Assignment 5 - Dimensionality Reduction

Abrar Faruque

November 20, 2025

The dataset can be accessed here: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

The phenotype attributions are as follows:

1) ID number
2) Diagnosis (M = malignant, B = benign) 3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Here is some code to read in the dataset

bCancer <- read.csv("breastCancer.csv", header = T)

1. Run a principal component analysis on the numeric variables of the bCancer dataset. How many principal components are there? [1]

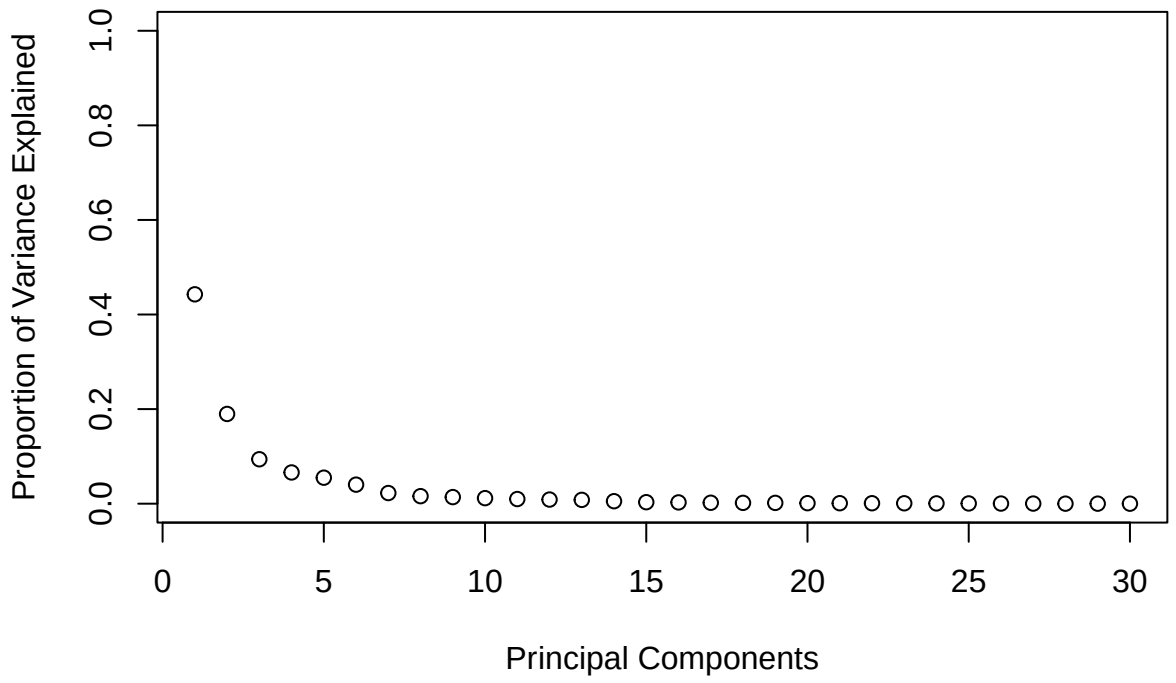
```
bCancer <- read.csv("breastCancer.csv", header = T)
pca_bCancer <- prcomp(bCancer[ , 3:32], scale = TRUE)
length(pca_bCancer$sdev)
```

```
## [1] 30
```

```
#there are 30 principal components
```

2. Visualize the percent variance explained by each component. How much of the total variance does the first and second components explain together? [1]

```r
variance_bCancer = pca_bCancer$sdev^2

# proportion variance
variance_bCancer / sum(variance_bCancer) -> prop_bCancer

# plot variance in a scree plot
plot(prop_bCancer,
     xlab = "Principal Components",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, 1))
```



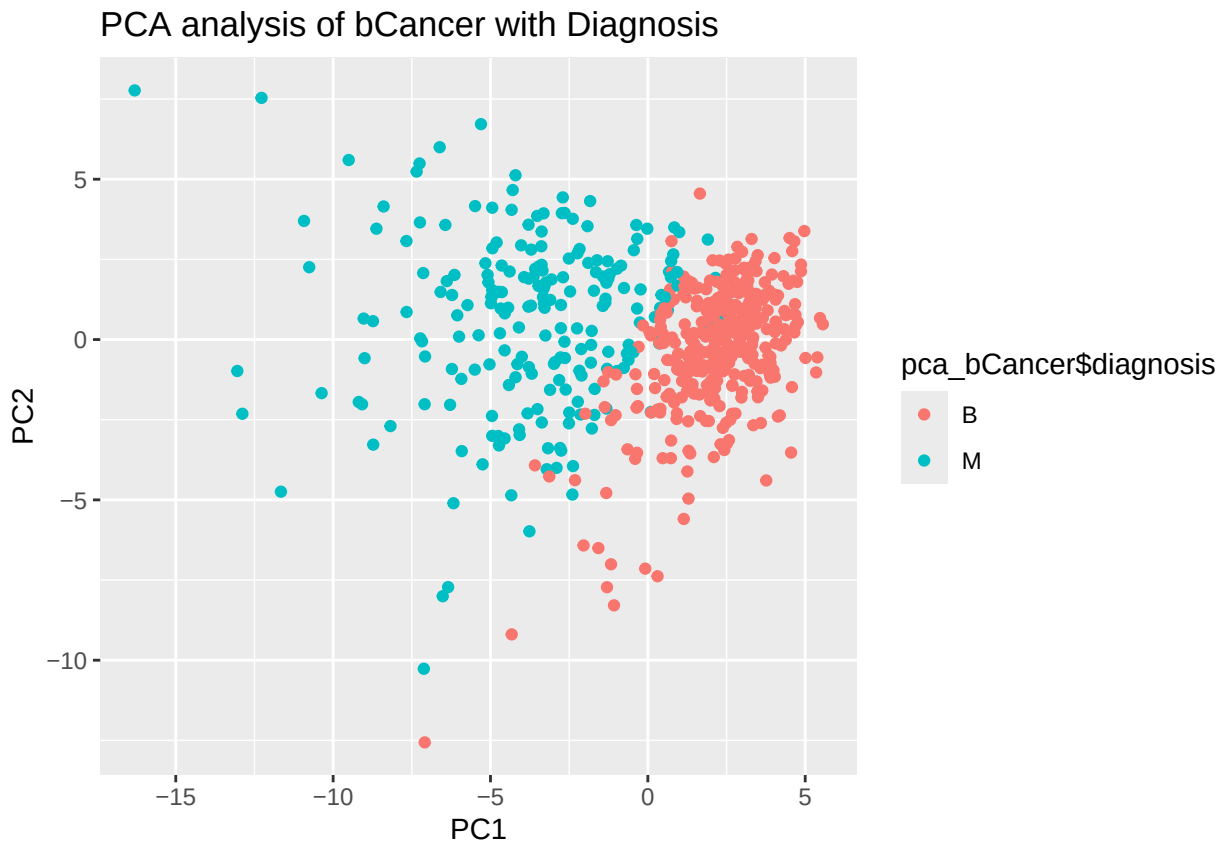

```r
#percentage variance for each PC
print((variance_bCancer[1]+variance_bCancer[2])/sum(variance_bCancer))
```

```
## [1] 0.6324321
```

```r
#'first and second component make up 63.2% of the total variance (explain 63.2% of
#'the variance), for further analysis, scaled data will be used because if gives a
#'better representation of PC1 and PC2 of the overall data
```

3. Visualize the PCA on a point plot with the samples colored according to their diagnosis. How do these two groups separate along PC1 and PC2? [1]

```
library(ggplot2)
pca_bCancer$diagnosis = bCancer$diagnosis
ggplot(pca_bCancer$x, aes(x = PC1, y = PC2)) +
  geom_point(aes(colour = pca_bCancer$diagnosis))+
  labs(title="PCA analysis of bCancer with Diagnosis")
```



4. What are the 5 variables most strongly contributing to the positive values in PC1? [1]

```
loadings_pc1 <- pca_bCancer$rotation[, "PC1"]
sorted_loadings_pc1 <- sort(abs(loadings_pc1), decreasing = TRUE)
print(head(sorted_loadings_pc1,n=5))
```

```
##  concave.points_mean       concavity_mean concave.points_worst
##            0.2608538            0.2584005            0.2508860
##     compactness_mean      perimeter_worst
##            0.2392854            0.2366397
```

```
#'the 5 variables most strongly contributing to the positive values in PC1 are:
#'1. concave.points_mean
#'2. concavity_mean
#'3. concave.points_worst
#'4. compactness_mean
#'5. perimeter_worst
#'in order from greatest to least
```

5. Create a heatmap of the top 10 most variable attributes. Discuss any trends you observe in the data in 5 sentences or less. [2]

Hint: this dataset is most similar to microarray data, not RNA-seq data. Hint: you may need to use `data.matrix` from base R to convert the original dataframe (minus select columns) to a matrix.

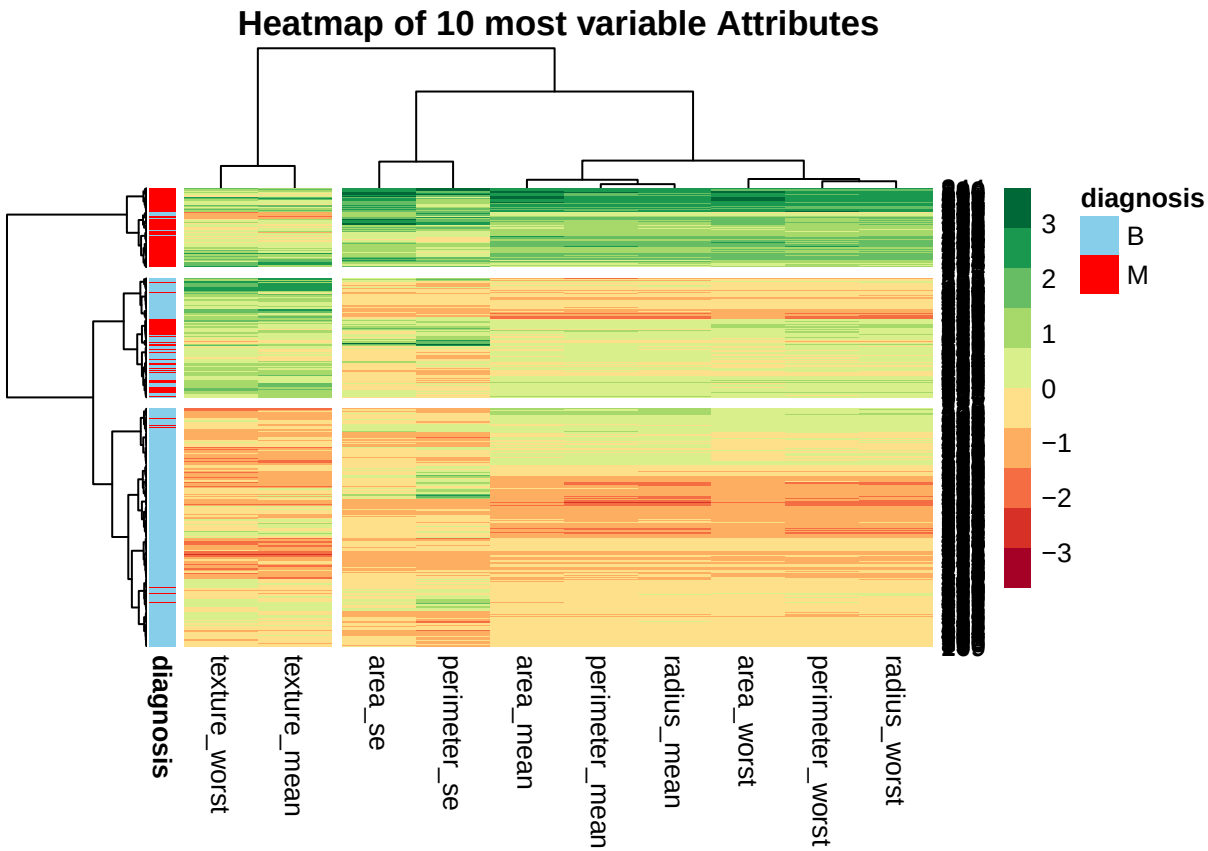
This is a tough question! Take your time, and move through the steps slowly. First, get the top 10 attributes, then make a heatmap that is coloured by diagnosis.

```
# Function to remove outliers in each column
remove_outliers_iqr <- function(x) {
  q1 <- quantile(x, 0.25)
  q3 <- quantile(x, 0.75)
  iqr <- q3 - q1
  lower <- q1 - 1.5 * iqr
  upper <- q3 + 1.5 * iqr
  x[x < lower | x > upper] <- NA
  return(x)
}

library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.5.2
```

```
library(RColorBrewer)

# Top 10 variables
ncol <- bCancer[, 3:32]
standard_deviations <- apply(ncol, 2, sd)^2
top10_names <- names(head(sort(standard_deviations, decreasing = TRUE),n=10))

# Extract top 10
top_10 <- bCancer[, top10_names]

# Apply to numerical columns
clean_top_10 <- apply(top_10, 2, remove_outliers_iqr)
rownames(clean_top_10) <- rownames(bCancer)

# Remove rows with any NA
clean_top_10_df <- as.data.frame(clean_top_10)
clean_top_10_df <- clean_top_10_df[complete.cases(clean_top_10_df), ]

# Convert to matrix
top_10_matrix <- as.matrix(clean_top_10_df)

# Row annotation
diagnosis_df <- data.frame(diagnosis = bCancer$diagnosis)
diagnosis_df <- diagnosis_df[rownames(clean_top_10_df), , drop = FALSE]

# Colors
anno_colors <- list(
  diagnosis = c(B = "skyblue", M = "red")
)
```

```
p <- pheatmap(
  top_10_matrix,
  col = brewer.pal(10, 'RdYlGn'),
  scale = "column",
  clustering_distance_cols = 'euclidean',
  clustering_distance_rows = 'euclidean',
  clustering_method = 'ward.D',
  annotation_row = diagnosis_df,
  annotation_colors = anno_colors,
  cutree_cols = 2,
  cutree_rows = 3,
  main = "Heatmap of 10 most variable Attributes"
)
```



Some noticable trends include 3 general groupings of bCancer specimines these are: B diagnosis, M diagnosis and a third group that is a mix of B and M diagnosis. M diagnosis tend to follow similar trends in area_worst, area_mean, area_se, perimeter_worst, perimeter_mean, radius_worst, radius_mean and perimeter_se than B diagnosis groups. M diagnosis also tends to have higher values in the previously mentioned groups than B diagnosis group, the third group (mix of M and B) have values that fall between M and B diagnosis groups. 'texture_worst' and texture_mean' do not appear to follow the trends of the other 8 groups. On average ignoring the 3rd group mentioned earlier, B diagnosis on average has lower values for the top 10 most variable attributes than M diagnosis according to the heat map