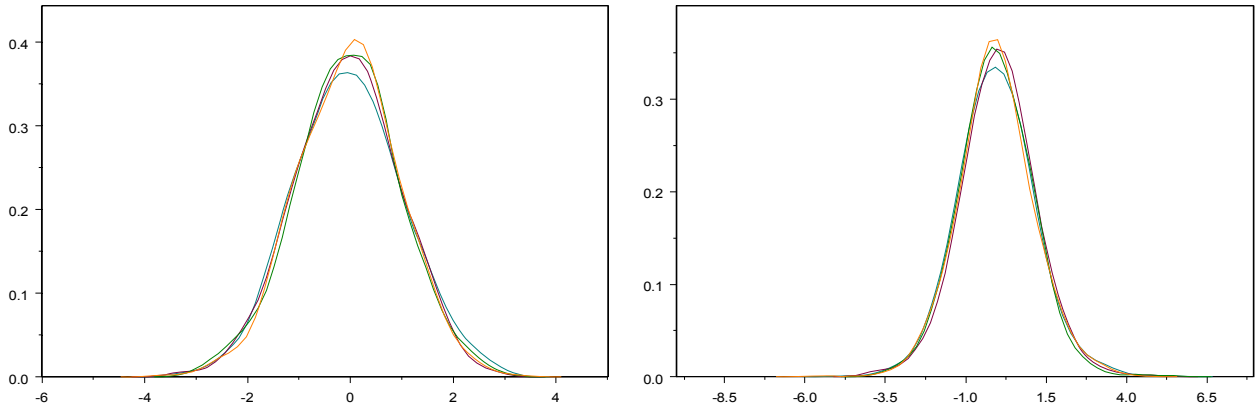
Homework 1 **Multivariate Analysis**

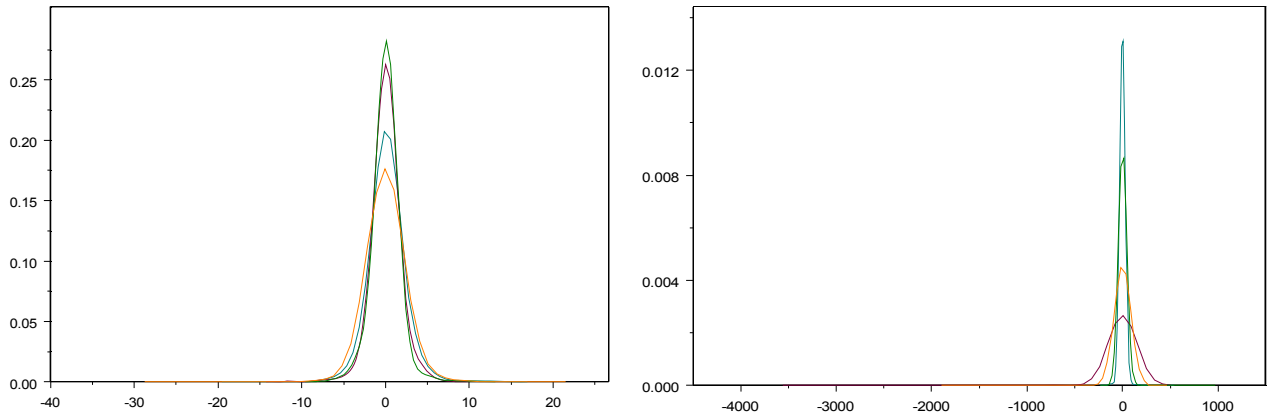
Date: 30/01/2005

Banerjee Soutrik
Lauwers Kris
Kishtammagari Manjula Rani
Rosius Wannes
Scheers Hans

Answer 2:



Figures above showing the density plot of computer-generated 4 standard normally distributed random variables of sample size 750 on the left and 4 t-distributed random variables with 8 degrees of freedom of sample size 750 on the right.



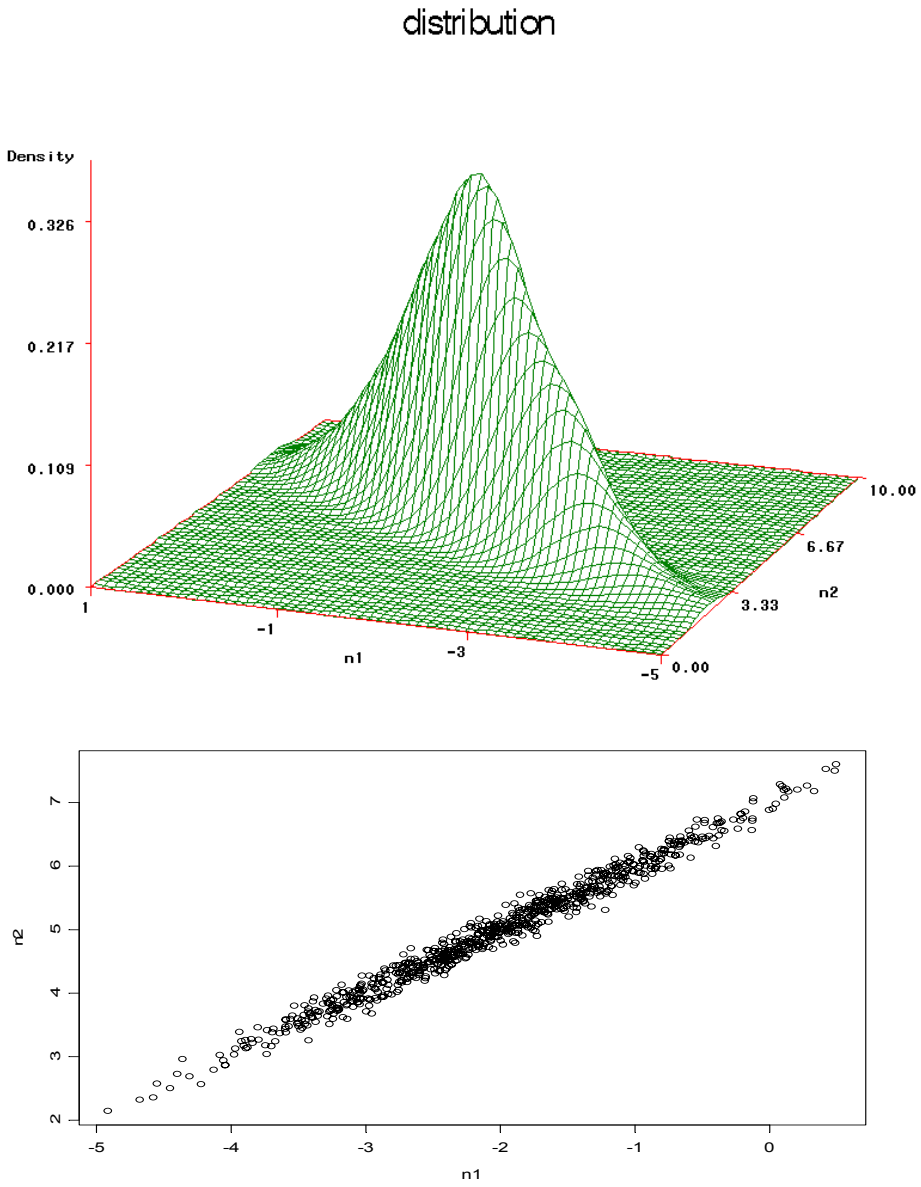
Figures above showing the density plot of computer-generated 4 t-distributed random variables with 3 degrees of freedom of sample size 750 on the left and 4 Cauchy distributed random variables (location 0, scale 1) of sample size 750 on the right.

Discussion: In the four graphs, one can observe that there is progressive bilateral thickening of the tails in an increasing order as follows : standard normal distribution > t-distribution (8 df) > t-distribution (3 df) > Cauchy distribution (0,1). This means that in the standard normal variables, the tails become flat most abruptly, whereas in the Cauchy distribution, the tails are presumably wider and become flat less abruptly.
Note : This looks paradoxically different in the figures, but this effect is visual as the scales of the X-axes get progressively wider in the above figures giving a false impression of progressive narrowing of the distributions in that order.

We did a pairwise (XY-axes) 2D plot for each type of distribution. The plots between two normally distributed random variables appeared circular in the scatter plots. The

same pattern was observed for the t_8-distributed random variables. For the t_3-distributed random variables, the scatter plot appeared elliptical in most of the cases, but one which appeared circular. Lastly, for the Cauchy distributed random variables, the scatter plot appeared cruciate in most of the plots.

Answer 3:



Figures showing 3D and 2D bivariate plots of computer-generated two normally distributed random variables (n1, n2) with $\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 5 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$. The

spreads in the two perpendicular axes are not same due to the presence of a strong correlation between these two variables. The eigenvalue and eigenvector pairs indicate the spread and direction in two perpendicular axes. The eigenvalue-eigenvector pairs are 1.99 (1, 1), where the spread is maximum, and 0.01 (1, - 1), where the spread is minimum. The directions (eigenvectors) of spread are easier to see in the 2D plot, which is maximum in the NE-SW directions.

Answer 4:

The Box-Cox power transformation (Box & Cox, 1964) is given by

$$y=\frac{x^{\lambda}-1}{\lambda} \quad \text{if} \quad \lambda \neq 0 \quad \text{and} \quad y=\log x \quad \text{if} \quad \lambda=0$$

where y has to be strictly positive. An often used solution incorporating negative and zero values is adding a constant to x where x is zero or negative. However problems arise when the result relates to the constant chosen.
To circumvent this problem several alternatives have been proposed. For instance John and Draper (1980)

$$\psi(x,\lambda)=\frac{\operatorname{sgn}(x+1)|x+1|^{\lambda}-1}{\lambda} \text{ for } \lambda \neq 0, \quad \operatorname{sgn}(x)\log(|x|+1) \text{ otherwise}$$

and Bickel and Doksum (1981) $\psi(\lambda,x)=\frac{\operatorname{sgn}(x)|x|^{\lambda}-1}{\lambda}$ proposed respectively their modulus transformation and signed power transformation. Both transformations are good to handle kurtosis problems. However, when applied to skewed distributions these transformations have their drawbacks (see Yeo & Johnson, 2000). Recently Yeo and Johnson (2000) proposed a new family of power transformations:

$$\begin{aligned}
\psi(\lambda,x) &= \frac{(x+1)^{\lambda}-1}{\lambda} && \text{if } x \geq 0, \lambda \neq 0 \\
\psi(\lambda,x) &= \log(x+1) && \text{if } x \geq 0, \lambda = 0 \\
\psi(\lambda,x) &= \frac{-[(-x+1)^{2-\lambda}-1]}{2-\lambda} && \text{if } x < 0, \lambda \neq 2 \\
\psi(\lambda,x) &= -\log(-x+1) && \text{if } x < 0, \lambda = 2
\end{aligned}$$

Here the constant 1 makes the transformed value have the same sign as the original data value.

For positive values of x their transformations are equivalent to the Box-Cox transformations. On the whole normal approximation in is much improved with these new transformations.

Answer 5:

(We opted discussing the Box-Cox transformation by adding a positive constant to the non positive values. For the constants to add, we looked at serial P-values to see which constant leads to a multivariate normal dataset.)

The radiotherapy data consists of 6 response variables, of which the last variable is categorical with values 0, 1, 2, 3. We decided not to include it for the Box-Cox transformation, since adding a constant would not render a sensible meaning to the variable. In addition, that the variable being categorical, we couldn't use it in the multivariate assumption of normality testing with other continuous variables.

The continuous variables were names as V1, V2, V3, V4, V5 respectively.

At the beginning, we checked the univariate normality assumption for each of the 5 variables. The variables 1, 2 and 4 were non-normal (Shapiro-Wilk's test, $P < .05$ for the 3 variables).

We then checked the multivariate normality of the entire set using the 'mshapiro.test' function in R, which was not significant, $P < .001$. Therefore, the assumption of multivariate normality for the 5 variables was rejected.

Next procedure was then to transform the variables using the multivariate Box-Cox power transformation on the variables (using the 'box.cox.powers' function in R).

Box-Cox family of power transformation transforms a variable X into Y as given below :

$$y = \frac{x^{\lambda} - 1}{\lambda} \quad \text{if} \quad \lambda \neq 0 \quad \text{and} \quad y = \log x \quad \text{if} \quad \lambda = 0 \quad \text{(by definition)}$$

The idea of this transformation is to make the multivariate distribution normal.

However, we found that the minimum value for the variable V1 was equal to 0. Therefore, we added a positive constant, 0.001, to all the values of this variable. We applied multivariate Box-Cox transformation to this data set of 5 variables (with V1 increased by a positive constant).

Next on this Box-Cox transformed data, we again checked the multivariate normality as before. This time the *P*-value was .028. Therefore, we had to reject the multivariate normality assumption.

Following, we checked how the *P*-value varies for different positive constants added to the variable V1. In the figure 1 below, we plot the different positive constant values (x-axis) against the *P*-values to see if for some positive constant, the significance level crossed the .05 level. We noted a maximum *P*-value of 0.04 (approx.) could be attained for the constants in the neighbourhood of value 1.1. However, the differences between

the *P*-values (with the constants added having values upto 3) were minimal as can be seen in the figure. We opine that this is not true in general, but only in this particular case.

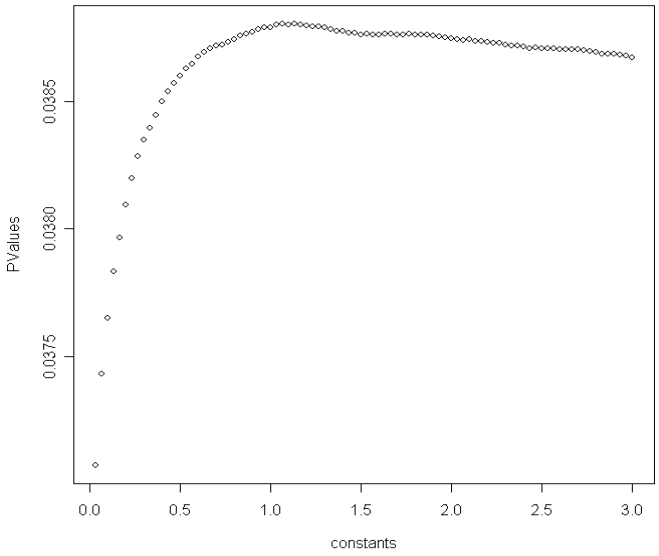


Figure 1 showing *P*-values plotted against positive constants added to the variable V1.

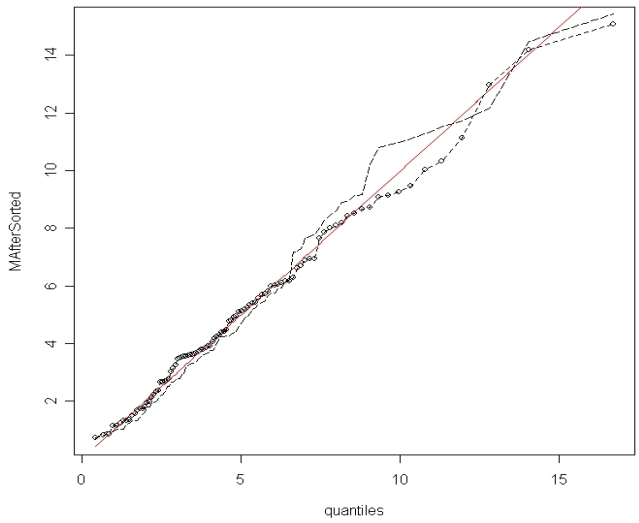


Figure 2 showing the Gamma plot with and without the Box-Cox power transformation. The line with circular hollow dots shows the plot after first Box-Cox transformation and the dashed line shows the plot before transformation.

In the second step, we did once more a Box-Cox transformation on the already Box-Cox transformed dataset (first iteration). This was done in order to get a significant *P*-value > .05 for the multivariate dataset. Before doing this, we again had to add a constant due to the same problem of negative values in the Box-Cox transformed dataset. This time the constants were added to the 2^{nd}, 3^{th} and 5^{th} variables, in order to make them positive.

Therefore, we added positive constants to make the values positive. For the second and third column, we added a value such that the minimum value was 0.1 for these two variables. For the fifth column, we add a constant in order that the minimum was 0.01 for this variable. Doing a Box-Cox transformation on this transformed dataset, it gives a *P*-value for multivariate normality of 0.112.
A word to add in this context, that we had also tried a few different constant values for the three variables during the iteration process, but the *P*-values didn't improve above our target value for constants other than mentioned before.

Concluding remark : The Box-Cox transformation yielded some improvement on the multivariate normality, however the threshold significance level of $P = .05$ couldn't be attained after the first transformation. This also shows that although we may apply multivariate Box-Cox transformation on the variables, but still multivariate normality criteria may not be achieved.
After doing a second Box-Cox, we attained a multivariate normal dataset. However, three points to add: firstly, in the second transformation step, we noticed that the lambda values were closer to 1 than those we obtained in the first step for the 5 variables. This might indicate closer approximation to multivariate normality with successive iteration by Box-Cox method. The second remark critical – by repeated transformation, the original data gets more and more transformed, hence there always remains questions to be answered regarding the significance of the results obtained by the iterative Box-Cox transformation of non-normal data. Finally, we cross checked the univariate normality of individual variables separately after the second Box-Cox transformation. It clearly shows that all the variables (two in our case) were not normal if analysed individually.

References:

Bickel, P. J. & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc*. **76**, 296-311

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Satist. Soc*. B. **26**, 211-311.

John, J. A. & Draper, N. R. (1980). An alternative family of transformations. *Appl. Statist.* **29**, 190-097.

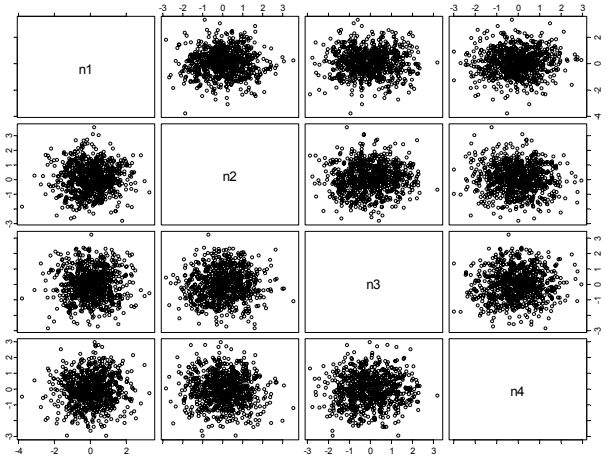
Yeo, I.-K. & Johnson, R. A. (2000). A new family pf power transformations to improve normality or symmetry. *Biometrica.* **87** (4), 954-959.

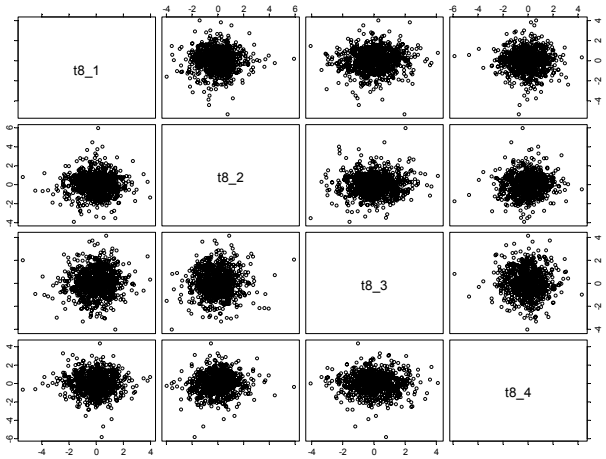
Appendix

Question 2:

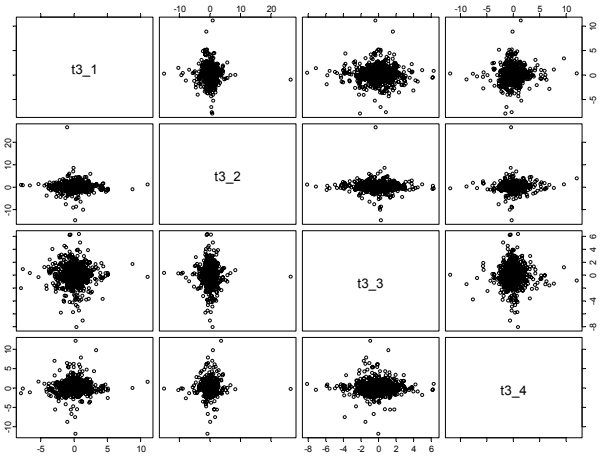
Plots:

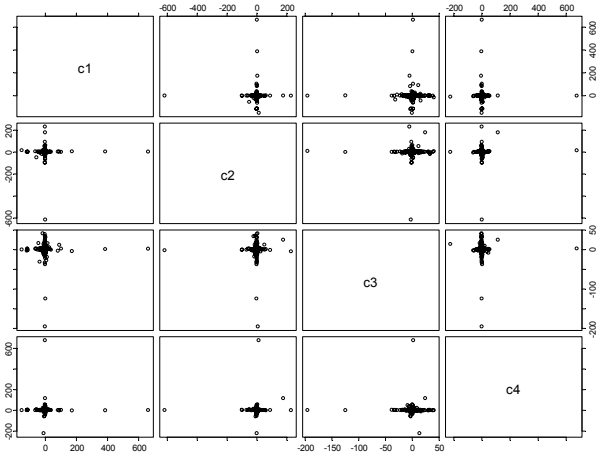
Plots of pairwise normally distributed random variables (left) and t_8-distributed random variables (right):





Plots of pairwise t_3-distributed random variables (left) and Cauchy distributed random variables (right):





S+ Code:

```
n<-750
nmat<-matrix(0,n,4)
t8mat<-matrix(0,n,4)
t3mat<-matrix(0,n,4)
cmat<-matrix(0,n,4)

for (i in 1:4)
```

```
   {
     nmat[,i]<-rnorm(n,0,1)
     t8mat[,i]<-rt(n,8)
     t3mat[,i]<-rt(n,3)
     cmat[,i]<-rcauchy(n,0,1)
   }

normal<-as.data.frame(nmat)
t8<-as.data.frame(t8mat)
t3<-as.data.frame(t3mat)
cauchy<-as.data.frame(cmat)

names(normal)<-c("n1","n2","n3","n4")
names(t8)<-c("t8_1","t8_2","t8_3","t8_4")
names(t3)<-c("t3_1","t3_2","t3_3","t3_4")
names(cauchy)<-c("c1","c2","c3","c4")

pairs(normal)
pairs(t8)
pairs(t3)
pairs(cauchy)

shapiro.test(normal$n1)
shapiro.test(normal$n2)
shapiro.test(normal$n3)
shapiro.test(normal$n4)
```

Question 3:

S+ Code:

```
sigma<-matrix(c(1,0.99,0.99,1),nrow=2,byrow=T)
eigen(sigma)
sample<-rmvnorm(750,mean=c(-2,5),cov=sigma)
dimnames(sample)<-list(1:750,c("n1","n2"))
plot(sample)
```

Question 5:

R Code:

```
# Reading the data, and deleting the 6th column

radio<-read.table("E:\\My Documents\\school\\UH\\2 de trim\\multivarial data
analysis\\HW1\\radio.dat", header=F, sep="")
radio2<-radio[,-6]

# loading the packages "CAR" and "mvnormtest"

library(car)
library(mvnormtest)


# Mahalanobis distance of the dataset

SigmaBefore<-var(radio2)
MeanBefore<-mean(radio2)
MBefore<-mahalanobis(radio2,MeanBefore,SigmaBefore)
MBeforeSorted<-sort(MBefore)
```

```
# test for multivariate normality of the dataset

mshapiro.test(t(radio2))


# make the plot to see which constant leads to the best P-value for
# multivariate normality

constants<-c(1:90)/30
PValues<-c()

for (c in 1:90)
{radio2$V1<-radio$V1+constants[c]
LValues<-box.cox.powers(radio2)$lambda
radionew<-array(0,c(98,5))
for (i in 1:5)
{radionew[,i]<-(radio2[,i]^LValues[i]-1)/LValues[i]}
PValues[c]<-mshapiro.test(t(radionew))$p}

plot(constants,PValues)

# adding a constant 1.1 (=the constant with the largest p-value) to the first
#column

radio2$V1<-radio$V1+1.1
box.cox.powers(radio2)
LValues<-box.cox.powers(radio2)$lambda
radionew<-array(0,c(98,5))
for (i in 1:5)
{radionew[,i]<-(radio2[,i]^LValues[i]-1)/LValues[i]}
mshapiro.test(t(radionew))

# The Mahalanobis distance for the new dataset

SigmaAfter<-var(radionew)
MeanAfter<-c()
for (i in 1:5)
{MeanAfter[i]<-mean(radionew[,i])}
MAfter<-mahalanobis(radionew,MeanAfter,SigmaAfter)
MAfterSorted<-sort(MAfter)

# Plotting the q-q plot, of the new, and of the old dataset

q=c()
for(i in 1:98){
q[i]=(i-0.5)/98}
quantiles=c()
for(i in 1:98){
quantiles[i]=qchisq(q[i], 5, ncp=0, lower.tail = TRUE, log.p = FALSE)}

plot(quantiles, MAfterSorted)
lines(quantiles,quantiles,col=2,lty=1)
lines(quantiles,MBeforeSorted,col=1,lty=5)
lines(quantiles,MAfterSorted,col=1,lty=2)


# Second prodedure

# Looking for minima in th columns, to see which are nonpositive, and look
# after which value to add

Minima<-c()
for (i in 1:5)
```

```
{Minima[i]<-c(min(radionew[,i]))}
Adding<-c()
for (i in 1:5)
{if (Minima[i]<=0) Adding[i]<- -Minima[i] else Adding[i]<-0 }
Adding<-Adding+c(0,0.1,0.1,0,0.01)

# Box-Cox on new dataset

radio3<-radionew
for (i in 1:5)
{radio3[,i]<-radionew[,i]+Adding[i]}
box.cox.powers(radio3)
LValues<-box.cox.powers(radio3)$lambda
radio2Transformations<-array(0,c(98,5))
for (i in 1:5)
{radionewnew[,i]<-(radio3[,i]^LValues[i]-1)/LValues[i]}
mshapiro.test(t(radionewnew))
```