6.3730 Pset 2

Marlond Criollo Collaborators: Office Hours

Problem 1

As discussed in class, it is becoming clear that the spatial organization of the DNA is crucial for making the different cell-type specific gene expression patterns possible. In this problem, we analyze the data collected by Rao et al. (2014), in particular the Hi-C data for IMR90 (fibrobloast cells). The data for this problem is separated into tab-separated value files, chrX chrY.txt, representing a sparse matrix of interactions between chromosome X and chromosome Y. The first column corresponds to a location on chromosome X (in base pairs). The second column corresponds to a location on chromosome Y (in base pairs). The third column gives an observed interaction frequency between these two regions, averaged over both copies of chromosomes. The resolution of this data is 250kb, so the locations are all multiples of 250k. (You should divide all of the locations by 250k to build your matrices.)

Part A

For the purposes of this question, we will model log(1 + interaction frequency) as a Gaussian random variable for each interaction site. Compute the mean and standard deviation of log(1+ interaction frequency) across all inter-chromosome sites (i.e. don't include intra-chromosome interaction matrices like chr7_chr7.txt). Recall each file represents a sparse matrix, so there are many more entries in the matrix (with value 0) that are not listed. Hint: you do not need to simultaneously store all of the matrices in memory.

Sample Mean

To get sample mean I first summed up the ln(1+*interaction frequency*) for each matrix individually while estimating the total number of entries for each matrix. To avoid intra-chromosome interactions I excluded them from the `for` loop that was being done over all matrices. To estimate the total number of matrix entries I took the largest entry for the x and y columns and divided it by the resolution to get the number of possible entries in their respective sparse matrices. Then I multiplied the possible entries in each chromosome by the other to get the total number of entries in the sparse matrix.

Using the following formula for sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=i}^{n} X_i$$

I ended up getting a mean of about 0.7087101

Standard Deviation

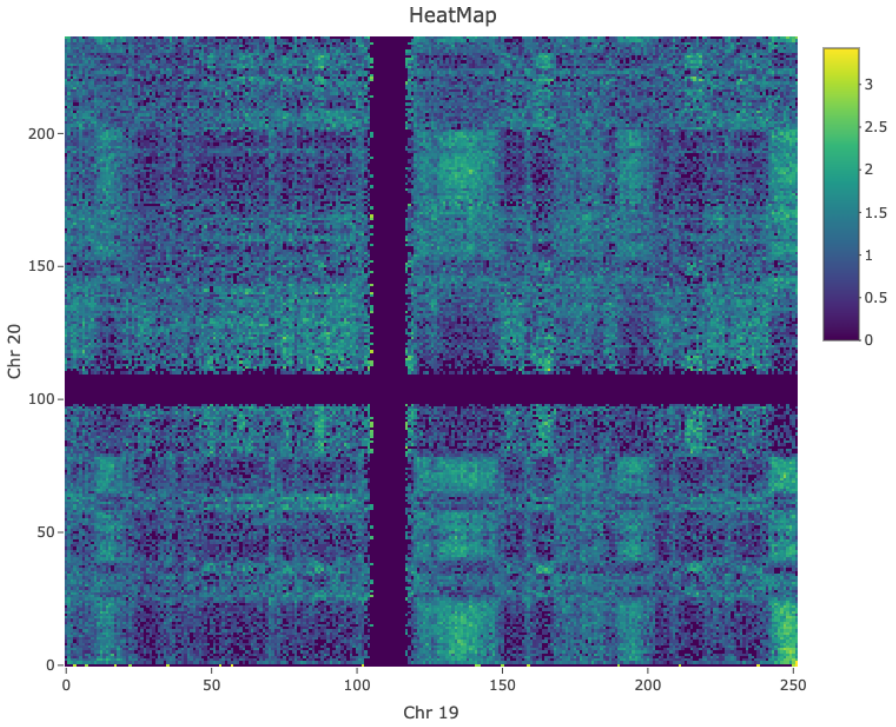
To get the the S.D. I first began by getting the variance of the data set. Given N is large I used the the simplified formula for variance. Which is as follows

$$S_n = \bar{X^2} - \bar{X}^2$$

After getting the sample variance I took the square root to get the standard deviation which is about .633799

Part B

Next, we will look at interactions between chromosomes 19 and 20. Our goal is to identify intermingling portions of these two chromosomes, as it has been shown experimentally that these regions are particularly active with respect to gene expression. These regions show up as contiguous sub-blocks of our interaction matrix with high average interaction values. To begin, plot a heat map of the interaction matrix (using the log(1 + x) transformation). What do you see? Do you see any interacting regions?



This was created by using the rows of the data set as entries into a empty matrix with size chromosome 19 length by chromosome 20 length .The darkest parts here represent no interaction with a value of 0. The axis were scaled as to not create a bigger spread out matrix. The heatmap can be seen in figure 1.

It seems as if the middle of each chromosome pair does not interact with each other, which might suggest that the middle parts of each chromosome do not interact with each other. It could also be a part of the data collection process or methodology.

Part C

Given kl variables and out null hypothesis we assume the variables are independent and identically distributed random variables with mean μ and σ from above. They also have a gaussian distribution. If we define our test for high interaction frequencies such that for a given α. $P(\delta = 1) \leq \alpha$. When $\delta = 0$ we fail to reject and when equal to 1 we reject. Using the central limit theorem we can note that:

Problem 2

Part A

The β made with ℓ_1 penalty will have more variables forces to 0 and capable of handling outliers. On the other hand the ℓ_2 penalty will never force any variable to 0. The focus of ℓ_2 penalty is to penalize larger coefficients more so than the smaller coefficients. The ℓ_2 penalty is also maintains uniqueness and differentiability while the ℓ_1 does not. Despite their differences both allow for some level of variance reduction and avoiding overfitting.

Part B

When looking at the formula we can see that it is a blend of both penalties. Which is to say it contributes to both enforcing sparisty as the ℓ_1 does, while also penalizes larger coefficients as the ℓ_2 does. It will also maintain uniqueness because of the ℓ_2 penalty.

Part C

The elastic net penalty might be useful in applications where we have large datasets with just as many variables and we are interested in just the most important or most important few. A case where this could be applicable would have to be genetic dataset as they are vary large with many different variables.

Problem 4

Part A

Part B