Homework 3

Due 11:59pm April 7th

Instructions on homework submission. Please upload your solutions to Gradescope by the due date. We accept both hand-written and typed solutions, as long as you upload them as a pdf file.
Instructions on programming assignments. You can use any programming languages you like, but we strongly encourage you to use Matlab, Octave (an open source version of Matlab), R or Python rather than C/C++/JAVA. Please attach your code at the end of your homework solutions as text in 'Appendix'. Include the code and a brief instruction for running the code. Not including the code will result in 0 points.

1. Consider the following Bayesian networks for a gene regulatory network in plant.

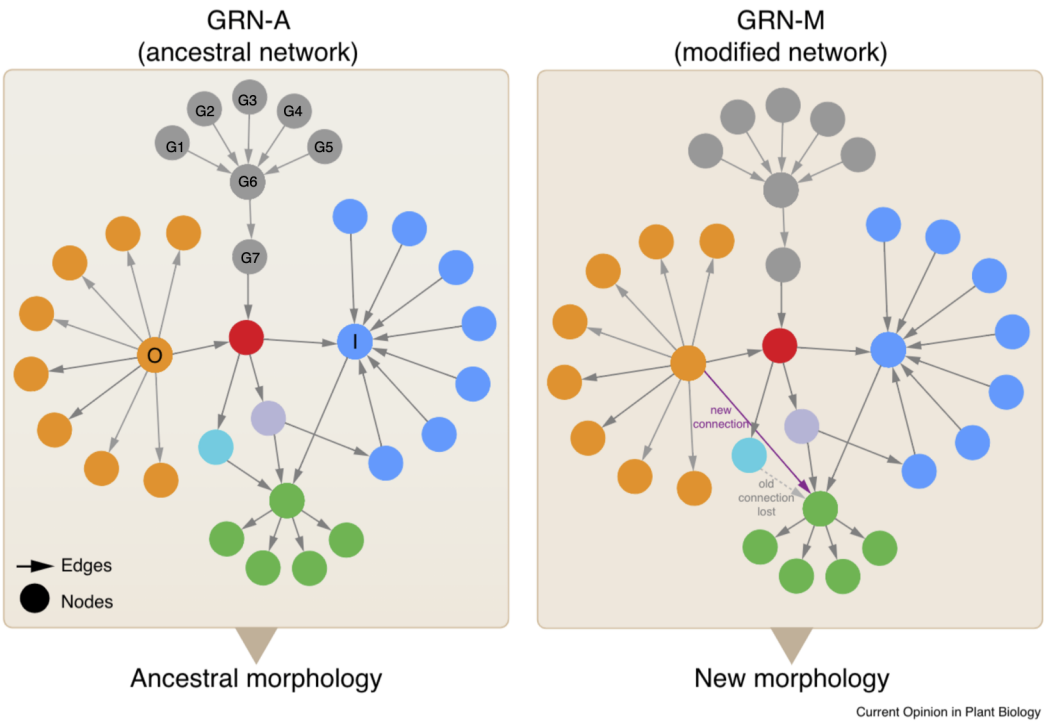


Figure 1: Plant gene regulatory networks. An ancestral gene regulatory network (left) and a gene regulatory network modified in evolution (right).

The conditional probability distribution for the expression level of gene i, denoted by G_i, is given as a linear regression model:

$$p(G_i|\text{Pa}(G_i)) = \mathcal{N}(\beta_{0i} + G_{\text{Pa}(G_i)}\beta_i, \sigma_i^2), \tag{1}$$

where $\text{Pa}(G_i)$ is the parents of G_i in the network, β_i is a vector of regresson coefficients corresponding to genes in $\text{Pa}(G_i)$, β_{0i} is an intercept, and σ_i^2 is the variance.

(a) (5 pts) Write down the local conditional probability distributions for each of the nodes $G_1, \ldots, G_7$ in the gray part of the ancestral network in Figure 1.

(b) (5 pts) Circle the nodes in the Markov blanket of node O. Circle the nodes in the Markov blanket of node I.

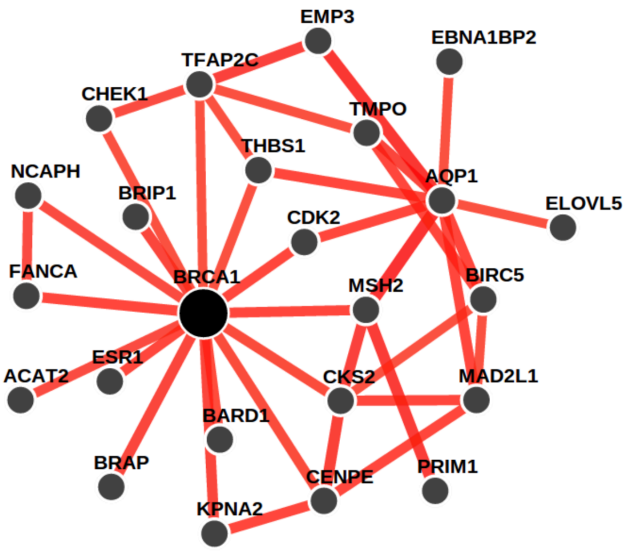
(c) (15 pts) Answer the following questions about the ancestral network in Figure 1. Provide a brief explanation for your answer.

- Are O and I d-separated by the red node?
- Are O and I d-separated by G_7?
- Are O and G_7 d-separated by the red node?
- Are O and G_5 d-separated by G_6?
- Are O and G_5 d-separated by the red node?

(d) (5 pts) The change of the network structure from left to right in Figure 1 can affect the local conditional probability distributions for individual nodes. Which node has its local conditional probability distribution affected by this structural change?

(e) (5 pts) Assume N samples are provided as training data. Describe how you would perform MLE to estimate the parameters of the ancestral network in Figure 1.

2. Consider a Gaussian graphical model $N(0, \Theta^{-1})$, where Θ is a 24 $\times$ 24 matrix, with the following undirected graph structure over 24 genes for BRCA gene regulation.



(a) (5 pts) What are the genes in the Markov blanket of BRCA1?

(b) (5 pts) Are BRCA1 and ELOVL5 conditionally independent given AQP1 and CDK2?

(c) (5 pts) Explain how you would obtain the marginal distribution of BRCA1.

(d) (5 pts) Assume you want to infer the conditional probability distribution of BRCA1 given all the other genes. Can you simplify this distribution?

(e) (5 pts) Assume you are performing an MLE with l_1 regularization. As you increase the regularization parameter, how would the graph structure be affected?

3. Implement the EM algorithm for Gaussian mixture models. Apply this to the expression data for mouse HIP brain tissue from Homework 2. Use only the first 10 mice (the first 10 rows in the data matrix) and cluster the genes.

(a) (10 pts) Assume $K = 3$ clusters. Use the initialization provided in Homework 2 to initialize the means (use the first 10 rows for 10 mice). Use a 10×10 diagonal matrix with 1.0 along the diagonals to initialize all covariance matrices. Use $[0.3, 0.3, 0.4]$ to initialize the mixing proportions. Plot the log-likelihood of data $\log p(\text{Data}) = \sum_{i=1}^{N} \log p(x_i)$, where N is the number of genes, over iterations. (Hint: The data log-likelihood should always go up over iterations. If this values goes down even slightly, this means your code has a bug!)

(b) (10 pts) Using the model you estimated in (a) above, compute the probability of the first gene to belong to each of the three clusters. (bonus question: do this for all genes and examine the cluster memberships.)

(c) (10 pts) For $K = 3$, try 10 different random initializations for all parameters. What is the data log-likelihood at convergence for each initialization? Which one do you think was the best initialization?

(d) (10 pts) Run the EM algorithm, assuming $K = 3, \ldots, 10$ clusters. Plot the log-likelihood of the data across different values for K. What do you think is the best choice for the number of clusters?