Lab #7
Dimensionality Reduction

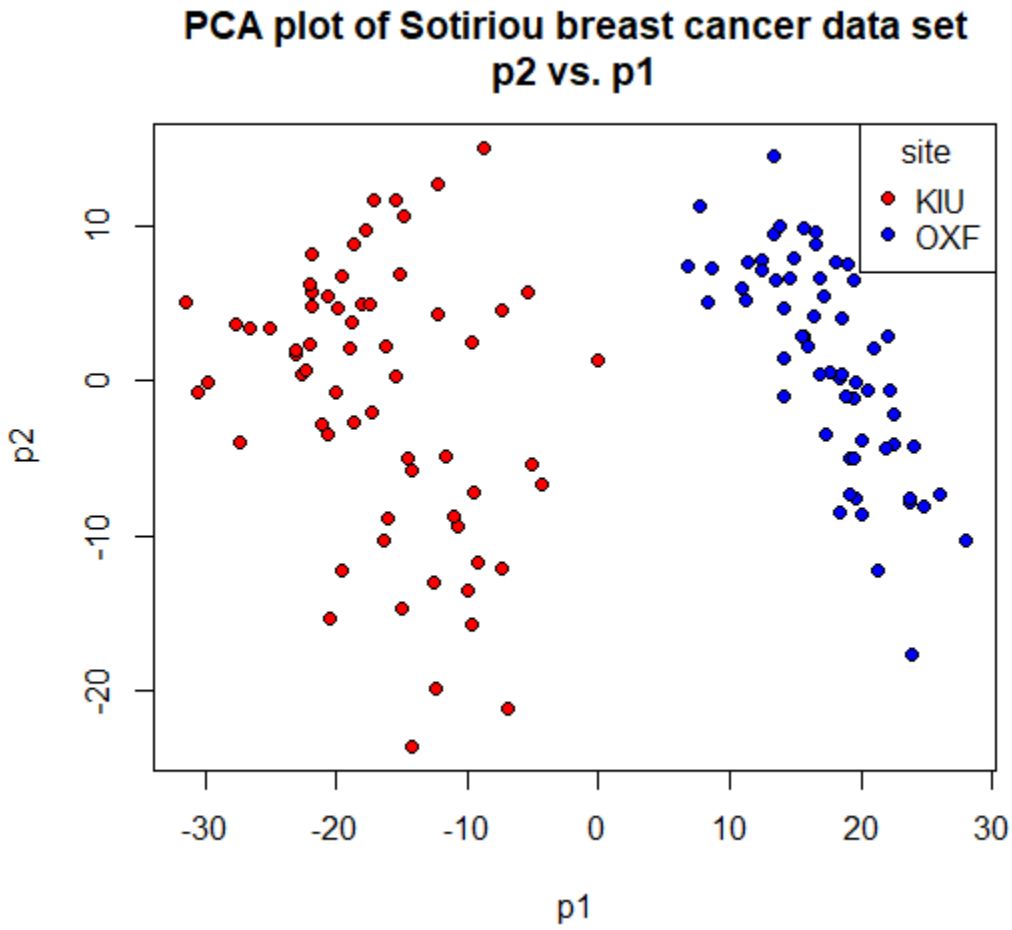
In this lab, we will be using an Affymetrix breast cancer data set that was run on the human HGU133A array. In the study, the investigators were interested in identifying transcripts that were differentially expressed in different histologic grade tumor samples to evaluate whether gene expression profiling 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.

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 Profi ling 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.

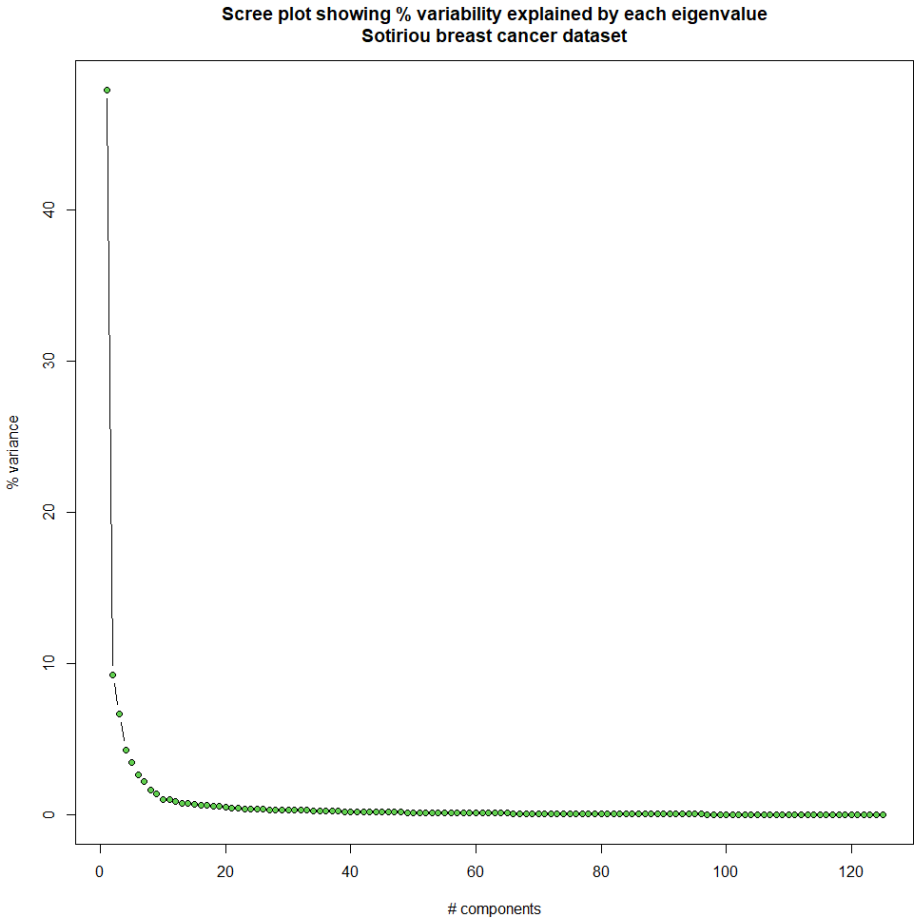
2.) Calculate and plot a PCA plot. Label the points based on the site ("site" column header in annotation file). Make sure to add a legend to denote the colors of the two sites.




```
plot(range(Sotiriou.pca.loading[,1]),range(Sotiriou.pca.loading[,2]),type="n",xlab='p1',ylab='p2',
     main='PCA plot of Sotiriou breast cancer data set\np2 vs. p1')
points(Sotiriou.pca.loading[,1][Sotiriou.ann$site=='KIU'], Sotiriou.pca.loading[,2][Sotiriou.ann$site=='KIU'],
       col=1,bg='red',pch=21,cex=1)
points(Sotiriou.pca.loading[,1][Sotiriou.ann$site=="OXF"], Sotiriou.pca.loading[,2][Sotiriou.ann$site=="OXF"],
       col=1,bg='blue',pch=21,cex=1)
legend('topright',title='site',c('KIU','OXF'),pch = 21,col=1,pt.bg  = c('red','blue'))
```

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?

```
Sotiriou.pca.var <- round(Sotiriou.pca$sdev^2 / sum(Sotiriou.pca$sdev^2)*100,2)
plot(c(1:length(Sotiriou.pca.var)),Sotiriou.pca.var,type="b",
     xlab="# components",ylab="% variance",pch=21,col=1,bg=3,cex=1)
title("Scree plot showing % variability explained by each eigenvalue\nSotiriou breast cancer dataset")
```

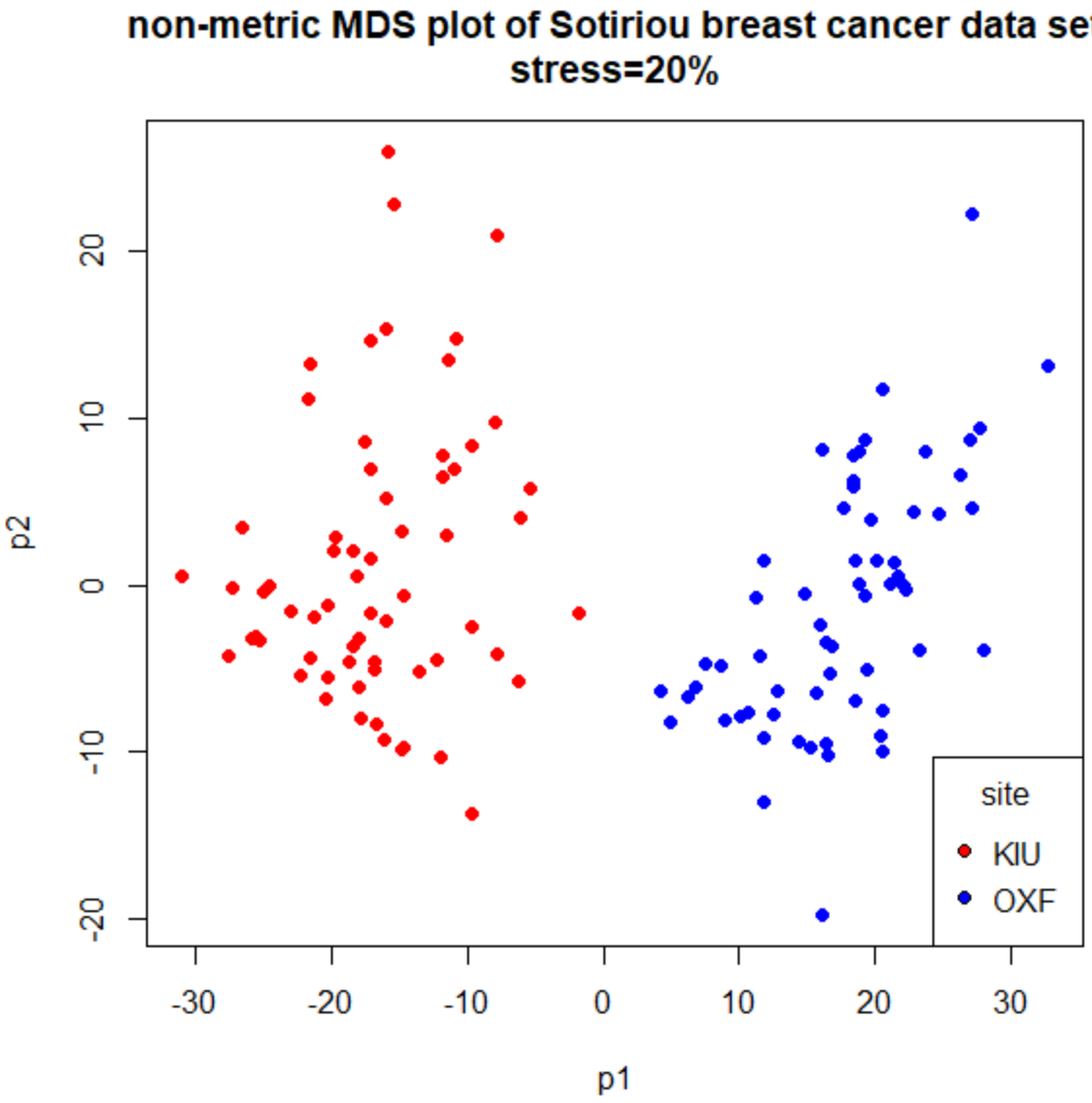


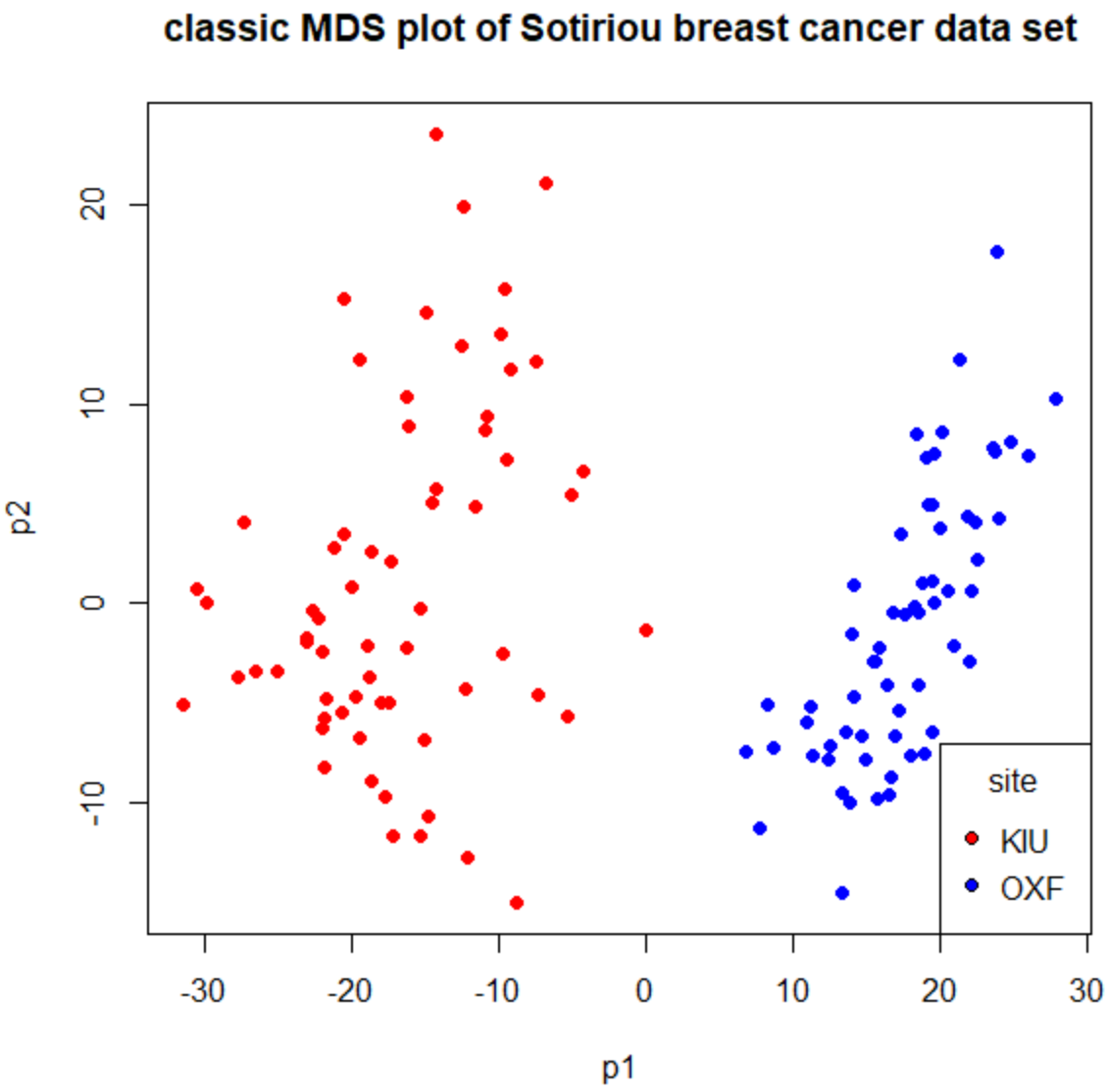
Nearly 60%

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
Sotiriou.dist <- dist(t(Sotiriou))
Sotiriou.mds <- isoMDS(Sotiriou.dist)
plot(Sotiriou.mds$points, type = "n",xlab='p1',ylab='p2')
points(Sotiriou.mds$points[,1][Sotiriou.ann$site=='KIU'], Sotiriou.mds$points[,2][Sotiriou.ann$site=='KIU'],
       col='red',pch=16,cex=1)
points(Sotiriou.mds$points[,1][Sotiriou.ann$site=="OXF"], Sotiriou.mds$points[,2][Sotiriou.ann$site=="OXF"],
       col='blue',pch=16,cex=1)
title(main='non-metric MDS plot of Sotiriou breast cancer data set\nstress=20%')
legend('bottomright',title='site',c('KIU','OXF'),col=1,pt.bg  = c('red','blue'),pch=21,cex=1,horiz=F)

#classic
Sotiriou.loc <- cmdscale(Sotiriou.dist)
plot(Sotiriou.loc, type = "n",xlab='p1',ylab='p2')
points(Sotiriou.loc[,1][Sotiriou.ann$site=='KIU'], Sotiriou.loc[,2][Sotiriou.ann$site=='KIU'],
       col='red',pch=16,cex=1.5)
points(Sotiriou.loc[,1][Sotiriou.ann$site=="OXF"], Sotiriou.loc[,2][Sotiriou.ann$site=="OXF"],
       col='blue',pch=16,cex=1.5)
title(main="MDS plot of breast cancer data set")
legend('bottomright',title='site',c('KIU','OXF'),col=1,pt.bg  = c('red','blue'),pch=21,cex=1,horiz=F)
```


non-metric MDS plot of Sotiriou breast cancer data set
stress=20%
p2
p1
20
10
0
-10
-20
-30
-20
-10
0
10
20
30
site
KIU
OXF

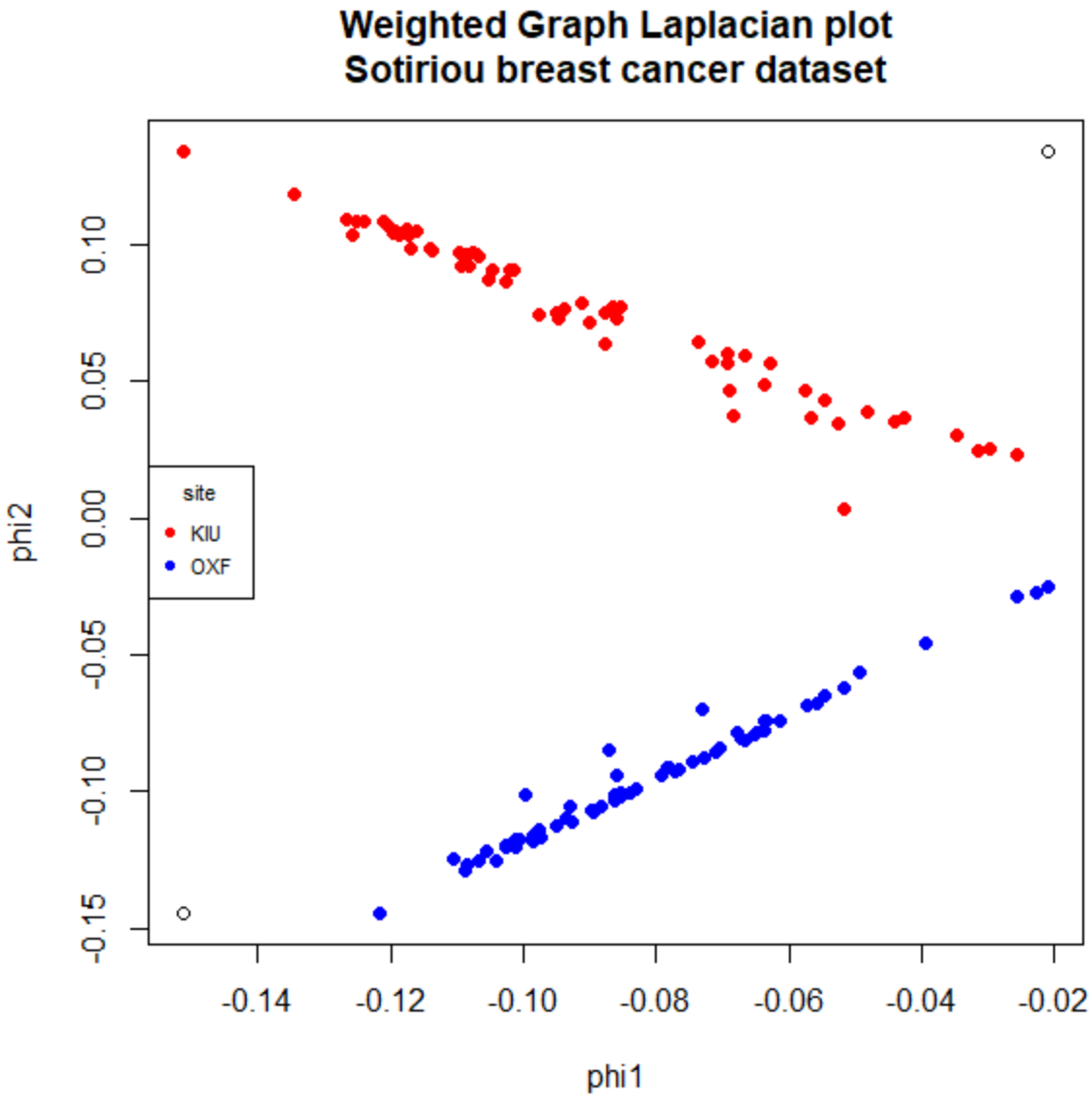


5.) Now, first center and scale the rows of the matrix with the commands below (assuming that dd is your data matrix):

```
> temp <- t(dd)
> temp <- scale(temp,center=T,scale=T)
```

```r
temp <- t(Sotiriou)
temp <- scale(temp,center=T,scale=T)

k.specClust2 <- function (X, qnt=NULL) {
  dist2full <- function(dis) {
    n <- attr(dis, "Size")
    full <- matrix(0, n, n)
    full[lower.tri(full)] <- dis
    full + t(full)
  }
  dat.dis <- dist(t(X),"euc")^2
  if(!is.null(qnt)) {eps <- as.numeric(quantile(dat.dis,qnt))}
  if(is.null(qnt)) {eps <- min(dat.dis[dat.dis!=0])}
  kernal <- exp(-1 * dat.dis/(eps))
  K1 <- dist2full(kernal)
  diag(K1) <- 0
  D = matrix(0,ncol=ncol(K1),nrow=ncol(K1))
  tmpe <- apply(K1,1,sum)
  tmpe[tmpe>0] <- 1/sqrt(tmpe[tmpe>0])
  tmpe[tmpe<0] <- 0
  diag(D) <- tmpe
  L <- D%*% K1 %*% D
  X <- svd(L)$u
  Y <- X / sqrt(apply(X^2,1,sum))
}
```




```
phi <- k.specClust2(t(temp),qnt=NULL)
plot(range(phi[,1]),range(phi[,2]),xlab="phi1",ylab="phi2",
     main="Weighted Graph Laplacian plot\nSotiriou breast cancer dataset")
points(phi[,1][Sotiriou.ann$site=='KIU'],phi[,2][Sotiriou.ann$site=='KIU'],col="red",pch=16,cex=1)
points(phi[,1][Sotiriou.ann$site=="OXF"],phi[,2][Sotiriou.ann$site=="OXF"],col="blue",pch=16,cex=1)
legend('left',title='site',c("KIU", "OXF"),col=c("red", "blue"),pch=16,cex=.7,horiz=F)
```

Then calculate and plot a two-dimensional embedding of the weighted graph Laplacian using t(temp) as the 'X' argument and 'NULL' for the qnt argument (don't use quotations to run the function). Label the points based on the site. Also add a legend.

Hint: for the colors, use the following command to get a color vector for the samples:
foo <- as.numeric(ann$site)