Correspondence Analysis & Related Methods

Michael Greenacre

SESSION 17: **Multiple correspondence analysis**

ISSP 1993: Environment

Q.4 SCIENCE AND ENVIRONMENT

How much do you agree or disagree with each of these statements?

Q.4a We believe too often in science, and not enough in feelings and faith.
Q.4b Over all, modern science does more harm than good.
Q.4c Any change humans cause in nature - no matter how scientific - is likely to make things worse.
Q.4d Modern science will solve our environmental problems with little change to our way of life.

Response categories

1. Strongly agree
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Strongly disagree
8. Can't choose, don't know
9. NA, refused

We are interested now in the relationship between the four variables, not so much as the differences between countries. Since the relationship between the four variables might change across the countries, we shall restrict our attention for the moment to one country, say Germany (data given on website). Also we have taken out the missing values in this initial example to make the problem simpler. The sample size is n = 916.

Original definition of MCA

Original responses (Q questions)

4a	4b	4c	4d
2	2	1	2
2	2	2	5
4	3	2	5
2	5	4	2
4	2	1	5
1	4	1	5
1	2	2	3
1	3	2	4
3	2	2	4
3	5	5	2
.	.	.	.
.	.	.	.
etc.	.	.	

Indicator matrix

A					B					C					D				
1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
etc.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

(916 rows)

MCA is the application of the CA algorithm to the indicator matrix. So:

- **Each profile is a series of zeros with a value of 1/Q indicating the response**
- **Each respondent has the same mass 1/n.**
- **Respondents (profiles) are at ordinary averages of their responses (vertices).**

"Covariances"

SPSS output V4a x V4b

Tabla de contingencia Science: believe too often in * Science: more harm than good

Recuento

		Science: more harm than good					
		Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Total
Science: believe too often in	Strongly Agree	33	51	41	53	10	188
	Agree	15	113	96	133	32	389
	Neither Agree nor Disagree	7	31	27	67	10	142
	Disagree	5	30	17	83	20	155
	Strongly Disagree	1	8	4	5	24	42
Total		61	233	185	341	96	916

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	185.192[a]	16	.000
Razón de verosimilitud	136.199	16	.000
Asociación lineal por lineal	62.837	1	.000
N de casos válidos	916		

a. 2 casillas (8.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 2.80.

V4b

V4a

33	51	41	53	10
15	113	96	133	32
7	31	27	67	10
5	30	17	83	20
1	8	4	5	24

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	.450	.000
	V de Cramer	.225	.000
	Coeficiente de contingencia	.410	.000
N de casos válidos		916	

Assembling the Burt matrix

	V4a					V4b					V4c					V4d				
V4a	188	0	0	0	0	33	51	41	53	10	54	73	29	26	6	10	42	33	64	39
	0	389	0	0	0	15	113	96	133	32	53	156	80	81	19	7	105	103	137	37
	0	0	142	0	0	7	31	27	67	10	17	51	37	34	3	5	36	41	41	19
	0	0	0	155	0	5	30	17	83	20	16	40	30	60	9	1	45	42	49	18
	0	0	0	0	42	1	8	4	5	24	2	12	11	9	8	6	6	12	6	12
V4b	33	15	7	5	1	61	0	0	0	0	37	17	4	1	2	4	4	12	21	20
	51	113	31	30	8	0	233	0	0	0	45	119	42	23	4	6	53	57	83	34
	41	96	27	17	4	0	0	185	0	0	17	76	58	31	3	4	50	59	51	21
	53	133	67	83	5	0	0	0	341	0	36	98	68	130	9	5	97	81	126	32
	10	32	10	20	24	0	0	0	0	96	7	22	15	25	27	10	30	22	16	18
V4c	54	53	17	16	2	37	45	17	36	7	142	0	0	0	0	8	25	24	49	36
	73	156	51	40	12	17	119	76	98	22	0	332	0	0	0	9	76	92	109	46
	29	80	37	30	11	4	42	58	68	15	0	0	187	0	0	4	60	58	48	17
	26	81	34	60	9	1	23	31	130	25	0	0	0	210	0	1	58	54	85	12
	6	19	3	9	8	2	4	3	9	27	0	0	0	0	45	7	15	3	6	14
V4d	10	7	5	1	6	4	6	4	5	10	8	9	4	1	7	29	0	0	0	0
	42	105	36	45	6	4	53	50	97	30	25	76	60	58	15	0	234	0	0	0
	33	103	41	42	12	12	57	59	81	22	24	92	58	54	3	0	0	231	0	0
	64	137	41	49	6	21	83	51	126	16	49	109	48	85	6	0	0	0	297	0
	39	37	19	18	12	20	34	21	32	18	36	46	17	12	14	0	0	0	0	125

MCA can be defined as the CA of the Burt matrix **B** , the "covariance matrix" between the 4 variables

Equivalent definition of MCA

MCA is (almost equivalently) the application of the CA algorithm to the Burt. This analysis gives exactly the same standard coordinates of the category points (columns), but the principal inertias are the squares of the indicator ones. This is because there is a simple relationship between the indicator matrix **Z** and the Burt matrix **B** : $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$.

Class exercize: show that the CA of **Z** and the CA of **B** give same column standard coordinates, and principal inertias (eigenvalues) of **B** are the squares of those of **Z**. Which analysis gives higher % of inertia?

N.B. XLSTAT gives results for **Z**

mjca in the **ca** package gives you several choices:

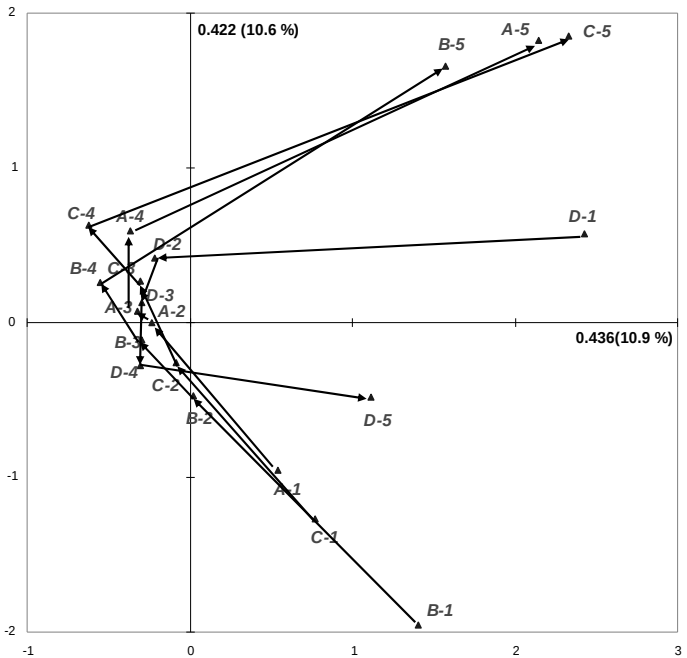
lambda="indicator"

lambda="Burt"

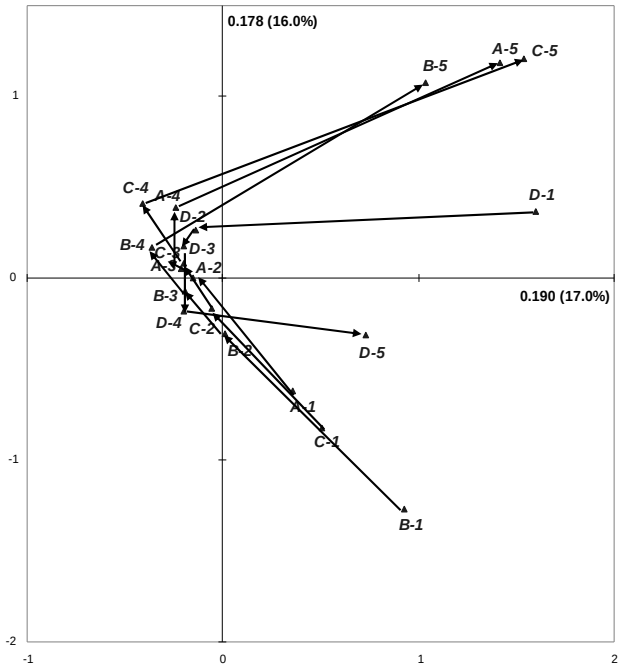
lambda="adjusted"

lambda="JCA"

Analysis of indicator matrix



Analysis of Burt matrix



Various measures of association

Chi-square

	V4b	V4c	V4d
V4a	185.2	86.8	59.4
V4b		326.9	73.8
V4c			101.5

χ^2

Phi coefficient

	V4b	V4c	V4d
V4a	.450	.308	.255
V4b		.597	.284
V4c			.333

$\sqrt{\chi^2/n}$

Cramer's V

	V4b	V4c	V4d
V4a	.225	.154	.127
V4b		.299	.142
V4c			.166

$\sqrt{(\chi^2/n)\,/\min(I\text{-}1, J\text{-}1)}$

Inertia

	V4b	V4c	V4d
V4a	.203	.095	.065
V4b		.356	.081
V4c			.111

χ^2/n

Burt matrix - table inertias

	V4a	V4b	V4c	V4d
V4a	188 0 0 0 0 0 389 0 0 0 0 0 1424 0 0 0 0 0 155 0 0 0 0 0 42	.203	.095	.065
V4b	33 15 7 5 1 51 113 31 30 8 41 96 27 17 4 53 133 67 83 5 10 32 10 20 24	61 0 0 0 0 0 233 0 0 0 0 0 1854 0 0 0 0 0 341 0 0 0 0 0 96	.356	.081
V4c	54 53 17 16 2 73 156 51 40 12 29 80 37 30 11 26 81 34 60 9 6 19 3 9 8	37 45 17 36 7 17 119 76 98 22 4 42 58 68 15 1 23 31 130 25 2 4 3 9 27	142 0 0 0 0 0 332 0 0 0 0 0 1874 0 0 0 0 0 210 0 0 0 0 0 45	.111
V4d	10 7 5 1 6 42 105 36 45 6 33 103 41 42 12 64 137 41 49 6 39 37 19 18 12	4 6 4 5 10 4 53 50 97 30 12 57 59 81 22 21 83 51 126 16 20 34 21 32 18	8 9 4 1 7 25 76 60 58 15 24 92 58 54 3 49 109 48 85 6 36 46 17 12 14	29 0 0 0 0 0 234 0 0 0 0 0 2314 0 0 0 0 0 297 0 0 0 0 0 125

Inertia of Burt matrix is the average of the 16 inertias: 1.114

Inertia of off-diagonal blocks of Burt matrix is the average of the 12 inertias: 0.152

Adjustment of principal inertias (eigenvalues)

We can rescale an existing MCA solution quite simply in order to best fit the off-diagonal tables. All we need is the total inertia of the Burt matrix, *inertia*(**B**), and the principal inertias λ_k^2 of the Burt matrix in the solution space.

If we have computed the solution on the indicator matrix **Z** (as in MCA module of XLSTAT), the eigenvalues calculated are λ_k so *all* the *squares* of the principal inertias of **Z** need to be summed in order to get *inertia*(**B**). If you have analysed the Burt matrix **B**, *inertia*(**B**) is the total inertia.

Here are the steps to rescale the solution:

1. Calculate the average off-diagonal inertia :

   $$\text{average off-diagonal inertia} = \frac{Q}{Q-1}\left(\textit{inertia}\,(\mathbf{B}) - \frac{J-Q}{Q^2}\right)$$

2. Calculate the adjusted principal inertias :

   $$\text{adjusted principal inertias} = \left(\frac{Q}{Q-1}\right)^2\left(\lambda_k - \frac{1}{Q}\right)^2 \quad \text{only for} \lambda_k > \frac{1}{Q}$$

3. Calculate adjusted percentages of inertia :

   $$\text{adjusted percentages of inertia} = \frac{\text{adjusted principal inertias}}{\text{average off-diagonal inertia}}$$

Adjustment of principal inertias - environment survey

average off-diagonal inertia $= (4/3)\,(1.114 - 16/16) = 0.1520$

adjusted principal inertias in first two dimensions

$= (4/3)^2(0.4355 - ¼)^2 = 0.06117$ (dimension 1)

$= (4/3)^2(0.4225 - ¼)^2 = 0.05290$ (dimension 2)

adjusted percentages of inertia

$= 0.06117 / 0.1520 = 0.4024$ i.e. 40.2 %

$= 0.05290 / 0.1520 = 0.3480$ i.e. 34.8 %

1. Calculate the average off-diagonal inertia :

   $$\text{average off-diagonal inertia} = \frac{Q}{Q-1}\left(\textit{inertia}\,(\mathbf{B}) - \frac{J-Q}{Q^2}\right)$$

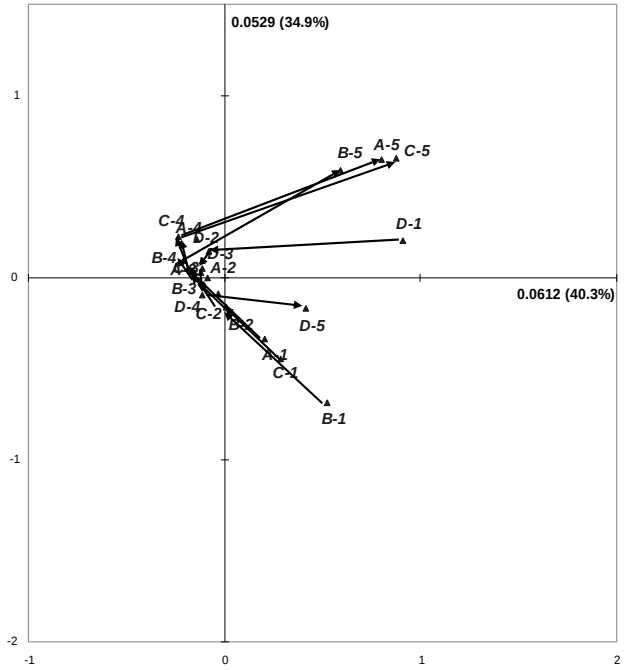
2. Calculate the adjusted principal inertias :

   $$\text{adjusted principal inertias} = \left(\frac{Q}{Q-1}\right)^2\left(\lambda_k - \frac{1}{Q}\right)^2 \quad \text{only for} \lambda_k > \frac{1}{Q}$$

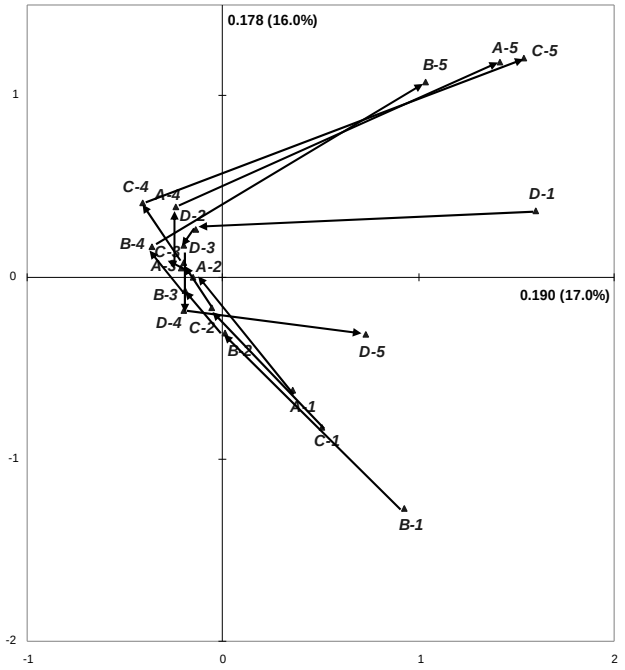
3. Calculate adjusted percentages of inertia :

   $$\text{adjusted percentages of inertia} = \frac{\text{adjusted principal inertias}}{\text{average off-diagonal inertia}}$$

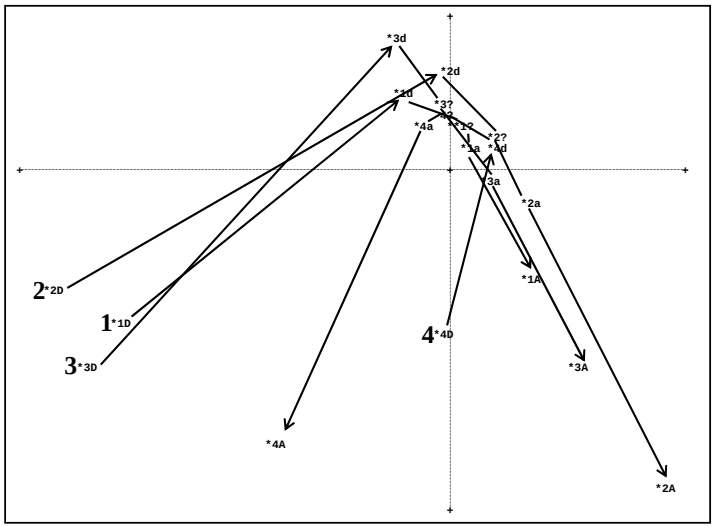
Analysis with adjusted eigenvalues



Analysis of Burt matrix



JCA



- The axes have “flipped” compared to previous maps: axis 1 ↔ axis 2
- Note: In above map labels are slightly different and arrows are drawn in reverse direction compared to previous maps...
- % of inertia explained by axes 1 & 2: 80.3% (75% before, with adjustments)
- Axes are not “nested”, as in CA and MCA.

Correspondence Analysis & Related Methods

Michael Greenacre

SESSION 18: **Multiple correspondence analysis & optimal scaling of categories**

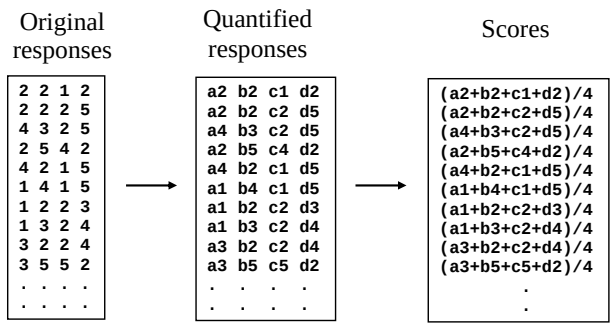
Output from SPSS-Homals (1)

```
 H O M A L S - VERSION 0.6
                                    BY
                          DEPARTMENT OF DATA THEORY
                      UNIVERSITY OF LEIDEN, THE NETHERLANDS

 The number of observations used in the analysis = 916

                          List of Variables
                          =================

   Variable     Variable label                          Number of
                                                        Categories
   V9         Science: believe too often in                5
   V10        Science: more harm than good                 5
   V11        Change in nature make things worse           5
   V12        Science: solve environmental problems        5

The iterative process stops because the convergence has been reached in 28
iteration(s).


Dimension   Eigenvalue
---------   ----------
    1          .4355
    2          .4225


Discrimination measures per variable per dimension
==================================================

Variable  Dimension

               1       2

V9          .335    .400
V10         .520    .627
V11         .468    .550
V12         .419    .113
```

Output from SPSS-Homals (2)

```
             Marginal Frequencies and Category Quantifications
             =================================================

Variable: V9           Science: believe too often in
---------
Category                       Marginal Frequency
--------                       ------------------
    1    Strongly Agree                188
    2    Agree                         389
    3    Neither Agree nor Di          142
    4    Disagree                      155
    5    Strongly Disagree              42


Category Quantifications
------------------------

Category   Dimensions
--------   ----------
               1        2

      1      .54     -.95
      2     -.23     -.00
      3     -.32      .08
      4     -.38      .60
      5     2.15     1.82

...and so on, for each variable (these are exactly the principal coordinates)
```

Quantification as a goal - Homogeneity Analysis (HOMALS)



Objective of HOMALS: to determine the set of J scale values **a1, a2, ..., b1, b2, ...,** etc... so that the implied scores for each individual are as "close" as possible to that individual's particular set of Q scale values.

"Closeness" is defined in terms of squared sum of differences, and the solution is obtained by least-squares – this is mathematically equivalent to maximising the sum of squared correlations between the scores and the quantified responses (cf. canonical correlation definition of CA...)

Optimal scale values and scores

Original responses	Quantified responses	Scores
2 2 1 2	a2 b2 c1 d2	(a2+b2+c1+d2)/4
2 2 2 5	a2 b2 c2 d5	(a2+b2+c2+d5)/4
4 3 2 5	a4 b3 c2 d5	(a4+b3+c2+d5)/4
2 5 4 2	a2 b5 c4 d2	(a2+b5+c4+d2)/4
4 2 1 5	a4 b2 c1 d5	(a4+b2+c1+d5)/4
1 4 1 5	a1 b4 c1 d5	(a1+b4+c1+d5)/4
1 2 2 3	a1 b2 c2 d3	(a1+b2+c2+d3)/4
1 3 2 4	a1 b3 c2 d4	(a1+b3+c2+d4)/4
3 2 2 4	a3 b2 c2 d4	(a3+b2+c2+d4)/4
3 5 5 2	a3 b5 c5 d2	(a3+b5+c5+d2)/4
. . . .	. . . .	.
. . . .	. . . .	.

MCA standard coordinates of the categories give the optimal scale values **a1,a2,..., b1,b2, ...,c1,c2,...** etc...

MCA principal coordinates of the respondents give the optimal scores.

Column contributions from MCA of indicator matrix Z

	frequency	mass	F1	F2	F3	F4	...
A-1	188	0.051	0.035	0.111	0.009	0.010	...
A-2	389	0.106	0.013	0.000	0.064	0.055	...
A-3	142	0.039	0.009	0.001	0.000	0.093	...
A-4	155	0.042	0.013	0.035	0.119	0.001	...
A-5	42	0.011	0.121	0.090	0.009	0.000	...
B-1	61	0.017	0.076	0.150	0.054	0.039	...
B-2	233	0.064	0.000	0.033	0.060	0.174	...
B-3	185	0.050	0.010	0.001	0.127	0.111	...
B-4	341	0.093	0.065	0.015	0.133	0.002	...
B-5	96	0.026	0.149	0.171	0.000	0.003	...
C-1	142	0.039	0.053	0.148	0.052	0.008	...
C-2	332	0.091	0.001	0.014	0.090	0.110	...
C-3	187	0.051	0.011	0.009	0.053	0.219	...
C-4	210	0.057	0.050	0.054	0.157	0.003	...
C-5	45	0.012	0.153	0.100	0.001	0.010	...
D-1	29	0.008	0.107	0.006	0.001	0.001	...
D-2	234	0.064	0.007	0.026	0.001	0.011	...
D-3	231	0.063	0.013	0.002	0.039	0.041	...
D-4	297	0.081	0.017	0.015	0.028	0.104	...
D-5	125	0.034	0.097	0.019	0.004	0.004	...
Eigenvalues			0.4355	0.4225	0.3450	0.2887	...

Column contributions from MCA of indicator matrix Z

	frequency	mass	F1	F2
A-1	188	0.051	0.035	0.111
A-2	389	0.106	0.013	0.000
A-3	142	0.039	0.009	0.001
A-4	155	0.042	0.013	0.035
A-5	42	0.011	0.121	0.090
B-1	61	0.017	0.076	0.150
B-2	233	0.064	0.000	0.033
B-3	185	0.050	0.010	0.001
B-4	341	0.093	0.065	0.015
B-5	96	0.026	0.149	0.171
C-1	142	0.039	0.053	0.148
C-2	332	0.091	0.001	0.014
C-3	187	0.051	0.011	0.009
C-4	210	0.057	0.050	0.054
C-5	45	0.012	0.153	0.100
D-1	29	0.008	0.107	0.006
D-2	234	0.064	0.007	0.026
D-3	231	0.063	0.013	0.002
D-4	297	0.081	0.017	0.015
D-5	125	0.034	0.097	0.019
Eigenvalues			0.4355	0.4225

.035+.013+.009+.013+.121 = .192

i.e. .192 of the first principal inertia

.192 x 0.4355 = 0.08344

Multiplying this part of inertia by the number of variables (4) we obtain:

0.08344 x 4 = 0.334

which is the discrimination value calculated by the program HOMALS:

Discrimination measures per variable per dimension

Variable	Dimension 1	2
V9	.335	.400
V10	.520	.627
V11	.468	.550
V12	.419	.113

Here's another example, the (West) German data set wg93 in the **ca** package: this is the indicator matrix solution.