MVN: An R Package for Assessing Multivariate Normality

Selcuk Korkmaz[1], Dincer Goksuluk and Gokmen Zararsiz

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

MVN version 4.0 (Last revision 2015-01-28)

Abstract

Assessing the assumption of multivariate normality is required by many parametric multivariate statistical methods, such as MANOVA, linear discriminant analysis, principal component analysis, canonical correlation, etc. It is important to assess multivariate normality in order to proceed with such statistical methods. There are many analytical methods proposed for checking multivariate normality. However, deciding which method to use is a challenging process, since each method may give different results under certain conditions. Hence, we may say that there is no best method, which is valid under any condition, for normality checking. In addition to numerical results, it is very useful to use graphical methods to decide on multivariate normality. Combining the numerical results from several methods with graphical approaches can be useful and provide more reliable decisions. Here, we present an R package, **MVN**, to assess multivariate normality. It contains the three most widely used multivariate normality tests, including Mardia's, Henze-Zirkler's and Royston's, and graphical approaches, including chi-square Q-Q, perspective and contour plots. It also includes two multivariate outlier detection methods, which are based on robust Mahalanobis distances. Moreover, this package offers functions to check the univariate normality of marginal distributions through both tests and plots. Furthermore, especially for non-R users, we provide a user-friendly web application of the package. This application is available at `http://www.biosoft.hacettepe.edu.tr/MVN/`.

1 Introduction

Many multivariate statistical analysis methods, such as MANOVA and linear discriminant analysis (**MASS**, [1]), principal component analysis (**FactoMineR**, [2], **psych**, [3]), canonical correlation (**CCA**, [4]), etc., require multivariate normality (MVN) assumption. If the data are multivariate normal (exactly or approximately), such multivariate methods provide more reliable results. The performance of these methods dramatically decreases if the data are not multivariate normal. Hence, researchers should check whether data are multivariate normal or not before continuing with such parametric multivariate analyses.

Many statistical tests and graphical approaches are available to check the multivariate normality assumption. Burdenski (2000) reviewed several statistical and practical approaches, including the Q-Q plot, box-plot, stem and leaf plot, Shapiro-Wilk and Kolmogorov-Smirnov tests to evaluate the univariate normality, contour and perspective plots for assessing bivariate normality, and the chi-square Q-Q plot to check the multivariate normality [5]. The author demonstrated each procedure using the real data from [6]. Ramzan et al. (2013) reviewed numerous graphical methods for assessing both univariate and multivariate normality and showed their use in a real life problem to check the MVN using chi-square and beta Q-Q plots [7]. Holgersson (2006) stated the importance of graphical procedures and presented a simple graphical tool, which is based on the scatter plot

of two correlated variables to assess whether the data belong to a multivariate normal distribution or not [8]. Svantesson and Wallace (2003) applied Royston's and Henze-Zirkler's tests to multiple-input multiple-output data to test MVN [9]. According to the review by Mecklin and Mundfrom (2005), more than fifty statistical methods are available for testing MVN [10]. They conducted a comprehensive simulation study based on type I and type II error and concluded that no single test excelled in all situations. The authors suggested using Henze-Zirkler's and Royston's tests among others for assessing MVN because of their good type I error control and power. Moreover, to diagnose the reason for deviation from multivariate normality, the authors suggested the use of Mardia's multivariate skewness and kurtosis statistics test as well as graphical approaches such as the chi-square Q-Q plot. Deciding which test to use can be a daunting task for researchers (mostly for non-statisticians) and it is very useful to perform several tests and examine the graphical methods simultaneously. Although there are a number of studies describing multifarious approaches, there is no single easy-to-use, up-to-date and comprehensive tool to apply various statistical tests and graphical methods together at present.

In this vignette, we introduce an R package, **MVN**, which implements the three most widely used MVN tests, including Mardia's, Henze-Zirkler's, and Royston's [11]. In addition to statistical tests, the **MVN** also provides some graphical approaches such as chi-square Q-Q, perspective and contour plots. Moreover, this package includes two multivariate outlier detection methods, which are based on Mahalanobis distance. In addition to multivariate normality, users can also check univariate normality tests and plots to diagnose the deviation from normality via package version 3.7 and later. Firstly, we discuss the theoretical background on the corresponding MVN tests. Secondly, two illustrative examples are presented in order to demonstrate the applicability of the package. Finally, we present a newly developed web interface of the **MVN**, which can be especially handy for non-R users. The R version of the **MVN** is publicly available in the Comprehensive R Archive Network (CRAN, `http://CRAN.R-project.org/package=MVN`).

2 Multivariate normality tests

2.1 Mardia's MVN test

Mardia (1970) proposed a multivariate normality test which is based on multivariate extensions of skewness ($\hat{\gamma}_{1,p}$) and kurtosis ($\hat{\gamma}_{2,p}$) measures as follows [12]:

$$\hat{\gamma}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3 \qquad \text{and} \qquad \hat{\gamma}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} m_{ii}^2 \tag{1}$$

where $m_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x})$, i.e the squared Mahalanobis distance, and p is the number of variables. The test statistic for skewness, $(n/6)\hat{\gamma}_{1,p}$, is approximately χ^2 distributed with $p(p+1)(p+2)/6$ degrees of freedom. Similarly, the test statistic for kurtosis, $\hat{\gamma}_{2,p}$, is approximately normally distributed with mean $p(p+2)$ and variance $8p(p+2)/n$.

For small samples, the power and the type I error could be violated. Therefore, Mardia (1974) introduced a correction term into the skewness test statistic, usually when $n < 20$, in order to control type I error [13]. The corrected skewness statistic for small samples is $(nk/6)\hat{\gamma}_{1,p}$, where $k = (p+1)(n+1)(n+3)/(n(n+1)(p+1)-6)$. This statistic is also distributed as χ^2 with degrees of freedom $p(p+1)(p+2)/6$.

2.2 Henze-Zirkler's MVN test

The Henze-Zirkler's test is based on a non-negative functional distance that measures the distance between two distribution functions. If data are distributed as multivariate normal, the test statistic

is approximately log-normally distributed. First, the mean, variance and smoothness parameter are calculated. Then, the mean and the variance are log-normalized and the p-value is estimated [14–18]. The test statistic of Henze-Zirkler's multivariate normality test is given in equation 2.

$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}} - 2\left(1+\beta^2\right)^{-\frac{p}{2}}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2\left(1+\beta^2\right)}D_i} + n\left(1+2\beta^2\right)^{-\frac{p}{2}} \tag{2}$$

where

$$\begin{aligned} p: &\quad \text{number of variables} \\ \beta &= \frac{1}{\sqrt{2}}\left(\frac{n\,(2p+1)}{4}\right)^{\frac{1}{p+4}} \\ D_{ij} &= (x_i - x_j)^{'} S^{-1}(x_i - x_j) \\ D_i &= (x_i - \bar{x})^{'} S^{-1}(x_i - \bar{x}) = m_{ii} \end{aligned}$$

From equation 2, D_i gives the squared Mahalanobis distance of i^{th} observation to the centroid and D_{ij} gives the Mahalanobis distance between i^{th} and j^{th} observations. If data are multivariate normal, the test statistic (HZ) is approximately log-normally distributed with mean μ and variance σ^2 as given below:

$$\begin{aligned} \mu =& 1 - \frac{a^{-\frac{p}{2}}\left(1 + p\beta^{\frac{2}{a}} + \left(p\,(p+2)\,\beta^4\right)\right)}{2a^2} \\ \sigma^2 =& 2\left(1+4\beta^2\right)^{-\frac{p}{2}} + \frac{2a^{-p}\left(1+2p\beta^4\right)}{a^2} + \frac{3p\,(p+2)\,\beta^8}{4a^4} \\ & - 4{w_\beta}^{-\frac{p}{2}}\left(1 + \frac{3p\beta^4}{2w_\beta} + \frac{p\,(p+2)\,\beta^8}{2{w_\beta}^2}\right) \end{aligned}$$

where $a = 1+2\beta^2$ and $w_\beta = (1+\beta^2)(1+3\beta^2)$. Hence, the log-normalized mean and variance of the HZ statistic can be defined as follows:

$$\log(\mu) = \log\left(\sqrt{\frac{\mu^4}{\sigma^2+\mu^2}}\right) \qquad \text{and} \qquad \log\left(\sigma^2\right) = \log\left(\frac{\sigma^2+\mu^2}{\sigma^2}\right) \tag{3}$$

By using the log-normal distribution parameters, μ and σ, we can test the significance of multivariate normality. The `Wald` test statistic for multivariate normality is given in equation 4.

$$z = \frac{\log(HZ) - \log(\mu)}{\log(\sigma)} \tag{4}$$

2.3 Royston's MVN test

Royston's test uses the `Shapiro-Wilk`/`Shapiro-Francia` statistic to test multivariate normality. If kurtosis of the data is greater than 3, then it uses the `Shapiro-Francia` test for leptokurtic distributions, otherwise it uses the `Shapiro-Wilk` test for platykurtic distributions [10, 15, 19–23].

Let W_j be the `Shapiro-Wilk`/`Shapiro-Francia` test statistic for the j^{th} variable ($j = 1, 2, \ldots, p$) and Z_j be the values obtained from the normality transformation proposed by [22].

$$\begin{array}{llll} \text{if } 4 \le n \le 11; & x = n & \text{and} & w_j = -\log\left[\gamma - \log\left(1 - W_j\right)\right] \\ \text{if } 12 \le n \le 2000; & x = \log(n) & \text{and} & w_j = \log\left(1 - W_j\right) \end{array} \tag{5}$$

As seen from equation 5, x and w_j's change with the sample size (n). By using equation 5, transformed values of each random variable can be obtained from equation 6.

$$Z_j = \frac{w_j - \mu}{\sigma} \tag{6}$$

where γ, μ and σ are derived from the polynomial approximations given in equation 7. The polynomial coefficients are provided by [22] for different sample sizes.

$$\begin{aligned} \gamma &= a_{0\gamma} + a_{1\gamma}x + a_{2\gamma}x^2 + \cdots + a_{d\gamma}x^d \\ \mu &= a_{0\mu} + a_{1\mu}x + a_{2\mu}x^2 + \cdots + a_{d\mu}x^d \\ \log(\sigma) &= a_{0\sigma} + a_{1\sigma}x + a_{2\sigma}x^2 + \cdots + a_{d\sigma}x^d \end{aligned} \tag{7}$$

The Royston's test statistic for multivariate normality as follows:

$$H = \frac{e\sum_{j=1}^{p}\psi_j}{p} \quad \sim \quad \chi_e^2 \tag{8}$$

where e is the equivalent degrees of freedom (edf) and $\Phi(.)$ is the cumulative distribution function for standard normal distribution such that,

$$\begin{aligned} e &= p/[1 + (p-1)\bar{c}] \\ \psi_j &= \left\{\Phi^{-1}\left[\Phi(-Z_j)/2\right]\right\}^2, \qquad j = 1, 2, \ldots, p. \end{aligned} \tag{9}$$

As seen from equation 9, another extra term $\bar{c}$ has to be calculated in order to continue with the statistical significance of Royston's test statistic given in equation 8. Let R be the correlation matrix and r_{ij} be the correlation between i^{th} and j^{th} variables. Then, the extra term $\bar{c}$ can be found by using equation 10.

$$\bar{c} = \sum_i \sum_j \frac{c_{ij}}{p(p-1)}, \qquad \{c_{ij}\}_{i \neq j} \tag{10}$$

where

$$c_{ij} = \begin{cases} g(r_{ij}, n) & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

with the boundaries of $g(.)$ as $g(0, n) = 0$ and $g(1, n) = 1$. The function $g(.)$ is defined as follows:

$$g(r, n) = r^{\lambda}\left[1 - \frac{\mu}{\nu}(1 - r)^{\mu}\right].$$

The unknown parameters, μ, λ and ν were estimated from a simulation study conducted by [24]. He found $\mu = 0.715$ and $\lambda = 5$ for sample size $10 \leq n \leq 2000$ and ν is a cubic function which can be obtained as follows:

$$\nu(n) = 0.21364 + 0.015124x^2 - 0.0018034x^3$$

where $x = \log(n)$.

3 Implementation of MVN package

The **MVN** package contains several functions in the `S4` class. The data to be analyzed should be given in the `"data.frame"` or `"matrix"` class. In this example, we will work with the famous `Iris` data set. These data are from a multivariate data set introduced by Fisher (1936) as an application of linear discriminant analysis [25]. 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 [26]. First of all, the **MVN** library should be loaded in order to use related functions.

```
# load MVN package
library(MVN)
```

Similarly, `Iris` data can be loaded from the R database by using the following R code:

```
# load Iris data
data(iris)
```

The `Iris` data set consists of 150 samples from each of the three species of Iris including `setosa`, `virginica` and `versicolor`. For each sample, four variables were measured including the length and width of the `sepals` and `petals`, in centimeters.

Example I: For simplicity, we will work with a subset of these data which contain only 50 samples of `setosa` flowers, and check MVN assumption using Mardia's, Royston's and Henze-Zirkler's tests.

```
# setosa subset of the Iris data
setosa <- iris[1:50, 1:4]
```

3.1 Mardia's MVN test: mardiaTest(...)

The `mardiaTest` function is used to calculate the Mardia's multivariate skewness and kurtosis coefficients as well as their corresponding statistical significance. This function can also calculate the corrected version of the skewness coefficient for small sample size ($n < 20$).

```
result <- mardiaTest(setosa, qqplot = FALSE)
result

##    Mardia's Multivariate Normality Test
## ---------------------------------------
##    data : setosa
##
##    g1p            : 3.08
##    chi.skew       : 25.66
##    p.value.skew   : 0.1772
##
##    g2p            : 26.54
##    z.kurtosis     : 1.295
##    p.value.kurt   : 0.1953
##
##    chi.small.skew : 27.86
```

```
##    p.value.small  : 0.1128
##
##    Result         : Data are multivariate normal.
## -------------------------------------
```

Here:

`g1p`: Mardia's estimation of multivariate skewness, i.e $\hat{\gamma}_{1,p}$ given in equation 1,
`chi.skew`: test statistic for multivariate skewness,
`p.value.skew`: significance value of skewness statistic,
`g2p`: Mardia's estimation of multivariate kurtosis, i.e $\hat{\gamma}_{2,p}$ given in equation 1,
`z.kurtosis`: test statistic for multivariate kurtosis,
`p.value.kurt`: significance value of kurtosis statistic,
`chi.small.skew`: test statistic for multivariate skewness with small sample correction,
`p.value.small`: significance value of small sample skewness statistic.

As seen from the results given above, both the skewness ($\hat{\gamma}_{1,p} = 3.0797, p = 0.1772$) and kurtosis ($\hat{\gamma}_{2,p} = 26.5377, p = 0.1953$) estimates indicate multivariate normality. Therefore, according to Mardia's MVN test, this data set follows a multivariate normal distribution.

3.2 Henze-Zirkler's MVN test: hzTest(...)

One may use the `hzTest` function in the **MVN** to perform the Henze-Zirkler's test.

```
result <- hzTest(setosa, qqplot = FALSE)
result

##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : setosa
##
##   HZ      : 0.9488
##   p-value : 0.04995
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

Here, `HZ` is the value of the Henze-Zirkler's test statistic at significance level 0.05 and `p-value` is the significance value of this test statistic, i.e the significance of multivariate normality. Since the p-value, which is derived from `hzTest`, is mathematically lower than 0.05, one can conclude that this multivariate data set deviates slightly from multivariate normality ($HZ = 0.9488$, $p = 0.05$). Since the p-value is very close to 0.05, researchers should also check the multivariate graphical approaches as well as univariate tests and plots to make a more reliable decision on multivariate normality.

3.3 Royston's MVN test: roystonTest(...)

In order to carry out the Royston's test, `roystonTest` function in the **MVN** can be used as follows:

```
##   Royston's Multivariate Normality Test
## --------------------------------------------
##   data : setosa
##
##   H       : 31.52
##   p-value : 2.188e-06
##
##   Result  : Data are not multivariate normal.
## --------------------------------------------
```

Here, `H` is the value of the Royston's test statistic at significance level 0.05 and `p-value` is an approximate significance value for the test with respect to edf. According to Royston's test, the `setosa` data set does not appear to follow a multivariate normal distribution ($H = 31.518$, $p < 0.001$).

3.4 Chi-square Q-Q plot

One can clearly see that different MVN tests may come up with different results. MVN assumption was rejected by Henze-Zirkler's and Royston's tests; however, it was not rejected by Mardia's test at a significance level of 0.05. In such cases, examining MVN plots along with hypothesis tests can be quite useful in order to reach a more reliable decision.

The Q-Q plot, where "Q" stands for quantile, is a widely used graphical approach to evaluate the agreement between two probability distributions. Each axis refers to the quantiles of probability distributions to be compared, where one of the axes indicates theoretical quantiles (hypothesized quantiles) and the other indicates the observed quantiles. If the observed data fit hypothesized distribution, the points in the Q-Q plot will approximately lie on the line $y = x$.

MVN has the ability to create three multivariate plots. One may use the `qqplot = TRUE` option in the `mardiaTest`, `hzTest` and `roystonTest` functions to create a chi-square Q-Q plot. We can create this plot for the `setosa` data set to see whether there are any deviations from multivariate normality. Figure 1 shows the chi-square Q-Q plot of the first 50 rows of `Iris` data, which are `setosa` flowers. It can be seen from Figure 1 that there are some deviations from the straight line and this indicates possible departures from a multivariate normal distribution.

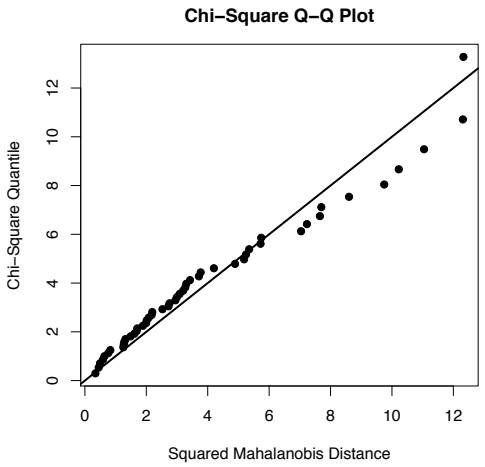


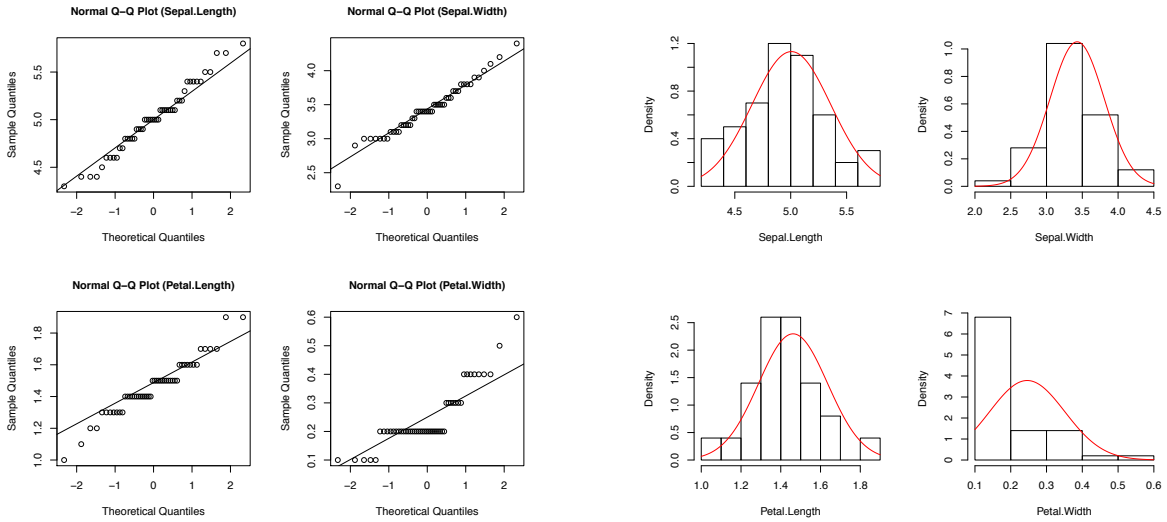
Figure 1: Chi-Square Q-Q plot for `setosa` data set.

As a result, we can conclude that this data set does not satisfy MVN assumption based on the fact that the two test results are against it and the chi-square Q-Q plot indicates departures from multivariate normal distribution.

3.5 Univariate plots and tests

As noted by several authors [5, 27, 28], if data have a multivariate normal distribution, then, each of the variables has a univariate normal distribution; but the opposite does not have to be true. Hence, checking univariate plots and tests could be very useful to diagnose the reason for deviation from MVN. We can check this assumption through `uniPlot` and `uniNorm` functions from the package. The `uniPlot` function is used to create univariate plots, such as Q-Q plots (Figure 2a), histograms with normal curves (Figure 2b), box-plots and scatterplot matrices.

```
uniPlot(setosa, type = "qqplot") # creates univariate Q-Q plots
uniPlot(setosa, type = "histogram") # creates univariate histograms
```



(a) Q-Q plots.

(b) Histograms with normal curves.

Figure 2: Univariate plots of `setosa`.

As seen from Figure 2, `Petal.Width` has a right-skewed distribution whereas other variables have approximately normal distributions. Thus, we can conclude that problems with multivariate normality arise from the skewed distribution of `Petal.Width`. In addition to the univariate plots, one can also perform univariate normality tests using the `uniNorm` function. It provides several widely used univariate normality tests, including Shapiro-Wilk, Cramer-von Mises, Lilliefors and Anderson-Darling. For example, the following code chunk is used to perform the Shapiro-Wilk's normality test on each variable and it also displays descriptive statistics including mean, standard deviation, median, minimum, maximum, 25th and 75th percentiles, skewness and kurtosis:

```
uniNorm(setosa, type = "SW", desc = TRUE)
```

```
## $`Descriptive Statistics`
##               n  Mean Std.Dev Median Min Max 25th  75th  Skew Kurtosis
## Sepal.Length 50 5.006   0.352    5.0 4.3 5.8  4.8 5.200 0.113   -0.451
## Sepal.Width  50 3.428   0.379    3.4 2.3 4.4  3.2 3.675 0.039    0.596
## Petal.Length 50 1.462   0.174    1.5 1.0 1.9  1.4 1.575 0.100    0.654
## Petal.Width  50 0.246   0.105    0.2 0.1 0.6  0.2 0.300 1.180    1.259
##
## $`Shapiro-Wilk's Normality Test`
##       Variable Statistic   p-value Normality
## 1 Sepal.Length    0.9777    0.4595    YES
## 2 Sepal.Width     0.9717    0.2715    YES
## 3 Petal.Length    0.9550    0.0548    YES
## 4 Petal.Width     0.7998    0.0000    NO
```

From the above results, we can see that all variables, except `Petal.Width` in the `setosa` data set, have univariate normal distributions at significance level 0.05. We can now drop `Petal.With` from `setosa` data and recheck the multivariate normality. MVN results are given in Table 1.

Test	Test Statistic	p-value
Mardia		
Skewness	11.249	0.338
Kurtosis	1.287	0.198
Henze-Zirkler	0.524	0.831
Royston	7.255	0.060

Table 1: MVN test results (`setosa` without `Petal.Width`).

According to the three MVN test results in Table 1, `setosa` without `Petal.Width` has a multivariate normal distribution at significance level 0.05.

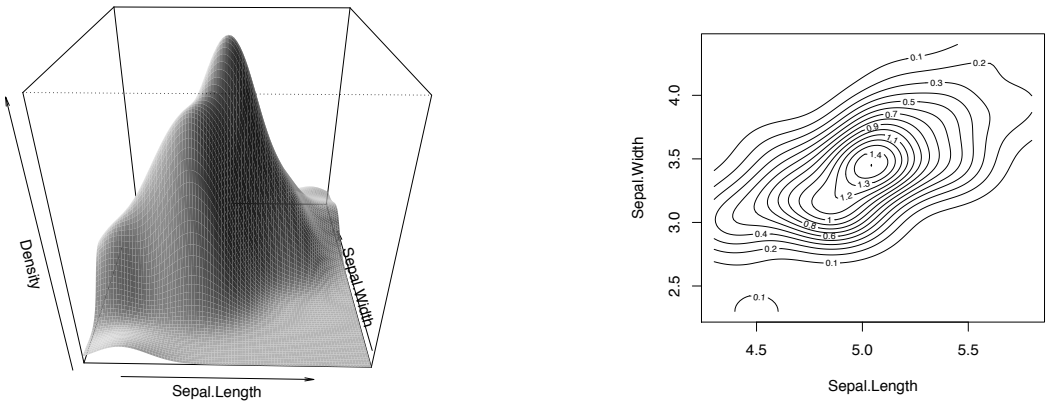
Example II: Whilst the Q-Q plot is a general approach for assessing MVN in all types of numerical multivariate datasets, perspective and contour plots can only be used for bivariate data. To demonstrate the applicability of these two approaches, we will use a subset of `Iris` data, named `setosa2`, including the `sepal length` and `sepal width` variables of the `setosa` species.

3.6 Perspective and contour plots

Univariate normal marginal densities are a necessary but not a sufficient condition for MVN. Hence, in addition to univariate plots, creating perspective and contour plots will be useful. The perspective plot is an extension of the univariate probability distribution curve into a 3·dimensional probability distribution surface related with bivariate distributions. It also gives information about where data are gathered and how two variables are correlated with each other. It consists of three dimensions where two dimensions refer to the values of the two variables and the third dimension, which is likely in univariate cases, is the value of the multivariate normal probability density function. Another alternative graph, which is called the "contour plot", involves the projection of the perspective plot into a 2·dimensional space and this can be used for checking multivariate normality assumption. For bivariate normally distributed data, we expect to obtain a three-dimensional bell-shaped graph from the perspective plot. Similarly, in the contour plot, we can observe a similar pattern.

To construct a perspective and contour plot for Example 2, we can use the `mvnPlot` function in the **MVN**. This function requires an object in the "MVN" class that is one of the results from **MVN** functions. In the following codes, the object from `hzTest` is used for the perspective plot given in Figure 3a. It is also possible to create a contour plot of the data. Contour graphs are very useful since they give information about normality and correlation at the same time. Figure 3b shows the contour plot of `setosa` flowers. As can be seen from the graph, this is simply a top view of the perspective plot where the third dimension is represented with ellipsoid contour lines. From this graph, we can say that there is a positive correlation among the `sepal` measures of flowers since the contour lines lie around the main diagonal. If the correlation were zero, the contour lines would be circular rather than ellipsoid.

```
setosa2 <- iris[1:50, 1:2]
result <- hzTest(setosa2, qqplot=FALSE)
mvnPlot(result, type = "persp", default = TRUE) # perspective plot
mvnPlot(result, type = "contour", default = TRUE) # contour plot
```



(a) Perspective plot

(b) Contour plot

Figure 3: Perspective and contour plot for bivariate `setosa2` data set.

Since neither the univariate plots in Figure 2 nor the multivariate plots in Figure 3 show any significant deviation from MVN, we can now perform the MVN tests to evaluate the statistical significance of bivariate normal distribution of the `setosa2` data set.

Test	Test Statistic	p-value
Mardia		
Skewness	0.760	0.944
Kurtosis	0.093	0.926
Henze-Zirkler	0.286	0.915
Royston	2.698	0.245

Table 2: MVN test results (`setosa` with `sepal` measures).

All three tests in Table 2 indicate that the data set satisfies bivariate normality assumption at the significance level 0.05. Moreover, the perspective and contour plots are in agreement with the test results and indicate approximate bivariate normality.

Figures 3a and 3b were drawn using a pre-defined graphical option by the authors. However, users may change these options by setting function entry to `default = FALSE`. If the `default` is `FALSE`, optional arguments from the `plot`, `persp` and `contour` functions may be introduced to the corresponding graphs.

3.7 Multivariate outliers

Multivariate outliers are the common reason for violating MVN assumption. In other words, MVN assumption requires the absence of multivariate outliers. Thus, it is crucial to check whether the data have multivariate outliers, before starting to multivariate analysis. The **MVN** includes two multivariate outlier detection methods which are based on robust Mahalanobis distances (rMD(x)). Mahalanobis distance is a metric which calculates how far each observation is to the center of joint distribution, which can be thought of as the centroid in multivariate space. Robust distances are estimated from minimum covariance determinant estimators rather than the sample covariance [29]. These two approaches, defined as Mahalanobis distance and adjusted Mahalanobis distance in the package, detect multivariate outliers as given below,

Mahalanobis Distance:

1. Compute robust Mahalanobis distances (rMD(x_i)),
2. Compute the 97.5 percent quantile (Q) of the chi-square distribution,
3. Declare rMD(x_i) $> Q$ as possible outlier.

Adjusted Mahalanobis Distance:

1. Compute robust Mahalanobis distances (rMD(x_i)),
2. Compute the 97.5 percent adjusted quantile (AQ) of the chi-Square distribution,
3. Declare rMD(x_i) $> AQ$ as possible outlier.

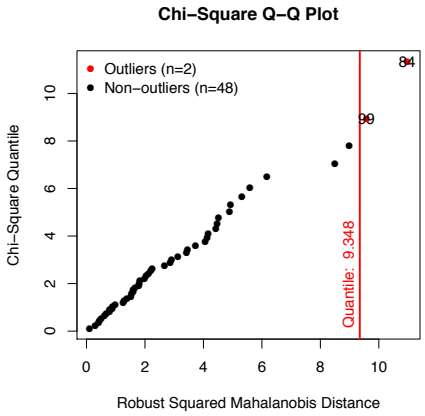
The `mvOutlier` function is used to detect multivariate outliers as given below. It also returns a new data set in which declared outliers are removed. Moreover, Q-Q plots can be created by setting `qqplot = TRUE` within `mvOutlier` for visual inspection of the possible outliers. For this example, we will use another subset of the `Iris` data, which is `versicolor` flowers, with the first three variables.

```
versicolor <- iris[51:100, 1:3]
# Mahalanobis distance
result <- mvOutlier(versicolor, qqplot = TRUE, method = "quan")
# Adjusted Mahalanobis distance
result <- mvOutlier(versicolor, qqplot = TRUE, method = "adj.quan")
```

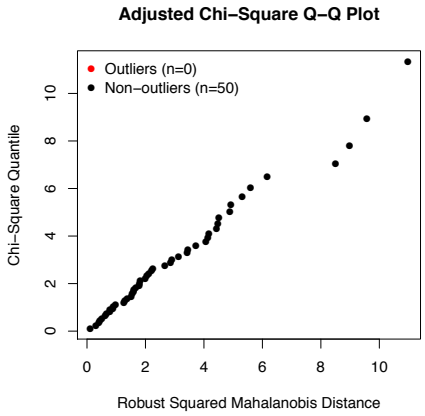
From Figure 4, Mahalanobis distance declares 2 observations as multivariate outlier whereas adjusted Mahalanobis distance declares none. See [30] for further information on multivariate outliers.

4 Web interface for the MVN package

The purpose of the package is to provide MVN tests along with graphical approaches for assessing MVN. Moreover, this package offers univariate tests and plots, and multivariate outlier detection for checking MVN assumptions through R. However, using R codes might be challenging for new R



(a) Mahalanobis Distance



(b) Adjusted-Mahalanobis Distance

Figure 4: Multivariate outlier detection.

users. Therefore, we also developed a user-friendly web application of **MVN** by using **shiny**[1] [31]. This web-tool, which is an interactive application, has all the features that the **MVN** package has. It is publicly available through `http://www.biosoft.hacettepe.edu.tr/MVN`.

5 Summary and further researches

As stated earlier, MVN is among the most crucial assumptions for most parametric multivariate statistical procedures. The power of these procedures is negatively affected if this assumption is not satisfied. Thus, before using any of the parametric multivariate statistical methods, MVN assumption should be tested first of all. Although there are many MVN tests, there is not a standard test for assessing this assumption. In our experience, researchers may choose Royston's test for data with a small sample size ($n < 50$) and Henze-Zirkler's test for a large sample size ($n > 100$). However, a more comprehensive simulation study is needed to provide more reliable inference. Instead of using just one test, it is suggested that using several tests simultaneously and examining some graphical representation of the data may be more appropriate. Currently, as we know, there is no such extensive tool to apply different statistical tests and graphical methods together.

In this vignette, we present the **MVN** package for multivariate normality checking. This package offers comprehensive flexibility for assessing MVN assumption. It contains the three most widely used MVN tests, including Mardia's, Henze-Zirkler's and Royston's. Moreover, researchers can create three MVN plots using this package, including the chi-square Q-Q plot for any data set and perspective and contour plots for bivariate data sets. Furthermore, since MVN requires univariate normality of each variable, users can check univariate normality assumption by using both univariate normality tests and plots with proper functions in the package. In the first example, different results on multivariate normality were achieved from the same data. When `sepal` and `petal` measures, i.e four variables, were considered, Mardia's test resulted in multivariate normality as well as Henze-Zirkler's test at the edge of type I error. However, Royston's test strongly rejected the null hypothesis in favor of non-normality. At this point, the only possible graphical approach is to use the chi-square Q-Q plot since there are more than two variables. The next step was to identify the cause of deviation from MVN by using univariate normality tests and plots. In the second example, all tests

[1] `http://www.rstudio.com/shiny/`

suggested bivariate normality, as did the graphical approaches. Although some tests can not reject null hypothesis, other tests may reject it. Hence, as stated earlier, selecting the appropriate MVN test dramatically changes the results and the final decision is ultimately the researcher's.

Currently, **MVN** works with several statistical tests and graphical approaches. It will continue to add new statistical approaches as they are developed. The package and the web-tool will be regularly updated based on these changes.

6 Acknowledgments

We would like to thank Izzet Parug Duru from Marmara University Department of Physics and Vahap Eldem from Istanbul University Department of Biology for making the web-tool version of the package possible.

References

[1] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[2] Francois Husson, Julie Josse, Sebastien Le, and Jeremy Mazet. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R*, 2014. R package version 1.26.

[3] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2014. R package version 1.4.8.

[4] Ignacio González and Sébastien Déjean. *CCA: Canonical correlation analysis*, 2012. R package version 1.2.

[5] Tom Burdenski. Evaluating univariate, bivariate, and multivariate normality using graphical and statistical procedures. *Multiple Linear Regression Viewpoints*, 26(2):15–28, 2000.

[6] D. George and P. Mallery. *SPSS for Windows step by step*. Allyn & Bacon, Boston, 1999.

[7] Shahla Ramzan, Faisal Maqbool Zahid, and Shumila Ramzan. Evaluating multivariate normality: A graphical approach. *Middle East Journal of Scientific Research*, 13(2):254–263, 2013.

[8] HET Holgersson. A graphical method for assessing multivariate normality. *Computational Statistics*, 21(1):141–149, 2006.

[9] Thomas Svantesson and Jon W Wallace. Tests for assessing multivariate normality and the covariance structure of mimo data. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages 656–659. IEEE, 2003.

[10] Christopher J. Mecklin and Daniel J. Mundfrom. 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(2):93–107, 2005.

[11] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. *MVN: Multivariate Normality Tests*, 2014. R package version 3.7.

[12] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[13] K. V. Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002)*, 36(2):115–128, 1974.

[14] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3617, 1990.

[15] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey, third edition, 1992.

[16] Norbert Henze and Thorsten Wagner. A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis*, 62(1):1–23, 1997.

[17] Christopher J Mecklin and Daniel J Mundfrom. On using asymptotic critical values in testing for multivariate normality. *InterStat*, 1:1–12, 2003.

[18] Reha Alpar. *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. Detay Yayıncılık, Ankara, Turkey, fourth edition, 2013.

[19] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1964.

[20] J. P. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115–124, 1982.

[21] J. P. Royston. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics*, 32(2):121–133, 1983.

[22] Patrick Royston. Approximating the Shapiro-Wilk W test for non-normality. *Statistics and Computing*, 2(3):117–119, 1992.

[23] Patrick Royston. Remark as r94: A remark on algorithm AS 181: The W test for normality. *Applied Statistics*, 44(4):547–551, 1995.

[24] Gavin J S Ross. MLP, Maximum Likelihood Program. *Harpenden: Rothamsted Experimental Station*, 1980.

[25] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[26] Edgar Anderson. The species problem in Iris. *Missouri Botanical Garden Press*, 23(3):457–509, 1936.

[27] James P Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[28] Robert E Kass, Uri T Eden, and Emery N Brown. *Analysis of Neural Data*. Springer, 2014.

[29] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[30] Peter Filzmoser, Robert G. Garrett, and Clemens Reimann. Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31(5):579–587, 2005.

[31] RStudio, Inc. *shiny: Web Application Framework for R*, 2014. R package version 0.10.1.