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", row.names=1)
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
```

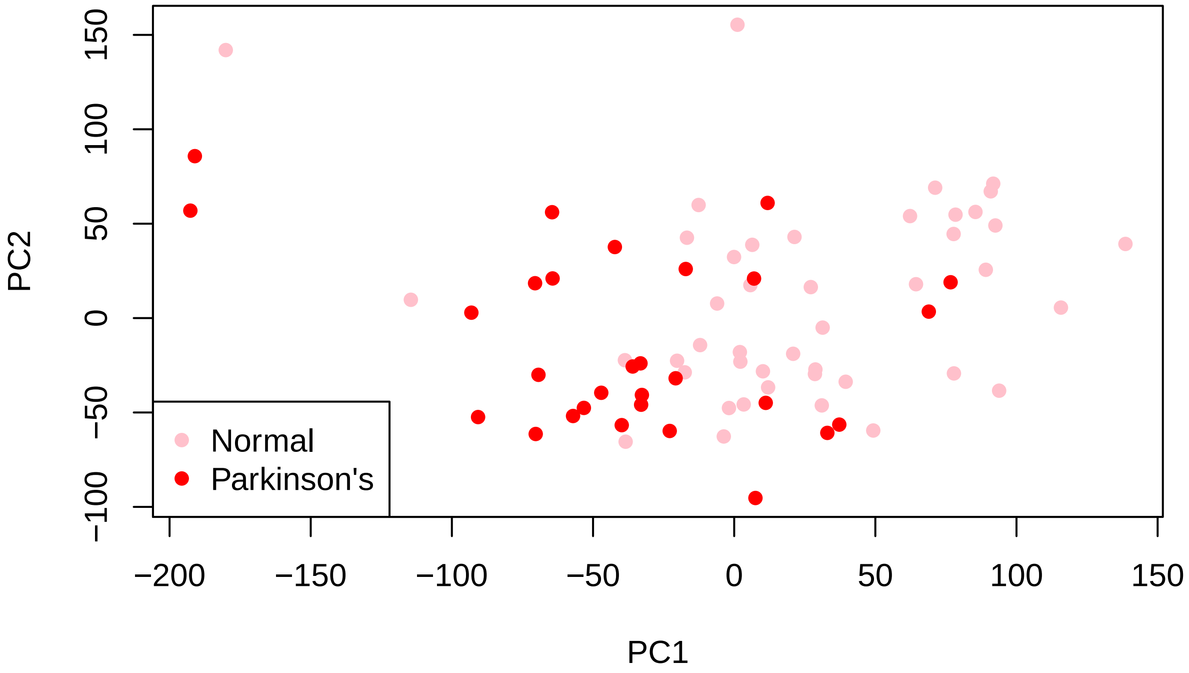
We now extract the variances of the components, and report the percent variance.

```
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 top 10 PCs are strongly correlated with any of the given covariates. Let's begin by looking at all the correlation estimates.

```
c_prime <- data.frame(t(c))
cor(pca$x[,1:10], c_prime, use="pairwise.complete.obs", method="pearson")
```

```
##      post_mortem_interval rna_integrity_number          age
## PC1            0.31373868           0.40541699 -0.414451027
## PC2            0.09390349           0.12251280 -0.261354434
## PC3           -0.07207742           0.23211063 -0.039000362
## PC4           -0.04857504           0.08878197  0.194298854
## PC5            0.02186293          -0.07913408  0.057051754
## PC6           -0.08107910          -0.15883359  0.123324186
## PC7            0.23782576          -0.15207238 -0.162206514
## PC8            0.39994010          -0.23834291 -0.429291440
## PC9            0.02768660          -0.02074695  0.093333472
## PC10          -0.11206010           0.00577553 -0.003989231
```

Let us say a PC is strongly correlated to a covariate if correlation estimate $|r| > 0.2$, and p-value < 0.05. Then, PCs strongly correlated to *post mortem interval* are

```
for (i in 1:10) { # For top 10 PCs
  c_test <- cor.test(pca$x[,i], c_prime[,1], use="pairwise.complete.obs", method="pearson")
  if (abs(c_test$estimate) > 0.2 & c_test$p.value < 0.05) {
    cat("PC", i, sep="")
    cat(" estimate =", c_test$estimate, "p-value =", c_test$p.value, "\n", sep=" ")
  }
}
```

```
## PC1 estimate = 0.3137387 p-value = 0.006873146
## PC7 estimate = 0.2378258 p-value = 0.04275626
## PC8 estimate = 0.3999401 p-value = 0.000455522
```

Similarly, those strongly correlated to *RNA integrity number* are

```
## PC1 estimate = 0.405417 p-value = 0.0003734216
## PC3 estimate = 0.2321106 p-value = 0.04815458
## PC8 estimate = -0.2383429 p-value = 0.04229354
```

and those to *age*

```
## PC1 estimate = -0.414451 p-value = 0.0002670358
## PC2 estimate = -0.2613544 p-value = 0.02551958
## PC8 estimate = -0.4292914 p-value = 0.0001507528
```

Pairwise Pearson correlation between disease status and covariates