HW1

Anwesha Bhattacharyya

September 18, 2017

Introduction

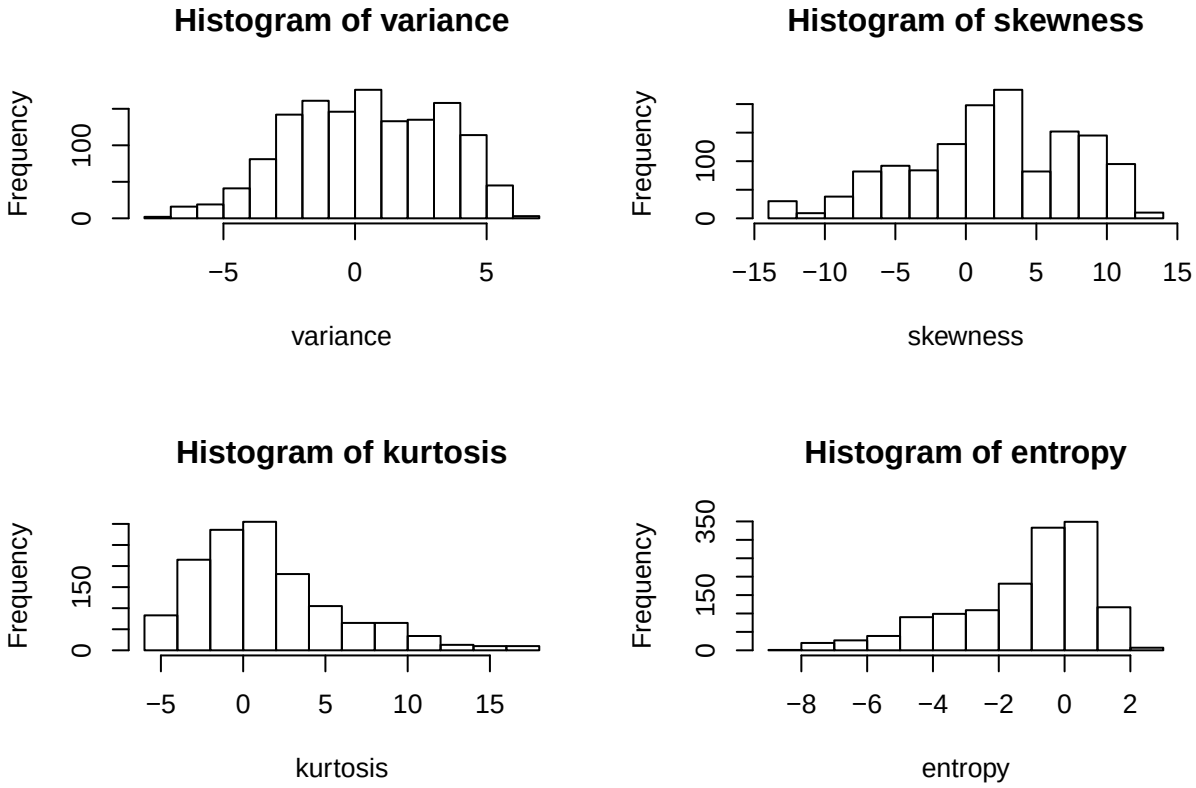
This homework is based on the "Banknote Authentication" dataset obtained in UCI laboratory. The owner of the dataset is Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe) and the donor of the dataset is Helene Darksen (University of Applied Sciences, Ostwestfalen-Lippe). Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images. There are five variables in the dataset. The objective is classification of the images as coming from a genuine banknote or a forged banknote like specimen. The first attribute is variance of the wavelet transformed image. The second is the skewness of the wavelet transformed image and the third is the kurtosis of the wavelet transformed image. The fourth variable dentes the entropy of images. All of the four attributes are continuous observations. The fifth variable is the classification variable which indicates whether its a genuine note or a forged specimen.
Data Summary

```
data = read.table("C:/Users/Anwesha/Desktop/700/banknote.txt", sep = ",")
n = dim.data.frame(data)[1]
class = data$V5
table(class)
```

```
## class
##   0   1
## 762 610
```

There is a total 1372 number of observations with no missing values and we want to build a supervised classifier. There are **610** genuine banknotes and **762** false ones. Following is the histogram of the four described attributes.

```
variance = data[,1]; skewness = data[,2]; kurtosis = data[,3]; entropy = data[,4]
par(mfrow = c(2,2))
hist(variance);hist(skewness);hist(kurtosis); hist(entropy)
```



The noticeable characteristics are the high variance of the "skewness" distribution, the positive skewness and variance of the "kurtosis" distribution, and the negative skewness of the "entropy" distribution. None of the distributions resemble a Gaussian or bell shaped distribution. however the flatness of the "variance" distribution indicates that it might be modelled as a mixture of two gaussian distribution.

A summary of the variables is presented below. One can quickly observe from the table and the histogram that the distribution of the variance and the skewness appears to be symmetric around 0 .

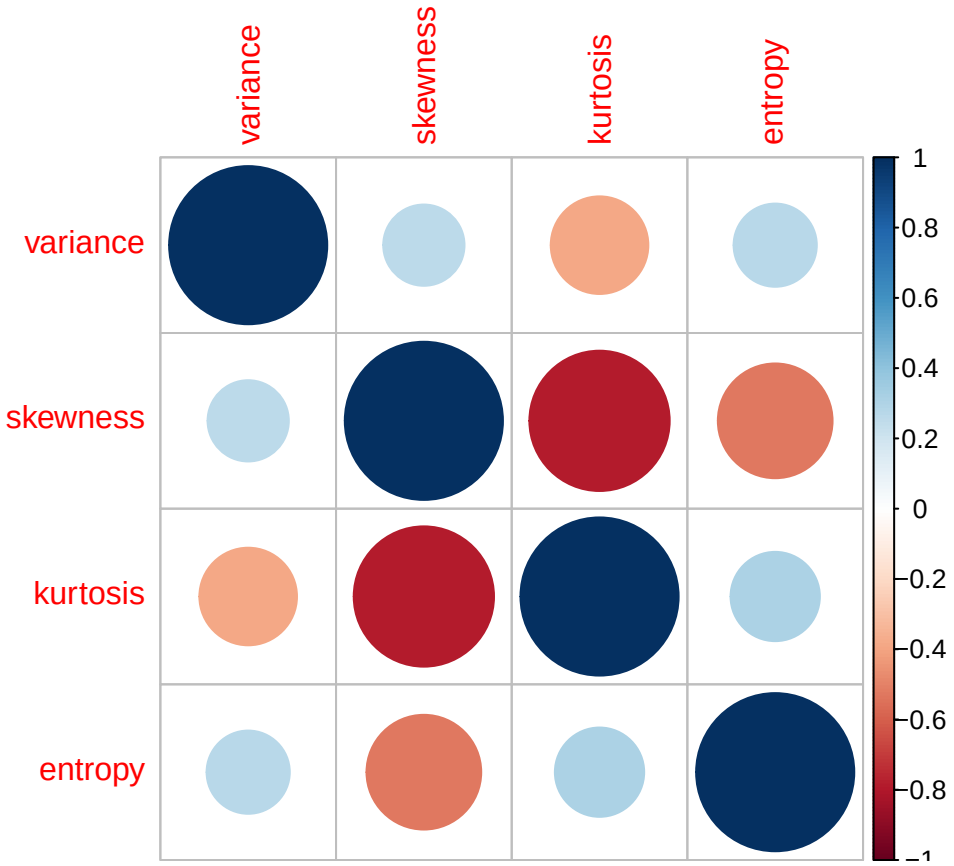
```
summary(cbind(variance,skewness,kurtosis,entropy,class))
```

```
##     variance          skewness          kurtosis          entropy
##  Min.   :-7.0421   Min.   :-13.773   Min.   :-5.2861   Min.   :-8.5482
##  1st Qu.:-1.7730   1st Qu.: -1.708   1st Qu.:-1.5750   1st Qu.:-2.4135
##  Median : 0.4962   Median :  2.320   Median : 0.6166   Median :-0.5867
##  Mean   : 0.4337   Mean   :  1.922   Mean   : 1.3976   Mean   :-1.1917
##  3rd Qu.: 2.8215   3rd Qu.:  6.815   3rd Qu.: 3.1793   3rd Qu.: 0.3948
##  Max.   : 6.8248   Max.   : 12.952   Max.   :17.9274   Max.   : 2.4495
##      class
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4446
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Correlation plots

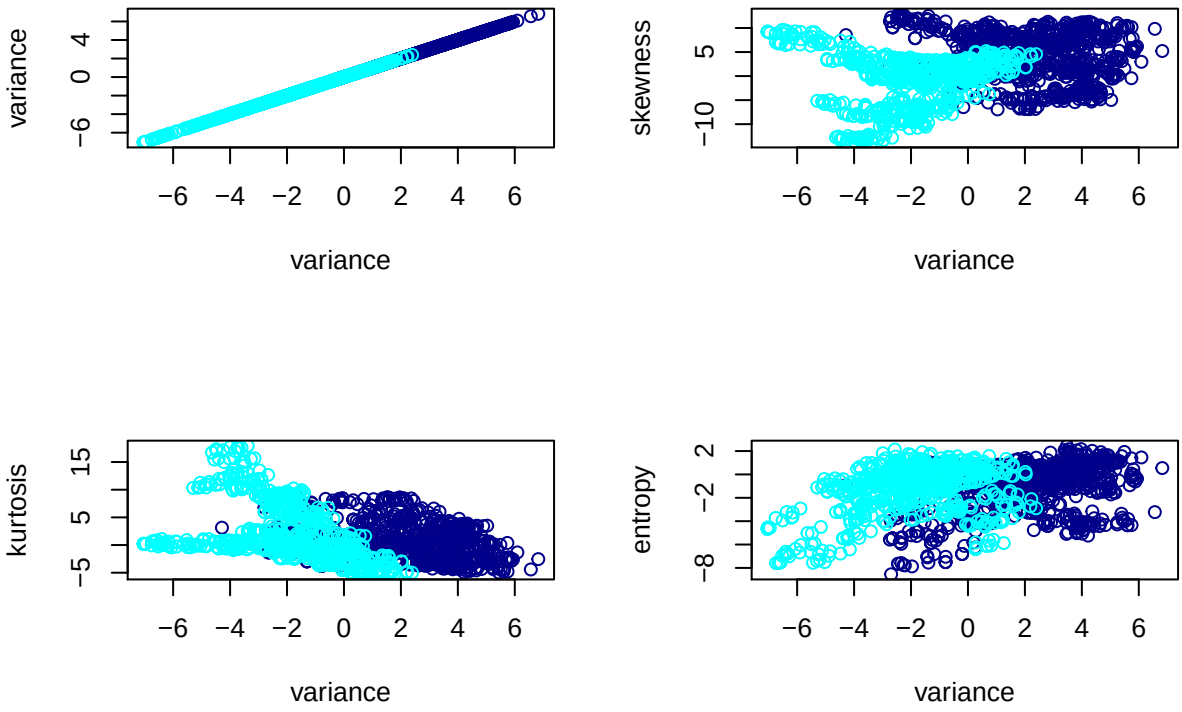
Next we study the association between the four attributes through a correlation plot. There is a high negative correlation between "skewness" and "kurtosis" of the images. The other associations are comparitively much less significant.

```
library(corrplot)
corrplot(corr = cor(cbind(variance,skewness,kurtosis,entropy)))
```



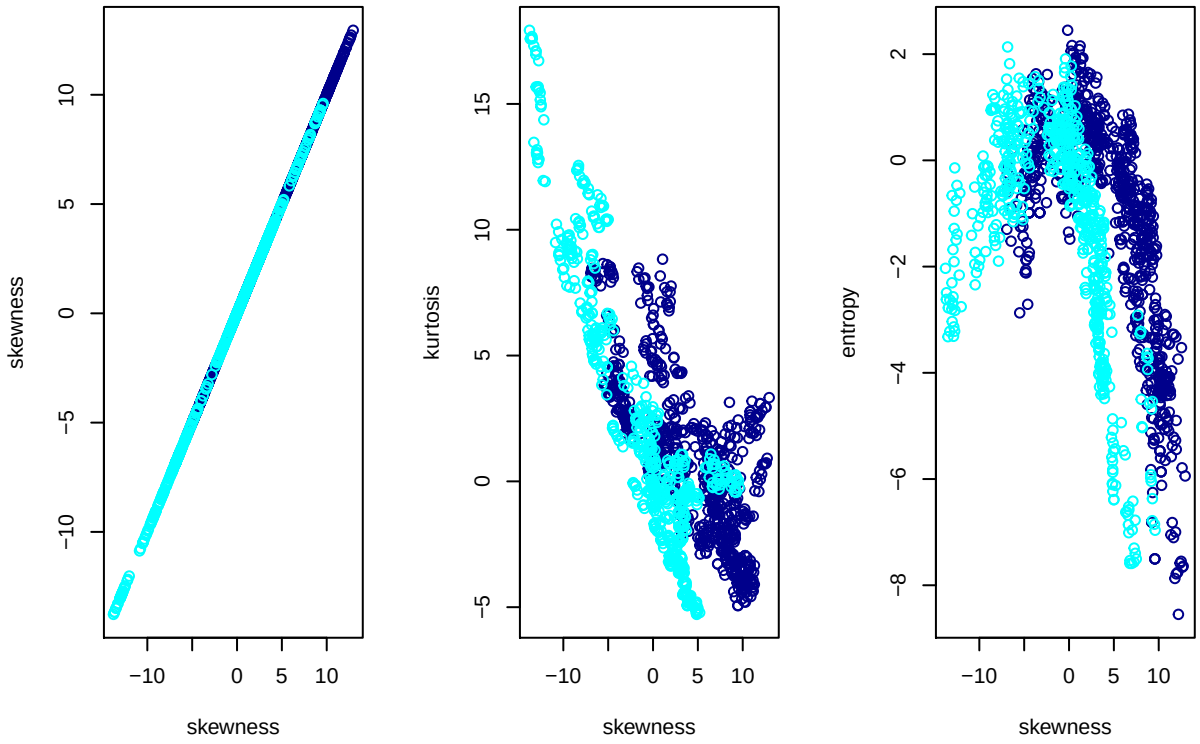
Classification plots for each attribute.

```
par(mfrow = c(2,2))
plot(data[,1],data[,1],col = c("darkblue","cyan")[data[,5]+1], xlab = "variance", ylab  = "variance")
plot(data[,1],data[,2],col = c("darkblue","cyan")[data[,5]+1], xlab = "variance", ylab = "skewness")
plot(data[,1],data[,3],col = c("darkblue","cyan")[data[,5]+1], xlab = "variance", ylab = "kurtosis")
plot(data[,1],data[,4],col = c("darkblue","cyan")[data[,5]+1], xlab = "variance", ylab = "entropy")
```



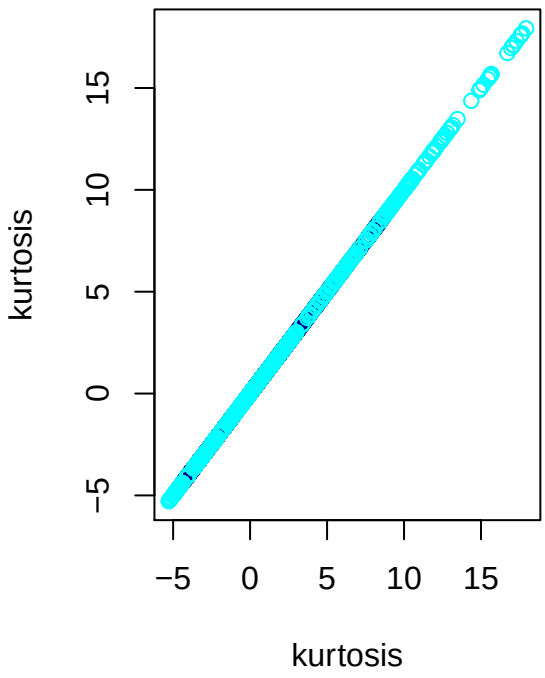
The variance attribute as well as the variance and skewness attribute jointly seems to be well seperated in the two classes. hence these two variates would be good classifiers. On the other hand kurtosis and entropy seems to introduce some overlap in the classes which indicates they are not the ideal candidates as classifiers as compared to variance and skewness.

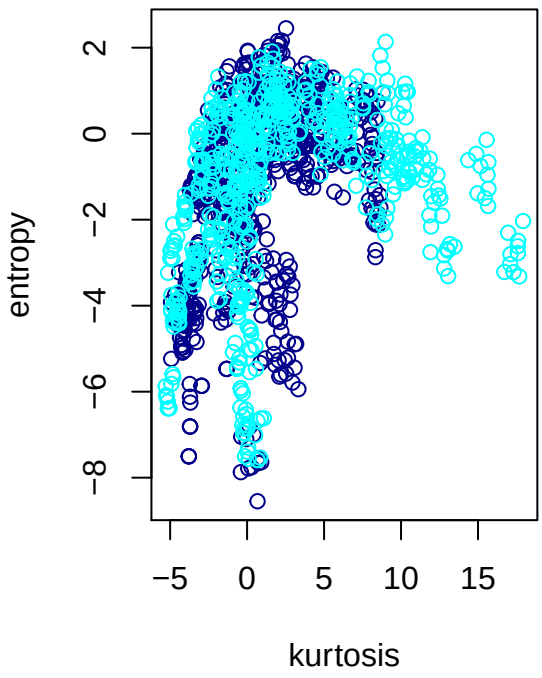
```
par(mfrow = c(1,3))
plot(data[,2],data[,2],col = c("darkblue","cyan")[data[,5]+1], xlab = "skewness", ylab = "skewness")
plot(data[,2],data[,3],col = c("darkblue","cyan")[data[,5]+1], xlab = "skewness", ylab = "kurtosis")
plot(data[,2],data[,4],col = c("darkblue","cyan")[data[,5]+1], xlab = "skewness", ylab = "entropy")
```



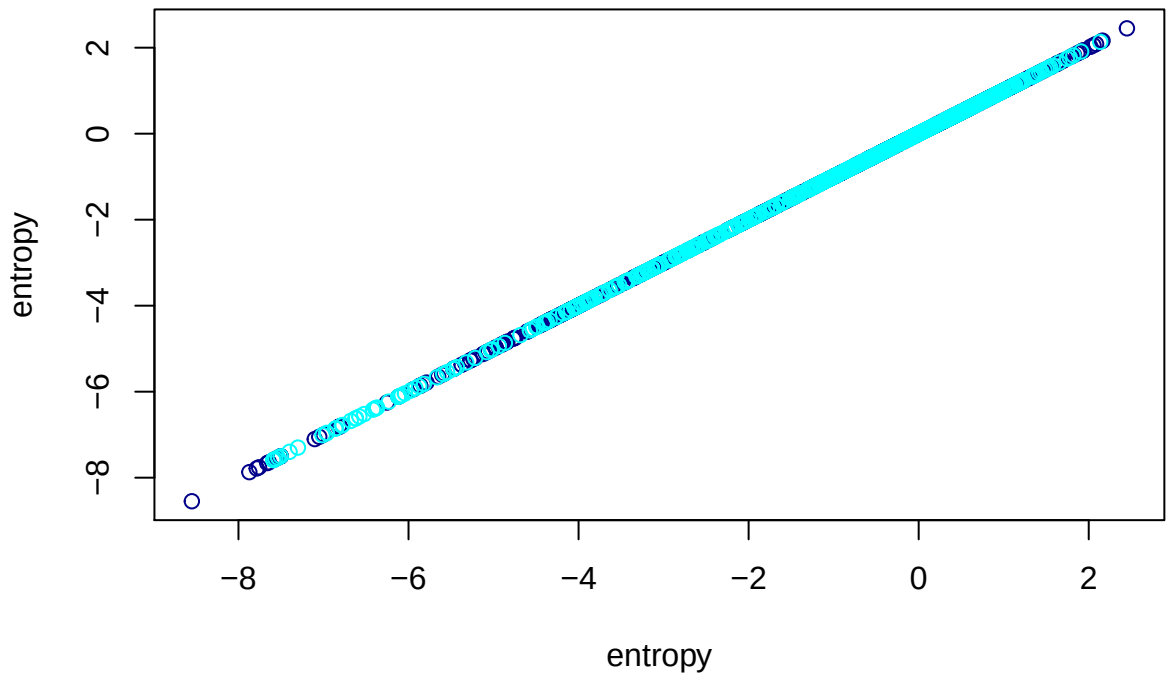
The skewness attribute seems to be a less efficient classifier than the variance attribute in the marginal sense. The plots clearly show the negative relation between skewness and kurtosis while they also indicate two different quadratic relationship between skewness and entropy for each of the two classes respectively.

```
par(mfrow = c(1,2))
plot(data[,3],data[,3],col = c("darkblue","cyan")[data[,5]+1], xlab = "kurtosis", ylab = "kurtosis")
plot(data[,3],data[,4],col = c("darkblue","cyan")[data[,5]+1], xlab = "kurtosis", ylab = "entropy")
```






```
plot(data[,4],data[,4],col = c("darkblue","cyan")[data[,5]+1], xlab = "entropy", ylab = "entropy")
```



The last set of three plots validates the initial idea that kurtosis and entropy should be poor classifiers as compared o variance and kurtosis of the images.

General linear model

We first fit a logistic regression model using general linear model with binomial family. We first fit a model without intercepts.

```
gl = glm(class ~ variance +skewness + kurtosis + entropy - 1,  family = "binomial")
summary(gl);
```

```
##
## Call:
## glm(formula = class ~ variance + skewness + kurtosis + entropy -
##     1, family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.92504  -0.00120  -0.00012   0.14632   2.62873
##
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## variance  -2.7285     0.2445 -11.159  < 2e-16 ***
## skewness  -1.6781     0.1837  -9.135  < 2e-16 ***
## kurtosis  -1.7833     0.1868  -9.548  < 2e-16 ***
## entropy   -0.8361     0.1322  -6.327  2.5e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1902.00  on 1372  degrees of freedom
## Residual deviance:  267.84  on 1368  degrees of freedom
## AIC: 275.84
##
## Number of Fisher Scoring iterations: 9
```

```
pred = gl$fitted.values; pred = round(pred)
errsq = sum((pred - class)^2)/n; errsq
```

```
## [1] 0.04154519
```

```
table(class,pred)
```

```
##      pred
## class   0   1
##     0 762   0
##     1  57 553
```

This shows all of the attributes to be significant. The prediction error sum of squares is **0.04**. There is a total number of 57 misclassification of genuine notes being classified as non-genuine.

We do the same fiiting next with an intercept.

```
gl = glm(class ~ variance +skewness + kurtosis + entropy ,  family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(gl);
```

```
##
## Call:
## glm(formula = class ~ variance + skewness + kurtosis + entropy,
##     family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.70001   0.00000   0.00000   0.00029   2.24614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.3218     1.5589   4.697 2.64e-06 ***
## variance     -7.8593     1.7383  -4.521 6.15e-06 ***
## skewness     -4.1910     0.9041  -4.635 3.56e-06 ***
## kurtosis     -5.2874     1.1612  -4.553 5.28e-06 ***
## entropy      -0.6053     0.3307  -1.830   0.0672 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1885.122  on 1371  degrees of freedom
## Residual deviance:   49.891  on 1367  degrees of freedom
## AIC: 59.891
```

```
##
## Number of Fisher Scoring iterations: 12
```

```
pred = gl$fitted.values; pred = round(pred)
errsq = sum((pred - class)^2)/n; errsq
```

```
## [1] 0.008017493
```

```
table(class,pred)
```

```
##      pred
## class   0   1
##     0 757   5
##     1   6 604
```

Two important change is that while the intercept becomes significant the attribute looses its significance at 5 /% level of significance. Also the predicion error square goes down to **.008**. The number of mis-classifications is greatly reduced to 11, with 6 images of genuine notes being mis-classified as non-genuine and 5 images being mis-classified the other way round.

Conclusion

The plot analysis indicate that the variance and skewness attributes themselves are sufficient as classifiers but the logistic regression analysis suggests that kurtosis is also a significant predictor. However as shown by the correlation plot analysis shows kurtosis and skewness are highly correlated which may lead to the significance of kurtosis as a predictor in logistic regresion.
This dataset has been widely used for studying vaious classification techniques and hence comparing the methods against those papers would give us a better analysis of the efficiency of the Bayesian methods. I would probably use this dataset for the rest of the homework problems. However I would like to work on a more challenging and interesting dataset for the project.