Package 'MVN'

May 28, 2025

Type Package

Title Multivariate Normality Tests

Version 6.0

Maintainer Selcuk Korkmaz <selcukorkmaz@gmail.com>

Depends R (>= 3.5.0)

Imports methods, nortest, moments, MASS, boot, car, dplyr, tidyr, purrr, stringr, tibble, ggplot2, viridis, cli, energy, plotly

Collate 'mvn.R' 'mardia.R' 'hz.R' 'royston.R' 'doornik_hansen.R' 'energy.R' 'descriptives.R' 'test_univariate_normality.R' 'multivariate_diagnostic_plot.R' 'mv_outlier.R' 'univariate_diagnostic_plot.R' 'box_cox_transform.R' 'arw_adjustment.R' 'plot.mvn.R' 'summary.mvn.R'

Description A suite of multivariate normality tests (Mardia, Henze–Zirkler, Royston, Doornik–Hansen, Energy), univariate diagnostics, robust outlier detection, bivariate density plots, and Box–Cox transformations (Korkmaz et al, (2014), <https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf>).

Encoding UTF-8

License MIT + file LICENSE

URL https://selcukorkmaz.github.io/mvn-tutorial/, https://github.com/selcukorkmaz/MVN

BugReports https://github.com/selcukorkmaz/MVN/issues

RoxygenNote 7.3.2

NeedsCompilation no

Author Selcuk Korkmaz [aut, cre] (ORCID: <https://orcid.org/0000-0003-4632-6850>), Dincer Goksuluk [aut], Gokmen Zararsiz [aut]

Repository CRAN

Date/Publication 2025-05-28 08:00:11 UTC

Contents

`arw_adjustment` *Atkinson–Riani–Welsh (ARW) Adjusted Cutoff for Robust Mahalanobis Distances*

Description

Implements the ARW procedure to compute an adjusted cutoff for squared Mahalanobis distances, then re-estimates location and scatter excluding points beyond the cutoff.

Usage

```
arw_adjustment(x, m0, c0, alpha, pcrit)
```

Arguments

`x`	A numeric matrix or data frame of observations (rows) by variables (columns), with at least 2 columns.
`m0`	A numeric vector of initial location estimates (length equal to number of columns in `x`).
`c0`	A numeric covariance matrix corresponding to `m0`.
`alpha`	Numeric; significance level for the chi-square threshold. Defaults to 0.05 if not provided.
`pcrit`	Numeric; minimal proportion for the adjusted cutoff. If not provided, it is computed as: $(0.24 - 0.003p)/\sqrt{n}$ if $p \leq 10$, or $(0.252 - 0.0018p)/\sqrt{n}$ if $p > 10$.

Value

A list with the following components: `m`, the updated location vector after excluding outliers; `c`, the updated covariance matrix; `cn`, the adjusted cutoff on Mahalanobis distances; `w`, a logical vector indicating which observations have distance less than or equal to `cn`.

---

box_cox_transform *Box-Cox Transformation for Numeric Data*

Description

Applies a Box-Cox power transformation to each numeric variable in the input, estimating or rounding the transformation parameter (lambda) using `car::powerTransform`.

Usage

```
box_cox_transform(data, type = c("optimal", "rounded"))
```

Arguments

`data`	A numeric vector, matrix, or data frame with observations in rows and variables in columns.
`type`	Character string; either `"optimal"` (use the estimated lambda) or `"rounded"` (use the rounded lambda). Default is `"optimal"`.

Value

A list with two components: `data`, a data frame (or vector) of transformed variables; and `lambda`, a numeric vector of the lambda values applied to each variable.

Examples

```
## Not run:
data <- iris[1:4]
# Apply Box-Cox to the first 4 numeric columns
res <- box_cox_transform(data, type = "rounded")
head(res$data)
res$lambda

## End(Not run)
```

descriptives *Descriptive Statistics for Numeric Data*

Description

Computes key descriptive statistics for each numeric variable in a vector, matrix, or data frame.

Usage

```
descriptives(data)
```

Arguments

data	A numeric vector, matrix, or data frame with observations in rows and variables in columns.

Value

A data frame where each row corresponds to a variable and each column represents a summary statistic: number of non-missing observations (n), arithmetic mean (Mean), standard deviation (Std.Dev), median (Median), minimum (Min), maximum (Max), first quartile (25th), third quartile (75th), sample skewness (Skew, from moments::skewness), and sample kurtosis (Kurtosis, from moments::kurtosis).

Examples

```
## Not run:
data <- iris[1:4]
descriptives(data)

## End(Not run)
```

doornik_hansen *Doornik-Hansen Test for Multivariate Normality*

Description

Performs the Doornik–Hansen omnibus test by transforming the data to approximate normality and combining skewness and kurtosis measures to test for multivariate normality.

Usage

```
doornik_hansen(data)
```

Arguments

data	A numeric matrix or data frame with observations in rows and variables in columns.

Value

A data frame with one row containing the following columns: Test, the name of the test ("Doornik-Hansen"); Statistic, the value of the test statistic; df, the degrees of freedom; and p.value, the p-value from a chi-square approximation.

Examples

```
## Not run:
data <- iris[1:50, 1:4]
dh_result <- doornik_hansen(data)
dh_result

## End(Not run)
```

---

energy *E-Statistic Test for Multivariate Normality (Energy Test)*

Description

Performs the E-statistic test for multivariate normality using a parametric bootstrap to estimate the null distribution of the test statistic.

Usage

```
energy(data, R = 1000, seed = 123)
```

Arguments

data	A numeric matrix or data frame with observations in rows and variables in columns.
R	Integer; number of bootstrap replicates to estimate the null distribution. Default is 1000.
seed	Optional integer to set the random seed for reproducibility.

Value

A data frame with one row containing the following columns: Test, the name of the test ("E-Statistic"); Statistic, the observed E-statistic; and p.value, the p-value obtained from the bootstrap procedure.

Examples

```
## Not run:
data <- iris[1:50, 1:4]
energy_result <- energy(data, R = 500)
energy_result

## End(Not run)
```

---

hz *Henze-Zirkler Test for Multivariate Normality*

Description

Performs Henze and Zirkler's test to assess multivariate normality based on a log-normal approximation of the test statistic.

Usage

```
hz(data, use_population = TRUE, tol = 1e-25)
```

Arguments

`data`	A numeric matrix or data frame with observations in rows and variables in columns.
`use_population`	Logical; if `TRUE`, uses the population covariance estimator $\frac{n-1}{n} \times \Sigma$; otherwise uses the sample covariance. Default is `TRUE`.
`tol`	Numeric tolerance passed to `solve` when inverting the covariance matrix. Default is `1e-25`.

Value

A data frame with one row, containing the following columns: `Test`, the name of the test ("Henze-Zirkler"); `HZ`, the test statistic (numeric); and `p.value`, the p-value computed from a log-normal approximation.

Examples

```
## Not run:
data <- iris[1:50, 1:4]
hz_result <- hz(data)
hz_result

## End(Not run)
```

mardia *Mardia's Test for Multivariate Normality*

Description

Performs Mardia's skewness and kurtosis tests to assess multivariate normality in a multivariate dataset.

Usage

```
mardia(data, use_population = TRUE, tol = 1e-25)
```

Arguments

data	A numeric matrix or data frame with observations in rows and variables in columns.
use_population	Logical; if TRUE, uses the population covariance estimator $\frac{n-1}{n} \times \Sigma$; otherwise uses the sample covariance. Default is TRUE.
tol	Numeric tolerance passed to solve when inverting the covariance matrix. Default is 1e-25.

Value

A data frame with two rows, one for Mardia's skewness test and one for the kurtosis test. Each row contains the name of the test (Test), the test statistic (Statistic), and the associated p-value (p.value).

Examples

```
## Not run:
data <- iris[1:50, 1:4]
mardia_result <- mardia(data)
mardia_result

## End(Not run)
```

multivariate_diagnostic_plot *Plot Multivariate Normal Diagnostics and Bivariate Kernel Density*

Description

Generates either a Mahalanobis Q-Q plot, an interactive 3D kernel density surface plot, or a 2D kernel density contour plot for exactly two numeric variables. The function is intended for assessing multivariate normality or exploring the bivariate distribution of the input data.

Usage

```
multivariate_diagnostic_plot(
  data,
  type = c("qq", "persp", "contour"),
  tol = 1e-25,
  use_population = TRUE
)
```

Arguments

data	A numeric vector, matrix, or data frame. Non-numeric columns are dropped with a warning; incomplete rows are removed. The input must contain exactly two numeric variables.
type	Character string specifying the type of plot to generate. Must be one of "qq" (Mahalanobis Q-Q plot), "persp" (3D KDE surface), or "contour" (2D KDE contour). Default is "qq".
tol	Numeric tolerance for matrix inversion passed to solve(). Default is 1e-25.
use_population	Logical; if TRUE, uses the population covariance estimator $\frac{n-1}{n} \times \Sigma$; otherwise uses the sample covariance. Default is TRUE.

Value

If type = "qq", returns a ggplot2 object representing a Mahalanobis Q-Q plot. If type = "persp" or "contour", returns an interactive plotly widget displaying the KDE surface or contour, respectively.

Examples

```
## Not run:
library(MASS)
data(iris)

# Mahalanobis Q-Q plot
multivariate_diagnostic_plot(iris[, 1:2], type = "qq")

# 3D KDE surface
multivariate_diagnostic_plot(iris[, 1:2], type = "persp")

# 2D KDE contour
multivariate_diagnostic_plot(iris[, 1:2], type = "contour")

## End(Not run)
```

mvn *Comprehensive Multivariate Normality and Diagnostic Function*

Description

Conduct multivariate normality tests, outlier detection, univariate normality tests, descriptive statistics, and Box-Cox transformation in one wrapper.

Usage

```
mvn(
  data,
  subset = NULL,
  mvn_test = "hz",
  use_population = TRUE,
  tol = 1e-25,
  alpha = 0.05,
  scale = FALSE,
  descriptives = TRUE,
  transform = "none",
  R = 1000,
  univariate_test = "AD",
  multivariate_outlier_method = "none",
  box_cox_transform = FALSE,
  box_cox_transform_type = "optimal",
  show_new_data = FALSE,
  tidy = TRUE
)
```

Arguments

data	A numeric matrix or data frame (rows = observations, columns = variables).
subset	Optional character; name of a grouping variable in data for subset analyses.
mvn_test	Character; one of "mardia", "hz", "royston", "doornik_hansen", or "energy". Default: "hz".
use_population	Logical; if TRUE, uses the population covariance estimator $\frac{n-1}{n} \times \Sigma$; otherwise uses the sample covariance. Default is TRUE.
tol	Numeric; tolerance for matrix inversion via solve(). Default: 1e-25.
alpha	Numeric; significance level for ARW outlier cutoff when multivariate_outlier_method = "adj". Default: 0.05.
scale	Logical; if TRUE, standardizes the data before analysis. Default: FALSE.
descriptives	Logical; if TRUE, compute descriptive statistics. Default: TRUE.
transform	Character; one of "none", "log", "sqrt", "square". Applies marginal transformations before analysis. Default: "none".

R Integer; number of bootstrap replicates for "energy" test. Default: 1000.

univariate_test
Character; one of "SW", "CVM", "Lillie", "SF", "AD". Default: "AD".

multivariate_outlier_method
Character; "none", "quan", or "adj". Default: "none".

box_cox_transform
Logical; if TRUE, applies Box-Cox transformation to all variables. Default: FALSE.

box_cox_transform_type
Character; either "optimal" or "rounded" lambda for Box-Cox. Default: "optimal".

show_new_data Logical; if TRUE, include cleaned data (non-outliers). Default: FALSE.

tidy Logical; if TRUE, returns the results as a tidy data frame with an added Group column. Default is TRUE.

Details

If mvn_test = "mardia", it calculates the Mardia's multivariate skewness and kurtosis coefficients as well as their corresponding statistical significance. It can also calculate corrected version of skewness coefficient for small sample size (n< 20). For multivariate normality, both p-values of skewness and kurtosis statistics should be greater than 0.05. If sample size less than 20 then p.value.small should be used as significance value of skewness instead of p.value.skew. If there are missing values in the data, a listwise deletion will be applied and a complete-case analysis will be performed.

If mvn_test = "hz", it calculates the 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 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. If there are missing values in the data, a listwise deletion will be applied and a complete-case analysis will be performed.

If mvn_test = "royston", it calculates the Royston's multivariate normality test. A function to generate the Shapiro-Wilk's W statistic needed to feed the Royston's H test 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. If there are missing values in the data, a listwise deletion will be applied and a complete-case analysis will be performed. 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.

If mvn_test = "doornik_hansen", it calculates the Doornik-Hansen's multivariate normality test. The code is adapted from asbio package (Aho, 2017).

If mvn_test = "energy", it calculates the Energy multivariate normality test. The code is adapted from energy package (Rizzo and Szekely, 2017).

Value

A named list containing:

multivariate_normality A data frame of the selected multivariate normality (MVN) test results.

univariate_normality A data frame of univariate normality test results.

descriptives (Optional) A data frame of descriptive statistics if `descriptives = TRUE`.

multivariate_outliers (Optional) A data frame of flagged multivariate outliers if `multivariate_outlier_method != "none"`.

new_data (Optional) Original data with multivariate outliers removed if `show_new_data = TRUE`.

box_cox_lambda (Optional) Estimated Box-Cox lambda values if `box_cox_transform = TRUE`.

data The processed data matrix used in the analysis (transformed and/or cleaned).

subset (Optional) The grouping variable used for subset analysis, if applicable.

Author(s)

Selcuk Korkmaz, <selcukorkmaz@gmail.com>

References

Korkmaz S, Goksuluk D, Zararsiz G. MVN: An R Package for Assessing Multivariate Normality. The R Journal. 2014 6(2):151-162. URL https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Mardia, K. V. (1970), Measures of multivariate skewnees and kurtosis with applications. Biometrika, 57(3):519-530.

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

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

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

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

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

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

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

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

Doornik, J.A. and Hansen, H. (2008). An Omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70, 927-939.

G. J. Szekely and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances, Journal of Statistical Planning and Inference, http://dx.doi.org/10.1016/j.jspi.2013.03.018

M. L. Rizzo and G. J. Szekely (2016). Energy Distance, WIRES Computational Statistics, Wiley, Volume 8 Issue 1, 27-38. Available online Dec., 2015, http://dx.doi.org/10.1002/wics.1375.

G. J. Szekely and M. L. Rizzo (2017). The Energy of Data. The Annual Review of Statistics and Its Application 4:447-79. 10.1146/annurev-statistics-060116-054026

Examples

```
result = mvn(data = iris[-4], subset = "Species", mvn_test = "hz",
             univariate_test = "AD",
             multivariate_outlier_method = "adj",
             show_new_data = TRUE)

### Multivariate Normality Result
summary(result, select = "mvn")

### Univariate Normality Result
summary(result, select = "univariate")

### Descriptives
summary(result, select = "descriptives")

### Multivariate Outliers
summary(result, select = "outliers")

### New data without multivariate outliers
summary(result, select = "new_data")
```

`mv_outlier` *Identify Multivariate Outliers via Robust Mahalanobis Distances*

Description

Computes robust Mahalanobis distances for multivariate data using the Minimum Covariance Determinant (MCD) estimator, flags outliers based on either a chi-square quantile cutoff or an adjusted cutoff using the Atkinson–Riani–Welsh (ARW) method, and optionally generates a Mahalanobis Q–Q plot.

Usage

```
mv_outlier(
  data,
  outlier = TRUE,
  qqplot = TRUE,
  alpha = 0.05,
  method = c("quan", "adj"),
  label = TRUE,
  title = "Chi-Square Q-Q Plot"
)
```

Arguments

`data`	A numeric matrix or data frame with observations in rows and at least two numeric columns.

outlier	Logical; if TRUE, includes the Mahalanobis distance values and outlier classification in the output. If FALSE, suppresses this component. Default is TRUE.
qqplot	Logical; if TRUE, a Chi-Square Q–Q plot is generated to visualize outlier detection. Default is TRUE.
alpha	Numeric; significance level used for the adjusted cutoff method (only applies if method = "adj"). Default is 0.05.
method	Character string specifying the outlier detection method. Must be either "quan" (quantile-based cutoff) or "adj" (adjusted cutoff via ARW). Default is "quan".
label	Logical; if TRUE and qqplot = TRUE, labels the detected outliers in the plot. Default is TRUE.
title	Optional character string specifying the title for the Q–Q plot. Default is "Chi-Square Q-Q Plot".

Value

A list containing the following components: outlier, a data frame of Mahalanobis distances with observation IDs and outlier flags (if outlier = TRUE); qq_outlier_plot, a ggplot object of the Mahalanobis Q–Q plot (if qqplot = TRUE); and newData, a data frame of non-outlier observations.

Examples

```
## Not run:
data <- iris[, 1:4]
res <- mv_outlier(data, method = "adj", alpha = 0.025)
res$outlier
res$qq_outlier_plot
head(res$newData)

## End(Not run)
```

---

plot.mvn *Plot Diagnostics for Multivariate Normality Analysis*

Description

Generates diagnostic plots for objects of class mvn, including multivariate Q-Q plots, 3D or contour kernel density plots, univariate plots (e.g., Q-Q, histograms, boxplots), and multivariate outlier detection plots. If a grouping variable (subset) was used in the mvn function, plots will be generated separately for each group.

Usage

```
## S3 method for class 'mvn'
plot(x, ...)
```

Arguments

x	An object of class mvn, as returned by the mvn function.
...	Additional arguments passed to internal plotting functions: diagnostic ("multivariate", "univariate", "outlier"), type (e.g., "qq", "boxplot", "persp"), interactive (logical; use **plotly**), and

Value

This function is called for its side effect of producing plots. It does not return a value.

Examples

```
## Not run:
data <- iris[1:4]
result <- mvn(data)

plot(result, diagnostic = "multivariate", type = "qq")
plot(result, diagnostic = "univariate", type = "boxplot")
plot(result, diagnostic = "outlier")

## End(Not run)
```

---

royston *Royston's Multivariate Normality Test*

Description

Performs Royston's test for multivariate normality by combining univariate W-statistics (Shapiro–Wilk or Shapiro–Francia) across variables and adjusting for the correlation structure.

Usage

```
royston(data, tol = 1e-25)
```

Arguments

data	A numeric matrix or data frame with observations in rows and variables in columns.
tol	Numeric tolerance passed to solve when inverting the covariance matrix. Default is 1e-25.

Value

A data frame with one row containing the test name (Test), the Royston test statistic (Statistic), and the associated p-value (p.value) from a chi-square approximation.

Examples

```
## Not run:
data <- iris[1:50, 1:4]
royston_result <- royston(data)
royston_result

## End(Not run)
```

---

summary.mvn *Summarize Multivariate Normality Analysis Results*

Description

Provides a structured summary of the results from an object of class mvn, including multivariate and univariate normality tests, descriptive statistics, and multivariate outlier detection (if applicable).

Usage

```
## S3 method for class 'mvn'
summary(
  object,
  select = c("mvn", "univariate", "descriptives", "outliers", "new_data"),
  ...
)
```

Arguments

object	An object of class mvn, as returned by the mvn function.
select	A character vector specifying which components to display. Must be one or more of "mvn", "univariate", "descriptives", "outliers", or "new_data". Defaults to showing all available sections.
...	Additional arguments (currently unused).

Value

Invisibly returns the input object.

Examples

```
## Not run:
data <- iris[1:4]
result <- mvn(data)

summary(result)  # Show all sections
summary(result, select = c("mvn", "outliers"))  # Show selected sections only

## End(Not run)
```

test_univariate_normality

Univariate Normality Tests

Description

Performs one of several common univariate normality tests on each numeric variable in a vector, matrix, or data frame.

Usage

```
test_univariate_normality(data, test = c("SW", "CVM", "Lillie", "SF", "AD"))
```

Arguments

`data`	A numeric vector, matrix, or data frame with observations in rows and variables in columns. Non-numeric columns are dropped with a warning. Each column is tested individually.
`test`	A character string specifying the normality test to use. Choices are: `"SW"` (Shapiro–Wilk), `"SF"` (Shapiro–Francia), `"AD"` (Anderson–Darling), `"CVM"` (Cramér–von Mises), and `"Lillie"` (Lilliefors test). Default is the first match from this list.

Value

A data frame with one row per variable and the following columns: `Test`, the name of the test used; `Variable`, the name of the tested variable; `Statistic`, the test statistic; and `p.value`, the associated p-value.

Examples

```
## Not run:
data(iris)
test_univariate_normality(iris[, 1:4], test = "AD")

## End(Not run)
```

univariate_diagnostic_plot

Diagnostic Plots for Univariate and Multivariate Data

Description

Generates QQ plots, histograms with density overlays, boxplots, or scatterplot matrices for numeric data (vector, matrix, or data frame).

Usage

```
univariate_diagnostic_plot(
  data,
  type = c("qq", "histogram", "boxplot", "scatter"),
  title = NULL,
  interactive = FALSE
)
```

Arguments

data	A numeric vector, matrix, or data frame with observations in rows and variables in columns.
type	Character; type of plot. One of: "qq", "histogram", "boxplot", "scatter". Default selects the first.
title	Character; plot title.
interactive	Logical; if TRUE, renders the plot interactively using plotly.

Examples

```
## Not run:
data <- iris[1:50, 1:3]
univariate_diagnostic_plot(data, type = "histogram")
univariate_diagnostic_plot(data, type = "qq")
univariate_diagnostic_plot(data, type = "boxplot")
univariate_diagnostic_plot(data, type = "scatter", interactive = TRUE)

## End(Not run)
```

Index