Principal Component 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 of gene expression data and calculate the Z-score.

```
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(L)
pca$sdev[1:10]
```

```
##  [1] 51.09799 33.10467 27.13555 26.05815 21.33622 18.10651 16.95858
##  [8] 14.69798 14.36086 13.52880
```

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

```
##                             PC1          PC2
## ENSG00000000003.10  0.004307860 -0.005476928
## ENSG00000000005.5  -0.011743743 -0.002597112
## ENSG00000000419.8  -0.003197778  0.003272388
## ENSG00000000457.8  -0.001744380  0.006200323
## ENSG00000000460.12  0.002952869  0.006566861
```

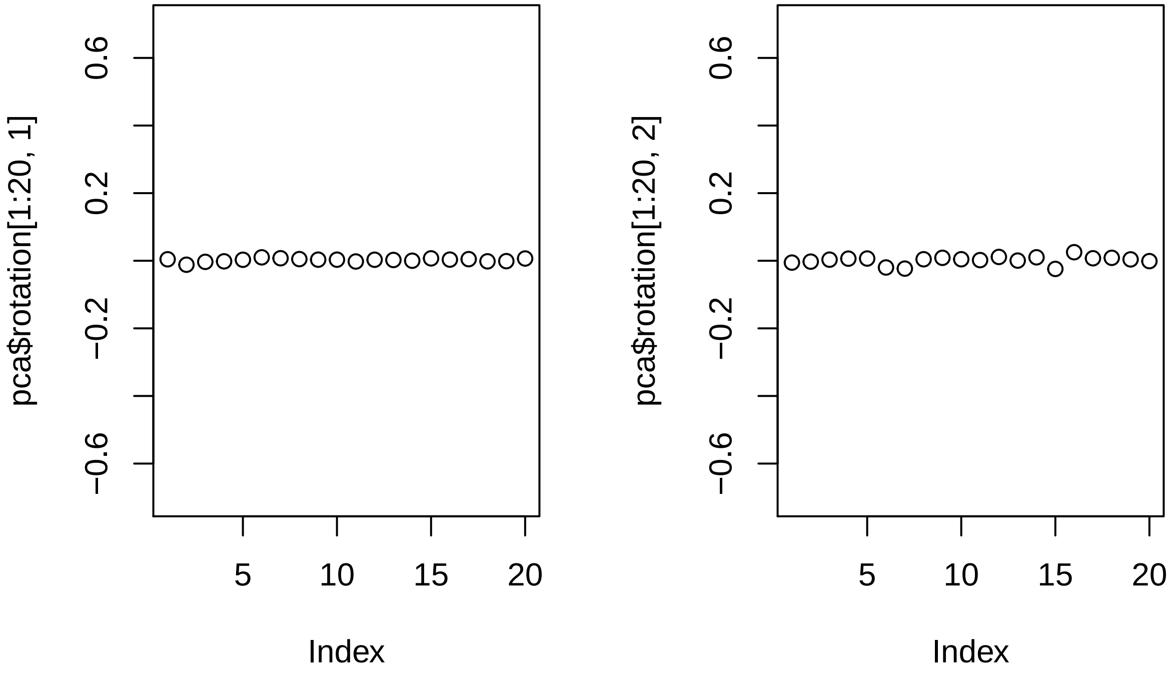
We now extract the variances of the components.

```
pca.var <- pca$sdev^2
pca.var[1:10]
```

```
##  [1] 2611.0045 1095.9189  736.3379  679.0271  455.2345  327.8456  287.5936
##  [8]  216.0307  206.2343  183.0283
```

Plots of first two PC loadings

```
par(mfrow = c(1, 2))
plot(pca$rotation[1:20, 1], ylim = c(-0.7, 0.7))
plot(pca$rotation[1:20, 2], ylim = c(-0.7, 0.7))
```


pca$rotation[1:20, 1]
Index
pca$rotation[1:20, 2]
Index