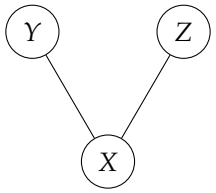
Stat 610 Homework 7

For this assignment:

- Due Friday, November 22, 11:59pm
- You may work in groups of up to three.

Background

In this homework, you will look into "structure estimation" using our convex optimization tools and Gaussian graphical models. The motivation behind Gaussian graphical models is as follows: if we have Gaussian random variables and the conditional dependencies between variables are encoded in a graph, the inverse of the covariance matrix will have zeros for pairs of variables that are not connected in the graph. For instance, if the relationships between variables X, Y, and Z are as depicted in the following graph:



then the inverse of the covariance matrix for (X, Y, Z) will be of the form $\begin{pmatrix} * & * & * \\ * & * & 0 \\ * & 0 & * \end{pmatrix}$, where * denotes some non-zero entry. The zeros in the $(3,2)$ and $(2,3)$ entries correspond to the lack of an edge between Y and Z in the graph. Long story short, the task of estimating conditional dependencies between variables is the same as the task of estimating zeros in the inverse covariance matrix.

Further, recall that for multivariate normal random vectors, the log likelihood is, up to a constant factor,

$$\ell(S;\Theta) = \log\det(\Theta) - \operatorname{tr}(S\Theta)$$

where S is the sample covariance matrix and Θ is the inverse of the true covariance matrix.

Assignment

1. Download the data from http://jfukuyama.github.io/teaching/stat610/assignments/hw7.csv. The rows of the matrix are the samples, and the columns are variable measurements. You should have 50 samples and 10 variables.
2. Estimate the inverse covariance matrix using the graphical lasso, that is, solve the following

optimization problem:

$$\text{minimize}_{\Theta} \quad -\log\det(\Theta) + \text{tr}(S\Theta) + \lambda \sum_{i=1}^{p}\sum_{j=1}^{p} |\Theta_{ij}| \tag{1}$$

where S is the sample covariance.

For appropriate values of λ, this will give you a "sparse" estimate of the inverse covariance matrix Θ, as we want for structure estimation.

You should use the functions `log_det` and `matrix_trace` (in the CVXR package).

3. Choose some subset of the elements of the inverse covariance and plot the estimates for a variety of values of λ.

4. We would like to pick a good value of λ: one way to do this is by cross validation. The idea is to choose the value of λ that gives the highest value of the likelihood on a held-out portion of the data.

   We will do 10-fold cross-validation.

   Choose a range of values of λ such that at one end you have a solution with only a few zeros and at the other end you have a solution with nearly all zeros. For each value of λ, perform the following:

   - Divide the samples into ten groups, and let I_i denote the indices of the samples corresponding to the ith group.
   - For each i, fit the model in 1 on the data excluding the indices I_i. Let the estimate of the inverse covariance based on this subset be $\hat{\Theta}^{(i)}$.
   - For each i, compute the negative log likelihood of the data on the hold-out sample, that is, compute

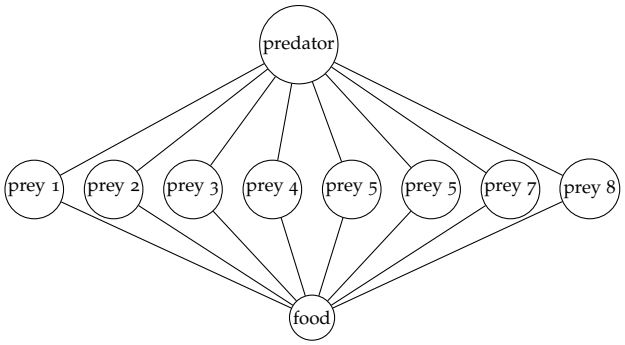
     $$-\log\det(\hat{\Theta}^{(i)}) + \text{tr}(S^{(i)}\hat{\Theta}^{(i)})$$

     where $S^{(i)}$ is the covariance computed just on the samples with indices I_i.
   - Compute the overall negative log likelihood for the held-out data: $-\ell_{cv} = \sum_{i=1}^{10} -\log\det(\hat{\Theta}^{(i)}) + \text{tr}(S^{(i)}\hat{\Theta}^{(i)})$

   Now you should have values for the negative log likelihood for the held-out data $-\ell_{cv}$ for each value of λ. Show the held-out negative log likelihoods against λ, and find the value of λ with the smallest held out negative log likelihood value.

5. It turns out that in this dataset, the first variable is a measure of the abundance of a keystone predator, the second through 9th variables are measures of the abundances of prey species, and the last variable is is a measure of the abundance of a food source.

   You suspect that the relationships between the variables might be encoded in the following graphical model:



that is, the predators and prey are have negative partial correlations, the prey and resources have positive partial correlations, and every other pair have zero partial correlation. The interpretation would be that when there is more of the food source, the prey species increase, when there is more of the predator, the prey species decrease, but the prey species do not directly effect each other.

Perform maximum likelihood estimation under the constraint that the partial correlations between the prey species are zero, that is

$$\begin{aligned} \text{minimize}_{\Theta} \quad & -\log\det(\Theta) + \text{tr}(S\Theta) \\ \text{subject to} \quad & \Theta_{ij} = 0 \quad (i,j) \text{ such that } i \neq j, i \in \{2,\ldots,9\}, j \in \{2,\ldots,9\} \end{aligned}$$

Report your estimate.

6. Obtain bootstrap confidence intervals for the non-zero elements of Θ: for some reasonably large number B, perform the following:
   - Sample the rows of X with replacement to make a new matrix, $X^{(b)}$, of the same size as X.
   - Make a new estimate of Θ, $\hat{\Theta}^{(b)}$ based on $X^{(b)}$ using the same strategy as in the previous part.

   Compute and report the .025 and .975 quantiles of $\hat{\Theta}_{ij}^{(b)}$ for each (i,j) pair. These are your bootstrap confidence intervals for Θ_{ij}.

Submission parameters

Submit two files:

- A pdf writeup containing the answers to the questions.
- A file containing the code you used.