MVN: An R Package for Assessing Multivariate Normality

Selcuk Korkmaz[1], Dincer Goksuluk and Gokmen Zararsiz

Trakya University, Faculty of Medicine, Department of Biostatistics, Edirne, TURKEY
[1]**selcukorkmaz@gmail.com**

MVN version 5.3 (Last revision 2018-05-06)

Abstract

We previously presented **MVN** (https://cran.r-project.org/web/packages/MVN/index.html) package to assess multivariate normality. We also published the paper of the package (https://journal.r-project.org/archive/2014/RJ-2014-031/RJ-2014-031.pdf). Now, we present an updated version of the package. The web-tool of the package available at http://opensoft.turcosa.com.tr/MVN/.

1 Implementation of MVN package

The **MVN** package contains functions in the `S3` class to assess multivariate normality. This package is the updated version of the **MVN** package [1]. 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 [2]. 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 [3]. First of all, the **MVN** library should be loaded in order to use related functions.

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

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

```r
# 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.

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

1.1 mvn function

In this section we will introduce our **mvn** function. This function includes all the arguments to assess multivariate normality through multivariate normality tests, multivariate plots, multivariate outlier detection, univariate normality tests and univariate plots.

```
mvn(data, subset = NULL, mvnTest = c("mardia", "hz", "royston", "dh",
  "energy"), covariance = TRUE, tol = 1e-25, alpha = 0.5,
  scale = FALSE, desc = TRUE, transform = "none", R = 1000,
  univariateTest = c("SW", "CVM", "Lillie", "SF", "AD"),
  univariatePlot = "none", multivariatePlot = "none",
  multivariateOutlierMethod = "none", showOutliers = FALSE,
  showNewData = FALSE)
```

Arguments	Definition
data	a numeric matrix or data frame
subset	define a variable name if subset analysis is required
mvnTest	select one of the MVN tests. Type ’mardia’ for Mardia’s test, ’hz’ for Henze-Zirkler’s test, ’royston’ for Royston’s test, ’dh’ for Doornik-Hansen’s test and energy for E-statistic. See details for further information.
covariance	this option works for ’mardia’ and ’royston’. If TRUE covariance matrix is normalized by n, if FALSE it is normalized by n-1
tol	a numeric tolerance value which isused for inversion of the covariance matrix (default = 1e-25)
alpha	a numeric parameter controlling the size of the subsets over which the determinant is minimized. Allowed values for the alpha are between 0.5 and 1 and the default is 0.5.
scale	if TRUE scales the colums of data
desc	a logical argument. If TRUE calculates descriptive statistics
transform	select a transformation method to transform univariate marginal via logarithm (’log’), square root (’sqrt’) and square (’square’)
R	number of bootstrap replicates for Energy test, default is 1000
univariateTest	select one of the univariate normality tests, Shapiro-Wilk (’SW’), Cramer-von Mises (’CVM’), Lilliefors (’Lillie’), Shapiro-Francia (’SF’), Anderson-Darling (’AD’)
univariatePlot	select one of the univariate normality plots, Q-Q plot (’qq’), histogram (’histogram’), box plot (’box’), scatter (’scatter’)
multivariatePlot	’qq’ for chi-square Q-Q plot, ’persp’ for perspective plot, ’contour’ for contour plot
multivariateOutlierMethod	select multivariate outlier detection method, ’quan’ quantile method based on Mahalanobis distance and ’adj’ adjusted quantile method based on Mahalanobis distance
showOutliers	if TRUE prints multivariate outliers
showNewData	if TRUE prints new data without outliers

1.2 Mardia's MVN test

`mvnTest = "mardia"` argument in the **mvn** 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 <- mvn(data = setosa, mvnTest = "mardia")
result$multivariateNormality

##              Test        Statistic           p value Result
## 1 Mardia Skewness 25.6643445196298 0.177185884467652    YES
## 2 Mardia Kurtosis 1.29499223711605 0.195322907441935    YES
## 3             MVN             <NA>              <NA>    YES
```

This function performs multivariate skewness and kurtosis tests at the same time and combines test results for multivariate normality. If both tests indicates multivariate normality, then data follows a multivariate normality distribution at the 0.05 significance level.

1.3 Henze-Zirkler's MVN test

One may use the `mvnTest = "hz"` in the **mvn** function to perform the Henze-Zirkler's test.

```
result <- mvn(data = setosa, mvnTest = "hz")
result$multivariateNormality

##            Test        HZ    p value MVN
## 1 Henze-Zirkler 0.9488453 0.04995356  NO
```

The last column indicates whether dataset follows a multivariate normality or not (i.e, YES or NO) at significance level 0.05.

1.4 Royston's MVN test

In order to carry out the Royston's test, set `mvnTest = "royston"` argument in the **mvn** function as follows:

```
result <- mvn(data = setosa, mvnTest = "royston")
result$multivariateNormality

##      Test        H      p value MVN
## 1 Royston 31.51803 2.187653e-06  NO
```

The last column indicates whether dataset follows a multivariate normality or not (i.e, YES or NO) at significance level 0.05.

NOTE: Do not apply Royston's test, if dataset includes more than 5000 cases or less than 3 cases, since it depends on Shapiro-Wilk's test.

1.5 Doornik-Hansen's MVN test

In order to carry out the Doornik-Hansen's test, set `mvnTest = "dh"` argument in the **mvn** function as follows:

```
result <- mvn(data = setosa, mvnTest = "dh")
result$multivariateNormality

##             Test        E df      p value MVN
## 1 Doornik-Hansen 126.5584  8 1.460761e-23  NO
```

The last column indicates whether dataset follows a multivariate normality or not (i.e, YES or NO) at significance level 0.05.

1.6 Energy test

In order to carry out the Doornik-Hansen's test, set `mvnTest = "energy"` argument in the **mvn** function as follows:

```
result <- mvn(data = setosa, mvnTest = "energy")
result$multivariateNormality

##          Test Statistic p value MVN
## 1 E-statistic  1.203397   0.023  NO
```

The last column indicates whether dataset follows a multivariate normality or not (i.e, YES or NO) at significance level 0.05.

1.7 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 `multivariatePlot = "qq"` option in the `mvn`, function 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.

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.

1.8 Univariate plots and tests

As noted by several authors [4–6], 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 `univariatePlot` and `univariateTest` arguments

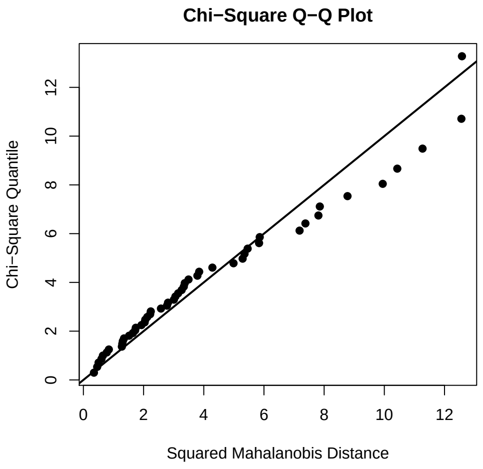
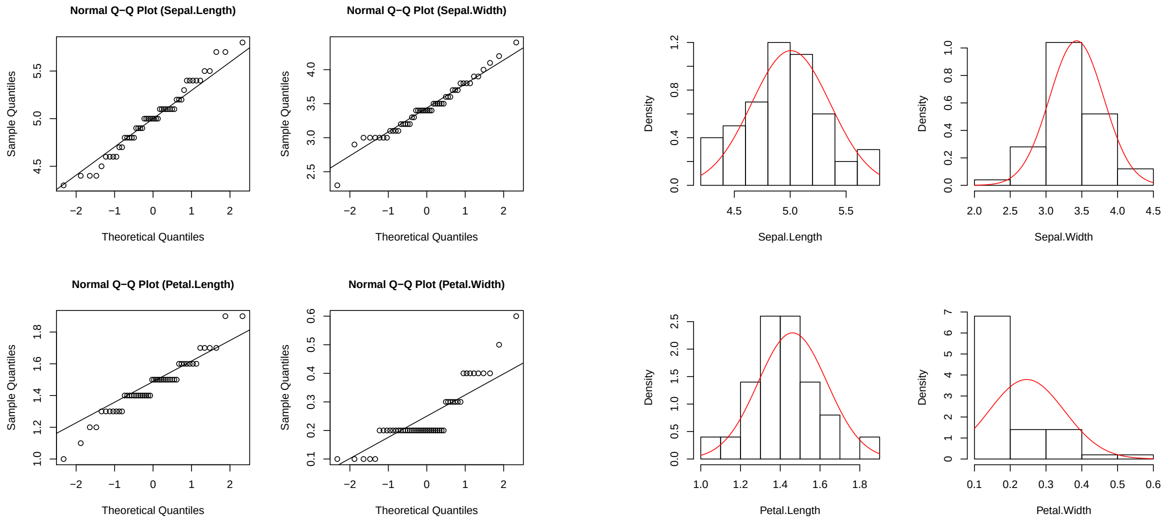


Figure 1: Chi-Square Q-Q plot for `setosa` data set.

from the `mvn` function. Set `univariatePlot` argument `"qq"` for Q-Q plots (Figure 2a), `"histogram"` for histograms with normal curves (Figure 2b), `"box"` for box-plots and `"scatter"` for scatterplot matrices.

```
# create univariate Q-Q plots
result <- mvn(data = setosa, mvnTest = "royston", univariatePlot = "qqplot")

# create univariate histograms
result <- mvn(data = setosa, mvnTest = "royston", univariatePlot = "histogram")
```



(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 `univariateTest` argument in the mvn `function`. It provides several widely used univariate normality tests, including `"SW"` (do not apply Shapiro-Wilk's test, if dataset includes more than 5000 cases or less than 3 cases.) for Shapiro-Wilk test, `"CVM"` for Cramer-von Mises test, texttt"Lillie" for Lilliefors test, `"SF"` for Shapiro-Francia test and `"AD"` Anderson-Darling test. 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:

```r
result <- mvn(data = setosa, mvnTest = "royston", univariateTest = "SW", desc = TRUE)
result$univariateNormalityResult
```

```
##               n  Mean   Std.Dev Median Min Max 25th  75th       Skew
## Sepal.Length 50 5.006 0.3524897    5.0 4.3 5.8  4.8 5.200 0.11297784
## Sepal.Width  50 3.428 0.3790644    3.4 2.3 4.4  3.2 3.675 0.03872946
## Petal.Length 50 1.462 0.1736640    1.5 1.0 1.9  1.4 1.575 0.10009538
## Petal.Width  50 0.246 0.1053856    0.2 0.1 0.6  0.2 0.300 1.17963278
##                 Kurtosis
## Sepal.Length -0.4508724
## Sepal.Width   0.5959507
## Petal.Length  0.6539303
## Petal.Width   1.2587179
```

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 2.

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
Doornik-Hansen	64.974	0.000
Energy	0.786	0.634

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

According to the three MVN test results in Table 2, `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.

1.9 Perspective and contour plots

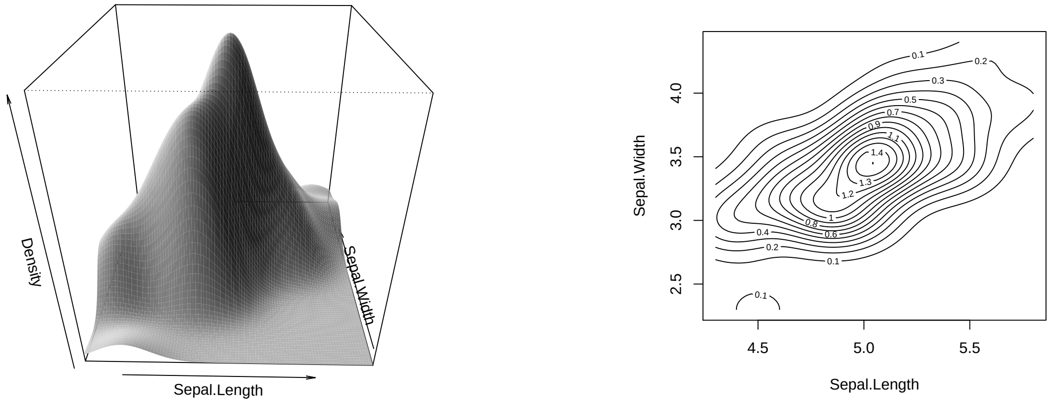
Univariate normal marginal densities are a necessary but not a sufficient condition for MVN. Hence, in addition to univariate plots, creating perspective and contour plots will be useful. The perspective plot is an extension of the univariate probability distribution curve into a 3·dimensional probability 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 `multivariatePlot` argument in the **mvn** function. In the following codes, we used `multivariatePlot = "persp"` to create perspective plot (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, when we set `multivariatePlot = "contour"`. 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.

```r
setosa2 <- iris[1:50, 1:2]

# perspective plot
result <- mvn(setosa2, mvnTest = "hz", multivariatePlot = "persp")

# contour plot
result <- mvn(setosa2, mvnTest = "hz", multivariatePlot = "contour")
```



(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.

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	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
Doornik-Hansen	11.570	0.021
Energy	0.527	0.776

Table 2: MVN test results (`setosa` without `Petal.Width`).

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.

1.10 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 [7]. These two approaches, defined as Mahalanobis distance and adjusted Mahalanobis distance in the package, detect multivariate outliers as given below,

Mahalanobis Distance:

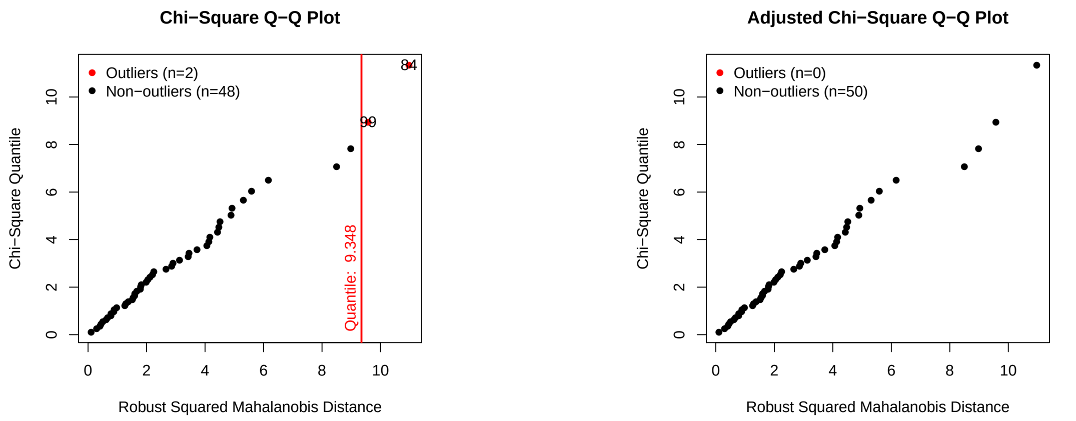
1. Compute robust Mahalanobis distances (rMD(x_i)),
2. Compute the 97.5 percent quantile (Q) of the chi-square distribution,
3. Declare rMD(x_i) $> Q$ as possible outlier.

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 `multivariateOutlierMethod` argument as `"quan"` for quantile method based on Mahalanobis distance and as `"adj"` for adjusted quantile method based on Mahalanobis distance to detect multivariate outliers as given below. It also returns a new data set in which declared outliers are removed. Moreover, this argument creates Q-Q plots 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 <- mvn(data = versicolor, mvnTest = "hz", multivariateOutlierMethod = "quan")
# Adjusted Mahalanobis distance
result <- mvn(data = versicolor, mvnTest = "hz", multivariateOutlierMethod = "adj")
```



(a) Mahalanobis Distance

(b) Adjusted-Mahalanobis Distance

Figure 4: Multivariate outlier detection.

From Figure 4, Mahalanobis distance declares 2 observations as multivariate outlier whereas adjusted Mahalanobis distance declares none. See [8] for further information on multivariate outliers.

1.11 Subset analysis

One may also perform sub-group analysis using `mvn` function. Let's use the `Iris` dataset once more for this purpose. In the dataset, there is a group variable (`Species`), which defines the specie of the flower.

```
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
```

```
result <- mvn(data = iris, subset = "Species", mvnTest = "hz")
result$multivariateNormality

## $setosa
##            Test        HZ    p value MVN
## 1 Henze-Zirkler 0.9488453 0.04995356  NO
```

```
##
## $versicolor
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.8388009 0.2261991 YES
##
## $virginica
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.7570095 0.4970237 YES
```

According to the Henze-Zirkler's test results, dataset for setosa does not follow a multivariate normal distribution, whereas dataset versicolor and virginica follow a multivariate normal distribution.

2 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 users. Therefore, we also developed a user-friendly web application by using **shiny**[1] [9]. 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/`.

References

[1] S Korkmaz, D Goksuluk, and G Zararsiz. Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

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

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

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

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

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

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

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

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

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