PCA

Parker Lambert

2024-01-30

Read in Data

```
# read in data
cancer <- read.table("PCA.example1.txt",header=TRUE)
```

Impute Mean

```
# fill in missing values
for(i in 1:ncol(cancer)){
cancer[is.na(cancer[,i]), i] <- round(mean(cancer[,i], na.rm = TRUE))
}
sum(is.na(cancer))
```

```
## [1] 0
```

```
# spectral decomp
date()
```

```
## [1] "Mon Aug 26 16:44:14 2024"
```

```
cancer.pc.eigen.cor <- prcomp(cancer)
date()
```

```
## [1] "Mon Aug 26 16:49:10 2024"
```

```
date()
```

```
## [1] "Mon Aug 26 16:49:10 2024"
```

```
cancer1 <- scale(cancer, scale = FALSE)
cancer.pc.scale.svd <- svd(cancer1)
date()
```

```
## [1] "Mon Aug 26 16:53:25 2024"
```

The singular value decomposition took less time by just under a minute.

Find Centroids

```
# our three populations
cancer.scores.eigen <- cancer.pc.eigen.cor$x
cancer.euro <- apply(cancer.scores.eigen[19662:19826,1:2], 2, mean)
cancer.asian <- apply(cancer.scores.eigen[19827:19963,1:2], 2, mean)
cancer.african <- apply(cancer.scores.eigen[19964:20166,1:2],2,mean)
```

```
# calculate centroid
(centroid <- rbind(cancer.asian, cancer.euro, cancer.african))
```

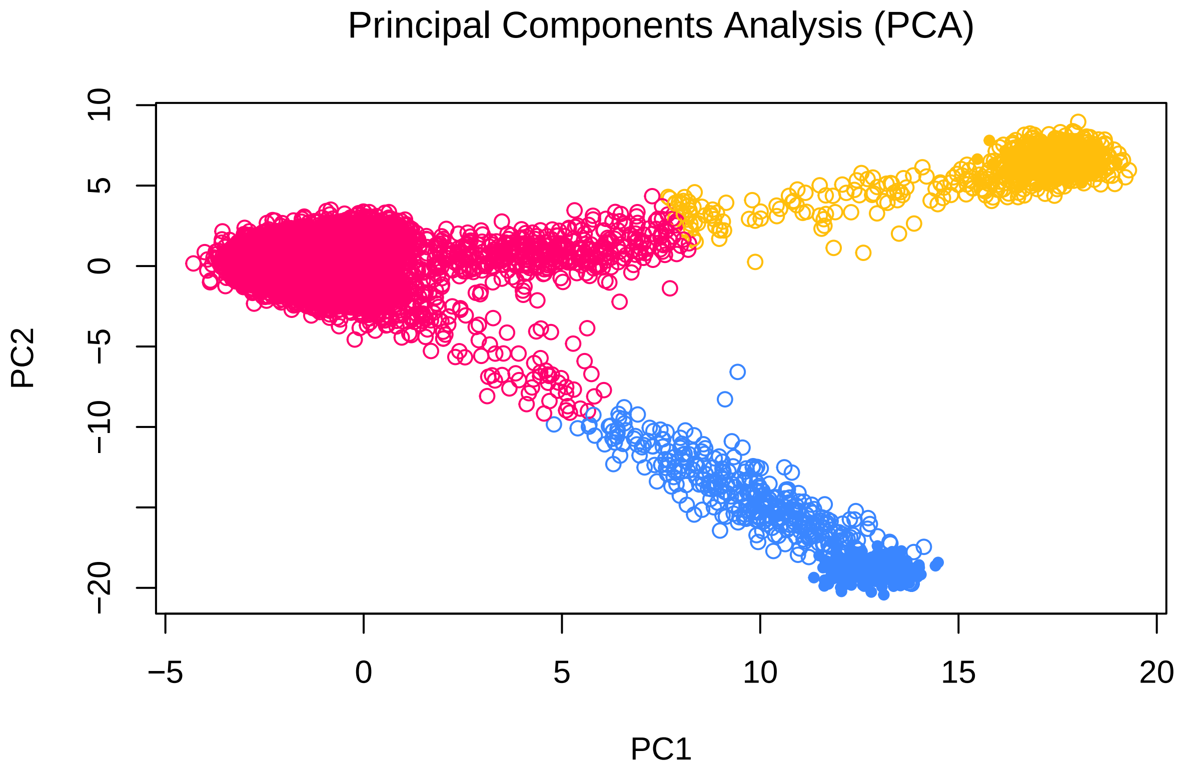
```
##                      PC1        PC2
## cancer.asian   17.177547   6.374995
## cancer.euro    -1.736351   0.390416
## cancer.african 12.804149 -18.833158
```

```
# cluster first 19661 samples into the closest cebtroid
sample <- cancer.scores.eigen[1:19661, 1:2]
# calculate distance from centroid
distances <- function(sample, centroids) {
  dist_matrix <- dist(rbind(sample, centroids))
  dist_vector <- as.vector(dist_matrix)[1:(length(centroids) - 1)]
  return(dist_vector)
}
# now calculate the distances
new <- t(apply(sample, 1, distances, centroid))
# find out which centroid our datapoint is closest too
(df <- apply(new, 1, which.min)) |>
  head(3)
```

```
## [1] 2 1 2
```

Graph PCA with Colored Clusters

```
# cluster points
colFirst <- c("#ffbe0b", "#ff006e", "#3a86ff")[df]
# plot
plot(cancer.scores.eigen[1:19661,1],cancer.scores.eigen[1:19661,2],type="p",col=colFirst,pch=1,xlab="PC
# colored points
points(cancer.scores.eigen[19662:19826,1],cancer.scores.eigen[19662:19826,2],col="#ff006e",pch=20)
points(cancer.scores.eigen[19827:19963,1],cancer.scores.eigen[19827:19963,2],col="#ffbe0b",pch=20)
points(cancer.scores.eigen[19964:20166,1],cancer.scores.eigen[19964:20166,2],col="#3a86ff",pch=20)
title(main="Principal Components Analysis (PCA)", col.main="black", font.main=1)
```


Principal Components Analysis (PCA)
PC2
10
5
0
−5
−10
−20
PC1
−5
0
5
10
15
20