# Lab 5: Descriptive Models

## Background

Briefly, "gene expression" is typically a measure of the amount of mRNA (copies) of a particular gene (before translated to a protein) that has been produced by transcription in a particular tissue type at a particular time. Many genes are up-regulated or down-regulated together through complex regulatory networks and cascades based on the specific needs of a cell or tissue. This data is a small random subset taken from a larger group of 20000 genes that were measured. Gene expression data can be used for many different purposes. Suites of genes can be co-regulated in specific cancer types and not others. This can be used purely for prediction; it can also be used to find to find heterogeneity in cancer types (i.e. not all "types" of breast cancer are the same) which may help identify why individuals with different "types" of cancer respond to drugs differently. Gene expression can be used for the inference of "types" of specific cancers by identifying which particular genes or classes of genes have co-expression.

## Instructions

Load the "Session9GeneExpression1000.csv" and "Session9labels.csv" files. "GeneExpression1000.csv" contains gene expression data for 801 tumor samples (rows) and 1000 genes (columns) it has a header row of gene names. "labels.csv" contains the "Class" variable (column 2) which identifies which type of cancer each of the 801 samples in "GeneExpression1000.csv" come from. Below are the labels for each tumor type in the "Class" variable of "labels.csv"

| Abbreviation | Tumor type |
|---|---|
| COAD | Colon Adenocarcinoma |
| KIRC | Kidney Renal Clear Cell Carcinoma |
| LUAD | Lung Adenocarcinoma |
| PRAD | Prostate Adenocarcinoma |
| BRCA | Breast Invasive Carcinoma |

After centering and scaling, do two types of analyses with the gene expression data:

1. Principle Component Analysis
2. Your choice of either k-means followed up by hierarchical clustering or kohonen SOMs followed up by hierarchical clustering

**Principle Component Analysis**

Address the following questions:

- Which gene has the largest influence/contribution to the first principle component?
- Is that gene positively or negatively correlated with the first principle component?
- How many principle components are required do explain 60% of the total variance?

**k-means or kohonen SOMs**

Regardless of which method you use, choose answer these two questions:

1. What number of nodes is recommended kohonen SOM with this data?
2. Based on the Hartigan method, what k should be used for a k-means model with this data (Note: compare k=2 to k=15), this could take a little while)?

Fit your preferred model (kmeans or Kohonen SOM) with the appropriate number of k or the square with the appropriate number of nodes and then use hierarchical clustering to cluster the kmeans clusters or SOM nodes. From the hierarchical cluster cut the tree in to 5 groups (use the cutree function). Determine which of the five groups each sample is assigned then make a table with the "Class" variable from "Session9labels.csv" and these new groupings.

Report the table. Is there a strong correspondence between cancer cell line and the 5 groups from the hierarchical clustering? Which cancer line seems to be best represented by the groups from the clustering?

## The report

Develop a report (I recommend a Word (or other text editor) document) for your problem set that includes answers to all of the questions posed above, showing plots where appropriate.

Save your report as a PDF file and submit your report through the course 2GW site. Clean up your code and submit it as a supplementary file, along with your main report.

## Due date

Day 7, Week 9