Package 'mda.biber'

October 7, 2025

Title Functions for Multi-Dimensional Analysis

Version 1.0.1

Date 2025-09-22

Description Multi-Dimensional Analysis (MDA) is an adaptation of factor analysis developed by Douglas Biber (1992) <doi:10.1007/BF00136979>. Its most common use is to describe language as it varies by genre, register, and use. This package contains functions for carrying out the calculations needed to describe and plot MDA results: dimension scores, dimension means, and factor loadings.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Imports dplyr, ggplot2, ggpubr, ggrepel, nFactors, stats, tidyr, viridis

Depends R (>= 2.10)

Suggests rmarkdown, knitr, corrplot, kableExtra, tidyverse, testthat (>= 3.0.0)

VignetteBuilder knitr

NeedsCompilation no

Author David Brown [aut, cre] (ORCID: <https://orcid.org/0000-0001-7745-6354>), Alex Reinhart [aut] (ORCID: <https://orcid.org/0000-0002-6658-514X>)

Maintainer David Brown <dwb2@andrew.cmu.edu>

Repository CRAN

Date/Publication 2025-10-07 18:00:02 UTC

Contents

`boxplot_mda` *Create boxplot for multi-dimensional analysis*

Description

Combine scaled vectors of the relevant factor loadings and boxplots of dimension scores.

Usage

```
boxplot_mda(mda_data, n_factor = 1)
```

Arguments

`mda_data`	An mda data.frame produced by the `mda_loadings()` function.
`n_factor`	The factor to be plotted.

Value

A combined plot of scaled vectors and boxplots.

See Also

`stickplot_mda()`

`mda_loadings` *Conduct multi-dimensional analysis*

Description

Multi-Dimensional Analysis is a statistical procedure developed by Biber and is commonly used in descriptions of language as it varies by genre, register, and task. The procedure is a specific application of factor analysis, which is used as the basis for calculating a 'dimension score' for each text.

Usage

```
mda_loadings(obs_by_group, n_factors, cor_min = 0.2, threshold = 0.35)
```

Arguments

obs_by_group	A data frame containing exactly 1 categorical (factor) variable and multiple continuous (numeric) variables. Each row represents one document/observation.
n_factors	The number of factors to be calculated in the factor analysis.
cor_min	The correlation threshold for including variables in the factor analysis. Variables whose (absolute) Pearson correlation with any other variable is greater than this threshold will be included in the factor analysis. Set to 0 to disable thresholding.
threshold	The loading threshold above which variables should be included in factor score calculations. Set to 0 to include all variables.

Details

MDA is fundamentally factor analysis using the promax rotation, applied to the numeric variables in obs_by_group. However, MDA adds two screening steps:

1. Only variables with a nontrivial correlation with any other variable are included; the correlation threshold is configurable with the cor_min argument.
2. The factor scores are based only on variables whose loadings are greater (in absolute value) than the threshold argument. (Variables are standardized to ensure loadings are comparable.)

These two choices eliminate variables that are uncorrelated with others, and essentially enforce sparsity in each factor, ensuring it is loaded only on a smaller set of variables.

Value

An mda data frame containing one row per document, containing factor scores for each document. Attributes include the number of factors (n_factors), the correlation threshold (threshold), the factor loadings (loadings), and the mean factor score for each group (group_means).

References

Biber (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber (1992). "The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings." *Computers and the Humanities* 26 (5/6), 331-345. doi:10.1007/BF00136979

See Also

screeplot_mda(), stickplot_mda(), boxplot_mda()

Examples

```
# Extract the subject area from each document ID and use it as the grouping
# variable
micusp_biber$doc_id <- factor(substr(micusp_biber$doc_id, 1, 3))

m <- mda_loadings(micusp_biber, n_factors = 2)

attr(m, "group_means")
```

```
heatmap_mda(m)
```

`micusp_biber` *MICUSP corpus tagged with pseudobibeR features*

Description

The Michigan Corpus of Upper-Level Student Papers (MICUSP) contains 828 student papers. Here each document is tagged with Biber features using the pseudobibeR package. Type-to-token ratio is calculated using the moving average type-to-token ratio (MATTR).

Usage

```
micusp_biber
```

Format

A data frame with 828 rows and 68 columns:

doc_id Document ID (from MICUSP)

f_01_past_tense Rate of past tense per 1,000 tokens

f_02_perfect_aspect Rate of perfect aspect per 1,000 tokens

f_03_present_tense Rate of present tense per 1,000 tokens

f_04_place_adverbials Rate of place adverbials per 1,000 tokens

f_05_time_adverbials Rate of time adverbials per 1,000 tokens

f_06_first_person_pronouns Rate of first person pronouns per 1,000 tokens

f_07_second_person_pronouns Rate of second person pronouns per 1,000 tokens

f_08_third_person_pronouns Rate of third person pronouns per 1,000 tokens

f_09_pronoun_it Rate of pronoun 'it' per 1,000 tokens

f_10_demonstrative_pronoun Rate of demonstrative pronouns per 1,000 tokens

f_11_indefinite_pronouns Rate of indefinite pronouns per 1,000 tokens

f_12_proverb_do Rate of proverb 'do' per 1,000 tokens

f_13_wh_question Rate of wh-questions per 1,000 tokens

f_14_nominalizations Rate of nominalizations per 1,000 tokens

f_15_gerunds Rate of gerunds per 1,000 tokens

f_16_other_nouns Rate of other nouns per 1,000 tokens

f_17_agentless_passives Rate of agentless passives per 1,000 tokens

f_18_by_passives Rate of by-passives per 1,000 tokens

f_19_be_main_verb Rate of 'be' as main verb per 1,000 tokens

f_20_existential_there Rate of existential 'there' per 1,000 tokens

f_21_that_verb_comp Rate of that-verb complements per 1,000 tokens

f_22_that_adj_comp Rate of that-adjective complements per 1,000 tokens

f_23_wh_clause Rate of wh-clauses per 1,000 tokens

f_24_infinitives Rate of infinitives per 1,000 tokens

f_25_present_participle Rate of present participles per 1,000 tokens

f_26_past_participle Rate of past participles per 1,000 tokens

f_27_past_participle_whiz Rate of past participle whiz-deletions per 1,000 tokens

f_28_present_participle_whiz Rate of present participle whiz-deletions per 1,000 tokens

f_29_that_subj Rate of that-subject clauses per 1,000 tokens

f_30_that_obj Rate of that-object clauses per 1,000 tokens

f_31_wh_subj Rate of wh-subject clauses per 1,000 tokens

f_32_wh_obj Rate of wh-object clauses per 1,000 tokens

f_33_pied_piping Rate of pied-piping per 1,000 tokens

f_34_sentence_relatives Rate of sentence relatives per 1,000 tokens

f_35_because Rate of 'because' per 1,000 tokens

f_36_though Rate of 'though' per 1,000 tokens

f_37_if Rate of 'if' per 1,000 tokens

f_38_other_adv_sub Rate of other adverbial subordinators per 1,000 tokens

f_39_prepositions Rate of prepositions per 1,000 tokens

f_40_adj_attr Rate of attributive adjectives per 1,000 tokens

f_41_adj_pred Rate of predicative adjectives per 1,000 tokens

f_42_adverbs Rate of adverbs per 1,000 tokens

f_43_type_token Type-token ratio (MATTR)

f_44_mean_word_length Mean word length

f_45_conjuncts Rate of conjuncts per 1,000 tokens

f_46_downtoners Rate of downtoners per 1,000 tokens

f_47_hedges Rate of hedges per 1,000 tokens

f_48_amplifiers Rate of amplifiers per 1,000 tokens

f_49_emphatics Rate of emphatics per 1,000 tokens

f_50_discourse_particles Rate of discourse particles per 1,000 tokens

f_51_demonstratives Rate of demonstratives per 1,000 tokens

f_52_modal_possibility Rate of possibility modals per 1,000 tokens

f_53_modal_necessity Rate of necessity modals per 1,000 tokens

f_54_modal_predictive Rate of predictive modals per 1,000 tokens

f_55_verb_public Rate of public verbs per 1,000 tokens

f_56_verb_private Rate of private verbs per 1,000 tokens

f_57_verb_suasive Rate of suasive verbs per 1,000 tokens

f_58_verb_seem Rate of 'seem' verbs per 1,000 tokens

f_59_contractions Rate of contractions per 1,000 tokens

f_60_that_deletion Rate of that-deletions per 1,000 tokens

f_61_stranded_preposition Rate of stranded prepositions per 1,000 tokens

f_62_split_infinitve Rate of split infinitives per 1,000 tokens

f_63_split_auxiliary Rate of split auxiliaries per 1,000 tokens

f_64_phrasal_coordination Rate of phrasal coordination per 1,000 tokens

f_65_clausal_coordination Rate of clausal coordination per 1,000 tokens

f_66_neg_synthetic Rate of synthetic negation per 1,000 tokens

f_67_neg_analytic Rate of analytic negation per 1,000 tokens

Source

Michigan Corpus of Upper-Level Student Papers, https://elicorpora.info/main, tagged with the pseudobibeR package.

`screeplot_mda` *Scree plot for multi-dimensional analysis*

Description

The scree plot shows each factor along the X axis, and the proportion of common variance explained by that factor on the Y axis. The proportion of common variance explained is given by the factor eigenvalue.

Usage

```
screeplot_mda(obs_by_group, cor_min = 0.2)
```

Arguments

`obs_by_group`	A data frame containing 1 categorical (factor) variable and continuous (numeric) variables.
`cor_min`	The correlation threshold for including variables in the factor analysis.

Details

A wrapper for the `nFactors:nScree()` and `nFactors::plotnScree()` functions.

Value

Nothing returned

See Also

`mda_loadings()`

`stickplot_mda` *Plots of MDA factor group means and loadings*

Description

Stick plots show each group's mean loading along a factor, plotted along a positive/negative cline. Heatmaps show each variable's loading on a factor. `stickplot_mda()` produces just a stick plot, while `heatmap_mda()` places a heatmap alongside the stick plot.

Usage

```
stickplot_mda(mda_data, n_factor = 1)

heatmap_mda(mda_data, n_factor = 1)
```

Arguments

`mda_data`	An mda data frame produced by the `mda_loadings()` function.
`n_factor`	Index of the factor to be plotted.

Value

ggplot object

See Also

`boxplot_mda()`

Index