Learning Causal Protein-Signaling Networks

Jin Tian JTIAN@CS.IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Akshay Deepak AKSHAYD@CS.IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Abstract

Graphical Models have been widely used for modelling causal relationships. We use causal Bayesian networks to model protein signaling networks and use the Bayesian approach to learn the network structure from mixed observational and experimental data. We compute the maximum a posteriori (MAP) network for a biological data set originally analyzed by Sachs et al. (2005).

Keywords: Causal discovery, Bayesian Network

1. Introduction

We study a biological data set CYTO presented by Sachs et al. (2005). The task is to learn a protein signaling network from multicolor flow cytometry data that recorded the molecular activity of 11 proteins under various experimental conditions. The CYTO data consists of roughly 700 to 900 samples per experimental condition, corresponding to various "interventions" on the system of interest. Sachs et al. (2005) modeled the protein signaling networks as causal Bayesian networks and inferred the network structure from the data using a Bayesian approach. More specifically, a random restart simulated annealing search is applied in the space of DAGs to find the networks with high posterior probabilities and a bootstrap method is used to find edges with high posteriors. The CYTO data was also analyzed by Ellis and Wong (2008) using MCMC in the space of node orderings, and by Eaton and Murphy (2007) using a dynamic programming algorithm that computes the exact edge posterior probabilities under a special graph prior.

In this paper, we also use the Bayesian approach to learn causal Bayesian networks from data. We compute the MAP network (the network with the globally highest posterior probabilities) using the dynamic programming (DP) algorithm in (Silander and Myllymaki, 2006). First we review the Bayesian approach to learning causal Bayesian network structures from interventional data.

2. The Bayesian Approach

Let our problem domain be a set of discrete random variables $V = \{V_1, \ldots, V_n\}$. We assume that a *causal model* (or a causal Bayesian network) over V is a pair $M = <G, \Theta_G>$, where G is a DAG over V, called a causal diagram, and Θ_G is a set of probability parameters. We assume that each variable V_i can take values from a finite domain, $Dm(V_i) = \{v_{i1}, \ldots, v_{ir_i}\}$, where r_i is the number of states of V_i. We use Pa_i to represent the set of parents of V_i in a causal diagram G and $Dm(Pa_i)$ to represent the set of states of Pa_i.

Assume that we have a set of random samples D generated from a causal model $M = <G, \Theta_G>$. In the Bayesian approach, we compute the posterior probability of a causal diagram G given the dataset D as:

$$P(G|D, \xi) = \frac{P(D|G, \xi)P(G|\xi)}{P(D|\xi)}, \tag{1}$$

where ξ represents our background knowledge. We can then compute the posterior probability of any hypothesis of interest by averaging over all possible causal models. Since the number of possible diagrams is exponential in the number of variables n, it is often impractical to sum over all diagrams unless for very small n. One way to deal with this problem is to use the relative posterior probability $P(D, G|\xi)$ as a *scoring metric* and search for diagrams with high scores. In this paper we will assume that $P(G|\xi)$ is a uniform distribution and use the following Bayesian score

$$score(G : D) = \ln P(D|G, \xi) \tag{2}$$

2.1 Bayesian Score

For the case that the dataset D is from a static distribution, closed form expressions for $P(D|G, \xi)$ have been derived (Cooper and Herskovits, 1992; Heckerman et al., 1995). Assuming global and local parameter independence, and parameter modularity, we have that the Bayesian score can be decomposed into the summation of local scores

$$score(G : D) = \sum_{i=1}^{n} score_i(V_i, Pa_i : D). \tag{3}$$

Further assuming Dirichlet parameter priors, we have

$$score_i(V_i, Pa_i : D) = aScore_i(A, D), \tag{4}$$

where the function $aScore_i(A, D)$ is defined as

$$aScore_i(A, D) = \ln \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + N_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + N_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})}, \tag{5}$$

where $\Gamma(\cdot)$ is the Gamma function, N_{v_i,pa_i} is the number of cases in data set D for which V_i takes the value v_i and its parents Pa_i takes the value pa_i, $\alpha_{v_i;pa_i}$ are Dirichlet hyper parameters, $A = \{\alpha_{v_i;pa_i} : v_i \in Dm(V_i), pa_i \in Dm(Pa_i)\}$, and

$$\alpha_{pa_i} = \sum_{v_i} \alpha_{v_i;pa_i}, \; N_{pa_i} = \sum_{v_i} N_{v_i,pa_i}.$$

We use $\prod_{v_i}$ as a shorthand for $\prod_{v_i \in Dm(V_i)}$ and $\prod_{pa_i}$ for $\prod_{pa_i \in Dm(Pa_i)}$. In this paper, we will use the BDe score and assume the following hyperparameters

$$\alpha_{v_i;pa_i} = 1/(r_i q_i) \tag{6}$$

where r_i is the number of states of V_i and q_i is the number of states of Pa_i.

2.2 Bayesian score with interventions

The Bayesian score described above assumes that the dataset D is drawn from a static distribution. We can adapt the score to deal with the situation where we have a number of data sets $D^1, D^2, \ldots$, generated from the same causal structure but under different experimental conditions (Tian and Pearl, 2001b). For example, assume that we have two data sets D^1 and D^2, and D^1 is generated from the causal model $M = <G, \Theta_G>$. Assume that D^2 is generated from M under an ideal intervention on variable V_k that set V_k to a fixed value. Then the Bayesian score is given by (Cooper and Yoo, 1999)

$$score_i(V_i, Pa_i : D^1, D^2) = \begin{cases} aScore_i(A, D^1 + D^2), & i \neq k \\ aScore_i(A, D^1), & i = k \end{cases} \tag{7}$$

If D^2 is generated from the same causal structure G but with different parameters Θ'_G, and we have no knowledge about how the two sets of parameters Θ_G and Θ'_G differ, we may assume that they are independent and we use the following Bayesian score (Tian and Pearl, 2001b):

$$score_i(V_i, Pa_i : D^1, D^2) = aScore_i(A, D^1) + aScore_i(A, D^2) \tag{8}$$

The CYTO data consists of 9 data sets under different conditions. Assume that the data set cd3cd28.xls is generated from the causal model $M = <G, \Theta_G>$. Then we assume that each of the data sets cd3cd28+aktinhib.xls, cd3cd28+g0076.xls, cd3cd28+psitect.xls, and cd3cd28+u0126.xls is generated from M under some ideal intervention. We will consider the data set cd3cd28+ly.xls as generated from a general perturbation rather than an ideal intervention on akt as the actual intervention is not directly on akt. We assume that each of the data sets cd3cd28icam2.xls, cd3cd28+ly.xls, pma.xls, and b2camp.xls is generated from the same causal structure G but with different parameters. In summary, we use the following Bayesian score. For those variables on which no intervention is performed, $V_i \in \{raf, plcg, PIP3, erk, P38, jnk\}$

$$\begin{aligned} score_i(V_i, Pa_i : D) = {} & aScore_i(A, D^{cd3cd28} + D^{u0126} + D^{g0076} + D^{psitect} + D^{aktinhib}) \\ & + aScore_i(A, D^{icam2}) + aScore_i(A, D^{ly}) + aScore_i(A, D^{pma} + D^{b2camp}). \end{aligned} \tag{9}$$

If a variable V_j is set by intervention in data set D^j, to compute the local score of V_j we will assume ideal intervention and simply drop the data set D^j from Eq. (9). For example, for $V_i = PKC$,

$$\begin{aligned} score_i(V_i, Pa_i : D) = {} & aScore_i(A, D^{cd3cd28} + D^{u0126} + D^{psitect} + D^{aktinhib}) \\ & + aScore_i(A, D^{icam2}) + aScore_i(A, D^{ly}) + aScore_i(A, D^{b2camp}). \end{aligned} \tag{10}$$

3. Finding Optimal Structures

Although finding the network structure with the maximum score is NP-hard, it is feasible for small networks. Given a decomposable score as in (3), a best network can be found in $O(n^2 2^n)$ time and $O(2^n)$ space using the DP algorithm in (Silander and Myllymaki, 2006). Note that several networks may have the same best scores and the DP algorithm simply returns one of them. There are $n = 11$ variables in the CYTO data, and a best network can be found in a few seconds.

Once we find a best network, then other best networks with the same scores can be easily identified. Two DAGs are Markov equivalent (or independence equivalent) if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (Verma and Pearl, 1990). Given observational data alone, two Markov equivalent structures are indistinguishable. In fact, the BDe score specified in (4) and (6) satisfies the property that if two neworks are Markov equivalent then they have the same scores (Heckerman et al., 1995). Interventional data can further increase our ability to recover the true causal structure. Specifically, an intervention on variable V_i determines the direction of the edges between V_i and its neighbors. Given a set of observational and interventional data, the extended BDe score as given in (9) satisfies the property that if two networks are Markov equivalent and have the same set of neighbors for each intervened variable then they have the same scores. In particular, we have the following

Property 1 *For the CYTO data, if two neworks have the same skeletons, the same sets of v-structures, and the same sets of neighbors for* $mek, PIP2, akt, PKA, PKC$*, then they have the same scores.*

Using Property 1, we can find all the best networks that are indistinguishable given the CYTO data by the DP algorithm.

4. Experimental Results

The CYTO data measured 11 variables under 9 experimental conditions. The original data were discretized into 3 states (low, medium, and high). We used the discretized data in Sachs et al. (2005) consisting of 600 samples per condition.

4.1 Learning Causation by Detecting Changes

Causal information may be learned by detecting changes in the probability distributions under different experimental conditions, in particular, changes in the marginal probability of each variable (Tian and Pearl, 2001a). It is obvious that an intervention on a variable X in a causal model $M = <G, \Theta_G>$ may potentially alter the marginal probabilities of the descendants of X in G and can not alter the marginals of nondescendants of X. Assume that we have two data sets D^1 and D^2 where D^1 is generated from M and D^2 is generated from M under an intervention on variable X. If the marginal of a variable Y has changed, we conclude that Y is a descendant of X, denoted by $X \rightarrow\rightarrow Y$.

We used the general perturbation data cd3cd28.xls as the base set and compared it with other 8 data sets to detect marginal probability changes in all 11 variables. We used χ^2 test

to detect distribution changes with significance level $\alpha = 0.0005$ (we used a small α value as there are close to 100 χ^2 tests to perform). If $\chi^2 > \chi^2_\alpha$ then we decide "change", else we do not make any conclusion. We drew the following conclusions

$$akt \rightarrow\rightarrow \{PIP2, PIP3, erk, p38, jnk\} \tag{11}$$

$$mek \rightarrow\rightarrow \{PKC, PKA, akt, erk, raf, jnk\} \tag{12}$$

$$PKC \rightarrow\rightarrow \text{all other 10 variables} \tag{13}$$

$$PIP2 \rightarrow\rightarrow \{akt, erk, raf, p38, jnk, plcg, PIP3\} \tag{14}$$

It is clear that there are cyclic causal relations. For example we have $mek \rightarrow\rightarrow PKC$ and $PKC \rightarrow\rightarrow mek$.

4.2 Finding Optimal Structures

The MAP network that maximizes the Bayesian score is shown in Figure 1, which turns out to be unique as there exists no other score equivalent networks. There is large agreement with the results obtained in (Sachs et al., 2005) if considering only skeletons (or undirected edges), and there are also many differences, in particular in the orientation of edges. There are a number of edges in the MAP network that are missing from the model in (Sachs et al., 2005). It appears that the skeleton of the MAP network agrees more with the model (Figure 6c) in (Eaton and Murphy, 2007) than that in (Sachs et al., 2005). It is likely that the orientation differences are due to the cyclic nature of the causal relations in protein signaling networks. We also noticed that the causal directions in the MAP network do not match with those obtained by detecting changes given in Eqs.(11)-(14).

5. Conclusion

We use causal Bayesian networks to model protein signaling networks and computed the MAP network for the CYTO data. It seems that inferring the causal directions using Bayesian network models may not be reliable due to the cyclic nature of the causal relations in protein signaling networks.

References

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

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

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *AI/Statistics*, 2007.

B. Ellis and W. H. Wong. Learning causal Bayesian network structures from experimental data. *J. Am. Stat. Assoc.*, 103:778–789, 2008.

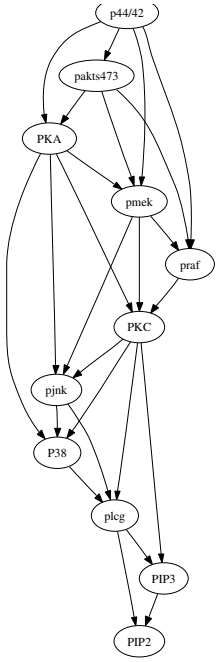


Figure 1: The MAP network.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In *UAI*, 2006.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–2001)*, pages 512–521, San Francisco, CA, 2001a. Morgan Kaufmann Publishers.

J. Tian and J. Pearl. Causal discovery from changes: a Bayesian approach. Technical Report R-285, Department of Computer Science, University of California, Los Angeles, 2001b.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In P. Bonissone et al., editor, *Uncertainty in Artificial Intelligence 6*, pages 220–227. Elsevier Science, Cambridge, MA, 1990.

Appendix B. Pot-luck challenge: FACT SHEET.

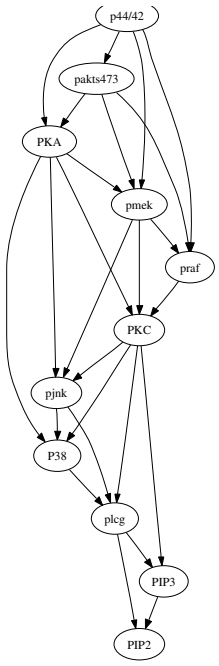
(for a task solved)

Title: Learning Causal Protein-Signaling Networks
Participant name, address, email and website:
Jin Tian and Akshay Deepak, jtian@cs.iastate.edu, akshayd@cs.iastate.edu
Department of Computer Science, Iowa State University, Ames, IA 50011, USA
Task(s) solved: CYTO
Reference:

Method:

- Preprocessing: We used the discretized data in Sachs et al. (2005) consisting of 5400 samples with 600 samples per condition.
- Causal discovery: We used the Bayesian approach to learn causal Bayesian networks from mixed observational and experimental data. We computed the maximum a posteriori (MAP) network using the dynamic programming algorithm in (Silander and Myllymaki, 2006).

Results: The MAP network.



Keywords:

- Causal discovery: Bayesian Network.