Skype Meeting Notes

Nooreen S Dabbish, Kaijun Wang, Dr. Deep

<2016-02-14 Sun>

Assignments/instructions from Dr. Deep are in **bold**.

1 math.sort

Kaijun: number is how many parts you want

- checks only the first factor by default
- `http://www.inside-r.org.packages/cran/psych/docs/mat.sort`
  - K: we want one or two clusters, so it regresses the columns of the

covariance matrix

  - How can I use the top three or four? (It sorts by the first factor by default)
  - fa( )
  - **Select f=1 and f=2 and display the correlation matrices**
  - Does the structure change?
  - f=2, first two or only the second?
  - K: it groups the columns
  - **dowload the function and unpack that function**

2 Lasso

- huge number of predictors (15,576)
- we want to identiify which correlations are changing with male and female

- lasso simultaneously picks variables and determines importance
- **Map the non-zero back to the brain**
- Matrix vectorization and adding response
- save all the data in a three D matrix X_{ijk} X[i,j,k] X[i, , ,] –> correlation matrix of ith invdividual c(X[i,,]) upper.tri()
- **Streamline data as .Rda** Y-vector and 3-d X matrix
- save(Y,X,file="<>.Rdata")
- **Any difference between lasso logistic and lasso linear** Are the selected coefficients different?

3 Paper to present on 2/22

- highly cited, very recent paper
- Main focus: one part is lasso logistic and linear regression Problem of how to select the threshold? The method we are doing takes all the correlation as a variable in the model. This method is not scalable for large dimentsional brain data with hundreds or thousands

of sensors. It is not an intelligent way for tackling the spatial correlation between nodes.

4 Kaijun: on graph kernel

- graph kernel is the relationship beween two graphs
- for example random walks gives a short path kernel
- the two-sample test: actuallyy tests the
- **More insight into definition of the kernel** I just want to know the formula.
- What people are doing right now: lasso approach, kernel approach. Will the

kernel approach help us to classify? How does it help us to answer the real question, if I have a Y vector and want to predict that?

- K: we look at the means and if they are the same, the graphs are the same.

- Kernel methods answers an important question of whehter the groups

have a different graph. But can you use this method to predict? If I reject and say they are different, the next step would be to build a classifier and tell us whether it is y=1 or y=2.

- *First, two-sample hypothesis testing. Second, if I reject that hypothesis (there is a difference)

I have to build a classifier.*

- The other method was lasso logistic, Nooreen can predict whether it is

male or female. We have to build that capacity for kernel. Nooreen answered the second problem without answering the first question.

- Kaijun is developing a test statistic. It is only meaningful to build a

classifier if the answer is reject.

- write down these questions and mention them in your 2/22 presentation.

We want to develop a single method that can answer both questions simultaneously.

- **Kaijun assignment: give me some numbers for the application of this method**

5 Summary

First point will be the two interesting questions 1, 2, The method to attack the first problem (5 min, Kaijun) will be the graph kernel part. Then (5 min, Nooreen) on the prediction lasso part.

Both of us present the paper (10 min, both) Last (10 min, focus on multiple testing thresholding) **new** understand that lasso logistic is not scalable. It is not the smartest way to attach that problem. We have to bring graph into the picture. What is the threshold that we should pick?

What is the right way to go from correlation to binary adjacency matrix? The paper will help you take a step in that direction.

You have all your correlations vectorized, take the Fisher's Z transformation 1/2 of log r ... After taking this transformation the numbers are very close to normal with variance $\frac{1}{(n-3)}$, so I can compute the P-values for all the correlation.

Now the multiple testing problem is waiting for us. We will find a cutoff and get a binary matrix 1 and 0.

- will send code for that
- this is an attempt at calculated the adjacency matrix
- the problem will be isolated node.