Statistical data analysis 2

Project 1

Learning Bayesian networks from gene expression data

The aim of this project is to apply the Bayesian network modeling paradigm to a concrete gene expression dataset.

For this you will use the R package “Deal" [1]. With Deal, learning a Bayesian network from data comprises three steps:

1. to generate a prior structure of the network.
2. to generate a prior distribution of the parameters of the local probability distributions. Deal does this by computing the joint distribution of the parameters using the prior network structure and decomposes it over the network nodes.
3. to find the optimal network structure using search heuristics. It locally searches the space of possible structures and finds a structure with (locally) highest score. This exploration is performed by modifying the prior structure through adding, deleting, and inverting one edge per step. One cycle consists of generating all graphs that differ from the prior by exactly one of these operations. For each new graph the posterior probability (the network score) is compared with the posterior probability of the prior structure. The best graph is selected as the new prior and a new cycle begins. The default number of cycles is 50.

The dataset you will analyze was generated in a pioneering gene expression experiment on yeast [2]. It consists of (normalized) expression levels of 800 genes measured at different time points during the cell cycle. Due to the prohibitive computational complexity of the structure space exploration, we focus on ten genes that show the highest variation in the experiments. The filtered dataset is a matrix where each of the ten columns corresponds to a gene and each of the 77 rows corresponds to an experiment.

In order to assess the uncertainty associated with network inference, you will compare the outcomes of the Bayesian network learning procedure (i.e., of the three steps stated above) based on the observed gene expression data with different perturbations of the original data. Each reported edge based on the original data is scrutinized with respect to its relative frequency among the networks reported based on the perturbed datasets. In addition, it should be established how often an edge that is not reported by the unperturbed data appears among the perturbed networks.

To begin, open the R script bayesianNetworks.R. It contains commands for using Deal and some other useful functions. The specific steps that you need to perform in order to master this project are:

1. Run the preprocessing steps of the R-script to filter out the 10 most variant genes.
2. Create a prior structure for the network by hand (see the instructions in the R-script). Regarding the prior structure, suppose that YOL007C has an effect on YBR088C, and that YNL327W has an effect on YER124C, YHR143W and YNR067C. Plot your prior structure and include it in your report. (**1/4 point**)
3. Inspect the local probability distributions either by clicking on the nodes or via the localprob command. What is the output for gene YBR088C? (**1/4 point**)
4. Generate a prior distribution for the parameters of the joint distribution using the jointprior command.
5. Learn the initial Bayesian network. What is its score? (**1/4 point**)
6. Perform a local search for an optimal network. Plot the network obtained and include it in your report. This optimal network is called BN*. What is the score now? (**1/4 point**)
7. For each of the 10 genes calculate the variance σ_i^2 among the experiments and include them in your report. (**1/2 point**)

8. Perturb the experimental values of each gene *i* by adding a noise term distributed as N(0,$\frac{\sigma_i^2}{10}$) to each entry in the column corresponding to gene *i*. Repeat this procedure 30 times for each gene such that you generate 30 perturbed datasets.
9. For gene YHR143W, generate the following box plot: The horizontal axis should display each experiment, i.e., the values 1 to 77. The vertical axis should display one box per experiment. In other words, the box for experiment j encodes the empirical distribution based on the 30 samples of experiment j for the gene YHR143W. Include this in your report. (**1 ½ points**)
10. Repeat steps 2-6 (this time ignore the requests for your report) with each of the 30 perturbed datasets. The optimal networks obtained are called PBN_1, …, PBN_{30}. Plot the graph corresponding to the network PBN5 and include it in your report. (**2 points**)
11. For each edge contained in BN_*, calculate the relative frequency of appearance among the PBN_i. Plot these frequencies and include this in your report. Which edges of your optimal network seem to be spurious? (**3 points**)
12. For each edge not contained in BN_*, establish its relative frequency among the PBN_i. Plot these frequencies and include this in your report. Are there edges that might be missing in your optimal network? (**2 points**)

References

[1] Bottcher, S G and Dethlefsen, C. deal: A package for learning Bayesian networks. Journal of Statistical Software, 2003

[2] Spellman, P T et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell, 1998

Report

Return by **November 15, 11:59 pm,** to rish.kaustav@gmail.com (please include the word SDA2, your name and last name, index and the word Project1 in the subject of your email and name the PDF file using your first name last name and Project1). Also sent the .R file of your code.