LAB: Multiple Correspondence Analysis (MCA)

2025 - 03 - 25

M1 MIDS/MFA/LOGOS Université Paris Cité Année 2024 Course Homepage Moodle



Besides the usual packages (tidyverse, ...), we shall require FactoMineR and related packages.

```
stopifnot(
  require(FactoMineR),
  require(factoextra),
  require(FactoInvestigate)
)
```

The GSS dataset

We will use data coming from the General Social Survey. The General Social Survey data can be retrieved using the gssr package. If needed, install the gssr package and its companion package gssrdoc.

```
stopifnot(
  require("gssr"),
  require("gssrdoc")
)
```

Loading required package: gssr

Package loaded. To attach the GSS data, type data(gss_all) at the console. For the codebook, type data(gss_dict). For the panel data and documentation, type e.g. data(gss_panel08_long) and data(gss_panel_ For help on a specific GSS variable, type ?varname at the console.

Loading required package: gssrdoc

The data we will use are panel data made available by the GSS. In order to explore them, it's very useful toload gss_dict and gss_panel_doc.

```
data(gss_dict)
data(gss_panel_doc)
```

Check the online help.

?gss_panel_doc

gss_dict and gss_panel_doc are dataframes that can be queried:

```
gss_panel_doc |>
  dplyr::sample_n(5) |>
  select(id, text) |>
  gt::gt()
```

id	text
relhh2	1632. What is (PERSON)'s relationship to head of household?
colrac	78. Or consider a person who believes that Blacks are genetically inferior. b. Should such
issp	1661. Status on International Social Survey Program Self-Completion Supplement.
rellife	Please tell me whether you strongly agree, agree, disagree, or strongly disagree with the fo
relate3	1621. Please tell me the names of the people who usually live in this household. Let's star

In the panel *questionnaire*, some questions have **race** in the field **id**. Check again the online help.

?race

The answer is something like that:

Question 24. What race do you consider yourself?

And it contains a brief summary of the answers given through the years.

Year	Black	Other	White	iap (NA)	Total
2010	311	183	1550	-	2044
2012	301	196	1477	-	1974
2014	386	262	1890	-	2538

(this is just an excerpt)

```
gss_panel_doc |>
filter(str_detect(id, "race")) |>
slice_sample(n=1, by=text) |>
select(id, description, text) |>
gt::gt()
```

id	description	text
race	RACE	24. What race do you consider yourself?
racecen1	RACECEN1	1602. What is your race? Indicate one or more races that you consider you
natrace	NATRACE	68. We are faced with many problems in this country, none of some of th
natracey	NATRACEY	69. We are faced with many problems in this country, none of some of th
intrace1	INTRACE1	What is your race? Indicate one or more races that you consider yourself

The data set we will use comes from year 2010 panel data.

data("gss_panel10_long")

gss_panel10_long

# A	tibble	: 6,132	2 x 1,200									
	firstid	wave	oversamp	sampcode	sa	ample		fc	orm	formwt	vpsu	vstrat
	<fct></fct>	<dbl></dbl>	<dbl></dbl>	<dbl+lbl></dbl+lbl>	<0	lbl+lb	L>	<0	lbl+lbl>	<dbl></dbl>	<dbl></dbl>	<dbl+></dbl+>
1	1	1	1	501	9	[2000	FP]	2	[ALTERNAT~	1	1	2240
2	1	2	1	501	9	[2000	FP]	2	[ALTERNAT~	1	1	2240
3	1	3	1	501	9	[2000	FP]	2	[ALTERNAT~	1	1	2240
4	2	1	1	501	9	[2000	FP]	1	[STANDARD~	1	1	2240
5	2	2	1	501	9	[2000	FP]	1	[STANDARD~	1	1	2240
6	2	3	1	501	9	[2000	FP]	1	[STANDARD~	1	1	2240
7	3	1	1	501	9	[2000	FP]	1	[STANDARD~	1	2	2240
8	3	2	1	501	9	[2000	FP]	1	[STANDARD~	1	2	2240
9	3	3	1	501	9	[2000	FP]	1	[STANDARD~	1	2	2240
10	4	1	1	501	9	[2000	FP]	2	[ALTERNAT~	1	2	2240
# i	6,122 m	nore ro	DWS									
# i	1,191 m	nore va	ariables:	samptype ·	<dł< td=""><td>ol+1bl></td><td>>, w</td><td>tpa</td><td>an12 <dbl+11< td=""><td>ol>,</td><td></td><td></td></dbl+11<></td></dł<>	ol+1bl>	>, w	tpa	an12 <dbl+11< td=""><td>ol>,</td><td></td><td></td></dbl+11<>	ol>,		
#	wtpan12	23 <db]< td=""><td>L+lbl>, wi</td><td>tpannr12 <</td><td>db]</td><td>L+1b1></td><td>, wt]</td><td>par</td><td>nr123 <dbl-< td=""><td>⊦lbl>,</td><td></td><td></td></dbl-<></td></db]<>	L+lbl>, wi	tpannr12 <	db]	L+1b1>	, wt]	par	nr123 <dbl-< td=""><td>⊦lbl>,</td><td></td><td></td></dbl-<>	⊦lbl>,		
#	id <db]< td=""><td>l+1b1>,</td><td>, mar1 <dł< td=""><td>ol+lbl>, ma</td><td>ar2</td><td>2 <dbl-< td=""><td>+lbl:</td><td>>,</td><td>mar3 <dbl+< td=""><td>Lbl>,</td><td></td><td></td></dbl+<></td></dbl-<></td></dł<></td></db]<>	l+1b1>,	, mar1 <dł< td=""><td>ol+lbl>, ma</td><td>ar2</td><td>2 <dbl-< td=""><td>+lbl:</td><td>>,</td><td>mar3 <dbl+< td=""><td>Lbl>,</td><td></td><td></td></dbl+<></td></dbl-<></td></dł<>	ol+lbl>, ma	ar2	2 <dbl-< td=""><td>+lbl:</td><td>>,</td><td>mar3 <dbl+< td=""><td>Lbl>,</td><td></td><td></td></dbl+<></td></dbl-<>	+lbl:	>,	mar3 <dbl+< td=""><td>Lbl>,</td><td></td><td></td></dbl+<>	Lbl>,		
#	mar4 <	ibl+lb]	L>, mar5 •	<dbl+lbl>,</dbl+lbl>	ma	ar6 <dl< td=""><td>51+1</td><td>bl></td><td>⊳, mar7 <db]< td=""><td>L+lbl>,</td><td></td><td></td></db]<></td></dl<>	51+1	bl>	⊳, mar7 <db]< td=""><td>L+lbl>,</td><td></td><td></td></db]<>	L+lbl>,		
#	mar8 <	ibl+lb]	L>, mar9 <	<dbl+lbl>,</dbl+lbl>	ma	ar10 <	ibl>	, n	nar11 <dbl+1< td=""><td>Lbl>, ma</td><td>ar12 <</td><td>ibl>,</td></dbl+1<>	Lbl>, ma	ar12 <	ibl>,
#	mar13 <	<dbl>,</dbl>	mar14 <dl< td=""><td>ol>, abany</td><td><0</td><td>lbl+lb</td><td>L>, a</td><td>abc</td><td>lefect <dbl-< td=""><td>+lbl>,</td><td></td><td></td></dbl-<></td></dl<>	ol>, abany	<0	lbl+lb	L>, a	abc	lefect <dbl-< td=""><td>+lbl>,</td><td></td><td></td></dbl-<>	+lbl>,		

At the beginning, the panel was made of roughly 2000 respondents. And the respondents were interviewed in 2010, 2012 and 2014 (the three waves).

```
gss_panel10_long |>
filter(wave==3, !is.na(id)) |>
select(firstid, wave, id, sex)
```

```
# A tibble: 1,304 x 4
```

	firstid	wave	id	se	ex
	<fct></fct>	<dbl></dbl>	<dbl+lbl></dbl+lbl>	• <c< td=""><td>lbl+lbl></td></c<>	lbl+lbl>
1	1	3	10001	1	[MALE]
2	2	3	10002	2	[FEMALE]
3	3	3	10003	2	[FEMALE]
4	6	3	10004	1	[MALE]
5	7	3	10005	2	[FEMALE]
6	9	3	10006	2	[FEMALE]
7	10	3	10007	2	[FEMALE]
8	11	3	10008	2	[FEMALE]
9	12	3	10009	1	[MALE]
10	13	3	10010	1	[MALE]
# i	i 1,294 m	nore r	ows		

, -

Some respondents left the survey. Attrition can be monitored within the next query.

```
gss_panel10_long |>
  select(wave, id) |>
  group_by(wave) |>
  summarize(observed = n_distinct(id),
            missing = sum(is.na(id)))
# A tibble: 3 x 3
  wave observed missing
  <dbl> <int> <int>
1 1 2044 0
```

2	2	1552	493
3	3	1305	740

The confidence topic

Do GSS surveys are made of a huge number of questions. Not all questions were asked to the respondents. Indeed, each question was asked to two thirds of the respondents. Questions were related to demographic features (Age, Sex, Level of education, Employment and so on). Answers to these questions can be used to determine whether the panel sample is representative. Data can be compared with last census data (there is a census every ten years in the US).

A number of questions are related to the *confidence* topic. Respondents were asked about the level of confidence they put in a wide array of institutions.

```
gss_panel_doc |>
filter(str_detect(text, "confidence")) |>
slice_sample(n=1, by=text) |>
select(id, description, text) |>
gt::gt()
```

id	description	text
confinan	CONFINAN	far as the people running these institutions are concerned, would you say
conbus	CONBUS	far as the people running these institutions are concerned, would you say
conclerg	CONCLERG	far as the people running these institutions are concerned, would you say
coneduc	CONEDUC	far as the people running these institutions are concerned, would you say
confed	CONFED	far as the people running these institutions are concerned, would you say
$\operatorname{conlabor}$	CONLABOR	far as the people running these institutions are concerned, would you say
conpress	CONPRESS	far as the people running these institutions are concerned, would you say
$\operatorname{conmedic}$	CONMEDIC	far as the people running these institutions are concerned, would you say
contv	CONTV	far as the people running these institutions are concerned, would you say
$\operatorname{conjudge}$	CONJUDGE	far as the people running these institutions are concerned, would you say
consci	CONSCI	far as the people running these institutions are concerned, would you say
$\operatorname{conlegis}$	CONLEGIS	far as the people running these institutions are concerned, would you say
conarmy	CONARMY	far as the people running these institutions are concerned, would you say

For institutions like *Science* or *Medicine*, or *Executive branch of federal government*, Respondents were asked whether they have: *Great deal of confidence*, *Only some confidence*, *Hardly any confidence* in the institution. The same Likert scale with 3 levels was used for all institutions.

i Question

From the $\tt gss_panel10_long$ dataset, extract columns corresponding to questions from the $confidence\ topic$

§ Solution

We first gather the id of the questions pertaining to the confidence topic.

```
confidence_ids <- gss_panel_doc |>
filter(str_detect(text, "confidence")) |>
slice_sample(n=1, by=text) |>
pull(id)
```

Then we protect the panel data on the questions pertaining to the confidence topic. We also keep the respondent key, sex, and the relevant wave. s

```
gss_panel10_long |>
filter(wave==1) |>
select(firstid, wave, id, sex, contains(confidence_ids)) |>
sample_n(10) |>
gt::gt()
```

-	firstid	wave	id_1: id	sex_1: RESPONDENTS SEX	confinan_1: CONFI	D IN BANKS &
-	286	1	286	1		3
	271	1	271	2		2
	647	1	647	1		NA
	1471	1	1471	2		2
	1739	1	1739	1		NA
	2016	1	2016	1		NA
	1618	1	1618	2		NA
	1849	1	1849	6 1		3
	1715	1	1715	2		2

panel_doc <- gssrdoc::gss_panel_doc</pre>

Table wrangling

Before proceeding to Multiple Correspondence Analysis (CA), let us select the set of active variables.

i Question

Project gss_panel10_long on firstid, wave, id, sex, and columns with names in

confinan conbus conclerg coneduc confed conlabor conpress conmedic contv conjudge co Filter so as to keep only wave 1.

Transform all relevant columns into factors.

🂡 Solution

```
confidence_panel10 <- gss_panel10_long |>
filter(wave==1) |>
select(firstid, wave, id, sex, contains(confidence_ids)) |>
mutate(across(starts_with("con"), as_factor)) |>
mutate(sex=as_factor(sex))
```

```
confidence_10 <- gss_panel_doc |>
  filter(id %in% confidence_ids) |>
  select(-ends_with("2"), -ends_with("3")) |>
  unnest(
    cols=c(properties_1), names_repair = "universal") |>
  unnest(
    cols=c(marginals_1), names_repair = "universal") |>
  mutate(text=str_squish(text)) |>
  mutate(text=str_match(text, "\\? [\\w]\\.(.*)")) |>
  mutate(text=text[,2]) |>
  filter(property == "Missing-data codes") |>
```

```
select(-starts_with("id"), -property, -`value...5`)
```

```
New names:
New names:
* `id` -> `id...1`
* `id` -> `id...6`
levels_table <- confidence_10 |>
filter(label!='Total' | is.na(label)) |>
distinct(`value...9`, label)
```

Elementary statistics

i Question

Use skimmr to summarize your dataset.

0

Solution

```
confidence_panel10 |>
skimr::skim(where(is.factor)) |>
select(-factor.ordered, -skim_type)|>
filter(str_starts(skim_variable,"con")) |>
gt::gt() |>
gt::fmt_number(decimals=3) |>
gt::tab_caption(
    caption = "GSS Panel data 2010 First wave, Topic confidence"
)
```

skim_variable	n_missing	$complete_rate$	$factor.n_unique$	factor.top_counts
confinan	680.000	0.667	3.000	ONL: 656, HAR: 564, AG: 14
conbus	705.000	0.655	3.000	ONL: 848, HAR: 317, AG: 17
conclerg	725.000	0.645	3.000	ONL: 720, HAR: 328, AG: 27
coneduc	681.000	0.667	3.000	ONL: 793, A G: 367, HAR: 20
confed	698.000	0.659	3.000	ONL: 620, HAR: 499, AG: 22
conlabor	744.000	0.636	3.000	ONL: 781, HAR: 373, AG: 14
conpress	689.000	0.663	3.000	ONL: 621, HAR: 594, AG: 14
conmedic	681.000	0.667	3.000	ONL: 647, A G: 558, HAR: 15
contv	681.000	0.667	3.000	ONL: 677, HAR: 521, AG: 16
conjudge	709.000	0.653	3.000	ONL: 706, A G: 400, HAR: 22
consci	738.000	0.639	3.000	ONL: 683, A G: 544, HAR: 79
conlegis	697.000	0.659	3.000	ONL: 635, HAR: 587, AG: 12
conarmy	688.000	0.663	3.000	A G: 720, ONL: 507, HAR: 12

```
confidence_panel10 |>
filter(is.na(conarmy)) |>
filter(is.na(conclerg)) |>
filter(is.na(conpress))
```

i Question

There are a lot of missing data in your data set. How are missing cells related? Drop rows with missing data in the confidence questions. What are the dimensions of your data set?

Solution

```
confidence_data <- confidence_panel10 |>
  mutate(across(starts_with("con"), .fns=\(x) !is.na(x), .names="not_{.col}_isna")) |>
  rowwise() |>
  filter(all(c_across(ends_with("isna")))) |>
  ungroup() |>
  select(-ends_with("isna"), -firstid, -wave, -econsci) |>
  mutate(across(starts_with("con"), fct_drop), sex=as_factor(sex))

confidence_data |>
  slice_sample(n=10) |>
```

```
gt::gt()
```

id 1: id	sex 1: RESPONDENTS SEX	Confinan 1: CONFID IN BANKS &	& FINANCIAL

453	MALE	ONLY SOME
1064	FEMALE	HARDLY ANY
19	MALE	ONLY SOME
1087	FEMALE	ONLY SOME
1585	FEMALE	HARDLY ANY
808	MALE	ONLY SOME
83	FEMALE	HARDLY ANY
67	MALE	ONLY SOME
973	FEMALE	A GREAT DEAL
1408	MALE	HARDLY ANY

i Question

In your explore possible associations between the answers to the different *confidence* questions?

How would you test possible independence between answer to *confidence in science* and *confidence in the Army*.

Solution

TODO

The case for using MCA

In order to construct a *field of ideological and political attitudes*, which will subsequently be used as a reference for locating the empirical typologies of response styles and survey compliance, we apply *multiple correspondence analysis* (MCA). MCA belongs to the family of techniques used in geometric data analysis (Le Roux and Rouanet 2004). It allows for the extraction of the most important dimensions in a set of categorical variables, and the graphical representation of variable categories and individuals relative to each other in a coordinate system. Distances between categories as well as individuals can be interpreted as a measure of (dis)similarity: If categories often co-appear in individual's responses, they are located close together in the space produced by MCA. Rare co-appearances, accordingly, result in a larger distance between the respective categories. Furthermore, *illustrative variables* can passively be projected into the field, a technique that has been termed *visual regression analysis* (Lebart et al. 1984). Whereas the space is determined by the distances between the categories of *active variables*, *passive variables* do not alter the structure of the constructed field, but appear in their average and hence most likely position.

From https://doi.org/10.1007/s11135-016-0458-3

In this lab, we won't look at the field of ideological and political attitudes, but rather at the field of confidence level in a variety of institutions.

Multiple Correspondance Analysis

MCA executive summary

The input of multiple correspondence analysis is a data frame X with n rows and p categorical columns. Multiple Correspondence Analysis starts by building the *indicator matrix*. The indicator matrix is built by *one-hot encoding* of each categorical variable.

- A categorical variable V_j (factor) with q levels is mapped to q $\{0,1\}$ -valued variables $V_{j,r}$ for $r \leq q$
- If levels are indexed by $\{1, ..., q\}$, if the value of the categorical variable V_j from row i is $k \in \{1, ..., q\}$, the bina^{\$}

$$k\mapsto \underbrace{0,\ldots,0}_{k-1},1,\underbrace{0,\ldots,0}_{q-k}$$

- The indicator matrix has as many rows as the data matrix
- The number of columns of the indicator matrix is the sum of the number of levels of the categorical variables/columns of the data matrix

• The indicator matrix is a numerical matrix. It is suitable for factorial methodss Recall X is the data matrix with n rows (individuals) and p categorical columns (variables)

For $j \in \{1, ..., p\}$, let J_j denote the number of levels (categories) of variable j Let $q = \sum_{j \leq p} J_j$ be the sum of the number of levels throughout the variables Let Z be the incidence matrix with n rows and q columns For $j \leq p$ and $k \leq J_j$, let $\langle j, k \rangle = \sum_{j' < j} J_{j'} + k$ Let $N = n \times p = \sum_{i \leq n} \sum_{j \leq p} X_{i,j}$ and

$$P = \frac{1}{N}Z$$

(the correspondence matrix for MCA)

➡ The row wise sums of correspondence matrix P are all equal to 1/n = p/NThe column wise sum of the correspondence matrix P for the kth level of the *j*th variable of X ($j \le p$) is

$$N_{\langle j,k\rangle}/N=f_{\langle j,k\rangle}/p$$

where $f_{(j,k)}$ stands for the relative frequency of level k of the jth variable

$$D_r = \frac{1}{n} \mathrm{Id}_n \qquad D_c = \mathrm{diag} \left(\frac{f_{\langle j, k \rangle}}{p} \right)_{j \leq p, k \leq J_j}$$

In MCA, we compute the SVD $U \times D \times V^{\top}$ of the standardized residuals matrix:

$$S = D_r^{-1/2} \times \left(P - r \times c^{\top}\right) \times D_c^{-1/2} = \sqrt{n} \left(P - r \times c^{\top}\right) \times D_c^{-1/2}$$

Coefficient $i, \langle j, k \rangle$ of S is

$$\frac{\mathbb{I}_{i,\langle j,k\rangle}-f_{\langle j,k\rangle}}{\sqrt{nf_{\langle j,k\rangle}/p}}$$

MCA consists in computing the SVD of the standardized residuals matrix $S = U \times D \times V^\top$

From the SVD, we get

• $D_r^{-1/2} \times U$ standardized coordinates of rows

- $D_c^{-1/2} \times V$ standardized coordinates of columns
- $D_r^{-1/2} \times U \times D$ principal coordinates of rows
- $D_c^{-1/2} \times V \times D$ principal coordinates of columns
- Squared singular values: the principal *inertia*

MCA and extended SVD

 As

$$D_r^{-1} \times P \times D_c^{-1} - \mathbf{II}^\top = (D_r^{-1/2} \times U) \times D \times (D_c^{-1/2} \times V)^\top$$

 $(D_r^{-1/2} \times U) \times D \times (D_c^{-1/2} \times V)^\top$

is the $extended\ SVD$ of

$$D_r^{-1} \times P \times D_c^{-1} - \mathbf{I}\mathbf{I}^{\top}$$

with respect to D_r and D_c

i Question

Perform MCA on the indicator matrix.

You may use FactoMineR::MCA(). It is interesting to compute the correspondence analysis in your own way, by preparing the matrix that is handled to svd() and returning a named list containing all relevant information.

? Solution

```
We lazyly rely on FactoMineR::MCA.
```

```
res.mca <- confidence_data |>
   select(starts_with("con")) |>
   FactoMineR::MCA(graph=F)
```

The result is an object of ${\tt S3}$ class ${\tt MCA}.$

\mathbf{S}

i Question

If you did use FactoMineR::MCA(), explain the organization of the result.

Solution

The result of FactoMineR::MCA(...) is a named and nested list with five elements: eig a matrix/array containing enough information to build a screeplot.

call a list of 9, containing the call to CA(), an object of type language, telling (in principle) the user how CA() was called. However, this is a *quoted expression*. Here we need to guess the value of y in the calling environment understand what's going on.

Element call also contains the table margin distributions marge.col and marge.row. The truncation rank ncp (number of components) can be assigned before computing the SVD (default value is 5). Element X stores the contingency table that was effectively used for computing Correpondence Analysis.

ind Information gathered from SVD to facilitate individuals (respindents) analysis.
var a list structured in the same way as element ind. Used for variables/categories
 profiles analysis

	Name	Description
1	\$eig	"eigenvalues"
2	\$var	"results for the variables (categories)"
3	<pre>\$var\$coord</pre>	"coord. of the categories"
4	\$var\$cos2	"cos2 for the categories"
5	<pre>\$var\$contrib</pre>	"contributions of the categories"
6	<pre>\$var\$v.test</pre>	"v-test for the categories"
7	\$ind	"results for the individuals"
8	<pre>\$ind\$coord</pre>	"coord. for the individuals"
9	<pre>\$ind\$cos2</pre>	"cos2 for the individuals"
10	<pre>\$ind\$contrib</pre>	"contributions of the individuals"
1	\$quali.sup	"results for the supplementary categorical variables"
2	\$quali.sup\$coord	"coord. for the supplementary categories"
3	\$quali.sup\$cos2	"cos2 for the supplementary categories"
4	\$quali.sup\$v.test	"v-test for the supplementary categories"
5	\$call	"intermediate results"
6	<pre>\$call\$marge.col</pre>	"weights of columns"
7	<pre>\$call\$marge.li</pre>	"weights of rows"

A res.mca\$svd is not part of the output!

Solution, output of print(res.mca)

- eig is computed from the singular values in res.mca\$svd
- var contains material for plotting information about categories and variables on factorial planes
- ind conatins material for plotting information about individuals on on factorial planes

▲ In principle, all relevant information can be gathered from components svd, call.marge.li, and call.marge.col.

Screeplots

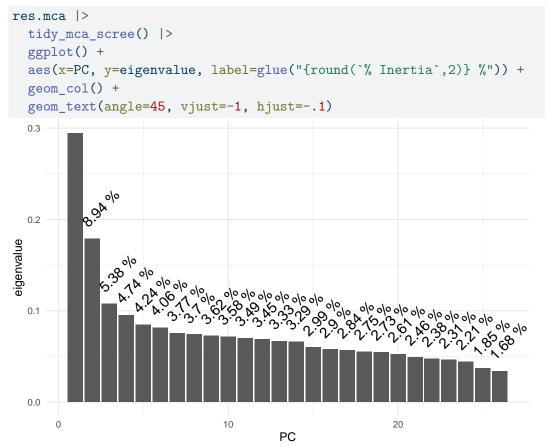
i Question

Draw screeplots. Why are they useful? Comment briefly.

Solution

We first define a tidying function for preparing the screeplot. It is designed along the lines of method tidy for objects of class prcomp. It consists in building a data frame from component eig of the MCA object.

We can now build a column plot.



Note that the share of cumulated inertia doesn't grow very fast with the number of components. Situation is different from what we met when performing Correspondence Analysis or Principal Component Analysis. This tells us that we will have to look beyond the first two axis. This is commonplace with multiple correspondence analysis. Look at the reference below where most of the interesting information can be found on the 3rd and 4th axis.

Individuals

i Question

Perform Individual profiles analysis. What are the classical plots? How can you build them from the output of FactoMiner::MCA? Build the table of row contributions (the so-called cos²)

Solution

Attribute ind of objects of class MCA (exported from FactoMineR) is the starting point of any row profiles analysis.

We design an augment method. For objects of class MCA. This is like what is done for objects of class prcomp and different from what we did with objects of class CA. Indeed, MCA on the indicator matrix handles a sample just like PCA and not a contingency table like CA.

```
augment_mca <- function(res_mca, data=NULL){
  result <- res_mca$ind$coord |>
    as_tibble()
  if (!is.null(data)) {
    result <- result |>
        bind_cols(confidence_data)
        # inner_join(confidence_data, by=c("id"))
  }
  return(result)
}
```

Solution 🕴

Attribute ind is a named list made of 4 components.

coord a matrix with *named* rows and columns. The number of rows of coord matches the number of rows of the table fed to MCA (here, the number of respondents in the panel). The number of columns matches the rank of the truncated SVD that underlies Multiple Correspondence Analysis. Here it is the default truncation 5.

The principal coordinates are the principal coordinates of each individual in terms of the principal component.

The columns of coord are pairwise orthogonal in the usual inner product space. Function augment_mca Bind the columns of coord with the columns of the original dataset.

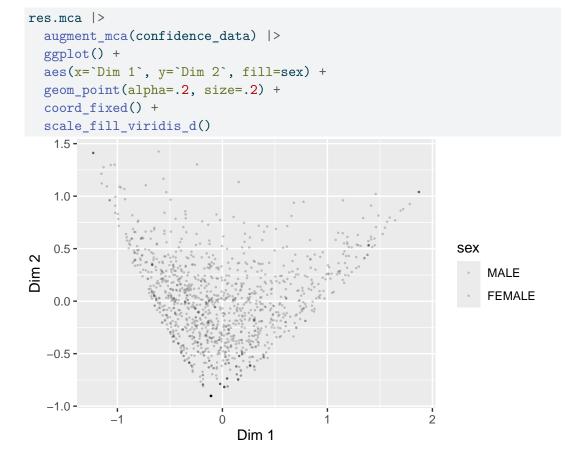
```
res.mca |>
#| label: augment_mca_glimpse
    augment_mca(confidence_data) |>
    glimpse()
```

```
Rows: 1,183
```

```
Columns: 20
                       <dbl> 1.35465733, -0.10861408, -0.33683136, 0.72450001, 0.72525336,~
   `Dim 1`
$
$ `Dim 2`
                        <dbl> 0.41190167, -0.90245225, -0.45183706, 0.25425970, 0.10225521,~
                        <dbl> 0.161371575, 0.010268886, 0.107542556, -0.009273224, -0.38139~
$
   `Dim 3`
   `Dim 4`
                        <dbl> 0.141129674, 0.310980059, -0.097880877, -0.085996832, 0.36106~
$
   `Dim 5`
                        <dbl> 0.227285260, 0.086562193, 0.016680275, -0.115998402, -0.06681~
$
$ id
                        <dbl+lbl> 4, 6, 7, 9, 11, 15, 19, 20, 21, 24, 26, 28, 31, 41, 4~
                        <dbl+lbl> 2, 1, 2, 2, 2, 2, 1, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, ~
$ sex
$ confinan <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, HARDLY ANY, ONLY SOME, ON~
                        <fct> ONLY SOME, 
$ conbus
$ conclerg <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, A GREAT DEAL, A GREAT DEA~
$ coneduc <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, ONLY SOME, A GREAT DEAL, ~
                        <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, A GREAT DEAL, A GREAT DEA~
$ confed
$ conlabor <fct> A GREAT DEAL, ONLY SOME, HARDLY ANY, ONLY SOME, HARDLY ANY, O~
$ conpress <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, HARDLY ANY, ONLY SOME, HA~
$ conmedic <fct> A GREAT DEAL, ONLY SOME, HARDLY ANY, A GREAT DEAL, ONLY SOME,~
$ contv
                        <fct> ONLY SOME, ONLY SOME, HARDLY ANY, ONLY SOME, ONLY ~
$ conjudge <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, A GREAT DEAL, A GREAT DEA~
                        <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, A GREAT DEAL, ONLY SOME, ~
$ consci
$ conlegis <fct> ONLY SOME, ONLY SOME, ONLY SOME, A GREAT DEAL, A GREAT DEAL, ~
$ conarmy <fct> A GREAT DEAL, ONLY SOME, ONLY SOME, A GREAT DEAL, A GREAT DEA~
Note that, we can recover ind$coord from the left singular vectors and the singular
```

Solution

cos2 Coefficients of matrix cos2 are the share of row inertia from the corresponding cell in coord



Solution

contrib

Not too surprisingly, $\tt coord,\ contrib,\ and\ cos2$ share the same row names and column names.

Variables/Categories

i Question

```
Perform column profiles (categories) analysis. You may use factoextra::fviz_mca_var()
```

i Question

What is the v.test component of the var component of an MCA object?

Solution 🕴

Here is a chatbot answer:

In Multiple Correspondence Analysis (MCA), the v-test (also known as the test-value or t-test value) is a statistical measure used to assess the significance of the association between a category and a dimension (factor). It helps determine which categories contribute the most to the definition of a given dimension.

Definition and Calculation

The v-test is computed as:

$$v = \frac{p - P}{\sqrt{P(1 - P)/n}}$$

where: - p = proportion of individuals in a given category for the specific dimension, - P = overall proportion of individuals in that category across all dimensions, - n = total number of individuals.

Alternatively, the v-test can also be expressed as:

 $v = \frac{\text{category coordinate}}{\text{standard deviation of the coordinate}}$

This essentially provides a **z-score**, which follows a standard normal distribution under the null hypothesis.

Interpretation

- A high absolute v-test value (typically above 2 or below -2) indicates that the category is strongly associated with the given dimension.
- *Positive v-test*: The category is overrepresented in that dimension.
- Negative v-test: The category is underrepresented in that dimension.
- If the absolute value of the v-test is below 2, the association is **not statistically significant** (at a 5% significance level).

Why is it useful?

- It helps *interpret MCA results quantitatively* rather than relying only on visual inspection.
- It aids in identifying *key categories* that structure the data on a given axis.
- It provides statistical validation for category contributions to dimensions.

Would you like a practical example or interpretation of a specific MCA result?

Symmetric plots

Question

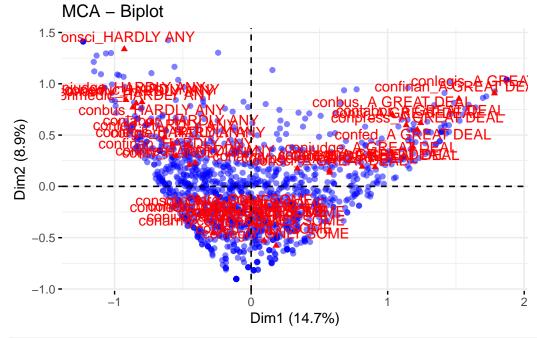
Build the symmetric plots (biplots) for multiple correspondence analysis.

From the shelf

We resort to factoextra. In principle, the biplot can be constructed from the coord matrices in the var and the ind components of the MCA object.

```
p <- res.mca |>
factoextra::fviz_mca_biplot(
    label= "var",
    alpha.ind=.5
)
```

р



p |> plotly::ggplotly()



Mosaicplots

MCA can be complemented by Mosaicplots, Double Decker plots, Chi-square tests, and Correspondence analyses between pair of variables.

i Question

Draw a mosaic plot to visualize association between confidence leves in Science and confidence level in Medicine.

Solution

```
confidence_data |>
  select(consci, conmedic) |>
  table() |>
  mosaicplot(color=T)
```

table(select(confidence_data, consci, conmedic))



Further references

Barth, Alice and Schmitz, Andreas. 2018. Response quality and ideological dispositions: an integrative approach using geometric and classifying techniques. Quality & Quantity

When analyzing survey data, response quality has consequential implications for substantial conclusions. Differences in response quality are usually explained by personality, or socio-demographic or cognitive characteristics. Little, however, is known about how respondents' political attitudes, values, and opinions impact on quality aspects. This is a striking analytical omission, as potential associations between political values and various forms of response biases and artefacts call into question surveys' ability to represent 'public opinion'. In this contribution, response quality is traced back to respondents' political and ideological dispositions. For this purpose, a relational understanding of response quality is applied that takes into account different aspects of response behaviors, as well as the interrelations between these indicators. Using data from the US General Social Survey (2010–2014), an empirical typology of response quality is created via finite mixture analysis. The resulting classes are then related to positions in the US field of ideological dispositions constructed via multiple correspondence analysis. The analyses reveal that there are (1) different combinations of response patterns and thus different empirical response types, and (2) that these types of response quality systematically vary with regard to the respondents' political and ideological (dis)positions. Implications of the findings for public opinion surveys are discussed.