Multiple Correspondence Analysis (MCA)

Hervé Abdi

www.utdallas.edu~herve herve@utdallas.edu

1. Refresher! and Menu

- What to do with more than 2 nominal variables?
- Answer: *Multiple* Correspondence Analysis.
- Back to punctuation . . .
- Another look at the individual / patterns
- MCA for 2 nominal variables and CA!
- The eigenvalue problem
- Correction formula. . .
- More than 2 variables
- MCA and ordinal data (ranks)

2. MCA . . . Remember the punctuation, 6 authors

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 3 of 55
Go Back
Full Screen
Close
Quit

	.	,	!?...	R	C	H	Z	P	G
mark # 1. Period Rousseau	1	0	0	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7836 marks later	1	0	0	1	0	0	0	0	0
mark # 7837. Comma Rousseau	0	1	0	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
53655 marks later	0	1	0	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
last mark: Other Giraudoux	0	0	1	0	0	0	0	0	1

Table 1: Multiple Correspondence Analysis, Punctuation example. The table of patterns.

	.	,	!?...	R	C	H	Z	P	G
P-R 7836	1	0	0	1	0	0	0	0	0
P-C 53655	1	0	0	0	1	0	0	0	0
P-H 115615	1	0	0	0	0	1	0	0	0
P-Z 161926	1	0	0	0	0	0	1	0	0
P-P 38177	1	0	0	0	0	0	0	1	0
P-G 46371	1	0	0	0	0	0	0	0	1
C-R 13112	0	1	0	1	0	0	0	0	0
C-C 102383	0	1	0	0	1	0	0	0	0
C-H 184541	0	1	0	0	0	1	0	0	0
C-Z 340479	0	1	0	0	0	0	1	0	0
C-P105101	0	1	0	0	0	0	0	1	0
C-G 58367	0	1	0	0	0	0	0	0	1
O-R 6026	0	0	1	1	0	0	0	0	0
O-C 42413	0	0	1	0	1	0	0	0	0
O-H 59226	0	0	1	0	0	1	0	0	0
O-Z 62754	0	0	1	0	0	0	1	0	0
O-P 12670	0	0	1	0	0	0	0	1	0
O-G 14299	0	0	1	0	0	0	0	0	1

Table 2: Multiple Correspondence Analysis, Punctuation example. The table of patterns.

The magic of the distributional equivalence!

	.	,	!?...	R	C	H	Z	P	G
P-R	7836	0	0	7836	0	0	0	0	0
P-C	53655	0	0	0	53655	0	0	0	0
P-H	115615	0	0	0	0	115615	0	0	0
P-Z	161926	0	0	0	0	0	161926	0	0
P-P	38177	0	0	0	0	0	0	38177	0
P-G	46371	0	0	0	0	0	0	0	46371
C-R	0	13112	0	13112	0	0	0	0	0
C-C	0	102383	0	0	102383	0	0	0	0
C-H	0	184541	0	0	0	184541	0	0	0
C-Z	0	340479	0	0	0	0	340479	0	0
C-P	0	105101	0	0	0	0	0	105101	0
C-G	0	58367	0	0	0	0	0	0	58367
O-R	0	0	6026	6026	0	0	0	0	0
O-C	0	0	42413	0	42413	0	0	0	0
O-H	0	0	59226	0	0	59226	0	0	0
O-Z	0	0	62754	0	0	0	62754	0	0
O-P	0	0	12670	0	0	0	0	12670	0
O-G	0	0	14299	0	0	0	0	0	14299

Table 3: Multiple Correspondence Analysis, Punctuation example. The table of patterns.

Table 4: Multiple Correspondence Analysis. Punctuation example. The Projections, Contributions, Squared Distances to Barycenter, and (Squared) Cosines.

	F_1	F_2	Ctr_1	Ctr_2	$m \times d^2$	cos_1	cos_2	Qual
PERIOD	−0.28	−1.10	3	51	.351	2	33	35
COMMA	0.55	0.36	39	17	.218	15	7	22
OTHER	−1.64	0.88	43	13	.431	33	10	43
Rousseau	−1.35	0.73	4	1	.491	3	1	4
Chateaubriand	−1.07	1.05	18	18	.430	14	14	28
Hugo	−0.58	−0.29	11	3	.374	8	2	10
Zola	0.52	0.02	18	0	.302	9	0	9
Proust	1.27	0.62	20	5	.445	15	4	19
Giraudoux	−0.27	−1.93	1	34	.458	1	29	29

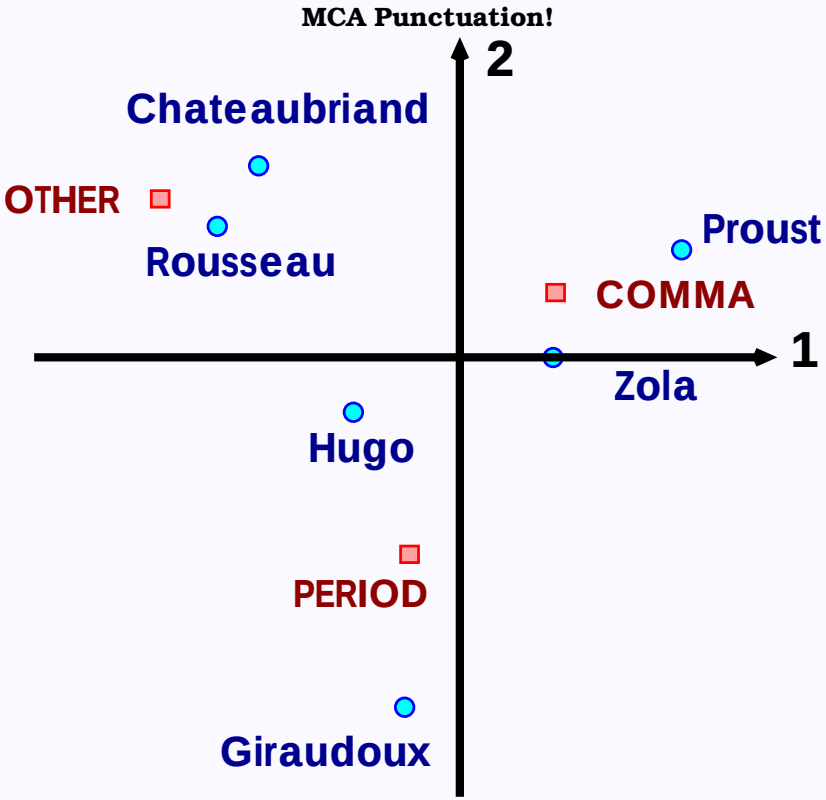


Figure 1: Multiple Correspondence Analysis: Punctuation

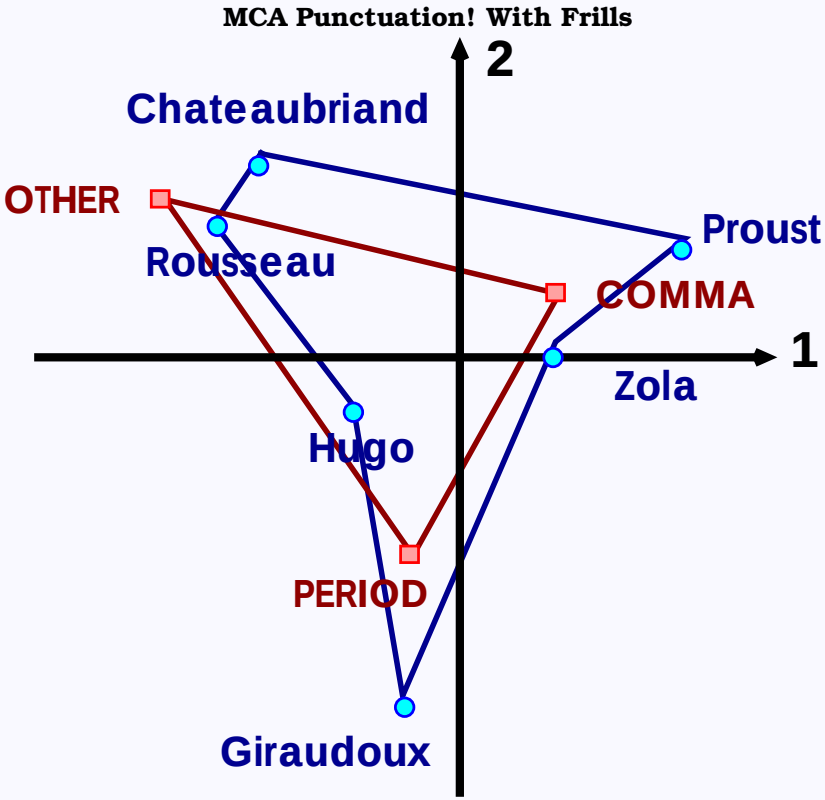


Figure 2: Multiple Correspondence Analysis: Punctuation

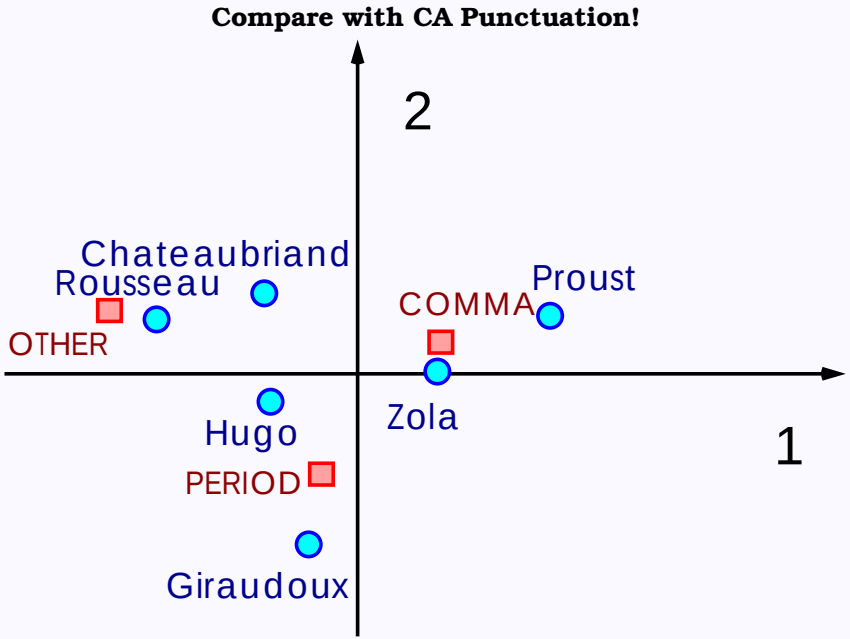


Figure 3: Punctuation: Correspondence Analysis

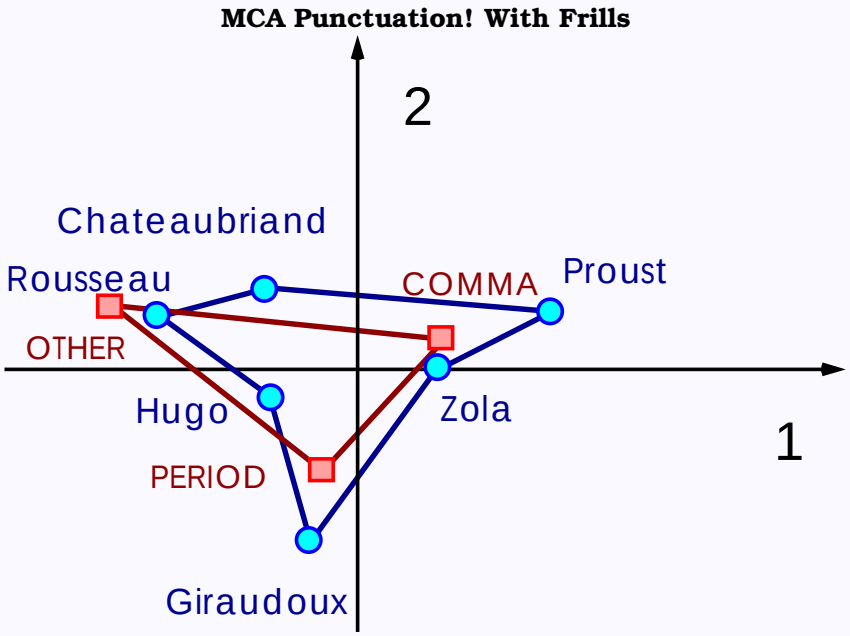


Figure 4: Punctuation Correspondence Analysis

The Eigenvalue Problem: Fictitious Eigenvalues (the hypercube problem) Compare

λ_1	λ_2
0.0178	0.0056

τ_1	τ_2
73	24

With

${}_M\lambda_1$	${}_M\lambda_2$	${}_M\lambda_3$	${}_M\lambda_4$	${}_M\lambda_5$	${}_M\lambda_6$	${}_M\lambda_7$
0.5667	0.5373	0.5000	0.5000	0.5000	0.4627	0.4333

τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
17	15	14	14	14	13	12

The Factors are the same but not the eigenvalues!

in MCA, the ${}_M\lambda \leq 1/K$ code the hypercube!

Correcting for The Eigenvalue Problem: Keep only the ${}_M\lambda > \frac{1}{2}$

$$\lambda = 4 \times \left({}_M\lambda - \frac{1}{2} \right)^2$$

Example

$$\lambda_1 = 4 \times \left({}_M\lambda_1 - \frac{1}{2} \right)^2$$
$$= 4 \times (0.0667)^2 = 4 \times 0045 = .0178$$

And Vice-Versa

$$_M\lambda = \frac{1}{2} \times (\lambda^{\frac{1}{2}} + 1)$$

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 14 of 55
Go Back
Full Screen
Close
Quit

The "Burt" Table

Table 5: Multiple Correspondence Analysis, Punctuation example. The Burt Table.

	.	,	!?...	R	C	H	Z	P	G
.	423580	0	0	7836	53655	115615	161926	38177	46371
,	0	803983	0	13112	102383	184541	340479	105101	58367
!?...	0	0	197388	6026	42413	59226	62754	12670	14299
R	7836	13112	6026	26974	0	0	0	0	0
C	53655	102383	42413	0	198451	0	0	0	0
H	115615	184541	59226	0	0	359382	0	0	0
Z	161926	340479	62754	0	0	0	565159	0	0
P	38177	105101	12670	0	0	0	0	155948	0
G	46371	58367	14299	0	0	0	0	0	119037

The "Burt" Table

Table 6: Multiple Correspondence Analysis, Punctuation example. The Burt Table.

	.	,	!?...	R	C	H	Z	P	G
.	423580	0	0	7836	53655	115615	161926	38177	46371
,	0	803983	0	13112	102383	184541	340479	105101	58367
!?...	0	0	197388	6026	42413	59226	62754	12670	14299
R	7836	13112	6026	26974	0	0	0	0	0
C	53655	102383	42413	0	198451	0	0	0	0
H	115615	184541	59226	0	0	359382	0	0	0
Z	161926	340479	62754	0	0	0	565159	0	0
P	38177	105101	12670	0	0	0	0	155948	0
G	46371	58367	14299	0	0	0	0	0	119037

Note: the inertia of the Burt Table is the average inertia (From Greenacre 2007)

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 17 of 55
Go Back
Full Screen
Close
Quit

Table 7: Multiple Correspondence Analysis. Burt Table Analysis. Punctuation example. The Projections, Contributions, Squared distance to barycenter, and Cosines

	F_1	F_2	ctr_1	ctr_2	$m \times d^2$	$\cos_1^2$	$\cos_2^2$	Qual
.	1.02	0.53	4	1	3.8562	3	1	4
,	0.80	0.77	21	19	0.8785	14	14	28
. . .	0.44	−0.21	13	3	0.1031	8	2	10
R	−0.39	0.01	20	0	0.0072	9	0	9
C	−0.95	0.45	22	5	0.2852	15	4	19
H	0.20	−1.41	1	36	0.6943	1	29	29
Z	0.21	−0.80	4	54	0.2258	2	33	35
P	−0.41	0.26	44	18	0.0224	15	7	22
G	1.24	0.65	48	13	0.1316	33	10	43

Compare

λ_1	λ_2
0.0178	00056

τ_1	τ_2
73	24

With

${}_M\lambda_1$	${}_M\lambda_2$	${}_M\lambda_3$	${}_M\lambda_4$	${}_M\lambda_5$	${}_M\lambda_6$	${}_M\lambda_7$
0.5667	0.5373	0.5000	0.5000	0.5000	0.4627	0.4333

τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
17	15	14	14	14	13	12

With

${}_B\lambda_1$	${}_B\lambda_2$	${}_B\lambda_3$	${}_B\lambda_4$	${}_B\lambda_5$	${}_B\lambda_6$	${}_B\lambda_7$
0.3212	0.2887	0.2500	0.2500	0.2500	0.2140	0.1877

τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
18	16	14	14	14	12	10

$$ {}_B\lambda = {}_M\lambda^2 \iff {}_M\lambda = {}_B\lambda^{\frac{1}{2}} $$

	CA		MCA		Burt		Normalized	
	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
.	−0.049	0.112	−0.276	−1.095	−0.208	0.802	−0.367	1.493
,	0.097	−0.037	0.549	0.360	0.413	−0.264	0.729	−0.491
'	−0.291	−0.090	−1.643	0.884	−1.237	−0.648	−2.183	−1.206
R	−0.240	−0.074	−1.352	0.727	−1.018	−0.533	−1.796	−0.992
C	−0.190	−0.107	−1.069	1.051	−0.805	−0.771	−1.420	−1.434
H	−0.103	0.030	−0.583	−0.292	−0.439	0.214	−0.774	0.398
Z	0.092	−0.002	0.518	0.016	0.390	−0.012	0.688	−0.022
P	0.224	−0.063	1.265	0.619	0.952	−0.454	1.680	−0.845
G	−0.048	0.196	−0.268	−1.926	−0.202	1.412	−0.356	2.628

Table 8: Comparison between three ways of analyzing the same contingency table. Correspondence analysis Multiple correspondence analysis, and Burt Table Analysis. Punctuation example. The last two rows are normalized to 1. They can be obtained by normalization of any of the analyses. The normalization is obtained by dividing each entry of a column by the square root of its eigenvalue.

A Small Example

Effect of Oak Species on Pinot Noirs

- Aged 3 Pinot with Oak Type I and 6 with Oak Type II.
- Three Assessors choose their own variables
- Answers coded as binary or ternary (0/1)

		Expert 1							Expert 2								Expert 3					
Wine	Oak Type	fruity		woody			coffee		red fruit		roasted		vanillin			woody		fruity		butter		woody	
W1	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1	0	1	0	1	0	1
W2	2	0	1	0	1	0	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	1	0
W3	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	0	1	1	0	1	0
W4	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0
W5	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1	1	0	0	1	0	1
W6	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0	1	0	1
W?	?	0	1	0	1	0	.5	.5	1	0	1	0	0	1	0	.5	.5	1	0	.5	.5	0	1

Figure 5: Wines. The Data.

2 corrections for the eigenvalues for MCA

First: Benzécri: Keep eigenvalue ¿ $\frac{1}{K}$

λ_ℓ: eigenvalues from the analysis indicator matrix.
${}_{\mathrm{c}}\lambda_\ell$ corrected eigenvalues

$$ {}_{\mathrm{c}}\lambda_\ell = \begin{cases} \left[\left(\frac{K}{K-1}\right)\left(\lambda_\ell - \frac{1}{K}\right)\right]^2 & \text{if } \lambda_\ell > \frac{1}{K} \\ 0 & \text{if } \lambda_\ell \leq \frac{1}{K} \end{cases} . \qquad (1) $$

Over-correct (but not much)!

2 corrections for the eigenvalues for MCA

Greenacre: Keep Benzécri's correction: Correct the Inertia

λ_ℓ: eigenvalues from indicator matrix
${}_{\mathrm{c}}\lambda_\ell$ corrected eigenvalues
Use average inertia, denoted $\overline{\mathcal{I}}$:

$$\overline{\mathcal{I}} = \frac{K}{K-1} \times \left(\sum_\ell \lambda_\ell^2 - \frac{J-K}{K^2} \right) \tag{2}$$

Percentage inertia:

$$\tau_{\mathrm{c}} = \frac{{}_{\mathrm{c}}\lambda}{\overline{\mathcal{I}}} \text{ instead of } \frac{{}_{\mathrm{c}}\lambda}{\sum {}_{\mathrm{c}}\lambda_\ell} \ . \tag{3}$$

Table 9: Eigenvalues, corrected eigenvalues, proportion of explained inertia and corrected proportion of explained inertia. The eigenvalues of the Burt matrix are equal to the squared eigenvalues of the indicator matrix; The corrected eigenvalues for Benzécri and Greenacre are the same, but the proportion of explained variance differ. Eigenvalues are denoted by λ, proportions of explained inertia by τ (note that the average inertia used to compute Greenacre's correction is equal to $\overline{\mathcal{I}} = .7358$).

	Indicator Matrix		Burt Matrix		Benzécri Correction		Greenacre Correction	
Factor	${}_{\mathrm{I}}\lambda$	τ_{I}	${}_{\mathrm{B}}\lambda$	τ_{B}	${}_{\mathrm{Z}}\lambda$	τ_{Z}	${}_{\mathrm{c}}\lambda$	τ_{c}
1	.8532	.7110	.7280	.9306	.7004	.9823	.7004	.9519
2	.2000	.1667	.0400	.0511	.0123	.0173	.0123	.0168
3	.1151	.0959	.0133	.0169	.0003	.0004	.0003	.0004
4	.0317	.0264	.0010	.0013	0	0	0	0
$\sum$	1.2000	1	.7822	1	.7130	1	.7130	.9691

Table 10: Factor scores, squared cosines, and contributions for the observations (I-set). The eigenvalues and proportions of explained inertia are corrected using Benzécri/Greenacre formula. Contributions corresponding to negative scores are in italic. The mystery wine (Wine ?) is a supplementary observation. Only the first two factors are reported.

		Wine 1	Wine 2	Wine 3	Wine 4	Wine 5	Wine 6	Wine ?
F	$_{c}\lambda$ $\%_{c}$			*Factor Scores*				
1	.7004 95	0.86	−0.71	−0.92	−0.86	0.92	0.71	0.03
2	.0123 2	0.08	−0.16	0.08	0.08	0.08	−0.16	−0.16
F				*Squared Cosines*				
1		.62	.42	.71	.62	.71	.42	.04
2		.01	.02	.01	.01	.01	.02	.96
F				*Contributions* ×1000				
1		177	*121*	*202*	*177*	202	121	–
2		83	*333*	83	83	83	*333*	–

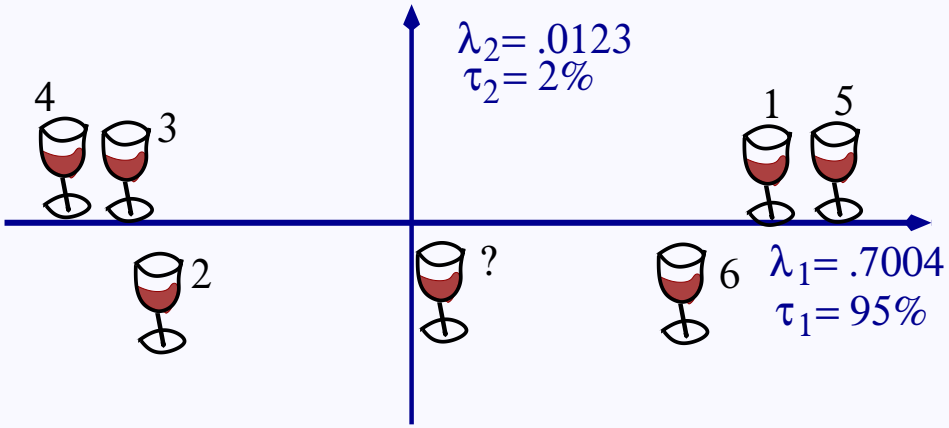


Figure 6: Wines. Map of the wines.

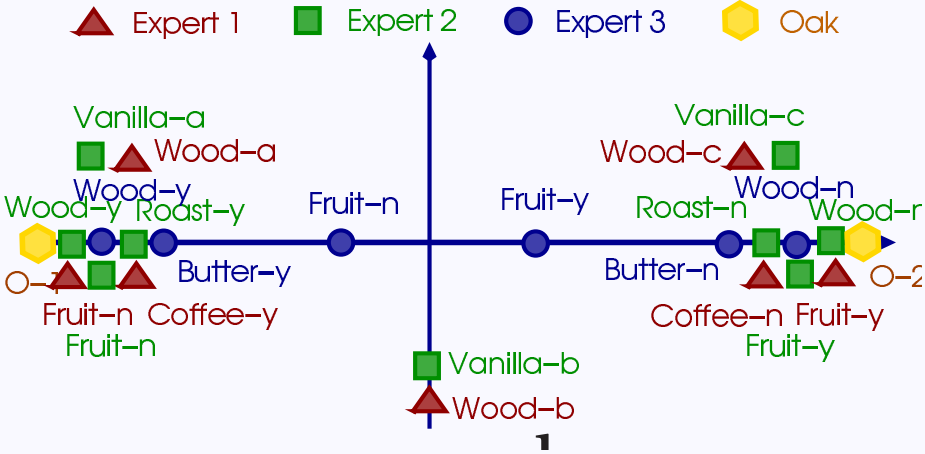


Figure 7: Wines. Map of the Variables (Assessors × Scales).

TAKS in TEXAS

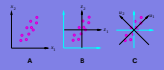
A Biggggggggggg Analysis:

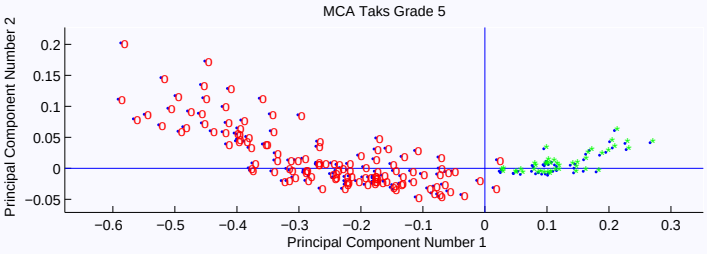
Effect of Correction on Big Data

Exemple: Math Test:

5th Grade (all of DISD).

(44 Questions $\times$ 4 answers each = 172 Columns).




MCA Taks Grade 5
Principal Component Number 2
0.2
0.15
0.1
0.05
0
−0.05
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
Principal Component Number 1

Correcting for The Eigenvalue Problem: With K Variables: Keep only the ${}_M\lambda > \frac{1}{K}$

$$\lambda = \left(\frac{K}{1-K}\right)^2 \times \left({}_M\lambda - \frac{1}{K}\right)^2$$

Example: TAKS

${}_M\lambda_1 = .1884$

$$\lambda_1 = \left(\frac{44}{43}\right)^2 \times \left(.1884 - \frac{1}{44}\right)^2$$
$$= .0287$$

(Goes from 7% to 96%!)

Ordinal Data

A Beer example

- Five judges
- Each judge rates five beer
- On a 6-point rating scale (0 to 5)

	Cain	Sims	Guy	Kinchen	Schalbs
Miller	3	2	2	2	2
RollRock	3	3	3	2.5	3
Lowenbrau	3.5	4	4	2.5	3
Coors	2.5	4	2.5	2.5	2.5
Budweiser	3.5	2	3	2	3

Table 11: Correspondence Analysis. Scoring Beers. The Original data.

Thermometer Coding: Use both sides of the scale

First Idea: Use the ratings

Not a very good idea!
Use Thermometer Coding.

Thermometer Coding: The Leverage Principle

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 34 of 55
Go Back
Full Screen
Close
Quit

	Cain		Sims		Guy		Kinchen		Schalbs	
	+	−	+	−	+	−	+	−	+	−
Miller	3	2	2	3	2	3	2	3	2	3
RollRock	3	2	3	2	3	2	2.5	2.5	3	2
Lowenbrau	3.5	1.5	4	1	4	1	2.5	2.5	3	2
Coors	2.5	2.5	4	1	2.5	2.5	2.5	2.5	2.5	2.5
Budweiser	3.5	1.5	2	3	3	2	2	3	3	2

Table 12: Correspondence Analysis. Scoring Beers. The data with thermometer coding.

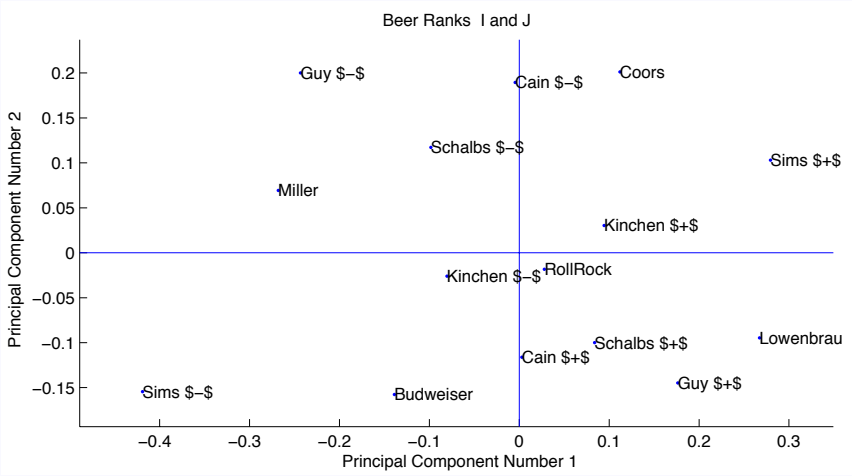


Figure 8: Ranking Beers: Correspondence Analysis . The Ugly plot: no frill at all! (Don't show that except to your close friends)

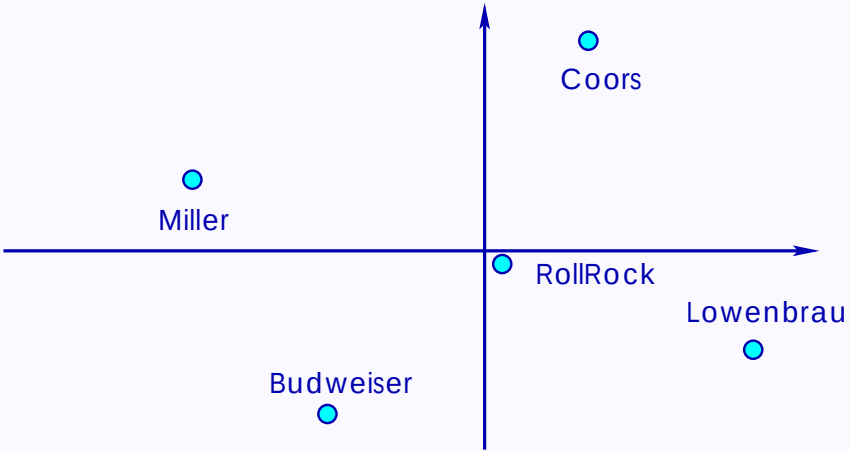


Figure 9: Ranking Beers: Correspondence Analysis (no frill, but prettier). The Beer Plot

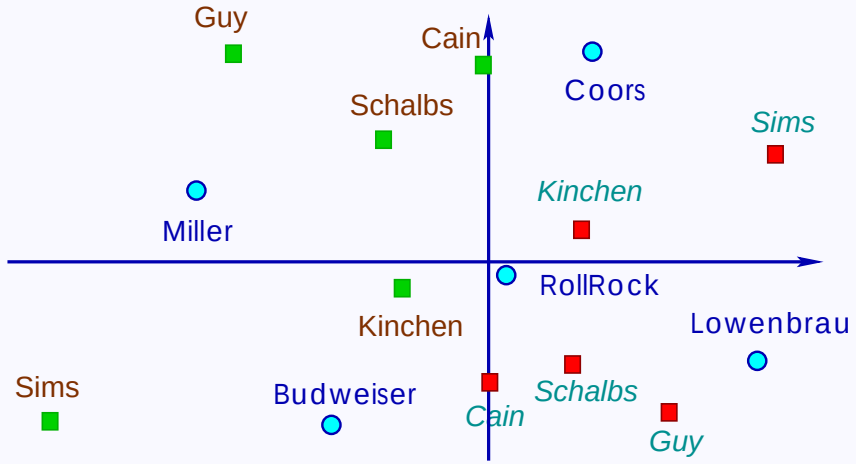


Figure 10: Ranking Beers: Correspondence Analysis. Beers and Raters (+ and -)

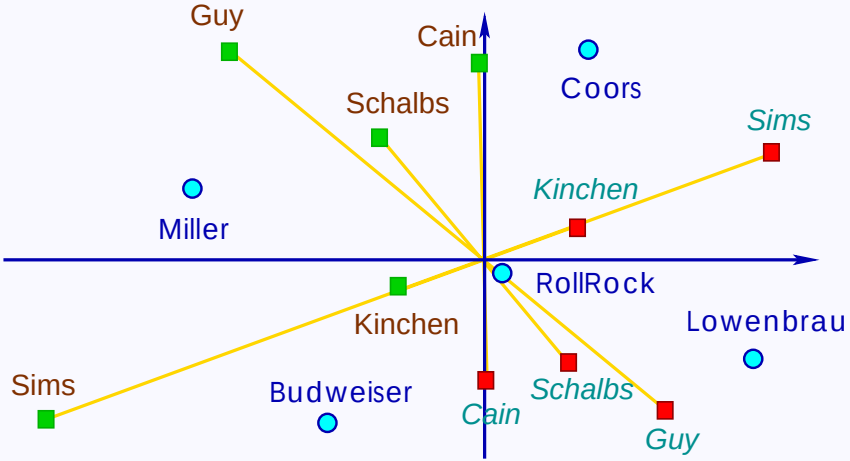


Figure 11: Ranking Beers: Correspondence Analysis. Beers and Raters (+ and -). Raters are lines

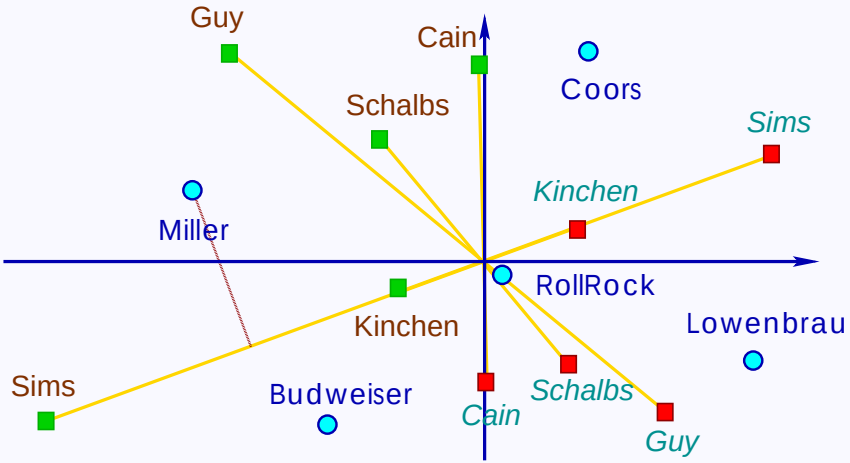


Figure 12: Ranking Beers: Correspondence Analysis. Projecting a Beer on a Rater.

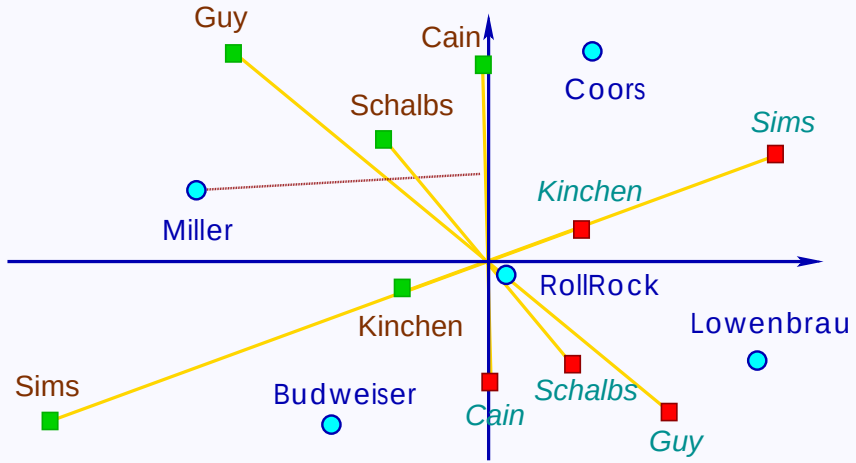


Figure 13: Ranking Beers: Correspondence Analysis. Projecting a Beer on another Rater.

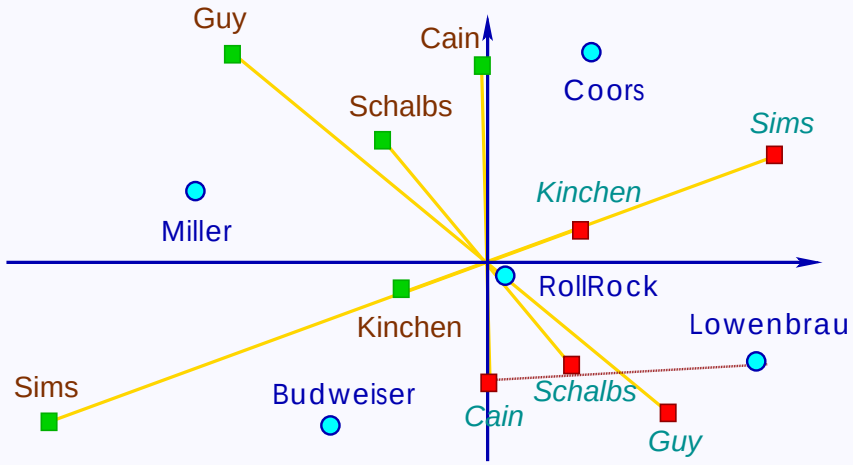


Figure 14: Ranking Beers: Correspondence Analysis. More projections.

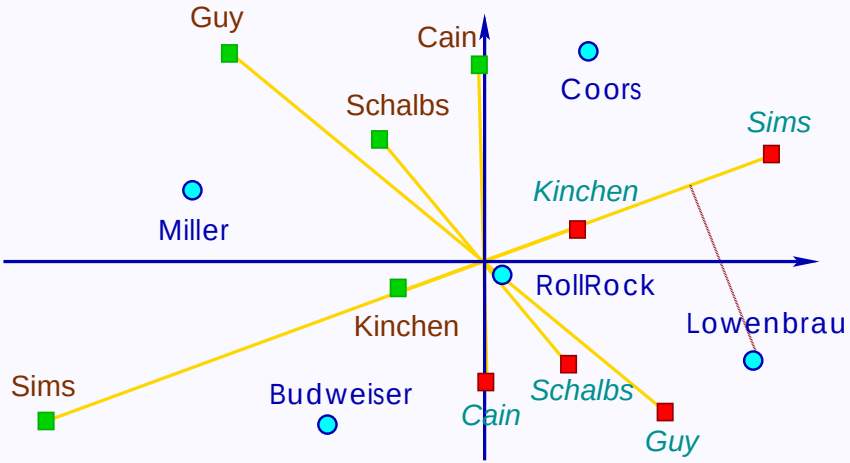


Figure 15: Ranking Beers: Correspondence Analysis. Even more projections

	F_1	F_2	ctr_1	ctr_2	$m \times d^2$	$\cos_1^2$	$\cos_2^2$	Qual
Cain +	0	−0.12	0	92	0	0	11	11
Cain −	−0	0.19	0	92	0	0	17	17
Sims +	0.28	0.10	88	12	0.01	27	8	35
Sims −	−0.42	−0.15	88	12	0.02	40	12	52
Guy +	0.18	−0.14	59	40	0.01	10	15	26
Guy −	−0.24	0.20	59	40	0.01	14	21	35
Kinchen +	0.09	0.03	79	8	0	2	1	3
Kinchen −	−0.08	−0.03	79	8	0	2	0	2
Schalbs +	0.08	−0.10	32	46	0	2	7	9
Schalbs −	−0.10	0.12	32	46	0	3	8	11

Table 13: Correspondence Analysis. Beers. J-set (Raters). The Projections, Contributions, Squared distance to barycenter, and (Squared) Cosines

	F_1	F_2	ctr_1	ctr_2	$m \times d^2$	$\cos_1^2$	$\cos_2^2$	Qual
Miller	−0.27	0.07	91	6	0.02	41	6	47
RollRock	0.03	−0.02	16	7	0	0	0	1
Lowenbrau	0.27	−0.09	86	11	0.02	41	11	52
Coors	0.11	0.20	23	76	0.01	7	51	58
Budweiser	−0.14	−0.16	43	55	0.01	11	31	42

Table 14: Correspondence Analysis. Beers: The I-set (Beers). The Projections, Contributions, Squared distance to barycenter, and Cosines

3. A Bigger Example: The Mask Fit Study

- 80 T/F statement, 5 point rating scale
- 102 participants

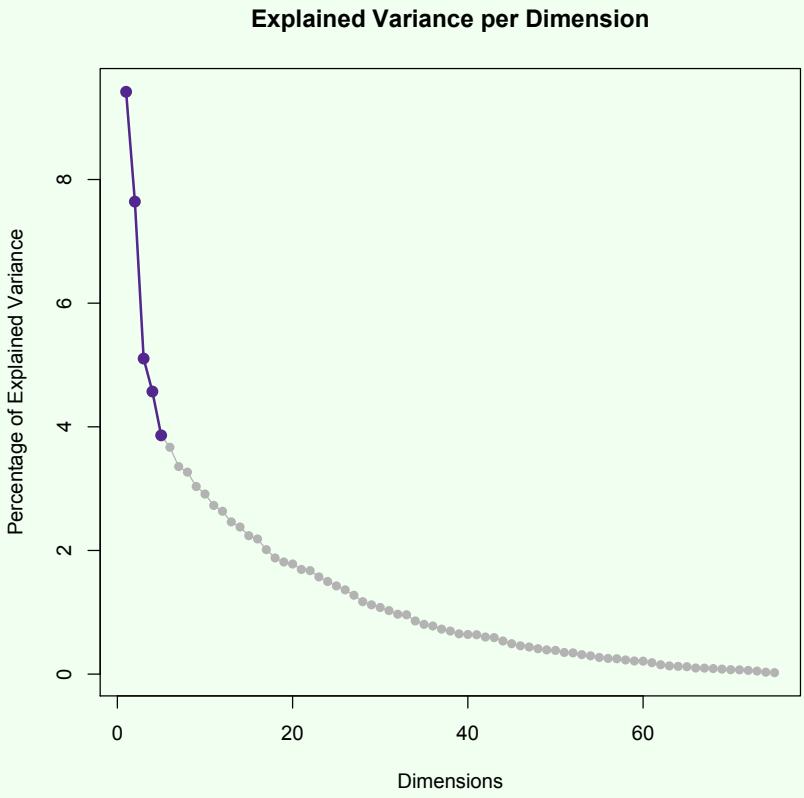


Figure 16: Mask Fit Study. The scree.

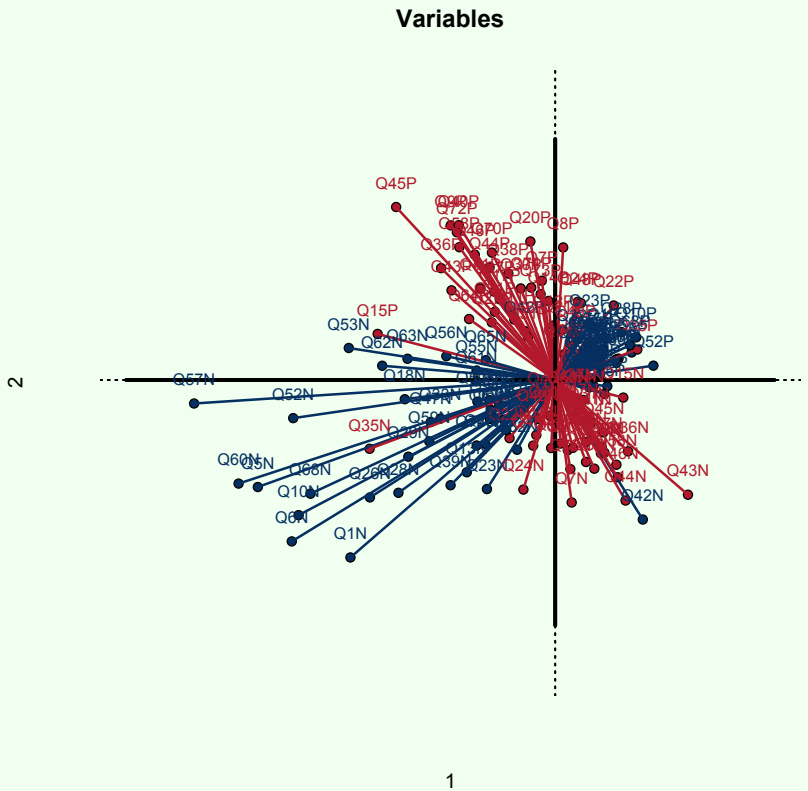


Figure 17: Mask Fit Study. The Variables. Colored. Omission vs Commission

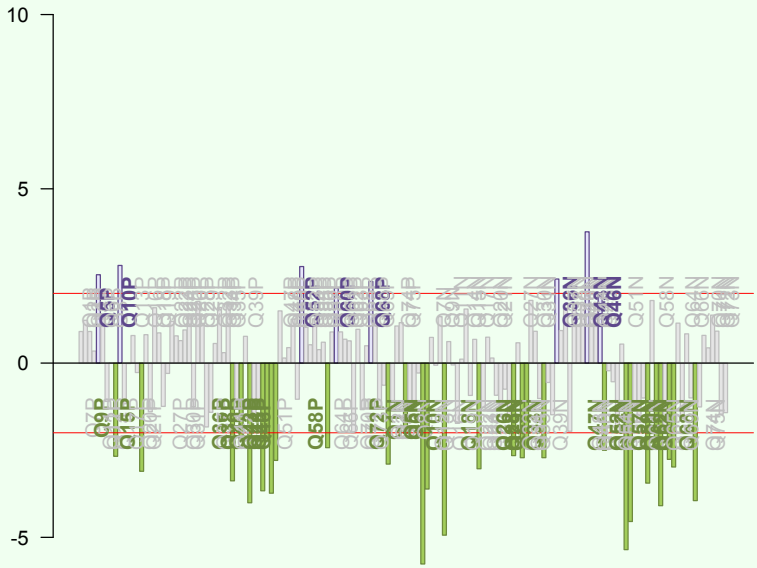


Figure 18: Mask Fit Study. The Variables. Bootstrap ratios. Dimension 1

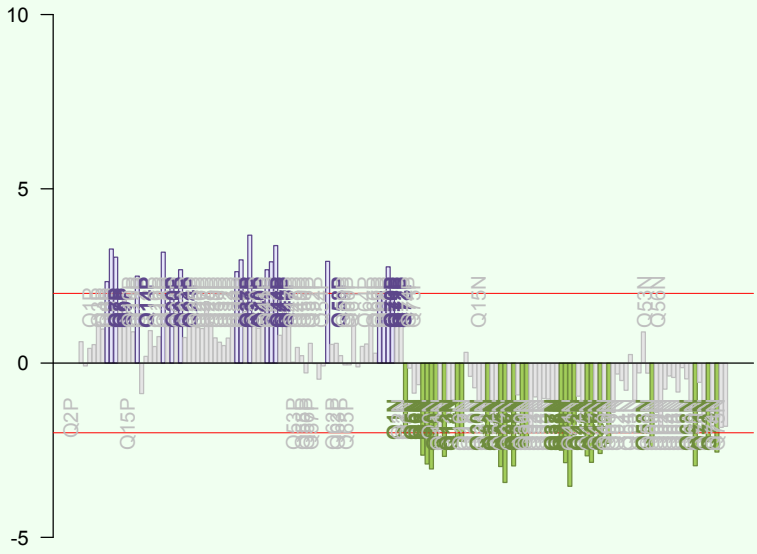


Figure 19: Mask Fit Study. The Variables. Bootstrap ratios. Dimension 2

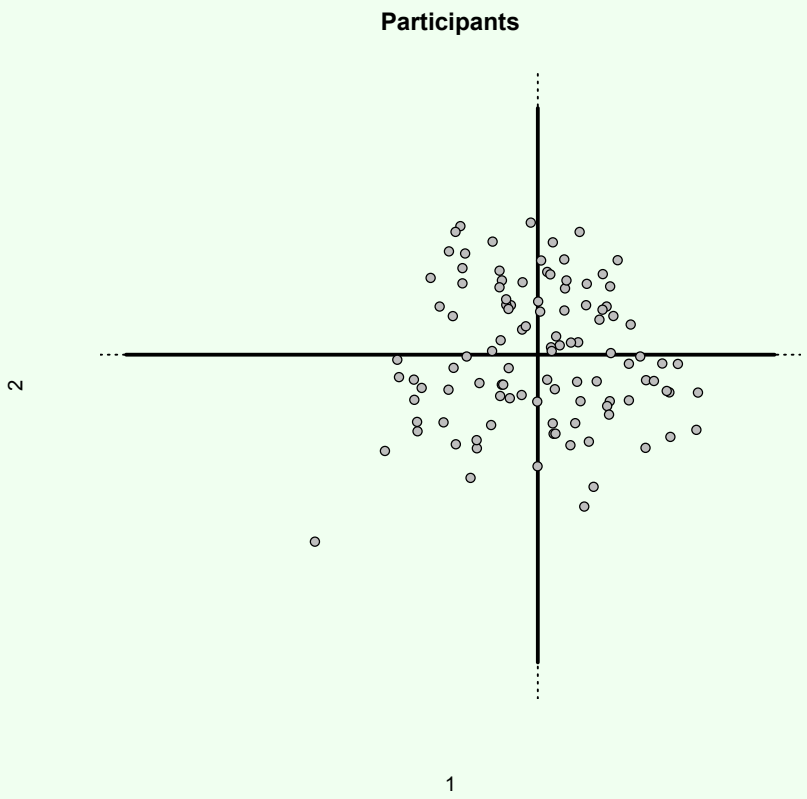


Figure 20: Mask Fit Study. The participants.

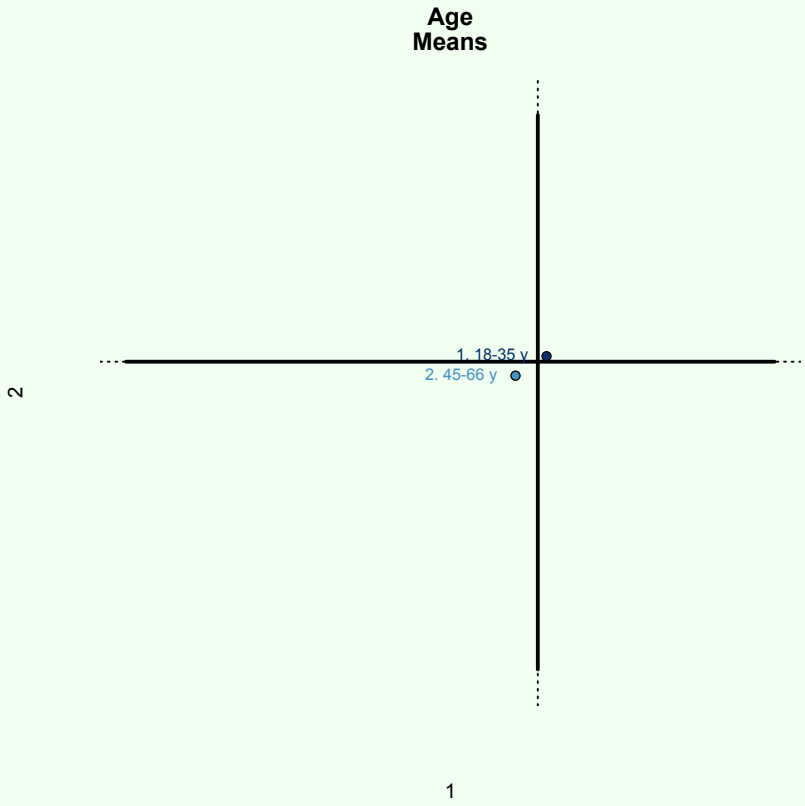


Figure 21: Mask Fit Study. Participants. Age Barycenters.

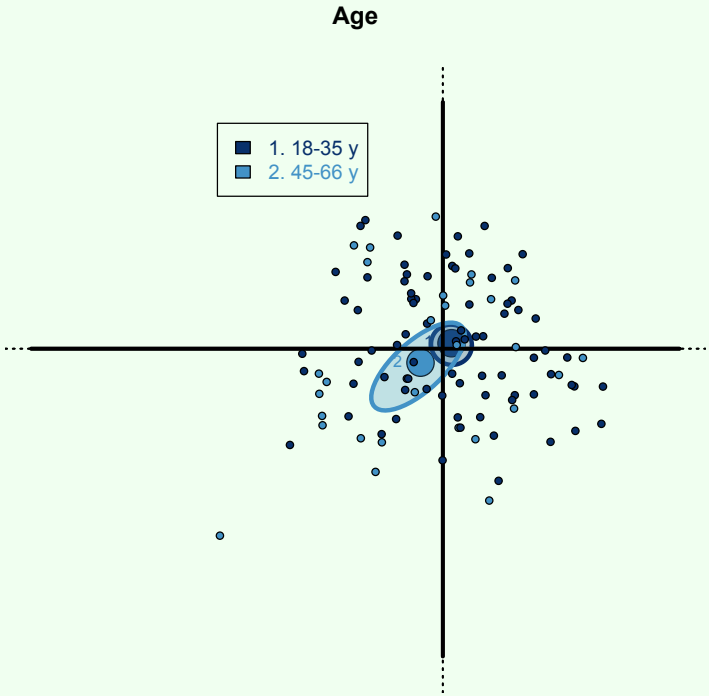


Figure 22: Mask Fit Study. Participants. Age Barycenters with observation and 95% CIs

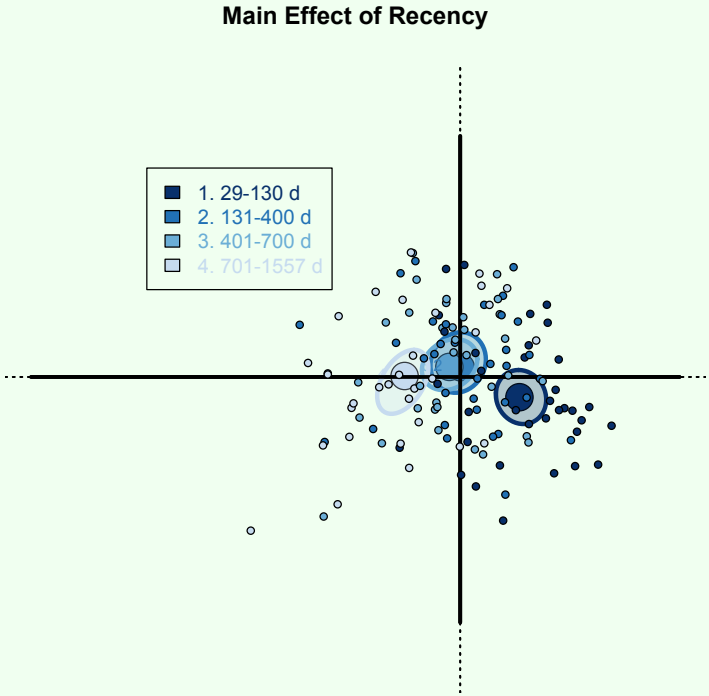


Figure 23: Mask Fit Study. Participants. Age Barycenters.

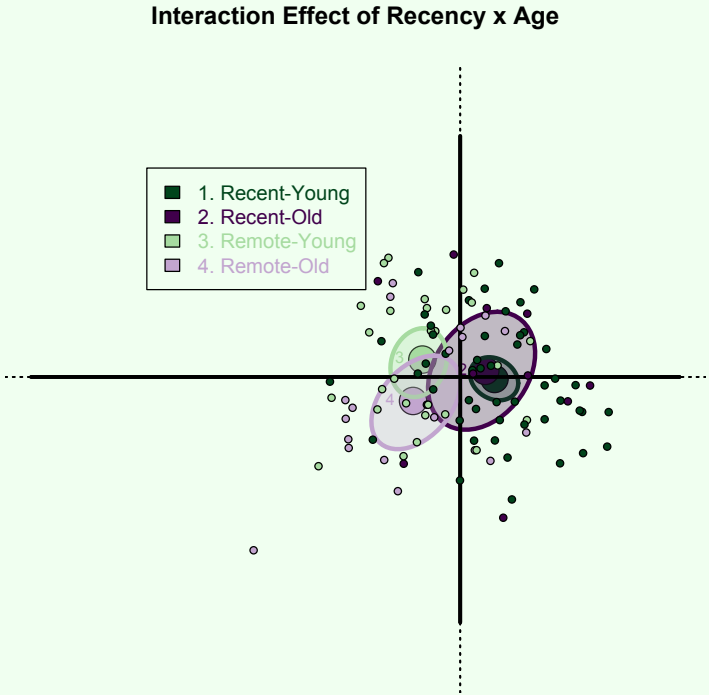


Figure 24: Mask Fit Study. Participants. Age Barycenters.