Principal Components Analysis

Pre-processing data

We'll be using the gene expression dataset for 17580 genes from 73 samples. There are two phenotypes, 0:no-diesase and 1:Parkinson's. We have an additional dataset containing 3 sample covariates.

```
library(rafalib)

e <- read.delim("data/counts.txt", row.names=1)
tab <- read.delim("data/phen.txt")
c <- read.delim("data/cov.txt", row.names=1)
```

We take the log-transform (normalize) of gene expression data and calculate the Z-score (standardize).

```
e_prime <- t(e) # Re-order data
L <- log2(1 + e_prime) # Log-transform data
head(L)[,1:2]
```

```
##        ENSG00000000003.10 ENSG00000000005.5
## C_0002           8.253656         0.7933007
## C_0003           8.207424         1.7152768
## C_0004           7.940356         2.9929645
## C_0005           7.760373         1.7993866
## C_0006           7.775682         2.2382442
## C_0008           8.086903         3.1276069
```

```
Z <- scale(L) # Z-score
head(Z)[,1:2]
```

```
##        ENSG00000000003.10 ENSG00000000005.5
## C_0002         -0.0964491       -1.27240600
## C_0003         -0.1742805       -0.40655202
## C_0004         -0.6238893        0.79336080
## C_0005         -0.9268921       -0.32756206
## C_0006         -0.9011194        0.08458153
## C_0008         -0.3771784        0.91980736
```

Computing principal components and percent variance

We use the `prcomp` function in the **`stats`** package to compute PCs of the scaled data.

```
pca <- prcomp(Z)
pca$sdev[1:10]
```

```
##  [1] 66.37047 49.97569 41.29392 36.34736 28.87607 25.43001 23.25561
##  [8] 18.73263 17.28147 16.18485
```

```
pca$rotation[1:5,1:2]
```

```
##                             PC1          PC2
## ENSG00000000003.10 -0.004452337  0.005470168
## ENSG00000000005.5   0.008304898 -0.002362205
## ENSG00000000419.8   0.007539490  0.011355355
## ENSG00000000457.8   0.003926745  0.007602980
## ENSG00000000460.12 -0.004979959  0.009618845
```

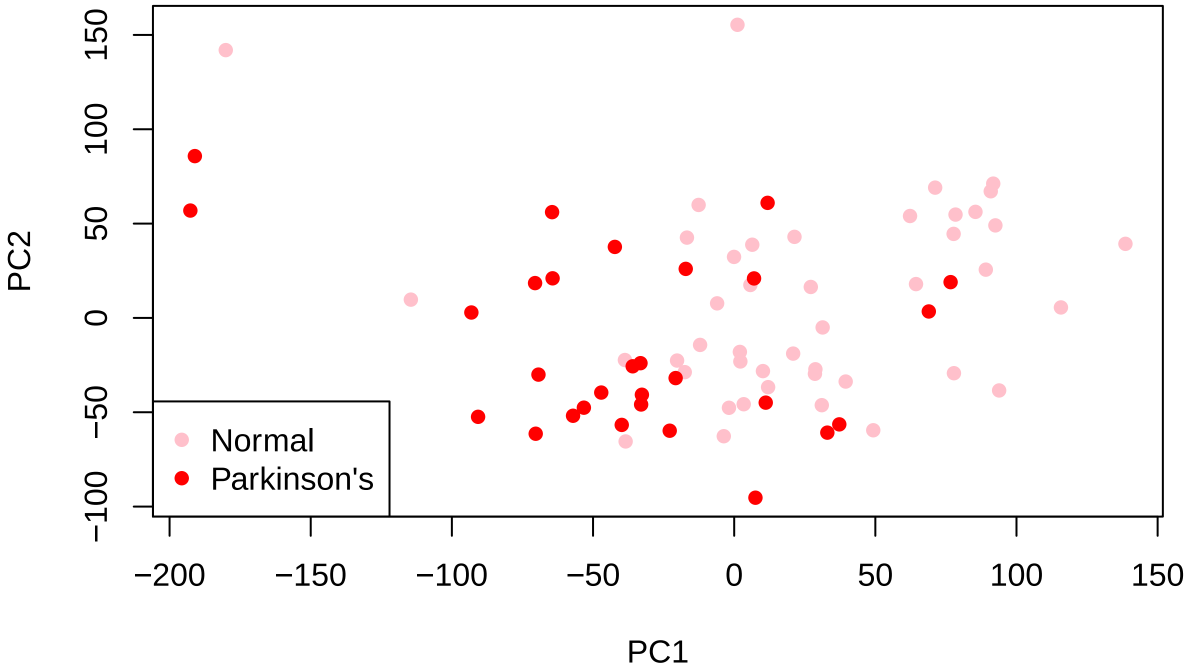
We now extract the variances of the components, and report the percent variance.

```
pca.var <- pca$sdev^2
pca.percent_var <- pca.var/sum(pca.var)
pca.percent_var[1:10]
```

```
##  [1] 0.25057109 0.14206880 0.09699589 0.07514963 0.04743045 0.03678529
##  [7] 0.03076356 0.01996083 0.01698802 0.01490042
```

Scatter plot of first two PC loadings

```
cols <- c('pink', 'red')
par(mfrow = c(1, 1))
plot(pca$x[, 1], pca$x[, 2], col=cols[tab$disease+1], pch=16, xlab="PC1", ylab="PC2")
legend("bottomleft", legend=c("Normal","Parkinson's"), col=cols, pch=16)
```



Pairwise Pearson correlation between PCs and covariates

Now, we want to see if any of the PCs are strongly correlated with any of the given covariates.