, D. Per, D. A. Lauffenburger, and G. P. Nolan. Causal Protein-Signalling Networks Derived From Multiparameter Single-Cell Data. *Science*, 308, 523C529, 2005.

4. Y. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. To appear in J. Machine Learning Research, 9, 2008.

5. Y. He, Z. Geng and X. Liang Learning causal structures based on Markov equivalence class. LNAI 3734, 92–106, ALT 2005, S. Jain, H. U. Simon and E. Tomita Eds. Springer-Verlag, Berlin, 2005.

Method:

At the preprocessing step, the original continuous data are discretized into 3 levels by using the information-preserving technique (Hartemink, 2001).

We propose an approach for discovering causal networks from multiple data bases with external interventions. In our approach, we first find a skeleton or a Markov equivalence class of networks, in which there are undirected and directed edges. Then we orient undirected edges in terms of information on causality from data sets with external interventions. Intuitively intervening a cause affects its effects, but intervening an

effect does not affect its causes. For each undirected edge, we determine its orientation using data sets with external interventions to its nodes.

Results:

The path matrix of the causal network obtained by using our approach for the CYTO data set is shown in Table 1, where '0' in cell (i, j) denotes no edge between nodes i and j, '1' denotes a directed edge $i \to j$, '-1' denotes a directed edge $i \leftarrow j$, and '2' denotes an unoriented edge. Our causal network includes 11 directed edges and 5 undirected edges. The network depicted in Figure 1 shows a comparison between our network G with the classic network G' mentioned by Sachs et al. (2005), where the black edges are consistent in both G and G', the red edges are those with reversed orientation in both networks, the blue edges are those in G but not in G', and the dashed ones are those not in G but in G. The network after drawing the dashed edges is our result G. Our network G is quite different to G'. It may be because our G is constructed as a whole network.

Table 1: The path matrix of the network obtained by our approach for the CYTO data set

	praf	pmek	plcg	PIP2	PIP3	Erk	Akt	PKA	PKC	P38	pjnk
praf	0	2	0	0	0	0	0	-1	0	0	0
pmek	2	0	0	0	1	0	0	0	0	0	0
plcg	0	0	0	-1	2	0	0	0	0	0	0
PIP2	0	0	1	0	1	0	0	0	0	0	0
PIP3	0	-1	2	2	0	2	-1	-1	0	2	2
Erk	0	0	0	0	2	0	1	-1	0	0	0
Akt	0	0	0	0	1	-1	0	1	0	0	0
PKA	1	0	0	0	1	1	-1	0	0	0	0
PKC	0	0	0	0	0	0	0	0	0	1	0
P38	0	0	0	0	2	0	0	0	-1	0	2
pjnk	0	0	0	0	2	0	0	0	0	2	0

Advantages of our approach: We propose an approach for structural learning from multiple data bases with external interventions. Comparing with the Bayesian approach via MCMC proposed by Sachs et al. (2005), our approach have higher computational efficiency. The Bayesian approach is a score-based method, and our approach is a constraint-based method. Since it is difficult to find posteriors for all possible causal networks in Bayesian approach, it uses MCMC to find posteriors of edges and then combines those edges with high posteriors together to construct a network, but such a network may not have the maximum posterior. Different to this, our approach determines orientation of undirected edges selecting data sets with external interventions, and thus this approach can also be used for intervention design. Our approach is based on conditional independence test, which can be easily executed when the number of

variables is small. According to our simulation, we find that our approach has high accuracy when the sample size is small.

Keywords: Active learning, Causal Network, Structural learning

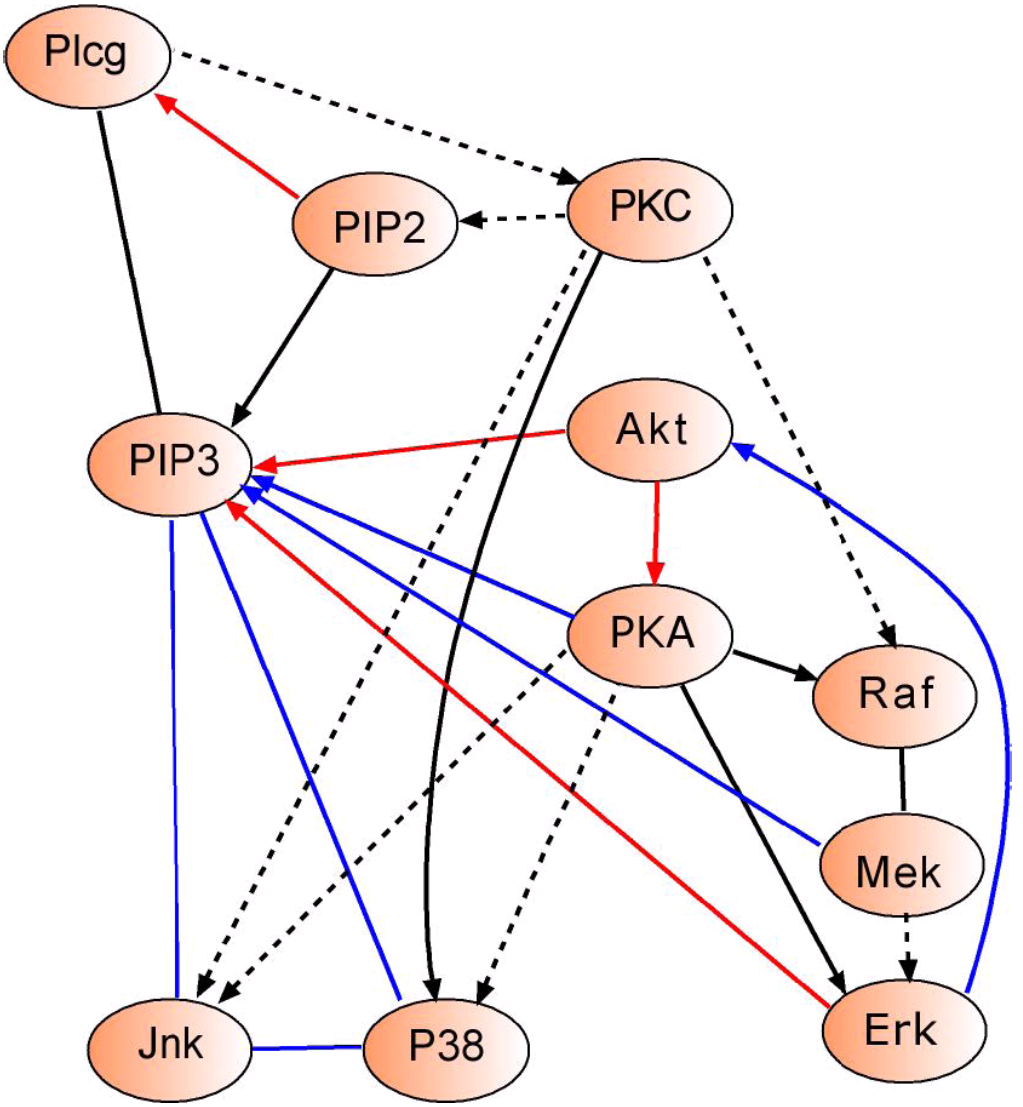


Figure 1: Results of our approach on CYTO data.

Learning Causal Protein-Signaling Networks

Jin Tian JTIAN@CS.IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
Akshay Deepak AKSHAYD@CS.IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Task(s) solved: CYTO

Reference:

Method:

1. Preprocessing: We used the discretized data in Sachs et al. (2005) consisting of 5400 samples with 600 samples per condition.

2. Causal discovery: We used the Bayesian approach to learn causal Bayesian networks from mixed observational and experimental data. We computed the maximum a posteriori (MAP) network using the dynamic programming algorithm in (Silander and Myllymaki, 2006).

Results:

The MAP network (Figure 1).

Keywords:

- Causal discovery: Bayesian Network.

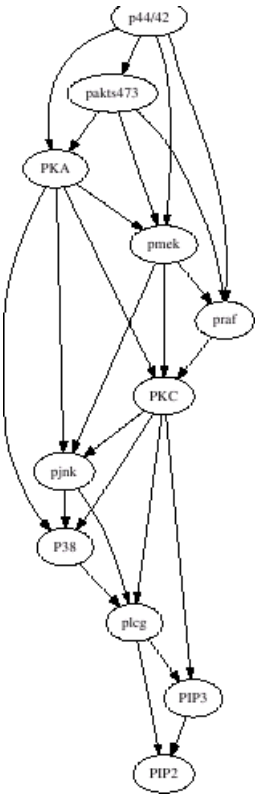


Figure 1: The MAP network

LOcal CAusal NETwork

Isabelle Guyon

Alexander Statnikov

Constantin Aliferis

Pot-luck challenge: FACT SHEET

Repository URLs:
LOCANET http://www.causality.inf.ethz.ch/data/LOCANET.html
REGED http://www.causality.inf.ethz.ch/repository.php?id=7
SIDO http://www.causality.inf.ethz.ch/repository.php?id=1
CINA http://www.causality.inf.ethz.ch/repository.php?id=6
MARTI http://www.causality.inf.ethz.ch/repository.php?id=8

Task name: LOCANET
Title: LOcal CAusal NETwork
Authors: Isabelle Guyon, Alexander Statnikov, Constantin Aliferis
Contact: Isabelle Guyon, isabelle@clopinet.com, http://clopinet.com/isabelle

Key facts:

Data sets for discovering the local structure around a target variable. Time independent tasks. Learning causal structure from observational data. Four semi-artificial datasets (two using re-simulated data and two using real data augmented with artificial probe variables):

Dataset	Domain	Type	Features	Feat. #	Train #	Test #
REGED	Genomics	Re-simulated	Numeric	999	500	20000
SIDO	Pharmacology	Real + probes	Binary	4932	12678	10000
CINA	Econometrics	Real + probes	Mixed	132	16033	10000
MARTI	Genomics	Re-simulated	Numeric	999	500	20000

Abstract:

We designed four datasets for the purpose of benchmarking local causal discovery algorithms. These include two "re-simulated" datasets obtained from artificially generated data from models trained with real data and two datasets including real variables intermixed with artificial variables (called probes). There is no time dependency in the samples. We chose applications in marketing, pharmacology and bio-medicine spanning

a high diversity of types of distributions. The datasets were used in two challenges in 2008 organized for the WCCI and NIPS conferences. A detailed technical report on the dataset design is available (Guyon et al., 2009). The website of the challenges remains open for post-challenge submissions (http://clopinet.com/causality).

Design:

We focused on some specific aspects of causal discovery:

Causality between random variables. We address causal relationships between random variables, as opposed to causal relationships between events, or objects.

No time dependency. Our everyday-life concept of causality is very much linked to time dependencies (the causes precede their effects). However, many machine learning problem are concerned with stationary systems or "cross-sectional studies", which are studies where many samples are drawn at a given point in time. Thus, sometimes the reference to time is replaced by the notion of "causal ordering". Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t + \delta t$, where δt can be made as small as we want. In practice, this means that the samples in our various training and test sets are drawn independently, according to a given distribution, which changes only between training and test set versions.[1]

Learning from observational data. Only training data from a "natural" pre-manipulation distribution (observational data) is available for training. In other settings, experimental data may be available as well. Relatively small training sets are provided, making it difficult to infer conditional independencies and learning distributions.

Discovering local causal relationships. We focus on one particular variable of interest called "target" and design tasks requiring to uncover the variables, which are most closely related (*e.g.*, direct causes and consequences, Markov blanket, depth 3 network). The problem of local causal relationships is closely related to that of variable selection: (1) variables closely related to the target in a causal graph may be highly predictive; (2) the knowledge of causal relationships is useful to select the variables, which will remain predictive in post-manipulation distributions.

Predicting the consequences of manipulations. There is no predictive task in the potluck challenge LOCANET tasks, but our datasets were previously used for prediction tasks in theWCCI 2008 "causation and prediction challenge" (Guyon et al., 2008). They include test samples drawn from a "natural" pre-manipulation distribution and test samples drawn from various post-manipulation distributions,

1. When manipulations are performed, we must specify whether we sample from the distribution before or after the effects of the manipulation have propagated. Here we assume that we sample after the effects have propagated.

which can be used to assess predictive performances of the target variable. Post-challenge submissions can be made online at http://www.causality.inf.ethz.ch/challenge.php.

The type of causal relationships under consideration have often been modeled as Bayesian causal networks or structural equation models (SEM) (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). In the graphical representation of such models, an arrow between two variables $A \rightarrow B$ indicates the direction of a causal relationship: A causes B. A node in of the graph, labeled with a particular variable X, represents a mechanism to evaluate the value of X given the parent node variable values. For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|Parents(X))$ while for structural equation models it is carried out by a function of the parent variables, plus some noise. Learning a causal graph can be thought of as a model selection problem: Alternative graph architectures are considered and a selection is performed, either by ranking the architectures with a global score (*e.g.*, a marginal likelihood, or a penalty-based cost function), or by retaining only graphs, which fulfill a number of constraints such as dependencies or independencies between subsets of variables. Bayesian networks and SEM provide a convenient language to talk about the type of problem we are interested in, but we made an effort to design tasks, which do not preclude of any particular model.

We have adopted two strategies to design datasets suitable for benchmarks:

- **Re-simulated data:** We train a causal model (a causal Bayesian network or a structural equation model) with real data. The model is then used to generate artificial training and test data for the challenge. Truth values of causal relationships are known for the data generating model and used for scoring causal discovery results.

- **Real data with probe variables:** We use a dataset of real samples. Some of the variables may be causally related to the target and some may be predictive but non-causal. The nature of the causal relationships of the variables to the target is unknown (although domain knowledge may allow us to validate the discoveries to some extent). We add to the set of real variables a number of distractor variables called "probes", which are generated by an artificial stochastic process, including explicit functions of some of the real variables, other artificial variables, and/or the target. All probes are non-causes of the target, some are completely unrelated to the target. The identity of the probes in concealed.

The LOCANET datasets include two re-simulated datasets and two real datasets with probes. They nicely complement each other: Re-simulated data provide us with full control over the data generative process and the truth values of all causal relationships, while real data with probes provide us with actual data distributions. The fact that truth values of causal relationships are known only for the probes affects the evaluation of causal discovery, which is less reliable than for artificial data.

Dataset description:

We formatted four datasets, including two re-simulated datasets (REGED and MARTI) and two real datasets with probes (CINA and SIDO). All datasets are thoroughly documented (including origin of the raw data, data preparation, past usage, and baseline results) in a Technical Report (Guyon et al., 2009). We briefly describe them:

REGED (REsimulated Gene Expression Dataset): The problem is to find genes, which could be responsible of lung cancer. The data are generated by a model derived from real human lung-cancer microarray gene expression data. From the causal discovery point of view, it is important to separate genes whose activity causes lung cancer from those whose activity is a consequence of the disease. The data include no hidden variable or missing data. The target variable is binary: it separates malignant samples (adenocarcinoma) from control samples (squamous).

SIDO (SImple Drug Operation mechanisms) contains descriptors of molecules which have been tested against the AIDS HIV virus. The target values indicate the molecular activity ($+1$ active, -1 inactive). The causal discovery task is to uncover causes of molecular activity among the molecule descriptors. This would help chemists in the design of new compounds, retaining activity, but having perhaps other desirable properties (less toxic, easier to administer). The molecular descriptors were generated programmatically from the three dimensional description of the molecule, with several programs used by pharmaceutical companies for QSAR studies (Quantitative Structure-Activity Relationship). For example, a descriptor may be the number of carbon molecules, the presence of an aliphatic cycle, the length of the longest saturated chain, etc.

CINA (Census Is Not Adult) is derived from census data (the UCI machine-learning repository Adult database). The data consists of census records for a number of individuals. The causal discovery task is to uncover the socio-economic factors affecting high income (the target value indicates whether the income exceeds 50K). The 14 original attributes (features) including age, workclass, education, marital status, occupation, native country, etc. are continuous, binary, or categorical. Categorical variables were converted to multiple binary variables (as we shall see, this preprocessing, which facilitates the tasks of some classifiers, complicates causal discovery).

MARTI (Measurement ARTIfact) is obtained from the same data generative process as REGED, a source of simulated genomic data. Similarly to REGED the data do not have hidden variables or missing data, but a noise model was added to simulate the imperfections of the measurement device. The goal is still to find genes, which could be responsible of lung cancer. The target variable is binary; it indicates malignant samples (adenocarcinoma) *vs.* control samples (squamous). The feature values representing measurements of gene expression levels are assumed to have been recorded from a two-dimensional microarray 32×32. The training set was perturbed by a zero-mean correlated noise model (?).

For the "causation and prediction challenge" (Guyon et al., 2008), the participants had to return predictions for the binary target variable on test data for three test set versions (version 0 from the unmanipulated distribution and versions 1, and 2 from the

manipulated distribution). For the "pot-luck challenge", the participants needed only the training data (the same in all three versions) to produce the local causal structure.

Task of the LOCANET challenge:

The participants were asked to provide a depth 3 causal network (oriented graph structure) around the target, using only training data only for causal discovery. The submission format is via a text file containing the list of parents of the features of interest. The target is numbered 0. All other features are numbered with their column number in the data tables. Provide a file named: `<yourlastname>_<dataname>_feat.localgraph`. Example `Guyon_LUCAS_feat.localgraph`:

```
0: 1 5
1: 3 4
2: 1
6: 5
8: 6 9
9: 0 11
11: 0 10
```

Evaluation:

The participants of LOCANET were ranked on the basis of an average edit distance to the true causal relationship between the target and variables in the depth three network. Specifically, we considered only local directed acyclic graphs and encoded the relationship of a variable to the target variable as a string of up (u) and down (d) arrows, from the target:
Depth 1 relatives: parents (u) and children (d).
Depth 2 relatives: spouses (du), grand-children (dd), siblings (ud), grand-parents (uu).
Depth 3 relatives: great-grand-parents (uuu), uncles/aunts (uud), nices/nephews (udd), parents of siblings (udu), spouses of children (ddu), parents in law (duu), children of spouses (dud), great-grand-children (ddd).

A confusion matrix C_{ij} was computed, recording the number of relatives confused for another type of relative, among the 14 types of relatives in depth 3 networks. A cost matrix A_{ij}, was applied to account for the distance between relatives (computed with an edit distance as the number of substitutions, insertion, or deletion to go from one string to the other, using the string description described above). The score of the solution was computed as:

$$S = \sum_{ij} A_{ij} C_{ij}$$

There are additional details on how to handle ties. We provide the Matlab code to compute this score (Guyon, 2009). For artificially generated data (REGED and MARTI), the ground truth for the target local neighborhood was determined by the generative

model. For real data with artificial "probe" variables (SIDO and CINA), we do not have ground truth for the relationships of the real variables to the target. The score was therefore computed on the basis of the artificial variables only.

After the challenge, we also computed other metrics of evaluation. For particular features subsets (parents, children, parents and children, Markov blanket[2], all relatives up to depth 2, all relatives up to depth 3), we computed precision and recall (*aka sensitivity or true positive rare*), defined as follow:
Precision: NumberGoodFound / NumberFound
Recall: NumberGoodFound / NumberGood.

We also evaluated the predictive power of the Markov blanket by training a reference classifier (linear ridge regression) and testing on unmanipulated test data.

Results and conclusions:

Ten participants entered the challenge. All the details of the analysis and fact sheets for some of the entries are available on-line at: http://www.causality.inf.ethz.ch/data/LOCANET.html.

The methods included: Structure learning using independence tests (Brown & Tsamardinos and Zhou, Wang, Yin & Geng), combinations of score-based and structure learning methods (de-Prado-Cumplido & Antonio Artes-Rodrigues and Tillman & Ramsey), combinations of feature selection and structure methods (Olsen, Meyer & Bontempi), and ensemble methods (Mwebaze & Quinn).

The edit distance scores of the participants were fairly poor. On REGED and MARTI, the best ranking entries were empty graphs. On CINA, the best ranking entry had results worse than the fully connected graph (with symmetric connections). On SIDO, the best result was barely better than that of the empty graph. From the point of view of the precision and recall metrics, structure learning methods gave the most promising results (highest precision), but all methods gave a poor recall, particularly for SIDO. We performed additional qualitative analyses in CINA using the semantics of the identifiers of the true variables to see whether the uncovered relationships made sense. It is unclear whether using the tools of causal discovery brought us a lot more information that simple correlation would have:

- most features cited as cause or effect of the target rank among the most correlated features,
- there is usually no consensus on the causal direction among the participants,
- when there is a large consensus on the causal direction, the result is sometimes suspicious given the semantics of the feature,
- a simple ranking in order of correlation yields nested feature subsets always more predictive than the Markov blanket.

2. We call Markov blanket the set of parents, children, and spouses of the target variable.

Overall these results point to the need to improve the reliability of causal discovery from observational data.

Acknowledgments

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional support was provided by the U.S. National Institute of Health under grant 2R56LM007948-04A1.

References

I. Guyon. Scoring code for the locanet tasks. http://www.causality.inf.ethz.ch/data/LocanetScoreCode.zip, October 2009.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*, volume 3, pages 1–33, WCCI2008 workshop on causality, Hong Kong, June 3–4 2008.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Datasets of the causation and prediction challenge. Technical Report, http://eprints.pascal-network.org/archive/00004566/, 2009.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

Judea Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, March 2000.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* The MIT Press, Cambridge, Massachusetts, London, England, 2000.

A Strategy for Making Predictions Under Manipulation

Laura Brown

Ioannis Tsamardinos

Participant name, address, email and website:

Laura Brown, Eskind Biomedical Library 4th floor, 2209 Garland Ave. Nashville, TN 37232 USA laura.e.brown@vanderbilt.edu http://www.dsl-lab.org

Ioannis Tsamardinos, FORTH-ICS, N. Plastira 100, Vassilika Vouton GR-700 13 Heraklion Crete, GREECE tsamard@ics.forth.gr

Task(s) solved: LOCANET

Reference:

- Bach's – Bach, F.R. and Jordan, M.I. NIPS, 2002
- Causality Challenge Submission – Brown, L.E. and Tsamardinos, I. "A Strategy for Making Predictions Under Manipulation" 2008
- MMHC – Tsamardinos, I. et al. Machine Learning, 2006
- MMPC, MMMB – Tsamardinos, I. et al. SIGKDD, 2003

Method:

Our submission for the LOCANET challenge relied on the results and procedures of the first causality challenge, from which the local networks were pruned. Details to the approach used for the first causality challenge are available in the paper for that challenge (available at the DSL website) but a general overview of the method and how the results are used for this task are presented.

Preprocessing: The preprocessing was tailored to each data set. For the REGED data set each variable was normalized so its mean was zero and standard deviation was one. For the SIDO data set, the variables were binary and no preprocessing was performed. For the CINA data set, variables that were not binary were treated as continuous and normalized; binary variables were all set to values of zero and one. For the MARTI data set, the preprocessed data by Dr. Guyon available on the challenge website was used.

Causal discovery: Once the initial data sets have been pre-processed, the next step of our procedure was to identify the skeleton structure of the Bayesian Network around the target variable recursively using the MMPC algorithm, up to three edges away from the target. This region of interest makes it practical to apply causal algorithms that cannot scale up to the sizes of all the networks in the challenge. The selection of a depth 3 in this case was for the previous challenge where we focused on identifying the Markov Blanket. For future work on the LOCANET challenge which focuses on learning a region out to depth 3, the MMPC recursion should run out to depth 4 and then be pruned back as a final step. In the next step of our analysis we tried to orient the edges of the region. For the case of continuous or mixed data, an adaptation of Bach's algorithm was used. For the case of binary data, MMHC was used to find the top scoring network. From the learned network, the region of depth 3 was extracted and submitted for analysis.

Feature selection: The recursive selection of variables to include in the region can be thought of as performing several iterations of feature selection.

Classification: None

Model selection/hyperparameter selection: Currently, a default set of parameters are used in the edge orientation procedure (parameters for score calculation either via BDeu score in MMHC or parameters for the kernel in Bach's algorithm). Future work for this challenge could also involve using many different parameters and perform model averaging over the results.

Results:

Table 1: Result table. The score of our method along with the top and lowest score for each data set are given. Three reference scores are also presented where applicable for comparison. The scores for REGED and MARTI are the second best submitted.

	REGED	SIDO	CINA	MARTI
Brown/Tsamardinos	0.27	3.46	2.23	0.36
Best Overall	0.22	3.31	1.70	0.21
Worst Overall	0.52	3.48	3.31	0.93
Reference A	0.01	0.64	0.64	0.02
Reference B	0.16	1.92	1.89	0.16
Reference C	3.08		1.67	3.01

Reference A: Truth graph with 20% of the edges flipped at random.
Reference B: Truth graph with connections symmetrized.
Reference C: Variables in the truth graph, fully connected.

Advantages:

The method gains in efficiency by rather than learning the entire network and pruning out the region it uses the recursive application of a local neighborhood identification method (MMPC) in a breadth-first search then orients the graph.

Limitations:

The results on CINA may be low because of the inappropriateness of the statistical tests used in MMPC for the mixed data. The MMPC algorithms have statistical tests provided for when the data is entirely binary or continuous (with a binary target); the mixed data set did not therefore match well to these methods. Also, as stated above the performance of the method may be improved by allowing the recursive procedure to run to a depth of 4 in order to better facilitate identification of all edges in the region of depth 3.

Implementation:

The methods are implemented in Matlab. The MMPC and MMHC algorithms are available from the Causal Explorer library, www.dsl-lab.org (please note, we were in part the developers of these methods and may have slightly extended or modified the code from the precise implementation available in Causal Explorer). Our method combined many algorithms and used the results from the previous challenge which are not available as a push-button application although the code and executables are available at the above website.

Keywords:

- **Preprocessing or feature construction:** normalization.
- **Causal discovery:** Bayesian Network,
- **Feature selection:** filter

PROMO Dataset

Jean-Philippe Pellet

Pot-luck challenge: Fact Sheet for the PROMO Dataset

Repository URL: http://www.causality.inf.ethz.ch/repository.php?id=2

Dataset name: PROMO

Title: Detecting simple causal effects in time series
Authors: Causality workbench team
Contact: Jean-Philippe Pellet, jep@zurich.ibm.com
Website: http://www.zurich.ibm.com/~jep/causality/promo.html

Key facts

This dataset contains artificial data about product sales and promotions as time series. There are 1000 binary promotions variables and 100 continuous product sales variables. The goal is to predict a 1000×100 boolean influence matrix, indicating for each (i, j) entry whether the ith promotion has a causal influence of the sales of the jth product.

Each of the 100 products has a defined seasonal baseline, repeating over the years. The seasonal effect can vary from almost inexistent to major. On top of this baseline are promotions. Each product is influenced by between 1 and 50 promotions out of the 1000 promotions available. Promotions usually increase the sales with respect to the baseline, but can occasionally reduce them (e.g., when a similar competing product is promoted, that promotion might have a negative effect on the sales of the current product). On top of that are daily variations.

Each of the 1000 promotions can be seasonal or not; i.e., they can have the same pattern from one year to another or be completely different. The average time a promotion stays active or inactive, however, is constant for each promotion.

The weighted normalized influence matrix is provided for result evaluation. It is normalized so that the maximum positive contribution is 1 and the maximum negative contribution is -1, and each nonzero (i, j) entry is weighted by how much promotion i affects product j.

Note that, as this matrix is provided, the participants are trusted to use it for evaluation purposes only, and not to tune potential hyperparameter of their approaches.

Keywords: time series, structural equation models

Data Generation

The data is generated in three steps:

1. Generate the 1000 promotion variables;
2. Generate the product baselines (without the promotion effect);
3. Generate the end product sales, including the promotion effect.

We denote promotion variables by P_i, $1 \leq i \leq 1000$, and the baselines and product sales by B_j and S_j, respectively, $1 \leq j \leq 100$. The value generated for variable P_i on day t is denoted by p_{it}.

The promotion variables are all generated according to a Markov chain whose parameters are randomly chosen. The Markov chain has two states, ON and OFF. The two transition probabilities are determined by the inverse of the average number of days in each state t_{ON} and t_{OFF}, which are drawn from a probability distribution covering from 1 day to 300 days This fully determines the Markov chain:

$$p_{\text{ON}\to\text{OFF}} = 1/t_{\text{ON}} \qquad p_{\text{OFF}\to\text{ON}} = 1/t_{\text{OFF}}$$
$$p_{\text{ON}\to\text{ON}} = 1 - 1/t_{\text{ON}} \qquad p_{\text{OFF}\to\text{OFF}} = 1 - 1/t_{\text{OFF}}.$$

Then, for each promotion variable, with probability 0.5, it is set to repeat each year in the same pattern as the previous year, and with probability 0.5, not to repeat automatically. In the former case, a full year (i.e., 365 values, one for each day) is sampled by determining the state of the variable according to the Markov chain, and then replicated twice, to obtain the time series over 1095 days. In the latter case, the full 1095 days are sampled with the Markov chain, resulting with high probability in different sequences for each year. In each case, the initial state is determined to be ON with probability $p_{\text{ON}} = t_{\text{ON}}/(t_{\text{ON}} + t_{\text{OFF}})$, and accordingly off with probability $p_{\text{OFF}} = 1 - p_{\text{ON}}$. This is show in Figure 1.

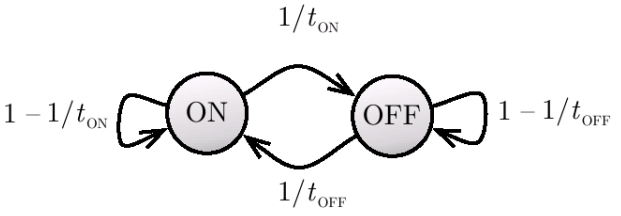


Figure 1: The Markov chain generating the promotion variables

$$\forall i : p_{it} = \text{time series sampled with Markov chain}$$

Product baselines are the sum of a constant factor c_j and of a seasonal effect. The seasonal effect repeats over the years. The baselines indicate what the sales would be, without promotions and without random noise. The constant factor is drawn randomly,

and the seasonal effect is determined as a superposition of n sines whose amplitude α_k phase ϕ_k, and pulse ω_k are drawn randomly. The number of sines n is drawn uniformly between 2 and 10. The seasonal effect is then shifted so that its minimum is 0. This is indicated with the shift$(\cdot)$ function, which we define as $\text{shift}(x_t) = x_t - \min_{t'} x_{t'}$.

$$\forall j : b_{jt} = c_j + \text{shift}\left(\sum_{k=1} n\alpha_k \sin(\omega_k \cdot t/365 - \phi_k)\right)$$

The end sales are generated as follows: for each product, a set $\mathbf{I}_j$ of influencing promotion variables is drawn at random, with its cardinality m uniformly distributed between 1 and 50. The influence f_{jl} of each influencing promotion I_{jl}, $1 \leq l \leq m$, is drawn randomly between 0.2 and 0.8, and negated with probability 0.1. For each day, the total promotion factor τ_{jt} is determined as the square root of the sum of the factors of all influencing promotions whose state is ON. Random Gaussian noise with mean 0 and standard deviation 0.1 is then added to this promotion factor. The end sales are then the product baseline multiplied by the total noisy promotion factor (not that this means that the promotion effect is thus multiplicative rather than additive).

$$\begin{aligned}
\forall j : \mathbf{I}_j &= \text{random set of } m \text{ promotion variables} \\
\forall j, m : f_{jl} &= \text{factor of influence for the } l\text{th promotion in } \mathbf{I}_j \\
\forall j : \tau_{jt} &= \sqrt{\sum_{l=1} m f_{jl} \cdot \mathbf{1}_{p_{\text{ind}_j(l),t}=1}} \\
\forall j : u_{jt} &= t \text{ realizations of a variable } U \sim \mathcal{N}(0,1) \\
\forall j : s_{jt} &= b_{jt} \cdot (\tau_{jt} + u_{jt})
\end{aligned}$$

The value of $\mathbf{1}_{p_{\text{ind}_j(l),t}=1}$ is 1 whenever the lth promotion for product j is ON on day t, and 0 otherwise (the notation $\text{ind}_j(l)$ just converts the product-specific promotion index l for product j to the global, product-independent promotion index).

The final data available to challenge participants are the end sales s_{it} and the promotion variables p_{it}; all other intermediary values remain hidden.

Discussion

There are several ambiguities in the data. For instance, all promotions that repeat year-to-year can be seen as seasonality. Further assumptions are needed here to tell if some observed recurring effect is due to seasonality or to a seasonal promotions. Another problem is that some promotion with a nonzero effect might be ON or OFF all the time, preventing learning algorithms from assessing its effect.

These points are deliberate and correspond to real-life scenarios. Often, products both have a seasonality, and often, the promotions applied to these products in the past also had a certain seasonality. It is therefore important to include an appropriate criterion for to tell these two effects apart. It is also necessary to have an algorithm that can correctly identify promotions whose effect cannot be assessed.

Note that the promotion effect is straightforwardly applied to the end sales: only the current day is used. A given promotion can only have an impact the day it is ON; the sale history has no memory of past promotions. This information was not given to the challenge participants.

Approaches Used by Participants

Two approaches were proposed to solve the PROMO task. They are briefly summarized here; more details can be found on their respective fact sheets at http://www.clopinet.com/isabelle/Projects/NIPS2008/home.html.

The first approach, A_1, first tries to extract the baseline by modeling it as an offset-plus-sine for each product, to which is then added the promotion effect:

$$\forall j : b_{jt} = c_j + \alpha_j \sin(\omega_j t + \phi_j)$$
$$\forall j : s_{jt} = b_{jt} + Bu_t,$$

where B is the influence matrix and u_t represents the state of the promotions. This is solved in two steps: first, the parameters c_j, α_j, ω_j, ϕ_j are estimated by fitting the data with the offset-plus-sine model; then, fixing those parameters to the obtained value, B is estimated solving j independent convex problems, subject to a sparsity constraint on B: for each promotion, the number of nonzero entries in B should not be greater than 50 (The number 50 is given in the problem description as upper bound on the number of relevant promotion variables). See Markovsky (2008) for more details as well as the whole source code to reproduce the results listed below.

The second approach, A_2, also consists of two steps, where first the seasonal component is removed, and then the relevant promotion variables are determined. The baseline is modeled as a constant plus a superposition of 16 sines and cosines with different frequencies. Denote a design matrix $Z = [z_1, z_2, \cdots, z_{1095}]^T$, where

$$z_t = (1 \quad \sin(2\pi t/365) \quad \cos(2\pi t/365) \quad \cdots \quad \sin(10\pi t/365) \quad \cos(10\pi t/365))^T,$$

then the baseline is estimated as $\hat{B} = (b_{jt}) = Z(Z^T Z)^{-1} Z^T S$, where $S = (s_{jt})$ is the matrix containing the end sales. The input to the second step of the method is the residuals of this regression, namely $Y = S - B$. The second step selects the relevant promotion variables for each product: this is done with an iterative stepwise selection. The hyperparameters of this selection is then chosen according to an EBIC criterion. See Yin et al. (2008) for more details about this method.

Results

To compare the results of the participants, we used the following metrics: for each of the 100 products, we determine the precision, recall, and F-score of the participants' solution.

The *precision* is a real value between 0 and 1 determining, out of the set of promotion variables proposed by a participant as influencing product j, what proportion of

them are actually promotion variables that were in $\mathbf{I}_j$; i.e., which were also used in the generating model to determine the end sales. The precision for product j is then:

$$pr_j = \frac{\text{number of correctly identified promotion variables}}{\text{total number of identified variables}}.$$

The *recall* is a also a real value between 0 and 1 determining how complete the participants' solution were. It is defined similarly as:

$$re_j = \frac{\text{number of correctly identified promotion variables}}{\text{total number of promotion variables used in the generating model}}.$$

A perfect solution has precision = recall = 1. A solution with precision = 1, recall = 0.5, for instance, means that all identified promotion variables were indeed correct, but that they only constituted 50% of those actually used in the generating model. Conversely, a solution with precision = 0.5 and recall = 1 is such that although all relevant variables were identified, 50% of all identified variables were not used by the generating model.

Finally, the F-score is the harmonic mean of precision and recall:

$$F_j = \frac{2 \cdot pr_j \cdot re_j}{pr_j + re_j}.$$

For the two participants, using approaches A_1 and A_2, the precision, recall, and F-score was evaluated for each product. Table 1 shows the mean and standard deviation of those measures aggregated over all products.

Table 1: Mean and standard deviation of the precision, recall, and F-score for the two participants

	A_1 (Markovsky, 2008)	A_2 (Yin et al., 2008)
Precision	0.38 ± 0.24	0.89 ± 0.14
Recall	0.32 ± 0.23	0.78 ± 0.17
F-score	0.31 ± 0.19	0.82 ± 0.13

Clearly, A_2 performs much better, getting twice as good both precision and recall. This can be due to a number of reasons: probably, extracting a baseline as a superposition of several sines and cosines rather than a single sine can better recover the original baseline as generated by the model, as the model used a superposition of sines with different amplitudes, phases, and pulses. The residuals obtained after baseline extraction by A_1 still contain a bigger part of the seasonal components than the residuals obtained by A_2. Taking in more promotion variables to try and compensate for a baseline detection that could be better then lowers the precision, while at the same time, not detecting the baseline correctly will tend to lower the recall, as it becomes less likely to be able to make out well the effect of the truly influencing promotion variables.

References

I. Markovsky. Results on the PASCAL challenge "Simple causal effects in time series". Technical report, University of Southampton, 2008.

J. Yin, S. Wang, W. Deng, Y. Hu, and Z. Geng. Iterative stepwise selection and threshold for learning causes in time series. Technical report, Peking University, 2008.

PASCAL PROMO Challenge

Ivan Markovsky

Pot-luck causality challenge: FACT

Title: Results on the PASCAL PROMO challenge

Participant name, address, email and website:
Ivan Markovsky
Building 1, Highfield campus
Southampton, SO17 1BJ, UK

Telephone: +44 (0)23 8059 8715
Fax: +44 (0) 23 8059 4498
Email: im@ecs.soton.ac.uk
WWW: http://users.ecs.soton.ac.uk/im/homepage.html

Task(s) solved: PASCAL PROMO challenge

Reference: http://eprints.ecs.soton.ac.uk/16779/

Method:

The data is modeled as a sum of a constant-plus-sin term and a term that is a linear function of a small number of inputs. The problem of identifying such a model from the data is nonconvex in the frequency and phase parameters of the sin and is combinatorial in the number of inputs. The proposed method is suboptimal and exploits several heuristics. First, the problem is split into two phases: 1) identification of the autonomous part and 2) identification of the input dependent part. Second, local optimization method is used to solve the problem in the first phase. Third, l1 regularization is used in order to find a sparse solution in the second phase.

Results:

Please refer to the technical report (http://eprints.ecs.soton.ac.uk/16779/) for table with results. In addition, the web page has a link to Matlab software that reproduces the presented results.

Comment about the following:

- **quantitative advantages** (e.g. compact feature subset, simplicity, computational advantages)

  The algorithm is computationally simple: the full model is identified in 3 hours on a standard PC.

- **qualitative advantages** (e.g. compute posterior probabilities, theoretically motivated, has some elements of novelty).

  The tools used to solve the subtasks (leading to the full identification method) are not new however their combination and application for causality detection is novel.

Briefly explain your implementation.

We use Matlab. The one variable nonconvex optimization problem is solved using the Optimization Toolbox (fminsearch function) and the L1 optimization problem is translated to a standard convex optimization problem, using CVX (http://www.stanford.edu/~boyd/cvx/).

Provide a URL for the code (if available).

http://eprints.ecs.soton.ac.uk/16779/2/challenge.tar

Precise whether it is a push-button application that can be run on benchmark data to reproduce the results, or resources such as modules or libraries.

1. Unpack the archive (it creates a directory called "challenge").
2. Download and unpack in the same directory the challenge data

   http://www.zurich.ibm.com/~jep/causality/PROMO.zip
3. If not already installed, download and install CVX

   http://www.stanford.edu/~boyd/cvx/
4. Make sure that the Optimization Toolbox of Matlab is installed.
5. Change directory to "challenge" and run the function "test" from the Matlab command line. The model is identified in approximately 3 hours and the results reported in paper (figures and numerical data) are available.

Keywords: Put at least one keyword in each category. Try some of the following keywords and add your own:

- **Preprocessing or feature construction:** redundant input removal.

- **Causal discovery:** prediction, least squares fitting.
- **Feature selection:** L1 norm regularization.

Iterative Stepwise Selection and Threshold for Learning Causes in Time Series

Jianxin Yin JIANXINYIN@MATH.PKU.EDU.CN
Shaopeng Wang WANGSHOP@GMAIL.COM
Wanlu Deng SHIRLEYPKU@GMAIL.COM
Ya Hu TERESAHU@PKU.EDU.CN
Zhi Geng ZGENG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing 100871, China

Abstract

When we explore the causal relationship among time series variables, we first remove the potential seasonal term then we deal with the problem in the feature selection framework. For a time series with seasonal term, we use several sequences of $\sin(t)$ and $\cos(t)$ functions with different frequencies to design a 'pseudo' design matrix, and the seasonal term is removed by getting the regression residual of the original series on this 'pseudo' design matrix. An iterative stepwise subset selection and threshold method are then applied. The cut-value for the threshold is selected by an EBIC criterion. Some simulations are performed to assess our method. In the PROMO task of the Potluck challenge, we apply our method and obtain a specificity of above 77% while keep the sensitivity of around 89% on the PROMO task.

Keywords: functional data, iterative threshold, linear model, seasonal term, stepwise subset selection, structural equation model, time series.

1. Introduction

In the Potluck challenge, we try to select the causal processes from the 1000 promotions for each product sales series separately. For a time series $Y(t)$ with seasonal terms, we can decompose it into three parts:

$$Y(t) = S(t) + T(t) + N(t) \tag{1}$$

where $S(t)$ denotes the seasonal term, $T(t)$ denotes the trend term which may be influenced by the other processes and $N(t)$ is the noise part. There exist many methods to model the seasonal term $S(t)$ in the literature (see Brockwell & Davis, 1991; Box et al., 1994). Here we treat $S(t)$ as a continuous periodic function and approximate it by a series of periodic functions bases in the sin and cos functions. We use a linear structure equation model (SEM) to model the other processes' causal influence on the target process. That is,

$$T(t) = \beta_0 + \beta_1 * x_1(t) + \cdots + \beta_p * x_p(t) \tag{2}$$

where $x_i(t)$, $i = 1, \cdots, p$ stands for the other processes which may also be influenced by the seasonal factor. So we also apply the same model of seasonal term on each $x_i(t)$ process. Equation (2) is then treated as a simple linear regression model and a stepwise selection (Weisberg, 1985) procedure is applied to screening out the influential independent variables. This stepwise selection is applied iteratively to eliminate the possible "boundary" variables (see the next section). To get a sparse model, we further use a kind of threshold on the regression coefficients to select the significant subset. This procedure is also applied iteratively to get a converged subset. And the hyper-parameter of the cut-point is selected by an extended BIC criterion. Some simulation study shows that our method can get the consistent result under different kinds of $S(t)$ and $N(t)$ process. This paper is organized as following. In Section 2, the preprocessing for the seasonal term is described, in Section 3 the stepwise selection procedure is mentioned and the iterative threshold method with its hyper-parameter selection method is introduced. Section 4 is the numerical study. Finally Section 5 gives some discussion on our method.

2. Preprocessing: Filtering the seasonal term

For a given time series, we use a series of continuous periodic functions to filter out the seasonal term. Suppose that the period length is T while in our problem, $T = 365$. Then we generate the periodical sequences $\sin(2\pi t/k), \cos(2\pi t/k)$, for $t = 1, \cdots, 1095$ and $k = T, T/2, T/3, T/4, T/5$. It is obvious that the period for each sequence is k. Denote a design matrix $Z = [z_1, z_2, \cdots, z_{1095}]^T$, where

$$z_t = (1 \quad \sin(2\pi t/T) \quad \cos(2\pi t/T) \quad \cdots \quad \sin(10\pi t/T) \quad \cos(10\pi t/T))^T$$

$S = (S(1), \cdots, S(1095))^T$ is estimated as $\hat{S} = Z(Z^TZ)^{-1}Z^TY$. Then we remove the seasonal term expressed in the regression value on this design matrix to get the residual as the input of our next analysis. $Y^* = Y - \hat{S}$, $x_i^* = x_i - Z(Z^TZ)^{-1}Z^Tx_i$ for $i = 1, \cdots, 1000$. Here Y is the realization of 1095 days of certain product sales in our problem and x_i is the realization of 1095 days for some promotion method. To simplify the notation, we still use the Y for Y^* and x_i for x_i^* respectively. The k is selected up to $T/5$ is determined by experience. We assume that the continuous periodic seasonal function can be approximated well enough by its Fourier expansion up to the fifth order.

3. Feature Selection

Since we have reduce our time series problem into a simple linear feature selection problem after removing the seasonal term, we can omit the subindex t and write our model

$$y = X_p\beta_p + \varepsilon \tag{3}$$

where p is the dimension of the original feature space.

3.1. Iterative stepwise subset selection

We use the stepwise selection (SW) to select the influential $x_i(t)$s for each $Y(t)$ through relation (1) and (2). When the significant level for entering (*penter*) is different from the one for removing (*premove*), there exists certain situation that some features on the "boundary" (here we mean that the significant level is between the *penter* and *premove*) can be dropped in the next round of stepwise selection on the remained feature set. For example, if x_2 is only significant for the response when x_1 is in the model; suppose x_1 enters first, then x_2 can enter, but later x_1 is removed and x_2 is not removed (if it is on the "boundary"). For the next round of SW selection, x_2 will be removed. So for the purpose of sparsity, we use the *stepwisefit* procedure in Matlab iteratively to select the feature set with *penter* = 0.05 and *premove* = 0.1.

3.2. Iterative threshold selection

Before the following analysis, each column of X_p is standardized to have zero mean and unit variance. Suppose that the true model has a dimension d and denoted as X_d, then the above relation (3) can be represented as

$$y = X_d \beta_d^* + \varepsilon \tag{4}$$

where $\beta_d^* = \{\beta_j : \beta_j \neq 0, 1 \leq j \leq p\}$ with a dimension $d < p$. Initialize $\mathcal{M}^{(0)}$ as the output of the iterative SW selection. $\mathcal{M}^{(i)}$ is obtained in an iteratively manner:

$$\mathcal{M}^{(i)} = \left\{ 1 \leq j \leq \|\mathcal{M}^{(i-1)}\| : |\hat{\beta}_j^{(i-1)}| \geq \alpha^* \max_{1 \leq k \leq \|\mathcal{M}^{(i-1)}\|} \left(|\hat{\beta}_k^{(i-1)}| \right) \right\} \tag{5}$$

where $\hat{\beta}^{(i-1)}$ is the least square estimate for regression coefficient vector of y on $X_{\mathcal{M}^{(i-1)}}$ and $\|\cdot\|$ denotes the cardinality of a set. Intuitively, we drop those features whose absolute values of regression coefficients are smaller than $\alpha * 100\%$ of the current largest one (in absolute value).

Denote the true feature set as $\mathcal{M}_T = \{1 \leq j \leq p : \beta_j \neq 0\}$. And use the note $|\beta|_{\min} = \min_{1 \leq j \leq p} |\beta_j|$. In order to justify our selection procedure, we need the following two assumptions on the underlying model.

- *Assumption1* There exists a constant number $c_0 > 0$, such that $|\beta^*|_{\min}/|\beta^*|_{\max} \leq c_0$
- *Assumption2* For any sub-model of the true model $\mathcal{M}^s \subset \mathcal{M}_T, |\beta^s|_{\min}/|\beta^s|_{\max}, c_0$, where c_0 is the same constant in assumption 1.

Remark 1 *Assumption 1 says that the ratio of the two extremes of the true coefficients is significantly apart from 0. Assumption 2 want to regulate the behavior of the load on every feature subset. It's not the possible weakest requirement.*

Under the above assumptions, we have the following results.

Theorem 2 *Suppose assumptions 1 and 2 are true, under our model setup (3)–(5), then with probability tending to one (as $n \longrightarrow \infty$) that there exists a constant α such that*
(I) If $\mathcal{M}^{(i)} \supsetneq \mathcal{M}_T$, then $\|\mathcal{M}^{(i+1)}\| < \|\mathcal{M}^{(i)}\|$.
(II) If $\mathcal{M}^{(i)} \subseteq \mathcal{M}_T$, then $\|\mathcal{M}^{(i+1)}\| = \|\mathcal{M}^{(i)}\|$.

Proof *The α can be chosen as a positive number that $0 < \alpha < c_0$, where $c_0 = |\beta^*|_{\min}/|\beta^*|_{\max}$. For case (I), since the estimate $\hat{\beta}^{(i)}$ is an unbiased consistent estimator for β_p, then with probability tending to one $|\hat{\beta}^{(i)}|_{\max} \approx |\beta^*|_{\max}$ and $|\hat{\beta}^{(i)}|_{\min} \approx 0$. Then $|\hat{\beta}^{(i)}|_{\min}/|\hat{\beta}^{(i)}|_{\max} < \alpha$, so at least we can drop one variable. Similarly, for case (II), with assumption 2, we can see that the same α is also appropriate here for $|\hat{\beta}^{(i)}|_{\min}/|\hat{\beta}^{(i)}|_{\max} > \alpha$* ■

From the above theorem, one can see that from an over-fit model including the true features, threshold on the regression coefficients can remove the unrelated features and the iteration can repeat this process until it converges to the true model or its subset. And it will not continue to delete variables as long as it is covered by the true subset. But what is the case when we begin from a subset that $\mathcal{M}^{(i)} \supsetneq \mathcal{M}_T$? There is no assertion can be made here, but from the experience we have, the iteration converges in finite steps under this case.

3.3. Extended-BIC criterion for model selection

Ordinary BIC is likely inconsistent when $p > \sqrt{n}$ (Chen & Chen, 2008). We used the extended-BIC (EBIC) criterion (Chen & Chen, 2008) to select the hyper-parameter α. From the simulation study in the next section, we can see that the EBIC outperforms the ordinary BIC and prediction MSE criterion in measure of SN. The extended-BIC is defined as:

$$EBIC(\mathcal{M}) = \log(\hat{\sigma}^2_{(\mathcal{M})}) + n^{-1}\|\mathcal{M}\| \times (\log n + 2\log p)$$

Then we search the minimum value for this EBIC ($\mathcal{M}_\alpha$) index by α in an interval. Denote $\hat{\mathcal{M}} = \arg\min_{\alpha\in[a,b]} EBIC(\mathcal{M}_\alpha)$. In practice, we choose $[a,b] = [0.1, 0.3]$.

4. Numerical Studies

In the simulation study, a linear additive model of (1) is considered. We consider three types of $S(t)$, extra lag-effect in the relationship between $T(t)$ and $x_i(t)$s rather than (2), and an ARMA noise with an acceptable signal-to-noise ratio for $N(t)$.The simulation result shows that our approaches have a good robust performance although we never take into account the lagged effect in $T(t)$ and ARMA in $N(t)$. We use the specificity (shorted as SP) and sensitivity (shorted as SN) to evaluate. To write them out explicitly,

$$SP = \frac{\#\{j : \hat{\beta}_j \neq 0 \& \beta_j \neq 0\}}{\#\{j : \beta_j \neq 0\}} \qquad SN = \frac{\#\{j : \hat{\beta}_j \neq 0 \& \beta_j \neq 0\}}{\#\{j : \hat{\beta}_j \neq 0\}}$$

4.1. Simulation method

We simulate the similar model configurations compared to our PROMO task under model (1). For the $S(t)$, we select three types of periodic continuous function to represent it.

Type(1) $S(t;n,m,\phi_1,\phi_2,T) = \sum_{i=1}^{n} \sin(2\pi it/T + \phi_1) + \sum_{j=0}^{m} \cos(2\pi jt/T + \phi_2)$

Type(2) $S(t;a_1,a_2,a_3,T) = t(T-t)[(t-a_1)(t-a_2)(t-a_3) - 200]$.

Type(3) Twice moving average of a random $N(0,\sigma^2)$ series with smoothing window width h which is sampled from [50, 80] uniformly.

We generate a total of 50 $S(t)$s for each seasonal type on the domain $t \in [0,365]$ and then extend them to span in $[0,1095]$ periodically. The parameters in the three types are randomly assigned except T which is set to 365.

For $T(t)$, firstly we generate 1000 binary series $x_i(t)$ for $t = 1,\cdots,1095$, $i = 1,\cdots,1000$ which are potential causes of each $Y(t)$ series through the relation (1) and (2). We generate 500 seasonal $x_i(t)$s (with period 365) and 500 non-seasonal $x_i(t)$s. For each point in $x_i(t)$, a binary number is generated from a binomial distribution $B(p)$ where p is sampled uniformly on $[0.2,0.8]$. The seasonal one is the triplication of the function defined in $[0,365]$. Then the coefficients β in (2) are defined as following: the number of non-zero β_j is uniform sampled from $[1,50]$ while the non-zero values are sampled uniformly from interval $[0.4,1]$ and assigned a negative sign with probability 0.2. Besides, we suppose covariates $x_i(t)$s may have lagged effect on the target $T(t)$, which often exists in the real world. Inspired by the power decay lagged effect model in Box & Tiao (1975), we use the backward operator $\omega * \sum_{lag}(\frac{1}{2}B)^{lag-1}$ on $x_i(t)$s, where ω represents the lagged effect coefficients randomly taking a smaller absolute value than the corresponding major effect but a possible opposite sign. The lag length lag is randomly selected in $\{0,1,2,3,4,5\}$. We add the lagged effect to the $T(t)$.

We model the noise term $N(t)$ as an ARMA model whose parameter (p,d) is randomly assigned. Finally, we get the simulated target series $Y(t)$ from the summation of the above three series multiplied with appropriate scale parameters, such that the signal-to-noise rate is around 4. Then we apply an iterative SW procedure followed by an iterative threshold procedure proposed above to select the significant subset. Different kinds of criterion for α are compared in their accuracy measures.

4.2. Simulation results

Table 1 tells us that the model selected by EBIC is comparably good in SP while is over-whelmingly better than BIC and MSE criterion in SN.

From Table 2 we can see that the iterative SW process is necessary for there are more than 30% of the case that there are 'boundary' variables in our selected model. Table 3 also supports that the iteration for Threshold process is necessary although they finally converge with large probability.

Table 1: Comparison of three criterions.

Seasonal Function	**criterion**	**SP**	**SN**
Type (1)	BIC	0.98	0.76
	EBIC	0.96	0.99
	MSE	0.99	0.59
Type (2)	BIC	0.96	0.57
	EBIC	0.92	0.94
	MSE	0.97	0.46
Type (3)	BIC	0.96	0.57
	EBIC	0.93	0.93
	MSE	0.96	0.47

Table 2: Iteration number distribution for iterative SW process.

Seasonal Function	1	2	3	$\geqslant 4$
Type (1)	0.7	0.3	0	0
Type (2)	0.58	0.38	0.04	0
Type (3)	0.62	0.36	0.02	0

Table 3: Iteration number distribution for the iterative Threshold process.

Seasonal Function	**criterion**	1	2	3	4	5	6	7	$\geqslant 8$
Type (1)	BIC	0	0.14	0.28	0.22	0.06	0.14	0.04	0.12
	EBIC	0	0.24	0.38	0.24	0.12	0	0	0.02
	MSE	0.38	0.26	0.14	0.08	0.04	0.06	0.02	0.02
Type (2)	BIC	0	0.06	0.26	0.28	0.12	0.08	0.02	0.18
	EBIC	0	0	0.44	0.34	0.16	0.04	0.02	0
	MSE	0.44	0.08	0.24	0.12	0.04	0.02	0.02	0.04
Type (3)	BIC	0	0.04	0.24	0.26	0.16	0.18	0.06	0.06
	EBIC	0	0.04	0.28	0.38	0.1	0.14	0.02	0.04
	MSE	0.42	0.12	0.12	0.16	0.02	0.1	0.04	0.02

4.3. Results on the PROMO task

We apply our algorithm to the PROMO task and get a specificity of above 77% and sensitivity around 89%.

5. Discussion

One may doubt that whether only the iterative threshold process can do the variable selection job well enough. When feature space is of high dimension, from our experience, only the iterative threshold without stepwise selection can lead to terrible results. The *stepwisefit* procedure in Matlab is very efficient in computation, and the computation for one time of iterative threshold can be negligible.

References

Box, G.E.P., and G.C. Tiao Intervention analysis with applications to economic and environmental problems, *J. Amer. Statist. Assoc.*, 70, 70–79, 1975.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C., *Time Series Analysis: Forecasting and Control, 3rd Edition*. Pearson Education Asia Ltd., 1994.

P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods (Second Edition)*. Springer-Verlag New York, Inc., 1991.

Chen, J. and Chen, Z., Extended Bayesian information criterion for model selection with large model spaces, *Biometrika*, 95, 759–771, 2008.

S. Weisberg, *Applied Linear Regression (Second Edition)*. John Wiley & Sons, Inc., 1985.

Manufacturing data: SEMI tool level fault isolation

Advanced Analytics, Intel, LTD EUGENE.TUV@INTEL.COM
5000 W. Chandler Blvd
CH5-295
Chandler, AZ, 85226, USA

Background

In semiconductor manufacturing process the basic manufacturing unit is a silicon disk called wafer. During the fabrication process each wafer goes through a product (chip type) specific sequence of operations (hundreds). Each operation in the sequence is identified by its operation number. Some of these operations include adding a layer to the wafer, drawing a pattern on the wafer, covering the pattern with a photo layer, etching the pattern, etc. Wafers travel through manufacturing line in batches or lots. Every lot goes through each operation in the sequence. At each operation a lot could go through only one of many tools performing the same function. Maximum number of tools ~ 25, and the number of tools could be different from operation to operation. At the end of the manufacturing line many performance metrics are measured to monitor deviations from the desired target specifications. Often observed variation of a performance metric is caused by a subset of tools with effects of the problematic tools potentially changing in time.

1. Problem statement

The simulated dataset closely reproduces the nature and complexity of the tool level fault isolation problem engineers face in the semiconductor manufacturing. It records every tool and time stamp at every operation every lot went through (predictors), and the corresponding numeric performance measure (target). The goal is to recover a small subset of influential/problematic operations/tools and the corresponding contributions in time (if the effect is not constant) to the variation of the numeric performance metric. Examples of problematic tools generating non-constant offsets are shown on the figures 1, 2.

1.1. Data generative model — regression

$$ObservedPerformanceMetric(t) = TargetedPerformance + \sum_{i_j \in I} OFFSET_{i_j}(t) + \sum_{k_l, r_s \in M} OFFSET_{k_l, r_s} + \varepsilon$$

where ObservedPerformanceMetric(t) is observed at the end of line performance metric; TargetedPerformance is targeted by the process specification performance metric; I is a subset of operations where tool j at operation i causes $OFFSET_{i_j}(t)$ from the performance target; M is a subset of operations (different from I) where tools l and s from operations k and r produce a constant $OFFSET_{k_l,r_s}$ (pure interactions). The noise ε was generated from the normal distribution with zero mean and variance adjusted to give a 1/1 signal-to-noise ratio for a tool (or combination of tools) with the weakest signal.

$$E|\varepsilon| = \min_i E_t |OFFSET_i(t) - \operatorname{median}_t OFFSET_i(t)|$$

1.2. Data description and desirable results

Commonalityx4000 dataset has 602 variables and 4000 observations (lots); *RES* is the target — the performance metric measured at the end of line; LOT coded as LOTID (to be ignored); the rest are predictors: $LOCN_i$ and $TDATE_i$. Every lot goes through each of 300 operations: $LOCN_i$ (operation ID) at time $TDATE_i$, i=1–300. At each operation it could go through only one of the tools. Hence $LOCN_i$ are categorical predictors with number of levels= number of tools used, $TDATE_i$ are numeric variables (coded times through operation-tool). Approximately 25% of the data is missing.

The desirable result of the study is to identify problematic operations/tools and the corresponding offset patterns in time. The performance metrics for the evaluation of submissions will include the number of correctly identified operations/tools and number of false positives. Furthermore, to quantify the accuracy of offset pattern predictions the following metric will be used

$$\sum_{i=1:300} \sum_{j=1:4000} |PredictedOffset_i(tool_j, time_j) - ActualOffset_i(tool_j, time_j))|$$

It is expected that submission would have at most 50 identified influential operations (the actual number is smaller), the rest of the operations will be assumed having no effect ($OFFSET(tool, time) \equiv 0$). The submitted prediction matrix would have at most 50 columns corresponding to $OFFSETS_i(tool, time)$ caused by a subset of tools at operation i calculated for 4000 observations (lots) from the provided dataset. Thus the metric above will be evaluated over union of actual $OFFSETS_a$ and predicted $OFFSETS_p$. Finally, submissions would include identified pairs of pure operation/tools interactions (no time effect).

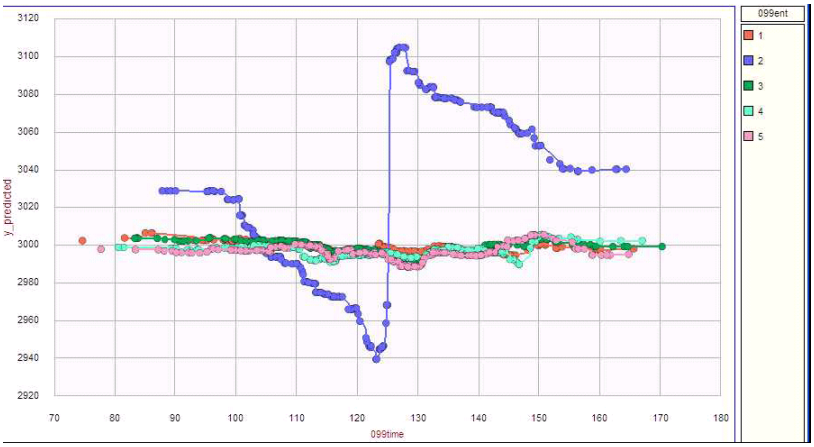


Figure 1: Sawtooth offset pattern caused by the tool=2 at operation=099. The rest of the tools stayed on target

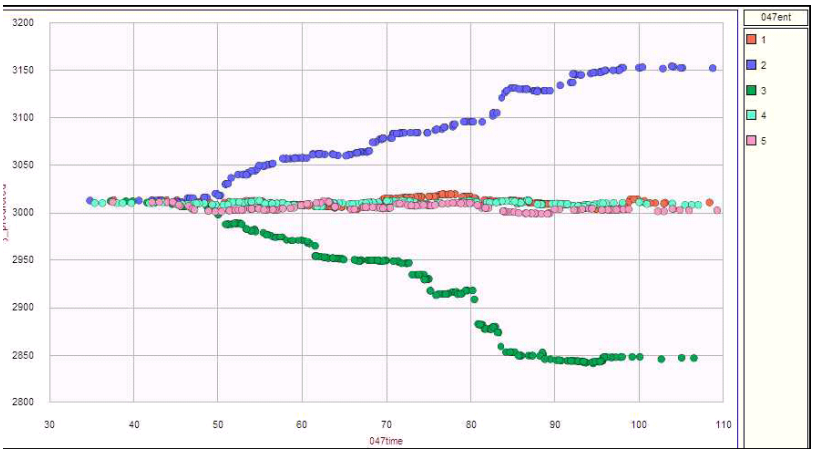


Figure 2: Trend offset patterns caused by the tools=2,3 at operation=047. The rest of the tools stayed on target

Pot-luck challenge: TIED

Advanced Analytics, Intel, LTD EUGENE.TUV@INTEL.COM
5000 W. Chandler Blvd
CH5-295
Chandler, AZ, 85226, USA

Task(s) solved:

- Using training data, find all minimal sets of features with optimal predictivity
- For each of the feature set identified, build a classifier model of the target variable using training data and apply it to the testing data.

Method:Rule induction on relevant features

Feature selection method (ACE - Artificial Contrasts with Ensembles) was used to remove irrelevant features. Two rule induction techniques were used to find sets of features with optimal predictability: CART with surrogate splits and a supervised APRIORI. Both point to the same optimal sets of features.

- Feature selection: ACE is a combination of three ideas: A) Estimating variable importance using RF ensemble of trees of a fixed depth (3–6 levels) with the split weight re-estimation on OOB samples (gives more accurate and unbiased estimate of variable importance in each tree), B) comparing variable importance against artificially constructed noise variables using a formal statistical test, and C) Iteratively removing the effect of identified important variables to allow detection of less important variables. ACE method is outlined in Tuv et al. (2006). The more comprehensive paper is submitted to JMLR (currently under review).

  The results of ACE applied to the TIED dataset are shown on the Figure 1. The algorithm stopped after 3 iterations (no new relevant features found), and the resulting set of selected relevant (strongly and weakly) features is shown in the last column.

- Classification tree (Breiman et al., 1984) built on selected features shown on Figure 1. Optimal tree has four terminal nodes, and gave CV BER ~ 0.02. The tree was used for the prediction on the test data. Figure 2 presents surrogate scores tables shown for each of the three splits. Note that for the first split on Column10 there are three surrogates with equivalent splits (Column1/2/3). Similarly for the second and the third splits equivalent splits are achieved by using Column11/12/13 and Column18/19/20 correspondingly.

- Supervised Apriori: we customized Apriori (Agrawal et al., 1993) algorithm to produce rules with known consequent - specific class of a categorical target. We use conditional support (fraction of the data from the specified class covered by the rule) to dramatically simplify APRIORI rule tree construction. As a preprocessing step numeric predictors are discretized, and levels of categorical predictors are optionally clustered with respect to the target class using decision tree with MDL based pruning. The preprocessing is done on each variable independently, and could result in suboptimal rules (this is the case for the target class=2, TIED). The set of the best rules found by the algorithm is shown on Figures 3–4, and involve the same set of variables $\{1,2,3,10\} \times \{11,12,13\} \times \{18,19,20\}$ found by a single tree (with surrogate splits).

Implementation:

All the methods described above are implemented in C++ within Intel Statistical Learning framework - IDEAL. It is not publicly available.

Results:

- Minimal sets of features with optimal predictivity: 36 sets of vars $\longrightarrow \{1,2,3,10\} \times \{11,12,13\} \times \{18,19,20\}$
- Model: Single 4-node classification tree built using any triple from the above cartesian product (see Figure 1) results in the equivalent model with CV BER ~ 0.02

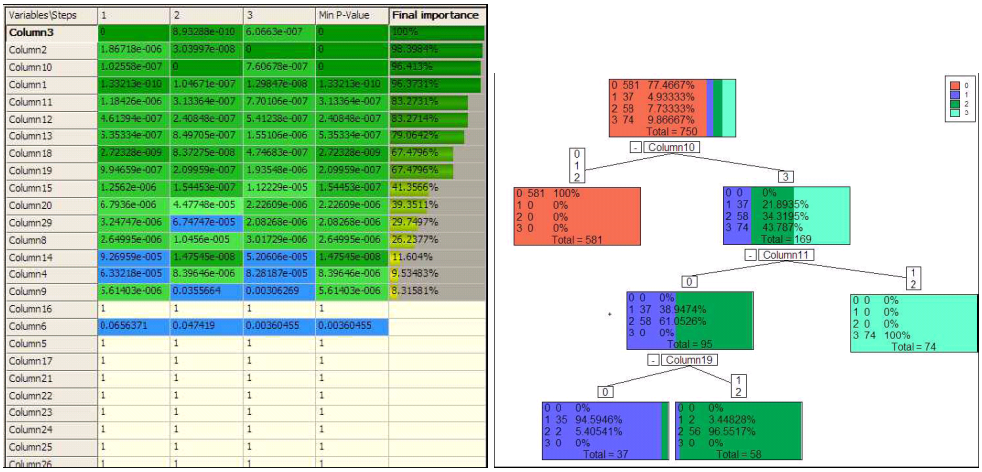
Keywords: feature selection, tree classifier, rule induction, supervised Apriori

References

R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington D.C., May 1993.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, MA, 1984.

E. Tuv, A. Borisov, and K. Torkkola. Feature selection using ensemble based ranking against artificial contrasts. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2006



Variables\Steps	1	2	3	Min P-Value	Final importance
Column3	0	8.93288e-010	6.0663e-007	0	100%
Column2	1.86718e-006	3.03997e-008	0	0	98.3984%
Column10	1.02558e-007	0	7.60678e-007	0	96.413%
Column1	1.33213e-010	1.04671e-007	1.29847e-008	1.33213e-010	96.3731%
Column11	1.18426e-006	3.13364e-007	7.70106e-007	3.13364e-007	83.2731%
Column12	4.61394e-007	2.40848e-007	5.41238e-007	2.40848e-007	83.2714%
Column13	3.35334e-007	8.49705e-007	1.55106e-006	5.35334e-007	79.0642%
Column18	2.72328e-009	8.37275e-008	4.74683e-007	2.72328e-009	67.4796%
Column19	9.94659e-007	2.09959e-007	1.93548e-006	2.09959e-007	67.4796%
Column15	1.2562e-006	1.54453e-007	1.12229e-005	1.54453e-007	41.3566%
Column20	6.7936e-006	4.47748e-005	2.22609e-006	2.22609e-006	39.3511%
Column29	3.24747e-006	6.74747e-005	2.08268e-006	2.08268e-006	29.7497%
Column8	2.64995e-006	1.0456e-005	3.01729e-006	2.64995e-006	26.2377%
Column14	9.26959e-005	1.47545e-008	5.20606e-005	1.47545e-008	11.604%
Column4	6.33218e-005	8.39646e-006	8.28187e-005	8.39646e-006	9.53483%
Column9	5.61403e-006	0.0355664	0.00306269	5.61403e-006	8.31581%
Column16	1	1	1	1	
Column6	0.0656371	0.047419	0.00360455	0.00360455	
Column5	1	1	1	1	
Column17	1	1	1	1	
Column21	1	1	1	1	
Column22	1	1	1	1	
Column23	1	1	1	1	
Column24	1	1	1	1	
Column25	1	1	1	1	
Column26	1	1	1	1	

Figure 1: Left graph: The results of ACE applied to the TIED dataset. The algorithm stopped after 3 iterations (no new relevant features found), and the resulting set of selected relevant (strongly and weakly) features sorted by relative importance is shown in the last column. Right Graph: Classification tree built on the set of the relevant features identified by ACE. For each split surrogate scores are calculated for each variable (see the Figure 2)

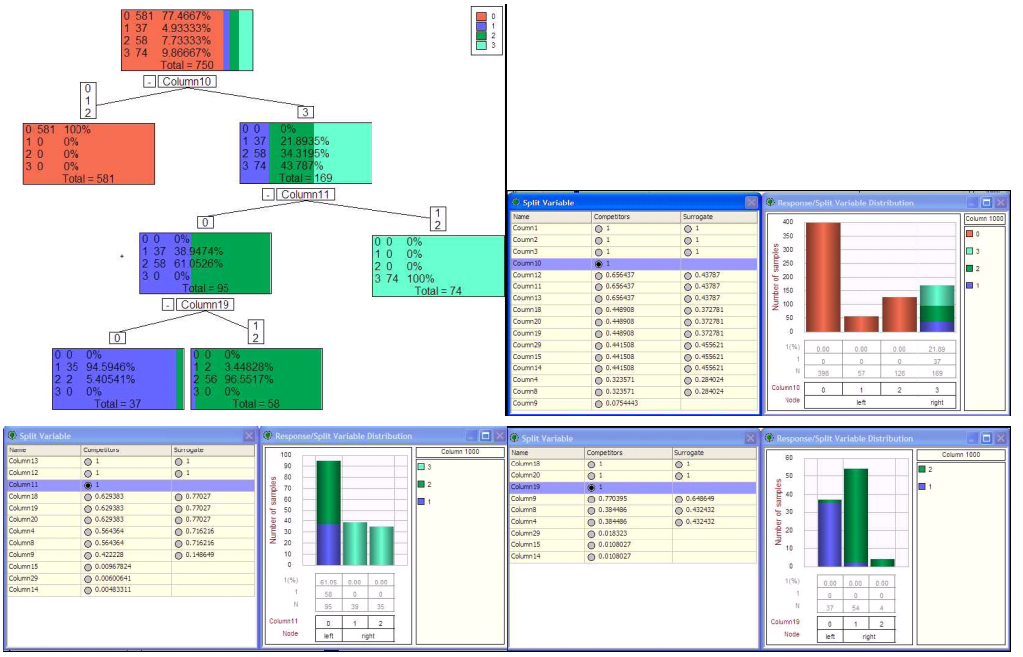


Figure 2: Surrogate scores tables shown for each of three splits for the tree model built to classify TIED target. Note that for the first split on Column10 there are three surrogates with equivalent splits (Column1/2/3). Similarly for the second and the third splits equivalent splits are achieved by using Column11/12/13 and Column18/19/20 correspondingly.

	Itemsets	Length	Confidence	Cond. Support	Support	Rank
☑	Column10 IN (0, 1, 2)	1	1	581 (100%)	581 (77.47%)	0
☑	Column3 IN (0, 1, 2)	1	1	581 (100%)	581 (77.47%)	1
☑	Column2 IN (0, 1, 2)	1	1	581 (100%)	581 (77.47%)	2
☑	Column1 IN (0, 1, 2)	1	1	581 (100%)	581 (77.47%)	3
☐	Column14 IN (0, 1)	1	0.934	526 (90.53%)	563 (75.07%)	4
☐	Column29 = 0	1	0.934	526 (90.53%)	563 (75.07%)	5
☐	Column15 = 0	1	0.934	526 (90.53%)	563 (75.07%)	6

	Itemsets	Length	Confidence	Cond. Support	Support	Rank
☑	Column11 IN (1, 2)	1	1	74 (100%)	74 (9.87%)	0
☑	Column12 IN (1, 2)	1	1	74 (100%)	74 (9.87%)	1
☑	Column13 IN (1, 2)	1	1	74 (100%)	74 (9.87%)	2
☐	Column3 = 3 AND Column18 = 1	2	0.938	61 (82.43%)	65 (8.67%)	3
☐	Column2 = 3 AND Column18 = 1	2	0.938	61 (82.43%)	65 (8.67%)	4
☐	Column1 = 3 AND Column19 = 2	2	0.938	61 (82.43%)	65 (8.67%)	5

Figure 3: Rules for the target class=0 (upper table). Perfect discrimination is achieved with one of the variables 1/2/3/10. Rules for the target class=3 (lower table). Perfect discrimination is achieved with one of the variables 11/12/13.

	Itemsets	Length	Confidence	Cond. Support	Support	Rank
☑	Column13 = 0 AND Column10 = 3 AND Column19 = 0	3	0.946	35 (94.59%)	37 (4.93%)	2
☑	Column13 = 0 AND Column10 = 3 AND Column20 = 0	3	0.946	35 (94.59%)	37 (4.93%)	3
☑	Column13 = 0 AND Column1 = 3 AND Column18 = 2	3	0.946	35 (94.59%)	37 (4.93%)	4
☑	Column11 = 0 AND Column3 = 3 AND Column19 = 0	3	0.946	35 (94.59%)	37 (4.93%)	5
☑	Column13 = 0 AND Column1 = 3 AND Column20 = 0	3	0.946	35 (94.59%)	37 (4.93%)	6
☑	Column11 = 0 AND Column3 = 3 AND Column20 = 0	3	0.946	35 (94.59%)	37 (4.93%)	7
☑	Column12 = 0 AND Column1 = 3 AND Column19 = 0	3	0.946	35 (94.59%)	37 (4.93%)	8

	Itemsets	Length	Confidence	Cond. Support	Support	Rank
☑	Column13 = 0 AND Column20 = 1	2	0.963	52 (89.66%)	54 (7.2%)	3
☑	Column12 = 0 AND Column20 = 1	2	0.963	52 (89.66%)	54 (7.2%)	4
☑	Column13 = 0 AND Column18 = 0	2	0.963	52 (89.66%)	54 (7.2%)	5
☑	Column13 = 0 AND Column19 = 1	2	0.963	52 (89.66%)	54 (7.2%)	6
☑	Column11 = 0 AND Column20 = 1	2	0.963	52 (89.66%)	54 (7.2%)	7
☑	Column11 = 0 AND Column18 = 0	2	0.963	52 (89.66%)	54 (7.2%)	8
☐	Column18 = 0 AND Column8 = 2	2	0.976	40 (68.97%)	41 (5.47%)	9

Figure 4: Rules for the target class=1 (upper table, a subset is shown). The best 36 equivalent rules found by the algorithm involve triples from the set $\{1,2,3,10\} \times \{11,12,13\} \times \{18,19,20\}$. Rules for the target class=2 (lower table). The best 9 equivalent rules found by the algorithm involve tuples from the set $\{11,12,13\} \times \{18,19,20\}$.