Analysis of Echocardiogram Dataset

STATS 4M03 Final Report

Rachel Kwan

04/13/2021

Abstract

We will analyze an echocardiogram dataset retrieved from the UCI Machine Learning Repository. From PCA, we see that the data can be reduced to three dimensions while still explaining around 80% of the total variance in the data. When we use all variables for classification, we obtain the lowest misclassification rate of 19.35%, even though these variables do not seem to follow a multivariate normal distribution. Removing the variable that seems to be causing the non-normality and using four variables for classification gives a misclassification rate of 24.19%. If misclassification costs are considered and we use all five variables, then the quadratic rule we obtain is better at correctly classifying individuals who will not survive at least one year after a heart attack, and has a misclassification rate of 20.97%. Finally, using a linear classification rule gives a misclassification cost of 22.58%, so LDA is also sufficient in this case.

Introduction

The Echocardiogram Data Set contains information from 132 patients who have all suffered a heart attack at some point in the past. It was collected by Dr. Evlin Kinney and donated to the UCI Machine Learning Repository by Steven Salzberg in 1989. The main problem presented by this dataset is to classify whether or not a patient will survive for at least one year after having a heart attack (Kinney, 1989). To prepare the data for classification, the variables *mult*, *name*, and *group* were removed since they are not useful for our analysis, as well as *wallmotion.score* since *wallmotion.index* is used instead. Finally, the variables *survival* and *alive* were used to derive our target variable (aliveat1), so these two variables will also be removed for analysis.

Analysis

Principal Component Analysis (PCA)

We will perform PCA on the remaining five continuous variables: *age*, *fractionalshortening*, *epss*, *lvdd*, *wallmotion.index*. Since these variables are on different scales, the correlation matrix will be used instead of the covariance matrix.

Table 1: Correlation matrix

age	fractionalshortening	epss	lvdd	wallmotion.index
1.0000	-0.0858	0.0385	0.1585	0.0537
-0.0858	1.0000	-0.3631	-0.3174	-0.2732
0.0385	-0.3631	1.0000	0.6282	0.4015
0.1585	-0.3174	0.6282	1.0000	0.2665
0.0537	-0.2732	0.4015	0.2665	1.0000

From the correlation matrix above, we observe that fractional shortening is negatively correlated with every other variable. This makes sense since low values of fractional shortening indicate more abnormal hearts. Also, age is weakly positively correlated with epss, lvdd, and wallmotion index, and is weakly negatively correlated with fracitonal shortening. This means that as the age at which the heart attack occured increases, there is a decrease in overall heart health. Using this matrix for PCA, we get the following principal components:

Table 2: Principal components

	PC1	PC2	PC3	PC4	PC5
age	0.1422	0.9685	0.1250	0.0953	-0.1307
fractionalshortening	-0.4309	0.0257	-0.3732	0.8204	-0.0375
epss	0.5706	-0.1564	-0.3052	0.1322	-0.7344
lvdd	0.5374	0.0712	-0.5421	0.0626	0.6390
wallmotion.index	0.4240	-0.1783	0.6768	0.5445	0.1841

The first principal component is almost equally composed of the four heart measurements, so this component can be seen as an overall measure of heart health (recall that lower values of fractional shortening indicates more abnormal hearts). The second principal component can be seen as a measure of a patient's age at which they had a heart attack.

Table 3: Summary of PCA

	PC1	PC2	PC3	PC4	PC5
standard deviation	1.5103	0.9865	0.8209	0.8186	0.63392
proportion of variance	0.4562	0.1946	0.1348	0.1340	0.08037
cumulative proportion	0.4562	0.6508	0.7856	0.9196	1.0000

The above summary table shows that our data can be reduced to 3 dimensions while still explaining around 80% of total variance.

Classification

First we will try to perform classification, again, using the five remaining continuous variables. Let π_1 be the class of individuals not surviving at least one year after a heart-attack and let π_2 be the class of individuals who do survive at least one year, corresponding to *aliveat1* values of 0 and 1, respectively. Before proceeding, we will check the assumption that these variables follow a multivariate distribution. Looking at the histogram of each variable individually, it is clear that *wallmotion.index* is not normally distributed (Figure 1). Further, looking at a matrix of scatter plots, all pairs of variables seem to follow a bivariate normal distribution (their points form an elliptical shape), except for pairs including *wallmotion.index* (Figure 2). Thus, these five variables do not appear to have a multivariate normal distribution. However, if we actually try to perform

classification using all five variables, we will see that we get a fairly low error rate. The mean vectors and covariance matrices are

$$\bar{x}_1 = \begin{bmatrix} 62.477 \\ 0.237 \\ 11.051 \\ 4.685 \\ 1.281 \end{bmatrix} \quad \bar{x}_2 = \begin{bmatrix} 69.196 \\ 0.174 \\ 15.380 \\ 5.140 \\ 1.714 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 63.186 & -0.054 & -1.612 & 0.997 & 0.0345 \\ -0.054 & 0.012 & -0.294 & -0.030 & -0.007 \\ -1.612 & -0.294 & 35.350 & 1.837 & 0.644 \\ 0.997 & -0.030 & 1.837 & 0.516 & 0.045 \\ 0.034 & -0.007 & 0.644 & 0.045 & 0.125 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 73.952 & 0.067 & 0.443 & -0.023 & -1.059 \\ 0.067 & 0.006 & -0.163 & -0.013 & -0.018 \\ 0.443 & -0.163 & 87.984 & 5.446 & 2.132 \\ -0.023 & -0.013 & 5.446 & 0.694 & 0.158 \\ -1.059 & -0.018 & 2.132 & 0.158 & 0.255 \end{bmatrix}$$

Since S_1 and S_2 are not equal, we will use the quadratic classification rule. Assuming equal misclassification costs and prior probabilities, we get the following classification rule:

Allocate x_0 to π_1 if

$$x_0 \begin{bmatrix} -0.0010 & -0.0152 & -0.0026 & 0.0186 & 0.0420 \\ -0.0152 & 55.236 & -0.3473 & -1.3856 & 7.1944 \\ -0.0026 & -0.3473 & -0.0090 & -0.0302 & 0.0181 \\ 0.0186 & -1.3856 & -0.0302 & 0.1133 & -0.0717 \\ 0.0420 & 7.1944 & 0.0181 & -0.0717 & -1.1946 \end{bmatrix} x_0' +$$
$$\begin{bmatrix} -0.217 & -15.557 & 0.95 & -2.231 & -7.083 \end{bmatrix} x_0' + 19.394 \geq 0$$

Otherwise, allocate x_0 to π_2.

This classification rule gives the confusion matrix

$$\begin{bmatrix} 37 & 7 \\ 5 & 13 \end{bmatrix}$$

so the apparent error rate is APER $= \frac{7+5}{44+18} \approx 0.1935$, meaning that around 20% of the observations were misclassified.

We would like to minimize the error rate as much as possible, so next we will try to do classification without *wallmotion.index* since we previously found that it is not normally distributed. We get the following mean vectors and covariance matrices:

$$\bar{x}_1 = \begin{bmatrix} 62.477 \\ 0.237 \\ 11.051 \\ 4.685 \end{bmatrix} \quad \bar{x}_2 = \begin{bmatrix} 69.196 \\ 0.174 \\ 15.380 \\ 5.140 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 63.186 & -0.054 & -1.612 & 0.997 \\ -0.054 & 0.012 & -0.294 & -0.030 \\ -1.612 & -0.294 & 35.350 & 1.837 \\ 0.997 & -0.030 & 1.837 & 0.516 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 73.952 & 0.067 & 0.443 & -0.023 \\ 0.067 & 0.006 & -0.163 & -0.013 \\ 0.443 & -0.163 & 87.984 & 5.446 \\ -0.023 & -0.013 & 5.446 & 0.694 \end{bmatrix}$$

Again, since S_1 and S_2 are not equal, we will use the quadratic classification rule, which is:

Allocate x_0 to π_1 if

$$x_0 \begin{bmatrix} -0.0015 & -0.1091 & -0.0019 & 0.0207 \\ -0.1091 & 37.2483 & -0.2275 & -0.9920 \\ -0.0019 & -0.2275 & -0.0087 & -0.0315 \\ 0.0207 & -0.9920 & -0.0315 & 0.1065 \end{bmatrix} x_0' + \begin{bmatrix} -0.019 & 20.942 & 0.828 & -2.872 \end{bmatrix} x_0' + 4.189 \geq 0$$

Otherwise, allocate x_0 to π_2.

Now we get the confusion matrix

$$\begin{bmatrix} 34 & 10 \\ 5 & 13 \end{bmatrix}$$

and so the apparent error rate is APER $= \frac{10+5}{44+18} \approx 0.2419$, meaning that around 24% of the observations were misclassified, which is higher than when all five variables were used.

Next, we will see what happens to our error rate when we consider misclassification costs, and using all five variables. Intuitively, it would be more costly to accidentally classify a patient who will not survive past 1 year as survivng past 1 year, i.e. it's more costly to mistakenly classify a π_1 individual into π_2. So let's suppose that the cost ratio is $\frac{c(1|2)}{c(2|1)} = 0.5$. (Johnson, 2007) Then if we continue to assume equal prior probabilities, we get the following classification rule:

Allocate x_0 to π_1 if

$$x_0 \begin{bmatrix} -0.001 & -0.015 & -0.003 & 0.019 & 0.042 \\ -0.015 & 55.24 & -0.347 & -1.386 & 7.194 \\ -0.003 & -0.347 & -0.009 & -0.030 & 0.018 \\ 0.019 & -1.386 & -0.030 & 0.113 & -0.072 \\ 0.042 & 7.194 & 0.018 & -0.072 & -1.195 \end{bmatrix} x_0' + \begin{bmatrix} -0.217 & -15.557 & 0.95 & -2.231 & -7.083 \end{bmatrix} x_0' + 19.39 \geq -0.693$$

Otherwise, allocate x_0 to π_2.

This rule gives the confusion matrix

$$\begin{bmatrix} 39 & 5 \\ 8 & 10 \end{bmatrix}$$

and so the apparent error rate is APER $= \frac{5+8}{44+18} \approx 0.2097$, meaning that around 21% of the observations were misclassified. This is slightly higher than the original misclassification rate of around 19% when we did not consider cost. However, by considering cost, we get a rule that is actually better at correctly classifying individuals who will not survive at least one year post heart attack, which is what we want. On the other hand, it is slightly worse at classifying the other population, but this may be a worthwhile trade-off.

Finally, we will do classification with a Fisher linear discriminant function to see how it compares to the previous quadratic classification rules. Our linear classification rule is:

Allocate x_0 to π_1 if

$$-0.1016x_1 + 2.6357x_2 - 0.0063x_3 - 0.1329x_4 - 2.2493x_5 + 10.2532 \geq 0$$

Otherwise, allocate x_0 to π_2.

Our classification rule is:

Allocate x_0 to π_1 if

$$-0.1016x_1 + 2.6357x_2 - 0.0063x_3 - 0.1329x_4 - 2.2493x_5 + 10.2532 \geq 0$$

Otherwise, allocate x_0 to π_2. Then we get the following confusion matrix,

$$\begin{bmatrix} 35 & 9 \\ 5 & 13 \end{bmatrix}$$

So the apparent error rate is APER $= \frac{9+5}{44+18} \approx 0.2258$, meaning that around 23% of the observations were misclassified. This is a relatively reasonable error rate, so linear discriminant analysis would be sufficient as well.

Conclusion

From PCA we see that the data can be reduced to three dimensions while still explaining around 80% of the total variance in the data. When we use all variables for classification, we obtain the lowest misclassification rate, even though these variables do not seem to follow a multivariate normal distribution. If we consider misclassification costs, then we get a rule that is better at correctly classifying individuals who will not survive at least one year after a heart attack. Even though the results of classification were relatively good, the classification performance may have been limited by sample size. This dataset originally has 132 observations, however many of these contain null values in the target variable *aliveat1*. After removing all null values, there are only 62 observations left to use for classification. Perhaps we would get an even better classification performance if there were more observations to work with.

References

Johnson, W., R.A. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Education.

Kinney, E. (1989). *Echocardiogram data set.* UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/echocardiogram

Appendix A (Figures)

Figure 1: Histogram of the variables age, fractionalshortening, epss, lvdd, and wallmotion.index

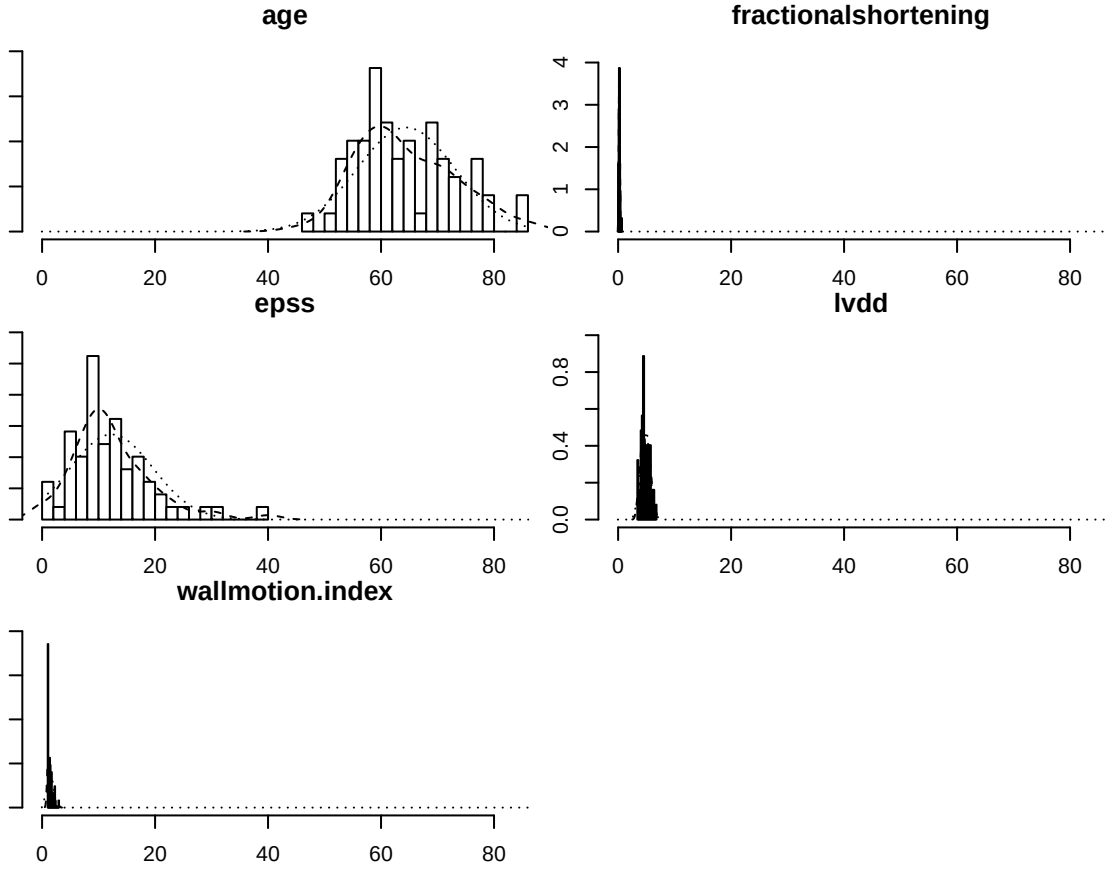
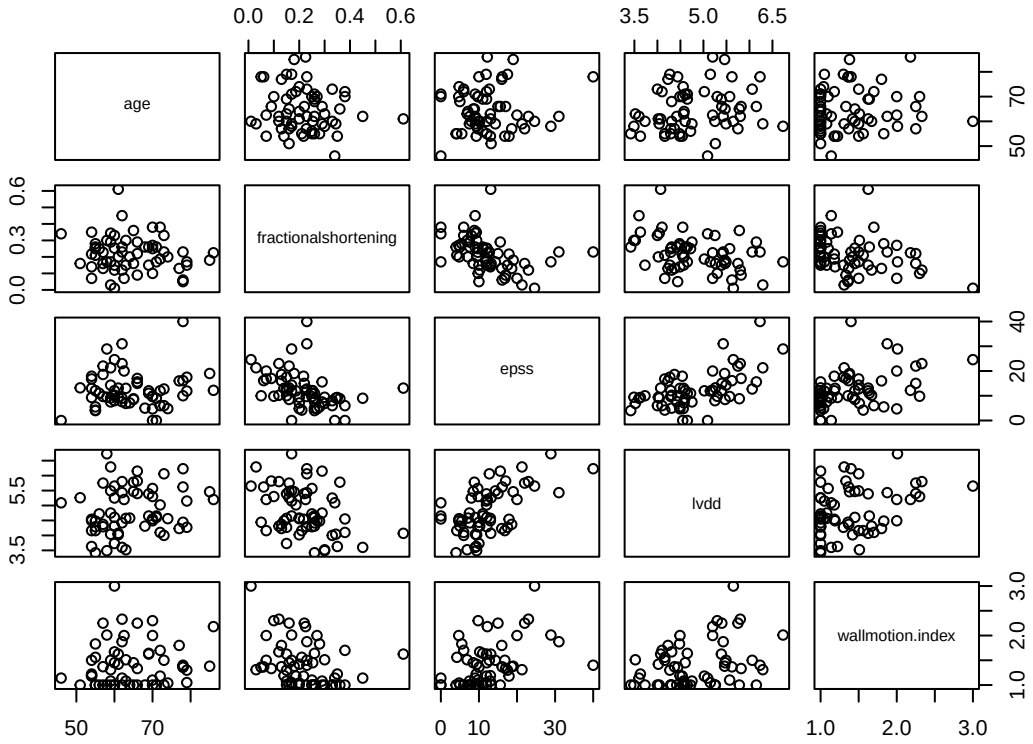


Figure 2: Matrix of scatter plots



Appendix B (Code)

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(magrittr)
library(tidyverse)
library(here)
echo <- read.csv(here("Echocardiogram Data Analysis", "echocardiogram.csv"))
echo1 <- echo %>%
  dplyr::select(-name, -mult, -group, -wallmotion.score, -survival, -alive) #%>%
  #drop_na(epss, lvdd, wallmotion.index, fractionalshortening)

head(echo1, 5)
echo5 <- echo1 %>%
  select(-pericardialeffusion) %>%
  drop_na()
echo6 <- echo5 %>%
  select(-aliveat1)
echo3 <- echo1 %>%
  select(-pericardialeffusion, -aliveat1) %>% #remove categorical variables
  drop_na()
#Using correlation matrix R
S <- cov(echo3)
```

```
R <- cov2cor(S); round(R, 4)
eig2 <- eigen(R)
eig2$values; sum(eig2$values)
eigvec2 <- eig2$vectors; round(eigvec2, 4)
screeplot <- prcomp(echo6, scale=TRUE)
summary(screeplot)
echo4 <- echo1 %>%
  dplyr::select(-pericardialeffusion) %>%
  drop_na() %>%
  arrange(aliveat1)
library(psych)
echo7 <- echo4 %>%
  dplyr::select(-aliveat1)
#classification using all variables
x1_bar <- matrix(c(mean(echo4$age[1:44]), mean(echo4$fractionalshortening[1:44]),
                   mean(echo4$epss[1:44]), mean(echo4$lvdd[1:44]),
                   mean(echo4$wallmotion.index[1:44]))) #not alive at 1
x2_bar <- matrix(c(mean(echo4$age[45:62]), mean(echo4$fractionalshortening[45:62]),
                   mean(echo4$epss[45:62]), mean(echo4$lvdd[45:62]),
                   mean(echo4$wallmotion.index[45:62]))) #alive at 1

pi1 <- matrix(c(echo4$age[1:44], echo4$fractionalshortening[1:44],
                echo4$epss[1:44], echo4$lvdd[1:44], echo4$wallmotion.index[1:44]), 44)
pi2 <- matrix(c(echo4$age[45:62], echo4$fractionalshortening[45:62],
                echo4$epss[45:62], echo4$lvdd[45:62], echo4$wallmotion.index[45:62]), 18)
S1 <- cov(pi1)
S2 <- cov(pi2)
x1_bar; x2_bar; S1; S2
S1_inv <- solve(S1)
S2_inv <- solve(S2)

k1 <- 0.5*log(det(S1)/det(S2)) + 0.5*(t(x1_bar)%*%S1_inv%*%x1_bar-
                                      t(x2_bar)%*%S2_inv%*%x2_bar)

A1 <- -0.5*(S1_inv-S2_inv)
B1 <- t(x1_bar)%*%S1_inv-t(x2_bar)%*%S2_inv
A1; B1; k1
classif <- data.frame(matrix(ncol = 62, nrow = 1))

for (i in 1:62){
  classif[i] <- as.matrix(echo7[i,]) %*% A1 %*% t(as.matrix(echo7[i,])) +
                B1 %*% t(as.matrix(echo7[i,])) - k1
}

classif_0 <- t(classif[1:44])
classif_1 <- t(classif[45:62])

matrix(c(sum(classif_0 >= 0), sum(classif_0 < 0),
         sum(classif_1 >= 0), sum(classif_1 < 0)), nrow = 2, byrow = TRUE)
echo8 <- echo7 %>%
  select(-wallmotion.index)
#classification using all variables - wallmotion.index
x1_bar_ <- matrix(c(mean(echo4$age[1:44]), mean(echo4$fractionalshortening[1:44]),
```

```
                  mean(echo4$epss[1:44]), mean(echo4$lvdd[1:44]))) #not alive at 1
x2_bar_ <- matrix(c(mean(echo4$age[45:62]), mean(echo4$fractionalshortening[45:62]),
                  mean(echo4$epss[45:62]), mean(echo4$lvdd[45:62]))) #alive at 1

pi1_ <- matrix(c(echo4$age[1:44], echo4$fractionalshortening[1:44],
                 echo4$epss[1:44], echo4$lvdd[1:44]), 44)
pi2_ <- matrix(c(echo4$age[45:62], echo4$fractionalshortening[45:62],
                 echo4$epss[45:62], echo4$lvdd[45:62]), 18)
S1_ <- cov(pi1_)
S2_ <- cov(pi2_)
x1_bar_; x2_bar_; S1_; S2_
S1_inv_ <- solve(S1_)
S2_inv_ <- solve(S2_)

k2 <- 0.5*log(det(S1_)/det(S2_)) + 0.5*(t(x1_bar_)%*%S1_inv_%*%x1_bar_-
                                        t(x2_bar_)%*%S2_inv_%*%x2_bar_)

A2 <- -0.5*(S1_inv_-S2_inv_)
B2 <- t(x1_bar_)%*%S1_inv_-t(x2_bar_)%*%S2_inv_
round(A2, 4); round(B2, 3); round(k2, 4)
classif <- data.frame(matrix(ncol = 62, nrow = 1))

for (i in 1:62){
  classif[i] <- as.matrix(echo8[i,]) %*% A2 %*% t(as.matrix(echo8[i,])) +
                B2 %*% t(as.matrix(echo8[i,])) - k2
}

classif_0 <- t(classif[1:44])
classif_1 <- t(classif[45:62])

matrix(c(sum(classif_0 >= 0), sum(classif_0 < 0),
         sum(classif_1 >= 0), sum(classif_1 < 0)), nrow = 2, byrow = TRUE)

classif <- data.frame(matrix(ncol = 62, nrow = 1))

for (i in 1:62){
  classif[i] <- as.matrix(echo7[i,]) %*% A1 %*% t(as.matrix(echo7[i,])) +
                B1 %*% t(as.matrix(echo7[i,])) - k1
}

classif_0 <- t(classif[1:44])
classif_1 <- t(classif[45:62])

matrix(c(sum(classif_0 >= -0.6931), sum(classif_0 < -0.6931),
         sum(classif_1 >= -0.6931), sum(classif_1 < -0.6931)), nrow = 2, byrow = TRUE)

S_pooled <- (44/62)*S1+(24/62)*S2;
S_pooled_inv <- solve(S_pooled)
y_hat <- t(x1_bar-x2_bar)%*%S_pooled_inv; y_hat
m <- 0.5*(t(x1_bar-x2_bar)%*%S_pooled_inv%*%(x1_bar+x2_bar)); m

classif <- data.frame(matrix(ncol = 62, nrow = 1))
```

```
for (i in 1:62){
  classif[i] <- y_hat %*% t(as.matrix(echo7[i,])) - m
}

classif_0 <- t(classif[1:44])
classif_1 <- t(classif[45:62])

matrix(c(sum(classif_0 >= 0), sum(classif_0 < 0),
         sum(classif_1 >= 0), sum(classif_1 < 0)), nrow = 2, byrow = TRUE)
multi.hist(echo7)
pairs(echo7)
```