Package ‘HotellingEllipse’

July 5, 2024

Title Hotelling’s T-Squared Statistic and Ellipse

Version 1.2.0

Description Functions to calculate the Hotelling’s T-squared statistic and corresponding confidence ellipses. Provides the semi-axes of the Hotelling’s T-squared ellipses at 95% and 99% confidence levels. Enables users to obtain the coordinates in two or three dimensions at user-defined confidence levels, allowing for the construction of 2D or 3D ellipses with customized confidence levels. Bro and Smilde (2014) <DOI:10.1039/c3ay41907j>. Brereton (2016) <DOI:10.1002/cem.2763>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

LazyDataCompression xz

RoxygenNote 7.3.1

URL `https://github.com/ChristianGoueguel/HotellingEllipse`

BugReports `https://github.com/ChristianGoueguel/HotellingEllipse/issues`

Imports dplyr, FactoMineR, ggforce, ggplot2, lifecycle, magrittr, purrr, rgl, stats, tibble

Depends R (>= 2.10)

Suggests rmarkdown, knitr, markdown, testthat (>= 3.0.0), spelling, covr, scales, viridisLite

VignetteBuilder knitr

Config/testthat/edition 3

Language en-US

NeedsCompilation no

Author Christian L. Goueguel [aut, cre] (<`https://orcid.org/0000-0003-0521-3446`>)

Maintainer Christian L. Goueguel <`christian.goueguel@gmail.com`>

Repository CRAN

Date/Publication 2024-07-04 23:10:09 UTC

Contents

`ellipseCoord` *Coordinate Points of the Hotelling's T-squared Ellipse*

Description

This function calculates the coordinate points for drawing a Hotelling's T-squared ellipse based on multivariate data. It can generate points for both 2D and 3D ellipses.

Usage

```
ellipseCoord(x, pcx = 1, pcy = 2, pcz = NULL, conf.limit = 0.95, pts = 200)
```

Arguments

`x` A matrix, data frame or tibble containing scores from PCA, PLS, ICA, or other dimensionality reduction methods. Each column should represent a component, and each row an observation.

`pcx` An integer specifying which component to use for the x-axis (default is 1).

`pcy` An integer specifying which component to use for the y-axis (default is 2).

`pcz` An integer specifying which component to use for the z-axis for 3D ellipsoids. If `NULL` (default), a 2D ellipse is computed.

`conf.limit` A numeric value between 0 and 1 specifying the confidence level for the ellipse (default is 0.95, i.e., 95% confidence).

`pts` An integer specifying the number of points to generate for drawing the ellipse (default is 200). Higher values result in smoother ellipses.

Details

The function computes the shape and orientation of the ellipse based on the Hotelling's T-squared distribution and the specified components. It then generates a set of points that lie on the ellipse's surface at the specified confidence level. For 2D ellipses, the function uses two components `pcx` and `pcy`. For 3D ellipsoids, it uses three components `pcx`, `pcy`, and `pcz`. The `conf.limit` parameter determines the size of the ellipse. A higher confidence level results in a larger ellipse that encompasses more data points.

Value

A data frame containing the coordinate points of the Hotelling's T-squared ellipse:

- For 2D ellipses: columns `x` and `y`
- For 3D ellipsoids: columns `x`, `y`, and `z`

Author(s)

Christian L. Gougeguel christian.goueguel@gmail.com

Examples

```
## Not run:
# Load required libraries
library(HotellingEllipse)
library(dplyr)

data("specData", package = "HotellingEllipse")

# Perform PCA
set.seed(123)
pca_mod <- specData %>%
  select(where(is.numeric)) %>%
  FactoMineR::PCA(scale.unit = FALSE, graph = FALSE)

# Extract PCA scores
pca_scores <- pca_mod$ind$coord %>% as.data.frame()

# Example 1: Calculate Hotelling's T-squared ellipse coordinates
xy_coord <- ellipseCoord(pca_scores, pcx = 1, pcy = 2)

# Example 2: Calculate Hotelling's T-squared ellipsoid coordinates
xyz_coord <- ellipseCoord(pca_scores, pcx = 1, pcy = 2, pcz = 3)

## End(Not run)
```

ellipseParam *Hotelling's T-squared Statistic and Ellipse Parameters*

Description

This function calculates Hotelling's T-squared statistic and, when applicable, the lengths of the semi-axes of the Hotelling's ellipse. It can work with a specified number of components or use a cumulative variance threshold.

Usage

```
ellipseParam(
  x,
  k = 2,
  pcx = 1,
  pcy = 2,
  threshold = NULL,
  rel.tol = 0.001,
  abs.tol = .Machine$double.eps
)
```

Arguments

`x` A matrix, data frame or tibble containing scores from PCA, PLS, ICA, or other similar methods. Each column should represent a component, and each row an observation.

`k` An integer specifying the number of components to use (default is 2). This parameter is ignored if `threshold` is provided.

`pcx` An integer specifying which component to use for the x-axis when `k = 2` (default is 1).

`pcy` An integer specifying which component to use for the y-axis when `k = 2` (default is 2).

`threshold` A numeric value between 0 and 1 specifying the desired cumulative explained variance threshold (default is `NULL`). If provided, the function determines the minimum number of components needed to explain at least this proportion of total variance. When `NULL`, the function uses the fixed number of components specified by `k`.

`rel.tol` A numeric value specifying the minimum proportion of total variance a component should explain to be considered non-negligible (default is 0.001, i.e., 0.1%).

`abs.tol` A numeric value specifying the minimum absolute variance a component should have to be considered non-negligible (default is `.Machine$double.eps`).

Details

When `threshold` is used, the function selects the minimum number of `k` components that cumulatively explain at least the specified proportion of variance. This parameter allows for dynamic component selection based on explained variance, rather than using a fixed number of components. It must be greater than `rel.tol`. Typical values range from 0.8 to 0.95.

The `rel.tol` parameter sets a minimum variance threshold for individual components. Components with variance below this threshold are considered negligible and are removed from the analysis. Setting `rel.tol` too high may remove potentially important components, while setting it too low may retain noise or cause computational issues. Adjust based on your data characteristics and analysis goals.

Note that components are considered to have near-zero variance and are removed if their relative variance is below `rel_tol` or their absolute variance is below `abs_tol`. This dual-threshold approach helps ensure numerical stability while also accounting for the relative importance of components. The default value for `abs.tol` is set to `.Machine$double.eps`, providing a lower bound for detecting near-zero variance that may cause numerical instability.

Value

A list containing the following elements:

- `Tsquare`: A data frame containing the T-squared statistic for each observation.
- `Ellipse`: A data frame containing the lengths of the semi-minor and semi-major axes (only when `k = 2`).
- `cutoff.99pct`: The T-squared cutoff value at the 99% confidence level.

- `cutoff.95pct`: The T-squared cutoff value at the 95% confidence level.
- `nb.comp`: The number of components used in the calculation.

Author(s)

Christian L. Gouegue christian.goueguel@gmail.com

Examples

```
## Not run:
# Load required libraries
library(HotellingEllipse)
library(dplyr)

data("specData", package = "HotellingEllipse")

# Perform PCA
set.seed(123)
pca_mod <- specData %>%
  select(where(is.numeric)) %>%
  FactoMineR::PCA(scale.unit = FALSE, graph = FALSE)

# Extract PCA scores
pca_scores <- pca_mod$ind$coord %>% as.data.frame()

# Example 1: Calculate Hotelling's T-squared and ellipse parameters using
# the 2nd and 4th components
T2_fixed <- ellipseParam(x = pca_scores, pcx = 2, pcy = 4)

# Example 2: Calculate using the first 4 components
T2_comp <- ellipseParam(x = pca_scores, k = 4)

# Example 3: Calculate using a cumulative variance threshold
T2_threshold <- ellipseParam(x = pca_scores, threshold = 0.95)

## End(Not run)
```

specData *LIBS spectra of 100 soil samples*

Description

Data set of the emission spectra of 100 soils measured in laboratory conditions. The samples were cleaned, dried, homogenized, sieved (10 Mesh size) and thereafter pelletized prior to LIBS measurements. LIBS spectra were preprocessed by performing baseline removal.

Usage

```
specData
```

Format

Data frame of 100 rows (soil samples) and 3152 columns (wavelengths).

Source

doi:10.1039/C9JA00090A

Index