Pooling vs Voting: An Empirical Study of Learning Causal Structures

Meghamala Sinha
Oregon State University
Corvallis, Oregon, 97330
sinham@oregonstate.edu

Prasad Tadepalli
Oregon State University
Corvallis, Oregon, 97330
tadepall@oregonstate.edu

Stephen A. Ramsey
Oregon State University
Corvallis, Oregon, 97330
stephen.ramsey@oregonstate.edu

Abstract

In this paper we present a novel way of combining information from multiple interventional experiments with observations to learn more accurate causal networks. While learning causal network by pooling data from different experiments is common, this paves the way for *false causal discoveries*, if the effects of interventions are *uncertain*. 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.

Introduction

The importance of causal modeling in science, engineering and humanities is remarkable due to its utility in action planning, prediction and diagnosis (Pearl 2003; Spirtes, Glymour, and Scheines 2000). A primary goal in causal modeling is to discover *"causal"* interactions of the form $A \rightarrow B$, where the arrow indicates that the state of A influences the state of B. Causal models can be fit to passive observational measurements (*"seeing"*) as well as measurements after performing external interventions (*"doing"*).

The inability of observational studies to discriminate between Markov-equivalent structures motivates studies that combine observational data with interventional data (Hagmayer et al. 2007). Despite this advantage, learning causal networks from a mix of observational and experimental studies is a significant challenge. Data collected after different experiments might not be identically distributed 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 (Eaton and Murphy 2007). 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 across experiments can lead to misleading changes in correlation. Eberhardt (2008) described two problems: a) *Independence to Dependence*: X and Y, two independent variables in a structure before and after an intervention, become dependent when the samples are pooled and b) *Dependence to Independence*: X and Y, dependent in an observational study, become independent when pooled with an interventional study. This problem occurs even when interventions are *perfect*. This generates a problem of *false causal discovery* which we address in our work.

Due to the above issues, it is important to consider how to handle *uncertain* interventions while learning causal structures. 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 assumes *perfect* interventions, which raises concerns 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 medical tests and panic. 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.

Motivation

Related Work

Popular *Constraint-based* causal learning methods like PC (Spirtes, Glymour, and Scheines 2000), FCI (Spirtes, Meek, and Richardson 1995), etc. uses the entire dataset to learn causal networks using conditional independence tests. Similarly, *Score based* methods like GES, GIES (Hauser and

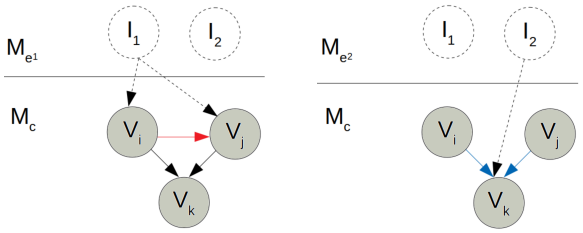
Bühlmann 2012) compute a score for the entire dataset to evaluate the best fitting candidate network. Both type of algorithms were originally designed to infer causal network using single observational dataset. They do not take into account the partitioning of the data based on the different experiments. In this section, we describe some extended works which address the different contexts behind the experiments.

- **Data Pooling Methods:** These methods learns a single causal graph by pooling data from different experiments. (Cooper and Yoo 1999) first provided a score-based causal learning algorithm by combining data from across various experiments, each having *perfect* and *known* targets of intervention. This idea was later re-defined by (Eaton and Murphy 2007) to handle *soft* interventions or *mechanism changes* (Tian and Pearl 2001). The causal invariance property across environment changes was used by (Claassen and Heskes 2010a). Some practical applications of these methods in biology were studied in flow cytometry data set (Sachs et al. 2005) (Cooper and Yoo 1999) and yeast transcriptional regulatory network (Chen, Emmert-Streib, and Storey 2007). A recent approach called Joint causal inference (JCI) (Mooij, Magliacane, and Claassen 2016) takes into account the data generated from different conditions and introduces additional context variables before pooling.
- **Network Combination Methods:** These methods learn information separately from each experiment and combine them to learn a single graph. The ION-algorithm (Danks, Glymour, and Tillman 2009) integrates locally learned causal networks having overlapping variables. (Triantafillou and Tsamardinos 2015) proposed a constraint based algorithm, COmbINE, which estimates dependencies and in-dependencies across separate experiments. However, both these methods assume a single underlying causal structure that accounts for all observed causal dependencies. It is difficult to achieve this in reality when the experimental conditions changes across experiments. The MCI (Claassen and Heskes 2010b) algorithm is a constraint based method that exploits the 'local' aspect of causal Y-structures (Mani, Spirtes, and Cooper 2012) which is sufficient to explain the independencies between two variables regardless different distribution.

In this work, we use the latter approach and describe *'Learn and Vote'*, a score based Bayesian method to learn causal network by learning causal arc-strengths from each experiments and combining the results.

False Causal Discoveries for Pooling

Our work is motivated by the fact that perturbations effecting two or more variables in a causal model M_c can lead to spurious dependencies or independencies. We show two such cases in Figure 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, as shown in Figure 1a. We represent the external perturbation caused by experiments as an external model M_e containing a set of unobserved policy variable



(a) False dependence (b) False independence

Figure 1: Problem with Pooling: Dashed arrows are interventional effects. Solid black are True Positives (TP). Red are False Positives (FP). Blue are False Negatives (FN).

$I_1, I_2 \ldots I_n \in I$. 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 1. *False causal dependence:* Two or more variables, say V_i, V_j, which are not causally related, are effected by a common intervention (I_1 in Figure 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 $P_1(V \subset M_{T_1})$. Pooling data from such different distributions may lead to spurious correlations between independent variables.

Definition 2. *False causal independence:* Intervening on a child node with causal parents removes all incident arrows and cuts off the causal influence. Pooling data from such models nullifies the true causal dependence of parents. This generates a new model $M_{T_2} = M_c + M_{e^2}$ with changed distribution $P_2(V \subset M_{T_2})$. Pooling various such experiments on V_k, shown in Figure 1b will dominate over other experiments having the causal relations $V_i \rightarrow V_k$ and $V_j \rightarrow V_k$.

These above mentioned cases could be shortcomings when pooling data to learn causal networks.

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.

Scoring Function

The interventional effect is incorporated in the score component of each node by modifying the Bayesian Dirichlet equivalent uniform score (BDeu) (Heckerman, Geiger, and Chickering 1995; Cooper and Yoo 1999; Pe'er et al. 2001).

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"*, (Pearl 2003) 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:

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

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 (Glover 1986) 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)
```

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 (Scutari 2009).

Application on Biological Signalling Networks

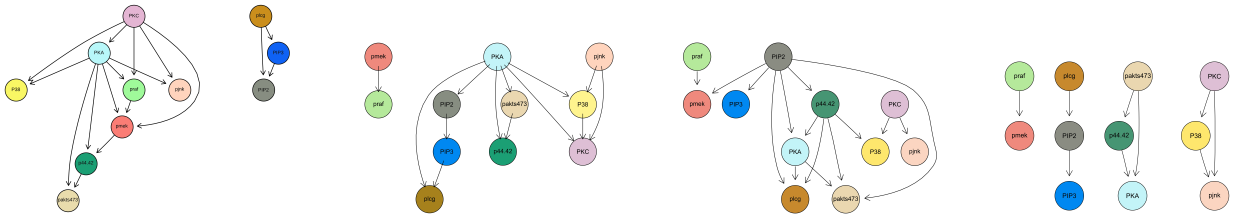
Cell signaling networks are a type of causal network in which proteins or other molecular species modify or influence the state of their *"child"* proteins or molecules. Such networks are amenable to both observational and interventional experiments. Sachs et al. (2005) inferred the signaling pathway and novel causal interactions in human CD4+ T-cells, using a Bayesian network approach (Figure 2a). They carried out nine experiments, two observational and seven interventional, 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). They did not have any *false positives* (FP=0). They also showed how including interventions into observations improves accuracy. We re-analyzed Sachs et al. approach twice, first using observational samples only (Figure 2b) and then using an equal number of samples comprising 50% observational and 50% interventional data. (Figure 2c) illustrates this point by being much closer to the ground truth.

However, like most causal discovery approaches, the methods (used in the Sachs et al.'s study and in our reanalysis) assume *perfect* intervention. Such a perfect intervention modelling is often not consistent in biological experiments like gene knockouts. In the Sachs et al.'s study, we know the nominal target of each of the reagents, but they might affect other variables. In such cases, using causal inference methods assuming *perfect* intervention with known targets can detect spurious interactions.

Comparative Studies

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. 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 (Sachs et al. 2005) 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 (Eaton and Murphy 2007).
- **Lizards:** This is a real-world dataset having 3 variables representing the perching behaviour of two species of lizards in the South Bimini island (Schoener 1968). We generated one observation and two interventional studies.



(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 2: (a) Network inferred by (Sachs et al. 2005) (b) Network inferred from two observational experiments (c) Network inferred from pooling data from an observational and an interventional experiment d) Network inferred from "Learn and Vote" using the same experiments as (c). The structure learning statistics used are True Positive (TP), False Positive (FP) and False Negative (FN).

- **Asia:** This is a synthetic network of 8 variables (Lauritzen and Spiegelhalter 1988) 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 (Beinlich et al. 1989). 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 (Binder et al. 1997). 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.

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 3a-3e.

- **PC:** The observational experiments were used to evaluate the equivalence class of a DAG using the PC algorithm (Spirtes, Glymour, and Scheines 2000). 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 (Hauser and Bühlmann 2012) to estimate Markov equivalence class of DAG from observational and interventional data. It works by maximizing a scoring function (l_0-*penalized Gaussian maximum likelihood estimator*) in 3 phases, i.e., *addition*, *removal* and *reversal* of an arrow, till the score improves.
- **GIES:** This algorithm (Hauser and Bühlmann 2012) extends the greedy equivalence search (GES) algorithm (Chickering 2002) to a generalized version that includes interventional data into observational data.
- **Globally optimal Bayesian Network:** This is a score-based dynamic programming approach (Silander and Myllymaki 2012) 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., (2016) 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.

Analysis of Results

Table 1 summarizes the results of the different structure learning algorithms over all the datasets.

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*.

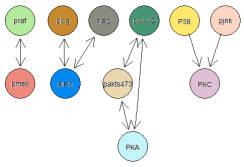
Our approach outperformed all the baselines in five out of nine studies in terms of *precision*, with the ICP method having second best performance. The positive predictive rate of our approach is higher for small or medium sized networks (less than 20 nodes) but comes down as the size of the network 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.

Network inference results on Sachs et al.'s dataset

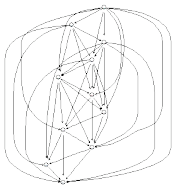
Here, we compare the graphs learned using our approach in Figure 3g with that of the Sachs et al. network inference method in Figure 3f on their cell signaling dataset.

Table 1: Comparative Results

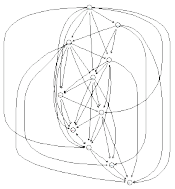
Dataset	*Metric*	*Causal Discovery Algorithms*						
		PC	GDS	GIES	ICP	simy	Sachs et al	Learn and Vote
Flow Cytometry	Precision	0.5714	0.4186	0.377	**1**	0.4222	0.68	0.89
	Recall	0.4	**0.9**	0.85	0.45	0.95	0.85	0.89
	F1 score	0.47	0.572	0.522	0.62	0.584	0.7558	**0.89**
Lizards	Precision	**1**	**1**	**1**	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**



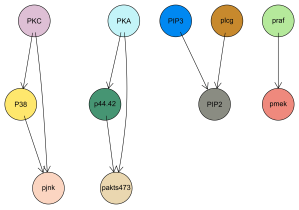
(a) TP:8, FP:6, FN:12



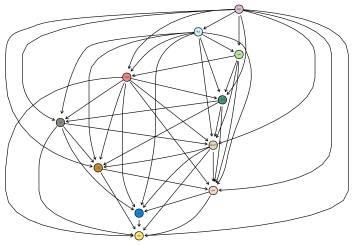
(b) TP:18, FP:25, FN:2



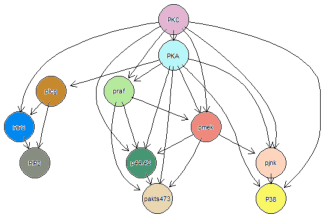
(c) TP:17, FP:28, FN:3



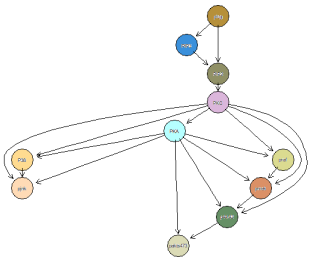
(d) TP:9, FP:0, FN:11



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

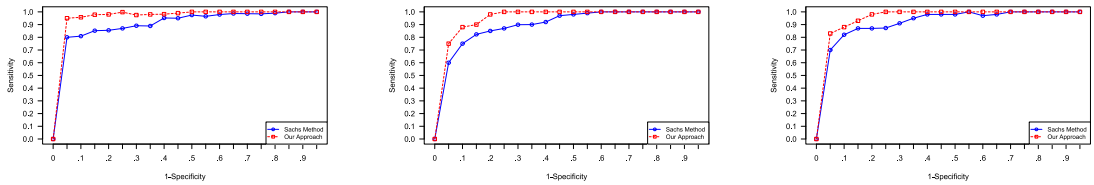


(f) TP:17, FP:8, FN:3



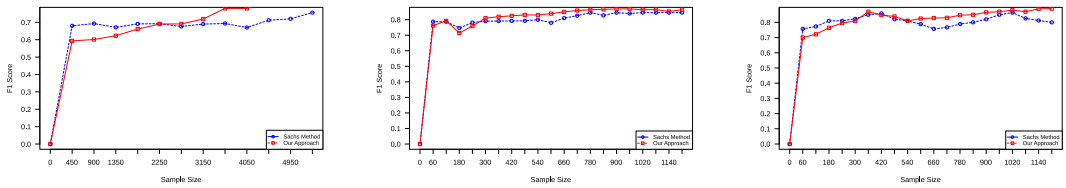
(g) TP:18, FP:2, FN:2

Figure 3: 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 4: ROC plot for comparing results over various datasets



(a) Flow Cytometry (b) Asia_mut1 (c) Asia_mut2

Figure 5: Sample size vs F1 score plot for comparing results over various datasets

The Sachs et al.'s method resulted in 8 *false positive* arcs, 3 *false negative* arcs, and 17 *true positive* arcs (Figure 3f). 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].

Figure 3 shows the networks inferred by the seven inference algorithms on the Sachs et al.'s dataset. The greedy algorithms (Figure 3b, fig. 3c, fig. 3e) 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 3d). We also contrast the performance of the PC algorithm by working only on the observational data. we can see from Figure 3a 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 2c and 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.

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 4a). 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.

Effect of Sample size

Figure 5 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 experiment, 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 5c). 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.

Conclusions

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

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

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.

Acknowledgements

We acknowledge the support of ARO under contract W911NF-19-1-0148, DARPA under contract N66001-17-2-4030, and NSF under grant number IIS-1619433.

References

Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; and Cooper, G. F. 1989. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*. Springer. 247–256.

Binder, J.; Koller, D.; Russell, S.; and Kanazawa, K. 1997. Adaptive probabilistic networks with hidden variables. *Machine Learning* 29(2-3):213–244.

Chen, L. S.; Emmert-Streib, F.; and Storey, J. D. 2007. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology* 8(10):R219.

Claassen, T., and Heskes, T. 2010a. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, 415–423.

Claassen, T., and Heskes, T. 2010b. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, 415–423.

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

Danks, D.; Glymour, C.; and Tillman, R. E. 2009. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, 1665–1672.

Eaton, D., and Murphy, K. 2007. Exact bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, 107–114.

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

Hagmayer, Y.; Sloman, S. A.; Lagnado, D. A.; and Waldmann, M. R. 2007. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation* 86–100.

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

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3):197–243.

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

Mani, S.; Spirtes, P. L.; and Cooper, G. F. 2012. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*.

Mooij, J. M.; Magliacane, S.; and Claassen, T. 2016. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*.

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

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(5):947–1012.

Pe'er, D.; Regev, A.; Elidan, G.; and Friedman, N. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17(suppl_1):S215–S224.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721):523–529.

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

Scutari, M. 2009. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*.

Silander, T., and Myllymaki, P. 2012. A simple approach for finding the globally optimal bayesian network structure. *arXiv preprint arXiv:1206.6875*.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. Causation, prediction, and search. adaptive computation and machine learning.

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

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

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