AN ABSTRACT OF THE THESIS OF

Meghamala Sinha for the degree of Master of Science in Computer Science presented on March 22, 2019.

Title: Causal Structure Learning from Experiments and Observations

Abstract approved: _____

Prasad Tadepalli

In this research, we address the problem of learning a single causal network structure from multiple dataset generated from different experiments. The experiments can be *observational* or *interventional*. We assume that each dataset is generated by an unknown causal network altered under different experimental conditions (interventions, manipulation or perturbation). As a result, we get a collection of heterogeneous datasets having different distributions. Manipulated distribution implies manipulated graphs over the variables. Combining all the data to learn a network might increase statistical power but only if it assumes a single encapsulating network that is true for all the datasets, which is not always the case under *uncertain* interventions. Pooling under uncertainty leads to spurious changes in correlations among variables. While learning causal network by pooling data from different experiments is common, we found by experimenting that this paves the way for *false causal discoveries*, if the effects of interventions are *uncertain*.

We address these issues and present a Bayesian approach of combining data from multiple experiments with observations to learn a single and accurate causal network. Our approach, called *'Learn and Vote'* learns causal links using data from each experiment and combines them by weighted averaging. We show through studies on synthetic and natural datasets that our method outperforms many state of the art approaches and is more robust with respect to modelling assumptions about the nature of the interventions.

Causal Structure Learning from Experiments and Observations

by

Meghamala Sinha

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented March 22, 2019
Commencement June 2019

Master of Science thesis of Meghamala Sinha presented on March 22, 2019.

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Meghamala Sinha, Author

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Prasad Tadepalli for his support during my pursuit of masters studies and research, and for his patience and immense knowledge.

I would like to convey my special thanks to Dr. Stephen Ramsey for his insightful directions and encouragement, which incented me to widen my research from various perspectives.

I would also like to thank the School of EECS for their constant support.

Last but not the least, I would like to thank my parents and my lab-mates and friends for supporting me and being with me through thick and thin of life in general.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

Chapter 1: Introduction

1.1 Significance of Causal Inference

> *"Where causation is concerned, a grain of wise subjectivity tells us more about the real world than any amount of objectivity."*
>
> Judea Pearl, The Book of Why (2018)

The importance of causal modeling in science, engineering and humanities is remarkable due to its utility in action planning, prediction and diagnosis [27, 37]. A primary goal in causal modeling is to discover *"causal"* interactions of the form $X \Rightarrow Y$, where X and Y are observable entities and the arrow indicates that the state of X causes the state of Y or vice-verse.

For example in medical domain, we wish to diagnose the conditions that lead to a certain disease. When doctors, while examining sample data from their patients, find a relationship between two or more variables, they conclude that they have found a **association** or correlation. Which means one has an effect on the other. A study made on the effects of marijuana [4] showed an association of weed smoking teenagers and their troubled relationship in future (late-20's). Furthermore, researchers observed that correlation can be positive or negative. For example as strength of one variable (marijuana) increases, the probability of the other variable (relationship trouble) increases as well. Obviously there might be other variables, like mental health, education, gender etc, which are causing the associations as well. We call these studies of the world as *observations*.

Although we can deduce that the variables of our interest are related to each other, we still can't say for sure that who caused the occurrence of the another. When we claim that causation is found, it means that a change or **intervention** in one of the variable is a direct cause for the change in the other. We can achieve this by fixing the value of one variable and measure the effect on the other. For example, inferring such causal relationship is of paramount importance in the fields of molecular biology. Knowing the causal structure of a molecular process allows

advanced reasoning on its behavior, and such knowledge could be valuable in therapeutic approaches, for example though predicting side effects of pharmaceutical drugs. There has been a lot of recent development of high throughput technologies to infer the structure and dynamics of complex biological networks, including protein signaling pathways [33], system biology cellular networks [15], and gene regulation networks [1] etc.

Causal models can also be further used to analyze "what-if" cases or **counterfactual** inferences. In economics and statistics, we wish to find out the effect of a number of proposed marketing strategies without actually implementing through counter-factual reasoning. Such queries can be answered once we have the true causal model of the world. Causal models can be fit to passive observational measurements (*"seeing"*) as well as measurements after performing external interventions (*"doing"*).

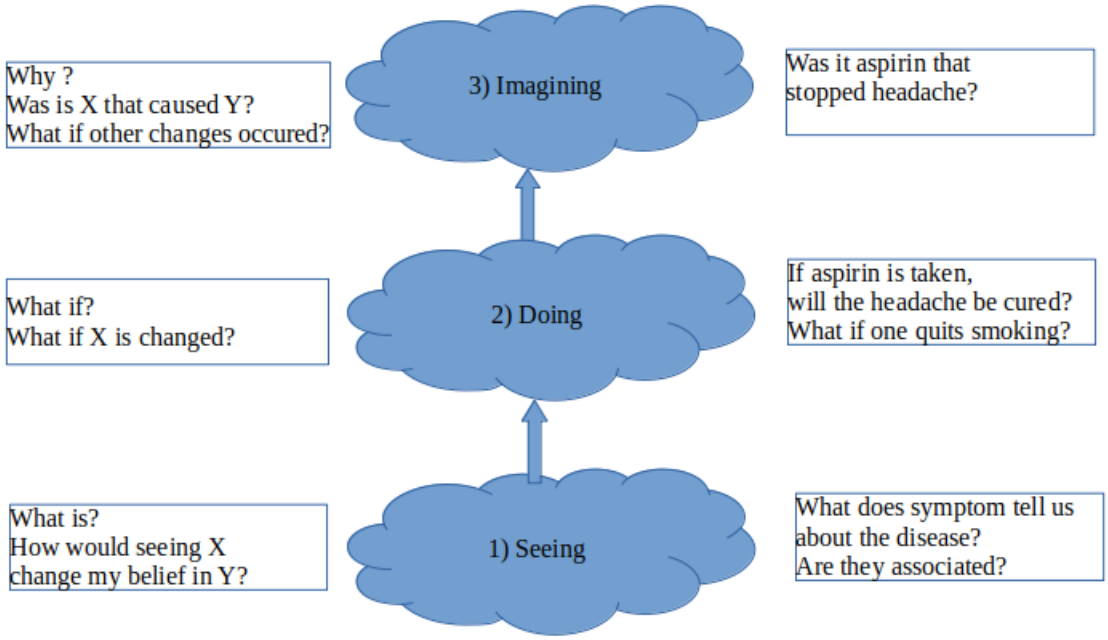


Figure 1.1: The three ladders of causality

1.2 Observation vs Intervention

The *"Principle of Common Cause"* [31] states that given two associated events V_i and V_j, there can be three possible explanations: V_i caused V_j, V_j caused V_i or both V_i and V_j have a common cause. For the third case, an inference of a causal relationship $V_i \Rightarrow V_j$ or $V_j \Rightarrow V_j$ would be spurious. In an observational study, we can determine whether other factors (besides V_i and V_j) are present in the underlying causal network.

Observational studies are useful in various contexts, for example, to retrospectively infer whether an exposure-outcome statistical relationship is causal. However, given only observational data we cannot distinguish between various *Markov equivalent* structures that are compatible with the data [22, 28]. For example, the three graphical models $V_i \rightarrow V_j \rightarrow V_k$, $V_i \leftarrow V_j \leftarrow V_k$, $V_i \leftarrow V_j \rightarrow V_k$ (where an arrow denotes an arc in the probabilistic graphical model) are *Markov equivalent*, since they encode the same $V_i \perp\!\!\!\perp V_k | V_j$ conditional independence statement.

Interventional studies, on the other hand, can be used to compare the likelihood of various *Markov equivalent* structures consistent with an observation [17]. For instance, if we intervene on A and find a change in B, we can conclude that A causes B. Unlike in observational studies, here events can be externally manipulated and become independent of their cause(s). Thus supplementing observational studies with interventions helps to discriminate between such models and determine the true causal directions of correlated relationships in the system [14]. The inability of observational studies to discriminate between Markov-equivalent structures motivated studies that combine observational data with interventional data [17].

Why do we need to include observation at all?

- Because its very crucial in these kinds of studies. We cannot simply assign people to intervention in real world in an unethical way, for example, children to consume marijuana to study their effects. We have to depend on the study of association. Through observations we learn the true nature of a relationship and rule out other interfering variables.

1.3 Key contribution of this work

The inability of observational studies to discriminate between Markov-equivalent structures motivates studies that combine observational data with interventional data [17]. Despite its potential advantages, learning causal networks from a mix of observational and experimental studies is a significant challenge. Data collected after different experiments might not be identically dis-

tributed as before making the results incoherent with one true causal structure. Such discrepancies could be due to *unknown* consequences of interventions. Different experiments might have different joint distributions due to *uncertain* effects of each intervention or condition [13]. For instance, in the case of a drug intervention on cells, a drug may have unintended direct effects on molecules other than the drug's intended target, i.e., "off-target" effects. Pooling data from different experiments can lead to misleading changes in correlation. Eberhardt (2008) described two such types of problems: a) *Independence to Dependence*: where V_i and V_j being two independent variables in a structure before and after an intervention, become dependent when the two samples are pooled and b) *Dependence to Independence*: V_i and V_j, which are dependent in a passive observational study, become independent when pooled with an interventional study. This problem occurs even though the interventions are *perfect*. This gives rise to a problem of *false causal discovery* which we address in this work.

In light of the above issue, it is important to carefully consider how to handle *uncertain* interventions while learning causal network structures. Relatively little attention has been given to the problem of how to assimilate data generated from many such interventional experiments. Given two or more datasets generated from different interventions, it is unclear how to combine the data for optimal efficiency of learning. Most of the popular causal learning algorithms assume that interventions are *perfect*, which raises questions about their applicability to real-world datasets that violate this assumption. While these algorithms might be able to learn most of the true arcs, significant false detection might lose the very purpose of learning such networks. For example in medical science, a *false positive* result giving an erroneous indication that a particular disease is present (when it isn't) can result in unnecessary panic and cost of medical tests. In such cases, learning a reliable causal network is more important than learning an accurate but low confident one. The key contributions of this paper are as follows:

1. We describe a way of handling *uncertain* interventions by learning causal information from different experiments separately and combining the resulting structures using a simple approach called *'Learn and Vote'*.

2. We compare our results with a baseline method on *Flow cytometry* data. We found that our approach gives a significant reduction of *false causal discovery*.

3. We performed a comparative study of prominent casual network discovery methods with *uncertain* interventions over various benchmark networks to validate our method's performance.

Chapter 2: Motivation

2.1 Related Study

While learning causal network from mix of observation and interventional data is useful, many of such experiments arise from different conditions. Even if the observed variables in all the experiments are the same, the conditions under which the experiments are conducted are different. A certain dependency occurring in one experiment might be absent in another. Collaborating such datasets might eradicate the dependency and should be taken extra care. How to correctly use causal information from such different experiments into a single efficient network is still an open question.

Popular *Constraint-based* methods like PC [37], FCI [38] etc use the entire dataset to learn causal networks using conditional Independence tests through examining the properties of Markov networks. Similarly, *Score based* methods like GES, GIES [19] etc uses a score based method to measure the best fitting candidate network using the entire dataset. Both type of algorithms are designed to fit a single dataset since they were originally designed to infer causal network from observational data. They do not take into account partitioning the data based on the different condition they have been experimented from. In this section, we describe some works which have extended these methods to take into account the different context behind the experiments.

- **Pooling:** These methods deals with learning a single causal graph by pooling data from across different experiments. The interventions can be perfect [33], imperfect [39], uncertain or even unknown. Cooper and Yoo (1999) [10] first provided a score-based causal learning algorithm by combining data from across various experiments each having one or more perfect known targets of intervention by adding additional context variables. This idea was later re-defined by Eaton and Murphy (2007) [13] to handle soft interventions or mechanism changes [39]. The causal invariance property across environments changes is exploited by Claassen and Heskes (2010) [7].

  Some practical application were studies in flow cytometry data set [10], yeast transcriptional regulatory network [5] etc. A recent approach called Joint causal inference (JCI)

[25] that takes into account data generating from different conditions and introduced additional context variables (representing different contexts like age, country etc that discriminates the different datasets) before pooling.

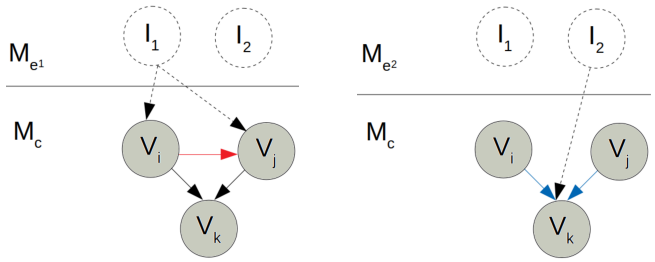
- **Not Pooling:** This method deals with learning causal information separately from each experiments and combining them to learn a single graph. The ION-algorithm [11] have shown a way to integrate locally learned causal networks having overlapping variables. Triantafillou and Tsamardinos (2015) [41] proposed a constraint based algorithm, COmbINE, which estimates dependencies and in-dependencies across separate experiments. However, both the above methods assume knowing a single causal structure that comprises of all observed causal dependencies. This might be hard to achieved in reality when the experimental conditions are changing across each experiments. The MCI [8] algorithm presents a constraint based method by exploiting the 'local' aspect of causal Y-structures [24] which is sufficient to explain the in-dependencies between two variable. This does not differs even if the information are learn from two separate datasets.

2.2 Spurious Dependencies and Independecies

External perturbation that effects two or more variables in a causal model M_c can lead to spurious dependencies or in-dependencies. We have shown two such cases in Figure 5.1. Each causal model M_c contains a pair *(V, E)*, where V is a set of vertices and E is a set of edges between pairs of nodes with *P(V)* representing the joint distribution. The causal arcs $V_i \Rightarrow V_k$ and $V_j \Rightarrow V_k$ are represented by black arrows, shown in Figure 2.1a. We represent the external perturbation caused by experiments as an external model M_e containing a set of unobserved policy variable $I_1,\ I_2\ ...\ I_n \in I$. Experiments can be both observational as well as controlled. The combined model is $M_T = M_c + M_e$ which includes all the external effects in the causal system.

Definition 2.2.1. *False causal dependence:* Two or more variable say V_i, V_j, where $V_i \nRightarrow V_j$, are effected by a common intervention (I_1 in Figure 2.1a) becomes $V_i \not\perp\!\!\!\perp V_j$ due to confounding effect. This gives rise to a new model $M_{T_1} = M_c + M_{e^1}$ with a changed distribution $P_1(\mathrm{V} \subset M_{T_1})$. Pooling data from such different distribution may lead to spurious correlation between independent variables.

Definition 2.2.2. *False causal independence:* Intervening on a child node having causal parents removes all the incident arrows and cuts of the causal influence. Pooling data from such models



(a) False dependence (b) False independence

Figure 2.1: Problem with Pooling: Dashed arrows represent intervention effect, Solid black arrows represents True positives (TP), Red arrows represents False positives (FP) and Blue arrows represents False negatives (FN).

nullifies the true causal dependence by causing $V_i Pa(V_i)$. This gives rise to a new model $M_{T_2} = M_c + M_{e^2}$ with a changed distribution $P_2(\mathrm{V} \subset M_{T_2})$. Pooling various such experiments on V_j, shown in Figure 2.1b will dominate over other experiments having the causal relations $V_i \Rightarrow \mathrm{V}_k$ and $V_j \Rightarrow \mathrm{V}_k$.

This potential for false causal dependencies and independencies is a shortcoming when pooling data to learn causal networks. Before describing how our approach aims to address these issues (compared in Figure 5.2d), we first define some causal network concepts, notation, and standard approaches.

Chapter 3: Causal Network learning

3.1 Brief Overview

A causal network [27, 37] is defined by a directed acyclic graph (DAG) $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ denotes the set of variables (nodes) and E denotes the causal relationships (edges), and P_i denotes a set of conditional probability distributions for each variable v_i. For an edge (v_i, v_j), we say v_i is a parent (cause) of v_j, and v_j is a child (effect) of v_i. We define $Pa(v_i)$ as the set of parents of v_i. The conditional probability distribution P_i defines the probability of v_i given its parents $Pa(v_i)$. A causal network represents a joint distribution P over variables V as long as it satisfies two main assumptions:

1. *Causal Markov Assumption:* Any given variable v_i is independent of its non-descendants, conditioned on all of its direct causes. The assumption implies that the joint distribution $P(V)$ can be factored as:
$$P(V) = \prod_{i=1}^{n} P_i(v_i|Pa(v_i)).$$

2. *Faithfulness Assumption:* $P(v_1, \ldots, v_n)$ is said to be faithful to G if every conditional independence relation that holds in P is entailed by the causal Markov Assumption applied to G [12].

Distinguishing networks via observations: Most of the popular causal learning algorithms are based on the Markov condition. These algorithms evaluates the statistical dependencies present in the given data samples and use them to build causal graphs, satisfying the Markov's constraints. The 'Constraint-based' method searches a space of all possible causal structure, but are limited due to the limited nature of the constraints. They are used to analyze whether two variables are statistically dependent and ignore specific indications —such as whether one variable is always present when the other is, or always takes on the same value as another variable etc.

Distinguishing networks via intervention: From Figure 3.1, we can see three networks are Markov equivalent with the common undirected graph but different direction of one or more

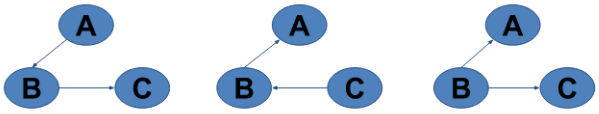


Figure 3.1: Markov equivalence class

arrows. Three networks which tells us the same statistical dependencies under observation. An intervention on one variable (say B) of this structures will be adding an external variable into the causal network as an additional parent of B. In an ideal intervention, the manipulated B becomes uniquely determined by the external parents an invalidates any other influences any other causes. The effect of this intervention tells us about the true underlying causal model. For example, if we intervene on B, and find both A and B turning on (under the hypothesis A $\leftarrow$ B $\rightarrow$ C , then we can conclude that the third network in the figure is the correct causal model.

3.2 Constructing a causal network

Let us suppose the dataset D consists of n observations and m variables. A common approach is to start by learning a reasonably high likelihood causal graph for D and then estimate model parameters based on the graph structure and data. The network learning approaches are categorized into mainly two groups:

- *Score-based:* This is based on a scoring function measuring the fit of the graph G to the data, while at the same time favoring simpler structures [9, 6]. The scoring function is combined with a search-heuristic that explores the space of possible graphs. Typical heuristic are greedy ones (hill-climbing or tabu search) [32, 16].

- *Constraint-based:* This approach relies on estimating some of the conditional (in)dependencies in the data distribution from the data by performing hypothesis tests [21]. The results of the hypothesis tests constrain the graph to reconstruct. Constraint-based methods start with a fully connected, undirected graph and progressively remove edges as a new conditional independence relationship is discovered.

3.3 Learning with interventions

Interventions are externally manipulate nodes *("targets")* in a network, which are assumed to be known. Interventions are important to detect causal relations that can help disambiguate *Markov equivalent* sub-networks. Let I_e represent the target nodes which are altered in the interventional experiment e and $O_e = V \backslash I_e$ be the complementary set of observational variables. Each intervention can have one *or more* targets whose conditional probabilities are changed. Hence, each intervention results in deletion of arrows pointing towards the intervened nodes. The joint distribution of P after intervention is given as:

$$P'(v_1, ..., v_n) = \prod_{v_i \in O_e} P_i(v_i|Pa(v_i)) \times \prod_{v_i \in I_e} P'_i(v_i|Pa'(v_i)),$$

where $P'(v_i|Pa'(v_i))$ is the post-intervention conditional probability of v_i given its new set of parents $Pa'(v_i)$. For a "perfect" intervention, we set $Pa'(v_i) = \emptyset$ [27].

Another way to model intervention is to introduce intervention nodes as *switching parents*, on the target nodes. These switches can have values 0 or 1. By setting them to 0, we assume a *normal* mechanism of the network. On setting it to 1, will model an *imperfect* or *"soft"* intervention [40], which will simply increase the likelihood of the node to enter its target state instead of cutting them off their parents. This is often referred as "mechanism change" and is more realistic than *perfect* interventions.

A further relaxation is to consider if all the targets of an intervention are *"uncertain"* [13]. Such cases (for example chemicals in molecular biology) occur when interventions have a *"fat hand"* which touches many nodes. In this study, we have considered the case where we are not aware of the consequence effects each intervention might be having other than its intended targets. The actual real-world setting would be a combination of the *imperfect* and *uncertain* case, which would be difficult to model.

Chapter 4: Our Approach: Learn and Vote

To avoid the problems arising from pooling data from different distributions, we propose an approach we call "Learn and Vote" (Algorithm 1) to learn causal networks. The approach is to learn a separate weighted causal network from the data generated in each experiment or observational study by ignoring the directed arcs into the intervened variables and then combine the results by weighted averaging. For each dataset, we have the observed variables (`N`) and the *known* targets (stored as list `intv`) if any intervention is performed. The details of our approach are as follows.

4.1 Scoring Function

The effect of intervention is incorporated in the score component associated with each node by modifying the standard Bayesian Dirichlet equivalent uniform score (BDeu) [20, 10, 30].

Given a dataset D_j from the j^{th} experiment G^j represents a DAG over a set of variables N learned from it (with conditional distributions $P(N_i|Pa_i^G)$, where Pa_i is parent of N_i). In case of an interventional experiment, we assume *perfect* intervention by fixing the values of $N_i[m]$ in *Int(m)*, which is the set of intervened entities in the m^{th} sample. Hence, we should no longer consider $P(N_i[m] \mid Pa_i[m])$ in the scoring function. But since the interventions are *"perfect"*, [27] all the other variables are unaffected and therefore we sample them from their original distributions. Here, the distribution D_j is per experiment and not a mixture of pooled data from different experiments like in Sachs et al.'s method. We define the score of $S(G^j : D_j)$ as a composition of the contributions of each local score (S_{local}) of variables N_i. The modified local score is as follows:

$$S_{local}(N_i, U \;:\; D_j) \;=\; \log P(Pa_i \;=\; U) \;+\; \log \int \prod_{m, N_i \notin \mathrm{Int}(m)} P(N_i[m]|U[m], \theta) dP(\theta),$$

4.2 Structure Learning

Due to limitations in data, the results of structure learning in most real-world setting are noisy. To overcome this we create $n = 100$ random DAGs using `createRandNet` over the set of given nodes

to learn an averaged network from each experiment. We learn the structure from each DAGs in `randomNet` using the *Tabu* search algorithm [16] which searches over the space of different structures and store them in a list `Net`. The list `intv` of *known* targets is passed as an argument which incorporates interventions in the search algorithm by preventing the arcs to be incident on the targets. Next, we measure the probabilistic arc strength and direction (using `arcStrength`) for each arc as its empirical frequency given the list of networks in `Net`. We average the arc strengths for every directed arc over the networks in which corresponding target node was not intervened and store them as `arcProb`.

ALGORITHM 1 Learn and Vote

Input: set of k experiments with dataset D_1,D_2...D_k
Output: DAG G^f = (E, V), final causal network

```
1:  procedure OUR APPROACH
2:      for j=1 to k do
3:          N=nodes In D_j
4:          intv=Intervened nodes in D_j
5:          randomNet=createRandNet(N, 100)
6:          for l=1 to 100 do
7:              Net[l]=Tabu(randomNet[l], intv)
8:          arcProb[j]=arcStrength(Net)
9:      avgArcs = avgNetwork(arcProb)
10:     G^f = learnDAG(avgArcs,Threshold)
```

4.3 Combining results from the experiments

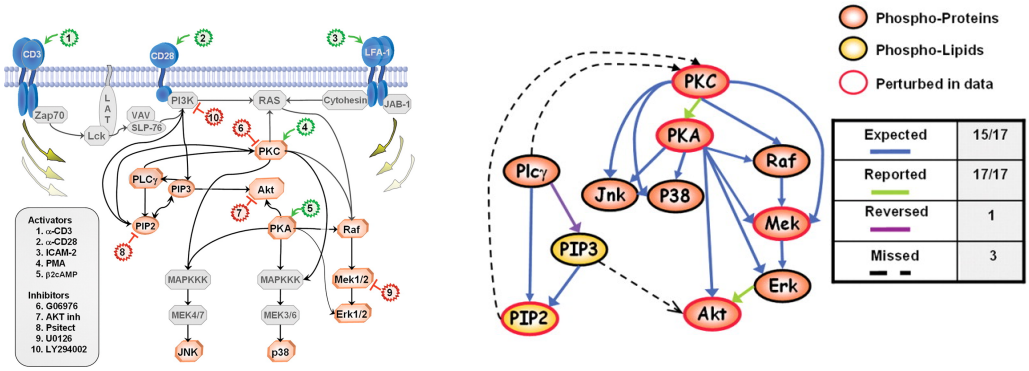
Given arc strengths from each experiment, we average their strengths and directions over the number of experiments the given arc is valid (using `avgNetwork`). Finally, we store the averaged arc strengths as `avgArcs` to build the final DAG (using `learnDAG`) containing only the significant arcs over a certain `Threshold`. We found our method performing best at a threshold of 0.5. We implemented our methods in the `bnLearn` R package [35].

Chapter 5: Comparative Studies

5.1 Biological Signaling Network

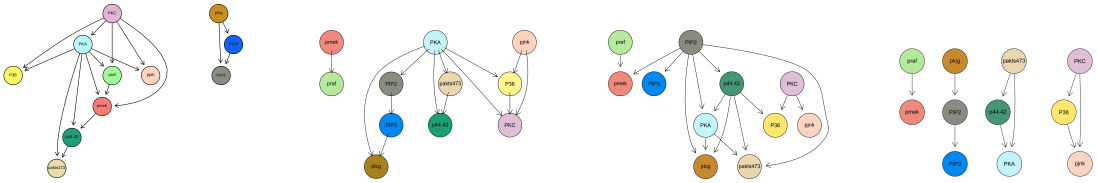
In this work, we have analyzed polychromatic flow cytometry data which examines the phosphorylation states of different proteins in cells. Flow cytometry is a process to examine the physical and chemical characteristics of cellular particles. The components are fluorescently labelled and are excited with the help of laser to discharge light at different wavelengths. These fluorescents are used to measure the different properties of cell molecules. Flow cytometry experiments have application in health-care science such as cell counting, cell sorting, detection of biomarkers, Protein engineering etc. In the experiments, the elements of the three Mitogen-Activated Protein Kinase (MAPK) pathway in human CD4+ T-cells are detected. The MAPK/ERK signal transduction pathway is a chain of proteins that emits a signal from a surface receptor to the DNA in the cell's nucleus and produces some changes in the cell. The proteins in the pathway communicate with each other by adding phosphorate groups to a neighbor protein acting like an on/off switch. Such pathways are very useful for cancer detections [26]. Since the protein states are found to get stuck in an *on or* off position on an occurrence of a mutation, reversing the effect of such switch can be useful for treatments. For this experiment, the 11 well-known proteins of the MAPK pathways were studied as highlighted in Figure 1a. The pathways were intervened by 9 external stimuli, whose effects targets are described in Table 1.

Sachs et al. [33] inferred the signaling pathway and novel causal interactions in human CD4+ T-cells, using a Bayesian network approach (Figure 5.2a). They carried out nine experiments, including two general (observations) and seven specific (interventional) perturbations to measure the expression levels of eleven phosphorylated proteins and phospholipids using multiparameter flow cytometry. They found 17 *true positives* (TP=17, with 15 from well-established literature and 2 with at least one citation) out of 20 expected arcs and missed 3 *false negatives* (FN=3) shown in blue dashed lines. They did not have any *false positives* (FP=0). They also showed that network inference from solely the observational measurements is inaccurate and that inclusion of interventions gives more accurate results. We re-analyzed Sachs et al. approach twice, first using observational samples only (Figure 5.2b) and then using an equal number of samples



(a) (b)

Figure 5.1: (a) Classic signaling network with points of interventions by reagents, (b) Model inference result by Sachs et al.



(a) TP: 17, FP: 0, FN: 3 (b) TP: 3, FP: 10, FN: 17 (c) TP: 5, FP: 10, FN: 15 (d) TP: 4, FP: 5, FN: 16

Figure 5.2: (a) Network inferred by [33] (b) Network inferred from two observational experiments (c) Network inferred from pooling data from a observational and a interventional experiment d) Network inferred from "Learn and Vote" using the same experiments as (c).

comprising 50% observational and 50% interventional data. (Figure 5.2c) illustrates this point by being much closer to the ground truth.

This analysis clearly shows the benefit of interventional experiments for causal network reconstruction. However, like most causal discovery approaches, the methods used in the Sachs et al.'s study and in our re-analysis, assume *perfect* intervention. Each of the seven specific perturbed experiments sets exactly one of the signalling molecule to a specific state. This is inspired by Pearl's *"do-calculus"* [27] in which all the parents' influences on the target node are removed from the graph structure. Such a perfect intervention modelling is often not consistent with interventions that are typical in biology, for example, due to off-target effects of an intervention. Moreover, the effects of interventions like a gene knockout may not be describable via

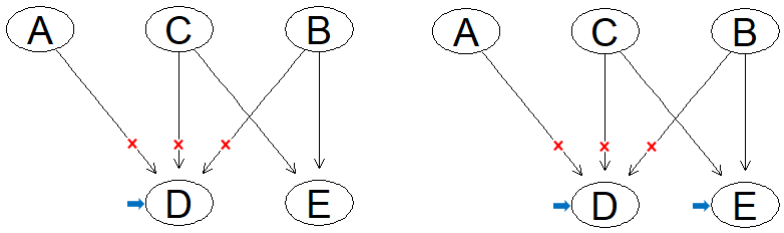


Figure 5.3: (a) ”perfect” intervention on D (b) ”uncertain” intervention on D with unknown target E

the forcing of a target node's state to a specific value in the observational network. In the Sachs et al.'s cell signaling data, we know the nominal target of each of the reagents used in the targeted interventions, but they might affect other (possibly hidden) variables. In such cases, using causal inference methods assuming *perfect* intervention with known targets can detect spurious interactions due to the following issues:

- For the cases where there are some intervention targets that are unknown, a problem of uncertainty is introduced. The current algorithms might make mistakes since the arcs pointing towards the unknown targets are not removed or handled properly, as shown in Figure 5.3. E being an unknown target of intervention, we need to remove the arcs from its parents too. So all the spurious arcs pointing towards the unknown targets still remains and are added up in the inferred causal network.

- Although combining data from different uncertain interventions adds more confidence into the true causal arcs, there are several spurious arcs which might get more weights and result in false detection.

- Apart from uncertain targets, each intervention might also cause a mechanism change or influence the local distribution in an unknown way [40].

Before describing how our approach aims to address these issues (compared in Figure 5.2d), we first define some causal network concepts, notation, and standard approaches.

We evaluated our algorithm on various synthetic and real world data ranging from small to medium size. For the synthetic networks we sample equal amount of data from observational and interventional experiments from each network. We simply draw the observation data as random

samples from each synthetic network. In the interventional experiments, to model uncertainty, we set the *known* target node of each *perfect* intervention to a certain value. Next we also set one or more of its children to different values (like *"fat-hands"*) which are assumed to be *unknown* and finally sample data from each of these mutilated networks.

5.2 Description of datasets used for experiments

Description of the datasets we used for this study are as follows:

- **Flow Cytometry:** This is a technique for obtaining multiparameter molecular measurements from individual cells. The original data, provided by [33] is collected from a series of 9 experiments. We use the raw data and replicate their data-processing procedure in R for our evaluation. Although, the interventions are assumed to be ideal, their effects are known to have unknown consequences as shown in [13].

Table 5.1: Summary of the stimulating reagent and their effects

REAGENT	CLASS
Anti-CD3/CD28	General Perturbation
ICAM-2	General Perturbation
b2cAMP	Activates PKA
AKT inhibitor	Inhibits AKT
U0126	Inhibits Mek1/2
PMA	Activates PKC
G06976	Inhibits PKC
Psitectorigenin	Inhibits PIP2
LY294002	Activates AKT

Data Preprocessing: We collected the raw data provided in [33] containing 9 experimental files. Next, the data was cleaned by removing all the outliers falling more than three standard deviations from the mean. Equal number of datapoints (600), from each experiments, were randomly sampled with replacement to prevent biasing to a particular experiment.

Data Discretization: Each dataset records all the proteins concentration levels after each

stimulatory cues. Since the data contains molecule concentrations, they are continuous and using a Gaussian Bayesian Network (GBN) seems more reasonable. However the result found is not as good as expected. Hence, data from each experiment are separately discretized to preserve maximum pairwise dependencies using a method described by Hartemink, et al [18]. First the data are binned into intervals, followed by subsequent collapsing due to reduction of mutual information among variables. The process continues until each variable has three levels (high, medium, low protein expression).

- **Lizards:** This is a real-world dataset having 3 variables representing the perching behaviour of two species of lizards in the South Bimini island [34]. We generated one observation and two interventional studies.

- **Asia:** This is a synthetic network of 8 variables [23] about occurrence of lung diseases and their relation with visits to Asia. For our experiment, we created two mutilated networks. *Asia_mut1* have one observation and one interventional study. *Asia_mut2* have one observation and two interventional studies.

- **Alarm:** This is a synthetic network of 37 variables representing an alarm messaging system for patient monitoring [2]. For our experiment, we created two mutilated networks. *Alarm_mut1* have three observational and six interventional studies. *Alarm_mut2* have five observational and ten interventional studies.

- **Insurance:** This is a synthetic network of 27 variables for evaluating car insurance risks [3]. We created two mutilated networks. *Insurance_mut1* have one observation and five interventional studies. *Insurance_mut2* have three observations and eight interventional studies.

- **gmInt:** This is a synthetic dataset containing a matrix of observational and interventional data from 8 Gaussian variables, provided in the `pcalg-R` package.

5.3 Popular Causal Structure Learning Methods

We evaluate the following algorithms (implemented in `R`) for our comparative analysis. The learned causal graphs on the *flow cytometry* datasets are shown in Figure 5.4a-5.4e.

- **PC:** The observational experiments were used to evaluate the equivalence class of a DAG using the PC algorithm [37]. Fisher's z-transformation conditional independence test was used by varying α from 0 to 1 in steps of 0.01.

- **GDS:** This is a greedy search method [19] to estimate Markov equivalence class of DAG from observational and interventional data. This is a greedy DAG search proposed by [19] to estimate the Markov Equivalence Class of a DAG. It estimates the observational or interventional Markov equivalence class of a DAG based on a mix of interventional and observational data. This algorithm performs a greedy DAG search by maximizing a score function (l_0-*penalized Gaussian maximum likelihood estimator*) in 3 phases: 1) forward phase (addition of an arrow in the space of DAGs, until improvement of score can be seen), 2) backward phase (removal of an arrow in the space of DAGs, until improvement of score can be seen) and 3) turning phase (reversal of an arrow in the space of DAGs, until improvement of score can be seen).

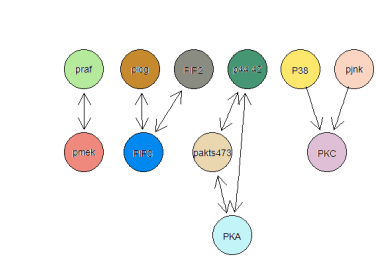
- **GIES:** This algorithm [19] extends the greedy equivalence search (GES) algorithm [6] to a generalized version that includes interventional data into observational data.

- **Globally optimal Bayesian Network:** This is a score-based dynamic programming approach [36] to find the optimum of any decomposable scoring criterion (like BDe, BIC, AIC). This function (simy) estimates the best Bayesian network structure given interventional and observational data but is only feasible up to about 20 variables.

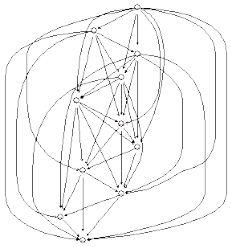
- **Invariant Causal Prediction:** This method by Peters et al., [29] calculates the confidence intervals for causal effects by exploiting the invariance property of a causal (vs. non-causal) relationship under different experimental settings. We implemented it using `InvariantCausalPrediction` R package.

Table 5.2: Comparative Results

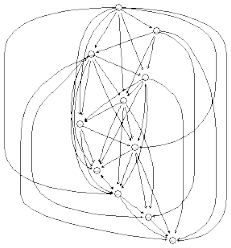
Dataset	*Metric*	*Causal Discovery Algorithms*						
		PC	GDS	GIES	ICP	simy	Sachs et al	Learn and Vot
Flow Cytometry	Precision	0.5714	0.4186	0.377	**1**	0.4222	0.68	0.89
	Recall	0.4	**0.9**	0.85	0.45	0.95	0.85	0.89
	F1 score	0.47	0.572	0.522	0.62	0.584	0.7558	**0.89**
Lizards	Precision	**1**	**1**	**1**	0	**1**	**1**	**1**
	Recall	**1**	**1**	**1**	0	**1**	0.5	0.5
	F1 score	**1**	**1**	**1**	0	**1**	0.667	0.667
*Asia_mut*1	Precision	**1**	0.625	0.625	**1**	0.31578	0.77	**1**
	Recall	0.75	0.625	0.625	0.5	0.75	**0.875**	0.75
	F1 score	**0.857**	0.625	0.625	0.666	0.4444	0.8237	**0.857**
*Asia_mut*2	Precision	**1**	0.85714	0.85714	**1**	0.3043	0.666	**1**
	Recall	0.75	0.75	0.75	0.5	**0.875**	0.75	0.75
	F1 score	**0.857**	0.8	0.8	0.666	0.4928	0.7058	**0.857**
gmInt	Precision	0.75	0.889	0.889	**1**	0.889	0.8571	**1**
	Recall	0.75	**1**	**1**	0.375	**1**	0.75	0.75
	F1 score	0.75	**0.94**	**0.94**	0.5454	**0.94**	0.8	0.857
*Alarm_mut*1	Precision	0.666	0.25	0.26	**0.7**	n/a	0.625	0.564
	Recall	0.434	0.217	0.26	0.26	n/a	**0.4464**	0.4
	F1 score	**0.526**	0.2325	0.26	0.38	n/a	0.52	0.468
*Alarm_mut*2	Precision	0.666	0.411	0.5128	0.6	n/a	0.725	**0.769**
	Recall	0.434	0.456	0.434	0.21	n/a	0.63	**0.642**
	F1 score	0.526	0.432	0.47	0.3115	n/a	0.675	**0.7**
*Insurance_mut*1	Precision	0.7143	0.36	0.3617	0.7	n/a	**0.857**	0.8
	Recall	0.288	0.3461	0.327	0.25	n/a	**0.577**	0.538
	F1 score	0.4107	0.352	0.3435	0.368	n/a	**0.689**	0.643
*Insurance_mut*2	Precision	**0.7143**	0.355	0.366	0.64	n/a	0.676	0.6857
	Recall	0.288	0.423	0.423	0.21	n/a	0.4423	**0.4615**
	F1 score	0.4107	0.386	0.392	0.316	n/a	0.535	**0.5517**



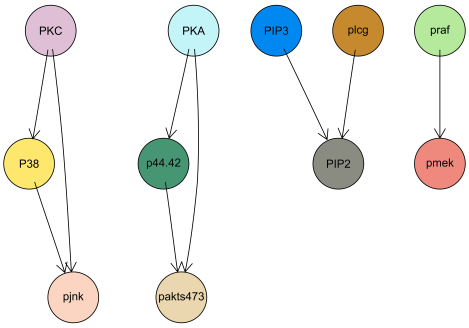
(a) TP:8, FP:6, FN:12



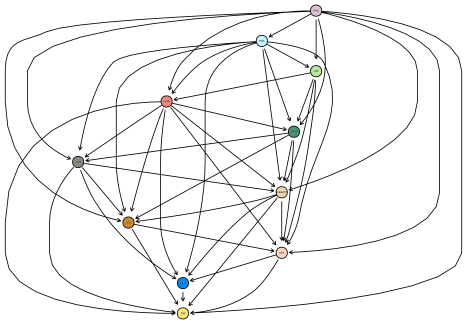
(b) TP:18, FP:25, FN:2



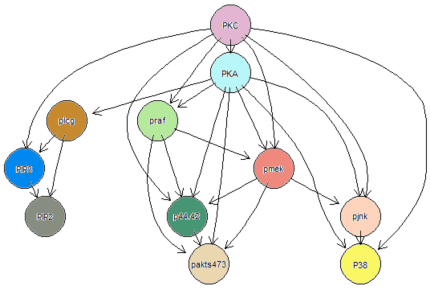
(c) TP:17, FP:28, FN:3



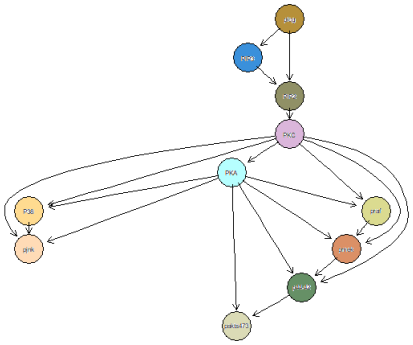
(d) TP:9, FP:0, FN:11



(e) TP:19, FP:26, FN:1

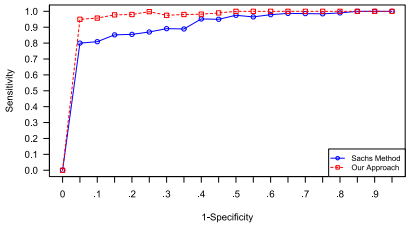


(f) TP:17, FP:8, FN:3

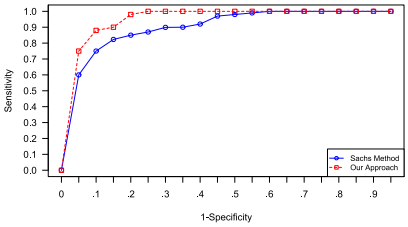


(g) TP:18, FP:2, FN:2

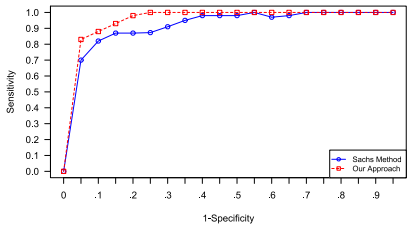
Figure 5.4: Network Inferred from various algorithms: (a) PC, (b) GDS, (c) GIES, (d) ICP, (e) simy, (f) Re-implemented Sachs et al. 2005 and (g) 'Learn and Vote'



(a) Flow Cytometry

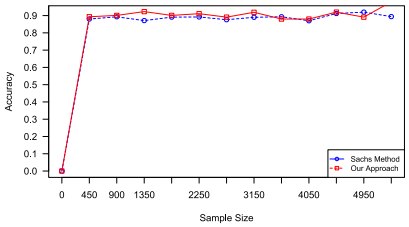


(b) Asia_mut1

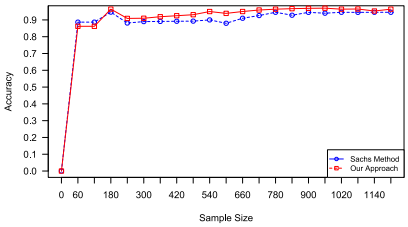


(c) Asia_mut2

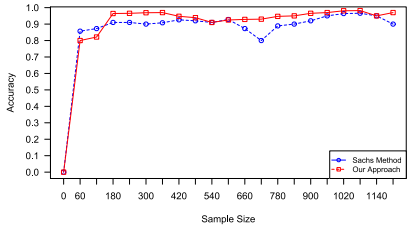
Figure 5.5: ROC plot for comparing results over various datasets



(a) Flow Cytometry



(b) Asia_mut1



(c) Asia_mut2

Figure 5.6: Sample size vs Accuracy plot for comparing results over various datasets

Chapter 6: Analysis of Results

Table 6.1 summarizes the results of the different structure learning algorithms over all the datasets.

6.1 Evaluation Metrics

We treat the presence of an arc in the ground-truth dataset as a *"positive"* example and its absence as a *"negative"* example. For each inferred network we compute the confusion matrix counts in the usual manner. For each of the nine datasets and each of the seven inference algorithms, we report the *precision*, the *recall*, and the *F1 score*.

Our approach outperformed all the baselines in five out of nine studies in terms of *precision*, with the ICP method having second best performance. The positive predictive rate of our approach is higher for small or medium sized networks (less than 20 nodes) but comes down as the size of the network increases. In terms of *recall*, although the performance of the greedy algorithms (GDS, GIES, simy) is better for smaller networks, it decreases as the network size increases. In terms of *F1*, our approach outperformed the others in five out of nine studies and is more stable even when the network size increases. The PC algorithm learns better in case of small networks (less than ten nodes), even with only observational data.

6.2 Network inference results on Sachs et al.'s dataset

Here, we compare the graphs learned using our approach in Figure 5.4g with that of the Sachs et al. network inference method in Figure 5.4f on their cell signaling dataset. The Sachs et al.'s method resulted in 8 *false positive* arcs, 3 *false negative* arcs, and 17 *true positive* arcs (Figure 5.4f). Our method detected all 17 arcs that were correctly detected by the Sachs et al. method plus another arc *(PIP2 → PKC)* that the Sachs et al. method missed. We detected two *false positives*. On further study, we found that both of the detected putative *false positives* by our method, *(P38 → pjnk)* and *(PKC → p44.42)*, are likely real interactions according to PCViz[1] and PubMed[2].

[1]PCViz: http://www.pathwaycommons.org/pcviz/
[2]PubMed : https://www.ncbi.nlm.nih.gov/pubmed/

Table 6.1: Results on Flow cytometry data

METHODS	Expected	Reported	Missed
Sachs et at.	15/17	17/17	3
Our Approach	16/20	20/20	2

Figure 5.4 shows the networks inferred by the seven inference algorithms on the Sachs et al.'s dataset. The greedy algorithms (Figure 5.4b, fig. 5.4c, fig. 5.4e) are able to find most of the *true positive* arcs at the cost of a large number of *false positives*. Hence such methods are not reliable in interventional studies having uncertain targets. ICP on other hand is restrictive due to its strict invariance property and helps reduce false causal arcs to a great extent, but at the cost of sensitivity (Figure 5.4d). We also contrast the performance of the PC algorithm by working only on the observational data. we can see from Figure 5.4a that most of the directions are undetermined and the overall performance improves by adding interventional data.

To show the effect on a smaller scale, we can refer back to Figure 5.2c and 5.2d. Here we used one general perturbation (*Anti-CD3/CD28*) and one specific perturbation experiment (*AKT inhibitor*). We can see how the number of *false positives* reduces by avoiding pooling data.

6.3 Sensitivity to Threshold

To analyze the sensitivity of our results to the threshold parameter (which was set to 0.5 in our experiments so far), we further compared *'Learn and Vote'* to the method of Sachs et al. using the threshold-independent performance visualization, the receiver operating characteristic (ROC) curve (Figure 5.5a). We can see that the area under ROC in our approach is more than theirs for the experiment on the *flow cytometry* data. The comparison on the two studies on Asia dataset (*asia_mut1* & *asia_mut2*) shows that including more experiments by *informative* targets improves the performance. However, choosing which intervention is *informative* in an *unknown* network structure is a challenging task, which will be a future extension of this work.

6.4 Effect of Sample size

Figure 5.6 shows the performance of our method vs Sachs et al.'s method by varying the sample sizes extracted from each experiments. We observe that in case of very small samples per exper-

iment, the learning from pooled data gives a better result. For the Asia network having 8 nodes, learning from 20 data points from each experiments gives a non-significant result (Figure 5.6c). Hence, in case of less number of available data it is a good idea to combine them irrespective of experimental conditions. However, for large enough sample data pooling will raise the issue of false discovery. In this work, we randomly sampled 'equal' number of data points from each experiments to prevent biasing towards a particular experiment. Future work will deal with the case of uneven samples of data from different experiments.

6.5 Conclusion

In this paper, we addressed the issue of *false causal discovery* which is observed when we pool data from two or more experiments having different joint distributions caused by *uncertain* interventions. We provided a benchmark for causal network learning methods with observational and interventional experiments having uncertain interventions. We showed by evaluating several state of the art causal learning algorithms that combining data from multiple experiments could result in a large number of false positive causal arcs.

We presented our new approach, *'Learn and Vote'*, which avoids pooling data from multiple experiments and instead combines the weighted graphs learned separately from each experiment. Our approach significantly reduces the number of false positive arcs and achieves superior F1 scores.

Our research motivates the need to focus on the uncertain and unknown effects of interventions to learn high precision causal networks from experimental data. Since most causal learning algorithms assume *perfect* interventions, they might not be working as well as we expect on real world domains having uncertain effects.

Bibliography

[1] Mark A Beaumont and Bruce Rannala. The bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251, 2004.

[2] Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, pages 247–256. Springer, 1989.

[3] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2-3):213–244, 1997.

[4] Judith S Brook, Kerstin Pahl, and Patricia Cohen. Associations between marijuana use during emerging adulthood and aspects of the significant other relationship in young adulthood. *Journal of child and family studies*, 17(1):1–12, 2008.

[5] Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8(10):R219, 2007.

[6] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

[7] Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, pages 415–423, 2010.

[8] Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, pages 415–423, 2010.

[9] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[10] Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125. Morgan Kaufmann Publishers Inc., 1999.

[11] David Danks, Clark Glymour, and Robert E Tillman. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, pages 1665–1672, 2009.

[12] Marek J Druzdzel. The role of assumptions in causal discovery. 2009.

[13] Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, pages 107–114, 2007.

[14] Ronald Aylmer Fisher. The design of experiments. 1935.

[15] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[16] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers & operations research*, 13(5):533–549, 1986.

[17] York Hagmayer, Steven A Sloman, David A Lagnado, and Michael R Waldmann. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, pages 86–100, 2007.

[18] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, and Richard A Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Biocomputing 2002*, pages 437–449. World Scientific, 2001.

[19] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

[20] 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.

[21] David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. In *Innovations in Machine Learning*, pages 1–28. Springer, 2006.

[22] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[23] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[24] Subramani Mani, Peter L Spirtes, and Gregory F Cooper. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*, 2012.

[25] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.

[26] Richard J Orton, Oliver E Sturm, Vladislav Vyshemirsky, Muffy Calder, David R Gilbert, and Walter Kolch. Computational modelling of the receptor-tyrosine-kinase-activated mapk pathway. *Biochemical Journal*, 392(2):249–261, 2005.

[27] Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

[28] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[29] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[30] Dana Pe'er, Aviv Regev, Gal Elidan, and Nir Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl_1):S215–S224, 2001.

[31] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

[32] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[33] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[34] Thomas W Schoener. The anolis lizards of bimini: resource partitioning in a complex fauna. *Ecology*, 49(4):704–726, 1968.

[35] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[36] Tomi Silander and Petri Myllymaki. A simple approach for finding the globally optimal bayesian network structure. *arXiv preprint arXiv:1206.6875*, 2012.

[37] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. adaptive computation and machine learning, 2000.

[38] Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506. Morgan Kaufmann Publishers Inc., 1995.

[39] Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521. Morgan Kaufmann Publishers Inc., 2001.

[40] Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521. Morgan Kaufmann Publishers Inc., 2001.

[41] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

APPENDICES

Appendix A: Redundancy

This appendix is inoperable.