Simple & Multiple Correspondence Analyses

Contingency, categorical, ordinal, continuous and mixed data

Derek Beaton

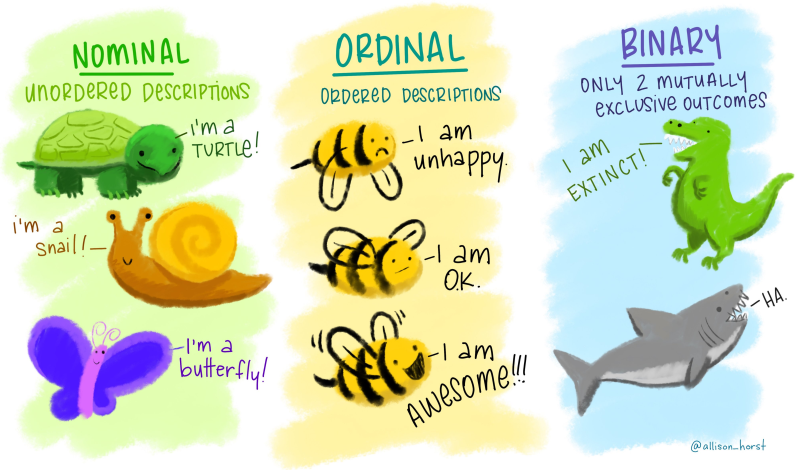
Rotman Research Institute

October 29, 2019

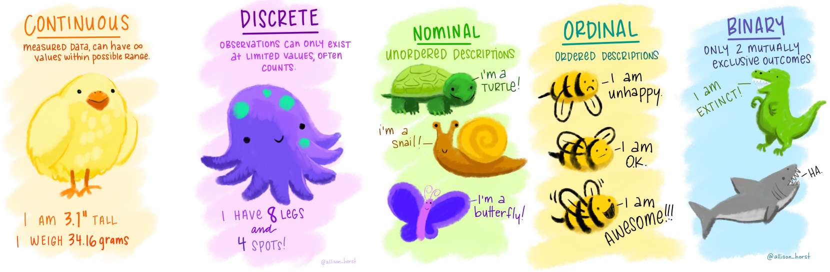
Before we get started

Our new best friends

via @allison_horst



via @allison_horst



- What do we do with all of these in a PCA like way?

- What do we do with all of these in a PCA like way?
- Some are *very* difficult and effectively ignored

- What do we do with all of these in a PCA like way?
- Some are *very* difficult and effectively ignored
  - We won't do that!

- What do we do with all of these in a PCA like way?
- Some are *very* difficult and effectively ignored
  - We won't do that!
- See SS Steven's typology: https://en.wikipedia.org/wiki/Level_of_measurement

Motivation & Objectives

- Not everything is a number

Motivation & Objectives

- Not everything is a number
    - Sometimes numbers aren't numbers!

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do
- Introduce CA, MCA, and tricks

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do
- Introduce CA, MCA, and tricks
  - Leave you overwhelmed, but knowing that

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do
- Introduce CA, MCA, and tricks
  - Leave you overwhelmed, but knowing that
  - PCA is sometimes the most wrong approach

Motivation & Objectives

- Not everything is a number
  - Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do
- Introduce CA, MCA, and tricks
  - Leave you overwhelmed, but knowing that
  - PCA is sometimes the most wrong approach
  - CA & MCA are suitably less wrong

Where to find everything

- Generally: https://github.com/derekbeaton/workshops

Where to find everything

- Generally: https://github.com/derekbeaton/workshops
- Today: https://github.com/derekbeaton/Workshops/tree/master/Misc/CA_MCA

Where to find everything

- Generally: https://github.com/derekbeaton/workshops
- Today: https://github.com/derekbeaton/Workshops/tree/master/Misc/CA_MCA
  - Built off of previous workshop

Where to find everything

- Generally: https://github.com/derekbeaton/workshops
- Today: https://github.com/derekbeaton/Workshops/tree/master/Misc/CA_MCA
  - Built off of previous workshop
  - Alzheimer's Disease Neuroimaging Initiative (ADNI) data

Overview

- Revisit PCA

Overview

- Revisit PCA
- Looking at some data

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses
  - Robustness, PLS, Software

Revisting PCA

What is PCA for?

- When we can compute a covariance or correlation matrix

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
  - Orthogonal

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
  - Orthogonal
  - Rank ordered

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
  - Orthogonal
  - Rank ordered
  - Made of bits & pieces of original measures

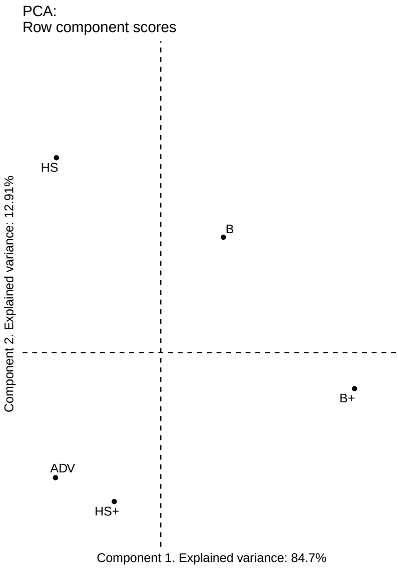
Some data

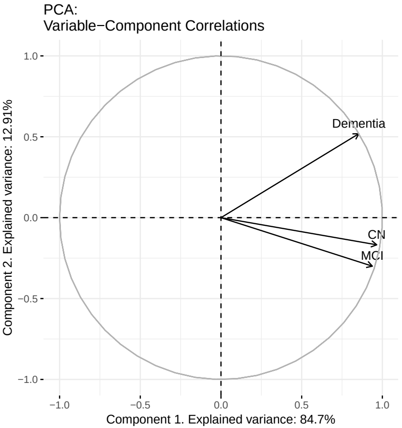
Diagnosis and education

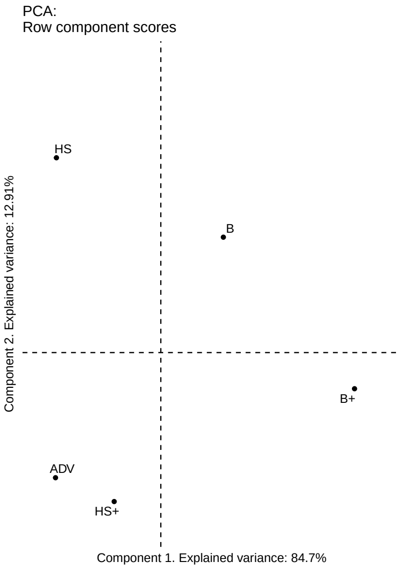
	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

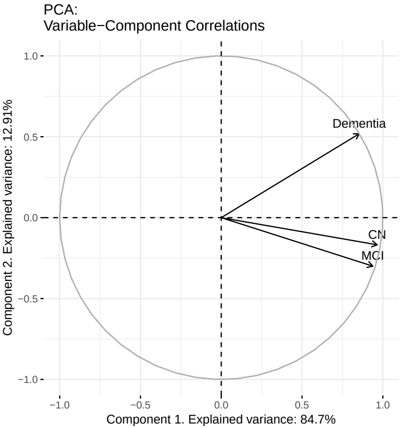
- Given a table, and asked for a multivariate analysis

- Given a table, and asked for a multivariate analysis
- We do what we know: PCA


PCA:
Row component scores
HS
B
B+
ADV
HS+
Component 1. Explained variance: 84.7%
Component 2. Explained variance: 12.91%


PCA:
Variable−Component Correlations
Component 2. Explained variance: 12.91%
Component 1. Explained variance: 84.7%
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Dementia
CN
MCI


PCA:
Row component scores
HS
B
B+
ADV
HS+
Component 1. Explained variance: 84.7%
Component 2. Explained variance: 12.91%


PCA:
Variable−Component Correlations
Dementia
CN
MCI
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Component 1. Explained variance: 84.7%
Component 2. Explained variance: 12.91%

What did we analyze?

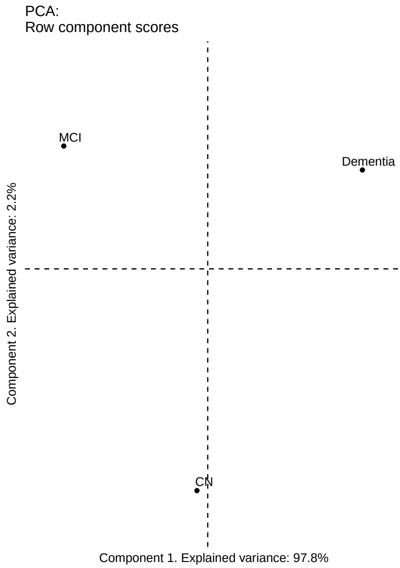
	CN	Dementia	MCI
CN	1.000	0.730	0.921
Dementia	0.730	1.000	0.652
MCI	0.921	0.652	1.000

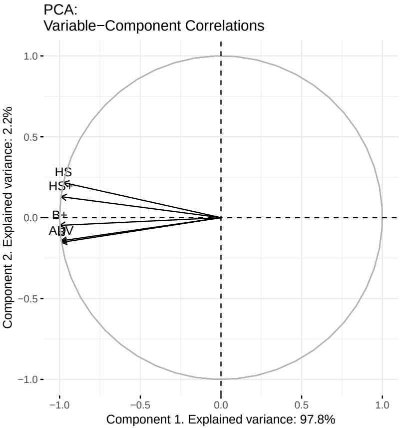
What did PCA detect?

	CN	Dementia	MCI	***Row sums***
ADV	39	7	54	***100***
B	57	17	75	***149***
B+	75	19	113	***207***
HS	25	13	46	***84***
HS+	39	9	77	***125***

Let's try something different!

	ADV	B	B+	HS	HS+
CN	39	57	75	25	39
Dementia	7	17	19	13	9
MCI	54	75	113	46	77


PCA:
Row component scores
MCI
Dementia
CN
Component 2. Explained variance: 2.2%
Component 1. Explained variance: 97.8%


PCA:
Variable−Component Correlations
HS
HS+
B+
ABV
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Component 2. Explained variance: 2.2%
Component 1. Explained variance: 97.8%

What did PCA analyze?

	ADV	B	B+	HS	HS+
ADV	1.000	1.000	0.995	0.935	0.963
B	1.000	1.000	0.994	0.932	0.960
B+	0.995	0.994	1.000	0.965	0.984
HS	0.935	0.932	0.965	1.000	0.996
HS+	0.963	0.960	0.984	0.996	1.000

What did PCA detect?

	ADV	B	B+	HS	HS+	***Row sums***
CN	39	57	75	25	39	***235***
Dementia	7	17	19	13	9	***65***
MCI	54	75	113	46	77	***365***

What is PCA for?

- When we can compute a *meaningful* covariance or correlation matrix

Let's take another look

	CN	Dementia	MCI	***Row sums***
ADV	39	7	54	***100***
B	57	17	75	***149***
B+	75	19	113	***207***
HS	25	13	46	***84***
HS+	39	9	77	***125***
Column sums	***235***	***65***	***365***	

- Tell me things about this matrix

Let's take another look

	CN	Dementia	MCI	***Row sums***
ADV	39	7	54	***100***
B	57	17	75	***149***
B+	75	19	113	***207***
HS	25	13	46	***84***
HS+	39	9	77	***125***
Column sums	***235***	***65***	***365***	

- Tell me things about this matrix
- What kind of problem does this look like?

Simple correspondence analysis

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)
  - Text (corpus) of philosphy, biblical passages, literature

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)
  - Text (corpus) of philosphy, biblical passages, literature
  - From Benzecri (1964) & Escofier (1965)

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)
  - Text (corpus) of philosphy, biblical passages, literature
  - From Benzecri (1964) & Escofier (1965)
  - Fully developed by Escofier (1969)

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)
  - Text (corpus) of philosphy, biblical passages, literature
  - From Benzecri (1964) & Escofier (1965)
  - Fully developed by Escofier (1969)
- Explosion of the technique in France

What is CA?

- Initially: *visualize contingency tables* (**a la PCA**, factor analyses)
  - Text (corpus) of philosphy, biblical passages, literature
  - From Benzecri (1964) & Escofier (1965)
  - Fully developed by Escofier (1969)
- Explosion of the technique in France
  - Across virtually every field (except psychology and neuroscience)

History

- Hotelling (1933) & Thurstone (1933)

History

- Hotelling (1933) & Thurstone (1933)
- Hirschfeld (1935) & Horst (1935)

History

- Hotelling (1933) & Thurstone (1933)
- Hirschfeld (1935) & Horst (1935)
- Guttman (1941)

History

- Hotelling (1933) & Thurstone (1933)
- Hirschfeld (1935) & Horst (1935)
- Guttman (1941)
- Burt (1950)

History

- Hotelling (1933) & Thurstone (1933)
- Hirschfeld (1935) & Horst (1935)
- Guttman (1941)
- Burt (1950)
- And then Benzecri (1964) & Escofier (1965)

History

- Hotelling (1933) & Thurstone (1933)
- Hirschfeld (1935) & Horst (1935)
- Guttman (1941)
- Burt (1950)
- And then Benzecri (1964) & Escofier (1965)
- Many more very important characters to re-discover CA

- See Lebart's History & Prehistory of CA

- See Lebart's History & Prehistory of CA
  - http://www.dtmvic.com/doc/About_the_History_of_CA.pdf

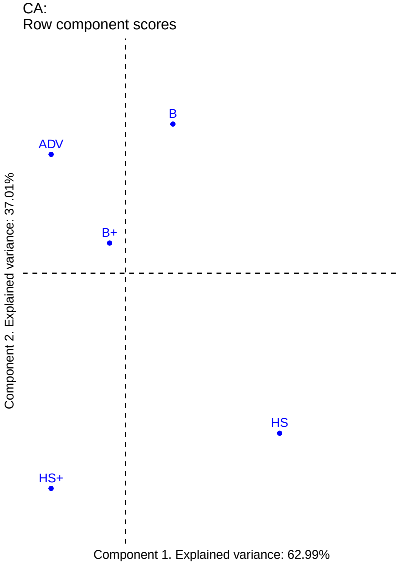
- See Lebart's History & Prehistory of CA
  - http://www.dtmvic.com/doc/About_the_History_of_CA.pdf
- And Beh & Lombardo's series

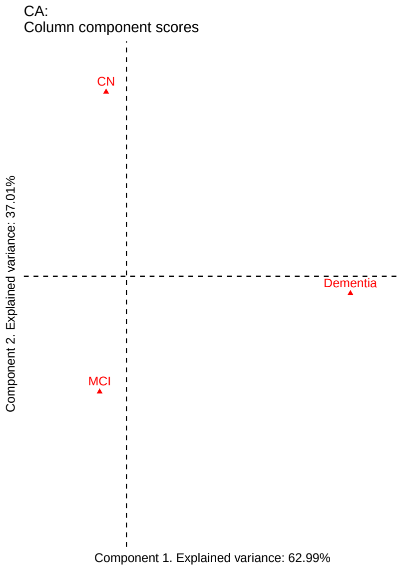
- See Lebart's History & Prehistory of CA
  - http://www.dtmvic.com/doc/About_the_History_of_CA.pdf
- And Beh & Lombardo's series
  - A geneaology of CA: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2012.00676.x

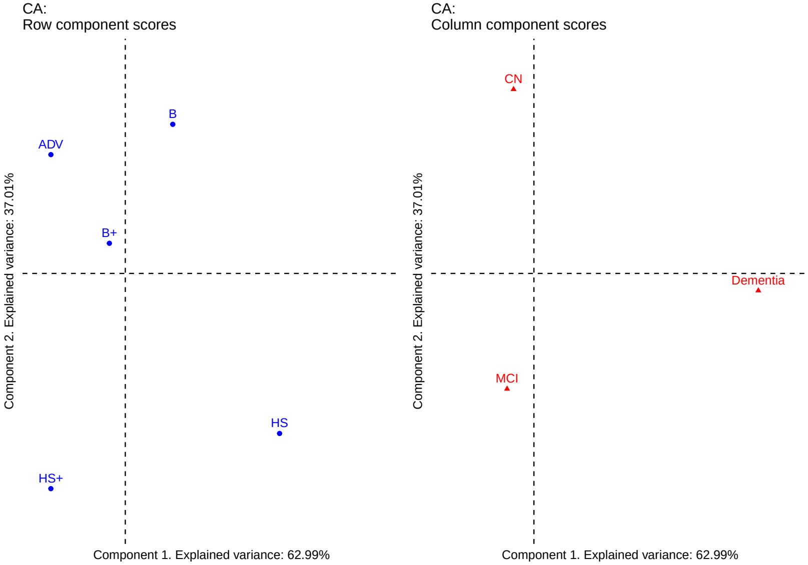
- See Lebart's History & Prehistory of CA
  - http://www.dtmvic.com/doc/About_the_History_of_CA.pdf
- And Beh & Lombardo's series
  - A geneaology of CA: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2012.00676.x
  - A geneaology of CA 2: http://siba-ese.unisalento.it/index.php/ejasa/article/view/19785

We're diving in

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77


CA:
Row component scores
B
ADV
B+
HS
HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%

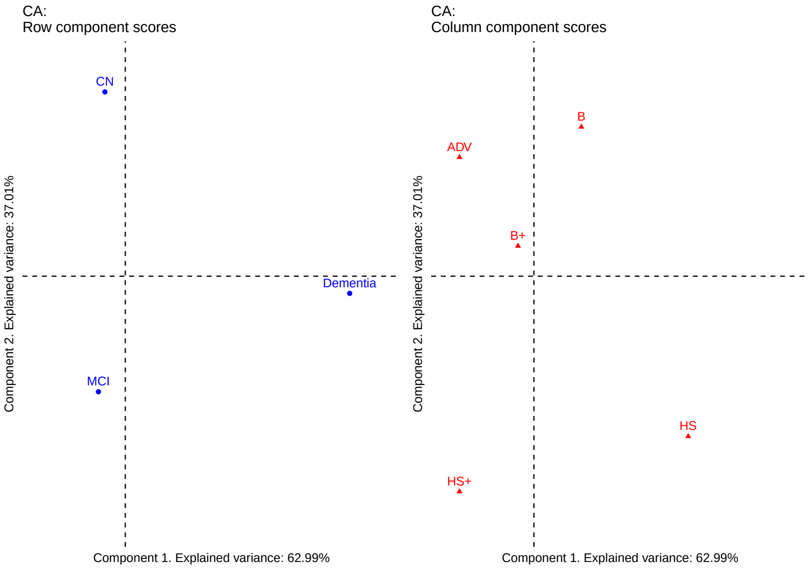

CA:
Column component scores
CN
Dementia
MCI
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%


CA:
Row component scores
ADV
B
B+
HS
HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%
CA:
Column component scores
CN
Dementia
MCI
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%

Want to see a cool trick?

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

	ADV	B	B+	HS	HS+
CN	39	57	75	25	39
Dementia	7	17	19	13	9
MCI	54	75	113	46	77


CA:
Row component scores
CN
Dementia
MCI
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%
CA:
Column component scores
B
ADV
B+
HS
HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%

How did that happen?

Table 1: Data

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

Table 2: Observed probabilites

	CN	Dementia	MCI
ADV	0.059	0.011	0.081
B	0.086	0.026	0.113
B+	0.113	0.029	0.170
HS	0.038	0.020	0.069
HS+	0.059	0.014	0.116

Table 3: Observed probabilites and margins

	CN	Dementia	MCI	***Row sums***
ADV	0.059	0.011	0.081	***0.150***
B	0.086	0.026	0.113	***0.224***
B+	0.113	0.029	0.170	***0.311***
HS	0.038	0.020	0.069	***0.126***
HS+	0.059	0.014	0.116	***0.188***
Column sums	***0.353***	***0.098***	***0.549***	

Table 4: Expected probabilites and margins

	CN	Dementia	MCI	***Row sums***
ADV	0.053	0.015	0.083	***0.150***
B	0.079	0.022	0.123	***0.224***
B+	0.110	0.030	0.171	***0.311***
HS	0.045	0.012	0.069	***0.126***
HS+	0.066	0.018	0.103	***0.188***
Column sums	***0.353***	***0.098***	***0.549***	

Table 5: Deviations: Observed - Expected

	CN	Dementia	MCI
ADV	0.006	-0.004	-0.001
B	0.007	0.004	-0.010
B+	0.003	-0.002	-0.001
HS	-0.007	0.007	0.000
HS+	-0.008	-0.005	0.013

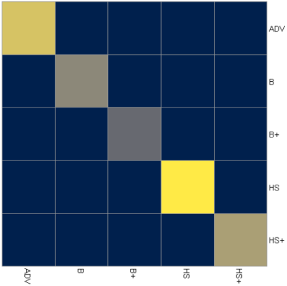
Table 6: Row constraints (inverse row margins)

	ADV	B	B+	HS	HS+
ADV	6.65	0.000	0.000	0.000	0.00
B	0.00	4.463	0.000	0.000	0.00
B+	0.00	0.000	3.213	0.000	0.00
HS	0.00	0.000	0.000	7.917	0.00
HS+	0.00	0.000	0.000	0.000	5.32

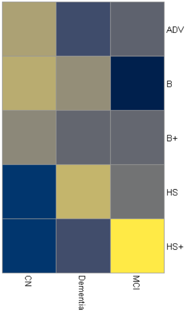
Table 7: Column constraints (inverse column margins)

	CN	Dementia	MCI
CN	2.83	0.000	0.000
Dementia	0.00	10.231	0.000
MCI	0.00	0.000	1.822

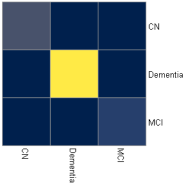
What CA needs



R: Row constraints
(inverse row probabilities)



Z: Deviations



C: Column constraints
(inverse column probabilities)

- GSVD($\mathbf{R}$, $\mathbf{X}$, $\mathbf{C}$)

- GSVD(**R**, **X**, **C**)
- Uses but generalizes the SVD

- GSVD(**R**, **X**, **C**)
- Uses but generalizes the SVD
  - Uses row & column weights (constraints)

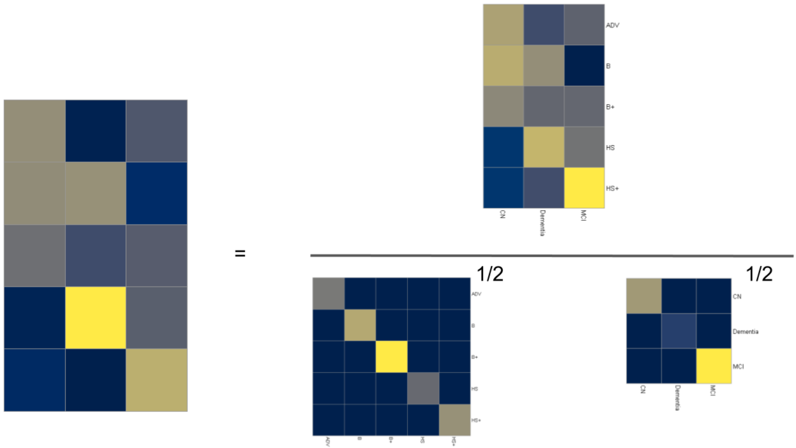
- GSVD($\mathbf{R}$, $\mathbf{X}$, $\mathbf{C}$)
- Uses but generalizes the SVD
  - Uses row & column weights (constraints)
- Gives back

- GSVD(**R**, **X**, **C**)
- Uses but generalizes the SVD
  - Uses row & column weights (constraints)
- Gives back
  - Component (factor) scores

- GSVD(**R**, **X**, **C**)
- Uses but generalizes the SVD
  - Uses row & column weights (constraints)
- Gives back
  - Component (factor) scores
  - Eigenvalues, singular values, & singular vectors

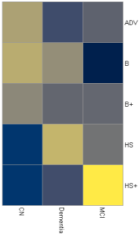
- GSVD(**R**, **X**, **C**)
- Uses but generalizes the SVD
  - Uses row & column weights (constraints)
- Gives back
  - Component (factor) scores
  - Eigenvalues, singular values, & singular vectors
  - *Generalized* singular vectors

What we really decompose



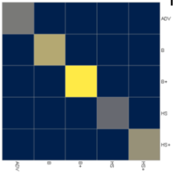
What we really decompose

- A rectangle
- Deviations: Observed - Expected
  - Expected from Observed's margins

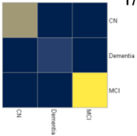


- Two squares
- Row margins and column margins

1/2



1/2



$$\frac{\mathbf{Z}}{\mathbf{R}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}}$$

$$\frac{(\mathbf{O} - \mathbf{E})}{\mathbf{E}^{\frac{1}{2}}}$$

$$\chi^2 = \Sigma \frac{(\mathbf{O} - \mathbf{E})^2}{\mathbf{E}}$$

CA's secrets

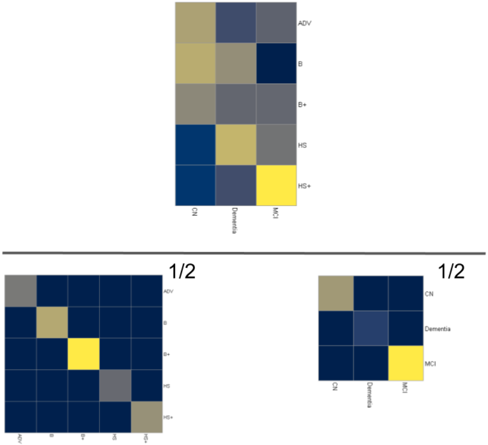
```r
EDU <- amerge_subset$PTEDUCAT
DX <- amerge_subset$DX
edu_dx_table <- table(EDU, DX)

chisq.test(edu_dx_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  edu_dx_table
## X-squared = 8.648, df = 8, p-value = 0.3729
```

```r
edu_dx_ca <- epCA(edu_dx_table, graphs = F)
sum(edu_dx_ca$ExPosition.Data$eigs) * sum(edu_dx_table)
```

```
## [1] 8.647979
```



Besides χ^2 this looks really familiar. What else are rectangles over squares?

$$r = \frac{cov(\mathbf{x}, \mathbf{y})}{\sigma_{\mathbf{x}} \times \sigma_{\mathbf{y}}}$$

More of CA's secrets

- CA generalizes canonical correlation analysis (CCA)

More of CA's secrets

- CA generalizes canonical correlation analysis (CCA)
- CA is the CCA between two *nominal* tables

More of CA's secrets

- CA generalizes canonical correlation analysis (CCA)
- CA is the CCA between two *nominal* tables
- How do we create a contingency table?

Nominal data

EDU	DX
B	Dementia
B	MCI
B+	Dementia
HS	Dementia
B+	CN

B	B+	ADV	HS+	HS
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	1	0	0	0

MCI	CN	Dementia
0	0	1
1	0	0
0	0	1
0	0	1
0	1	0

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

B	B+	ADV	HS+	HS
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	1	0	0	0

MCI	CN	Dementia
0	0	1
1	0	0
0	0	1
0	0	1
0	1	0

B	B+	ADV	HS+	HS
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	1	0	0	0

MCI	CN	Dementia
0	0	1
1	0	0
0	0	1
0	0	1
0	1	0

How to analyze this *on its own*?

How to analyze nominal data?

- “coding categorical variables with the indicator matrix of dummy variables and considering them as Gaussian, for instance, is almost a crime.” in *Jan de Leeuw and the French School of Data Analysis* (Husson, Josse, Saporta)

How to analyze nominal data?

- "coding categorical variables with the indicator matrix of dummy variables and considering them as Gaussian, for instance, is almost a crime." in *Jan de Leeuw and the French School of Data Analysis* (Husson, Josse, Saporta)
- We *could* perform PCA on nominal data, but what would we get?

	B	B+	ADV	HS+	HS	MCI	CN	Dementia
B	149	0	0	0	0	75	57	17
B+	0	207	0	0	0	113	75	19
ADV	0	0	100	0	0	54	39	7
HS+	0	0	0	125	0	77	39	9
HS	0	0	0	0	84	46	25	13
MCI	75	113	54	77	46	365	0	0
CN	57	75	39	39	25	0	235	0
Dementia	17	19	7	9	13	0	0	65

	B	B+	ADV	HS+	HS	MCI	CN	Dementia
B	1	−0.361	−0.226	−0.259	−0.204	−0.049	0.033	0.03
B+	−0.361	1	−0.283	−0.323	−0.256	−0.004	0.013	−0.013
ADV	−0.226	−0.283	1	−0.202	−0.16	−0.008	0.032	−0.039
HS+	−0.259	−0.323	−0.202	1	−0.183	0.065	−0.042	−0.042
HS	−0.204	−0.256	−0.16	−0.183	1	−0.001	−0.044	0.073
MCI	−0.049	−0.004	−0.008	0.065	−0.001	1	−0.815	−0.363
CN	0.033	0.013	0.032	−0.042	−0.044	−0.815	1	−0.243
Dementia	0.03	−0.013	−0.039	−0.042	0.073	−0.363	−0.243	1

Multiple correspondence analysis

Multiple correspondence analysis

- Two perspectives:

Multiple correspondence analysis

- Two perspectives:
  - *Weighted* PCA for nominal data

Multiple correspondence analysis

- Two perspectives:
  - *Weighted* PCA for nominal data
  - Generalized CA for N-way contingency tables

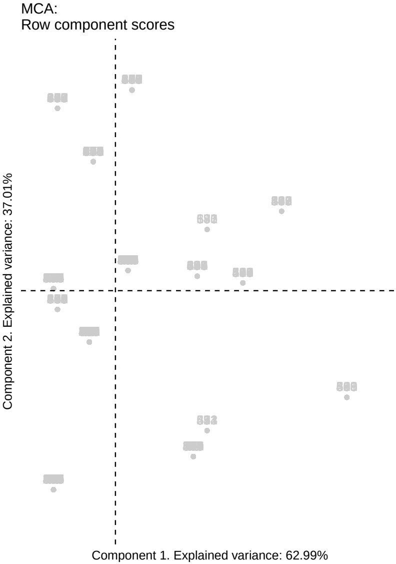
Multiple correspondence analysis

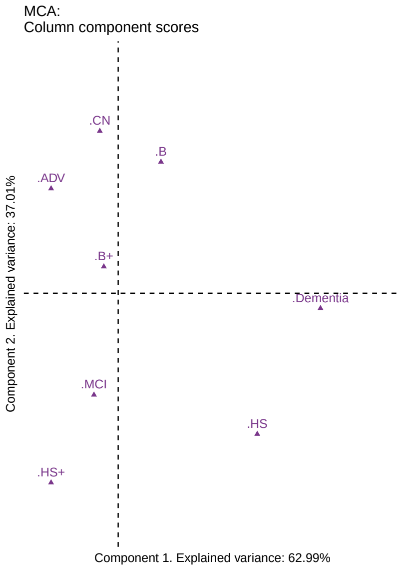
- Two perspectives:
  - *Weighted* PCA for nominal data
  - Generalized CA for N-way contingency tables
- So much more than nominal

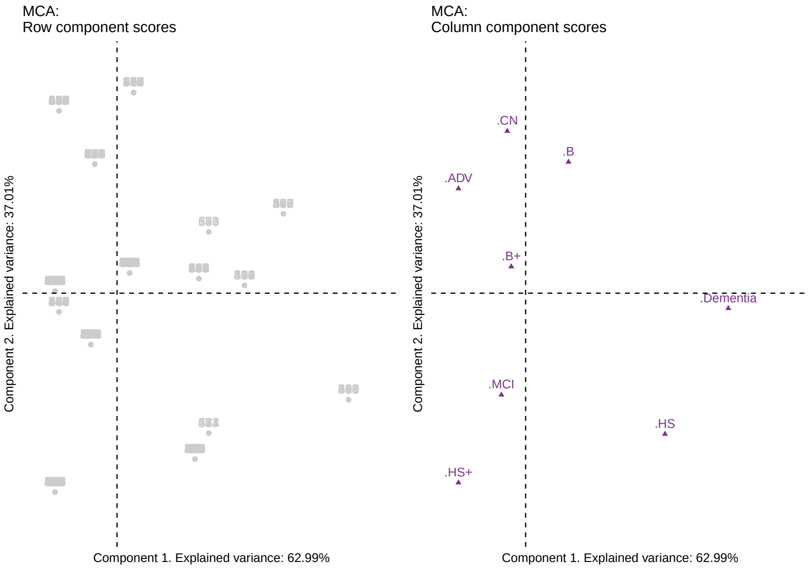
We're diving in

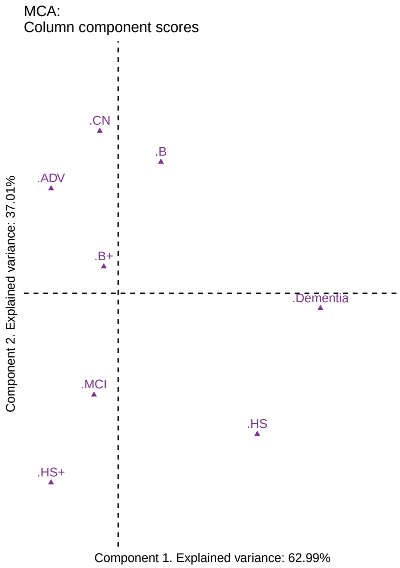
B	B+	ADV	HS+	HS	MCI	CN	Dementia
1	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1
0	0	0	0	1	0	0	1
0	1	0	0	0	0	1	0

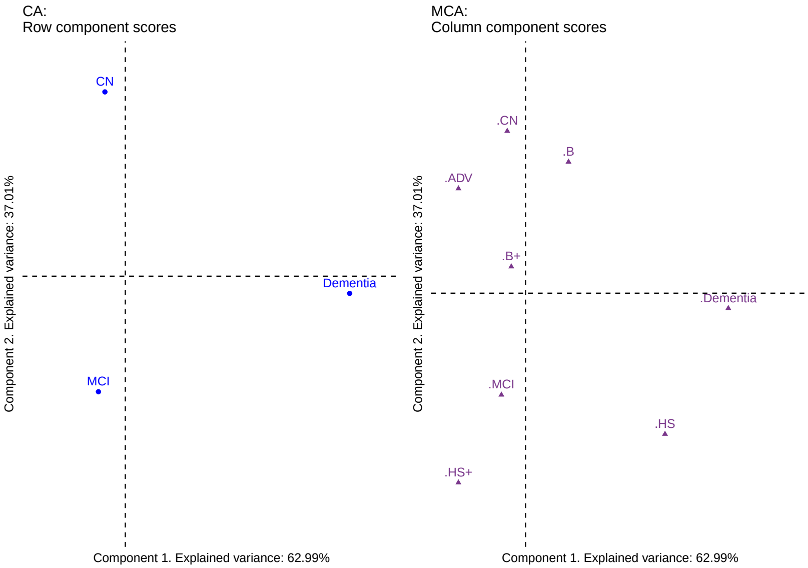
This is the kind of table we're analyzing. It has $N = 665$.


MCA:
Row component scores
Component 2. Explained variance: 37.01%
Component 1. Explained variance: 62.99%

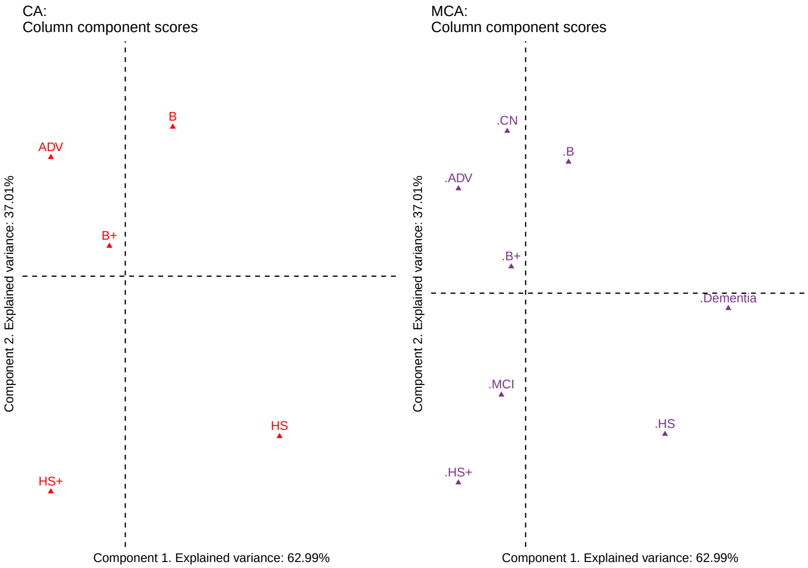

MCA:
Column component scores
.CN
.B
.ADV
.B+
.Dementia
.MCI
.HS
.HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%


MCA:
Row component scores
Component 2. Explained variance: 37.01%
Component 1. Explained variance: 62.99%
MCA:
Column component scores
.CN
.B
.ADV
.B+
.Dementia
.MCI
.HS
.HS+
Component 2. Explained variance: 37.01%
Component 1. Explained variance: 62.99%


MCA:
Column component scores
.CN
.B
.ADV
.B+
.Dementia
.MCI
.HS
.HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%



CA:
Row component scores
CN
Dementia
MCI
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%
MCA:
Column component scores
.CN
.B
.ADV
.B+
.Dementia
.MCI
.HS
.HS+
Component 1. Explained variance: 62.99%
Component 2. Explained variance: 37.01%


CA:
Column component scores
Component 2. Explained variance: 37.01%
Component 1. Explained variance: 62.99%
B
ADV
B+
HS
HS+
MCA:
Column component scores
Component 2. Explained variance: 37.01%
Component 1. Explained variance: 62.99%
.CN
.B
.ADV
.B+
.Dementia
.MCI
.HS
.HS+

CA & MCA Magic!

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

B	B+	ADV	HS+	HS	MCI	CN	Dementia
1	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1
0	0	0	0	1	0	0	1
0	1	0	0	0	0	1	0

Same technique on two *different* tables: same result

Scaling up

- Let's bring in ApoE

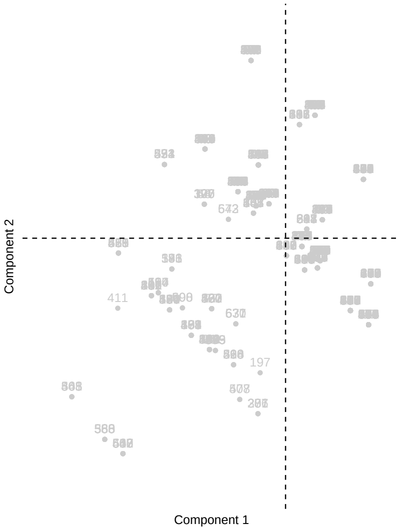
Scaling up

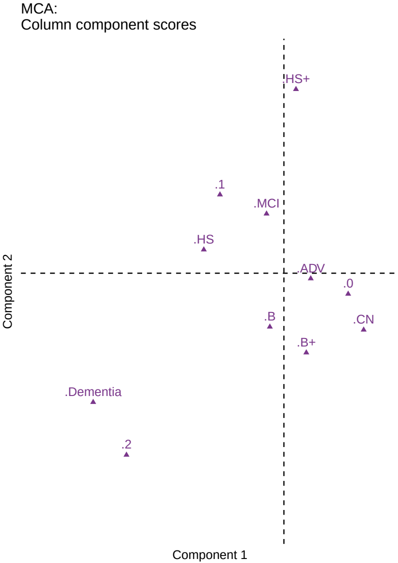
- Let's bring in ApoE
- It has 3 levels: 0 copy, 1 copy, 2 copies

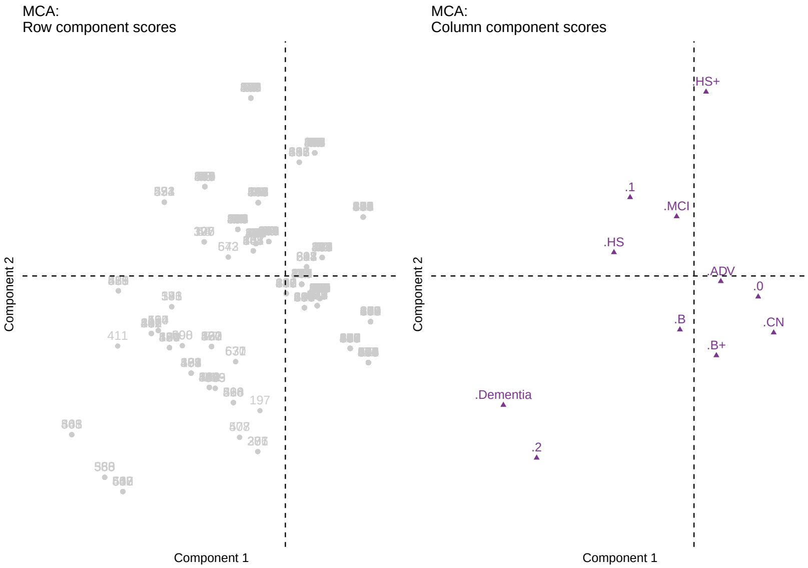
EDU	DX	APOE
B	Dementia	2
B	MCI	0
B+	Dementia	2
HS	Dementia	2
B+	CN	0

B	B+	ADV	HS+	HS	MCI	CN	Dementia	0	1	2
1	0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	1	0	0
0	1	0	0	0	0	0	1	0	0	1
0	0	0	0	1	0	0	1	0	0	1
0	1	0	0	0	0	1	0	1	0	0

MCA:
Row component scores




MCA:
Column component scores
.HS+
.1
.MCI
.HS
.ADV
.0
.B
.CN
.B+
.Dementia
.2
Component 2
Component 1


MCA:
Row component scores
Component 2
Component 1
411
197
MCA:
Column component scores
Component 2
Component 1
.HS+
.1
.MCI
.HS
.ADV
.0
.B
.CN
.B+
.Dementia
.2

Crisp vs. fuzzy coding

EDU	DX	APOE
B	Dementia	2
B	MCI	0
B+	Dementia	2
HS	Dementia	2
B+	CN	0

B	B+	ADV	HS+	HS	MCI	CN	Dementia	0	1	2
1	0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	1	0	0
0	1	0	0	0	0	0	1	0	0	1
0	0	0	0	1	0	0	1	0	0	1
0	1	0	0	0	0	1	0	1	0	0

Crisp vs. fuzzy coding

EDU	DX	APOE
B	Dementia	2
B	MCI	0
B+	Dementia	2
HS	Dementia	2
B+	CN	0

B	B+	ADV	HS+	HS	MCI	CN	Dementia	0	1	2
1	0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	0.5	0	0.5	1	0	0
0	1	0	0	0	0	0	1	0	0	1
0	0	0	0	1	0	0	1	0	0	1
0	1	0	0	0	0	1	0	1	0	0

Our first fuzzy friend

- Modified Hachinksi

- Modified Hachinksi
  - 0, 1, 2, 3 (in these data)

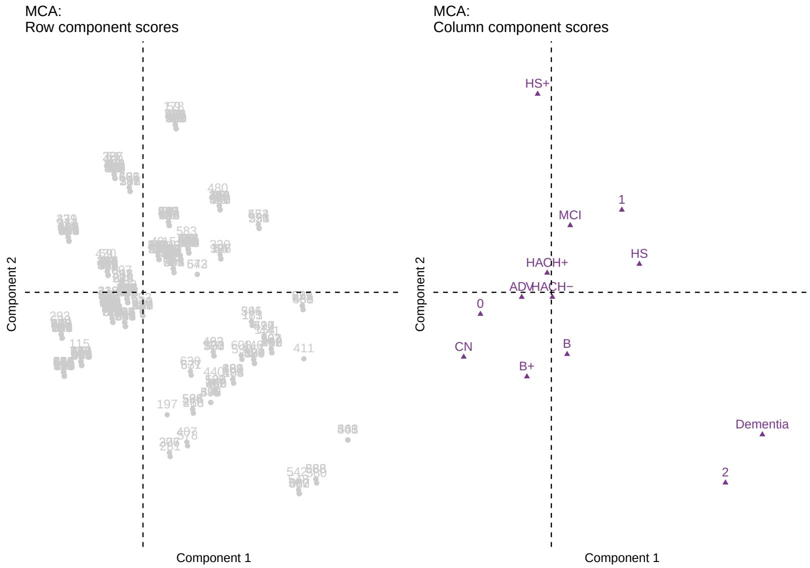
- Modified Hachinksi
    - 0, 1, 2, 3 (in these data)
- Specific form of fuzzy coding: “bipolar”

HACH
0
3
1
2
1

HACH−	HACH+
1	0
0	1
0.667	0.333
0.333	0.667
0.667	0.333

EDU	DX	APOE	HACH
B	Dementia	2	1
B	MCI	0	1
B+	Dementia	2	0
HS	Dementia	2	0
B+	CN	0	0

B	B+	ADV	HS+	HS	MCI	CN	Dementia	0	1	2	HACH−	HACH+
1	0	0	0	0	0	0	1	0	0	1	0.667	0.333
1	0	0	0	0	1	0	0	1	0	0	0.667	0.333
0	1	0	0	0	0	0	1	0	0	1	1	0
0	0	0	0	1	0	0	1	0	0	1	1	0
0	1	0	0	0	0	1	0	1	0	0	1	0


MCA:
Row component scores
Component 1
Component 2
MCA:
Column component scores
Component 1
Component 2
HS+
1
MCI
HS
HACH+
ADV
HACH−
0
CN
B
B+
Dementia
2

Our second fuzzy friend

- Age: 55.00 - 89.60

- Age: 55.00 - 89.60
  - But we need to scale it (Z-score)

- Age: 55.00 - 89.60
    - But we need to scale it (Z-score)
- We use two columns again:

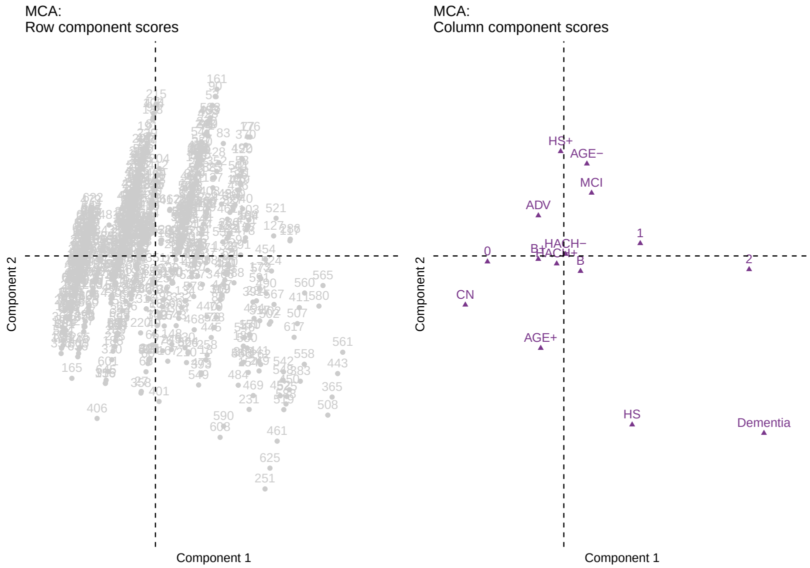
- Age: 55.00 - 89.60
  - But we need to scale it (Z-score)
- We use two columns again:
  - $\frac{(1-x)}{2}$ & $\frac{(1+x)}{2}$

AGE	AGE (Z)
76.3	0.637
76.5	0.666
64.4	−1.095
62.9	−1.314
63.9	−1.168

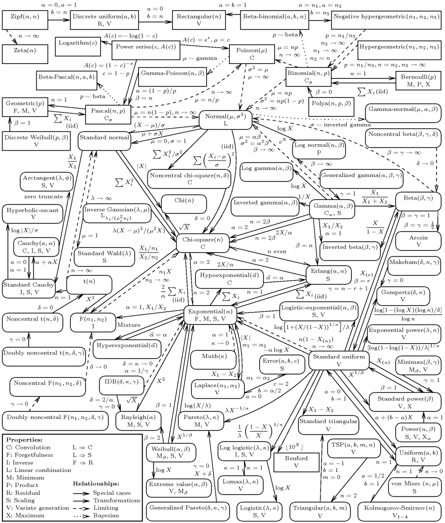
AGE−	AGE+
0.181	0.819
0.167	0.833
1.048	−0.048
1.157	−0.157
1.084	−0.084

EDU	DX	APOE	HACH	AGE
B	Dementia	2	1	76.3
B	MCI	0	1	76.5
B+	Dementia	2	0	64.4
HS	Dementia	2	0	62.9
B+	CN	0	0	63.9

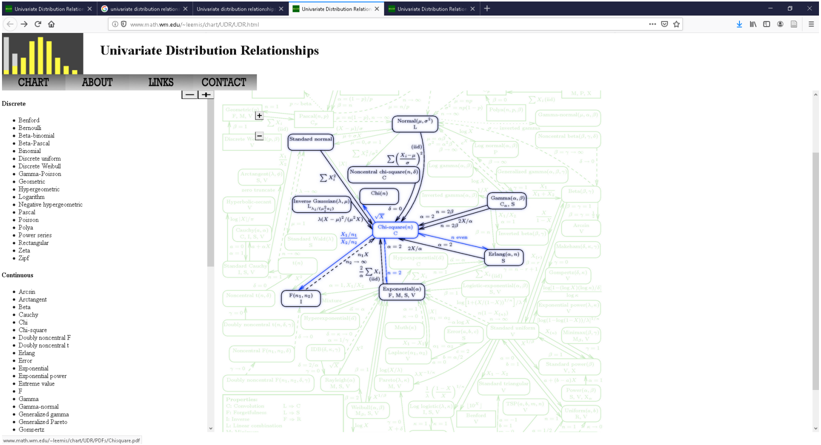
B	B+	ADV	HS+	HS	MCI	CN	Dementia	0	1	2	HACH−	HACH+	AGE−	AGE+
1	0	0	0	0	0	0	1	0	0	1	0.667	0.333	0.181	0.819
1	0	0	0	0	1	0	0	1	0	0	0.667	0.333	0.167	0.833
0	1	0	0	0	0	0	1	0	0	1	1	0	1.048	−0.048
0	0	0	0	1	0	0	1	0	0	1	1	0	1.157	−0.157
0	1	0	0	0	0	1	0	1	0	0	1	0	1.084	−0.084


MCA:
Row component scores
Component 2
Component 1
MCA:
Column component scores
HS+
AGE−
MCI
ADV
1
0
B+
HACH−
HACH+
B
2
CN
AGE+
HS
Dementia
Component 2
Component 1

How is this magic possible?!


Zipf(α, n)
Discrete uniform(a, b) R, V
Rectangular(n) V
Beta-binomial(a, b, n)
Negative hypergeometric(n1, n2, n3)
Zeta(α)
Logarithm(c)
Power series(c, A(c))
Poisson(μ) C
Hypergeometric(n1, n2, n3)
Beta-Pascal(n, a, b)
Gamma-Poisson(α, β)
Binomial(n, p) Cp
Bernoulli(p) M, P, X
Geometric(p) F, M, V
Pascal(n, p) Cp
Normal(μ, σ²) L
Polya(n, p, β)
Gamma-normal(μ, α, β)
Discrete Weibull(p, β) V
Standard normal
Noncentral beta(β, γ, δ)
Log normal(α, β) P
Log gamma(α, β)
Noncentral chi-square(n, δ) C
Generalized gamma(α, β, γ)
Arctangent(λ, φ) S, V
Chi(n)
Inverted gamma(α, β)
Gamma(α, β) Cα, S
Beta(β, γ)
Hyperbolic-secant V
Inverse Gaussian(λ, μ)
Arcsin V
Chi-square(n) C
Cauchy(a, α) C, I, S, V
Standard Wald(λ) S
Inverted beta(β, γ)
Makeham(δ, κ, γ)
Hypoexponential(α⃗) C
Erlang(α, n) S
Standard Cauchy I, S, V
t(n)
Gompertz(δ, κ) V
Noncentral t(n, δ)
F(n1, n2) I
Exponential(α) F, M, S, V
Logistic-exponential(α, β) S, V
Exponential power(λ, κ) V
Doubly noncentral t(n, δ, γ)
Hyperexponential(α⃗)
Muth(κ)
Error(a, b, c) S
Standard uniform V
Minimax(β, γ) Mβ, V
Noncentral F(n1, n2, δ)
IDB(δ, κ, γ)
Laplace(α1, α2) V
Standard power(β) V, X
Doubly noncentral F(n1, n2, δ, γ)
Rayleigh(α) M, S, V
Pareto(λ, κ) M, V
Standard triangular V
Power(α, β) S, V, Xα
Weibull(α, β) Mβ, S, V
Log logistic(λ, κ) I, S, V
Benford V
TSP(a, b, m, n) V
Uniform(a, b) R, V
Extreme value(α, β) V, Mβ
Lomax(λ, κ) V
von Mises (κ, μ) S
Generalized Pareto(δ, κ, γ)
Logistic(λ, κ) S, V
Triangular(a, b, m) V
Kolmogorov-Smirnov(n) V1–4
Properties:
C: Convolution
F: Forgetfulness
I: Inverse
L: Linear combination
M: Minimum
P: Product
R: Residual
S: Scaling
V: Variate generation
X: Maximum
L ⇒ C
L ⇒ S
F ⇒ R
Relationships:
Special cases
Transformations
Limiting
Bayesian

See here


www.math.wm.edu/~leemis/chart/UDR/UDR.html
Univariate Distribution Relationships
CHART
ABOUT
LINKS
CONTACT
Discrete
Benford
Bernoulli
Beta-binomial
Beta-Pascal
Binomial
Discrete uniform
Discrete Weibull
Gamma-Poisson
Geometric
Hypergeometric
Logarithm
Negative hypergeometric
Pascal
Poisson
Polya
Power series
Rectangular
Zeta
Zipf
Continuous
Arcsin
Arctangent
Beta
Cauchy
Chi
Chi-square
Doubly noncentral F
Doubly noncentral t
Erlang
Error
Exponential
Exponential power
Extreme value
F
Gamma
Gamma-normal
Generalized gamma
Generalized Pareto
Gompertz
Normal(μ, σ²)
Standard normal
Noncentral chi-square(n, δ)
Chi(n)
Inverse Gaussian(λ, μ)
Chi-square(n)
Gamma(α, β)
Erlang(α, n)
F(n₁, n₂)
Exponential(α)
www.math.wm.edu/~leemis/chart/UDR/PDFs/Chisquare.pdf

See here

Conclusions

- There's so much I'm not telling you

Conclusions

- There's so much I'm not telling you
  - I wish I could!

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point
  - CA is a *massive* world

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point
  - CA is a *massive* world
  - Solves many problems we typically ignore

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point
  - CA is a *massive* world
  - Solves many problems we typically ignore
  - And the *G*SVD needs to be your other new best friend

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point
  - CA is a *massive* world
  - Solves many problems we typically ignore
  - And the GSVD needs to be your other new best friend
- Learn to recognize data types

Conclusions

- There's so much I'm not telling you
  - I wish I could!
- But that's kind of the point
  - CA is a *massive* world
  - Solves many problems we typically ignore
  - And the *G*SVD needs to be your other new best friend
- Learn to recognize data types
  - Learn what to do with them

Conclusions

- There's so much I'm not telling you
    - I wish I could!
- But that's kind of the point
    - CA is a *massive* world
    - Solves many problems we typically ignore
    - And the *G*SVD needs to be your other new best friend
- Learn to recognize data types
    - Learn what to do with them
    - But

Conclusions

- Ordinal—and ordinal like—are very difficult

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?
  - Or are these counts, discrete, imprecise continuous?

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?
  - Or are these counts, discrete, imprecise continuous?
- What about:

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?
  - Or are these counts, discrete, imprecise continuous?
- What about:
  - 0, 1, 2, 3, 4?

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?
  - Or are these counts, discrete, imprecise continuous?
- What about:
  - 0, 1, 2, 3, 4?
  - Never, Rarely, Sometimes, Frequently, Always?

Conclusions

- Ordinal—and ordinal like—are very difficult
- Practical advice
  - Small number of levels? Treat them as categorical
  - HUGE number of levels? Inspect thoroughly
  - In between? Probably ordinal
- Are they *really* ordinal?
  - Ranked, ordered? No inherent value?
  - Or are these counts, discrete, imprecise continuous?
- What about:
  - 0, 1, 2, 3, 4?
  - Never, Rarely, Sometimes, Frequently, Always?
  - Use a mixture

	ASKS_HELP	NO_TIME_SELF
Ordinal responses		
PARTICIPANT 1	1	2
PARTICIPANT 2	2	3
PARTICIPANT 3	0	1
PARTICIPANT *n*	3	0
PARTICIPANT *N*	1	4

	ASKS_HELP			NO_TIME_SELF		
	NEVER	RARELY	ALWAYS	NEVER	RARELY	ALWAYS
Ordinal responses						
PARTICIPANT 1	0	1	0	0	.667	.333
PARTICIPANT 2	0	.667	.333	0	.333	.667
PARTICIPANT 3	1	0	0	0	1	0
PARTICIPANT *n*	0	.333	.667	1	0	0
PARTICIPANT *N*	0	1	0	0	0	1

Some many bonuses!

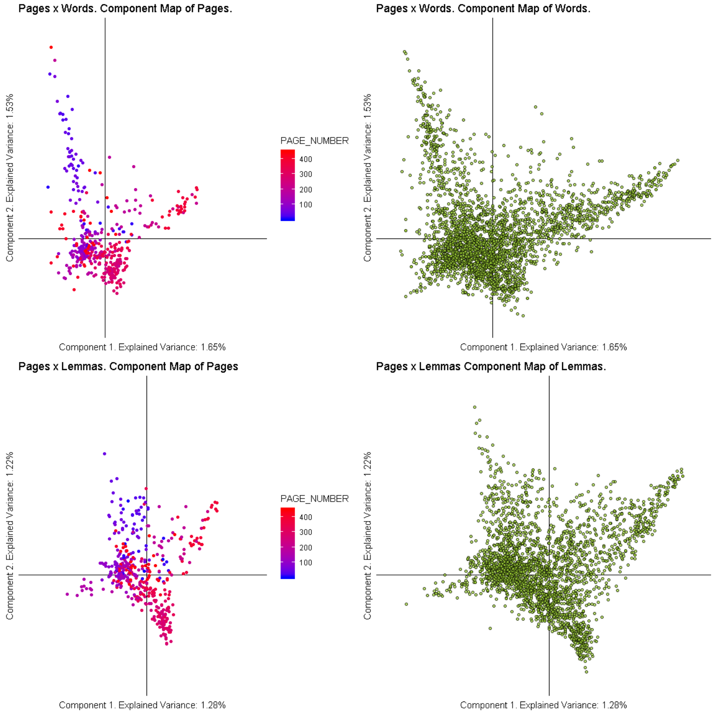
A workshop

- https://github.com/derekbeaton/Workshops/tree/master/RTC/PCA_MCA_Resampling

A workshop

- https://github.com/derekbeaton/Workshops/tree/master/RTC/PCA_MCA_Resampling
  - Extraordinary detail on all of this

The Mueller report


Pages x Words. Component Map of Pages.
Component 2. Explained Variance: 1.53%
Component 1. Explained Variance: 1.65%
PAGE_NUMBER
400
300
200
100
Pages x Words. Component Map of Words.
Component 2. Explained Variance: 1.53%
Component 1. Explained Variance: 1.65%
Pages x Lemmas. Component Map of Pages
Component 2. Explained Variance: 1.22%
Component 1. Explained Variance: 1.28%
PAGE_NUMBER
400
300
200
100
Pages x Lemmas Component Map of Lemmas.
Component 2. Explained Variance: 1.22%
Component 1. Explained Variance: 1.28%

The Marvel Cinematic Universe

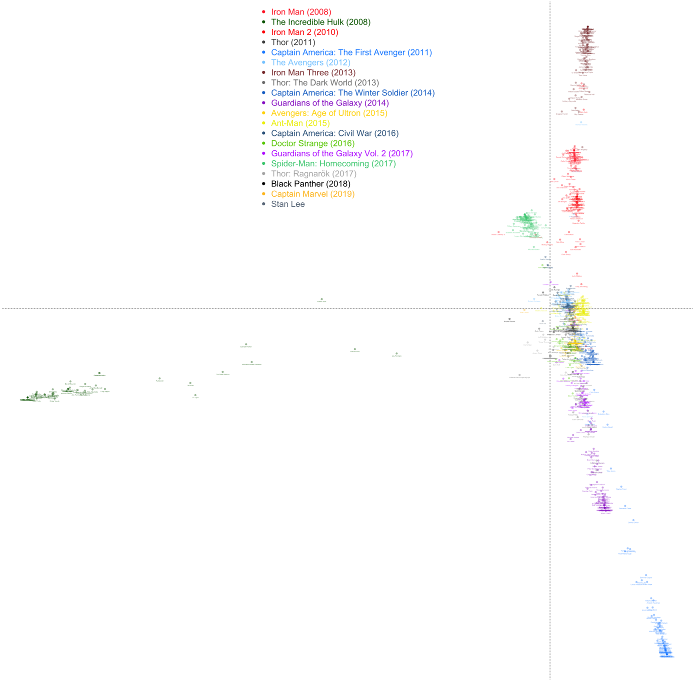
- Actually super cool

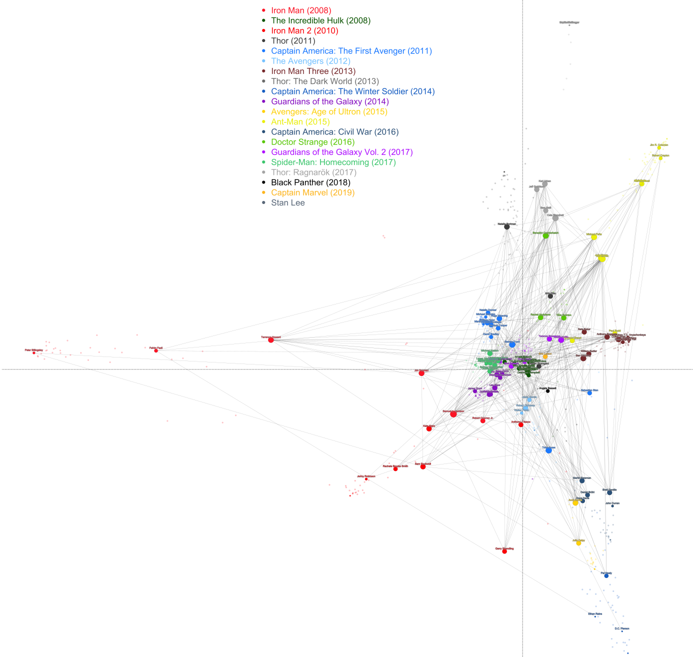
The Marvel Cinematic Universe

- Actually super cool
- CA naturally handles networks

The Marvel Cinematic Universe

- Actually super cool
- CA naturally handles networks
- https://github.com/derekbeaton/Marvel-Cinematic-Universe_Network/


Iron Man (2008)
The Incredible Hulk (2008)
Iron Man 2 (2010)
Thor (2011)
Captain America: The First Avenger (2011)
The Avengers (2012)
Iron Man Three (2013)
Thor: The Dark World (2013)
Captain America: The Winter Soldier (2014)
Guardians of the Galaxy (2014)
Avengers: Age of Ultron (2015)
Ant-Man (2015)
Captain America: Civil War (2016)
Doctor Strange (2016)
Guardians of the Galaxy Vol. 2 (2017)
Spider-Man: Homecoming (2017)
Thor: Ragnarök (2017)
Black Panther (2018)
Captain Marvel (2019)
Stan Lee


Iron Man (2008)
The Incredible Hulk (2008)
Iron Man 2 (2010)
Thor (2011)
Captain America: The First Avenger (2011)
The Avengers (2012)
Iron Man Three (2013)
Thor: The Dark World (2013)
Captain America: The Winter Soldier (2014)
Guardians of the Galaxy (2014)
Avengers: Age of Ultron (2015)
Ant-Man (2015)
Captain America: Civil War (2016)
Doctor Strange (2016)
Guardians of the Galaxy Vol. 2 (2017)
Spider-Man: Homecoming (2017)
Thor: Ragnarök (2017)
Black Panther (2018)
Captain Marvel (2019)
Stan Lee

GMCD & ours

- Generalized minimum covariance determinant

GMCD & ours

- Generalized minimum covariance determinant
- ours (another R package)

GMCD & ours

- Generalized minimum covariance determinant
- ours (another R package)
  - New package for outliers

GMCD & ours

- Generalized minimum covariance determinant
- ours (another R package)
  - New package for outliers
  - Has some important bells-and-whistles

GMCD & ours

- Generalized minimum covariance determinant
- ours (another R package)
  - New package for outliers
  - Has some important bells-and-whistles
  - https://github.com/derekbeaton/ours

GPLS

- First: a PLS for mixed data types

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here
- Second: unify the “two-table” techniques

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here
- Second: unify the “two-table” techniques
  - PLS, CCA, RRR/RDA

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here
- Second: unify the “two-table” techniques
  - PLS, CCA, RRR/RDA
- Package & preprint

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here
- Second: unify the "two-table" techniques
  - PLS, CCA, RRR/RDA
- Package & preprint
  - https://github.com/derekbeaton/gpls

GPLS

- First: a PLS for mixed data types
  - Including those not discussed here
- Second: unify the “two-table” techniques
  - PLS, CCA, RRR/RDA
- Package & preprint
  - https://github.com/derekbeaton/gpls
  - Github issues where I routinely call myself a “dummy”

ExPosition

- ExPosition

ExPosition

- ExPosition
  - Family of packages

ExPosition

- ExPosition
  - Family of packages
  - Includes resampling

ExPosition

- ExPosition
  - Family of packages
  - Includes resampling
  - Lots of PCA & CA techniques

Some alternatives

- FactoMineR

Some alternatives

- FactoMineR
- ade4

Some alternatives

- FactoMineR
- ade4
- ca

Some alternatives

- FactoMineR
- ade4
- ca
- MASS

Some alternatives

- FactoMineR
- ade4
- ca
- MASS
- psych

Some alternatives

- FactoMineR
- ade4
- ca
- MASS
- psych
- factoextra (visualization)

Some alternatives

- FactoMineR
- ade4
- ca
- MASS
- psych
- factoextra (visualization)
- So many others

(Some) References

Expansions & data details

- Beaton, D., Saporta, G., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. bioRxiv, 598888.

Expansions & data details

- Beaton, D., Saporta, G., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. bioRxiv, 598888.
- Beaton, D., Sunderland, K. M., Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., ... & Strother, S. C. (2019). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. bioRxiv, 333005.

And these

- Beaton, D., Fatt, C. R. C., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. Computational Statistics & Data Analysis, 72, 176-189.

And these

- Beaton, D., Fatt, C. R. C., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. Computational Statistics & Data Analysis, 72, 176-189.
- Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological methods, 21(4), 621.

Techniques

- Greenacre, M. (2017). Correspondence analysis in practice. CRC press.

Techniques

- Greenacre, M. (2017). Correspondence analysis in practice. CRC press.
- Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis. Retrieved from http://books.google.com/books?id=LsPaAAAAMAAJ

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley.

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley.
- Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLOS Computational Biology, 15(6), e1006907.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.
- Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse Des Données, 4(2), 137–146.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.
- Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse Des Données, 4(2), 137–146.
- Greenacre, M. (2014). Data Doubling and Fuzzy Coding. In J. Blasius & M. Greenacre (Eds.), Visualization and Verbalization of Data (pp. 239–253). Philadelphia, PA, USA: CRC Press.

History

- Holmes S, Josse J. Discussion of “50 Years of Data Science”. Journal of Computational and Graphical Statistics. 2017, V26(4) 768-769. https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1385471