Correspondence Analysis and Related Methods

Michael Greenacre
Universitat Pompeu Fabra
Barcelona

www.econ.upf.edu/~michael

www.globalsong.net

10-26 May 2010

PROGRAM

Monday	May 10,	12:00 pm – 4:00 pm – SR Statistik
Tuesday	May 11,	2:00 pm – 6:00 pm – 2H564 (PC Labor Statistik)
Wednesday	May 12,	12:00 pm – 2:00 pm – SR Statistik
Wednesday	May 12,	2:00 pm – 4:00 pm – 2H564 (PC Labor Statistik)
Monday	May 17,	12:00 pm – 2:00 pm – SR Statistik
Tuesday	May 18,	2:00 pm – 4:00 pm – 2H564 (PC Labor Statistik)
Wednesday	May 19,	2:00 pm – 4:00 pm – SR Statistik
Tuesday	May 25,	10:00 am – 12:00 – 2H564 (PC Labor Statistik)
Wednesday	May 26,	10:00 am – 12:00 – SR Statistik

COURSE CONTENTS: main themes

Theme 1: Introduction to multivariate data and multivariate analysis

Theme 2: Geometric concepts of correspondence analysis and related methods

Theme 3: Theory of correspondence analysis and related methods: the SVD

Theme 4: Biplots

Theme 5: Diagnostics for interpretation

Theme 5: Multiple & joint correspondence analysis

Theme 6: Extension to other types of data: ratings, rankings, square matrices

Theme 7: Investigating stability using bootstrap; testing hypotheses using permutation test

BIBLIOGRAPHY and SUPPORTING MATERIAL

Greenacre, M. and Blasius, J. (2006). Multiple Correspondence Analysis and Related Methods. Chapman & Hall /CRC Press.

Greenacre, M. (2007). Correspondence Analysis in Practice, 2nd edition. Chapman & Hall/ CRC Press.

Some PDFs of selected articles…

Web page of course material and R scripts:

www.econ.upf.edu/~michael/CARME

Introduction to multivariate data and multivariate analysis

Introduction to multivariate data

- Let's start with some simple trivariate data...

Continuous variables

X1 – Purchasing power/capita (euros)
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6

Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

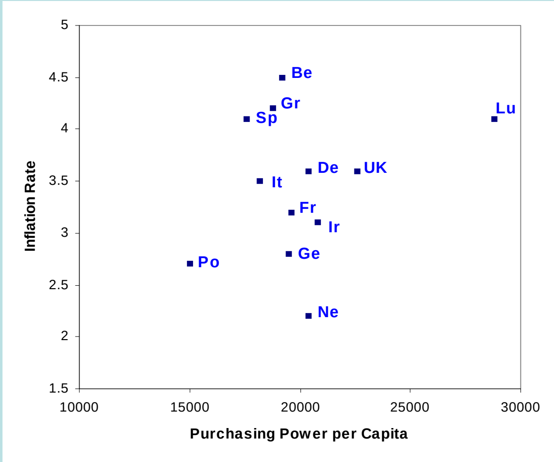
Education		*C1*	*C2*	*C3*
Some primary	E1	5	7	2
Primary completed	E2	18	46	20
Some secondary	E3	19	29	39
Secondary completed	E4	12	40	49
Some tertiary	E5	3	7	16

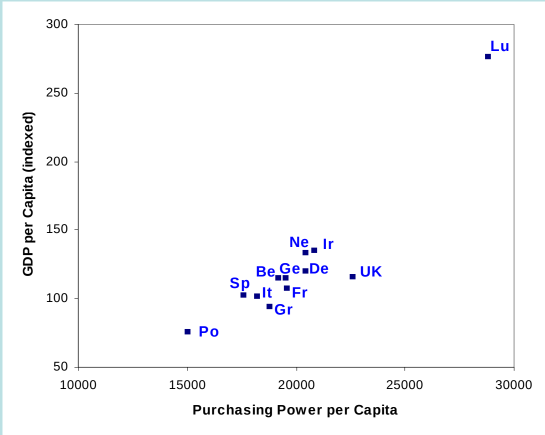
Visualizing trivariate continuous data

Continuous variables

X1 – Purchasing power/capita (euros)
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6



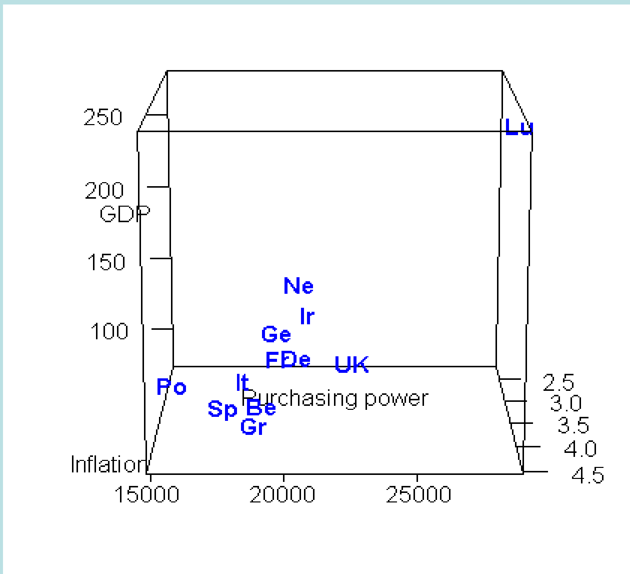


Visualizing trivariate continuous data

Continuous variables

X1 – Purchasing power/capita (euros)
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6

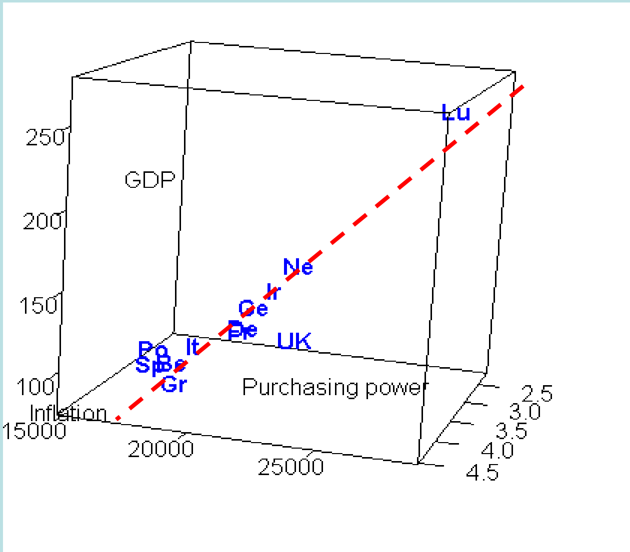


Visualizing trivariate continuous data

Continuous variables

X1 – Purchasing power/capita (euros
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6



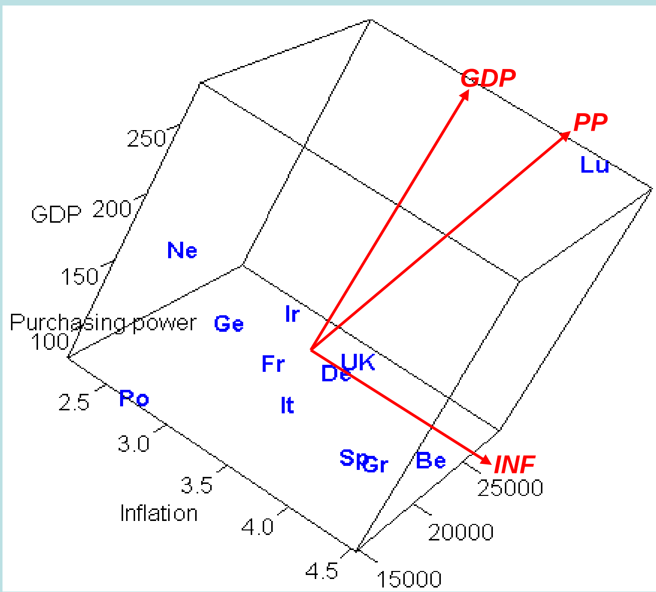
Visualizing trivariate continuous data

Continuous variables

X1 – Purchasing power/capita (euro
X2 – GDP/capita (index)
X3 – inflation rate (%)

cor	*X1*	*X2*	*X3*
X1	1.000	0.929	0.243
X2	0.929	1.000	0.207
X3	0.243	0.207	1.000

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6



Visualizing trivariate count data

Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

Education

		C1	C2	C3	
Primary incomplete	E1	5	7	2	*14*
Primary completed	E2	18	46	20	*84*
Secondary incomplete	E3	19	29	39	*87*
Secondary completed	E4	12	40	49	*101*
Some tertiary	E5	3	7	16	*263*

→

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*

Visualizing trivariate count data

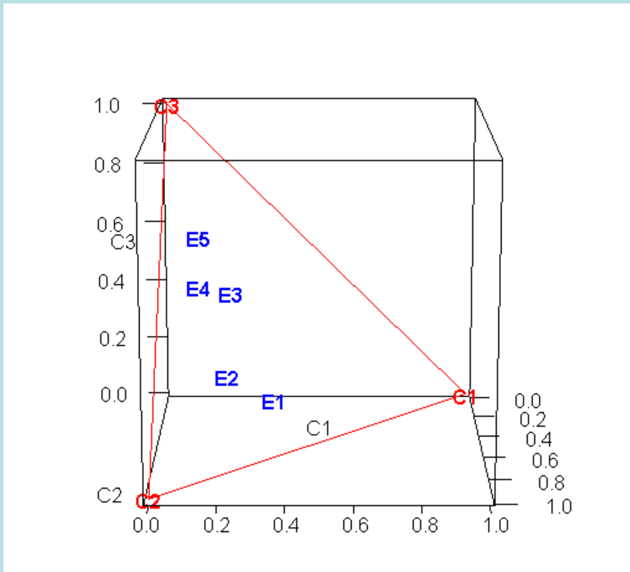
Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

row profiles

Education

		C1	C2	C3	
Some primary	E1	.36	.50	.14	*1*
Primary completed	E2	.21	.55	.24	*1*
Some secondary	E3	.22	.33	.45	*1*
Secondary completed	E4	.12	.40	.49	*1*
Some tertiary	E5	.12	.27	.62	*1*
	C1	1	0	0	*1*
	C2	1	0	0	*1*
	C3	1	0	0	*1*



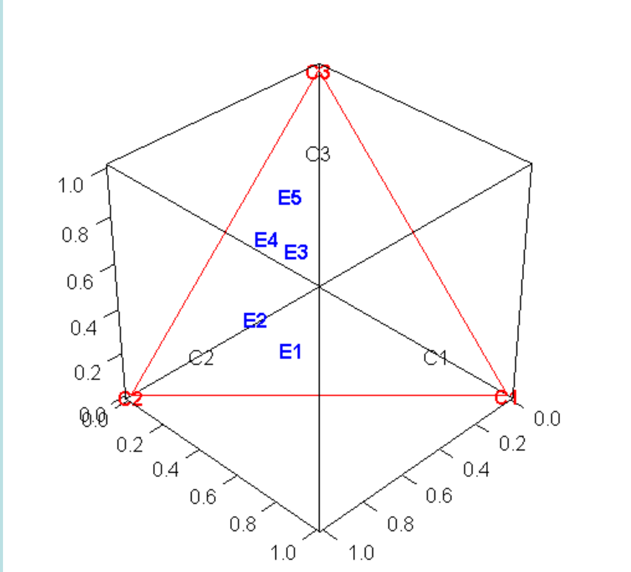
Visualizing trivariate count data

Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

row profiles

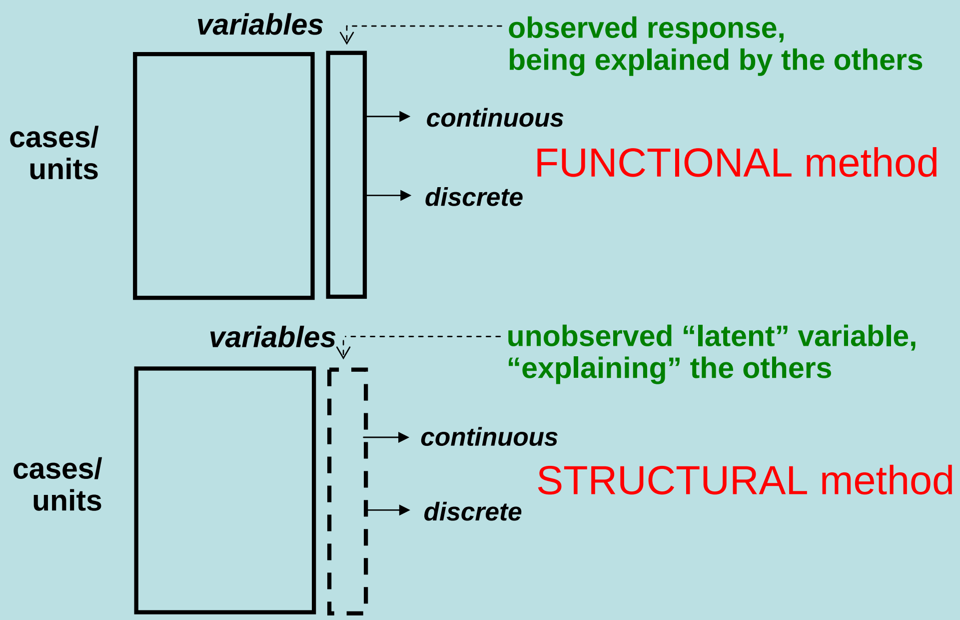
Education		*C1*	*C2*	*C3*	
Some primary	E1	.36	.50	.14	*1*
Primary completed	E2	.21	.55	.24	*1*
Some secondary	E3	.22	.33	.45	*1*
Secondary completed	E4	.12	.40	.49	*1*
Some tertiary	E5	.12	.27	.62	*1*
	C1	1	0	0	*1*
	C2	1	0	0	*1*
	C3	1	0	0	*1*



This is almost a correspondence analysis!

A basic scheme of multivariate analysis

All multivariate methods fall basically into two types, depending on the data structure and the question being asked:



Four corners of multivariate analysis

Structural equation models

Partial least squares

Canonical correspondence analysis

FUNCTIONAL
method

Generalized linear & additive models (GLMs & GAMs)

Classification and regression trees

- **regression**
  - *regression*
  - *analysis of (co)variance*
  - *general linear model*

- **classification**
  - *discriminant analysis*
  - *logistic regression*
  - *pattern recognition*

CONTINUOUS
variable of interest

DISCRETE
variable of interest

- **scaling**
  - *principal component analysis*
  - *factor analysis*
  - *correspondence analysis*
  - *multidimensional scaling*

- **cluster analysis**
  - *hierarchical clustering*
  - *nonhierarchical clustering*
  - *latent class analysis*

STRUCTURAL
method

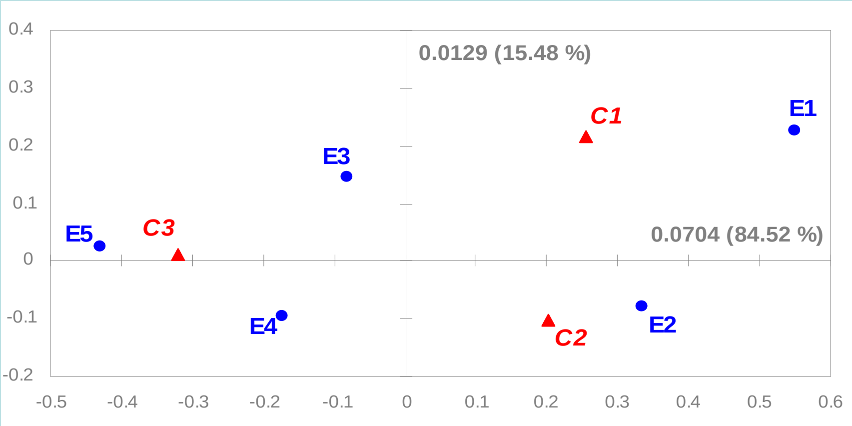
Basic geometric concepts of correspondence analysis and related methods

(principal component analysis, log-ratio analysis, discriminant analysis, multidimensional scaling…

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

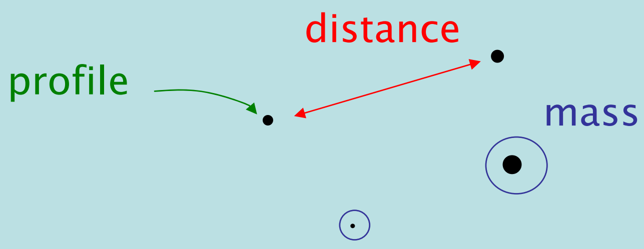
	C1	*C2*	*C3*
E1	5	7	2
E2	18	46	20
E3	19	29	39
E4	12	40	49
E5	3	7	16



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

- We use this simple example to explain the three basic concepts of CA: **profile**, **mass** and (chi-square) **distance**

Three basic geometric concepts



profile – *the coordinates (position) of the point*

mass – *the weight given to the point*

(chi-square) distance – *the measure of proximity between points*

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*

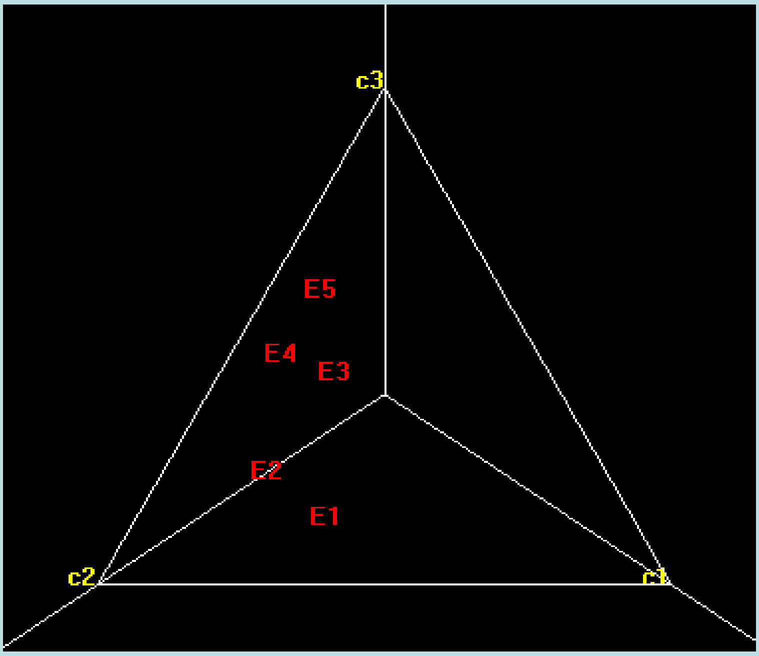
row profiles

	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*
	.18	*.41*	*.40*	*1*

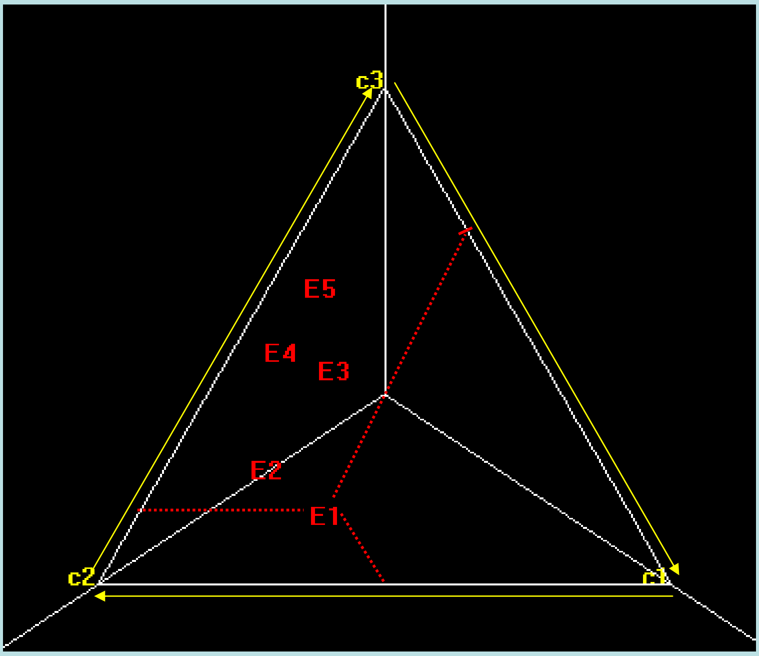
column profiles

	C1	C2	C3	
E1	.09	.05	.02	*.05*
E2	.32	.37	.16	*.27*
E3	.33	.22	.31	*.28*
E4	.21	.31	.39	*.32*
E5	.05	.05	.13	*.08*
	1	*1*	*1*	*1*

Row profiles viewed in 3-d



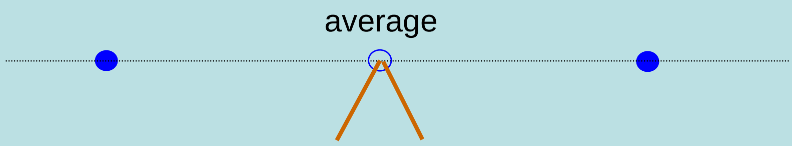
Plotting profiles in profile space (triangular coordinates)



E1 :

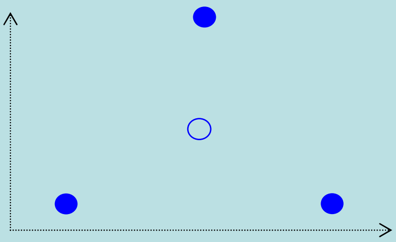
0.36 0.50 0.14

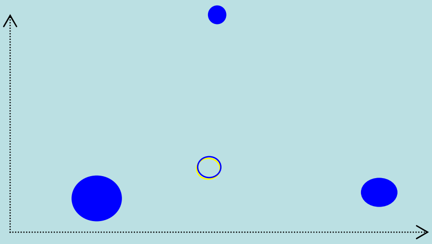
Weighted average (centroid)



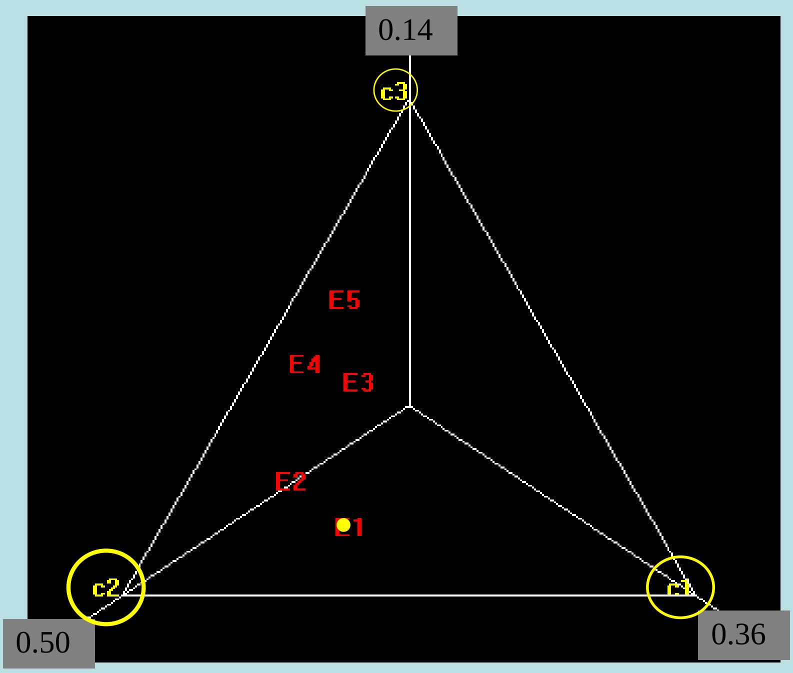
The average is the point at which the two points are balanced.

The situation is identical for multidimensional points...





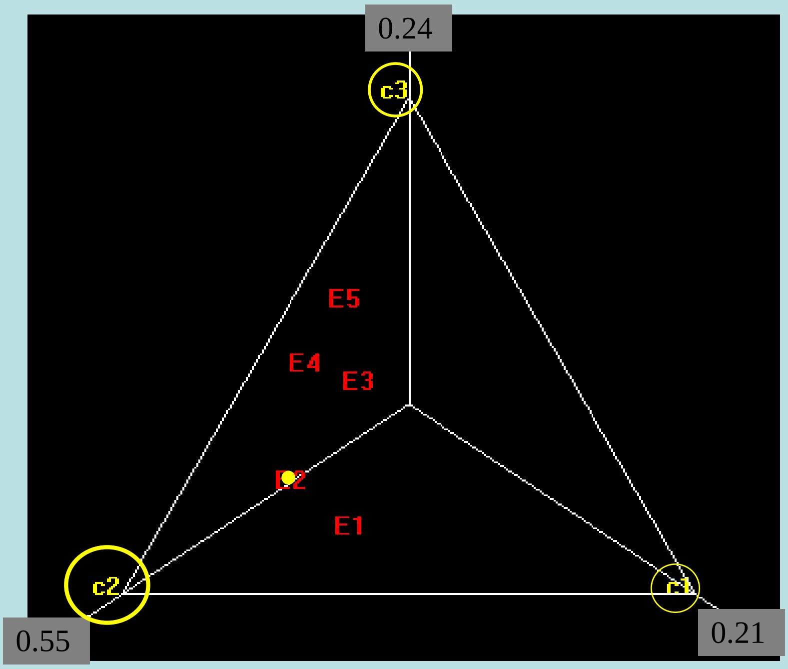
Plotting profiles in profile space (barycentric – or weighted average – principle)



E1:

0.36   0.50   0.14

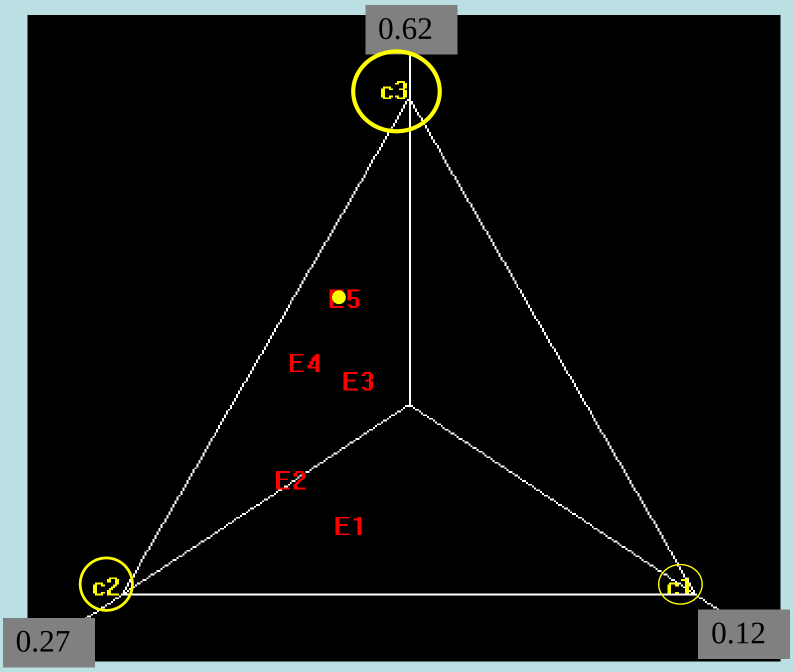
Plotting profiles in profile space (barycentric – or weighted average – principle)



E2:

0.21 0.55 0.24

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



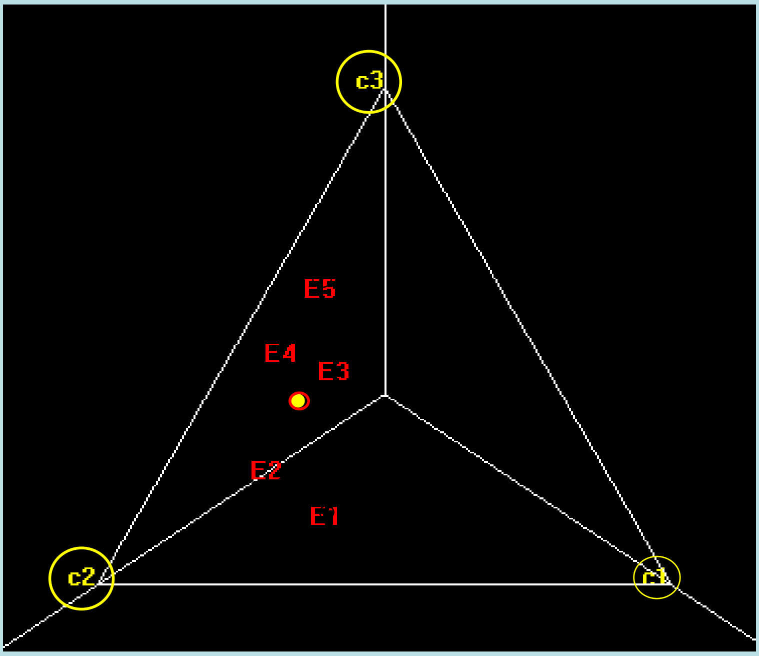
E5:

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

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

$$+ \frac{(3 - 4.76)^2}{4.76} + \frac{(7 - 10.74)^2}{10.74} + \frac{(16 - 10.50)^2}{10.50}$$

$$= \quad 26.0$$

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

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

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

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

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

	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

$$\text{Inertia} = \chi^2 / 312 = \textit{similar terms for first four rows ...}$$

$$+ \underbrace{0.083}_{\substack{\textbf{\textit{mass}} \\ \text{(of row } \textbf{E5}\text{)}}} \underbrace{\left[\frac{(0.115-0.183)^2}{0.183} + \frac{(0.269-0.413)^2}{0.413} + \frac{(0.615-0.404)^2}{0.404} \right]}_{\substack{\textbf{\textit{squared chi-square distance}} \\ \text{(between the profile of } \textbf{E5} \text{ and the} \\ \text{average profile)}}}$$

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

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

EUCLIDEAN

WEIGHTED

How can we see chi-square distances?

$$\text{Inertia} = \chi^2 / 312 = \textit{similar terms for first four rows ...}$$

$$+ \underbrace{0.083}_{\substack{\textbf{\textit{mass}} \\ \text{(of row } \textbf{E5}\text{)}}} \underbrace{\left[\frac{(0.115-0.183)^2}{0.183} + \frac{(0.269-0.413)^2}{0.413} + \frac{(0.615-0.404)^2}{0.404} \right]}_{\substack{\textbf{\textit{squared chi-square distance}} \\ \text{(between the profile of } \textbf{E5} \text{ and the} \\ \text{average profile)}}}$$

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

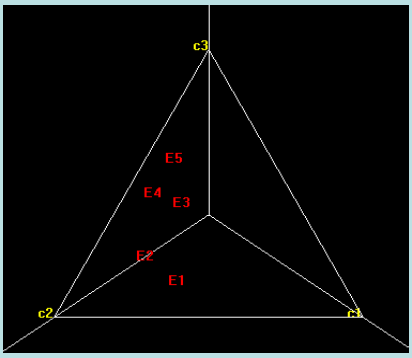
EUCLIDEAN

WEIGHTED

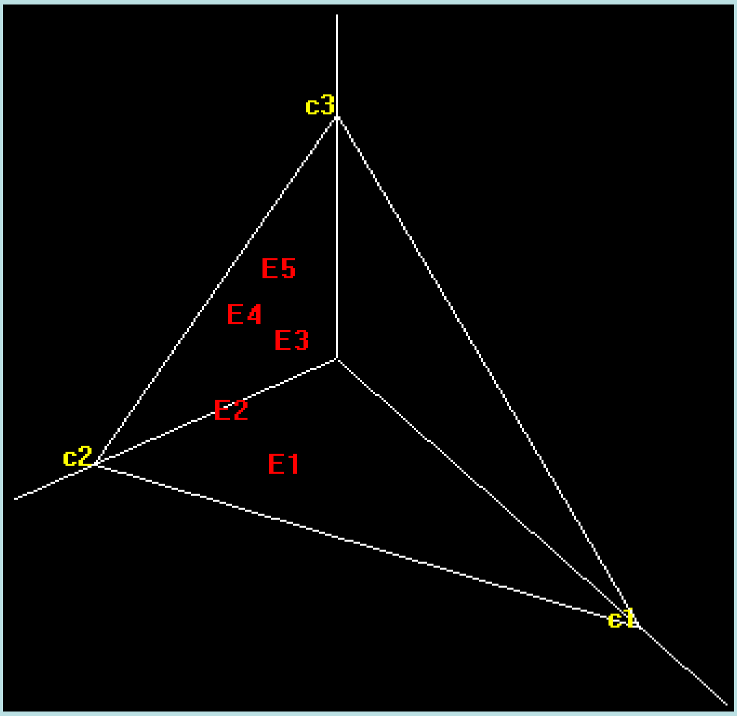
$$\left(\frac{0.115}{\sqrt{0.183}} - \frac{0.183}{\sqrt{0.183}} \right)^2 + \left(\frac{0.269}{\sqrt{0.413}} - \frac{0.413}{\sqrt{0.413}} \right)^2 + \left(\frac{0.615}{\sqrt{0.404}} - \frac{0.404}{\sqrt{0.404}} \right)^2$$

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

“Stretched” row profiles viewed in 3-d chi-squared space

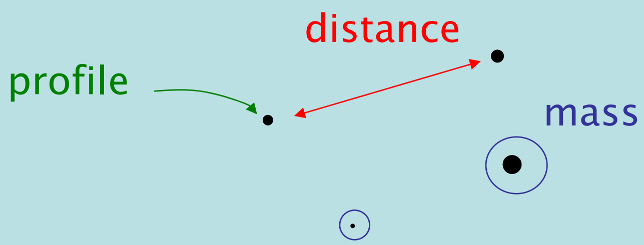


“Pythagorian” – ordinary Euclidean distances



Chi-square distances

Three basic geometric concepts

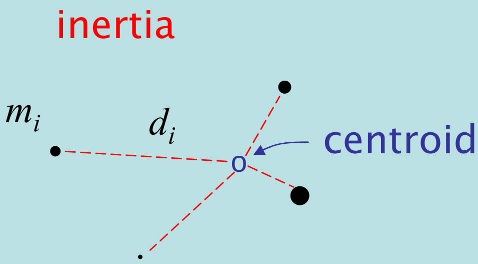


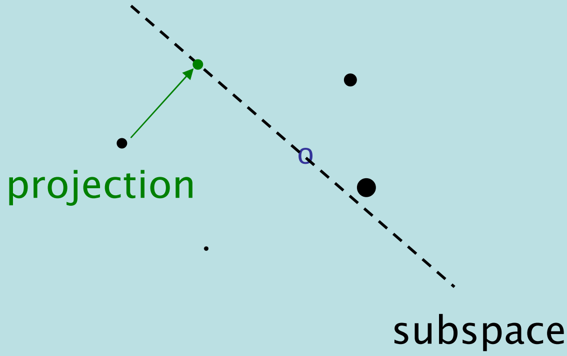
profile – *the coordinates (position) of the point*

mass – *the weight given to the point*

(chi-square) distance – *the measure of proximity between points*

Four derived geometric concepts





$\boxed{\text{inertia} = \sum_i m_i d_i^2}$

centroid – *the weighted average position*

inertia – *the weighted sum-of-squared distances to centroid*

subspace – *space of reduced dimensionality within the space (it will go through the centroid)*

projection – *the closest point in the subspace*

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 $I \times J$ 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 one-minute CA course

- The 'famous' smoking data.

staff group		smoking class none	light	medium	heavy	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*	*193*

- Now for the one-minute course in correspondence analysis, possible thanks to dynamic graphics!

One minute CA course: slide 1

3 columns

	light	medium	heavy	sum
SM	2	3	2	7
JM	3	7	4	14
SE	10	12	4	26
JE	24	33	13	70
SC	6	7	2	15

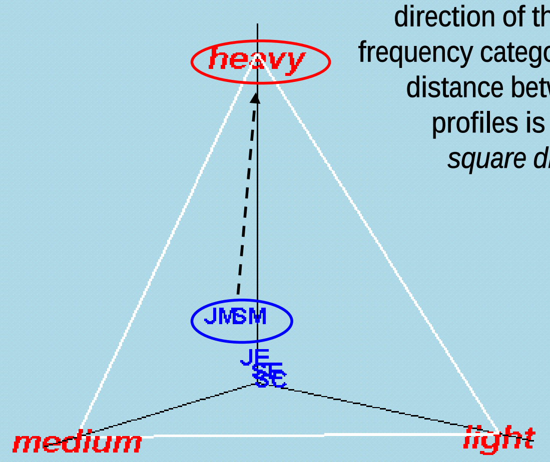
express relative to row sums

These are called "row profiles"

	light	medium	heavy	sum
SM	0.29	0.43	0.29	1
JM	0.21	0.50	0.29	1
SE	0.38	0.46	0.15	1
JE	0.34	0.47	0.19	1
SC	0.40	0.47	0.13	1

plot

N.B. The triangle is not equilateral: it has been stretched more in the direction of the lowest frequency category. The distance between the profiles is the "*chi-square distance*".



One minute CA course: slide 2

Relative values of row sums are used to weight the row profiles

4 columns

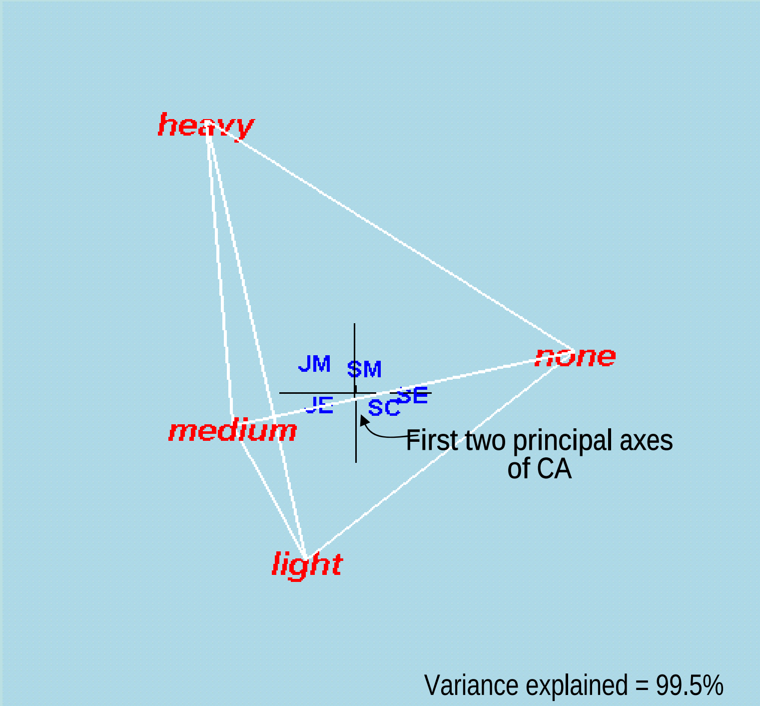
	none	*light*	*medium*	*heavy*	sum
SM	4	2	3	2	*11*
JM	4	3	7	4	*18*
SE	25	10	12	4	*51*
JE	18	24	33	13	*88*
SC	10	6	7	2	*25*
sum	*61*	*45*	*62*	*25*	*193*

express relative to row sums

	none	*light*	*medium*	*heavy*	sum
SM	0.36	0.18	0.27	0.18	*1*
JM	0.22	0.17	0.39	0.22	*1*
SE	0.49	0.20	0.24	0.08	*1*
JE	0.20	0.27	0.38	0.15	*1*
SC	0.40	0.24	0.28	0.08	*1*
sum	*0.32*	*0.23*	*0.32*	*0.13*	*1*

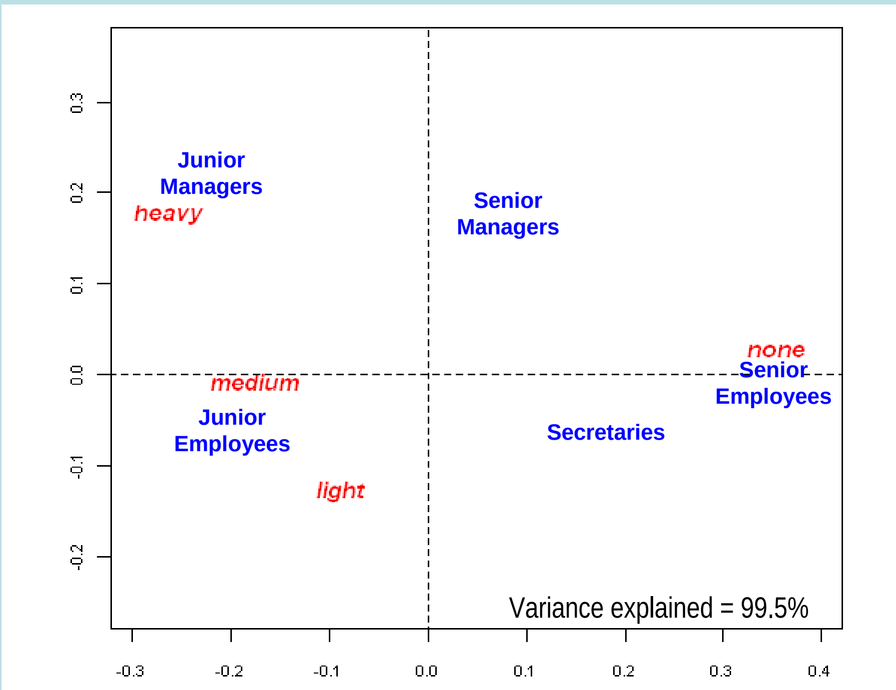
Average profile is the centre of the CA map

plot



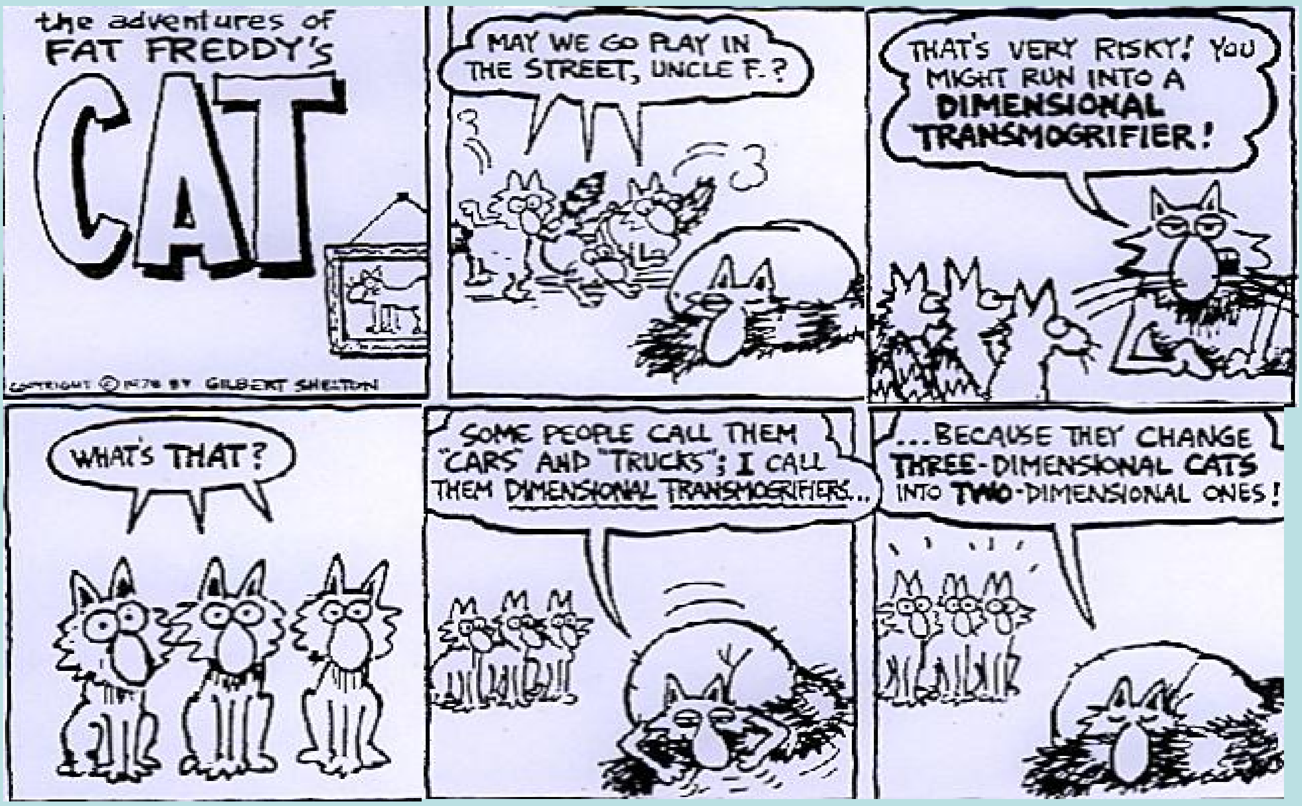
On minute CA course: slide 3

often rescale result so that rows and columns have same dispersions along the axes



Dimension reduction
Joint display of rows and columns

Dimensional Transmogrifier



with thanks to Jörg Blasius

The “famous” smoking data: row problem

- 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

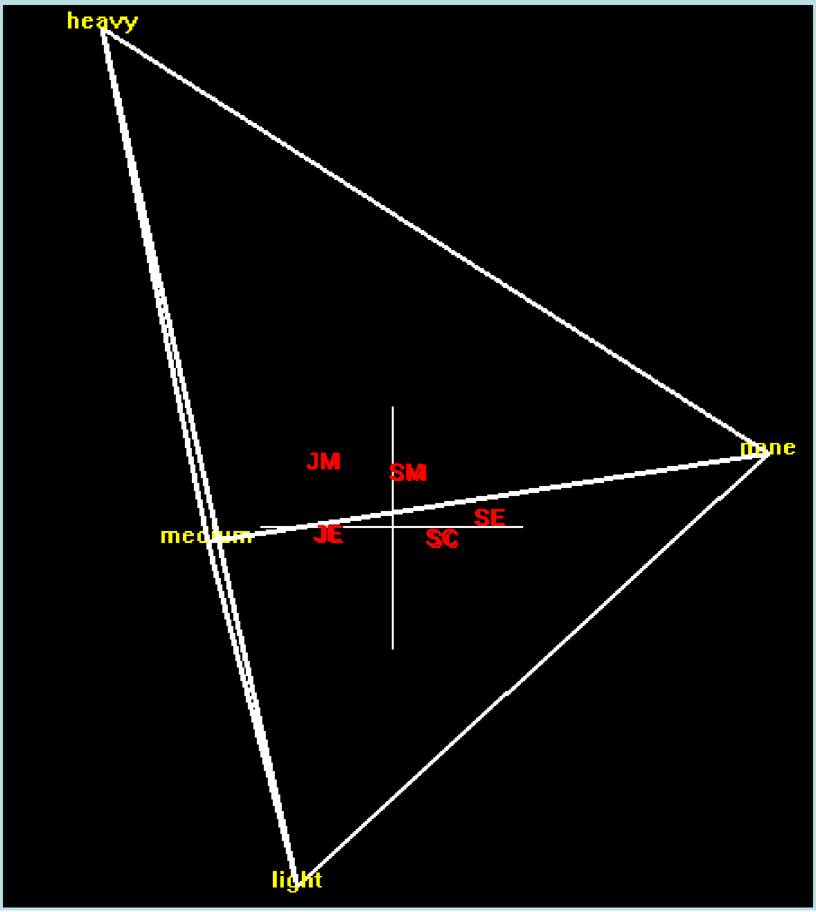
→

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

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

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

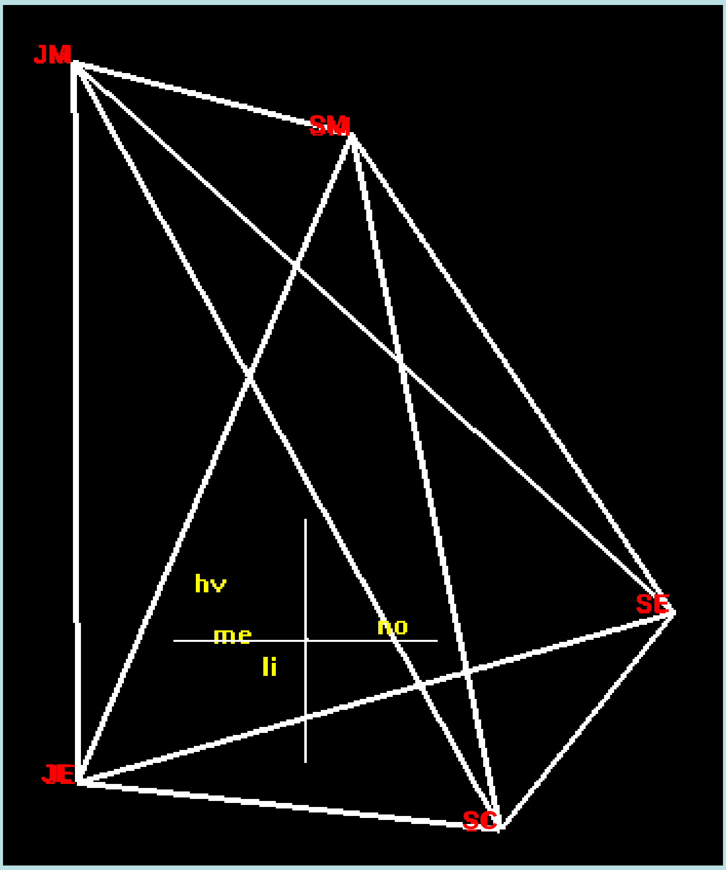
It seems like the column profiles, with 5 elements, are 4-dimensional, BUT there are only 4 points and 4 points lie exactly in 3 dimensions.

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

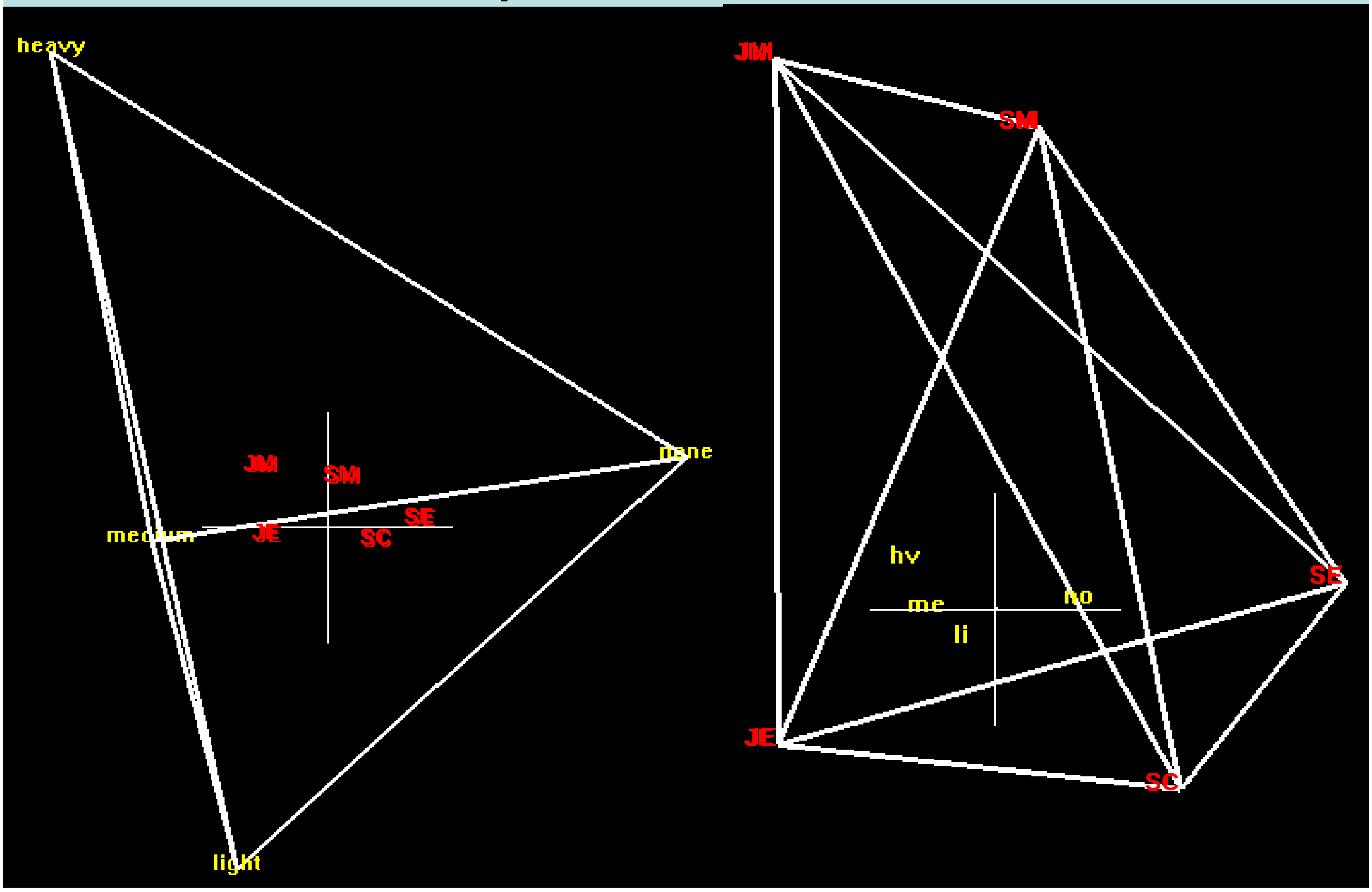
column profiles

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



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.

Singular value decomposition

Generalized principal component analysis

Generalized SVD

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

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

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

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

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

Generalized principal component analysis

- Suppose we want to represent the (centred) rows of a matrix **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 <u>*correspondence matrix*</u> $\mathbf{P} = (1/n)\,\mathbf{N}$
- Let the row and column marginal totals of **P** be the vectors **r** and **c** respectively, that is the vectors of row and column <u>*masses*</u>, and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses

⋮ *(to be derived algebraically in class)*

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

or equivalently

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

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

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

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	

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

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

	no	li	me	hv
masses	.32	.23	.32	.13

Relationship between row and column solutions

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

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

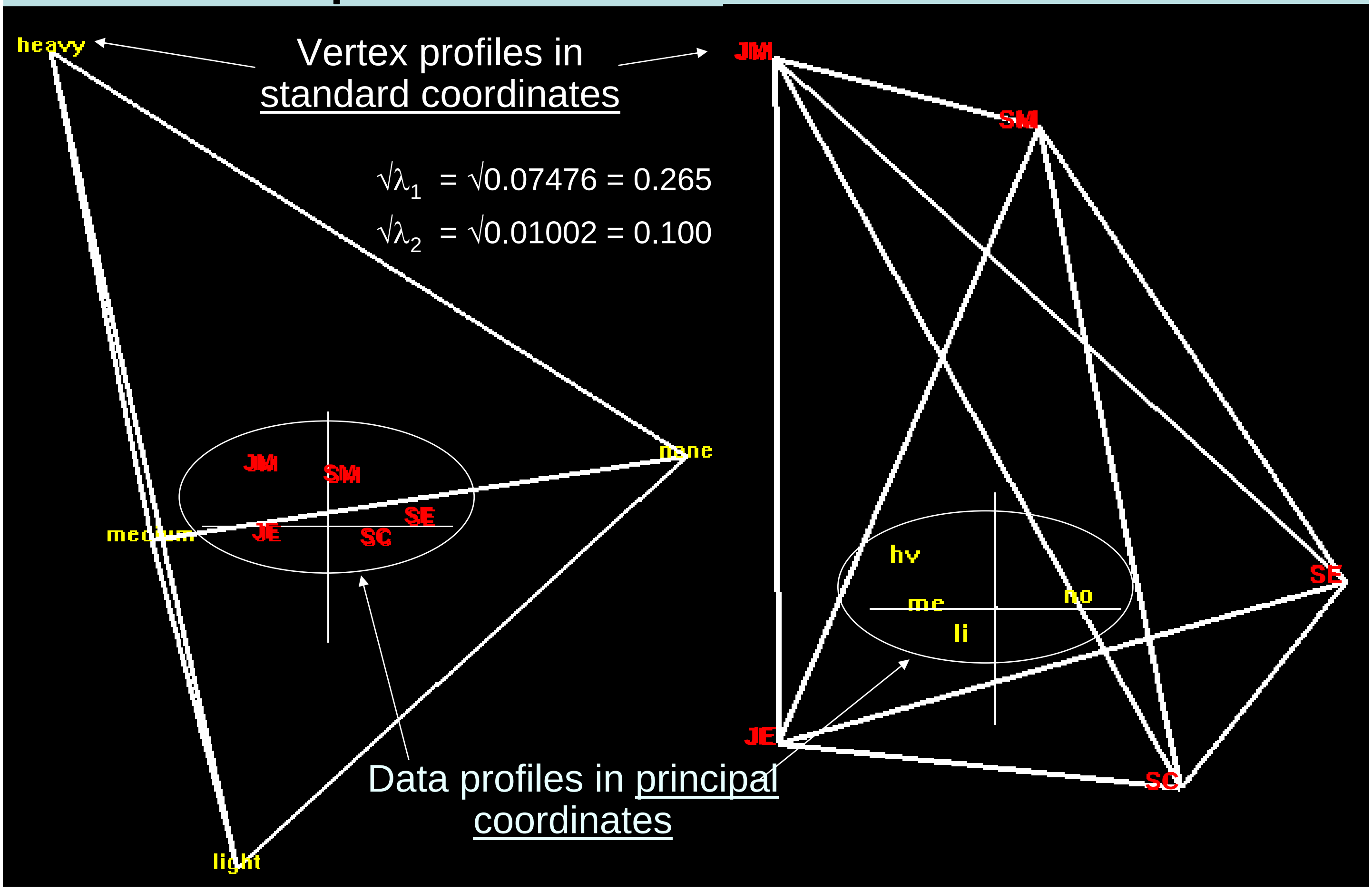
principal = standard $\times\ \alpha_k$

standard = principal / α_k

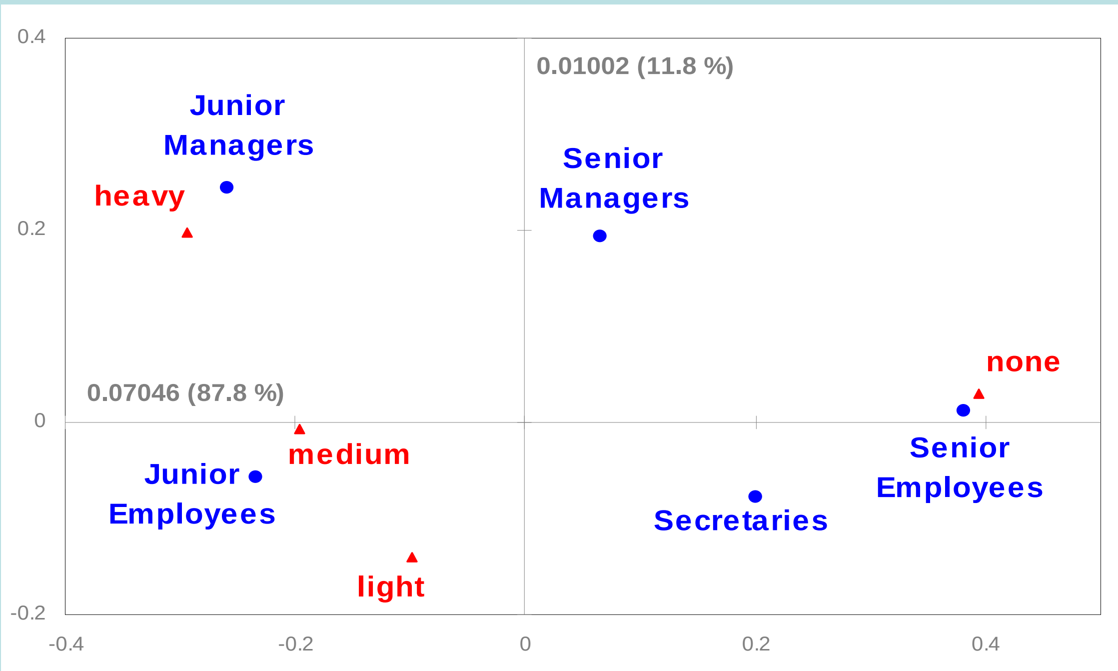
Data profiles in <u>principal coordinates</u>

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

Relationship between row and column solutions



Symmetric map using XLSTAT



Summary:
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