Correspondence Analysis & Related Methods

Michael Greenacre

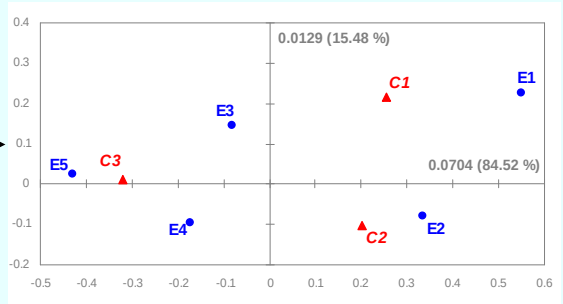
SESSION 9:

(SIMPLE) CORRESPONDENCE ANALYSIS: basic geometric concepts

Overview of CA and basic geometric concepts

- 312 respondents, all readers of a certain newspaper, cross-tabulated according to their education group and level of reading of the newspaper

	C1	C2	C3
E1	5	7	2
E2	18	46	20
E3	19	29	39
E4	12	40	49
E5	3	7	16



- **E1**: some primary **E2**: primary completed **E3**: some secondary **E4**: secondary completed **E5**: some tertiary
- ***C1***: glance ***C2***: fairly thorough ***C3***: very thorough

Profile

- A profile is a set of relative frequencies, that is a set of frequencies expressed relative to their total (often in percentage form).
- Each row or each column of a table of frequencies defines a different profile.
- It is these profiles which CA visualises as points in a map.

original data

	C1	C2	C3	
E1	5	7	2	14
E2	18	46	20	84
E3	19	29	39	87
E4	12	40	49	101
E5	3	7	16	26
	57	129	126	312

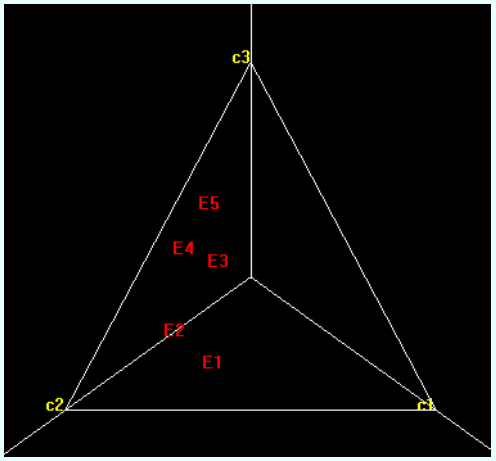
row profiles

	C1	C2	C3	
E1	.36	.50	.14	1
E2	.21	.55	.24	1
E3	.22	.33	.45	1
E4	.12	.40	.49	1
E5	.12	.27	.62	1

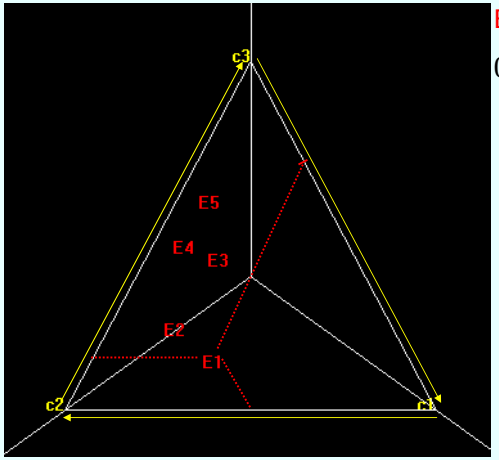
column profiles

	C1	C2	C3
E1	.09	.05	.02
E2	.32	.37	.16
E3	.33	.22	.31
E4	.21	.31	.39
E5	.05	.05	.13
	1	1	1

Row profiles viewed in 3-d



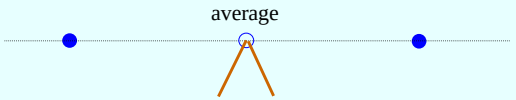
Plotting profiles in profile space (triangular coordinates)



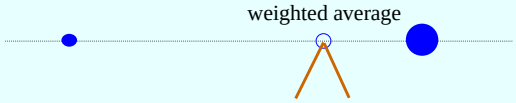
E1 :

0.36 0.50 0.14

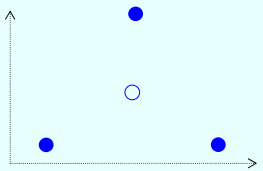
Weighted average (centroid)

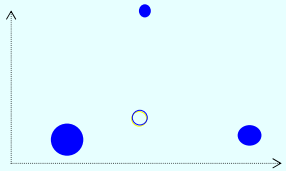


The average is the point at which the two points are balanced.

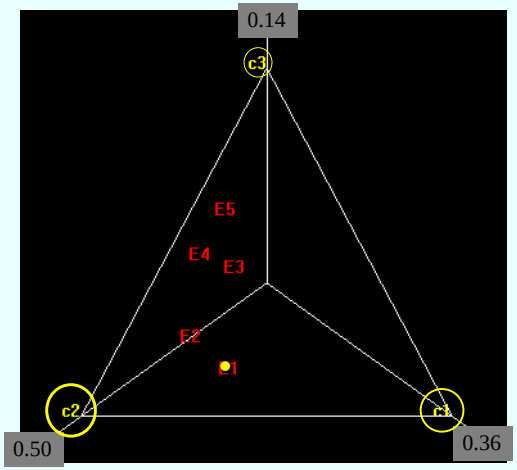


The situation is identical for multidimensional points...





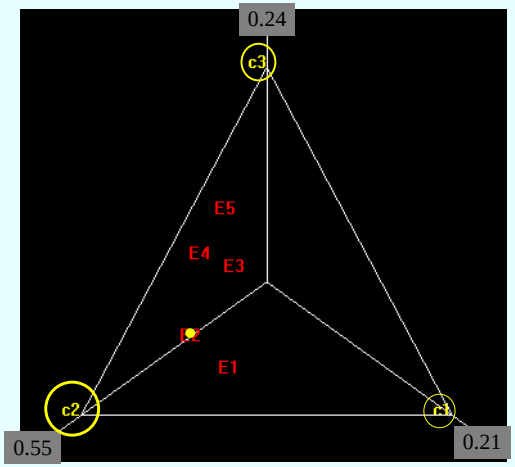
Plotting profiles in profile space (barycentric - or weighted average - principle)



E1:

0.36 0.50 0.14

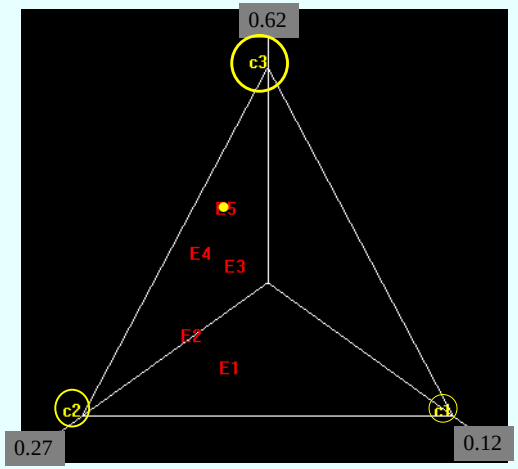
Plotting profiles in profile space (barycentric - or weighted average - principle)



E2:

0.21 0.55 0.24

Plotting profiles in profile space (barycentric - or weighted average - principle)



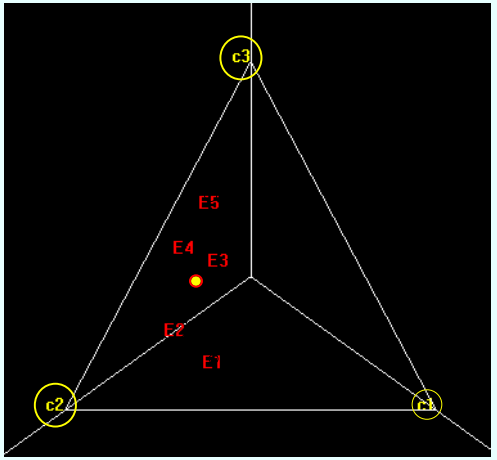
E5:

0.12 0.27 0.62

Masses of the profiles

original data

	C1	C2	C3		masses
E1	5	7	2	14	.045
E2	18	46	20	84	.269
E3	19	29	39	87	.279
E4	12	40	49	101	.324
E5	3	7	16	26	.083
	57	129	126	312	1
average row profile	.183	.413	.404	1	



Readership data

	Education Group	C1	C2	C3	Total	Mass
E1	Some primary	5 (0.357)	7 (0.500)	2 (0.143)	14	**0.045**
E2	Primary completed	18 (0.214)	46 (0.548)	20 (0.238)	84	**0.269**
E3	Some secondary	19 (0.218)	29 (0.333)	39 (0.448)	87	**0.279**
E4	Secondary completed	12 (0.119)	40 (0.396)	49 (0.485)	101	**0.324**
E5	Some tertiary	3 (0.115)	7 (0.269)	16 (0.615)	26	**0.083**
	Total	57 **(0.183)**	129 **(0.413)**	126 **(0.404)**	312	

***C1**: glance **C2**: fairly thorough **C3**: very thorough*

Calculating chi-square

$$\chi^2 = \text{12 similar terms ....} + \frac{(3 - 4.76)^2}{4.76} + \frac{(7 - 10.74)^2}{10.74} + \frac{(16 - 10.50)^2}{10.50}$$

$$= 26.0$$

	Education Group	C1	C2	C3	Total	Mass
....	.....	....	....	....	**14**	**....**
....	.....	....	....	....	**84**	**....**
....	.....	....	....	....	**87**	**....**
....	.....	....	....	....	**101**	**....**
E5	*Observed Frequency* Some tertiary *Expected Frequency*	**3** (0.115) ***4.76***	**7** (0.269) ***10.74***	**16** (0.615) ***10.50***	**26**	***0.083***
	Total	57 **(0.183)**	129 **(0.413)**	126 **(0.404)**	**312**	

*For example, expected frequency of (**E5**,**C1**):*

0.183 x **26** = ***4.76***

Calculating chi-square

$$\chi^2 = \textit{12 similar terms ....}$$

$$+ 26 \left[\frac{(3/26 - 4.76/26)^2}{4.76/26} + \frac{(7/26 - 10.74/26)^2}{10.74/26} + \frac{(16/26 - 10.50/26)^2}{10.50/26} \right]$$

$$\chi^2 / \mathbf{312} = \textit{12 similar terms ....}$$

$$+ 0.083 \left[\frac{(0.115 - 0.183)^2}{0.183} + \frac{(0.269 - 0.413)^2}{0.413} + \frac{(0.615 - 0.404)^2}{0.404} \right]$$

	Education Group	C1	C2	C3	Total	Mass
….	…….	….	….	….	**14**	**….**
….	…….	….	….	….	**84**	**….**
….	…….	….	….	….	**87**	**….**
….	…….	….	….	….	**101**	**….**
E5	*Observed Frequency* Some tertiary *Expected Frequency*	**3** *(0.115)* ***4.76***	**7** *(0.269)* ***10.74***	**16** *(0.615)* ***10.50***	**26**	***0.083***
	Total	57 ***(0.183)***	129 ***(0.413)***	126 ***(0.404)***	**312**	

Calculating inertia

Inertia = $\chi^2 / 312$ = *similar terms for first four rows ...*

$$+ 0.083 \left[\frac{(0.115 - 0.183)^2}{0.183} + \frac{(0.269 - 0.413)^2}{0.413} + \frac{(0.615 - 0.404)^2}{0.404} \right]$$

mass (of row **E5**)

squared chi-square distance (between the profile of **E5** and the average profile)

$$\text{Inertia} = \sum \textit{mass} \times (\textit{chi-square distance})^2$$

$$\frac{(0.115 - 0.183)^2}{0.183} + \frac{(0.269 - 0.413)^2}{0.413} + \frac{(0.615 - 0.404)^2}{0.404}$$

EUCLIDEAN

WEIGHTED

How can we see chi-square distances?

Inertia = $\chi^2 / 312$ = *similar terms for first four rows ...*

$$+ 0.083 \left[\frac{(0.115 - 0.183)^2}{0.183} + \frac{(0.269 - 0.413)^2}{0.413} + \frac{(0.615 - 0.404)^2}{0.404} \right]$$

mass (of row **E5**)

squared chi-square distance (between the profile of **E5** and the average profile)

$$\frac{(0.115 - 0.183)^2}{0.183} + \frac{(0.269 - 0.413)^2}{0.413} + \frac{(0.615 - 0.404)^2}{0.404}$$

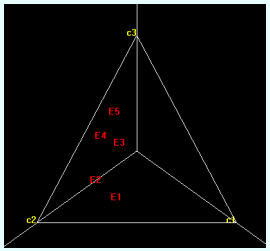
EUCLIDEAN

WEIGHTED

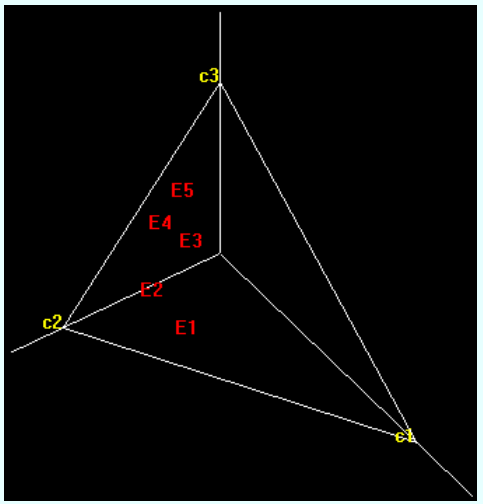
$$\left(\frac{0.115}{\sqrt{0.183}} - \frac{0.183}{\sqrt{0.183}} \right)^2 + \left(\frac{0.269}{\sqrt{0.413}} - \frac{0.413}{\sqrt{0.413}} \right)^2 + \left(\frac{0.615}{\sqrt{0.404}} - \frac{0.404}{\sqrt{0.404}} \right)^2$$

So the answer is to divide all profile elements by the √ of their averages

"Stretched" row profiles viewed in 3-d chi-squared space



"Pythagorian" – ordinary Euclidean distances



Chi-square distances

Summary: Basic geometric concepts

- **Profiles** are rows or columns of relative frequencies, that is the rows or columns expressed relative to their respective marginals, or bases.
- Each profile has a weight assigned to it, called the **mass**, which is proportional to the original marginal frequency used as a base .
- The **average profile** is the the centroid (weighted average) of the profiles.
- **Vertex profiles** are the extreme profiles in the profile space ("simplex").
- Profiles are weighted averages of the vertices, using the profile elements as weights.
- The **dimensionality** of an I x J matrix = min{I – 1, J – 1}
- The **chi-square distance** measures the difference between profiles, using an Euclidean-type function which standardizes each profile element by dividing by the square root of its expected value.
- The **(total) inertia** can be expressed as the weighted average of the squared chi-square distances between the profiles and their average.

The "famous" smoking data: row problem

- (see "Correspondence Analysis in Practice") – artificial example designed to illustrate two-dimensional maps

	no	li	me	hv
Senior managers SM	4	2	3	2
Junior managers JM	4	3	7	4
Senior employees SE	25	10	12	4
Junior employees JE	18	24	33	13
Secretaries SC	10	6	7	2

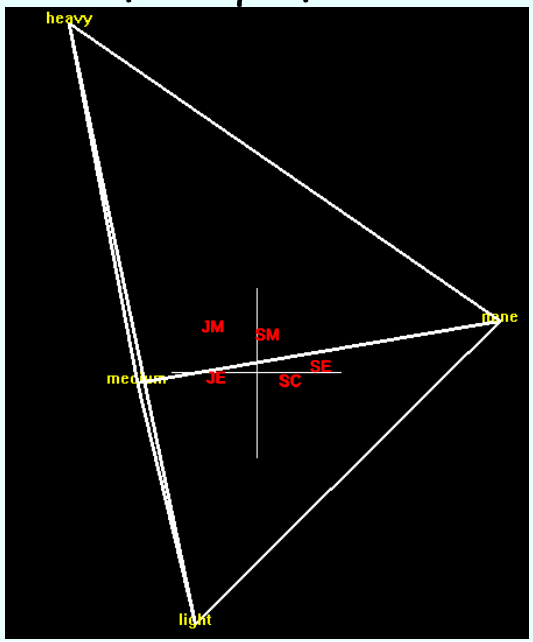
→ *row profiles*

SM	.36	.18	.27	.18
JM	.22	.17	.39	.22
SE	.49	.20	.24	.08
JE	.20	.27	.38	.15
SC	.40	.24	.28	.08
ave	.32	.23	.32	.13

- 193 employees of a firm
- 5 categories of staff group
- 4 categories of smoking (none,light,medium,heavy)

none	1	0	0	0
light	0	1	0	0
medium	0	0	1	0
heavy	0	0	0	1

View of row profiles in 3-d



The "famous" smoking data: column problem

	no	li	me	hv
Senior managers SM	4	2	3	2
Junior managers JM	4	3	7	4
Senior employees SE	25	10	12	4
Junior employees JE	18	24	33	13
Secretaries SC	10	6	7	2

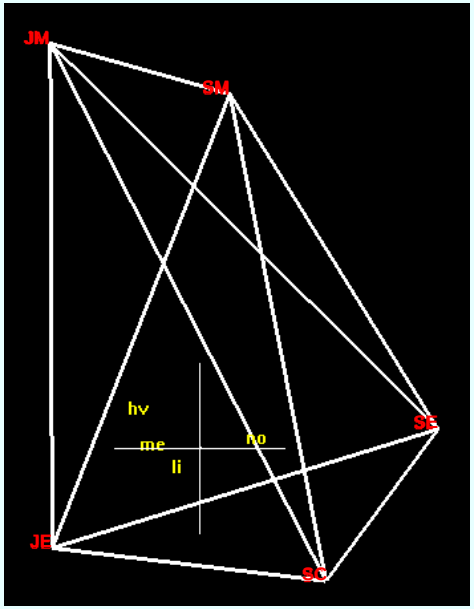
It seems like the column profiles, with 5 elements, are 4-dimensional, BUT there are only 4 points and 4 points lie exactly in 3 dimensions.

So the dimensionality of the columns is the same as the rows.

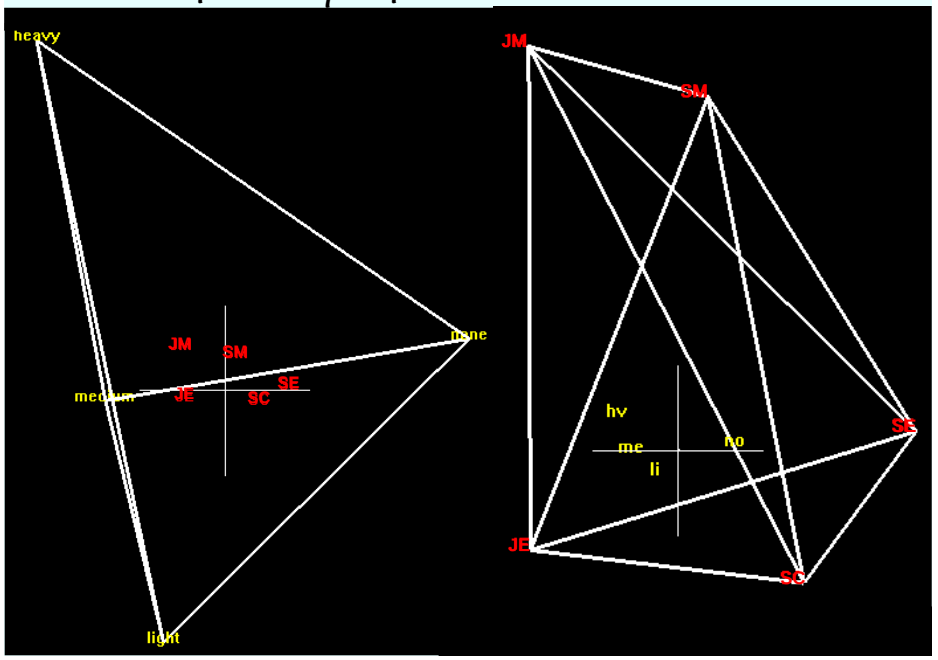
column profiles

no	li	me	hv	ave	SM	JM	SE	JE	SC
.07	.04	.05	.08	.06	1	0	0	0	0
.07	.07	.11	.16	.09	0	1	0	0	0
.41	.22	.19	.16	.26	0	0	1	0	0
.30	.53	.53	.52	.46	0	0	0	1	0
.16	.13	.11	.08	.13	0	0	0	0	1

View of column profiles in 3-d



View of both profiles and vertices in 3-d



Correspondence Analysis & Related Methods

Michael Greenacre

SESSION 10:

(SIMPLE) CORRESPONDENCE ANALYSIS: SVD theory

What CA does…

- … centres the row and column profiles with respect to their average profiles, so that the origin represents the average.
- … re-defines the dimensions of the space in an ordered way: first dimension "explains" the maximum amount of inertia possible in one dimension; second adds the maximum amount to first (hence first two explain the maximum amount in two dimensions), and so on… until all dimensions are "explained".
- … decomposes the total inertia along the principal axes into principal inertias, usually expressed as % of the total.
- … so if we want a low-dimensional version, we just take the first (principal) dimensions

The row and column problem solutions are closely related, one can be obtained from the other; there are simple scaling factors along each dimension relating the two problems.

Generalized SVD (repeat)

We often want to associate weights on the rows and columns, so that the fit is by weighted least-squares, not ordinary least squares, that is we want to minimize

$$\text{RSS} = \sum_{i=1}^{n}\sum_{j=1}^{p} r_i c_j (x_{ij} - x^*_{ij})^2$$

$$\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{D}_c^{1/2} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \quad \text{where } \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I},\ \alpha_1 \geq \alpha_2 \geq \cdots \geq 0$$

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha(\mathbf{D}_c^{-1/2}\mathbf{V})^\mathsf{T}$$

$$\mathbf{X}^* = etc...$$

Weighted metric multidimensional scaling (repeat)

- Suppose we want to represent the (centred) rows of a matrix $\mathbf{Y}$, weighted by (positive) elements down diagonal of matrix $\mathbf{D}_r$, where distance between rows is in the (weighted) metric defined by matrix $\mathbf{D}_m^{-1}$.
- Total inertia = $\sum_i\sum_j q_i (1/m_j) y_{ij}^2$
- $\mathbf{S} = \mathbf{D}_q^{½}\,\mathbf{Y}\,\mathbf{D}_m^{-½} = \mathbf{U}\,\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}$ where $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$
- Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_q^{-½}\,\mathbf{U}\,\mathbf{D}_\alpha$
- Principal axes of the rows: $\mathbf{D}_m^{½}\,\mathbf{V}$
- Standard coordinates of columns: $\mathbf{G} = \mathbf{D}_m^{-½}\,\mathbf{V}$
- Variances (inertias) explained: $\lambda_1 = \alpha_1^2,\ \lambda_2 = \alpha_2^2, \ldots$

Correspondence analysis

Of the rows:

- **Y** is the centred matrix of row profiles
- row masses in $\mathbf{D}_q$ are the relative frequencies of the rows
- column weights in $\mathbf{D}_w$ are the inverses of the relative frequencies of the columns
- Total inertia = χ^2/n

Of the columns:

- **Y** is the centred matrix of column profiles
- column masses in $\mathbf{D}_q$ are the relative frequencies of the columns
- row weights in $\mathbf{D}_w$ are the inverses of the relative frequencies of the rows
- Total inertia = χ^2/n

Both problems lead to the SVD of the same matrix

Correspondence analysis

- Table of nonnegative data **N**
- Divide **N** by its grand total n to obtain the so-called *correspondence matrix* $\mathbf{P} = (1/n)\,\mathbf{N}$
- Let the row and column marginal totals of **P** be the vectors **r** and **c** respectively, that is the vectors of row and column *masses*, and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses

⋮ *(to be derived algebraically in class)*

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2} \qquad \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

or equivalently

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}^{\mathsf{T}})\mathbf{D}_c^{1/2} \qquad \sqrt{r_i}\left(\frac{p_{ij}}{r_i c_j} - 1\right)\sqrt{c_j}$$

Principal coordinates $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$, $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$

Standard coordinates $\mathbf{\Phi} = \mathbf{D}_r^{-1/2}\mathbf{U}$, $\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$

Decomposition of total inertia along principal axes

	I rows	*(smoking I=5)*		*J* columns	*(smoking J=4)*
Total inertia	in(*I*)	0.08519		in(*J*)	0.08519
Inertia axis 1	λ_1	0.07476	(87.8%)	λ_1	0.07476
Inertia axis 2	λ_2	0.01002	(11.8%)	λ_2	0.01002
Inertia axis 3	λ_3	0.00041	(0.5%)	λ_3	0.00041

Duality (symmetry) of the rows and columns

	no	li	me	hv	sum
Senior managers SM	4	2	3	2	11
Junior managers JM	4	3	7	4	18
Senior employees SE	25	10	12	4	51
Junior employees JE	18	24	33	13	88
Secretaries SC	10	6	7	2	25
sum	61	45	62	25	

	row profiles				masses
SM	.36	.18	.27	.18	.06
JM	.22	.17	.39	.22	.09
SE	.49	.20	.24	.08	.26
JE	.20	.27	.38	.15	.46
SC	.40	.24	.28	.08	.13
ave	.32	.23	.32	.13	

column profiles	no	li	me	hv	ave
	.07	.04	.05	.08	.06
	.07	.07	.11	.16	.09
	.41	.22	.19	.16	.26
	.30	.53	.53	.52	.46
	.16	.13	.11	.08	.13
masses	.32	.23	.32	.13	

	SM	JM	SE	JE	SC
	1	0	0	0	0
	0	1	0	0	0
	0	0	1	0	0
	0	0	0	1	0
	0	0	0	0	1

	no	li	me	hv
no	1	0	0	0
li	0	1	0	0
me	0	0	1	0
hv	0	0	0	1

Relationship between row and column solutions

	rows	columns
standard coordinates	$\Phi = [\ \phi_{ik}\ ]$	$\Gamma = [\ \gamma_{jk}\ ]$
principal coordinates	$\mathbf{F} = [\ f_{ik}\ ]$	$\mathbf{G} = [\ g_{jk}\ ]$
relationships between coordinates	$\mathbf{F} = \Phi \mathbf{D}_\alpha$	$\mathbf{G} = \Gamma \mathbf{D}_\alpha$
	$f_{ik} = \alpha_k\, x_{ik}$	$g_{jk} = \alpha_k\, y_{jk}$

where $\alpha_k = \sqrt{\lambda_k}$ is the square root of the principal inertia on axis k

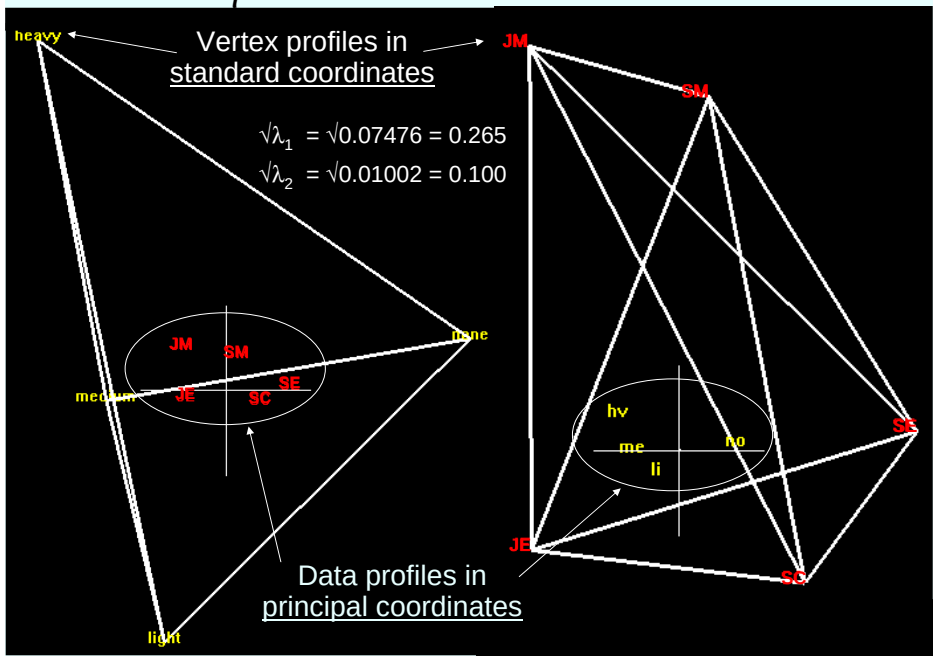
principal = standard $\times\ \alpha_k$

standard = principal / α_k

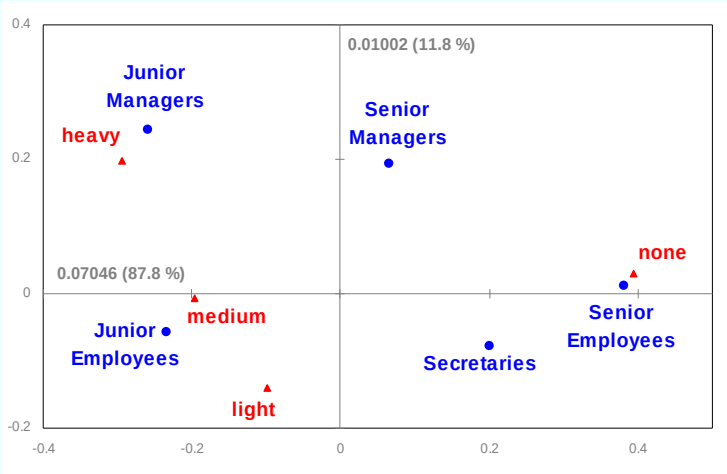
Data profiles in <u>principal coordinates</u>

Vertiex profiles in <u>standard coordinates</u>

Relationship between row and column solutions



Symmetric map using XLSTAT



<u>Summary:</u> Relationship between row and column solutions

1. same dimensionality (*rank*) = min{I –1, J – 1}
2. same total inertia and same principal inertias λ_1, λ_2, …, on each dimension (i.e., same decomposition of inertia along principal axes), hence same percentages of inertia on each dimension
3. "same" coordinate solutions, up to a scalar constant along each principal axis, which depends on the square root $\sqrt{\lambda_k} = \alpha_k$ of the principal inertia on each axis:
   principal = standard $\times\ \sqrt{\lambda_k}$
   standard = principal / $\sqrt{\lambda_k}$
4. Asymmetric map: one set principal, other standard
5. Symmetric map: both sets principal