Gene Shaving: A clustering technique applied to DNA microarray data

Author: 91673

Major: Statistics

Courses: Regression Analysis, Theoretical Statistics, Multivariate Statistics, Computational Statistics

Abstract: Clustering is one of the most popular techniques used in the analysis of DNA microarray experiments. It does contain some drawbacks in this setting however. Genes are clustered in a global manner, meaning that entire cell lines are clustered instead of subsets of the genes. This is the motivation behind the technique of *gene shaving,* which seeks to identify subsets of genes that are highly correlated and have high variance across tissue sample. This technique also allows us to cluster genes into more than one group, and thus contains more flexibility than traditional clustering methods used in microarray analysis. In this paper we use gene clustering to analyze a publicly available microarray dataset and analyze the results.

1 Introduction

A DNA microarray experiment is a technology that allows researchers to measure the relative activity of a large number of genes from different tissue samples. This is useful, for instance, to determine which genes are relatively more active in diseased tissue compared to normal tissue, allowing biologists to isolate the possible genetic causes of a disease. The process is illustrated in Figure 1.

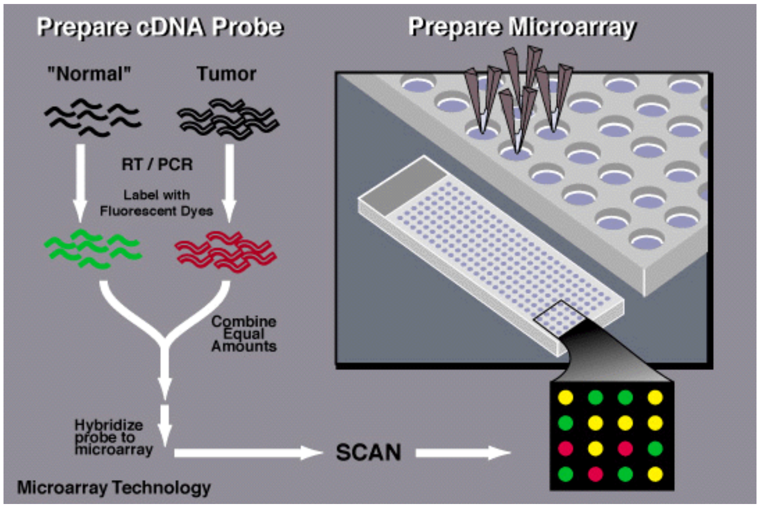


Figure 1: *Microarray Process. A target sample (tumor) and a reference sample (normal) are taken and labeled with green and red dye and then hybridized with DNA on the microarray slide. The log (red/green) intensities of RNA hybridizing at each site is measured.* **(source: http://www.genome.gov/10000533)**

Thousands of DNA samples, each corresponding to a specific gene, are spotted in an ordered array on a glass slide. A target sample and a reference sample of mRNA are labeled with red and green dyes and then hybridized with the DNA on the slide. Using fluoroscopy, the log(red/green) intensities of RNA hybridizing with DNA are measured at each site. Positive values indicate higher expression in the target versus the reference, vice versa for negative values. We thus obtain a data matrix X, with rows being the different genes and columns being the different cell lines (sample tissues). In entry x_{ij} is the expressed relative intensity measure of gene *i* in cell line *j*. We are now in a position to bring the power of quantitative tools to the analysis of the microarray. What is of interest is finding if and what certain genes are relatively more active in certain cell lines, for instance in a cancer tumor tissue compared to a healthy tissue sample, or in several different types of cancerous tissue. Hierarchical clustering, K-means, and other methods have been applied in the past to successfully group genes and cell lines, both one-way (by either cell line or gene) and two-way (by both cell lines and genes) (Tibshirani, et al. 1999). It have been noted in the literature, however, that these methods, while informative, are not able to identify more complex patterns that may exist. For example, one set of genes may divide the cell lines in one way, while another set might divide them in a very different way. With this motivation, Hastie et al. (2000) introduced a method they call *gene shaving* for "extracting coherent and typically small clusters of genes that vary as much as possible across the cell lines." This project will employ the gene shaving method to a publicly available microarray dataset in order to try to cluster genes in an informative manner. One problem from the outset was implementing the geneshaving algorithm. Fortunately, a software implementation of the method, in the form of a program called *Geneclust*, was obtained from researchers at the University of Texas MD Anderson Cancer

Center.[1] The software consists of a frontend GUI programmed in Java and a backend program consisting of C code for computation and interfacing with R for graphics and statistical analysis.

2 Data Source

The data analyzed in this project came from the Stanford NCI60 Cancer Micorarray Project.[2] NCI60 is a dataset of gene expression profiles of 61 National Cancer Institute (NCI) cell lines. These 60 human tumour cell lines are derived from patients with leukemia, melanoma, along with, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers. For each cell line there are 5244 genes in the dataset. Thus our data matrix is 5244 x 61. The data is displayed in the form of a heatmap in Figure 2 below.

3 Gene Shaving Procedure

The gene shaving algorithm can be described in the following steps:

1. Starting with the full data, find the first principal component of the genes. We are seeking a function of the genes in the direction of maximal variation across tissue samples.
2. Compute the absolute value of the correlation of each gene with the first principal component and remove a fraction α of genes having smallest absolute correlation.
3. Iterate on the reduced data matrix until there are only 2 genes left.
4. Select the optimal cluster size. This is based on a *Gap statistic* (Tibshirani et al. (2000)) which is based on the variances between and within the gene blocks computed from the raw data matrix and it's permutations. The Gap statistic first breaks down the total variance in an ANOVA-type decomposition: $V_T = V_B + V_W$, where V_B is the between cluster variance and V_W is the within cluster variance. Then we compute the statistic $R^2(S_k) = \frac{V_B}{V_W} * 100\%$, which is the percentage of total variance explained by cluster S_k. Now randomly permute the elements of each row of the data matrix, say b times, apply the shaving algorithm to the permuted matrix and compute the associated R^2 for it. Then the Gap statistic is Gap(k) = $R^2(S_k)$ - $avg(R^2(S_k^*))$, the last term being the average of all R^2 values of the permutations. The cluster size k is the value that maximizes the Gap. (The number of permutations is specified in advance of running the algorithm.)

[1] although obtained through correspondence with the researchers, the software is publicly available at their website

[2] http://genome-www.stanford.edu/NCI60/

5. Remove the effect of the genes in the optimal cluster. This is done by computing a vector of column averages of the genes in this cluster and regressing each row of X on this vector, then replacing the row with the regression residuals. Hastie et al. calls this modified data matrix X_{ortho}
6. Repeat above steps with X_{ortho} in place of X until desired number of clusters is reached.

   This number is specified in advance.

4 Application

As mentioned earlier, Geneclust was the software used to perform gene shaving on the NCI60 dataset. Besides the data matrix, three other inputs were required: the number of clusters, the number of permutations used by the gap statistics when computing the optimal cluster sizes, and α, the shaving parameter. The latter two were selected to be 5 and .10, respectively, because these were found to be commonly in use and because other values did not seem to effect the result very much, but mostly just the run time. The algorithm was run for k = 1,2,3,4, and 5. Four clusters were selected because the fourth cluster appeared to significantly group an additional set of genes while the 5th cluster did not. Figures 2-6 below show, respectively, the variance graphs, principal component graphs, gap curve graphs, and heat maps for the four clusters. For the variance graphs, the variance ratio is near 1 for awhile before dropping off for each cluster, indicating a coherency of the cluster. The Gap statistic graphs clearly display the values k that we choose to be the cluster size. For the first cluster, the graph is flat near the maximum, making the value k not clear. The other cluster sizes are more easily distinguished by the gap graphs. The heatmaps for each cluster display what appears to be good coherency within the clusters, especially in clusters 3 and 4. This is expected due to the small size of the clusters. The cluster sizes were: 107, 56, 9, and 4, for clusters 1, 2, 3, and 4 respectively. The genes in these clusters show both high variation across samples and correlation across genes. Once these genes are identified, further biological and chemical study can be undertaken in an attempt to understand their significance in the diseases in the samples.

GeneShave Variance Graphs

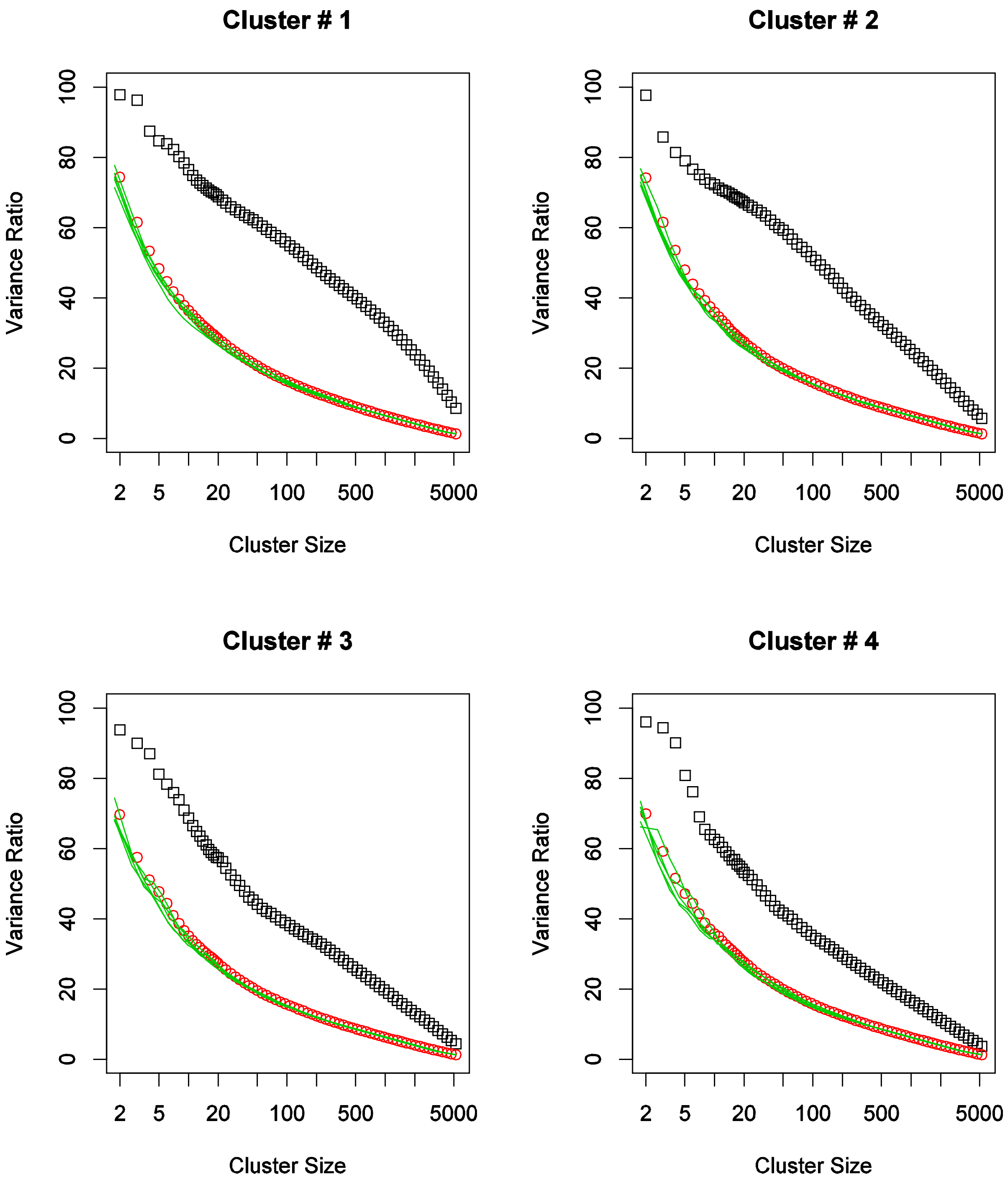


Figure 2: Black line is for Real Data, Colored line for Randomized

GeneShave Principal Component Graphs

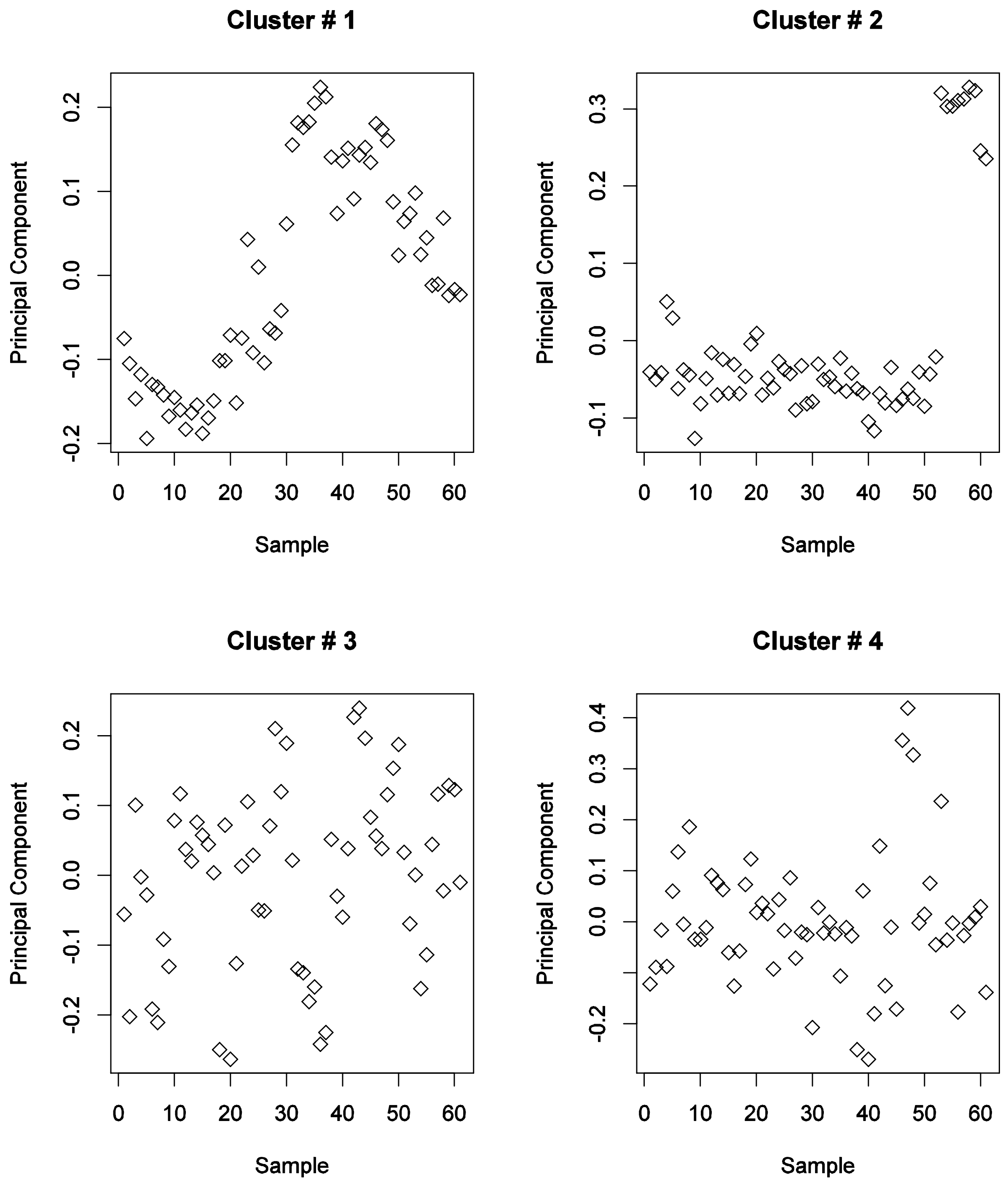


Figure 3

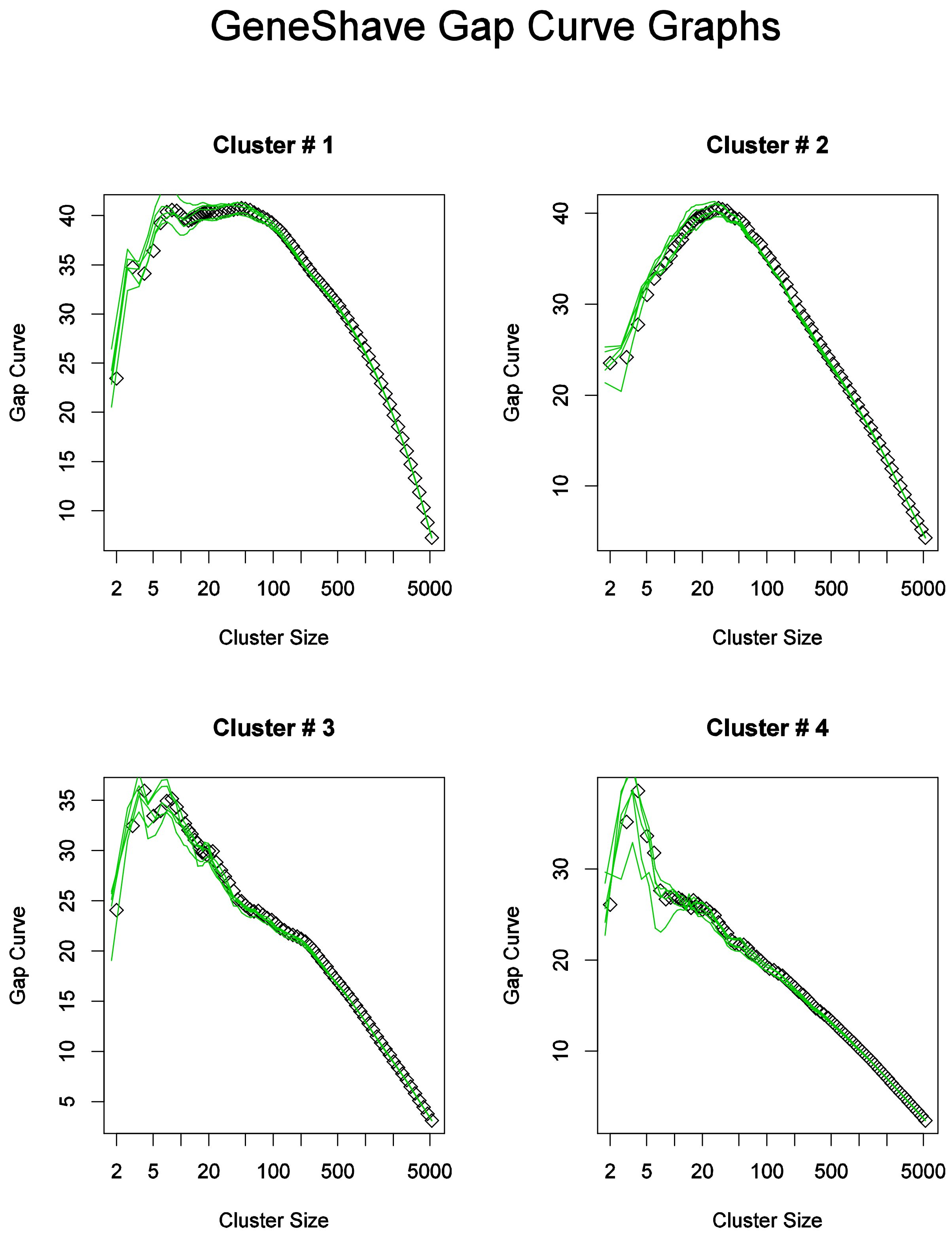
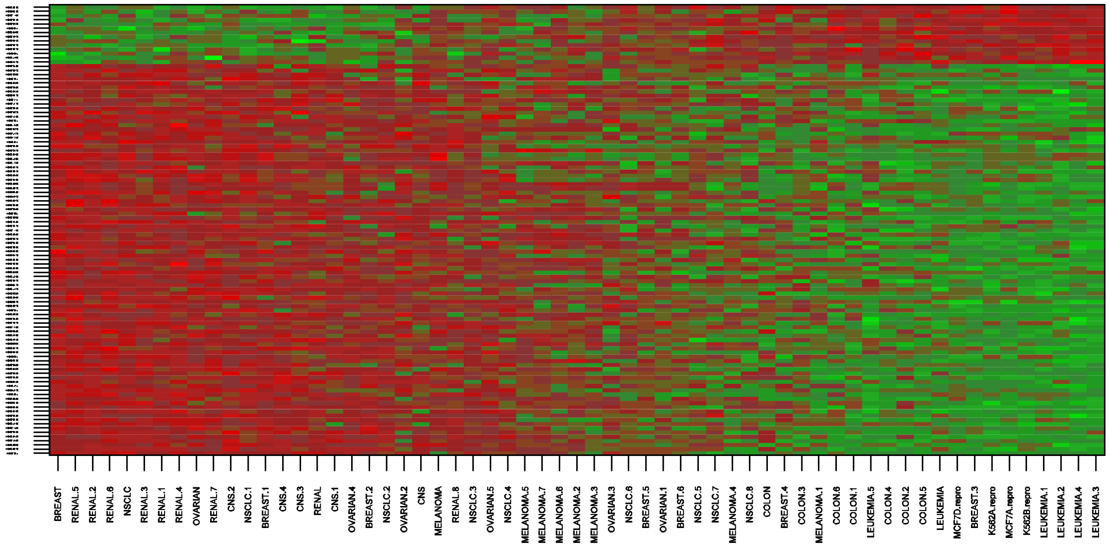

GeneShave Gap Curve Graphs
Cluster # 1
Cluster # 2
Cluster # 3
Cluster # 4
Gap Curve
Cluster Size
2
5
20
100
500
5000
10
15
20
25
30
35
40

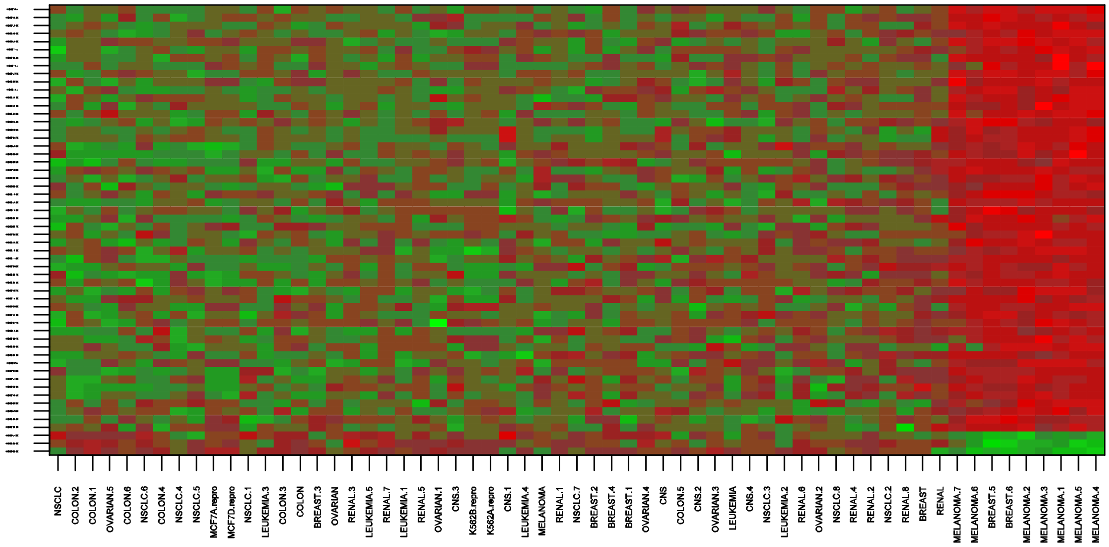
Figure 4

GeneShave Cluster Plots



Cluster # 1

eigenvalue= 3534.1692 %variance= 54.8819 VT= 0.9836 VB= 0.5398 VB/VW= 1.2164



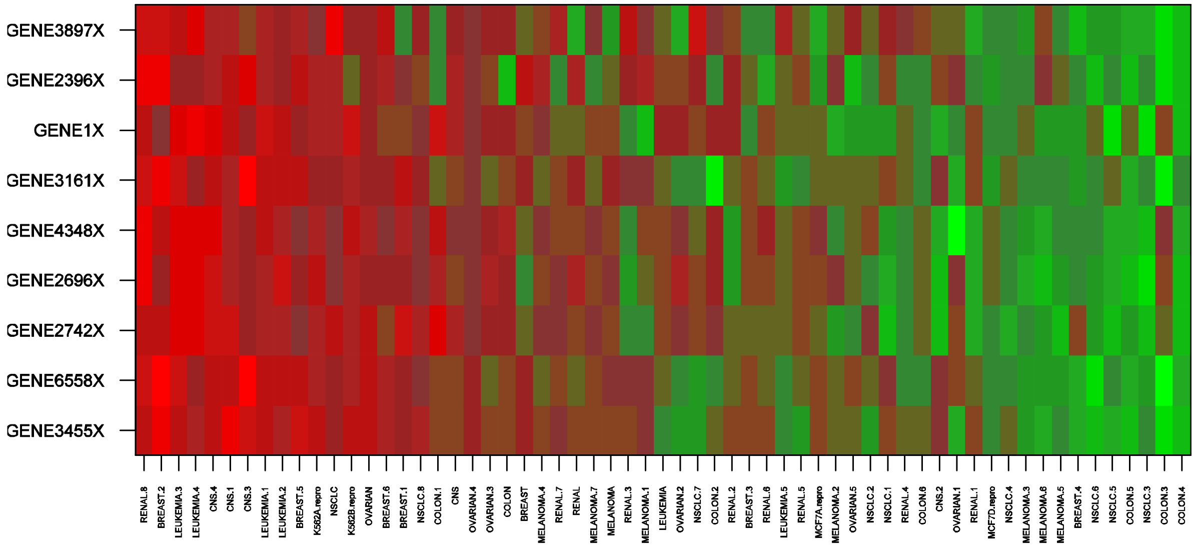
Cluster # 2

eigenvalue= 1882.3873 %variance= 58.0403 VT= 0.9452 VB= 0.5486 VB/VW= 1.3832

Figure 5. Cluster Plots for Clusters 1 and 2

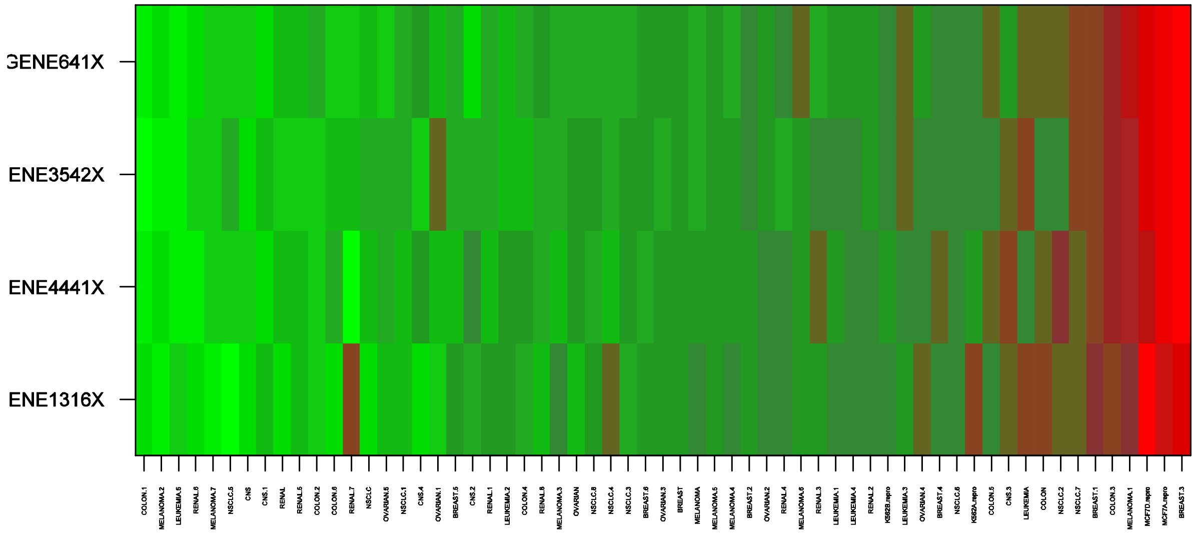
GeneShave Cluster Plots

Cluster # 3



eigenvalue= 368.2091 %variance= 70.9868 VT= 0.9414 VB= 0.6682 VB/VW= 2.4467

Cluster # 4



eigenvalue= 191.0393 %variance= 90.1119 VT= 0.8686 VB= 0.7827 VB/VW= 9.1132

Figure 6. Cluster Plots for Clusters 3 and 4

5 Conclusion

DNA microarray studies are undertaken with the goal of understanding the genetic nature of a biological problem. In the most familiar case this is the study of the genetic roots of a disease such as cancer. Once the experiments are done and data is collected, it is the role of statisticians, working with biologists, to understand the structure of the data and try to extract information. In the past, statisticians have mostly used various clustering techniques to group the data. These have been successful, but they are global methods that seek a single reordering of the cell lines for all genes, and thus may miss any finer structure that exists in the data. This project has used a recently developed technique called gene shaving to analyze data from a microarray experiment. Results were four clusters of genes that showed both high variation across samples and correlation across genes. This technique can also be used in a supervised fashion, with cell line labels or survival times. These were not considered here. It is also possible that the techniques could be used on data similar to microarray data; it is not necessarily exclusive to genetic data. This direction is left to future research.

The main lesson learnt in this project was that the nature of the data sometimes demands that existing statistical techniques be “tweeked” to produce more meaninful information. While clustering algorithms had been successful in microarray data analysis, gene shaving added a further dimension to these by introducing principal components to measure the largest direction of variation in the data. The gap statistic was also introduced as a way to measure the coherency of the cluster and thus choose an optimal cluster size. Finally, the implementation aspect was made clear, as I worked with programs written in R exclusively at first, and then a version written in Java with a C and R backend. The implementation of new algorithms is a significant process, but is made much easier with freely available, and powerful, software such as R.

References

1. Tibshirani, R., Hastie, T. Eisen, M., Ross, D. , Botstein, D. and Brown, P. "Clustering methods for the analysis of DNA microarray data". (compressed postscript 4.8mb) Tech. report Oct. 1999
2. Tibshirani, R. Walther, G. and Hastie, T. "Estimating the number of clusters in a dataset via the Gap statistic". *Journal of the Royal Statistical Society, B, 63:411-423,2001.*
3. Hastie T., Tibshirani R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., Botstein, D. "Gene Shaving: a New Class of Clustering Methods for Expression Arrays". Postscript (2.9mb) or Adobe pdf (5.4mb) Tech. report. Jan 2000.