Package 'caplot'

June 6, 2023

Title Correspondence Analysis with Geometric Frequency Interpretation

Version 0.2

Description Performs Correspondence Analysis on the given dataframe and plots the results in a scatterplot that emphasizes the geometric interpretation aspect of the analysis, following Borg-Groenen (2005) and Yelland (2010). It is particularly useful for highlighting the relationships between a selected row (or column) category and the column (or row) categories. See Borg-Groenen (2005, ISBN:978-0-387-28981-6); Yelland (2010) <doi:10.3888/tmj.12-4>.

Depends R (>= 4.0.0)

Imports ca (>= 0.71), ggplot2 (>= 3.4.0), ggrepel (>= 0.9.0)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

NeedsCompilation no

Author Gianmarco Alberti [aut, cre]

Maintainer Gianmarco Alberti <gianmarcoalberti@gmail.com>

Repository CRAN

Date/Publication 2023-06-06 11:10:10 UTC

R topics documented:

borggroenen *Dataset: Cross-tabulation of crime rate vs. 10 US states*

Description

Cross-tabulation (7x10) of crime rate across 10 US states.
After: Borg-Groenen 2005 (Table 24.5).

Usage

```
data(borggroenen)
```

Format

dataframe

caplot *Correspondence Analysis with Geometric Frequency Interpretation*

Description

This function performs Correspondence Analysis on the given data frame and plots the results in a scatterplot that emphasizes the geometric interpretation aspect of the analysis (Borg-Groenen 2005; Yelland 2010). It is particularly useful for highlighting the relationships between a selected row (or column) category and the column (or row) categories.
Visit this LINK to access the package's vignette.

Usage

```
caplot(
  df,
  dims = c(1, 2),
  ref.category,
  dot.size = 2,
  label.size = 3,
  axis.title.size = 8,
  equal.scale = TRUE
)
```

Arguments

df	A cross-tabulation (dataframe) for which the correspondence analysis is performed.
dims	A numeric vector specifying the dimensions to be plotted (default: c(1,2)).
ref.category	The reference category for interpreting the plot. Must be a row or column name of the input dataframe. Note that, to enhance visual focus on the selected category, the other categories are rendered in grey.
dot.size	A numerical value representing the size of dots in the plot (default: 2)
label.size	A numerical value representing the size of labels in the plot (default: 3).
axis.title.size	A numerical value specifying the size of the axis titles (default: 8).
equal.scale	Logical value indicating whether to use the same scale for both axes (default: TRUE).

Details

Overview

The function follows a visualization approach outlined, e.g., by Borg-Groenen 2005 and Yelland 2010. This method allows the relative frequencies of categories in the dataset to be intuitively read off the plot. The function first draws a line through the origin and the point corresponding to the selected reference category. Perpendicular lines are then dropped from each category's position on the plot to the line connecting the reference category and the origin.

Interpretation

The relative frequencies of the categories can be inferred by looking at the positions at which the perpendiculars from the categories intersect this line. Categories with an intersection on the same side of the origin as the reference category occur more often than the average; the further from the origin an intersection occurs, the higher the frequency. Conversely, categories with an intersection on the opposite side occur less frequently; in this case, the further from the origin an intersection occurs, the smaller the frequency.

Limitations

The method implemented here complements, but does not replace, the nuanced interpretation of a correspondence analysis output gleaned from a typical scatterplot of row and column categories. While this approach may not be ideally suited for handling large tables, it provides a valuable tool for the visual interpretation of small-to-medium size tables, particularly for non-expert audiences. This method simplifies the understanding of some aspects of the data and can make the analytical results more accessible.

Value

A scatterplot visualizing the results of the Correspondence Analysis with geometric frequency interpretation.

References

Borg, I., & Groenen, P. J. F. (2005). Modern Multidimensional Scaling: Theory and Applications. Springer Science & Business Media.

Yelland, P. (2010). An Introduction to Correspondence Analysis. In The Mathematica Journal (Vol. 12). Wolfram Research, Inc.

Examples

```
# EXAMPLE 1
# Build a toy dataset (the famous Greenacre's "smoke" dataset).

mytable <- structure(list(none = c(4, 4, 25, 18, 10), light = c(2, 3, 10,
24, 6), medium = c(3, 7, 12, 33, 7), heavy = c(2, 4, 4, 13, 2
)), row.names = c("SM", "JM", "SE", "JE", "SC"), class = "data.frame")

# Run the function, using the "heavy" smoking as reference category

caplot(mytable, ref.category="heavy")

# In the returned scatterplot, it can be seen that the JM and SM categories
# feature a larger-than-average proportion of heavy smokers, whereas SC,
# SE, and JE feature a smaller-than-average proportion.
# Also, JM intersects the reference category line at a larger
# distance compared to the SM category, which indicates that the proportion
# of heavy smokers in JM is larger than in SM.
# Finally, SC features the smallest proportion of heavy smokers since it intersects the
# reference line at the furthest distance, on the side opposite to the plot's origin.
# This can be cross-checked by inspecting the table of row profiles:

row_props <- round(prop.table(as.matrix(mytable), margin = 1),3)


# EXAMPLE 2
# Run the function using the "yelland" in-built dataset and "MT2" (Mark Twain 2)
# as reference category, so reproducing the scatterplot in Yelland 2010, figure 4.

caplot(yelland, ref.category="MT2", label.size=2)


# EXAMPLE 3
# Run the function using the "borggroenen" in-built dataset and "MA"
# as reference category, so reproducing the scatterplot in Borg-Groenen 2005, figure 24.9.

caplot(borggroenen, ref.category="MA", label.size=2)

# As noted by Borg-Groenen 2005:
# "the projections on the line through the origin and MA (Massachusetts)
# show that auto theft and robbery happen more often than average.
# Because larceny and burglary project almost on the origin,
# they occur at an average rate in Massachusetts, whereas murder,
# rape, and assault are below average."
```

`yelland` *Dataset: Cross-tabulation of authors vs. alphabet letters*

Description

Cross-tabulation (15x16) of the frequency of alphabet letters across multiple authors.
After: Yelland 2010 (page 3).

Usage

```
data(yelland)
```

Format

dataframe

Index