PCA

Tasheena Narraidoo, Michael Shi, Reynaldo Pena

12/08/16

Contents


```
summary(wine[,-13])
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.08     Min.   :0.000   Min.   : 0.6
##  1st Qu.: 6.40   1st Qu.:0.23     1st Qu.:0.250   1st Qu.: 1.8
##  Median : 7.00   Median :0.29     Median :0.310   Median : 3.0
##  Mean   : 7.22   Mean   :0.34     Mean   :0.319   Mean   : 5.4
##  3rd Qu.: 7.70   3rd Qu.:0.40     3rd Qu.:0.390   3rd Qu.: 8.1
##  Max.   :15.90   Max.   :1.58     Max.   :1.660   Max.   :65.8
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  1.0       Min.   :  6          Min.   :0.987
##  1st Qu.:0.038   1st Qu.: 17.0       1st Qu.: 77          1st Qu.:0.992
##  Median :0.047   Median : 29.0       Median :118          Median :0.995
##  Mean   :0.056   Mean   : 30.5       Mean   :116          Mean   :0.995
##  3rd Qu.:0.065   3rd Qu.: 41.0       3rd Qu.:156          3rd Qu.:0.997
##  Max.   :0.611   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH         sulphates        alcohol        quality
##  Min.   :2.72   Min.   :0.220   Min.   : 8.0   Min.   :3.00
##  1st Qu.:3.11   1st Qu.:0.430   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.21   Median :0.510   Median :10.3   Median :6.00
##  Mean   :3.22   Mean   :0.531   Mean   :10.5   Mean   :5.82
##  3rd Qu.:3.32   3rd Qu.:0.600   3rd Qu.:11.3   3rd Qu.:6.00
##  Max.   :4.01   Max.   :2.000   Max.   :14.9   Max.   :9.00
```

All our variables are numerical except wine type, so we will omit this from our exploratory factor analysis.

```
sapply(1:10, function(f) fa(wine[,-13], nfactors=f,rotate="varimax")$PVAL)
```

```
##  [1]  0.00e+00  0.00e+00  0.00e+00  0.00e+00  0.00e+00 9.28e-264  1.58e-35
##  [8]        NA        NA        NA
```

Now we run the factor analysis on 1-10 factors to see which would work best for our purposes. Unfortunately, it looks as though the P value of any number of factors is below .05. The maximum number of factors according to the degrees of freedom for our data set is 7. Even at 7 factors, the p value of the factor analysis is below .05, which means that we can reject the null hypothesis that the model is a good fit to the data.

```r
sapply(1:10, function(f) fa(whitewine, nfactors=f,rotate="varimax")$PVAL)
```

```
##  [1] 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 5.74e-82 1.08e-21
##  [8]       NA       NA       NA
```

```r
sapply(1:10, function(f) fa(redwine, nfactors=f,rotate="varimax")$PVAL)
```

```
##  [1]  0.00e+00  0.00e+00  0.00e+00  0.00e+00 5.23e-215  1.16e-83  3.47e-45
##  [8]        NA        NA        NA
```

The factor analysis also does not yield a p value below .05 for either red wine alone or white wine alone.

```r
FAWine = fa(whitewine, nfactors=7,rotate="varimax")
FAWine
```

```
## Factor Analysis using method =  minres
## Call: fa(r = whitewine, nfactors = 7, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                        MR1   MR3   MR5   MR2   MR4   MR6   MR7   h2    u2
## fixed.acidity         0.06  0.07 -0.03  0.97 -0.19  0.00 -0.02 1.00 0.005
## volatile.acidity      0.02  0.09 -0.02 -0.04 -0.01  0.75 -0.06 0.58 0.423
## citric.acid           0.07  0.01  0.08  0.26 -0.19 -0.16  0.22 0.19 0.812
## residual.sugar        0.96  0.08  0.23  0.01 -0.13  0.05 -0.05 1.00 0.005
## chlorides             0.03  0.41  0.08 -0.03 -0.11  0.04  0.16 0.22 0.781
## free.sulfur.dioxide   0.14  0.03  0.72 -0.05 -0.03 -0.12  0.05 0.56 0.437
## total.sulfur.dioxide  0.21  0.27  0.81  0.09  0.03  0.12  0.18 0.83 0.165
## density               0.79  0.51  0.24  0.20  0.06 -0.03  0.10 1.00 0.005
## pH                   -0.07 -0.11  0.01 -0.27  0.80 -0.02  0.23 0.78 0.221
## sulphates            -0.01  0.03  0.07  0.01  0.10 -0.03  0.39 0.17 0.834
## alcohol              -0.35 -0.87 -0.22 -0.05 -0.03  0.21  0.09 0.98 0.020
## quality              -0.04 -0.52  0.00 -0.07  0.01 -0.20  0.04 0.32 0.678
##                       com
## fixed.acidity         1.1
## volatile.acidity      1.1
## citric.acid           4.0
## residual.sugar        1.2
## chlorides             1.6
## free.sulfur.dioxide   1.2
## total.sulfur.dioxide  1.6
## density               2.1
## pH                    1.5
## sulphates             1.2
## alcohol               1.6
## quality               1.3
##
##                         MR1  MR3  MR5  MR2  MR4  MR6  MR7
## SS loadings            1.74 1.57 1.36 1.15 0.76 0.70 0.34
## Proportion Var         0.15 0.13 0.11 0.10 0.06 0.06 0.03
## Cumulative Var         0.15 0.28 0.39 0.49 0.55 0.61 0.63
## Proportion Explained   0.23 0.21 0.18 0.15 0.10 0.09 0.04
## Cumulative Proportion  0.23 0.43 0.61 0.76 0.86 0.96 1.00
##
```

```
## Mean item complexity =  1.6
## Test of the hypothesis that 7 factors are sufficient.
##
## The degrees of freedom for the null model are  66  and the objective function was  5.41 with Chi Squa
## The degrees of freedom for the model are 3  and the objective function was  0.02
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.06
##
## The harmonic number of observations is  4898 with the empirical chi square  102  with prob <  5.6e-21
## The total number of observations was  4898  with MLE Chi Square =  101  with prob <  1.1e-21
##
## Tucker Lewis Index of factoring reliability =  0.918
## RMSEA index =  0.082  and the 90 % confidence intervals are  0.068 0.096
## BIC =  75.2
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 MR1  MR3  MR5  MR2  MR4
## Correlation of scores with factors             0.99 0.97 0.90 0.99 0.85
## Multiple R square of scores with factors       0.98 0.95 0.81 0.98 0.72
## Minimum correlation of possible factor scores  0.96 0.90 0.62 0.97 0.44
##                                                 MR6  MR7
## Correlation of scores with factors             0.81 0.78
## Multiple R square of scores with factors       0.66 0.61
## Minimum correlation of possible factor scores  0.32 0.21
```

PCA

Red Wine

```
cor(redwine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000         -0.25613      0.6717
## volatile.acidity            -0.2561          1.00000     -0.5525
## citric.acid                  0.6717         -0.55250      1.0000
## residual.sugar               0.1148          0.00192      0.1436
## chlorides                    0.0937          0.06130      0.2038
## free.sulfur.dioxide         -0.1538         -0.01050     -0.0610
## total.sulfur.dioxide        -0.1132          0.07647      0.0355
## density                      0.6680          0.02203      0.3649
## pH                          -0.6830          0.23494     -0.5419
## sulphates                    0.1830         -0.26099      0.3128
## alcohol                     -0.0617         -0.20229      0.1099
## quality                      0.1241         -0.39056      0.2264
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                0.11478   0.09371            -0.15379
## volatile.acidity             0.00192   0.06130            -0.01050
## citric.acid                  0.14358   0.20382            -0.06098
## residual.sugar               1.00000   0.05561             0.18705
## chlorides                    0.05561   1.00000             0.00556
```

```
## free.sulfur.dioxide         0.18705   0.00556             1.00000
## total.sulfur.dioxide        0.20303   0.04740             0.66767
## density                     0.35528   0.20063            -0.02195
## pH                         -0.08565  -0.26503             0.07038
## sulphates                   0.00553   0.37126             0.05166
## alcohol                     0.04208  -0.22114            -0.06941
## quality                     0.01373  -0.12891            -0.05066
##                       total.sulfur.dioxide density      pH sulphates
## fixed.acidity                      -0.1132  0.6680 -0.6830   0.18301
## volatile.acidity                    0.0765  0.0220  0.2349  -0.26099
## citric.acid                         0.0355  0.3649 -0.5419   0.31277
## residual.sugar                      0.2030  0.3553 -0.0857   0.00553
## chlorides                           0.0474  0.2006 -0.2650   0.37126
## free.sulfur.dioxide                 0.6677 -0.0219  0.0704   0.05166
## total.sulfur.dioxide                1.0000  0.0713 -0.0665   0.04295
## density                             0.0713  1.0000 -0.3417   0.14851
## pH                                 -0.0665 -0.3417  1.0000  -0.19665
## sulphates                           0.0429  0.1485 -0.1966   1.00000
## alcohol                            -0.2057 -0.4962  0.2056   0.09359
## quality                            -0.1851 -0.1749 -0.0577   0.25140
##                       alcohol quality
## fixed.acidity        -0.0617  0.1241
## volatile.acidity     -0.2023 -0.3906
## citric.acid           0.1099  0.2264
## residual.sugar        0.0421  0.0137
## chlorides            -0.2211 -0.1289
## free.sulfur.dioxide  -0.0694 -0.0507
## total.sulfur.dioxide -0.2057 -0.1851
## density              -0.4962 -0.1749
## pH                    0.2056 -0.0577
## sulphates             0.0936  0.2514
## alcohol               1.0000  0.4762
## quality               0.4762  1.0000
```

There are a number of moderately high correlations in the red wine data set (above .3). This means that there is some sort of underlying structure in the data and PCA could work on this data set.

```
PCWineRed <- prcomp(redwine[,-13], scale=TRUE)
PCWineRed$rotation
```

```
##                             PC1      PC2     PC3      PC4     PC5     PC6
## fixed.acidity         0.48788 -0.00417  0.1648  0.23110 -0.0788  0.0555
## volatile.acidity     -0.26513  0.33897  0.2271 -0.04186  0.2994  0.2973
## citric.acid           0.47334 -0.13736 -0.1002  0.05674 -0.1201  0.1366
## residual.sugar        0.13915  0.16774 -0.2436  0.38304  0.7094  0.1093
## chlorides             0.19743  0.18979  0.0266 -0.65478  0.2662  0.3373
## free.sulfur.dioxide  -0.04588  0.25948 -0.6161  0.03371 -0.1594 -0.0426
## total.sulfur.dioxide  0.00407  0.36397 -0.5407  0.02846 -0.2185  0.1160
## density               0.37030  0.33078  0.1687  0.20069  0.2088 -0.4257
## pH                   -0.43272 -0.06544 -0.0698  0.00547  0.2576 -0.4804
## sulphates             0.25454 -0.10933 -0.2129 -0.56050  0.2148 -0.4037
## alcohol              -0.07318 -0.50271 -0.2250  0.09170  0.2597  0.3922
## quality               0.11249 -0.47317 -0.2234  0.03667  0.1376 -0.1418
```

```
##                              PC7     PC8     PC9     PC10    PC11     PC12
## fixed.acidity             -0.307  0.2005 -0.1746  0.18296 -0.2564  0.63858
## volatile.acidity          -0.626  0.1461 -0.0602 -0.15511  0.3772  0.00466
## citric.acid                0.244  0.2963 -0.2210 -0.34609  0.6243 -0.07004
## residual.sugar             0.284 -0.1706  0.2782  0.05224  0.0881  0.18365
## chlorides                  0.231 -0.1869 -0.4199  0.00386 -0.2086  0.05393
## free.sulfur.dioxide       -0.138 -0.0194 -0.3180  0.58539  0.2379 -0.05192
## total.sulfur.dioxide      -0.110  0.0899  0.1218 -0.58919 -0.3550  0.06979
## density                   -0.123  0.0795 -0.2491 -0.04354 -0.2315 -0.56664
## pH                         0.186  0.3147 -0.4619 -0.20761 -0.0056  0.34123
## sulphates                 -0.233  0.2755  0.4527  0.07192  0.0976  0.06779
## alcohol                   -0.122  0.4712 -0.0965  0.11061 -0.3199 -0.31764
## quality                   -0.412 -0.6122 -0.2402 -0.26024  0.0525  0.00847
```

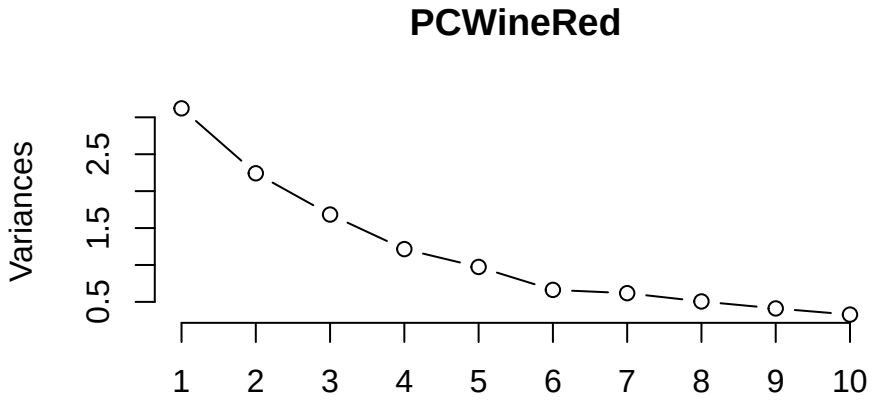
```
summary(PCWineRed)
```

```
## Importance of components:
##                           PC1   PC2   PC3   PC4    PC5    PC6    PC7    PC8
## Standard deviation     1.77 1.497 1.297 1.102 0.9865 0.8140 0.7863 0.7112
## Proportion of Variance 0.26 0.187 0.140 0.101 0.0811 0.0552 0.0515 0.0422
## Cumulative Proportion  0.26 0.447 0.587 0.688 0.7695 0.8247 0.8763 0.9184
##                           PC9   PC10  PC11    PC12
## Standard deviation     0.6413 0.5726 0.425 0.24396
## Proportion of Variance 0.0343 0.0273 0.015 0.00496
## Cumulative Proportion  0.9527 0.9800 0.995 1.00000
```

We run our PCA on correlations over covariances because our variables are on different scales and we do not want to weight variables with higher covariances differently. Now we would like to choose how many principal components to keep. Looking at the variance explained explained by each principal component, we can see we need 5 principal components to explain 70% of the variation in the data. According to the eigenvalue >0.7 rule, we need 5 principal components as well.

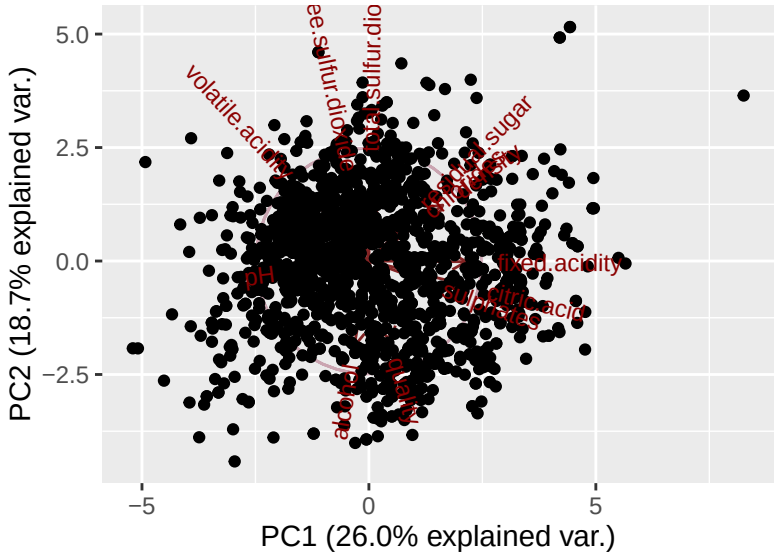
Looking at the loadings, it looks like the first PC represents the negative relationship between fixed acidity, citric acid, and density and pH. The second represents the positive relationship between alcohol and quality. The third represents free and total sulfur dioxide and the fourth PC represents chlorides and sulphates. The fifth PC represents residual sugar levels.

```
screeplot(PCWineRed, type="l") #it is an L
```

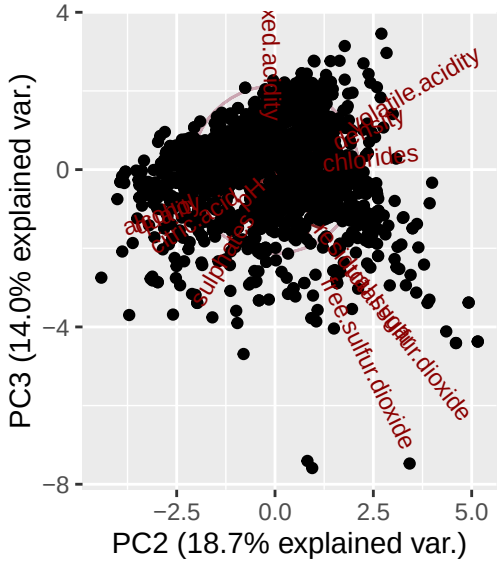


There is no real clear elbow in the scree plot. The closest thing to an elbow occurs at the 6th principal component. Because two of three guidelines reccomend 5 principal components, we will continue our PCA using 5 PCs.

```
g <- ggbiplot(PCWineRed, choices=1:2, obs.scale = 1, var.scale = 1,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```




```
g <- ggbiplot(PCWineRed, choices=2:3, obs.scale = 1, var.scale = 1,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```



White Wine

```
cor(whitewine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000          -0.0227     0.28918
## volatile.acidity            -0.0227           1.0000    -0.14947
## citric.acid                  0.2892          -0.1495     1.00000
## residual.sugar               0.0890           0.0643     0.09421
## chlorides                    0.0231           0.0705     0.11436
## free.sulfur.dioxide         -0.0494          -0.0970     0.09408
## total.sulfur.dioxide         0.0911           0.0893     0.12113
## density                      0.2653           0.0271     0.14950
## pH                          -0.4259          -0.0319    -0.16375
## sulphates                   -0.0171          -0.0357     0.06233
## alcohol                     -0.1209           0.0677    -0.07573
## quality                     -0.1137          -0.1947    -0.00921
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 0.0890    0.0231           -0.049396
## volatile.acidity              0.0643    0.0705           -0.097012
## citric.acid                   0.0942    0.1144            0.094077
## residual.sugar                1.0000    0.0887            0.299098
## chlorides                     0.0887    1.0000            0.101392
## free.sulfur.dioxide           0.2991    0.1014            1.000000
## total.sulfur.dioxide          0.4014    0.1989            0.615501
## density                       0.8390    0.2572            0.294210
## pH                           -0.1941   -0.0904           -0.000618
## sulphates                    -0.0267    0.0168            0.059217
## alcohol                      -0.4506   -0.3602           -0.250104
## quality                      -0.0976   -0.2099            0.008158
##                       total.sulfur.dioxide density        pH sulphates
## fixed.acidity                      0.09107  0.2653 -0.425858   -0.0171
## volatile.acidity                   0.08926  0.0271 -0.031915   -0.0357
## citric.acid                        0.12113  0.1495 -0.163748    0.0623
## residual.sugar                     0.40144  0.8390 -0.194133   -0.0267
## chlorides                          0.19891  0.2572 -0.090439    0.0168
## free.sulfur.dioxide                0.61550  0.2942 -0.000618    0.0592
## total.sulfur.dioxide               1.00000  0.5299  0.002321    0.1346
## density                            0.52988  1.0000 -0.093591    0.0745
## pH                                 0.00232 -0.0936  1.000000    0.1560
## sulphates                          0.13456  0.0745  0.155951    1.0000
## alcohol                           -0.44889 -0.7801  0.121432   -0.0174
## quality                           -0.17474 -0.3071  0.099427    0.0537
##                       alcohol  quality
## fixed.acidity         -0.1209 -0.11366
## volatile.acidity       0.0677 -0.19472
## citric.acid           -0.0757 -0.00921
## residual.sugar        -0.4506 -0.09758
## chlorides             -0.3602 -0.20993
## free.sulfur.dioxide   -0.2501  0.00816
## total.sulfur.dioxide  -0.4489 -0.17474
## density               -0.7801 -0.30712
## pH                     0.1214  0.09943
```

```
## sulphates            -0.0174  0.05368
## alcohol               1.0000  0.43557
## quality               0.4356  1.00000
```

There are a number of moderately high correlations in the white wine data set (above .3). This means that there is some sort of underlying structure in the data and PCA could work on this data set.

```
PCWineWhite <- prcomp(whitewine[,-13], scale=TRUE)
PCWineWhite$rotation
```

```
##                              PC1     PC2     PC3     PC4     PC5      PC6
## fixed.acidity            -0.1569  0.5607 -0.2074  0.0337 -0.2441  0.10586
## volatile.acidity         -0.0243  0.0161  0.5249 -0.1312 -0.7030 -0.12370
## citric.acid              -0.1329  0.2894 -0.4464  0.3295 -0.0651 -0.13196
## residual.sugar           -0.4061 -0.0388 -0.0338 -0.4162  0.0161  0.28992
## chlorides                -0.2175  0.0369  0.2147  0.5096  0.1783 -0.40932
## free.sulfur.dioxide      -0.2747 -0.3455 -0.3130 -0.1489 -0.1112 -0.48809
## total.sulfur.dioxide     -0.3904 -0.2723 -0.1248 -0.0216 -0.2714 -0.27249
## density                  -0.5013 -0.0177  0.0320 -0.1039  0.0783  0.32601
## pH                        0.1300 -0.5671  0.0685  0.2041  0.1127  0.19269
## sulphates                -0.0336 -0.2483 -0.2270  0.5192 -0.4562  0.47981
## alcohol                   0.4428  0.0170 -0.1589 -0.1344 -0.3086 -0.13544
## quality                   0.2271 -0.1460 -0.4888 -0.2782 -0.0411 -0.00552
##                              PC7     PC8     PC9    PC10    PC11     PC12
## fixed.acidity             0.2236  0.1304 -0.6315  0.2009 -0.1041  0.17079
## volatile.acidity         -0.2236 -0.2296 -0.0316 -0.1418 -0.2700  0.01338
## citric.acid              -0.1204 -0.6914  0.2495 -0.1063 -0.0540  0.00965
## residual.sugar           -0.3386 -0.1133  0.1773  0.3743  0.1799  0.49357
## chlorides                -0.5523  0.2114 -0.1792  0.2355  0.0911  0.02517
## free.sulfur.dioxide       0.2241  0.1288  0.1018  0.3273 -0.4992 -0.02948
## total.sulfur.dioxide      0.2038  0.0129 -0.1780 -0.3474  0.6436  0.03506
## density                  -0.1231 -0.0867 -0.1254  0.0435 -0.0669 -0.76118
## pH                        0.0770 -0.4780 -0.5203  0.1838 -0.0791  0.14184
## sulphates                -0.0446  0.3364  0.2366  0.0552 -0.0410  0.04279
## alcohol                  -0.0980 -0.0890  0.0128  0.5753  0.4190 -0.35016
## quality                  -0.5843  0.1444 -0.2997 -0.3677 -0.1462 -0.01607
```

```
summary(PCWineWhite)
```

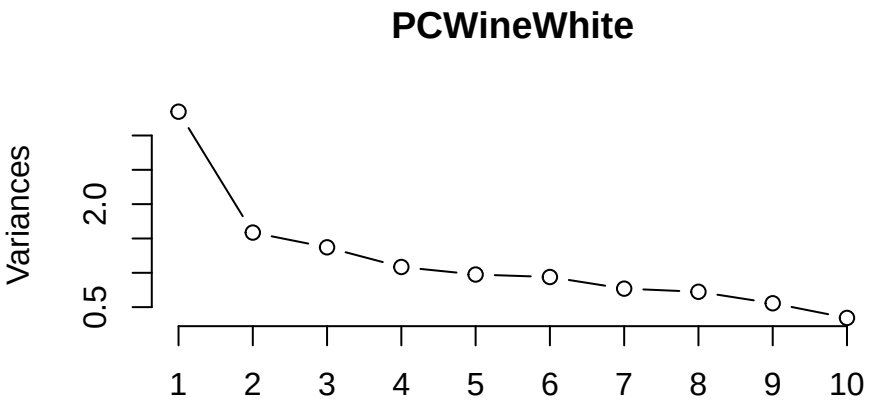
```
## Importance of components:
##                           PC1   PC2   PC3    PC4    PC5    PC6    PC7
## Standard deviation     1.829 1.259 1.171 1.0416 0.9876 0.9689 0.8771
## Proportion of Variance 0.279 0.132 0.114 0.0904 0.0813 0.0782 0.0641
## Cumulative Proportion  0.279 0.411 0.525 0.6157 0.6970 0.7752 0.8393
##                           PC8    PC9   PC10   PC11    PC12
## Standard deviation     0.8508 0.7460 0.5856 0.5330 0.14307
## Proportion of Variance 0.0603 0.0464 0.0286 0.0237 0.00171
## Cumulative Proportion  0.8997 0.9460 0.9746 0.9983 1.00000
```

We run our PCA on correlations over covariances because our variables are on different scales and we do not want to weight variables with higher covariances differently. Now we would like to choose how many principal components to keep. Looking at the variance explained explained by each principal component, we can see

we need 6 principal components to explain 70% of the variation in the data. According to the eigenvalue >0.7 rule, we need 8 principal components as well.

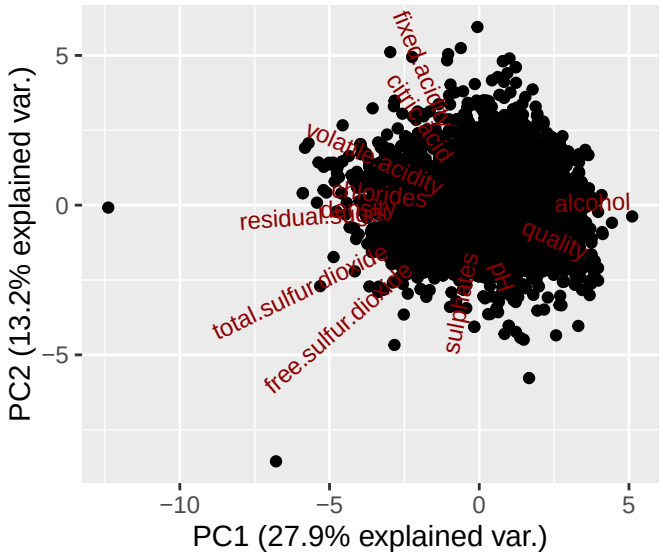
Looking at the loadings for the principal components, it looks like the first principal component represents the negative relationship between residual sugar, total sulfur dioxide, and density and alcohol. The second PC mainly represents the negative relationship between acidity and pH. The third PC captures the negative relationship between volatile acidity and citric acid and quality. The fourth PC mainly represents the amount of chlorides and sulphates. The 5th PC represents the volatile acidity and sulphate levels. The 6th PC represents cloride and free sulfur dioxide levels and the 7th chlorides and quality. Finally, the 8th PC represents citric acid and pH levels.

```
screeplot(PCWineWhite, type="l") #it is an L
```

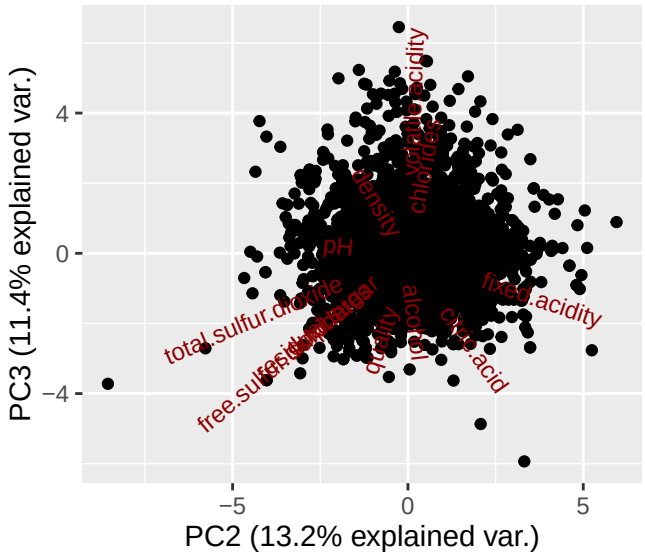


There is no real clear elbow in the scree plot. The closest thing to an elbow occurs at the 2nd principal component. We will use 6 PCs because our overall goal is dimension reduction. We will not follow the guideline of 2 from the scree plot since that explains just 41% of the variance.

```
g <- ggbiplot(PCWineWhite, choices=1:2, obs.scale = 1, var.scale = 1,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```



```
g <- ggbiplot(PCWineWhite, choices=2:3, obs.scale = 1, var.scale = 1,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```



Combined

```
cor(wine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000           0.2190      0.3244
## volatile.acidity             0.2190           1.0000     -0.3780
## citric.acid                  0.3244          -0.3780      1.0000
## residual.sugar              -0.1120          -0.1960      0.1425
## chlorides                    0.2982           0.3771      0.0390
## free.sulfur.dioxide         -0.2827          -0.3526      0.1331
## total.sulfur.dioxide        -0.3291          -0.4145      0.1952
## density                      0.4589           0.2713      0.0962
## pH                          -0.2527           0.2615     -0.3298
## sulphates                    0.2996           0.2260      0.0562
## alcohol                     -0.0955          -0.0376     -0.0105
## quality                     -0.0767          -0.2657      0.0855
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 -0.112    0.2982             -0.2827
## volatile.acidity              -0.196    0.3771             -0.3526
## citric.acid                    0.142    0.0390              0.1331
## residual.sugar                 1.000   -0.1289              0.4029
## chlorides                     -0.129    1.0000             -0.1950
## free.sulfur.dioxide            0.403   -0.1950              1.0000
```

```
## total.sulfur.dioxide          0.495   -0.2796              0.7209
## density                       0.553    0.3626              0.0257
## pH                           -0.267    0.0447             -0.1459
## sulphates                    -0.186    0.3956             -0.1885
## alcohol                      -0.359   -0.2569             -0.1798
## quality                      -0.037   -0.2007              0.0555
##                       total.sulfur.dioxide density      pH sulphates
## fixed.acidity                      -0.3291  0.4589 -0.2527   0.29957
## volatile.acidity                   -0.4145  0.2713  0.2615   0.22598
## citric.acid                         0.1952  0.0962 -0.3298   0.05620
## residual.sugar                      0.4955  0.5525 -0.2673  -0.18593
## chlorides                          -0.2796  0.3626  0.0447   0.39559
## free.sulfur.dioxide                 0.7209  0.0257 -0.1459  -0.18846
## total.sulfur.dioxide                1.0000  0.0324 -0.2384  -0.27573
## density                             0.0324  1.0000  0.0117   0.25948
## pH                                 -0.2384  0.0117  1.0000   0.19212
## sulphates                          -0.2757  0.2595  0.1921   1.00000
## alcohol                            -0.2657 -0.6867  0.1212  -0.00303
## quality                            -0.0414 -0.3059  0.0195   0.03849
##                        alcohol quality
## fixed.acidity        -0.09545 -0.0767
## volatile.acidity     -0.03764 -0.2657
## citric.acid          -0.01049  0.0855
## residual.sugar       -0.35941 -0.0370
## chlorides            -0.25692 -0.2007
## free.sulfur.dioxide  -0.17984  0.0555
## total.sulfur.dioxide -0.26574 -0.0414
## density              -0.68675 -0.3059
## pH                    0.12125  0.0195
## sulphates            -0.00303  0.0385
## alcohol               1.00000  0.4443
## quality               0.44432  1.0000
```

Looking at the correlations between variables in our data, it looks like there are a number of correlations above .3, indicating that there is some sort of underlying structure in the data and fulfills the assumptions required for principal components analysis.

```
PCWine <- prcomp(wine[,-13],scale=TRUE)
PCWine$rotation
```

```
##                          PC1    PC2     PC3     PC4      PC5     PC6
## fixed.acidity        -0.2569  0.262 -0.4675  0.1440 -0.16536  0.0300
## volatile.acidity     -0.3949  0.105  0.2797  0.0801 -0.14777 -0.3827
## citric.acid           0.1465  0.144 -0.5881 -0.0555  0.23462  0.3622
## residual.sugar        0.3189  0.343  0.0755 -0.1125 -0.50792 -0.0633
## chlorides            -0.3134  0.270 -0.0468 -0.1653  0.39390 -0.4254
## free.sulfur.dioxide   0.4227  0.111  0.0990 -0.3033  0.24845 -0.2832
## total.sulfur.dioxide  0.4744  0.144  0.1013 -0.1322  0.22397 -0.1068
## density              -0.0924  0.555  0.0516 -0.1506 -0.33036  0.1546
## pH                   -0.2081 -0.153  0.4068 -0.4715  0.00146  0.5609
## sulphates            -0.2999  0.120 -0.1687 -0.5880  0.19325 -0.0201
## alcohol              -0.0589 -0.493 -0.2129 -0.0800 -0.11602 -0.1695
## quality               0.0875 -0.297 -0.2958 -0.4724 -0.45913 -0.2779
```

```
##                              PC7      PC8      PC9    PC10     PC11      PC12
## fixed.acidity            -0.3934  0.00116  0.42417 -0.2724 -0.27693 -0.335093
## volatile.acidity         -0.4451  0.31008 -0.12323  0.4939  0.14080 -0.082421
## citric.acid              -0.0477  0.44496 -0.24623  0.3304  0.22928  0.001347
## residual.sugar            0.0958  0.08194 -0.48802 -0.2072  0.00514 -0.451215
## chlorides                 0.4733  0.37553 -0.04405 -0.2389 -0.19340 -0.043278
## free.sulfur.dioxide      -0.3627  0.12010  0.30140 -0.3034  0.48616 -0.000905
## total.sulfur.dioxide     -0.2348  0.01128  0.00181  0.2948 -0.72016  0.064063
## density                  -0.0133  0.04294  0.07108 -0.0768 -0.00332  0.715667
## pH                       -0.0793  0.36228  0.13666 -0.1124 -0.13908 -0.206763
## sulphates                -0.1702 -0.59222 -0.29740  0.0855  0.04722 -0.078200
## alcohol                  -0.3389  0.22604 -0.41706 -0.4161 -0.19129  0.332012
## quality                   0.2732  0.09305  0.35665  0.3078 -0.01808  0.008288
```

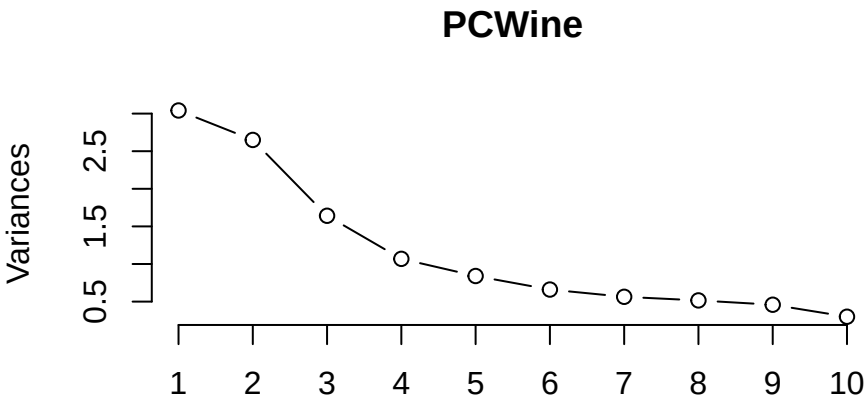
We run our PCA on correlations over covariances because our variables are on different scales and we do not want to weight variables with higher covariances differently.

```
summary(PCWine) #prop variance explained by each component
```

```
## Importance of components:
##                             PC1   PC2   PC3    PC4   PC5   PC6   PC7   PC8
## Standard deviation     1.744 1.628 1.281 1.0337 0.917 0.813 0.751 0.718
## Proportion of Variance 0.253 0.221 0.137 0.0891 0.070 0.055 0.047 0.043
## Cumulative Proportion  0.253 0.474 0.611 0.7001 0.770 0.825 0.872 0.915
##                              PC9   PC10  PC11    PC12
## Standard deviation     0.6770 0.5468 0.477 0.18107
## Proportion of Variance 0.0382 0.0249 0.019 0.00273
## Cumulative Proportion  0.9534 0.9783 0.997 1.00000
```

Now we would like to choose how many principal components to keep. Looking at the variance explained explained by each principal component, we can see we need 4 principal components to explain 70% of the variation in the data. According to the eigenvalue >0.7 rule, we need 5 principal components.

```
screeplot(PCWine, type="l") #it is an L
```



Looking at the scree plot, there is a slight elbow at 4 principal components.

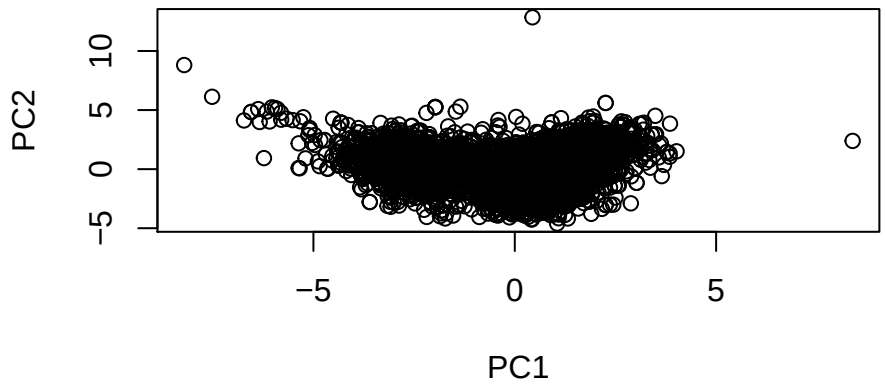
Since our overall goal is to reduce the number of variables and 2 of the 3 guidelines recommend 4 principal components, we will use 4 principal components.

```r
cor(wine[,-13],PCWine$x) #loadings
```

```
##                         PC1    PC2     PC3     PC4      PC5     PC6
## fixed.acidity        -0.448  0.426 -0.5989  0.1488 -0.15160  0.0244
## volatile.acidity     -0.689  0.171  0.3583  0.0828 -0.13548 -0.3110
## citric.acid           0.255  0.235 -0.7535 -0.0574  0.21510  0.2944
## residual.sugar        0.556  0.558  0.0967 -0.1163 -0.46566 -0.0515
## chlorides            -0.547  0.439 -0.0599 -0.1709  0.36112 -0.3457
## free.sulfur.dioxide   0.737  0.181  0.1268 -0.3135  0.22778 -0.2301
## total.sulfur.dioxide  0.827  0.234  0.1298 -0.1367  0.20533 -0.0868
## density              -0.161  0.903  0.0661 -0.1557 -0.30287  0.1256
## pH                   -0.363 -0.249  0.5212 -0.4874  0.00134  0.4558
## sulphates            -0.523  0.195 -0.2161 -0.6079  0.17717 -0.0164
## alcohol              -0.103 -0.802 -0.2728 -0.0827 -0.10637 -0.1377
## quality               0.153 -0.483 -0.3790 -0.4884 -0.42092 -0.2258
##                          PC7      PC8      PC9    PC10     PC11      PC12
## fixed.acidity        -0.29542  0.00083  0.28718 -0.1490 -0.13211 -0.060674
## volatile.acidity     -0.33423  0.22273 -0.08343  0.2701  0.06717 -0.014924
## citric.acid          -0.03582  0.31963 -0.16671  0.1806  0.10938  0.000244
## residual.sugar        0.07191  0.05886 -0.33041 -0.1133  0.00245 -0.081700
## chlorides             0.35539  0.26975 -0.02982 -0.1306 -0.09226 -0.007836
## free.sulfur.dioxide  -0.27236  0.08627  0.20406 -0.1659  0.23193 -0.000164
## total.sulfur.dioxide -0.17632  0.00810  0.00123  0.1612 -0.34356  0.011600
## density              -0.00998  0.03085  0.04812 -0.0420 -0.00159  0.129583
## pH                   -0.05956  0.26023  0.09252 -0.0615 -0.06635 -0.037438
## sulphates            -0.12783 -0.42540 -0.20135  0.0467  0.02253 -0.014159
## alcohol              -0.25448  0.16237 -0.28236 -0.2275 -0.09126  0.060116
## quality               0.20512  0.06684  0.24146  0.1683 -0.00863  0.001501
```

Looking at the loadings, the first PC represents a negative relationship between the variables fixed and volatile acidity, chlorides, pH, and sulphates and the variables residual sugar and sulfur dioxide. The second PC mainly capures the negative relationship between the varibles density, fixed acidity, residual sugar, and chlorides and the variables alcohol and quality. The third PC represents the negative relationship between the variables citric acid and fixed acidity and the variable pH. The 4th PC represents the positive relationship between pH, sulphates, and quality.

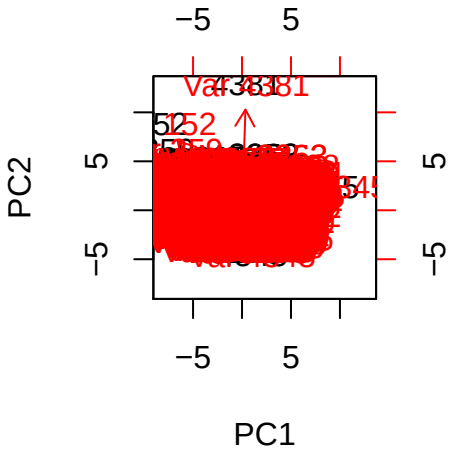
```r
plot(PCWine$x)
```



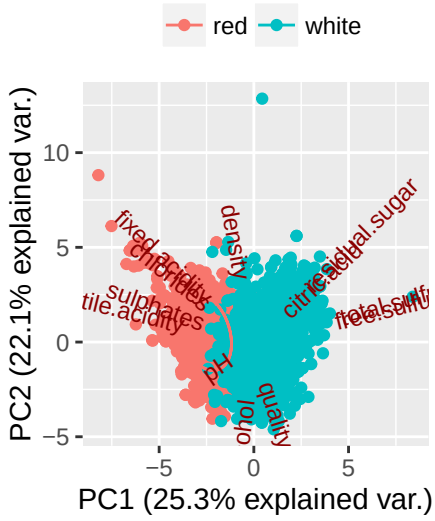
```
biplot(PCWine$x[,1:2],PCWine$x[,1:2])
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length
## = arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length
## = arrow.len): zero-length arrow is of indeterminate angle and so skipped
```




```
g <- ggbiplot(PCWine, choices=1:2, obs.scale = 1, var.scale = 1, groups = wine$type,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```

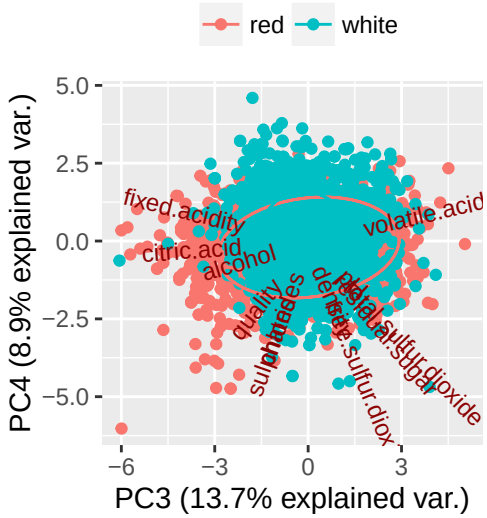


Looking at the biplot for all the points, we can see a clear separation of groups when looking at the first 2 principal components. The red wines have a lower value for PC1, which would indicate higher levels of fixed acidity, volatile acidity, chlorides, and sulphates and lower levels of residual sugar and free and total sulfur dioxide.

```
g <- ggbiplot(PCWine, choices=2:3, obs.scale = 1, var.scale = 1, groups = wine$type,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```

Looking at the biplot for the second and third principal components, we can see that there is much less clear separation. It does look like red wines have a higher spread in PC3, which indicates a greater variance in pH levels. There doesn't look to be much difference in red and white wines for PC2.

```
g <- ggbiplot(PCWine, choices=3:4, obs.scale = 1, var.scale = 1, groups = wine$type,
              ellipse = TRUE,
               circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```



Looking at PC3 and PC4, we see again that there is a larger spread in PC3 for red wines. There does not look to be any significant difference between red and white wines in PC4.

<–! CLASSIFICATION –>

After creating a random forest, we can see that it does a very good job of classifying wine type. The AER is just .03% and the estimated TER is only .48%. This indicates that our random forest does a very good job at classifying wine type.

It seems that total.sulfur.dioxide and chlorides are very important in determining the difference between red and white wines. Volatile acidity is also important in classifying wines to a lesser degree.