Applied Machine Learning – Exercise 3 (12.05.17)

Ben Wulf, Lee Hong, Amnon Bleich

Task 1.1

For the plot, we removed all genes which has NA values. We selected all genes with a variance bigger than 5 (7) and ploted their density.

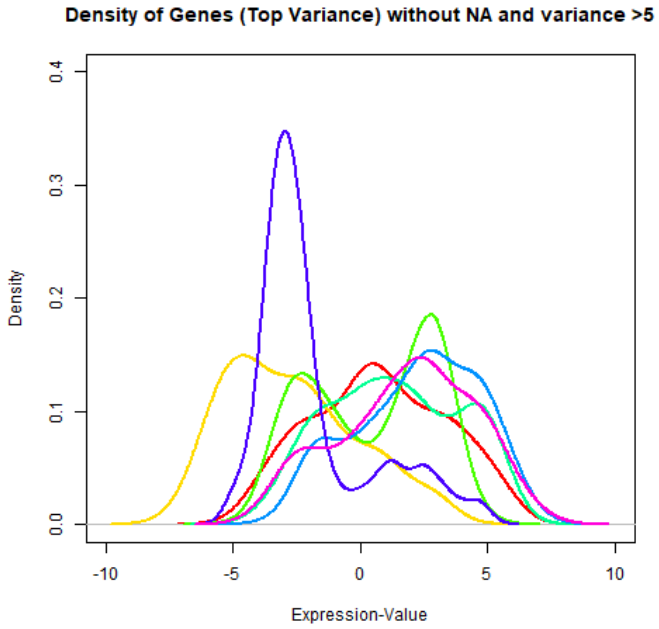
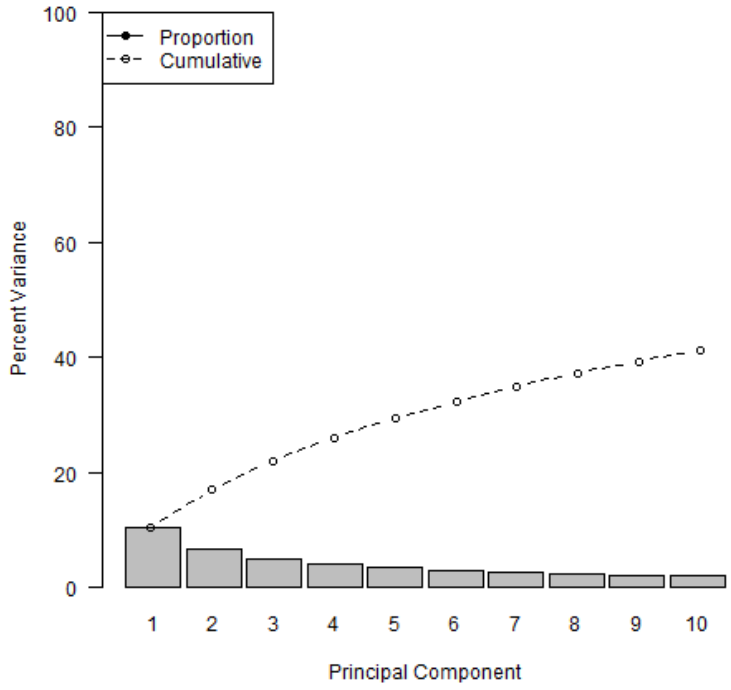


Figure 1

Figure 1 shows the density of the genes with the biggest variance. The distributions are not centered and the variance differs. To get them comparable we must center the distribution to 0 and normalize the standard deviation to 1.

Task 1.2

It seems that there is no M<<p components that explains most of the variance. However, the first component explains around 10% of the variance, the second around 5%, and the rest around 2% each. The 10 most important components declare just around 40% of the Variances.



Task 1.3

contend differs by 0.1% but the N-content is 10times smaller than without trimming.

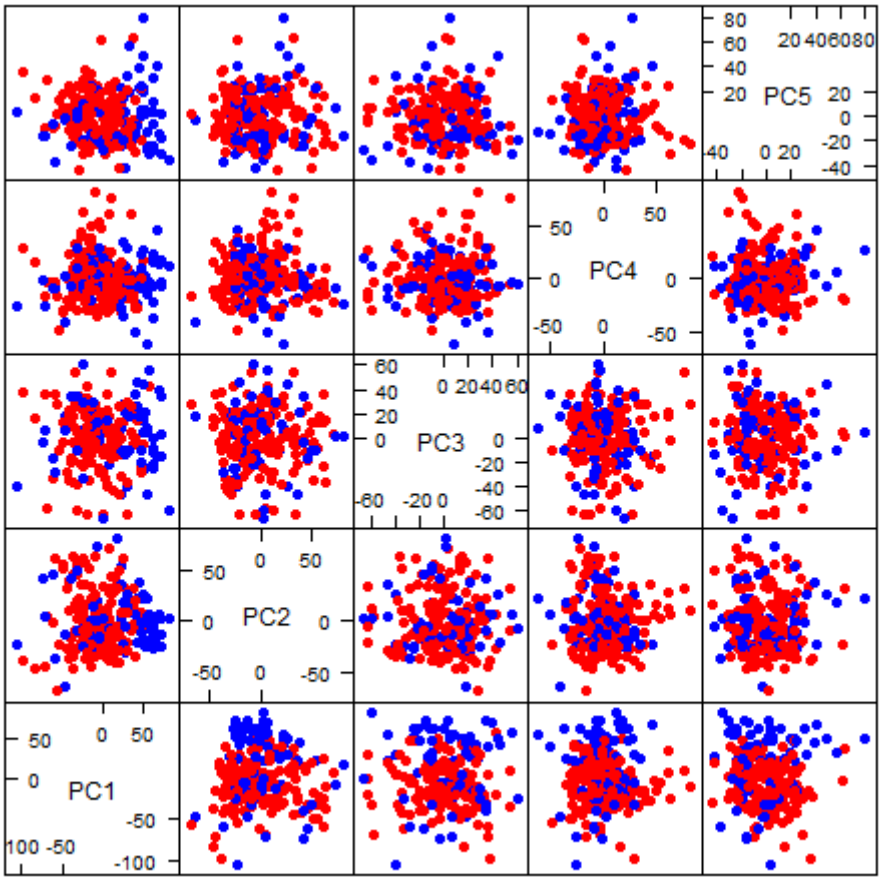
```
pc1[1:5]
A_23_P211600 A_23_P201529  A_23_P78976 A_32_P138556 A_23_P250044
   0.02452318   0.02408349   0.02336386   0.02319608  -0.02319463

pc2[1:5]
  A_23_P211600  A_23_P201529   A_23_P78976  A_32_P138556  A_23_P250044
  4.219083e-05  5.781348e-04  4.194760e-04 -2.799018e-03 -7.689996e-03
```

The absolute values specify the proportion of contribution to pc1 (/pc2 respectively) of the gene.

Task 1.4

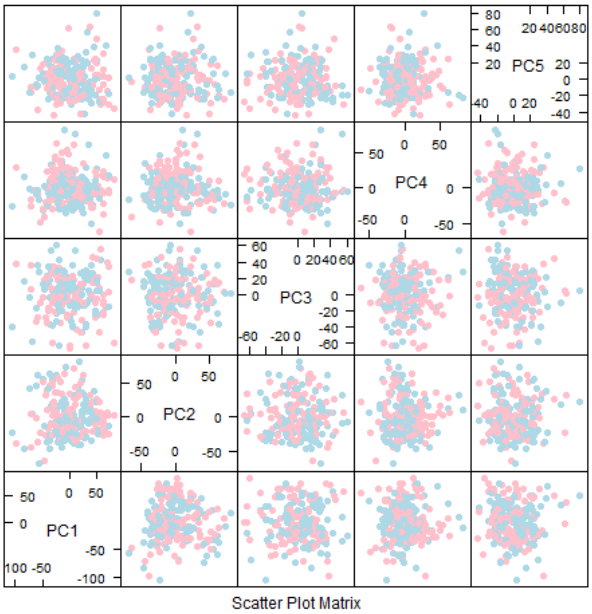
The classes are not really separated, but there is a trend between PC1 and the others. The rectal cancer (blue) is tendential on the right side (1st column).



Scatter Plot Matrix

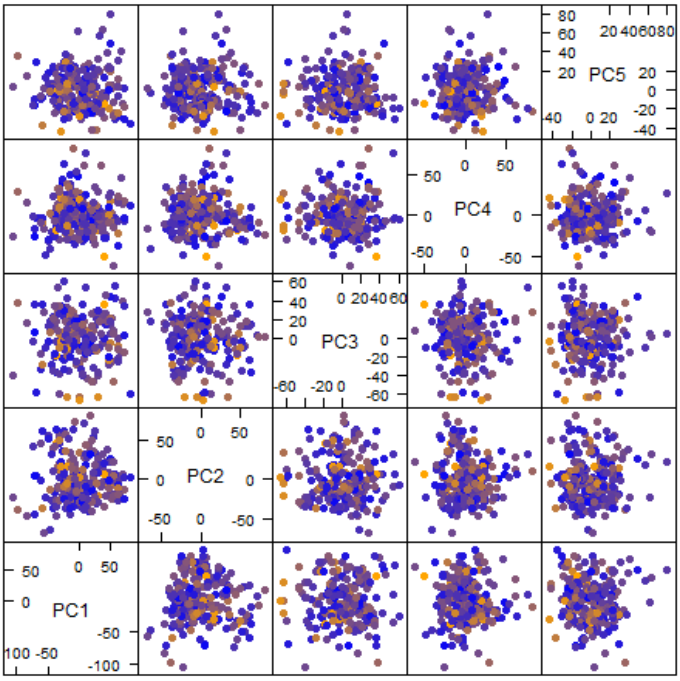
Task 1.5

We try to correlate the variance of the PC's with the gender. It seems to be also random.



Scatter Plot Matrix

Also, the relation between the age and the variance in the PC's seems to be random.



Scatter Plot Matrix

Task 2.1

This is a classification task. While we use regression to predict value of an unknown instance based on values of previous, instances, here we try to classify an unknown instance into one of two classes (colon vs rectum) based on known classes of previous instances.

Task 2.2

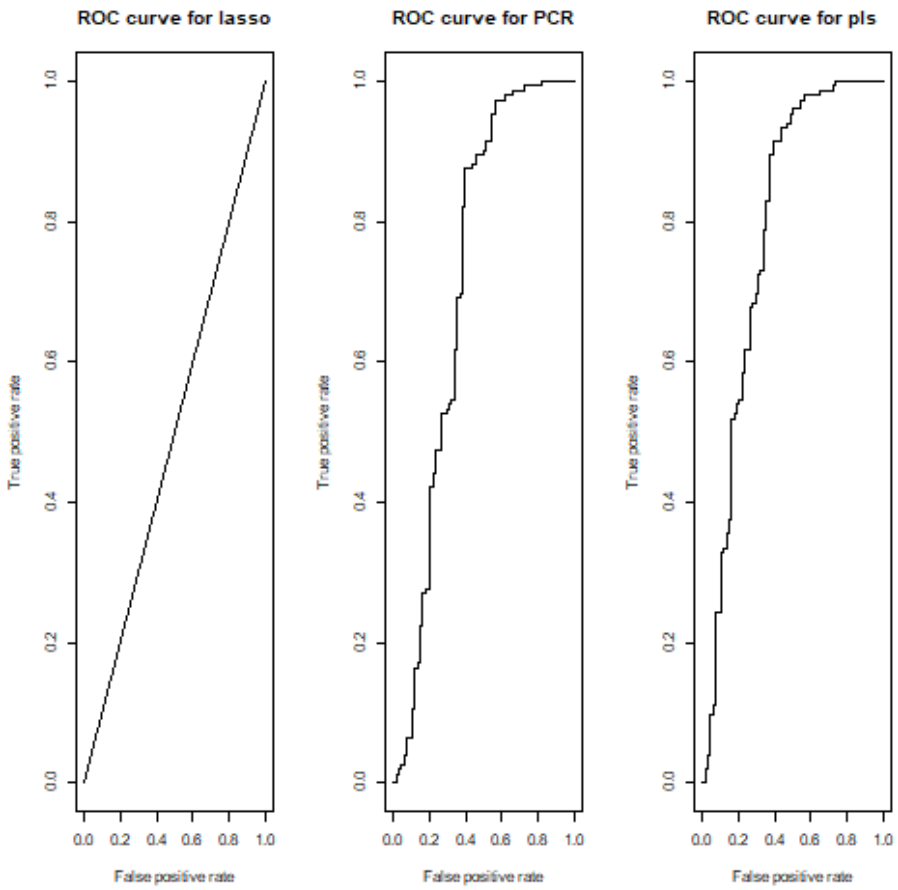
With a look on the correlation between genes and the cancer type, we can see there is a correlation between some genes and the cancer type (0.5). So PLSR can be used to convert a set of correlated variables to a set of independent variables with usage of linear transformations.

Task 2.3

We performed the partial least squares regression with 10 components.

Task 2.4

To evaluate the performance of the methods we decide to plot the ROC curves and the curve of sensitivity vs specify. All these ROC curves are not 'perfect', but pls has the best performance and lasso performed worst.



Same for the senility vs. specify. Pls shows the best results, lasso is still worse