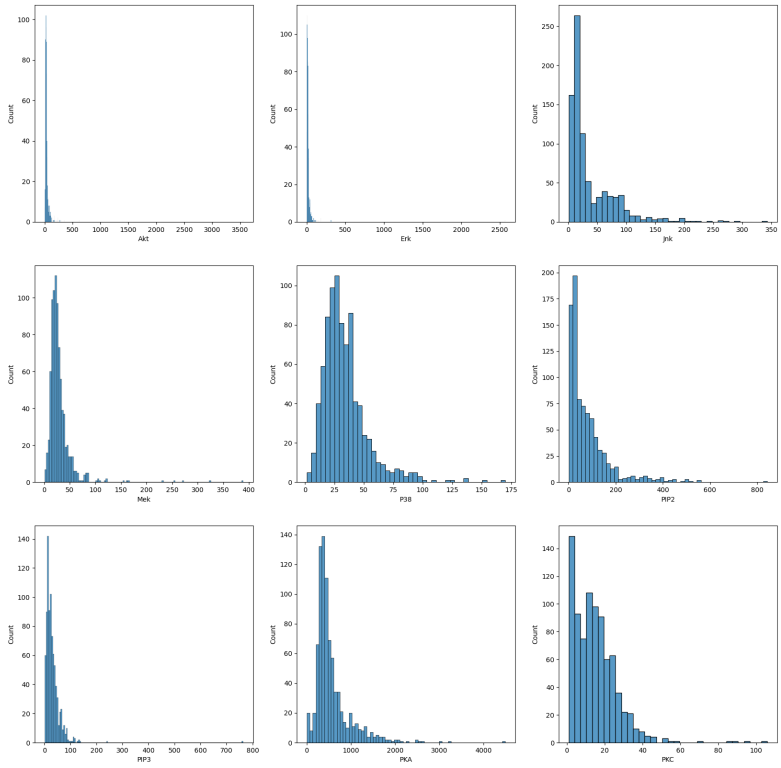
A BAYESIAN NETWORK MODEL FOR CAUSAL PROTEIN-SIGNALING NETWORKS

Exam project for the course *Models for Complex Systems*



1. Introduction

Cells use the presence or absence of signalling molecules as information cues. Extracellular input can trigger a cascade of information flow, in which signaling molecules are chemically, physically, or locationally modified and affect other molecules in the cascade. Different inputs result in different cellular responses. Historically mapping such signalling cascades has involved making intuitive inferences by aggregating results from experiments that study the relationship between single pairs of signalling molecules. It is now understood that signalling cascades networks are complex and as a result cannot be described well by aggregating individual pathways. To properly understand the cellular responses a global multivariate approach is required. One natural approach is to model the whole signalling cascade with a Bayesian network.

In this project we consider an intracellular signalling network among human primary naïve $CD4^+$ T-cells. We apply Bayesian network analysis to multivariate flow cytometry data. The data was collected after the T-cells were exposed to the chemical reagent anti-CD3/CD28. Fifteen minutes after stimulation flow cytometry measurements of 11 phosphorylated proteins and phospholipids were taken from individual T-cells. Figure shows a histogram of the measurements across the different cells. For simplicity we will only consider the data in binned form, were for each protein we divide the data into two states "low" and "high" depending on whether the amount of the protein in an individual measurement is above or below the average across all measurements.

With the Bayesian Network Model we have a tool to explore data of this form and a systematic way to *describe* the signalling pathway.

We consider data of the 9 proteins `Akt`, `Erk`, `Jnk`, `Mek`, `P38`, `PIP2`, `PIP3`, `PKA`, and `PKC`, with data missing for `Plcg` and `Raf`. The data consists of 853 measurements. For ease of notation we denote one observation by

$$\mathbf{X}_j = (X_{1,j}, ..., X_{9,j})$$

where the variables are ordered as above, i.e., $X_{1,j}$ is the j'th observation of `Akt`, $X_{2,j}$ is the j'th observation of `Erk` and so forth. We assume that each observation is an independent draw from the Bayesian network in Figure 1.

2. A generative model

The generative model we will consider is a Bayesian network whose graph is given by Figure 1. All the random variables are binary taking values in $\{0, 1\}$. Also, $X_{10,j}$ and $X_{11,j}$ are unobserved for all $j = 1, ..., n$. We shall simply write $X_1, ..., X_{11}$ and omit the second subscript when we are not talking about any one observation in particular

The CPDs are parametrized by the success probability of a node given its parents. For each $i = 1, ..., 11$ let Pa_i be the parents of X_i and define d_i as the number of nodes in Pa_i. Then for some outcome, $(k_1, ..., k_{d_i}) \in \{0, 1\}^{d_i}$, of its parents, we define

$$p^{(i)}_{k_1,...,k_{d_i}} = P(X_i = 1 | \mathrm{Pa}_i = (k_1, ..., k_{d_i})).$$

Thus, for each node there are 2^{d_i} parameters resulting in a model with a total of $\sum_{i=1}^{11} 2^{d_i}$ parameters. Note that the CPDs can be represented as tables. In particular, if $d_i = 2$ we can define the 2×2 matrix $\Psi^{(i)}$ whose entries are given by

$$\Psi^{(i)}_{k,l} = p^{(i)}_{k,l}$$

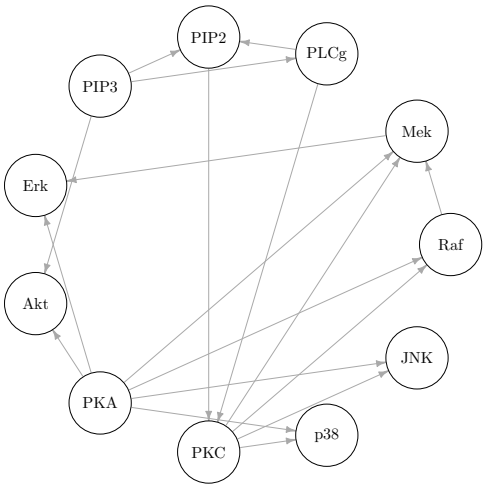


FIGURE 1. Causal DAG representing intracellular signalling network among human primary naïve $CD4^+$ T-cells.

for $k, l = 0, 1$. This then completely characterises the CPD of X_i given its parents. The largest number of parents of any node in our network is 3. The biggest CPD table will therefore be $2 \times 2 \times 2$. On the other hand, we also see that there are nodes without any parents, e.g. X_7 corresponding to `PIP3` in the graph. The CPD table for X_7 will simply by a scalar.

In this part we want to simulate from our model for a given set of parameters $\psi^{(1)}, ..., \psi^{(11)}$. To make the task simpler you can for now assume that the parameters are given by

$$p^{(i)}_{k_1,\ldots,k_{d_i}} = 1 - \alpha^{1+\sum_{j=1}^{d_i} k_j}$$

for $\alpha = \frac{1}{2}$, but you are welcome to try other values of $\alpha \in (0, 1)$ as well.
Note, an earlier version of this project description (before 2023-03-13) instead stated

$$p^{(i)}_{k_1,\ldots,k_{d_i}} = 1 - \alpha^{\sum_{j=1}^{d_i} k_j}$$

which leads to degenerate distributions for nodes without parents. If you have implemented the simulation using this old formula for the CPDs of nodes with parents and, for example, $P(X_i = 1) = 1 - \alpha$ or $P(X_i = 1) = \frac{1}{2}$ for nodes X_i without parents, there is no need to change that now (it will not impact your grade). In this case, please just add a small note in your project report how you have implemented the CPDs of nodes with or without parents.

For Part I you are asked to:

- Load the graph from `consensus_adj_mat.csv` and check that it is the same as in Figure 1.
- Implement forward simulation from the Bayesian network.
- Illustrate the implementation by generating example data and present them visually.

- Fit logistic regression models of X_{10} and X_{11} against $X_1, ..., X_9$ using (lots of) simulated data.

3. Inference of hidden nodes

In the given data, X_{10} and X_{11} are not observed and prediction of these is an inference problem. In part II of the project you need to:

- Implement inference algorithms for computing the conditional distribution of X_{10} and X_{11} given the observed variables for any choice of parameters $\psi^{(1)}, ..., \psi^{(11)}$.
- Test the inference algorithm using simulated data (e.g. using the parameters given in part I).
- Apply the inference algorithm on the data in the data file and present the results.

You can also test the implementations by comparing the results to the results from the logistic regression model found using forward simulation in Part I.

Your should explore ways of making the inference algorithm as efficient as possible.

4. Learning of the parameters

The objective of this part is to learn the parameters $\psi^{(1)}, ..., \psi^{(11)}$ from data. The ultimate goal is to learn the parameters from observing only $X_1, ..., X_9$, but the problem is broken down so you first consider learning from a complete observation of all variables. Part III contains the following objectives:

- Suppose first that all variables, $X_1, ..., X_{11}$ are observed. Learn the parameters for each CPD by fitting a logistic regression of the child against its parents.
- Test the algorithm using simulated data.
- Proceed to implement learning with only $X_1, ..., X_9$ observed. One simple solution is the hard-assignment EM algorithm, which combines the inference from Part II with the logistic regression above. Thus, after initializing the parameters at some value (possibly stochastic), iterate the following steps:
  (1) Compute:
  $$\hat{X}_{10,j} = \arg\max_{x \in \{0,1\}} P(X_{10} = x | X_{1,j}, ..., X_{9,j})$$
  $$\hat{X}_{11,j} = \arg\max_{x \in \{0,1\}} P(X_{11} = x | X_{1,j}, ..., X_{9,j}).$$
  (2) Update the parameters by fitting logistic regression models as above using $\hat{\mathbf{X}}_j = (X_{1,j}, ..., X_{9,j}, \hat{X}_{10,j}, \hat{X}_{11,j})$ for $j = 1, ..., n$ as observations.

  You should check that your implementation converges.

It might be possible to speed up the EM-algorithm. Only $\hat{X}_{10}$ and $\hat{X}_{11}$ are updated each step and you might therefore consider which parameters you need to re-learn.

Alternatives to the hard-assignment EM algorithm are gradient ascent and the soft-assignment EM algorithm. You are welcome to explore such alternative algorithms, but this is not required.

5. Data

Data for this project comes in the file `proj_CPSN.zip`, which is a zip-file. It contains two files: `sachs_bin.csv` and `consensus_adj_mat.csv`. The latter is the adjacency matrix of the graph in Figure 1. The (i, j)'th element in the adjacency matrix is 1 if there is a directed edge from X_i to X_j and 0 otherwise. Use this file to load the graph. The other file contains the data from the experiments. There are 9 columns corresponding to the observed nodes of the network: `Akt`, `Erk`, `Jnk`, `Mek`, `P38`, `PIP2`, `PIP3`, `PKA`, and `PKC`. Each row in the data table is one observation $\mathbf{X}_j$.

6. Postscript

The model is based on the paper: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data, 2005, *Science*, by Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P.