Multivariate Final Project Proposal: Gene Clustering

Aguilar Oscar
Deng Guo
Zerwick Andrew

March 5, 2012

Abstract

DNA stands for deoxyribonucleic acid, and it is the basic material that makes up human chromosomes. DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present in for that gene. These microarrays are large dataset that are represented as matrices, where the rows represent the gene type and the columns represent different samples. One of the reasons why people analyze these datasets is because they want to find which genes are most similar to each other in terms of their expression across genes. The result of this project might help cancer researchers to develop new treatments for different cancers due to the different clusters we might find.

1 Type of Project

This project is a blend of multivariate analysis techniques and data analysis. The techniques that will be used to carry out the data analysis are Clustering Analysis and Principal Component Analysis.

1.1 Cluster Analysis

Cluster Analysis is one of the most common statistical tool used to analyze multivariate data. The goal of Cluster Analysis is grouping a collection of observations into subsets or "cluster" such that all the observations within a cluster are more closely related to each other than other from other cluster. Clustering involves the following steps:

- First, a suitable distance between objects must be defined, based on relevant features.
- Then, a clustering algorithm must be selected and applied.
- Finally, analyze results of clustering.

There are two types of clustering:

- **Hierarchical methods**, these methods provide a hierarchy of clusters, from the smallest, where all objects are in one cluster, through to the largest set, where each observation is in its own cluster.
- **Non hierarchical methods**, these usually require the specification of the number of clusters. Then, a technique for apportioning objects to clusters must be determined.

2 Data

A gene expression data of G genes and n mRNA samples has the following form:

$$X_{G\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{G1} & x_{G2} & \cdots & x_{Gn} \end{bmatrix}$$

where:

$$x_{gi} = \text{expression measure for gene g in mRNA sample } i$$

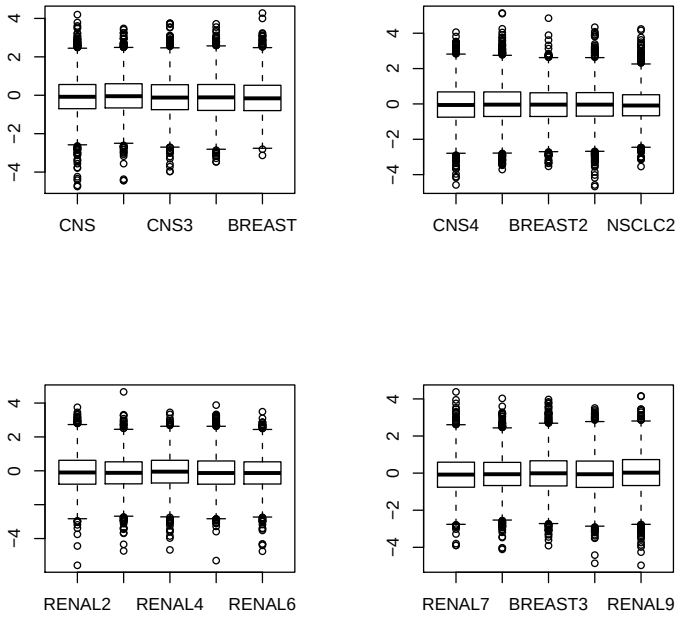
2.1 Exploratory Analysis

The dataset that will be used is a Gene Expression Data that comes from the National Cancer Institute. This dataset is represent as a matrix that 5244 rows and 62 columns. The rows represent the different genes, and the column represent the mRNA samples. As mentioned before, the entries of the dataset represent the measure of a gene in a mRNA samples. Here is a small sample of the data:

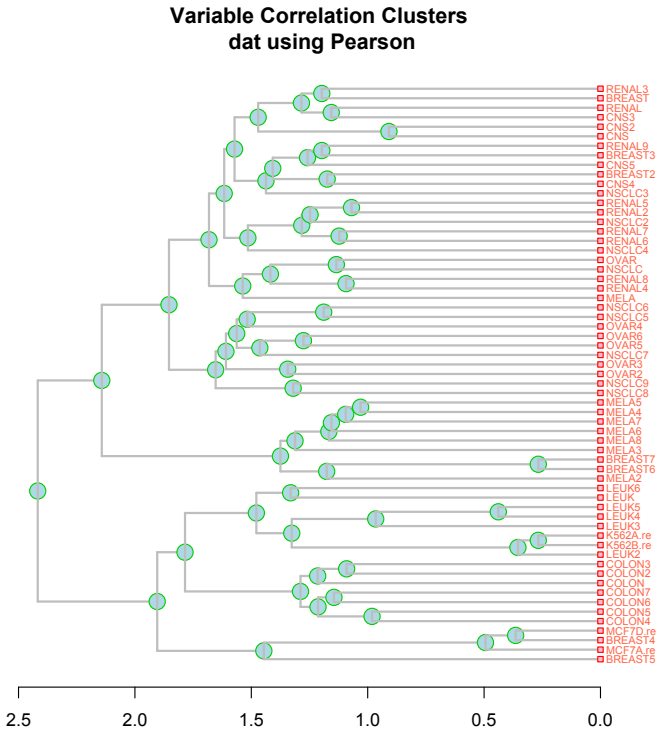
```
> head(dat)
      Aname    CNS  CNS2  CNS3 RENAL BREAST  ...
1 GENE6683X  0.71  1.51  2.41  0.52   0.70  ...
2 GENE2811X  1.41  0.36 -0.46  1.41   0.58  ...
3 GENE3496X -0.76  0.88 -1.56  1.11   0.13  ...
4 GENE4158X -0.35  1.62 -0.08 -0.08  -0.24  ...
5 GENE1771X -1.11  2.10  2.12  0.58   0.04  ...
6 GENE4499X -1.24 -0.17  1.15  0.03   0.29  ...

...  BREAST6 BREAST7 MELA3 MELA4 MELA5 MELA6 MELA7 MELA8
...   -0.80    0.10  1.50 -0.22 -0.74  0.40 -0.17  0.84
...   -0.09    0.93  0.36  0.04 -0.26 -0.42 -0.23  0.05
...   -0.26   -0.24 -0.23 -0.58  2.41 -0.05 -0.01 -1.70
...    0.49    0.51  0.32  0.91  0.20 -1.38 -0.10 -1.14
...   -1.28   -0.37  0.91  0.10 -2.27 -2.19 -0.93  1.47
...    0.04    0.89 -0.22 -1.03 -0.84 -0.95  0.68  0.58
```

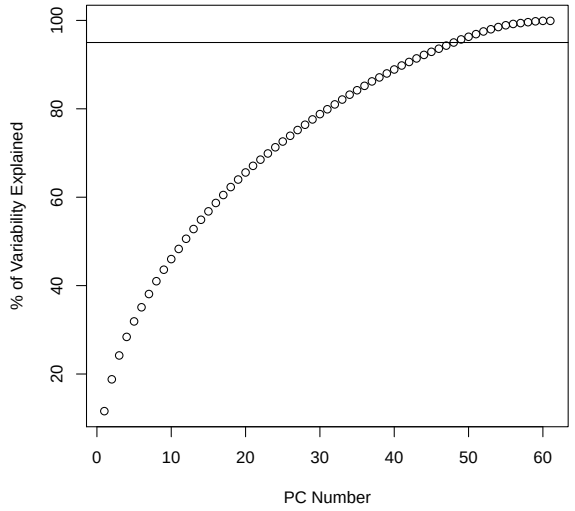
The names of the columns represent the different kind of cancer classes. The following plots are relevant to the data:



The below plot represent box plots for some of the samples. As we can see, there are outliers in each sample. After that, we want to see if there is any correlation in the data. The below plot groups cancer classes that are highly correlated.

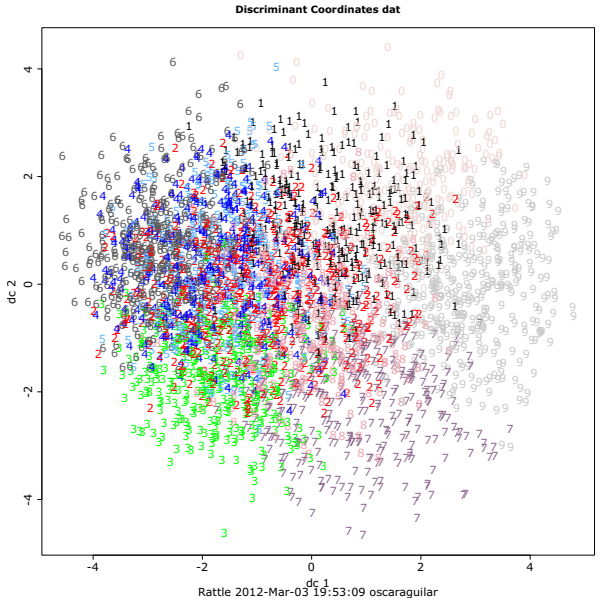


Since we are dealing with a big dataset, the principal component approach may be helpful in this case. The below plot is a plot that shows us the percentage of variability that is covered by the principal components.



As we can see, 48 principal components cover 95% of the variabilty, so we could reduce the data. Finally, we used the k-means algorithm to start clustering the data. The number of cluster was 10. The below plot is a discriminant plot that gives us an idea of how the

data is grouped in 10 clusters. The numbers represent the cluster number.



3 Data Source

The data was obtained from the website **http://astro.ocis.temple.edu/alan/MMST**, which is the website of the book called **Modern Multivariate Statistical Techniques** by **Alan Julian Izenman**.

4 Importance of the this project

The optimal treatment of patients with cancer depends on establishing accurate diagnoses. If we can find there are certain cluster of genes correspond with certain cancer, then this result may help doctors to identify if a patient has cancer and which kind of cancer as early as possible. So they can start the correct treatment early, thus increase patient's chance of cure. The result of this project might help cancer researchers to develop new treatments for different cancers due to the different clusters we find. Also, clustering can be helpful for identifying gene expression patterns in time or space.

5 Research Plan

Date	Objective	Person(s) in charge
Mar 5-9	Literature Review	Oscar, Guo, Andrew
Mar 12-16	Selection of potential clustering algorithm	Guo, Andrew
Mar 19-23	Start with the analysis	Oscar, Andrew
Mar 26-30	Start coding the algorithm that was chose	Oscar, Guo
Apr 2-6	Fix any potential code problem and start writing the report	Oscar, Andrew
Apr 9-13	Meet with Dr. Hearing to get ideas for completion of project	Oscar, Guo, Andrew
Apr 16-20	Keep working with the analysis and keep writing the final report	Oscar, Guo
Apr 23-27	Complete work on project and edition of final report	Andrew, Guo
May 2-4	Preparation of poster and practice for verbal presentation	Oscar, Guo, Andrew

References

[1] Alan Julian Izenman. 2008. **Modern Multivariate Statistical Techniques**, chapter 12.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. **The Elements of Statistical Learning**, chapter 14.

[3] Richard A. Johnson, and Dean W. Wichern. 2000. **Applied Multivariate Statistical Anlaysis**. Sixth ed. chaper 8, 12.

[4] Trevor Hastie, Robert Tibshirani, Michael B. Eisen, Ash Alizadeh, Ronald Levy, Louis Staudt, Wing C. Chan, David Botstein, and Patrick Brown. "Gene shaving as a method for identifying distinct set of genes with similar expressions patterns". *Genome Biology*. August 2000.

[5] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander, and Todd R. Golub. "Multiclass cancer diagnosis using tumor gene expression signatures". *PNAS*, vol. 98, no. 26, pp. 15149–15154. December 18, 2001.