MVN package: Multivariate Normality Tests

Selcuk Korkmaz[1], Dincer Goksuluk

Hacettepe University, Faculty of Medicine, Department of Biostatistics, Ankara, TURKEY
[1]selcuk.korkmaz@hacettepe.edu.tr

MVN version 2.0 (Last revision 2014-03-02)

Abstract

Assessing the assumption of multivariate normality is required by many parametric multivariate statistical methods, such as discriminant analysis, principal component analysis, MANOVA, etc. Here, we present an R package to asses multivariate normality. The MVN package contains three most widely used multivariate normality tests, including Mardia's, Henze-Zirkler's and Royston's multivariate normality tests.

Contents

1 Preparation of input data

`MVN` package expects a numeric matrix or a data frame that contains minimum two variables. In this vignette, we will work with the `Iris` data set. This data set is a multivariate data set introduced by Ronald A. Fisher (1936) as an application of discriminant analysis [1]. It is also called Anderson's Iris data set because Edgar Anderson collected the data to measure the morphologic variation of Iris flowers of three related species [2]. The data set consists of 50 samples from each of three species of Iris including `setosa`, `virginica` and `versicolor`. For each sample, four variables were measured including the length and the width of the sepals and petals, in centimeters. We will check the multivariate normality of the Iris data set by using three multivariate normality tests, including Mardia's, Royston's and Henze-Zirkler's multivariate normality tests.

First, we can call our data set using `data` function and display it using `head` function as follows:

```
data(iris)
head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

The Iris data is in `data.frame` format which consists of 5 variables (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`) and 150 samples.

```
class(iris)

## [1] "data.frame"

dim(iris)

## [1] 150   5
```

For simplicity, we will work with a subset of the Iris data with first 50 samples and leave out the class label.

```
data = iris[1:50, 1:4]
head(data)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

2 Multivariate Normality Tests

We will introduce three multivariate normality tests below, including Mardia's Multivariate Normality Test, Henze-Zirkler's Multivariate Normality Test and Royston's Multivariate Normality Test.

Before using our multivariate normality tests, we need to load our `MVN` package as follows:

```
library(MVN)
```

2.1 Mardia's Multivariate Normality Test

Mardia's test is based on multivariate extensions of skewness and kurtosis measures [3]. Now, we will check the multivariate normality of the Iris data using `mardia.test` function in the `MVN` package. This function calculates the Mardia's multivariate skewness and kurtosis coefficients as well as their corresponding statistical tests. For large sample size the multivariate skewness is asymptotically distributed as a Chi-square random variable; here it is corrected for small sample size. Likewise, the multivariate kurtosis is distributed as a unit-normal [4–6].

```
result <- mardia.test(data, cov = TRUE, qqplot = FALSE)

##
##  Mardia's Multivariate Normality Test
##
## data:  data
## g1p = 3.0797, skew = 25.6643, p.value.skew = 0.1772, small.skew =
## 27.8597, p.value.small = 0.1128, g2p = 26.5377, kurtosis = 1.2950,
## p.value.kurt = 0.1953
```

Here, `g1p`: Mardia's estimation of multivariate skew, `skew`: Mardia's skew statistic, `p.value.skew`: p-value of skew statistic, `small.skew`: Mardia's small sample skew statistic, `p.value.small`: p-value of small sample skew statistic, `g2p`: Mardia's g2p estimate of multivariate kurtosis, `kurtosis`: Mardia's multivariate kurtosis statistic and `p.value.kurt`: p-value of kurtosis statistic.

As seen above results, both skewness ($p = 0.1772$) and kurtosis ($p = 0.1953$) values indicate multivariate normality.

2.2 Henze-Zirkler's Multivariate Normality Test

The Henze-Zirkler test is based on a non-negative functional distance that measures the distance between two distribution functions. If the data is multivariate normal, the test statistic 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. We can use `HZ.test` function in the `MVN` package to calculate the Henze-Zirkler's Multivariate Normality Test [7–11].

```
result <- HZ.test(data, cov = TRUE, qqplot = FALSE)

##
##  Henze-Zirkler's Multivariate Normality Test
##
```

```
## data:  data
## HZ = 0.9488, p-value = 0.04995
```

Here, `HZ` is the value of Henze-Zirkler statistic at significance level 0.05 and `p-value` is a p-value for the Henze-Zirkler's Multivariate Normality Test.

Since the p-value, which optain from the `HZ.test`, lower than 0.05, one can conclude that this multivariate data set deviates from multivariate normality.

2.3 Royston's Multivariate Normality Test

Royston's H test uses Shapiro-Wilk's W statistic for multivariate normality. However, if kurtosis of the data greater than 3 then `Shapiro-Francia` test is used for leptokurtic samples else `Shapiro-Wilk` test is used for platykurtic samples [10, 12–18].

```
result <- royston.test(data, qqplot = FALSE)

##
##  Royston's Multivariate Normality Test
##
## data:  data
## H = 31.52, p-value = 2.188e-06
```

Here, `H` is the value of Royston's H statistic at significance level `0.05` and `p-value` is an approximate p-value for the test with respect to equivalent degrees of freedom (`edf`).

According to the Royston's Multivariate Normality Test, the Iris data set does not appear to follow a multivariate normal distribution ($p < 0.001$).

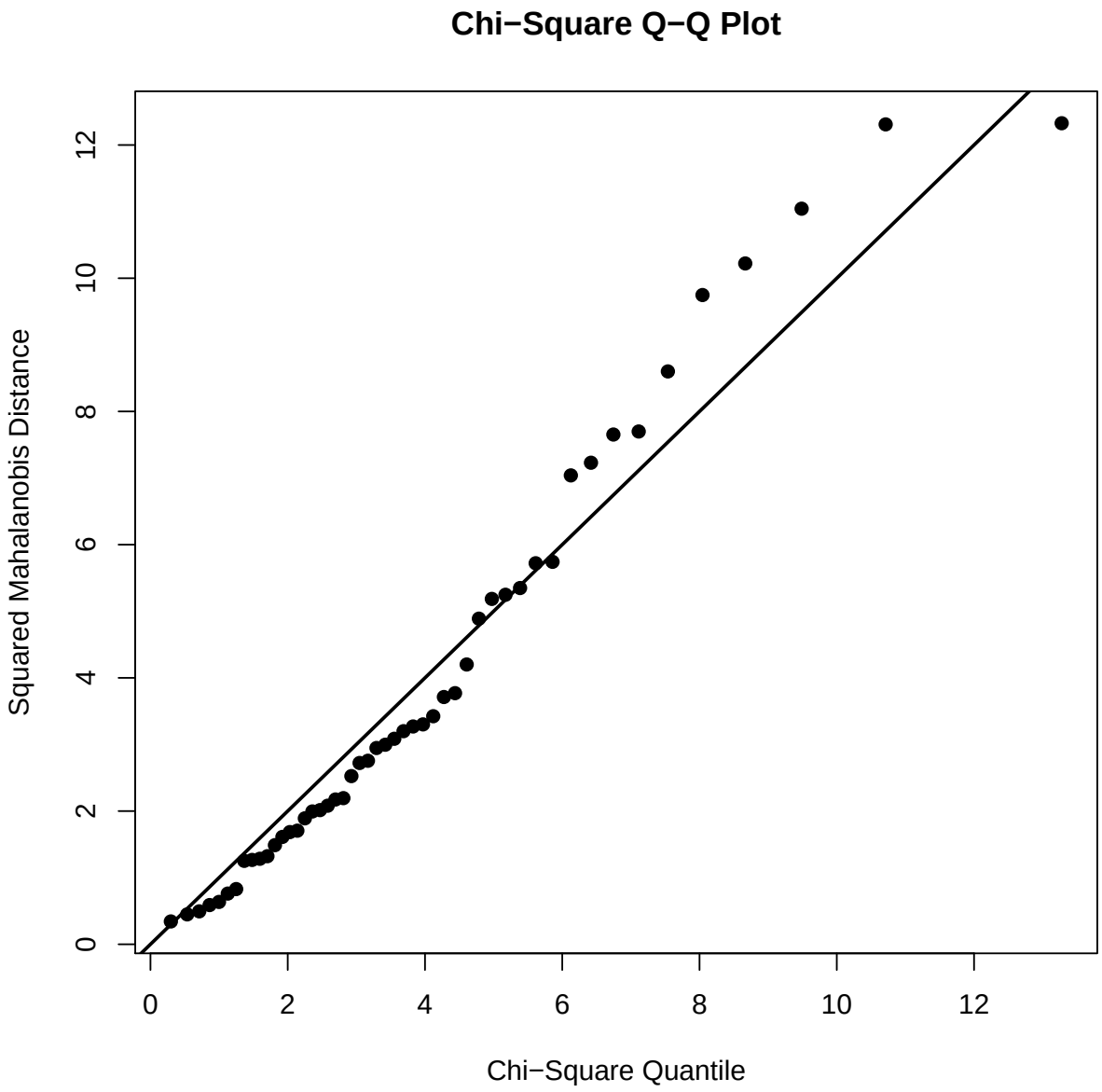
3 Multivariate Normality Plots

Our `MVN` package has ability to draw three multivariate plots. We can use `qqplot = TRUE` option in the `mardia.test`, `HZ.test` and `royston.test` functions to draw a Q-Q plot. Furthermore, we can use `mvn.plot` function in our `MVN` package to draw perspective and contour plots for binary data sets.

3.1 Q-Q Plot

We can draw a Q-Q plot for our Iris data set to see whether there is any deviation from multivariate normality.

```
result <- royston.test(data, qqplot = TRUE)
```




```
##
##  Royston's Multivariate Normality Test
##
## data:  data
## H = 31.52, p-value = 2.188e-06
```

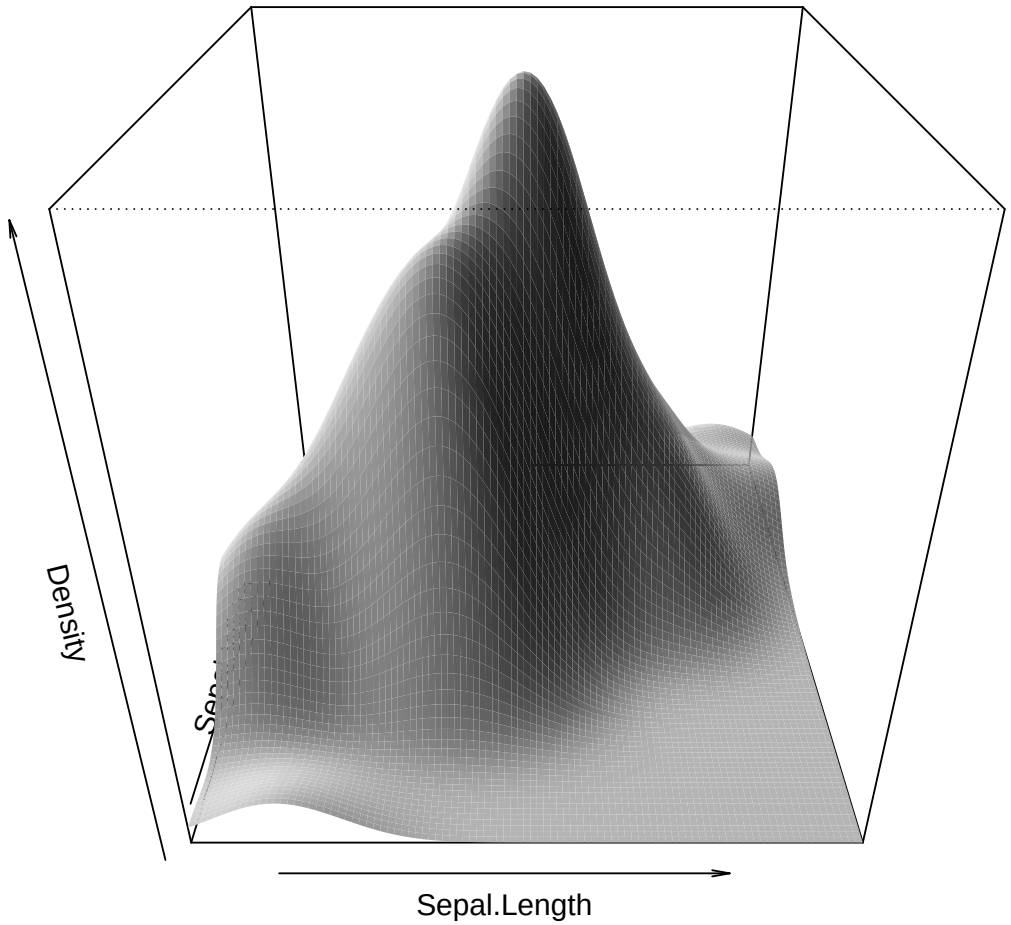
If the data set follows approximately a multivariate normal distribution, the resulting plot should be roughly straight line. As you can see from the Q-Q plot above, there are some deviations from the straight line and this indicates possible departures from a multivariate normal distribution.

3.2 Perspective and Contour Plots

We can use the `mvn.plot` function in the `MVN` package to draw a perspective plot for a binary data set. In order to get a perspective plot, we should continue with two variables, i.e., bivariate normal

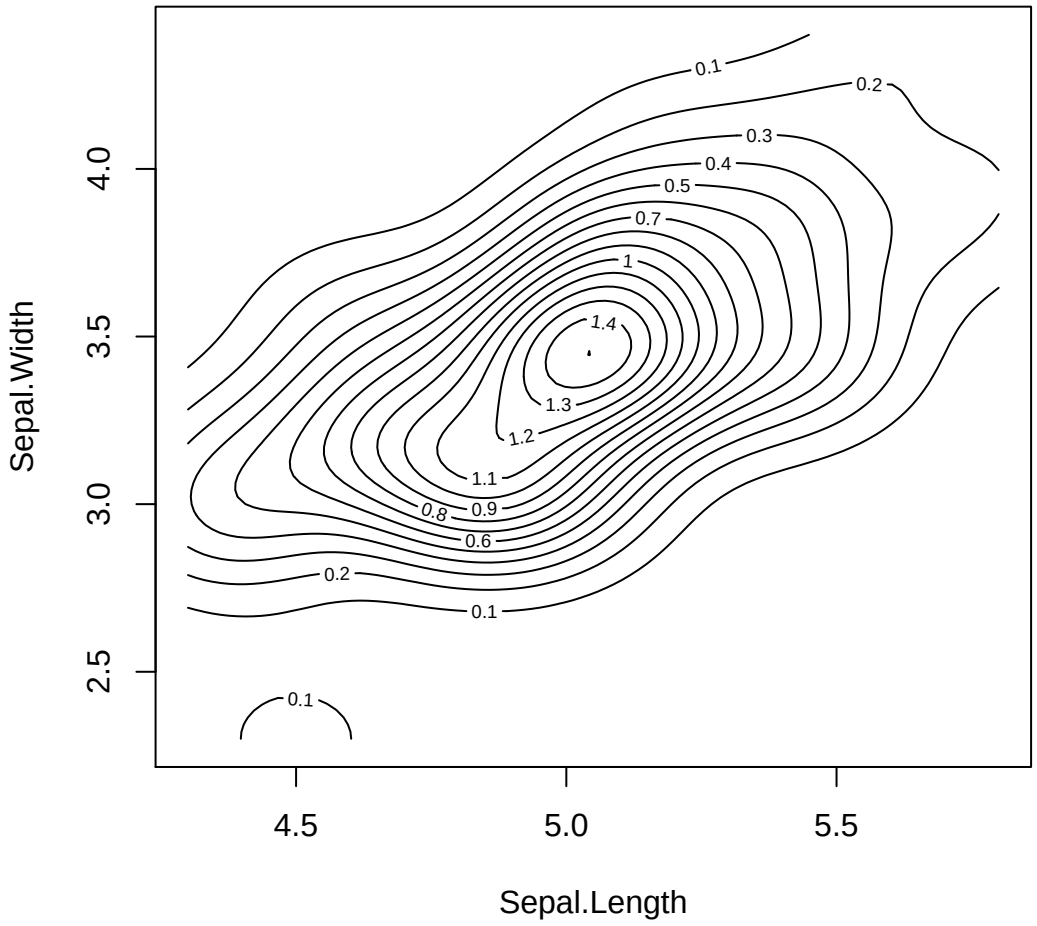
distribution. As an example, we subset first 50 rows and Sepal measures of Iris data. Sepal measures of first 50 samples are bivariate normal. We can see that from the perspective plot. Perspective plot produces 3-dimensional bell-shaped graph when data is bivariate normal.

```
data = iris[1:50, 1:2]
result = HZ.test(data)

##
##  Henze-Zirkler's Multivariate Normality Test
##
## data:  data
## HZ = 0.2856, p-value = 0.9146

mvn.plot(result, type = "persp", default = TRUE)
```



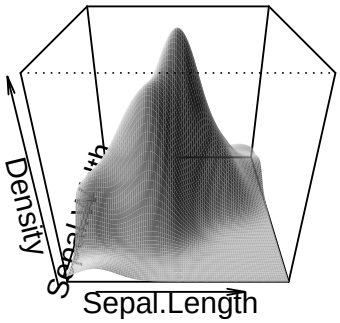
Another alternative is to use 2-dimensional contour graphs. We can use the `mvn.plot` function in the `MVN` package to draw a contour plot for a binary data set. Contour graphs are very useful since it gives information about normality and correlation at the same time. From contour graph below, we can say that there is a positive correlation among sepal measures of flowers since contour lines lie around main diagonal.

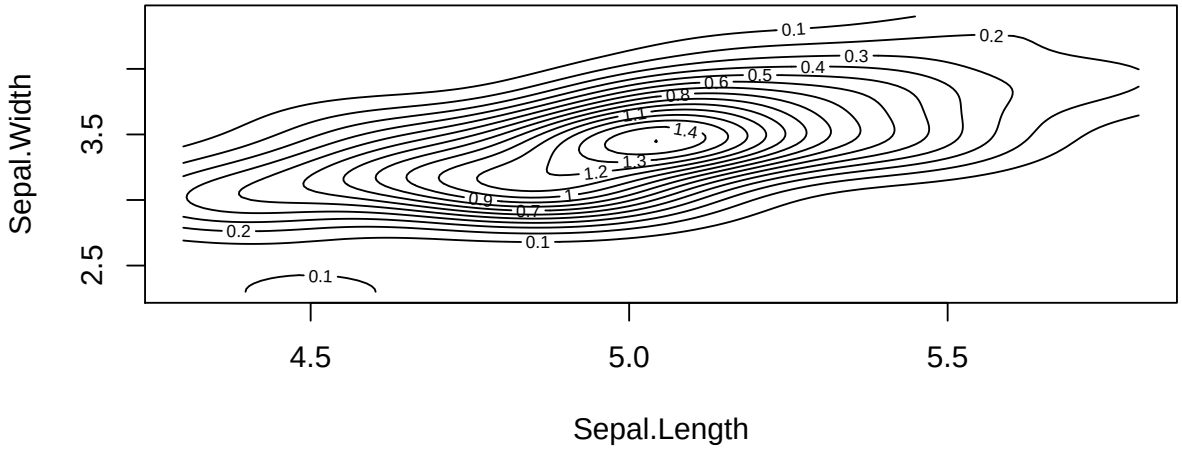
```
data = iris[1:50, 1:2]
result = mardia.test(data)

##
##  Mardia's Multivariate Normality Test
##
## data:  data
## g1p = 0.0911, skew = 0.7595, p.value.skew = 0.9438, small.skew =
## 0.8379, p.value.small = 0.9333, g2p = 8.1057, kurtosis = 0.0935,
## p.value.kurt = 0.9255

mvn.plot(result, type = "contour", default = TRUE)
```



Moreover, we can draw two plots at the same time as follows:

```
mvn.plot(result, type = "both", default = TRUE)
```





Using two graphs together may not be very useful since plot margins for perspective and contour graphs are different. Contour graph uses larger graph area comparing to the perspective plot. However, it might be a better solution to draw graphs separately and combine them manually.

4 Session info

```
sessionInfo()

## R version 3.0.2 (2013-09-25)
## Platform: i386-w64-mingw32/i386 (32-bit)
##
## locale:
## [1] LC_COLLATE=C
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] MVN_2.0       MASS_7.3-29   moments_0.13  nortest_1.0-2 knitr_1.5
##
## loaded via a namespace (and not attached):
## [1] codetools_0.2-8 digest_0.6.4    evaluate_0.5.1  formatR_0.10
## [5] highr_0.3       stringr_0.6.2   tools_3.0.2
```

References

[1] R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics 7 (2): 179-188. doi:10.1111/j.1469-1809.1936.tb02137.x

[2] Edgar Anderson (1936). "The species problem in Iris". Annals of the Missouri Botanical Garden 23 (3): 457-509. JSTOR 2394164.

[3] Mardia, K. V. (1970). "Measures of multivariate skewness and kurtosis with applications". Biometrika 57 (3): 519-530. doi:10.1093/biomet/57.3.519.

[4] Mardia, K. V. (1974), Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. Sankhy A, 36:115-128.

[5] Trujillo-Ortiz, A. and R. Hernandez-Walls. (2003). Mskekur: Mardia's multivariate skewness and kurtosis coefficients and its hypotheses testing. A MATLAB file. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3519

[6] Stevens, J. (1992), Applied Multivariate Statistics for Social Sciences. 2nd. ed. New-Jersey:Lawrance Erlbaum Associates Publishers. pp. 247-248.

[7] Trujillo-Ortiz, A., R. Hernandez-Walls, K. Barba-Rojo and L. Cupul-Magana. (2007). HZmvntest:Henze-Zirkler's Multivariate Normality Test. A MATLAB file. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17931

[8] Henze, N. and Zirkler, B. (1990), A Class of Invariant Consistent Tests for Multivariate Normality. Commun. Statist.-Theor. Meth., 19(10): 35953618.

[9] Henze, N. and Wagner, Th. (1997), A New Approach to the BHEP tests for multivariate normality. Journal of Multivariate Analysis, 62:1-23.

[10] Johnson, R.A. and Wichern, D. W. (1992). Applied Multivariate Statistical Analysis. 3rd. ed. New-Jersey:Prentice Hall.

[11] Mecklin, C. J. and Mundfrom, D. J. (2003), On Using Asymptotic Critical Values in Testing for Multivariate Normality.

[12] Mecklin, C.J. and Mundfrom, D.J. (2005). A 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.

[13] Royston, J.P. (1982). An Extension of Shapiro and Wilks W Test for Normality to Large Samples. Applied Statistics, 31(2):115124.

[14] Royston, J.P. (1983). Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W. Applied Statistics, 32(2).

[15] Royston, J.P. (1992). Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and Computing, 2:117-119.121133.

[16] Royston, J.P. (1995). Remark AS R94: A remark on Algorithm AS 181: The W test for normality. Applied Statistics, 44:547-551.

[17] Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality. Biometrika, 52:591611.

[18] Trujillo-Ortiz, A., R. Hernandez-Walls, K. Barba-Rojo and L. Cupul-Magana. (2007). Roystest:Royston's Multivariate Normality Test. A MATLAB file. URL http://www.mathworks.com/matlabcentral/fileexchange/17811