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
```

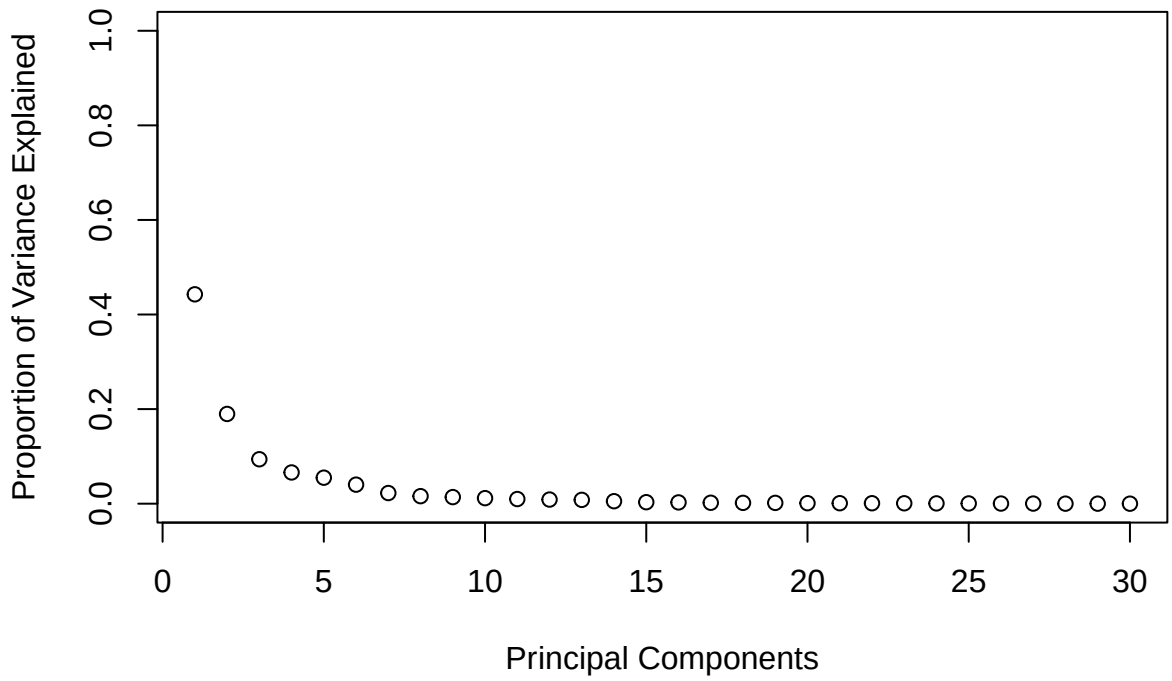
```
#there are 30 principal components
```

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))
```



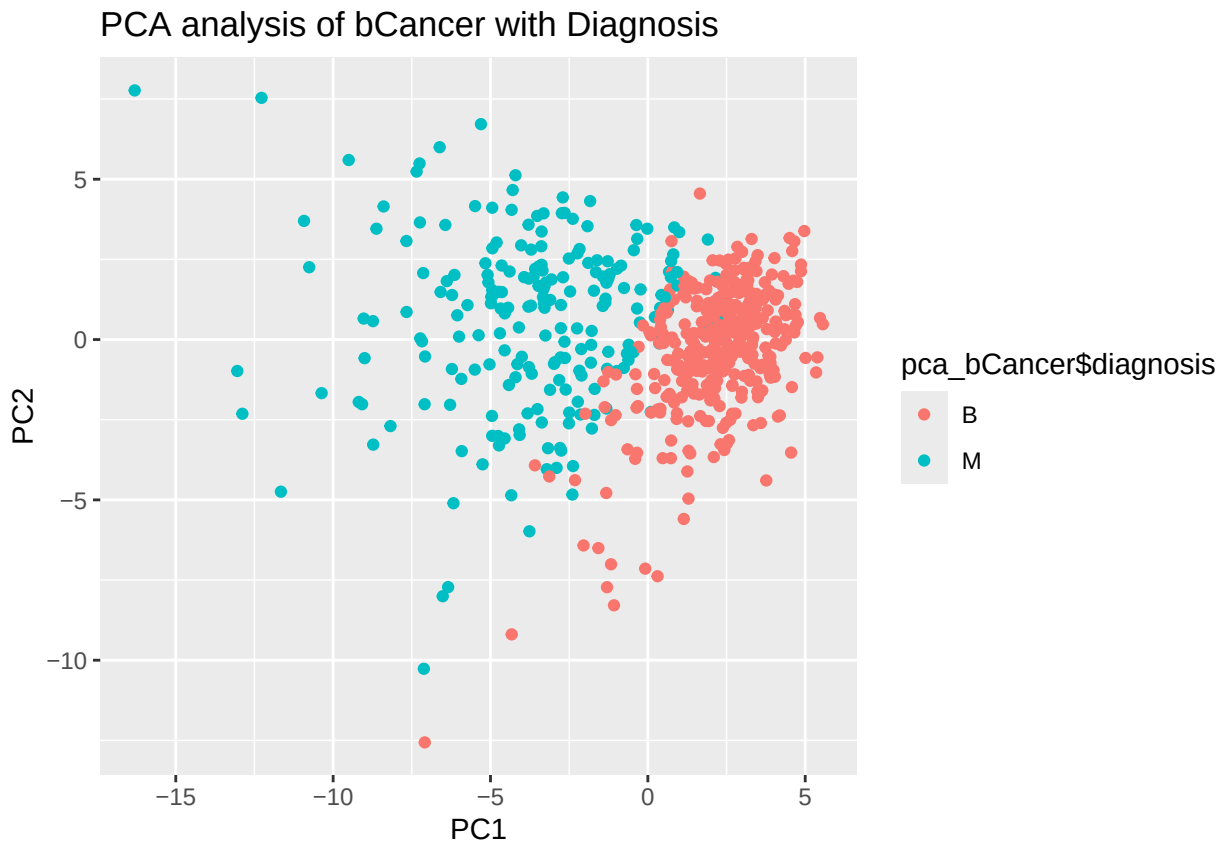

```r
#percentage variance for each PC
print((variance_bCancer[1]+variance_bCancer[2])/sum(variance_bCancer))
```

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

```r
#'first and second component make up 63.2% of the total variance (explain 63.2% of
#'the variance)
```

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(loadings_pc1, decreasing = TRUE)
print(head(sorted_loadings_pc1,n=5))
```

```
##          smoothness_se             texture_se           symmetry_se
##            -0.01453145            -0.01742803            -0.04249842
## fractal_dimension_mean   fractal_dimension_se
##            -0.06436335            -0.10256832
```

```
#'the 5 variables most strongly contributing to the positive values in PC1 are:
#'1. smoothness_se
#'2. texture_se
#'3. symmetry_se
#'4. fractal_dimension_mean
#'5. fractal_dimension_se
#'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.

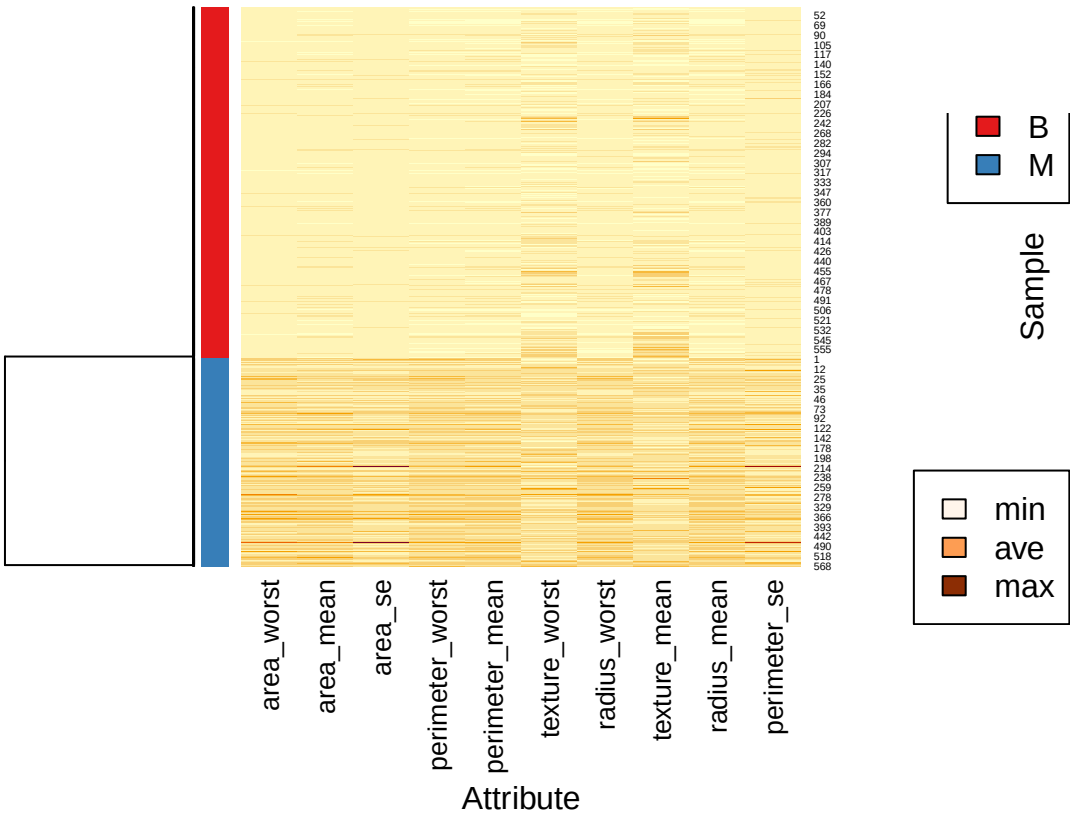
This is a tough question! Take your time, and move through the steps slowly. First, get the top 10 attributes, then make a heatmap that is coloured by diagnosis.

```
#finding top 10 greatest variable attributes
ncol = bCancer[,3:32]
standard_deviations <- sapply(ncol, sd)^2
print(head(sort(standard_deviations, decreasing = TRUE),n=10))
```

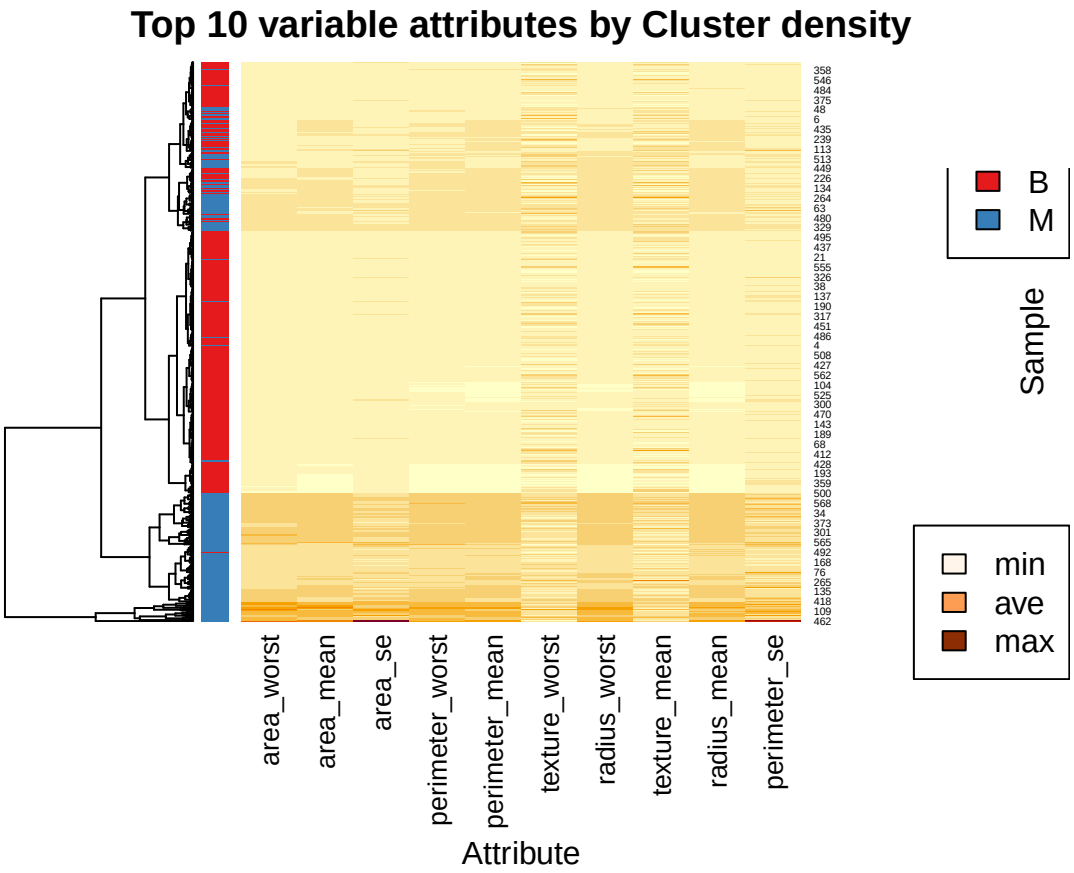
```
##      area_worst       area_mean         area_se perimeter_worst  perimeter_mean
##    3.241674e+05    1.238436e+05    2.069432e+03    1.129131e+03    5.904405e+02
##   texture_worst    radius_worst    texture_mean     radius_mean    perimeter_se
##    3.777648e+01    2.336022e+01    1.849891e+01    1.241892e+01    4.087896e+00
```

```
#generating heat map using top 10 most variable attributes
library(RColorBrewer)
row_distance <- dist(as.numeric(as.factor(bCancer$diagnosis)))
row_cluster <- hclust(row_distance, method = "complete")
row_dendrogram <- as.dendrogram(row_cluster)

#sorted_abs_loading <- sort(abs(loadings_pc1), decreasing = TRUE)
MostVarAtributes <- bCancer[names(head(sort(standard_deviations, decreasing = TRUE),n=10))]
MostVarAtributes_matrix <- data.matrix(MostVarAtributes)
my_group <- as.numeric(as.factor(bCancer$diagnosis))
colSide <- brewer.pal(9, "Set1")[my_group]
heatmap(MostVarAtributes_matrix, Colv = NA, Rowv = row_dendrogram,
        RowSideColors=colSide, scale="column",
        margins = c(10, 10),
        main='Top 10 variable attributes by bCancer diagnosis',
        ylab = 'Sample',
        xlab = 'Attribute'
        )
legend(x="topright", legend=c("B", "M"),
     fill=brewer.pal(9, "Set1"),
     inset=c(0,-0.05))
legend(x="bottomright", legend=c("min", "ave", "max"),
     fill=colorRampPalette(brewer.pal(8, "Oranges"))(3))
```




```
heatmap(MostVarAtributes_matrix, Colv = NA, RowSideColors=colSide, scale="column",
        main='Top 10 variable attributes by Cluster density',
        ylab = 'Sample',
        xlab = 'Attribute',
        margins = c(10, 10),
        )
legend(x="topright", legend=c("B", "M"),
     fill=brewer.pal(9, "Set1"),
     inset=c(0,-0.05))
legend(x="bottomright", legend=c("min", "ave", "max"),
     fill=colorRampPalette(brewer.pal(8, "Oranges"))(3))
```



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