R In-Class Lab: Principal Component Analysis with mtcars

Using PCA to Reduce Dimensionality

Using the functions discussed above, we will now show how PCA can be used to reduce the dimension of the mtcars dataset, which is a readily available dataset in base R. See https://www.datacamp.com/community/tutorials/pca-analysis-r for more information about the mtcars dataset.

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

There are two categorical features, namely ‘vs’ and ‘am’, which will be excluded, since PCA should only be used on numerical features. Also, we partition the mtcars dataset into two separate datasets, so we can have some “new” data to show how new observations are handled.

```
mtcars <- mtcars[, c(1:7, 10:11)]

indices <- sample(1:32, size = 27)

mtcars_newdata <- mtcars[-indices, ]
mtcars <- mtcars[indices, ]
```

Let’s try reducing the 9 numerical features to a smaller set using the principal components. First, as above, we use the prcomp function to calculate the principal components. Here, we both center and scale the features, since they are measured in different units.

```
pca_mtcars <- prcomp(mtcars, center = TRUE, scale = TRUE)
summary(pca_mtcars)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.3817 1.4679 0.67436 0.52508 0.38499 0.35156 0.30408
## Proportion of Variance 0.6303 0.2394 0.05053 0.03063 0.01647 0.01373 0.01027
## Cumulative Proportion  0.6303 0.8697 0.92024 0.95088 0.96734 0.98108 0.99135
##                            PC8     PC9
## Standard deviation     0.24809 0.12766
## Proportion of Variance 0.00684 0.00181
## Cumulative Proportion  0.99819 1.00000
```

From these results, we see that the first 3 principal components explain approximately 92% of the variation in the mtcars dataset. Thus, we can try reducing the 9 original features by projecting onto the first 3 principal components. To do this, we first center and scale the mtcars dataset, since we applied both of these actions when we calculated the principal components. Then, we rotate the data, which gives us the projection of the (centered and scaled) mtcars dataset onto the 9 principal components. Only the first 3 resulting columns are kept, which correspond to the projection onto the first 3 principal components.

```
mtcars_reduced <- scale(as.matrix(mtcars),
                        center = pca_mtcars$center,
                        scale = pca_mtcars$scale) %*% pca_mtcars$rotation
mtcars_reduced <- mtcars_reduced[, 1:3]

head(mtcars_reduced)
```

```
##                            PC1        PC2        PC3
## Datsun 710          -2.316827 -0.1740157  0.2337354
## Ford Pantera L       1.476906  2.9695487  0.2222096
## Fiat 128            -3.440017 -0.2035849  0.1191914
## Merc 230            -2.291891 -1.1093910 -1.8467126
## Lincoln Continental  3.580864 -0.9567075 -0.8468006
## Camaro Z28           2.468255  0.7046585  0.3867386
```

Now, we have 3 new features called PC1, PC2, and PC3, since they are the projections of the data onto the first 3 principal components. To handle new observations, we perform the same steps as above:

```
mtcars_newdata_reduced <- scale(mtcars_newdata,
                                center = pca_mtcars$center,
                                scale = pca_mtcars$scale) %*% pca_mtcars$rotation
mtcars_newdata_reduced <- mtcars_newdata_reduced[, 1:3]

head(mtcars_newdata_reduced)
```

```
##                          PC1        PC2         PC3
## Mazda RX4         -0.6877430  1.1506869  0.27206660
## Hornet Sportabout  1.4845162 -0.8099558  0.99772716
## Merc 280          -0.4143797  0.5818249 -0.78259167
## Honda Civic       -3.8399477  0.9664247 -0.06592879
## Pontiac Firebird   1.7691452 -0.9668468  0.82543849
```

It is important to notice that we use the center and scale, as well as rotation, based on the mtcars dataset and not the new data!