Causal Inference from Protein Expression Data
CS6347

Harsha Kokel hxk171430

Michael Skinner mas140130

1 Introduction

In the field of machine learning, causal discovery is concerned with the etiologic dependence among variables in a set of data (Eberhardt, 2017). Whereas these causal relationships are typically learned from direct intervention and experimentation, our aim here is to discover such relationships directly from observed data. That "correlation does not imply causation" suggests that standard analysis of Bayesian graphical models (BGMs), in which the correlation of variables is learned, must be modified to infer etiologic connections. To illustrate that we cannot simply infer causal relationships from our probability densities, consider Figure 1 (obtained from the example in (Eberhardt, 2017), which we repeat here to clarify).

Suppose we are investigating a causal relationship between wine drinking and heart disease. In panel (a), we postulate a direct causal relationship between these variables, which

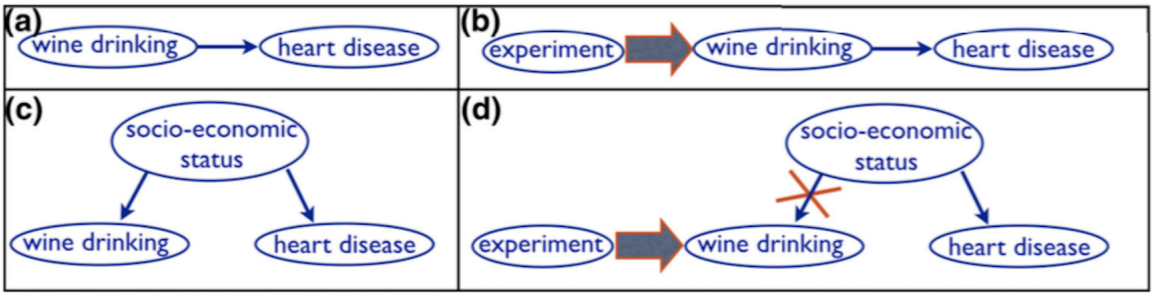

Fig. 1: Inference and Causation

is consistent with the data obtained from the literature. If we intervene experimentally on wine drinking (by say, encouraging moderate wine drinking in a population as in panel (b)), and this is in fact the causal relation that holds, we will see the expected reduction in heart disease; we can conclude with some reliability that there is a causal relationship. However, suppose there is a confounding factor, such as socio-economic status, that mutually causes both an increase in wine consumption and a reduction in heart disease, as depicted graphically in panel (c). Although Bayesian inference in this model demonstrates a probabilistic relationship between the variables, when we intervene on wine drinking in this model, there will be no change in heart disease. We see from this example very clearly how in the setting of confounding factors, although we can infer *correlative* relationships between variables, we cannot easily make *causal* inferences.

Owing to the inability of the probability distribution to specify the causal relations, a separate causal graph must be defined. The directed graph $G = \{V, E\}$ is defined over the variables V and edges E. A causal graph is not designed to represent independence relationships between variables, as in standard Bayes networks; rather, we define causal relations such that an edge from a parent vertex to a child vertex implies a direct causal relation from parent variable to the child variable. The notions of *ancestor* and *descendant* variables are defined as in the case of Bayesian probabilistic networks, and can be inferred from the graph.

The task of causal discovery over a set of variables is the unveiling from the data as much as possible about whatever causal relations hold. This may be restated as the inference of a causal graph from the probability distribution. Whereas we hold that there is a difference between causation and correlation, we certainly use correlation results in this task. Indeed, one of our objectives is to know under what circumstances can we infer causal influence from the local probability distributions available in the Bayesian graph. There are additional concepts connecting the notions of correlation and causation. For example, it is generally the case that if correlation is absent between variables, then causation will also be absent.

In causal analysis, there are the *causal Markov* and *causal faithfulness* conditions. The *causal Markov* condition states that every vertex V is probabilistically independent of its non-descendants given its parents. *Causal faithfulness* is a simplifying assumption stating that if a variable X is independent from Y given a conditioning set of variables S, then X and Y are d-separated in the graph. These assumptions assist in the causal task, since they allow us to use d-separation (in the causal graph) to exclude causal relations between

variables; whether there is d-separation must be inferred from the conditional probability tables obtained from the data. Finally, we assume that the causal graph is acyclic, as in the directed Bayesian network.

Although there are important connections between the joint probability distribution and the causal structure, there are significant differences as well. For example, whereas *statistical* information flows in any direction through the graph, *causal* information only flows one way from parent to child. So if we want to estimate the causes from effects within a given graph, we need to exclude non-causal flows of information. If we want to learn the causal structures, we must learn which ones are consistent with the conditional dependencies supported by the data.

In view of this, Pearl conceived the "do operator" to distinguish the inference and causal tasks (Pearl, 2009). In particular, the inference, or *observational*, conditional probability between sets of variables X and Y is $\Pr(Y|X)$, and in the causal setting is denoted $\Pr(Y|do(X)$ to indicate we are trying to infer what effect on Y follows after intervening on X. This is termed the *interventional* probability. This nomenclature emphasizes the fact that simply obtaining $\Pr(Y|X)$ from our observational data (that is, filtering the data on a particular value of X) is not the same as determining how the value of Y responds to an alteration in the value of X; it is possible that in this case, X may have been set to a value by a process that also effects the value of Y. The task is to know how to use the local probability distributions inferred from the data to learn the interventional (sometimes termed *counterfactual*) probabilities, since we are unable to actually perturb X experimentally.

When computing $\Pr(Y|do(X))$ from observational data, the main objective is to control for confounding variables, such as an observed or unobserved common cause illustrated in Figure 1. To do this, we first perform "graph surgery" on the Bayesian directed graph by removing all of the edges incident on X, to ablate the flow of information through any vertices that may be mutually causal to both X and Y. Then, we have to control for any information paths between the variables of interest other than the direct edge between them, commonly termed *back-door paths*. The back-door phenomenon is illustrated in Figure 2, (obtained from the textbook by Shalizi (Shalizi, 2013)).

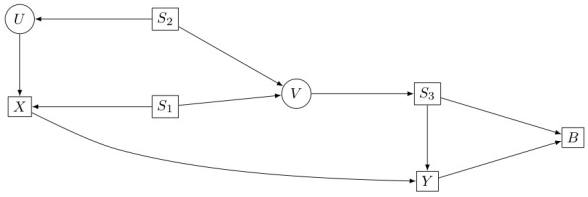

Fig. 2: Illustration of the back-door criterion for identifying the causal effect of X on Y.

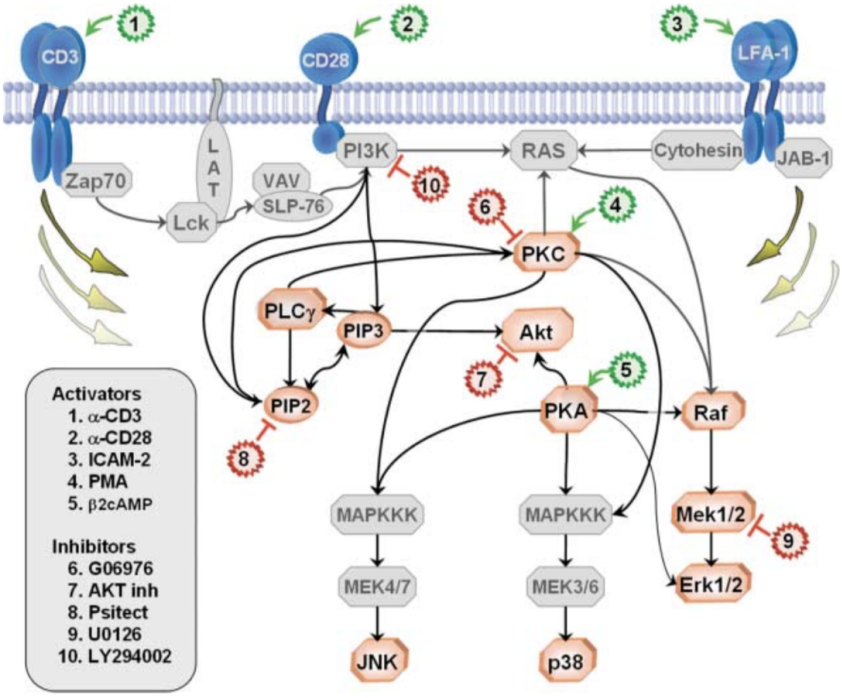

Fig. 3: Network of the proteins evaluated by Sachs *et al*, with the inhibitory and stimulatory interventions labeled with red and green respectively (figure from (Sachs et al., 2005)).

Here, we see that there are several alternate paths by which information may flow between X and Y. The back-door criterion is satisfied when there exists a set of variables S where S blocks every back-door path from X to Y, and no node in S is a descendant of X. When these conditions are met, the variable set S can be summed over, and the causal probability is computed by equation 1.

$$\Pr(Y|do(X=x)) = \sum_{S} Pr(Y|X=x, S=s)Pr(S=s) \tag{1}$$

.

This expression contains only observational conditional probabilities, not counterfactuals, and can be obtained from the observational data. This expression is derived by Pearl (Pearl, 2009).

In this report, we applied these techniques of causal inference from observational data to investigate the interaction between proteins in a biological system.

2 Cellular protein activation

Biological cells commonly interact with their environment through a system of protein receptors on the cell membrane that bind to their cognate extracellular protein ligands. This interaction alters the receptor structure to activate its intracellular partners by phosphorylating some of the amino acid residues, similarly activating downstream intracellular proteins and ultimately creating the desired biological response.

These complicated signaling cascades have been dissected by classical techniques of chemistry and molecular biology, usually requiring the chemical lysis of cells in culture and measuring the average phosphorylation status of a population of cells grown in culture; that only a population of cells can be evaluate has hindered the precision of the findings. For this reason, newer assays of intracellular protein receptor phosphorylation status have been devised to allow the automated investigation of individual cells, so many more experiments can be run not requiring the consideration of an entire population of cells. Each cell may be regarded as an example of the data set, with a vector of phosphorylation levels for a set of intracellular proteins, each of which is regarded as a random variable.

The availability of these data led to the natural consideration of BGMs to represent and analyze the intracellular signaling network, in which the vertices of the graph represent variables whose values correspond to the activation level of particular proteins and where edges correspond to interactions between proteins. A depiction of the consensus phosphorylation circuitry in a population of human immune cells, obtained experimentally using the classical assay methods, is presented in F

The interaction between protein levels in biological cells can be described as a network in which some molecules, when their phosphorylation status is altered, regulate the phosphorylation of other molecules in the system.

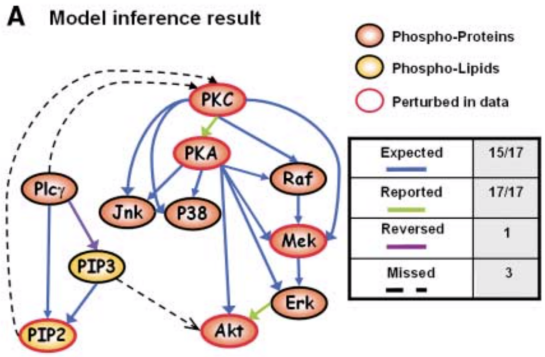

Fig. 4: Inferred protein network.

A landmark report demonstrated the usefulness of the BGM framework in eliciting such biologic signaling pathways (Sachs et al., 2005). In this study, Sachs *et al* experimentally altered the phosphorylation of proteins in human immune cells using chemical inhibitors or stimulators to selectively activate or ablate regions of the circuit. The resulting

graph is depicted in Figure 4. (Peer, Regev, Elidan, & Friedman, 2001).

Impressively, they largely confirmed the results obtained using the classical techniques. They elicited 17 edges, of which 15 were expected. They missed 3 edges described in the literature. These results confirm the value of Bayesian networks for investigating biologic signalling pathways.

3 Experiments

We applied Pearl's causal techniques to investigate the causative interactions between protein phosphorylation in human immune cells. In particular, we studied a subset of the molecules investigated by Sachs *et al*, using their protein expression data (Sachs et al., 2005). The subset of proteins studied are shown in Figure 5. Note that this subset is a causally sufficient set. We can see that some proteins fulfil the backdoor criterion, we analyze those interactions using back-door adjustment formula mentioned in 1.

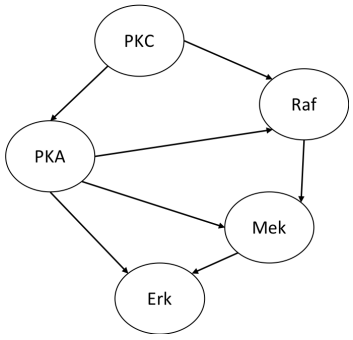

Fig. 5: Protein sub-network

3.1 Data preparation

The protein phosphorylation and expression data[1] used by Sachs and her colleagues consist of information from individual human immune cells cultured in the presence of chemical phosphorylation inhibitors or stimulators, and then exposed to fluorescent antibodies binding selectively to particular phosphorylated moities, which stimulates fluorescence. Then, when analyzed by a high-speed Fluorescence Activated Cell Sorting (FACS) machine, the phosphorylation signals for specific proteins from individual cells (see Figure 3) are obtained.

As with Sachs and her colleagues, we also discretized the fluorescence levels into "low","medium", and "high" categories, but due to lack of details in the paper, we were not able replicate their discretization strategy. So, we pooled of all the data from all of the different experiments, to obtain probability distribution over all of the possible fluorescence levels. We excluded the cells exhibiting fluorescence levels greater than 3 standard deviations from the mean, as mentioned in Sachs et al., 2005. Then, under the simplifying

[1] Downloaded from http://www.sciencemag.org/cgi/content/full/308/5721/523

assumption that the data were distributed according to a mixture of Gaussian distributions, we used MATLAB to evaluate for the best parameters for mixtures (see Figure 5 for a representative histogram). At times, insisting on a mixture of three distributions seemed to return a mixture in which one of the mixture parameter was too low; in these cases we looked for four mixtures (essentially splitting one of the distributions) to obtain a more balanced distribution. We chose the discrete boundaries by inspection.

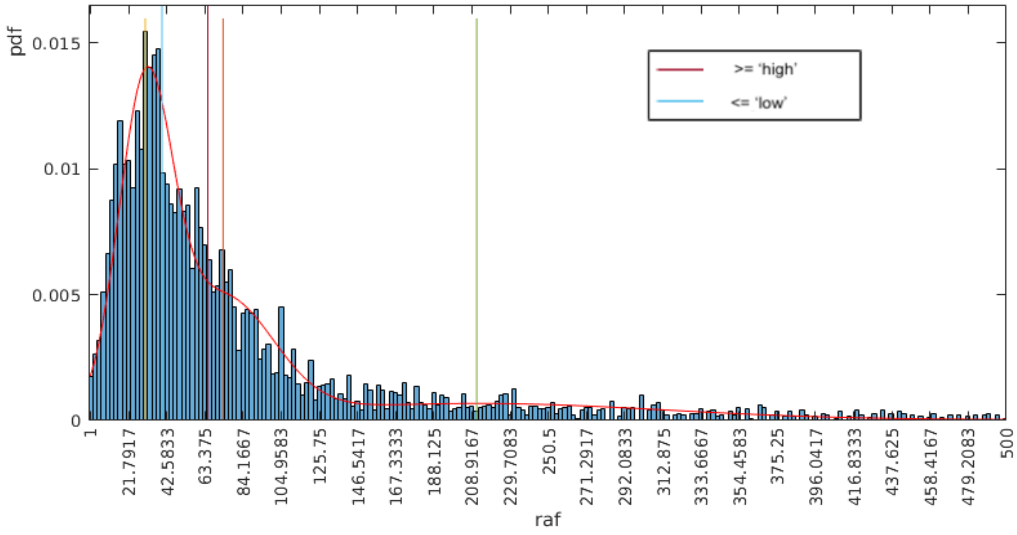

Fig. 6: Probability distribution of fluorescence values of Raf. Yellow, Orange and Green vertical lines indicate the means of the 3-gaussian mixtures. Blue vertical line indicates the upper boundary selected for level 'low', and Red vertical line indicates the lower boundary selected for the level 'high'. Any fluorescence value between these two boundaries was assigned to level 'medium.'

3.2 Causal analysis

An early description of a technique to derive causal inferences from genetic expression data was reported by Friedman *et al* (Friedman, Linial, Nachman, & Pe'er, 2000). Here, we performed causative inference in a biological system using biological network models described in their report, as well as in Pe'er *et al* (Peer et al., 2001), comparing our results using the techniques of causal discovery discussed above with the published experimental results.

We compared the interventional probability as mentioned in equation 1 with the con-

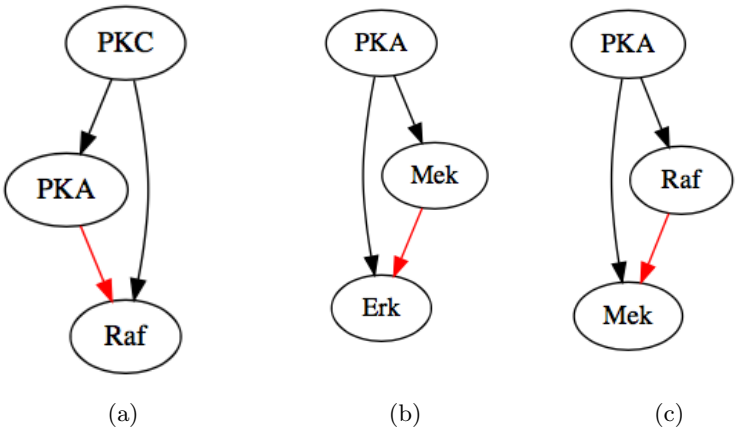

Fig. 7: Backdoor analysis, red arrow indicates the hypothesized edge.

ditional probability for the hypothesized causal edge for the graphs shown in Figure 7(a). If a causal edge exists then the interventional probability should not equal the conditional probability. However, on running the experiments for the observational data we see that the interventional probability and conditional probability are almost equal for all of these relationships, shown in Tables 1.

From these results, we conclude that **PKA** is independent of **PKC**; we did not confirm the dependency between these molecules discovered by Sachs and demonstrated in Figure 4. So, we remove the edge from **PKA** to **PKC** in figure 5 and we analyze two other back-door phenomenon, shown in figure 7(b) and 7(c). Tables 2 and 3 shows all of these probabilities are equal, suggesting that there really is no back-door communication between **PKA** and **Erk**; neither is there a back-door communication between **PKA** and **Mek**.

To further investigate these findings, we employed the PC algorithm for causal analysis, another technique to infer the causal graph (Spirtes & Glymour, 1991). The resulting graph is presented in Figure 8. We can see that the our data do not allow the inference of causal relations other than those existing between **PKA** and **Erk**, and between **Raf** and **Mek**; again, we are unable to reconstruct the causal structure discovered through experimental interventions by Sachs *et al.*

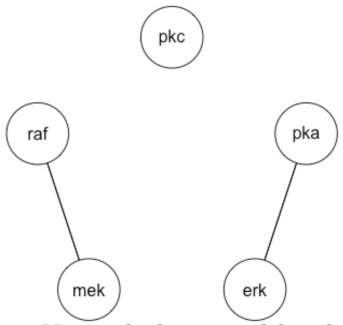

Fig. 8: Network discovered by the PC algorithm.

PKA	RAF	P(RAF|PKA)	P(RAF|do(PKA))
low	low	0.3036	0.2984
low	medium	0.3929	0.4031
low	high	0.3036	0.2984
medium	low	0.3469	0.3466
medium	medium	0.3010	0.3022
medium	high	0.3520	0.3512
high	low	0.3012	0.2998
high	medium	0.3901	0.3909
high	high	0.3086	0.3094

Tab. 1

Mek	Erk	Pr(Erk|Mek)	Pr(Erk|do(Mek))
low	low	0.6912	0.6921
low	medium	0.2158	0.2171
low	high	0.0930	0.0909
medium	low	0.6462	0.6413
medium	medium	0.2358	0.2371
medium	high	0.1179	0.1215
high	low	0.6479	0.6475
high	medium	0.2254	0.2182
high	high	0.1268	0.1343

Tab. 2

Raf	Mek	Pr(Mek|Raf)	Pr(Mek|do(Raf))
low	low	0.9382	0.9385
low	medium	0.0582	0.0576
low	high	0.0036	0.0039
medium	low	0.7483	0.7463
medium	medium	0.2282	0.2326
medium	high	0.0235	0.0210
high	low	0.3179	0.3169
high	medium	0.4571	0.4599
high	high	0.2250	0.2232

Tab. 3

4 Conclusion

There are a number of reasons to explain our failure to duplicate the results of Sachs *et al.* First, for the sake of computational tractability, we rather arbitrarily discretized the phosphorylation levels for the proteins, and our "high", "medium", and "low" levels probably did not conform to those of Sachs. Moreover, whereas those investigators in their analysis imputed the phosphorylation levels to the "low" and "high" categories for molecules under inhibition or stimulation, respectively, we used the levels as they were presented in the data. Finally, it is certain that some of the assumptions we made in performing the causal do-calculus do not hold in biologic systems. For example, there are multiple feedback circuits between the molecules, so that the assumption of acyclicity does not hold. The drawing in Figure 3 is necessarily limited to the molecules that are easily manipulated and easily measured. In truth, there are thousands of molecules involved in these circuits; thus, our "graph surgery" and management of the known alternate pathways between the parent and child molecules are almost certainly failing to control all of the confounding processes.

For these reasons, we were unsuccessful at accurately discovering causal molecular relationships solely from the evaluation of observational data; there remains the great need for interventional experiments for the elicitation of causal relationships between proteins in biological systems.

References

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, *3*(2), 81–91.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601–620.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peer, D., Regev, A., Elidan, G., & 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., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523–529.

Shalizi, C. (2013). *Advanced data analysis from an elementary point of view.* Citeseer.

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, *9*(1), 62–72.