Problem 9.5

Chen Bo Calvin Zhang

01/01/2021

Firstly, let us load the data set and analyse it.

```
library("whitening")
```

```
## Loading required package: corpcor
```

```
data(forina1986)
wine.attrib = forina1986$attrib
wine.type = forina1986$type
print(dim(wine.attrib))
```

```
## [1] 178  27
```

```
print(levels(wine.type))
```

```
## [1] "Barolo"     "Grignolino" "Barbera"
```

```
print(table(wine.type))
```

```
## wine.type
##     Barolo Grignolino    Barbera
##         59         71         48
```

Here are two helper functions we will need for this problem.

```
# function to compute the feature ranking
# diagonal = TRUE: use t-scores for ranking
# diagonal = FALSE: ZCA-cor whiten the data, then use t-scores
library("sda")
```

```
## Loading required package: entropy
```

```
## Loading required package: fdrtool
```

```
featureRanking = function(train.x, train.y, diagonal=TRUE)
{
  return( sda.ranking(train.x, train.y, diagonal=diagonal,
                      verbose=FALSE, fdr=FALSE, lambda=0, lambda.var=0) )
}

library("crossval")
# predictor function for LDA (using sda)
predfun.lda = function(train.x, train.y, test.x, test.y)
{
  # fit sda with zero shrinkage and full covariance (=classic LDA)
  sda.fit = sda(train.x, train.y, diagonal=FALSE, lambda=0, lambda.var=0, verbose=FALSE)
  ynew = predict(sda.fit, test.x, verbose=FALSE)$class
  # compute accuracy
  out = mean( ynew == test.y)
}
```

We want to have a ranking of the predcitors based on the t-scores and the decorrelates t-scores.

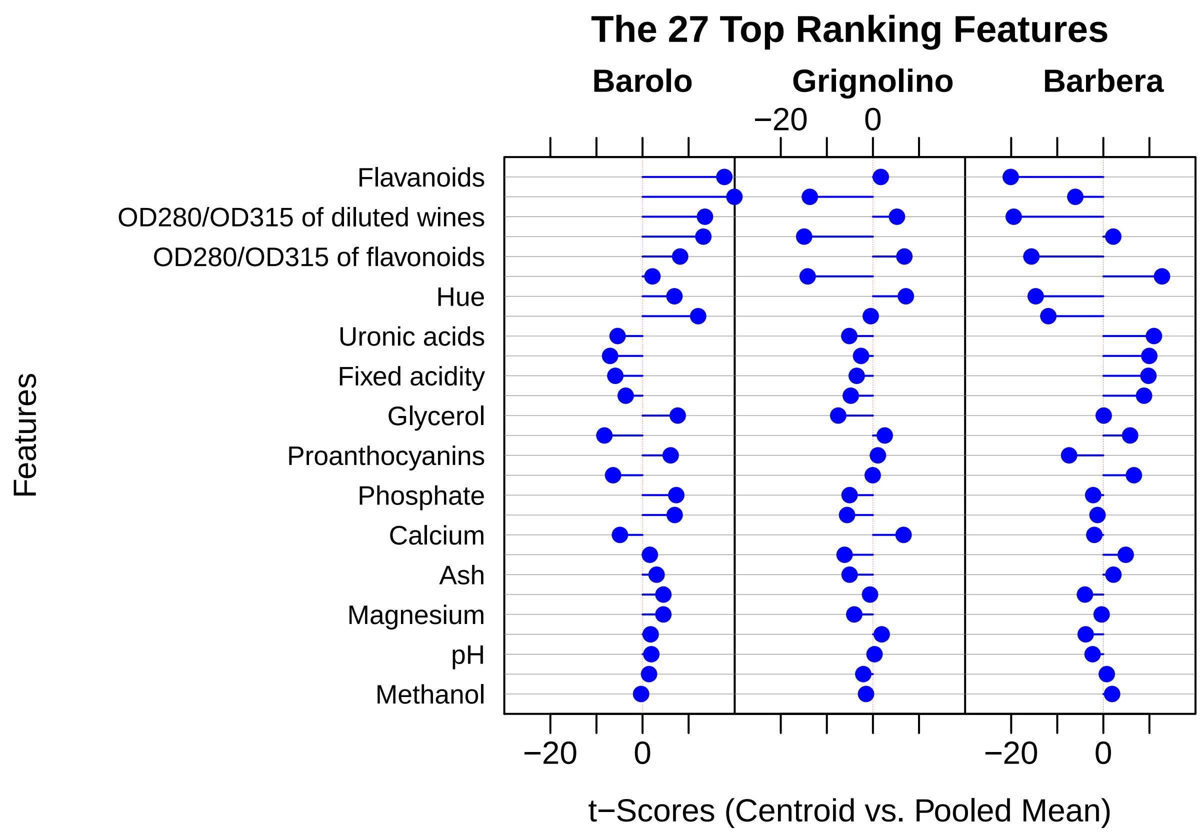
```
ranking = featureRanking(wine.attrib, wine.type, diagonal=TRUE)
ordering = ranking[, "idx"]

print(ranking)
```

```
##                              idx      score  t.Barolo t.Grignolino    t.Barbera
## Flavanoids                    16 496.049773 17.763256   1.70891910 -20.09700604
## Proline                       26 409.725828 19.941877 -13.69483060  -6.08360654
## OD280/OD315 of diluted wines  21 405.223816 13.544644   5.21348593 -19.45118841
## Alcohol                        1 260.130345 13.202908 -14.92837519   2.15837217
## OD280/OD315 of flavonoids     22 244.801398  8.159856   6.82518254 -15.60666178
## Color intensity               19 243.378065  2.154400 -14.16590217  12.73192500
## Hue                           20 215.460680  6.926708   7.14777656 -14.67717272
## Total phenols                 15 197.446430 12.073472  -0.47375500 -11.93382004
## Uronic acids                   6 120.857666 -5.409200  -5.13118529  10.98607231
## Tartaric acid                  4 107.190286 -7.042827  -2.58077884   9.97677008
## Fixed acidity                  3  98.284167 -5.900362  -3.51616130   9.78740546
## Malic acid                     5  78.289417 -3.651041  -4.80961551   8.83605567
## Glycerol                      23  74.709047  7.636684  -7.53264589   0.08530927
## Alcalinity of ash              9  73.603436 -8.276934   2.57224647   5.80923206
## Proanthocyanins               18  64.360601  6.102795   1.08474563  -7.42977547
## Nonflavanoid phenols          17  58.241044 -6.396111  -0.04749024   6.63713106
## Phosphate                     13  55.423311  7.328132  -5.06337808  -2.20298333
## Sugar-free extract             2  53.450943  6.974636  -5.61286082  -1.25901577
## Calcium                       11  46.602930 -4.903978   6.64580636  -1.96360953
## 2-3-butanediol                24  42.123250  1.588802  -6.15368623   4.85834914
## Ash                            8  25.773670  3.055571  -5.06247517   2.19614415
## Total nitrogen                25  25.114413  4.525427  -0.63333946  -3.99206987
## Magnesium                     12  24.104206  4.518372  -4.05728341  -0.37119026
## Chloride                      14  14.679526  1.763293   1.90967844  -3.83138615
## pH                             7   6.329852  1.908744   0.34801429  -2.33300508
## Potassium                     10   4.498390  1.413799  -2.09423359   0.75424914
## Methanol                      27   4.031202 -0.314787  -1.49856389   1.90575937
## attr(,"class")
```

```
## [1] "sda.ranking"
## attr(,"diagonal")
## [1] TRUE
## attr(,"cl.count")
## [1] 3
```

```r
plot(ranking)
```




```r
print(ordering)
```

```
##                    Flavanoids                       Proline
##                            16                            26
## OD280/OD315 of diluted wines                       Alcohol
##                            21                             1
##     OD280/OD315 of flavonoids               Color intensity
##                            22                            19
##                           Hue                Total phenols
##                            20                            15
##                  Uronic acids                 Tartaric acid
##                             6                             4
##                 Fixed acidity                   Malic acid
##                             3                             5
##                      Glycerol             Alcalinity of ash
##                            23                             9
##               Proanthocyanins          Nonflavanoid phenols
```

```
##                        18                       17
##                 Phosphate        Sugar-free extract
##                        13                        2
##                   Calcium          2-3-butanediol
##                        11                       24
##                       Ash            Total nitrogen
##                         8                       25
##                 Magnesium                 Chloride
##                        12                       14
##                        pH                Potassium
##                         7                       10
##                  Methanol
##                        27
```

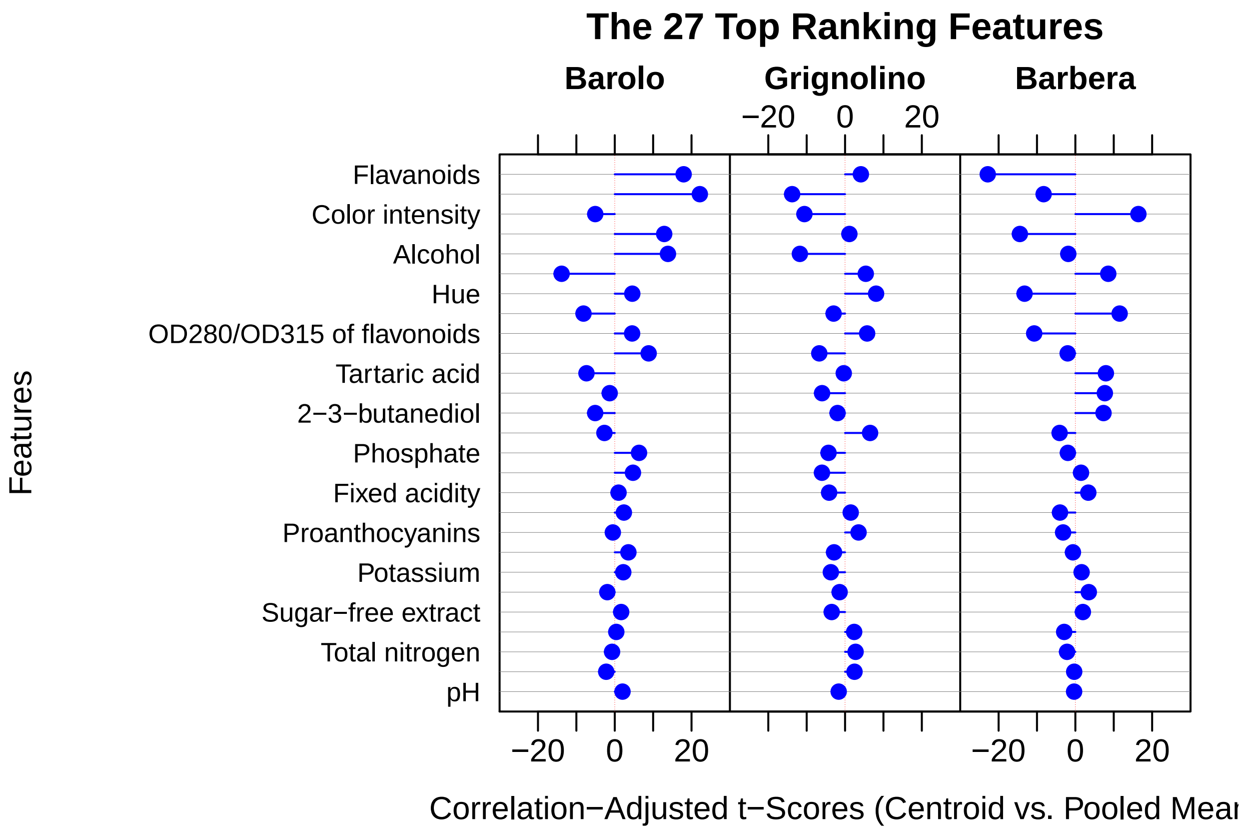
```r
ranking.decor = featureRanking(wine.attrib, wine.type, diagonal=FALSE)
ordering.decor = ranking.decor[, "idx"]

print(ranking.decor)
```

```
##                             idx      score  cat.Barolo cat.Grignolino
## Flavanoids                   16 591.359348  17.9418978      4.1193824
## Proline                      26 496.025763  22.1668240    -13.7922340
## Color intensity              19 277.144648  -5.0867058    -10.5906489
## OD280/OD315 of diluted wines 21 258.340397  12.8843495      1.1248338
## Alcohol                       1 218.037792  13.8606497    -11.7727286
## Alcalinity of ash             9 197.936837 -13.8489691      5.3966212
## Hue                          20 178.245837   4.5702171      8.0867041
## Uronic acids                  6 143.251437  -8.1422345     -2.9828616
## OD280/OD315 of flavonoids    22 115.322301   4.5255555      5.7499418
## Glycerol                     23  83.338415   8.8307919     -6.7184241
## Tartaric acid                 4  80.666244  -7.3666693     -0.3434994
## Malic acid                    5  65.163841  -1.3165301     -5.9906114
## 2-3-butanediol               24  57.557784  -5.1073112     -1.9614847
## Calcium                      11  43.493599  -2.6955113      6.5211998
## Phosphate                    13  41.131135   6.3243292     -4.3129666
## Ash                           8  39.419417   4.7436774     -6.0239484
## Fixed acidity                 3  19.502072   0.9875740     -4.1555597
## Total phenols                15  16.402248   2.3783706      1.4734698
## Proanthocyanins              18  15.129412  -0.4968765      3.5147357
## Magnesium                    12  13.905612   3.5537829     -2.8706534
## Potassium                    10  13.805547   2.2213433     -3.7064698
## Methanol                     27  12.369497  -1.9407558     -1.4196153
## Sugar-free extract            2  12.207048   1.6716833     -3.4848419
## Chloride                     14   9.639717   0.4033013      2.3726176
## Total nitrogen               25   8.372663  -0.6987125      2.7403376
## Nonflavanoid phenols         17   7.183395  -2.2231467      2.4608190
## pH                            7   4.542749   2.0210763     -1.6614184
##                             cat.Barbera
## Flavanoids                  -22.8249712
## Proline                      -8.2721603
## Color intensity              16.4158450
## OD280/OD315 of diluted wines -14.4560421
## Alcohol                      -1.8494550
## Alcalinity of ash             8.5667393
```

```
## Hue                           -13.2412897
## Uronic acids                   11.5333469
## OD280/OD315 of flavonoids    -10.7290872
## Glycerol                       -2.0037642
## Tartaric acid                   7.9490637
## Malic acid                      7.6782863
## 2-3-butanediol                  7.3298865
## Calcium                        -4.1064813
## Phosphate                      -1.9612016
## Ash                             1.4723887
## Fixed acidity                   3.3687088
## Total phenols                  -4.0044488
## Proanthocyanins                -3.1977290
## Magnesium                      -0.6301757
## Potassium                       1.6241488
## Methanol                        3.4969352
## Sugar-free extract              1.9563090
## Chloride                       -2.9193876
## Total nitrogen                 -2.1725730
## Nonflavanoid phenols           -0.3076376
## pH                             -0.3279446
## attr(,"class")
## [1] "sda.ranking"
## attr(,"diagonal")
## [1] FALSE
## attr(,"cl.count")
## [1] 3
```

```
plot(ranking.decor)
```




```r
print(ordering.decor)
```

```
##                 Flavanoids                        Proline
##                         16                             26
##            Color intensity OD280/OD315 of diluted wines
##                         19                             21
##                    Alcohol               Alcalinity of ash
##                          1                              9
##                        Hue                   Uronic acids
##                         20                              6
##     OD280/OD315 of flavonoids                    Glycerol
##                         22                             23
##              Tartaric acid                     Malic acid
##                          4                              5
##             2-3-butanediol                        Calcium
##                         24                             11
##                  Phosphate                            Ash
##                         13                              8
##              Fixed acidity                 Total phenols
##                          3                             15
##            Proanthocyanins                      Magnesium
##                         18                             12
##                  Potassium                       Methanol
##                         10                             27
##         Sugar-free extract                       Chloride
```

```
##                        2                          14
##           Total nitrogen         Nonflavanoid phenols
##                       25                          17
##                       pH
##                        7
```

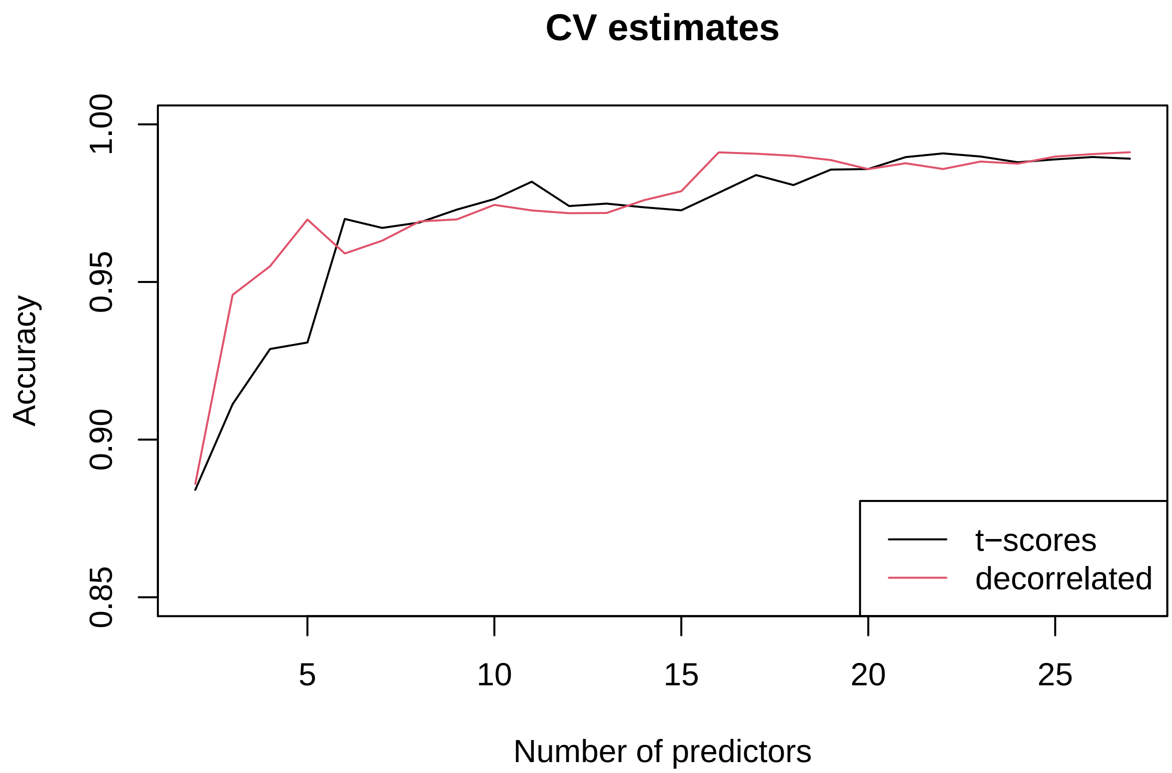
Now we compute the accuraracy using all subsets of 2 to 27 best features in the LDA predictor.

```r
cvmat = matrix(0, 26, 2)
cvmat.decor = matrix(0, 26, 2)

for (i in 2:27)
{
  cv.out = crossval(predfun.lda, wine.attrib[, ordering[1:i]], wine.type,
                    K=5, B=50, verbose=FALSE)
  cvmat[i-1,] = c(cv.out$stat, cv.out$stat.se)

  cv.out = crossval(predfun.lda, wine.attrib[, ordering.decor[1:i]], wine.type,
                    K=5, B=50, verbose=FALSE)
  cvmat.decor[i-1,] = c(cv.out$stat, cv.out$stat.se)
}

plot(2:27, cvmat[, 1], type="l", xlab="Number of predictors",
     ylab="Accuracy", ylim=c(0.85, 1), main="CV estimates")
lines(2:27, cvmat.decor[, 1], col=2)
legend("bottomright", c("t-scores", "decorrelated"), col=1:2, lty=1)
```



We can see that we can obtain good predictions using 5 or 16 predictors using decorrelated t-scores.

```
cv.out = crossval(predfun.lda, wine.attrib[, ordering.decor[1:5]], wine.type,
                  K=5, B=50, verbose=FALSE)
print(c(cv.out$stat, cv.out$stat.se))
```

```
## [1] 0.971243703 0.001627228
```

```
cv.out = crossval(predfun.lda, wine.attrib[, ordering.decor[1:16]], wine.type,
                  K=5, B=50, verbose=FALSE)
print(c(cv.out$stat, cv.out$stat.se))
```

```
## [1] 0.9918045424 0.0008545267
```