Automatical Learning and Data Mining
HW1 Report

Yifan Zhao

February 7, 2018

1 Problem 1

I choose a data set consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians as my classification task. One case is a vector of 11 respective diagnostic features of fetal cardiotocograms. Thus there are 11 descriptors. The meaning of descriptors are as follows:
LB - FHR baseline (beats per minute)
AC - number of accelerations per second
FM - number of fetal movements per second
UC - number of uterine contractions per second
DL - number of light decelerations per second
DS - number of severe decelerations per second
DP - number of prolongued decelerations per second
ASTV - percentage of time with abnormal short term variability
MSTV - mean value of short term variability
ALTV - percentage of time with abnormal long term variability
MLTV - mean value of long term variability
There are 5 classes, each class represents a pattern of FHR(fetal heart rate). Class1 has 963 cases; class2 has 134 cases; class3 has 404 cases; class4 has 359 cases; class5 has 266 cases.
The correct classification of each case was originally obtained by 3 expert obstetricians and it is in the 12th column of the data set(valued as integer 1 to 10).
There is a publication analyzing the classification task associated to this data set:
Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318

2 Problem 2

Although some descriptors seems to discrete(such as AC, the number of accelerations per second. But it is continuous since it is a ratio of the number of accelerations,an integer, and time, a real number), So all 11 descriptors are numeric and continuous.
Their respective max value, min value, range, mean and standard deviation are as follows:

1	1.600000e+02	106.0	54.000000000	1.333039e+02	9.840844e+00
2	1.928375e-02	0.0	0.019283747	3.169682e-03	3.859772e-03
3	4.806338e-01	0.0	0.480633803	9.473980e-03	4.666983e-02
4	1.492537e-02	0.0	0.014925373	4.356799e-03	2.940320e-03
5	1.538462e-02	0.0	0.015384615	1.884572e-03	2.962264e-03
6	1.353180e-03	0.0	0.001353180	3.584711e-06	6.292393e-05
7	5.347594e-03	0.0	0.005347594	1.565723e-04	5.796968e-04
8	8.700000e+01	12.0	75.000000000	4.699012e+01	1.719281e+01
9	7.000000e+00	0.2	6.800000000	1.332785e+00	8.832413e-01
10	9.100000e+01	0.0	91.000000000	9.846660e+00	1.839688e+01
11	5.070000e+01	0.0	50.700000000	8.187629e+00	5.628247e+00

And here is their histogram:

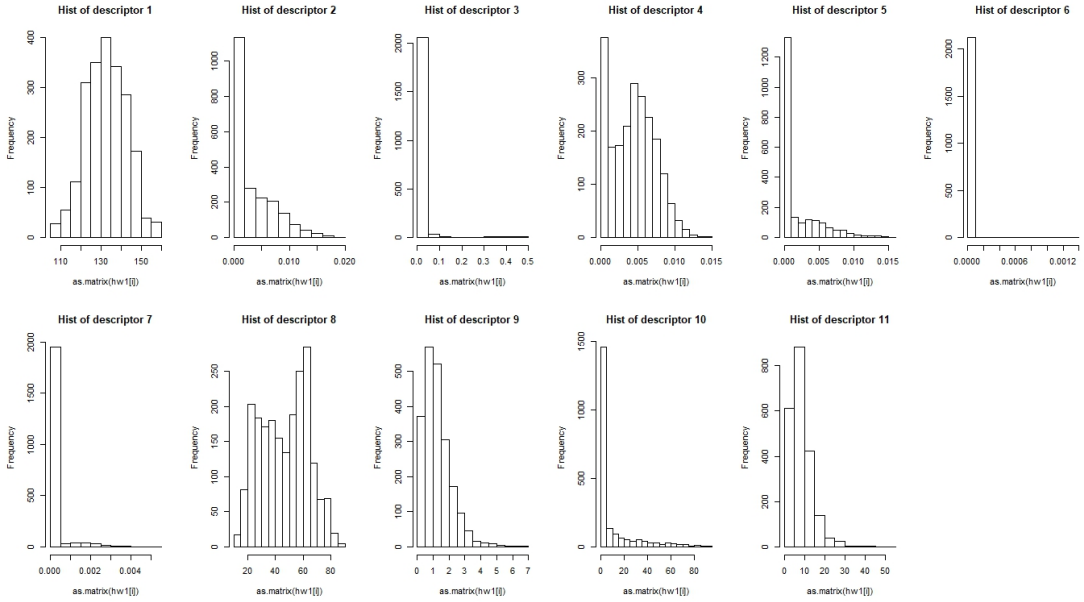

Figure 1: Histogram of descriptors

3 Problem 3

Here are the histograms for each descriptor with 5 classes overlapped:

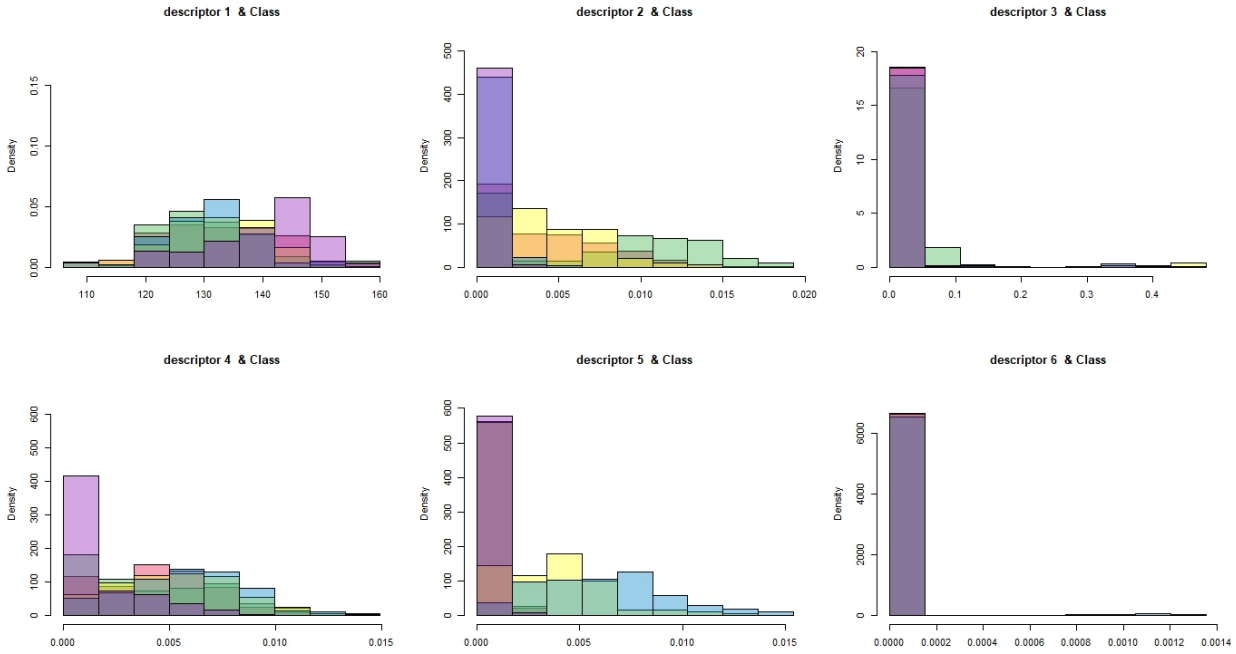

Figure 2: Histogram of classes of descriptors 1-6

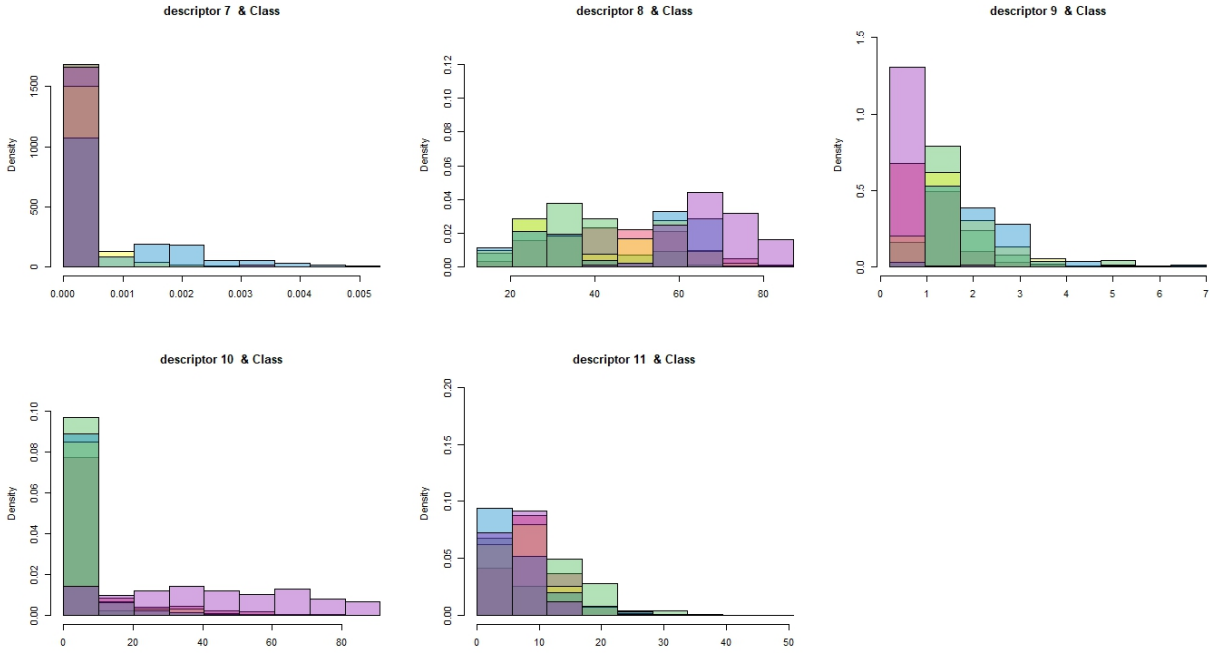

Figure 3: Histogram of classes of descriptors 7-11

According to the histograms, descriptor 2,3,5,8,9 and 10 are relatively more helpful to discriminate between some pairs of classes than the others.
The means and standard deviations are as follows:

1	1.326397e+02	1.319403e+02	1.323366e+02	1.311671e+02	1.407481e+02
2	3.975202e-03	7.074086e-03	4.419545e-03	4.312716e-04	8.410957e-05
3	3.517883e-03	1.627938e-02	1.745995e-02	1.483818e-02	8.239824e-03
4	4.315537e-03	4.146653e-03	5.275595e-03	5.690293e-03	1.416860e-03
5	2.462328e-04	2.131042e-04	3.689374e-03	6.182784e-03	1.157693e-04
6	0.000000e+00	0.000000e+00	2.204138e-06	1.874826e-05	0.000000e+00
7	0.000000e+00	0.000000e+00	1.201562e-04	7.585755e-04	4.511662e-05
8	4.530010e+01	3.500746e+01	4.164356e+01	4.569359e+01	6.901504e+01
9	1.111526e+00	1.637313e+00	1.706436e+00	2.086630e+00	3.954887e-01
10	7.555556e+00	5.223881e-01	3.950495e+00	2.779944e+00	4.133083e+01
11	9.273832e+00	1.018881e+01	7.538861e+00	6.301950e+00	6.777444e+00

Figure 4: The means

1	1.034639e+01	8.816220e+00	8.699593e+00	8.168609e+00	8.827941e+00
2	3.847144e-03	5.929581e-03	3.104287e-03	9.094156e-04	3.913130e-04
3	1.395493e-02	2.676686e-02	7.697444e-02	6.468489e-02	3.802604e-02
4	2.600082e-03	3.374936e-03	2.813684e-03	2.747351e-03	2.105249e-03
5	6.648174e-04	5.862644e-04	2.642867e-03	3.161186e-03	4.067137e-04
6	0.000000e+00	0.000000e+00	4.430264e-05	1.449519e-04	0.000000e+00
7	0.000000e+00	0.000000e+00	3.667894e-04	1.119647e-03	4.354813e-04
8	1.447897e+01	1.092455e+01	1.654100e+01	1.830852e+01	8.043265e+00
9	6.386072e-01	8.219427e-01	8.473038e-01	9.493835e-01	1.932306e-01
10	1.395746e+01	2.286661e+00	9.367730e+00	7.499649e+00	2.481941e+01
11	5.433388e+00	7.871846e+00	5.121002e+00	6.148672e+00	3.320096e+00

Figure 5: The standard deviations

With the t-test and the Kolmogorov-Smirnov test we can compare pairs of histograms and pairs of means. Because there are 11 descriptors, there are 220 tests in total.So here are only the p-values of these tests in descriptor1 given the $\alpha = 0.05$, when $p < \alpha$, two histograms or two means are strongly distinct:

1	1.000000e+00	4.013005e-01	5.792696e-01	7.039397e-03	2.472030e-32
2	4.013005e-01	1.000000e+00	6.513899e-01	3.779398e-01	2.125188e-18
3	5.792696e-01	6.513899e-01	1.000000e+00	5.594646e-02	2.888540e-30
4	7.039397e-03	3.779398e-01	5.594646e-02	1.000000e+00	1.416339e-37
5	2.472030e-32	2.125188e-18	2.888540e-30	1.416339e-37	1.000000e+00

Figure 6: p-value of t-tests

1	1.000000e+00	1.577054e-02	0.0007964438	4.548742e-09	0.000000e+00
2	1.577054e-02	1.000000e+00	0.0880266476	5.561977e-01	2.220446e-16
3	7.964438e-04	8.802665e-02	1.0000000000	2.063542e-03	0.000000e+00
4	4.548742e-09	5.561977e-01	0.0020635423	1.000000e+00	0.000000e+00
5	0.000000e+00	2.220446e-16	0.0000000000	0.000000e+00	1.000000e+00

Figure 7: p-value of K-S tests

According to the results, some pairs are strongly distinct while some are not. Relevant R codes are as follows:

```
class=list(class1,class2,class3,class4,class5)
t=array(0,c(5,5,11))
ks=array(0,c(5,5,11))
mc=array(0,c(11,5))
sdc=array(0,c(11,5))
for (i in 1:11){
```

```
  for (k in 1:5){
    mc[i,k]=mean(class[[k]][[i]])
    sdc[i,k]=sqrt(var(class[[k]][[i]]))
    for (m in 1:5){
      t[k,m,i]=t.test(as.matrix(class[[k]][[i]]),as.matrix(class[[m]][[i]]))$p.val
      ks[k,m,i]=ks.test(as.matrix(class[[k]][[i]]),as.matrix(class[[m]][[i]]))$p.v
      }
  }
}
```

4 Problem 4

The thing that principal component analysis(PCA) does is to transfer the original data to a basis that the features or characteristics of the data be relatively most reflected, and to decrease the redundancy of pairs of descriptors of the data. Follow this idea, we choose the directions that are linearly independent and the data's projections on which(the directions) are variance-maximized. Then this new basis is the eigenvector matrix of the original covariance matrix, and respective eigenvalues are the maximized variances. And if we think that to include relatively small eigenvalues and respective eigenvectors to the new basis is unnecessary, we could calculate the ratio:

$$RAT_j = \frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

where $j \leq n$ and $\lambda_1, \cdots \lambda_n$ are all the eigenvalues rank descendingly.
We then choose the first j eigenvalues and eigenvectors as our new basis according to a given ratio level like 0.85,0.95,etc such that RAT_j be grater than that given ratio level. This is my brief explanation of PCA.
The eigenvalues of this data set is 480.5307 171.3118 83.05912 27.88628 0.5497105 0.002137986 1.309506e-05 7.119464e-06 3.837938e-06 2.382227e-07 3.873138e-09.
Display graphically as follow:

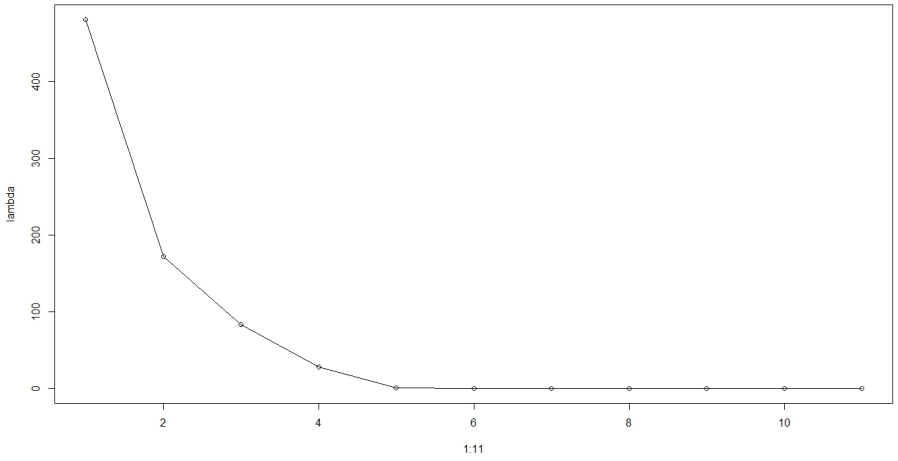

Figure 8: eigenvalues

The samllest j such that $RAT_j \geq 0.95$ is 3, $RAT_j = 0.9627451$. Display graphically as follow:

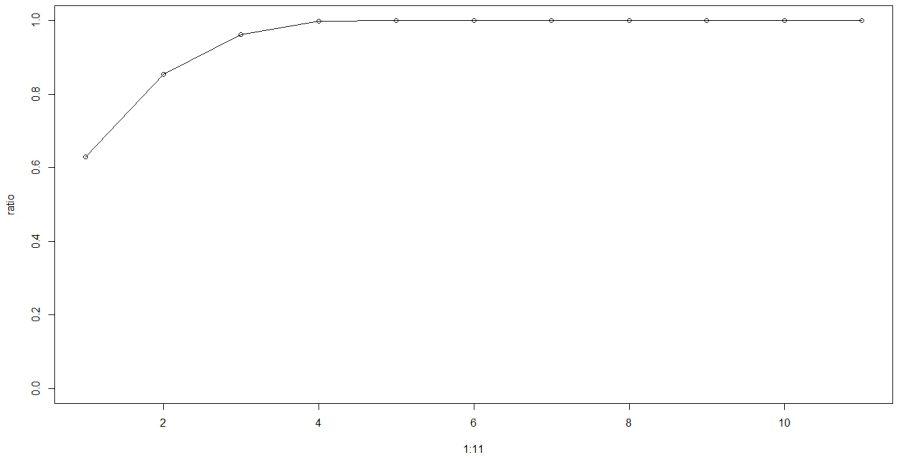

Figure 9: ratio

5 Problem 5

The scatter plot of basis {W1,W2,W3} is as follow:

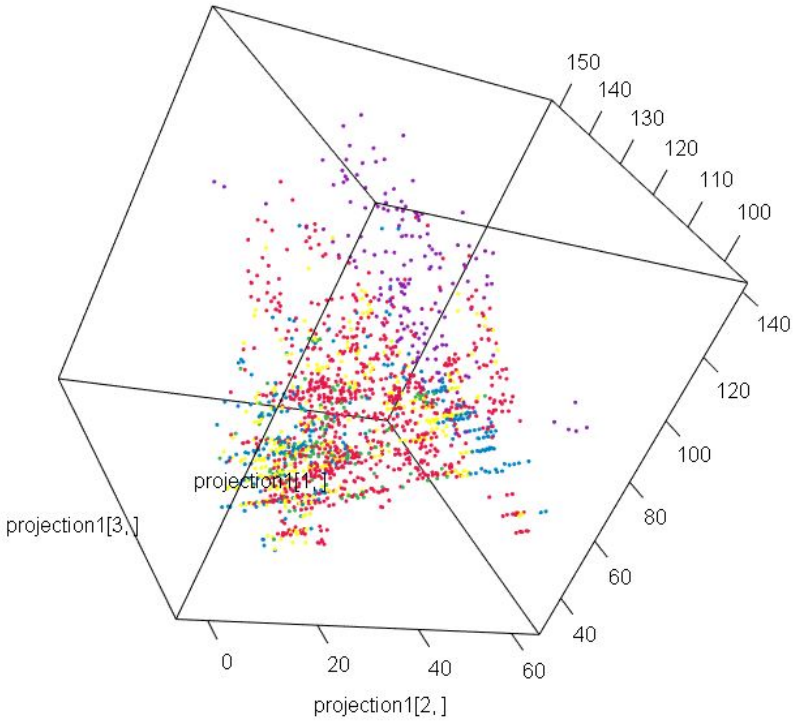

Figure 10: ratio

As we can see from the 3-dimensional scatter, points from different classes are separate mutually, and points from the same class are relatively concentrated. Although there are still some areas of mixed different color points, this is not a PCA could handle, but the classification by deep learning algorithms. The scatter plots of basis {W1,W2} {W1,W3} {W2,W3} are as follows:

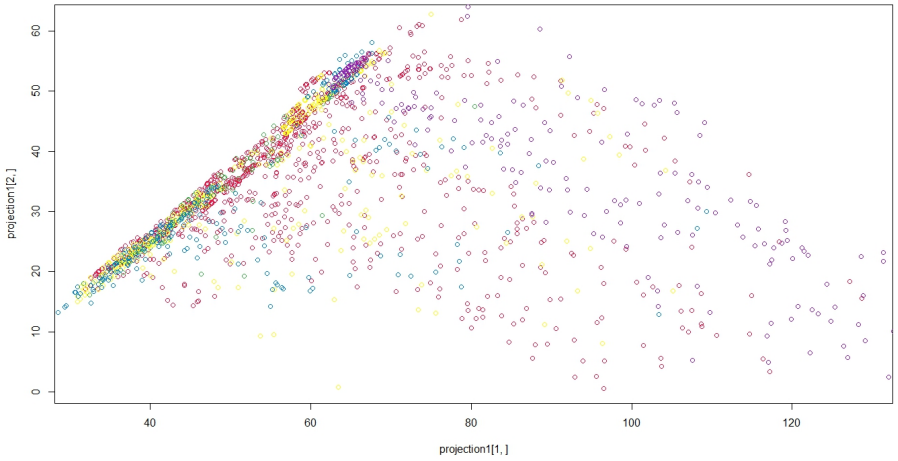

Figure 11: ratio

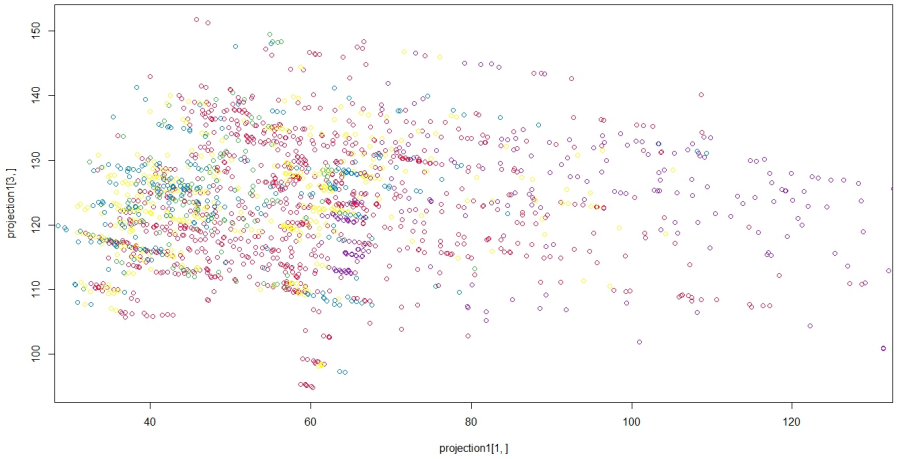

Figure 12: ratio

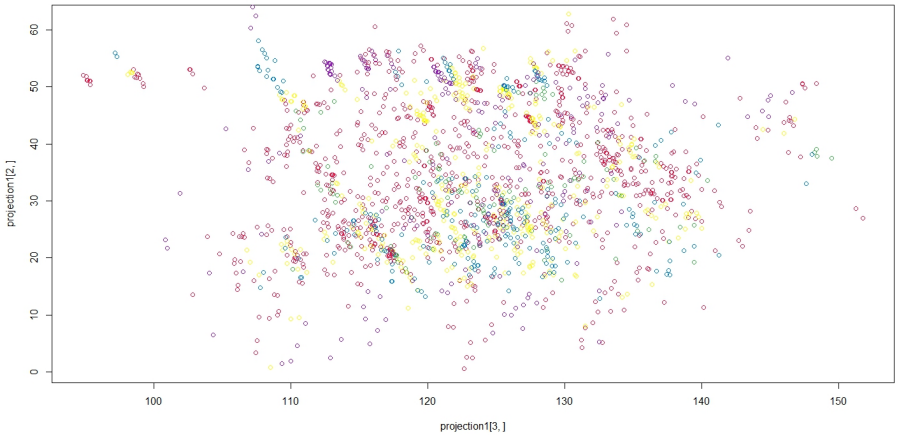

Figure 13: ratio

As we can see from the scatters, the basis {W1,W2} is the best 2d basis that reflects the classification(different colors of points are relatively most separate).