Dimensionality Reduction

In this lab, we will be using an Affymetrix breast cancer data set that was run on the human HGU133A array. In the study, the investigators were interested in identifying transcripts that were differentially expressed in different histologic grade tumor samples to evaluate whether gene expression profiling could be used to improve histologic grading. Our interest in this data set does not focus on this same biological question. Rather, we would like to assess the processing variability in the data, since this type of variability can many times confound the biological variability. **(links 28,29, or 30)**

Since the array data was generated at different sites, we are interested in how this factor affects the variability in the samples. Specifically, we would like to use dimensionality reduction (DR) methods to evaluate the amount of variance that is explained by differences in processing sites. You will compute 4 different DR methods on this data set with the objectives of 1) summarizing the amount of variability is explained by differences in processing sites, and 2) understanding the visual differences in how the data structure is embedded when using difference methods of DR.

The paper is entitled "Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis" and is available on the course website.

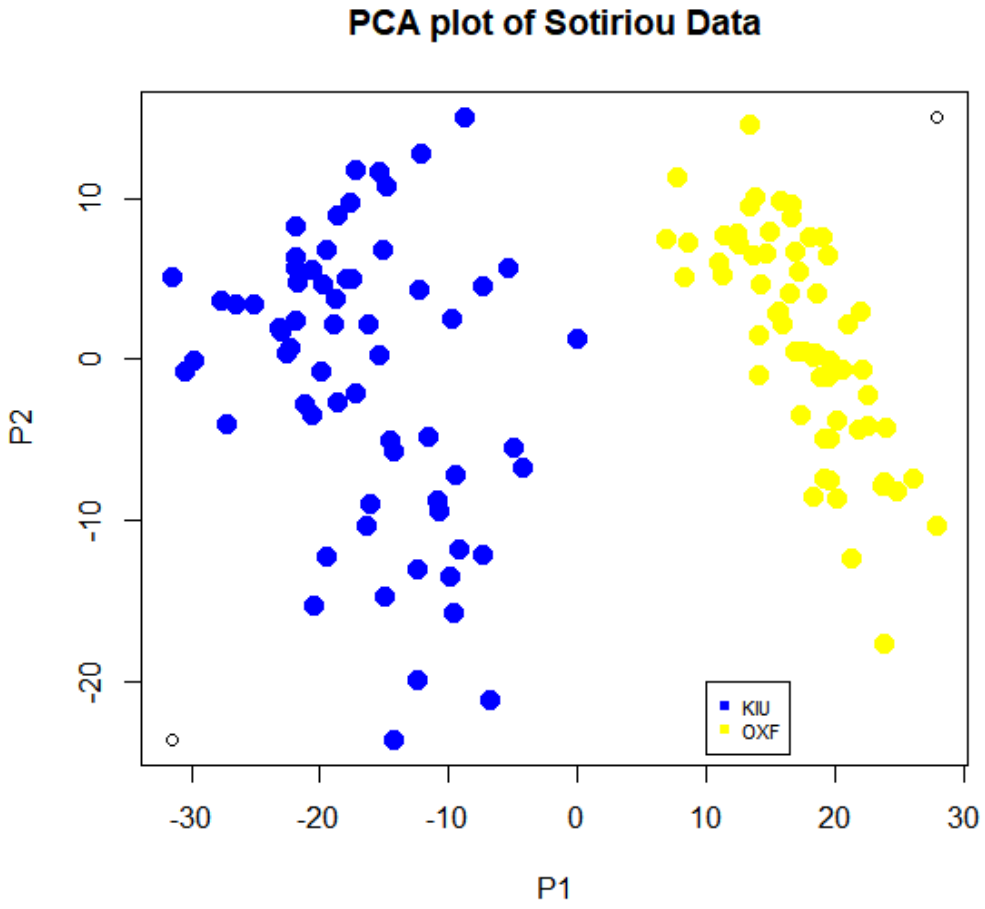
1.) Load the Sotiriou breast cancer data set from the class website as well as the annotation file.

```
>FilePath <- "C:/Users/fermi/Documents/Fall 2020/Lab 7/Sotiriou.txt"
>AnnoPath <- "C:/Users/fermi/Documents/Fall 2020/Lab
7/Sotiriou_annotations.txt"

>dat<- read.table(FilePath, header=T, row.names=1)
>anno<-read.table(AnnoPath, header=T, row.names=1)
```

2.) Calculate and plot a PCA plot. Label the points based on the site ("site" column header in annotation file). Make sure to add a legend to denote the colors of the two sites. - 2D PC1 vs PC2

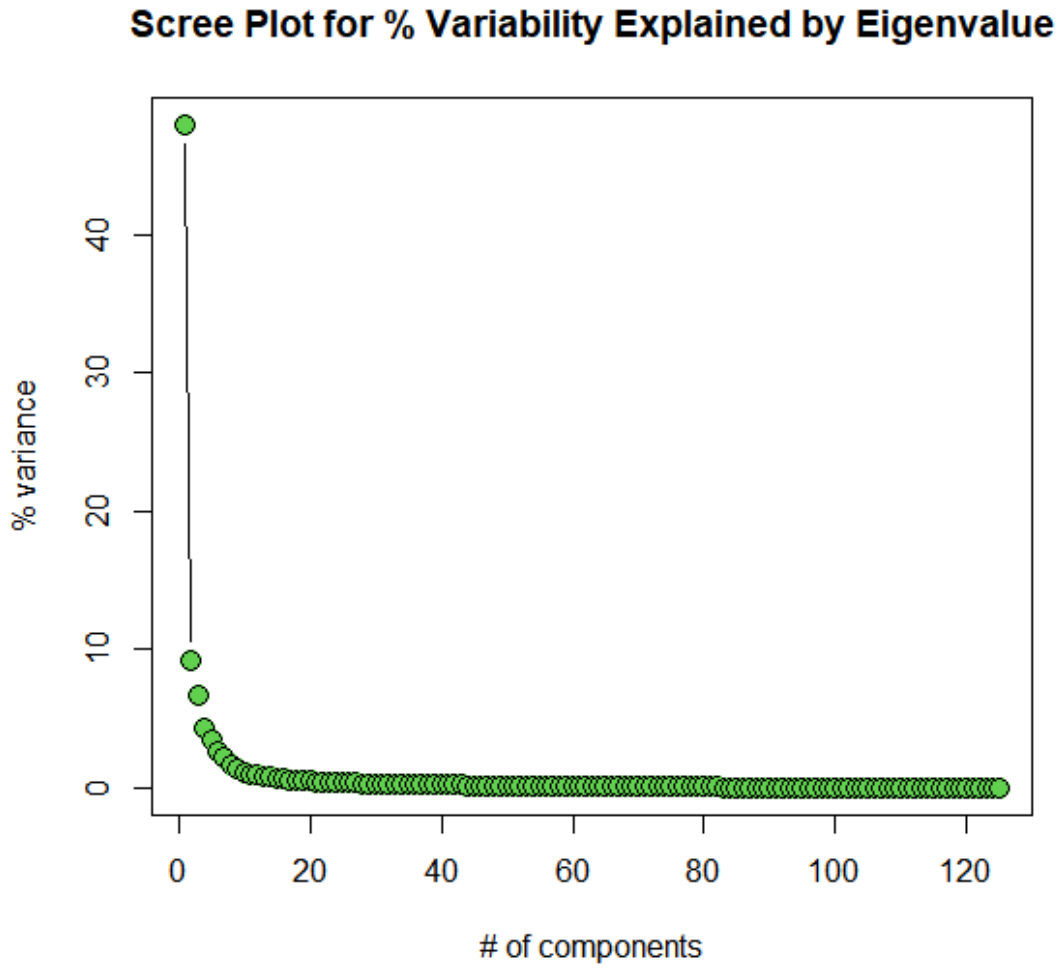
```
>dat.pca<-prcomp(t(dat),cor=F)
>newcols<-anno[,1]
>colnames(dat)<- newcols

>dat.loadings<-dat.pca$x[,1:3]

>plot(range(dat.loadings[,1]),range(dat.loadings[,2]),xlab='P1',
ylab='P2',main='PCA plot of Sotiriou Data')
```

```
>points(dat.loadings[,1][anno=="KIU"],dat.loadings[,2][anno=="KIU"],col='blue', pch=16, cex=1.5)
>points(dat.loadings[,1][anno=="OXF"],dat.loadings[,2][anno=="OXF"],col='yellow', pch=16, cex=1.5)
>legend(10,-20,c('KIU', 'OXF'),col=c("blue", "yellow"), pch=15, cex=.7, horiz=F)
```



3.) Calculate and plot the scree plot that corresponds to the PCA from question #2. Using only the first two eigenvalues, approximately how much variability in the data is explained? - Scree plot to understand how much variability is explained in eigen values

```
>dat.pca.var<-round(dat.pca$sdev^2/sum(dat.pca$sdev^2)*100,2)
> plot(c(1:length(dat.pca.var)),dat.pca.var,type="b", xlab="# of components", ylab="% variance", pch=21, col=1, bg=3, cex=1.5)
> title("Scree Plot for % Variability Explained by Eigenvalue")
```

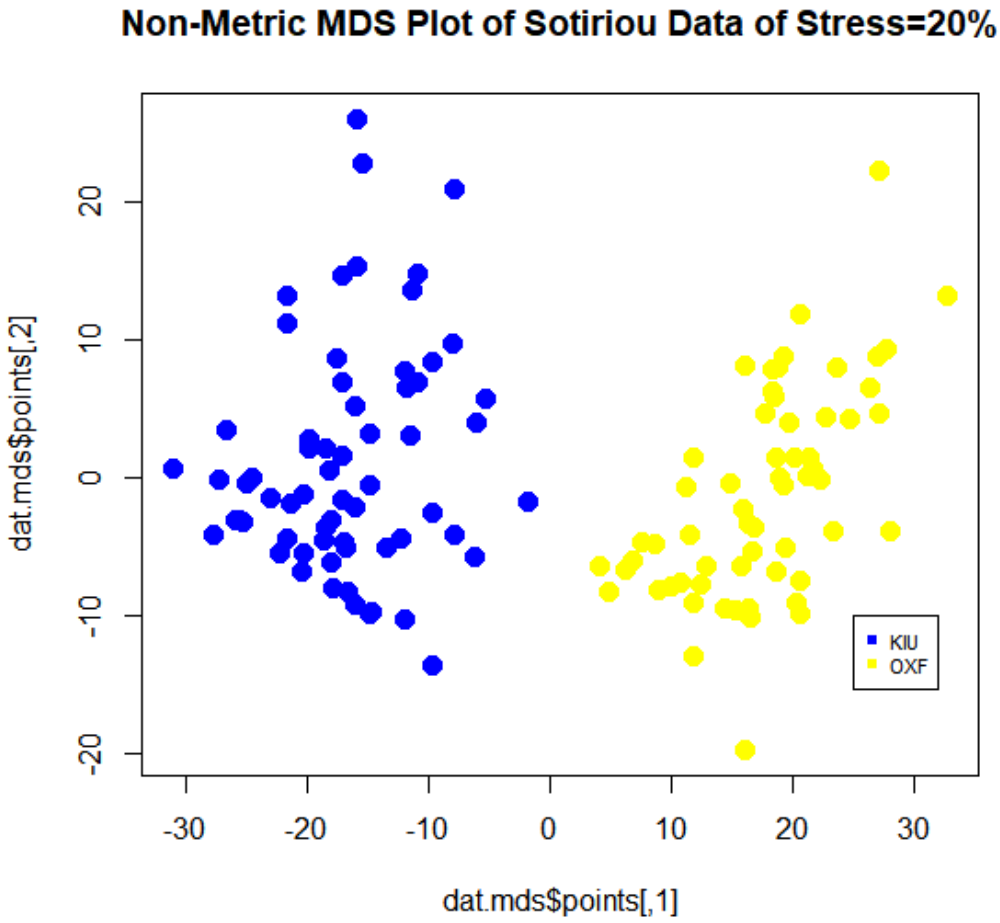


#Using only the first two eigenvalues, we see that by far the largest amount of variance lies in the first eigenvalue, with a bit more in the second-- together amounting to about 80% of variability. However, we do not see the desired "elbow" shape for a plateau in variance by the 2nd eigenvalue.

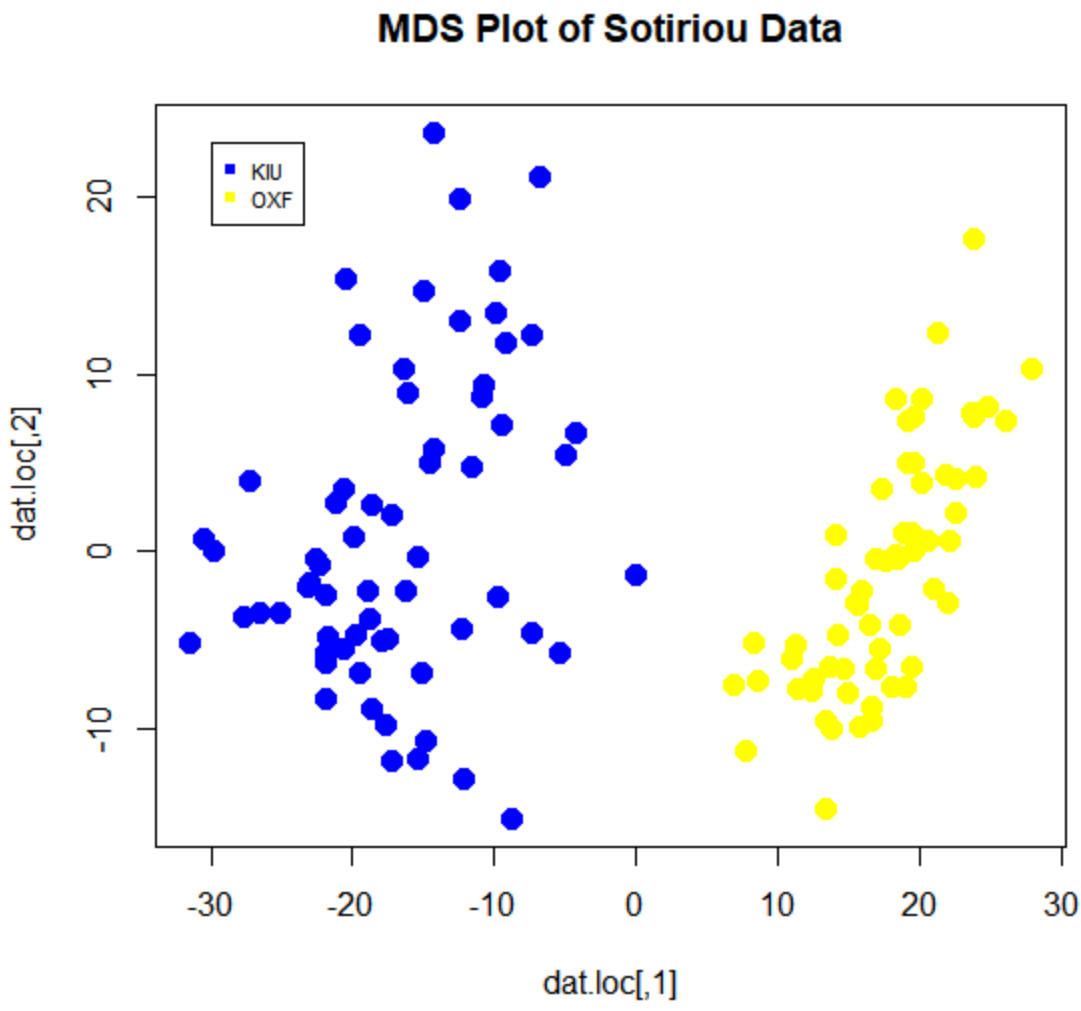
4.) Now calculate and plot 2 different MDS plots: 1) classic MDS and 2) nonmetric MDS. Label the points based on the site. Make sure to load the MASS library for the nonmetric MDS plot function. Also add a legend to both plots.

```
#non-metric MDS
>dat.dist<-dist(t(dat))
>dat.mds<-isoMDS(dat.dist)
>plot(dat.mds$points,type="n")
> points(dat.mds$points[,1][anno=="KIU"], dat.mds$points[,2][anno=="KIU"],
col="blue", pch=16, cex=1.5)
> points(dat.mds$points[,1][anno=="OXF"], dat.mds$points[,2][anno=="OXF"],
col="yellow", pch=16, cex=1.5)
> title(main="Non-Metric MDS Plot of Sotiriou Data of Stress=20%")
```

```
> legend(25,-10,c("KIU", "OXF"), col=c('blue', 'yellow'), pch=15, cex=.7,horiz=F)
```




```
#classical MDS
> dat.loc<-cmdscale(dat.dist)
> plot(dat.loc,type="n")
>
points(dat.loc[,1][anno=="KIU"],dat.loc[,2][anno=="KIU"],col="blue",pch=16,cex
=1.5)
>
points(dat.loc[,1][anno=="OXF"],dat.loc[,2][anno=="OXF"],col="yellow",pch=16,c
ex=1.5)
> title(main="MDS Plot of Sotiriou Data")
> legend(-30,23,c("KIU","OXF"), col=c('blue', 'yellow'),pch=15,cex=.7,horiz=F)
```


MDS Plot of Sotiriou Data
KIU
OXF
dat.loc[,2]
20
10
0
-10
-30
-20
-10
0
10
20
30
dat.loc[,1]