Hasselt University

Master of Statistics

multivariate data analysis

HomeWork1: Multivariate Normality Assessment

Students:
Armel Maurice Cheugoua Zanetsie
Anthony Agyapong Adomah
Mbukam Edward Chongsi
Melvis Emade Ngeme-Ndie

Lecturer:
Prof. dr. Christel Faes

November 9, 2016

Contents

1 Introduction

The multivariate normal distribution is undoubtedly one of the most well-known and useful distribution in statistics, playing a predominant role in many areas of application such as representing a natural extension of the univariate normal distribution and provides a suitable model for many real-life problems concerning vector-valued data. Also, for the bivariate normal distribution, positive and negative dependence properties of the components of the random vector are completely determined by the sign and size of the correlation coefficient (Tong, 1990). In Statistics, many problems concerning data analysis as well as inference and interpretation, give more reliable results if samples are collected randomly. The tests and methods used for this analysis are usually more powerful if the sample follows a normal distribution.
In this report, univariate and multivariate random samples were taken from different distributions. The univariate random samples were generated from a standard normal distribution and Cauchy distribution. The multivariate random sample was generated from the bivariate normal distribution. Also two other techniques to assess multivariate normality were discussed.

2 Methods

2.1 Data

In this study, four different random samples each of size 250, from the standard normal (N(0,1) and Cauchy (location=0, scale=1) distributions were generated. A bivariate normal random sample of size 250 with mean μ and variance covariance matrix $\sum$ was also generated, and normality was assessed by the quantile-quantile (QQ) plot and the Shapiro Wilk test. Where

$$\mu = \begin{pmatrix} 0.0 \\ 2.0 \end{pmatrix} and \sum = \begin{pmatrix} 1.0 & 0.95 \\ 0.95 & 1.0 \end{pmatrix}$$

.

2.2 Assessing Normality

These samples were generated for the specified distribution and summary statistics were then done to get the descriptive measures. Normality was assessed using histograms, skewness, kurtosis, quantile-quantile plots. Scatter plot and gamma plots were used to graphically assess bivariate normality. Furthermore the Shapiro Wilk's test was used to test the normality assumption.
The software used was R 3.1.1

3 Results and Discussion

3.1 Sampling from the Normal distribution

From the summary statistics shown in table 1 below, the means of the generated random samples are slightly different. The sample means ranges from -0.075 to 0.008 and from the shapiro-wilk's test, the different samples generated with a normal distribution presented a

p-value > 0.05 confirming that univariate normality holds for each of the four samples.

Table 1: Summary statistics for samples from normal distribution with normality test

Measures	**Sample1**	**Sample2**	**Sample3**	**Sample4**
mean	-0.004	0.008	-0.075	-0.034
median	-0.043	0.019	-0.026	-0.083
min	-3.233	-3.396	-2.864	-2.906
max	3.043	3.195	2.356	2.706
Shapiro-Wilk(Value)	0.991	0.997	0.996	0.994
Shapiro-Wilk(p-Value)	0.138	0.924	0.698	0.383

From figure 1, the QQ plots suggest normality since there is little deviation from the QQ line and the box plots in figure 2 show possible outliers in some samples though the samples seem to be symmetrical with respect to the histogram in figure 3.
The skewness of the four samples was found to be 0.147, -0.047, -0.149, -0.031 for sample 1 to sample4 respectively. The skewness value of sample 1 indicated that the distribution of the data was slightly skewed to the right and those of sample 2 to 4 indicated that the distribution of the data is slightly skewed to the left because of the negative value and because the value is close to zero.Normally, skewness values below -1 and above +1 are clear indications for skewed distributions.
The kurtosis values of samples 1,2 and 4 were > 3 indicating leptoykurtic distribution, shaper peaks than a normal distribution, The kurtosis value for sample 3 was < 3 indicating platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.

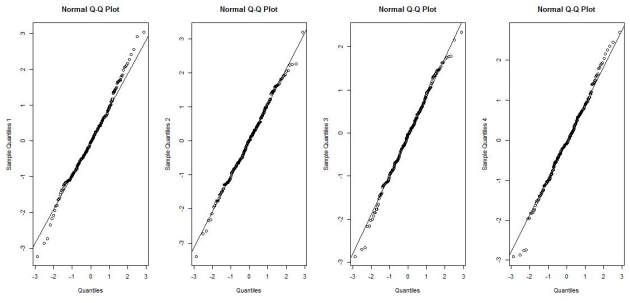

Figure 1: QQ-plots of 4 samples

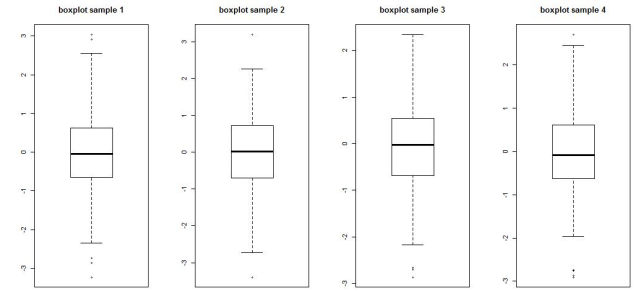

Figure 2: Box plots of 4 samples

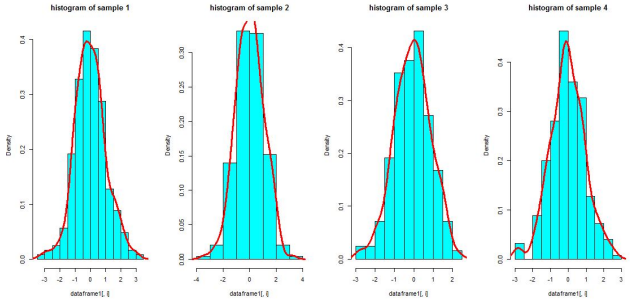

Figure 3: Histograms of 4 samples

3.2 Sampling from the Cauchy distribution

The random samples generated from Cauchy distribution did not present a normal distribution as verified by the Shapiro-Wilk's test for normality ($p < 0.05$). We also saw a great difference in the means of the samples which may indicate they maybe properply some outlying observations. This is shown in table 2 below

Table 2: Summary statistics for samples from cauchy distribution with normality test

Measures	Sample1	Sample2	Sample3	Sample4
mean	1.224	-1.198	0.727	-0.838
median	0.036	-0.071	-0.078	-0.043
min	-159.354	-177.670	-36.594	-54.485
max	701.718	26.270	152.439	19.833
Shapiro-Wilk(Value)	0.172	0.254	0.306	0.533
Shapiro-Wilk(p-Value)	2.2e-16	2.2e-16	2.2e-16	2.2e-16

As shown graphically in figure 4 below, no insight of symmetric and bell-shaped imparted and shows heavy tails from the Cauchy distribution and the random samples generated deviate from a straight line, which is an indication of non-normality.

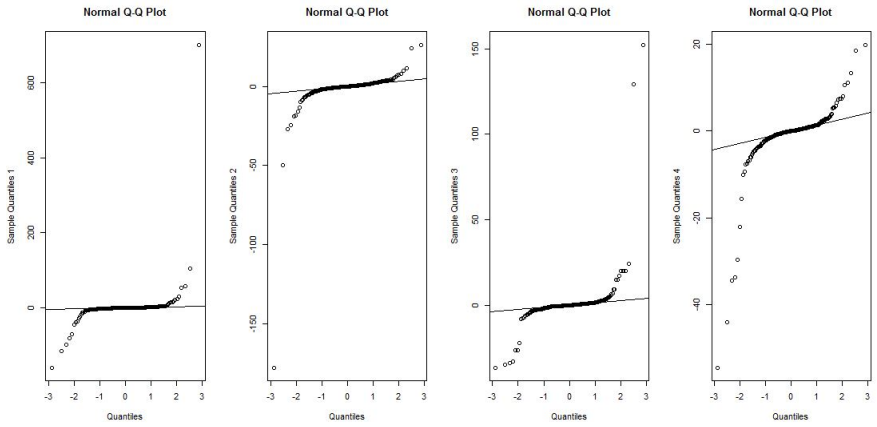

Figure 4: QQ-plots of 4 samples from Cauchy distribution

3.3 Bivariate Random Sample

A random sample of size 250 bivariate normal with means μ and variance covariance matrix $\sum$ was generated and the histograms for both variables in the univariate context showing each distribution to be symmetry with histogram for variable 2 been bimodal. The QQ plots for both variables also suggested normality since only very few points seems to deviate from the line as seen in figure 5 and 6 below. Shapiro-Wilk's test shows the variables are univariately normal and this was strongly corroborated by the Shapiro-Wilk's test for multivariate normality (p-value = 0.6667) as shown in table 3 below.

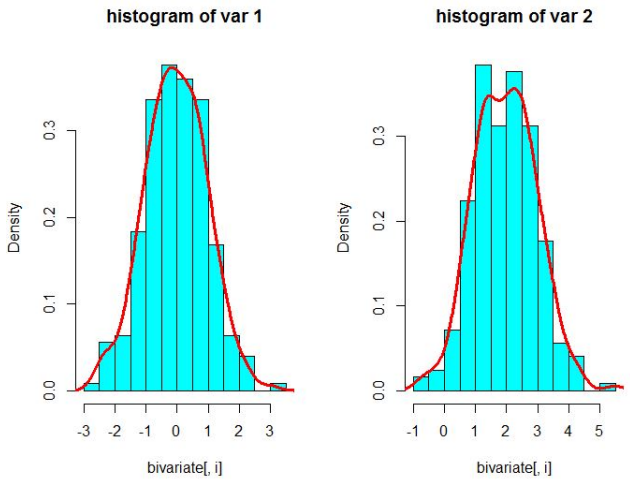

Figure 5: Histogram for bivariate samples

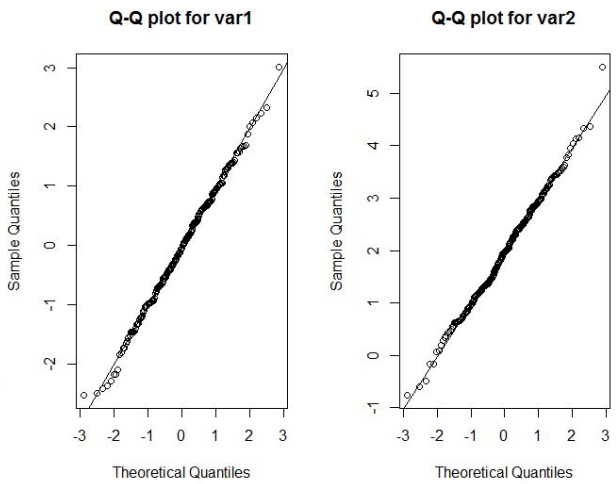

Figure 6: QQ-plots for bivariate samples

Table 3: Shapiro Wilk's test for normality

Random Variable	W	P-value
Variable1	0.997	0.938
Variable2	0.996	0.774
Jointly multivariate normality	0.995	0.667

To assess bivariate normality, a scatter plot of both variables done as well as a gamma plot which was done by plotting the squared mahalanobis distance against chi-square quantiles. The scatter plot of both variables in figure 7 shows an ellipsoidal plot though there seems to be possible outliers but we could not base a conclusion on the contours. Instead, we compared the mahalanobis square distance with the chi-square distribution and the multivariates gamma plot in figure 8 shows very few points deviating from the straight line of the gamma plot which may indicate a bivariate normality.

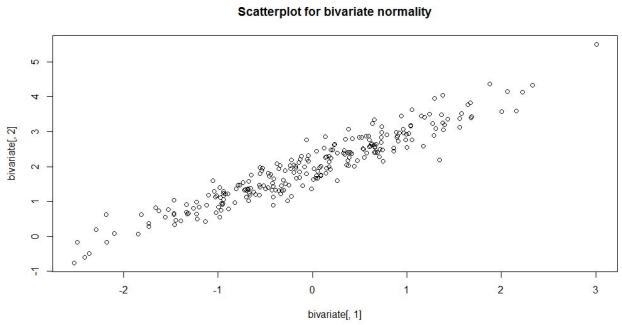

Figure 7: Scatter bivaraite normality

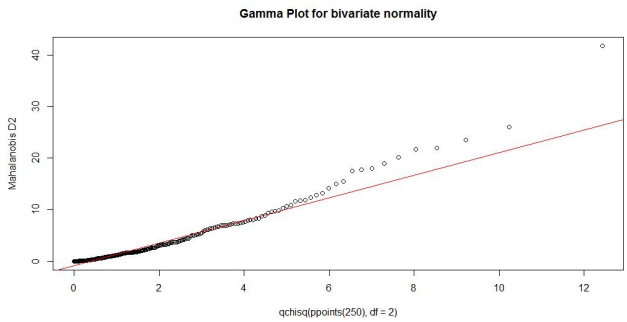

Figure 8: Multivariate gamma plot

3.4 Other Multivariate Normality techniques

Generally in literature, there are many tests to assess multivariate normality (Mecklin and Mundfrom, 2003) but Henze-Zirkler (1990) and Royston (1982) tests are among the best and most popular methods used by many Researchers.

Henze and Zirkler (1990) introduce a multivariate version of the univariate. This test is based on a functional distance that measures the distance between two distribution functions. If the data is multivariate normal, the test statistic HZ is approximately lognormally distributed. It proceeds to calculate the mean, variance and smoothness parameter. Then, mean and variance are lognormalized and the p-value is estimated. Usually, this test has some desirable strong points such as consistency against each fixed nonnormal alternative distribution, asymptotic power against contiguous alternatives of order $n^{-1/2}$, feasibility for any dimension and any sample size and affine invariance. As a weakness the Henze-Zirkler test statistic, like other test statictics does not help in indicating the reason for the rejection of normality, a test rejection should be complemented with graphical procedures such as a chi-square plot and multivariate descriptive statistics such as Mardia's skewness and kurtosis to arrive at the right conclusion.

The Shapiro-Wilk test (Shapiro and Wilk, 1965), is generally considered to be an excellent test of univariate normality. It is only natural to extend it to the multivariate case, as done by Royston (1982). Royston's (1983) marginal method first tests each of the variates for univariate normality with a Shapiro-Wilk statistic, then combines all the dependent tests into one omnibus test statistic for multivariate normality. Royston transforms the Shapiro-Wilk statistics into an approximate Chi-squared random variable, with degrees of freedom estimated by taking into account possible correlation structures between the original test statistics. As a strong point, simulation results have shown that Royston's test has very good Type I error control and power against many different alternative distributions. Further, Royston's test involves a rather ingenious correction for the correlation between the variables in the sample. As weakness, Royston's test has been found to behave well when the sample size is small and the variates are relatively uncorrelated (Mecklin and Mundfrom, 2005).

4 Conclusion

From combining graphical methods and test statistics helped us improved our judgment on the normality of the data. The test of normality for the generated standard normal samples indicated that the samples are normally distributed after been tested univariately. And from the random Cauchy samples, all the samples did not fulfilled the Shapiro-Wilk's test implying these samples were from a non normal population or the cauchy distribution is far from normality due to high skewness as compared to normal.
In the bivariate case, both the marginal and joint multivariate Shapiro-Wilk's test suggested normality and this was also confirmed by the multivariate gamma plot.

References

1. Christel, F. (2015/2016) *Project: Multivariate Data Analysis*. Course notes. University of Hasselt.

2. Henze, N., Zirkler, B. (1990). *A Class of Invariant Consistent Tests for Multivariate Normality* Commun. Statist.-Theor. Meth., vol. 19, no. 10, pp. 3595–3618.

3. Johnson, R. A., Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. 6^{th} edi. Englewood Cliffs:Prentice-Hall.

4. Mecklin, C.J., Mundfrom, D.J. (2005). *Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality*. Journal of Statistical Computation and Simulation, 75, 93 - 107.

5. Royston, J.P. (1983b). *Some techniques for assessing multivariate normality based on the Shapiro-Wilk* . Applied Statistics, 32, 121-133.

6. Thomas, S., Jon W. W., (no date). *Tests for Assessing Multivariate Normality and the Covariance Structure of MIMO Data*. [Online] Brigham Young University. Available: http://www.wireless.groups.et.byu.net/pubs/wallace/icassp_03.pdf [Accessed 12 November 2015]

7. Tong, Y. L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

8. Trujillo, A. (2007). *Royston's Multivariate Normality Test*. [Online] MATLAB release. Available: http://www.mathworks.com/matlabcentral/fileexchange/17811-roystest [Accessed 12 November,2015]

APPENDIX

```
##### Question1(normal)######

set.seed(1234)
dataframe=replicate(4,rnorm(250,0,1))
dataframe1=dataframe
colnames(dataframe1)=c("X1","X2","X3","X4")
str(dataframe1)
summary(dataframe1)
dataframe1

######checking normality
#qq-plot for normal distribution
library(MASS)
par(mfrow=c(1,4))
for(i in 1:4){
dataframe1[,i]
qqnorm(dataframe1[,i],xlab = "Quantiles", ylab = paste("Sample Quantiles",i))
qqline(dataframe1[,i])
}

###Histograms
par(mfrow=c(1,4))
for(i in 1:4){
hist(dataframe1[,i],prob=TRUE,col="cyan",main=paste("histogram of sample",i))
lines(density(dataframe1[,i]),lwd=3,col="red")
}

####Box plots
par(mfrow=c(1,4))
for(i in 1:4){
boxplot(dataframe1[,i],main=paste("boxplot sample",i))
}

#Checking both the univariate and the multivariate normality from the four
standard normal samples using Shapiro test.
library(mvnormtest)
for(i in 1:4){
print(shapiro.test(dataframe1[1:250,i]))
}
mshapiro.test(t(dataframe1[1:250,1:4]))

#####skewness and kurtosis
library(moments)
print(skewness(dataframe1[1:250,1:4]))
print(kurtosis(dataframe1[1:250,1:4]))
```

```
########Q2: CAUCHY VARIATES ########################################
set.seed(1234)
Cauchy=data.frame(replicate(4, rcauchy(250, scale=1, location=0)))
Cauchy
summary(Cauchy)

######checking normality
#qq-plot for normal distribution
library(MASS)
par(mfrow=c(1,4))
for(i in 1:4){
dataframe1[,i]
qqnorm(Cauchy[,i],xlab = "Quantiles", ylab = paste("Sample Quantiles",i))
qqline(Cauchy[,i])
}

###Shapiro test###
for(i in 1:4){
print(shapiro.test(Cauchy[1:250,i]))
}
mshapiro.test(t(Cauchy[1:250,1:4]))


#####Q3: Bivariate random sample#####
##To check normality the "mvnormtest" package should be installed then we can use it
for checking normality##

library(mvnormtest)
library(MASS)
set.seed(5238)
mu<- matrix(c(0,2), nrow =2, ncol = 1, byrow = TRUE, dimnames = list(c("mu1", "mu2"),
c("means")))
sigma<- matrix(c(1.0,0.95,0.95,1.0), nrow = 2, ncol = 2,byrow = FALSE,
dimnames = NULL)
bivariate<-mvrnorm(n =250, mu, sigma, tol = 1e-1, empirical = FALSE)
bivariate<-as.matrix(data.frame(bivariate))
summary(bivariate[,1])
summary(bivariate[,2])

###qqplot
par(mfrow=c(1,2))
qqnorm(bivariate[,1],main="Q-Q plot for var1")
qqline(bivariate[,1])
qqnorm(bivariate[,2],main="Q-Q plot for var2")
qqline(bivariate[,2])
```

```
###Histograms
par(mfrow=c(1,2))
for(i in 1:2){
hist(bivariate[,i],prob=TRUE,col="cyan",main=paste("histogram of var",i))
lines(density(bivariate[,i]),lwd=3,col="red")
}

#checking normality for each variates univariately and multivariately
for(i in 1:2){
  print(shapiro.test(bivariate[,i]))
}
mshapiro.test (t(bivariate[1:250,1:2]))

###Scatter plot###
plot(bivariate[,1],bivariate[,2],main="Scatterplot for bivariate normality")

#creating gamma plot
require(graphics)
covariance <-cov(bivariate,method = c("pearson"))
d2 <- mahalanobis(bivariate, colMeans(bivariate), covariance, inverted=TRUE)
qqplot(qchisq(ppoints(250),df=2),d2, main="Gamma Plot for bivariate normality",
ylab="Mahalanobis D2")
abline(a=-0.9,b=2.2, col='red')
```