Correspondence Analysis (CA)

François Husson & Magalie Houée-Bigot

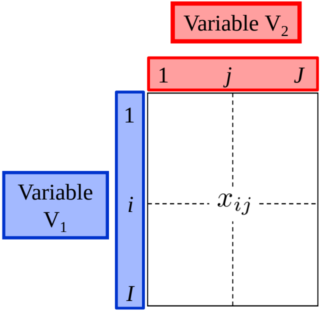
Department of Applied Mathematics - Rennes Agrocampus

husson@agrocampus-ouest.fr

Correspondence Analysis (CA)

Correspondence Analysis (CA)

Contingency tables



x_{ij}: number of individuals with
category i of the variable I
category j of the variable J

Characters in Phèdre (Racine)	Words	Number of times character i uses the word j
Perfume	Descriptors	Number of times perfume i was described with the word j
Biodiversity	Species	Abundance of species j in place i

$\Longrightarrow$ Examples where a χ^2 test for independence can be applied

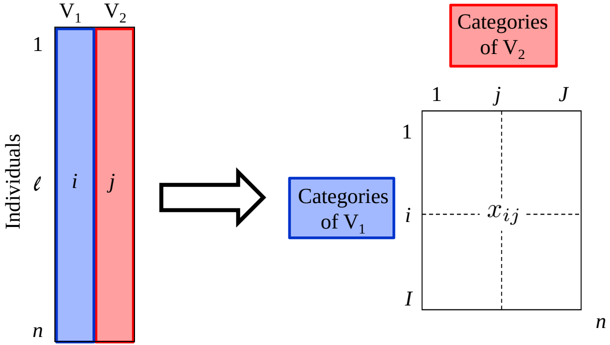
Nobel prize data

	Chemistry	Economic sciences	Literature	Medicine	Peace	Physics	Sum
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
Germany	24	1	8	18	5	24	80
Italy	1	1	6	5	1	5	19
Japan	6	0	2	3	1	11	23
Russia	4	3	5	2	3	10	27
UK	23	6	7	26	11	20	93
USA	51	43	8	70	19	66	257
Sum	121	60	49	140	51	149	570

Is there a link between countries and prize categories?
Do some countries specialize in certain prizes?

Data

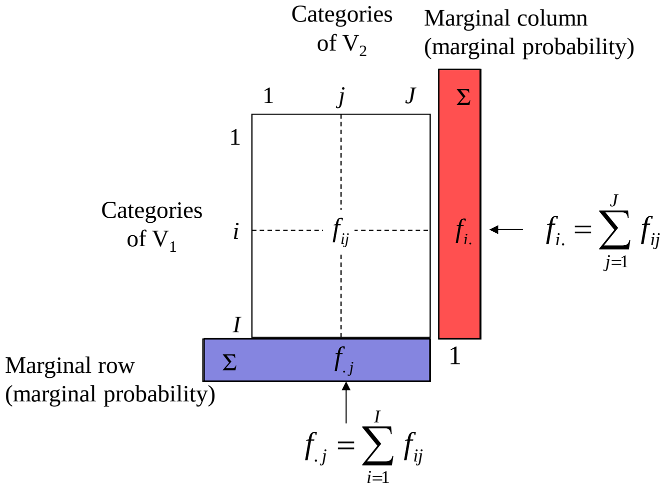
n individuals and 2 qualitative variables



Distribution of the n individual in the $I \times J$ boxes

From contingency tables to probability tables

$$f_{ij} = \frac{x_{ij}}{n}$$



Link between V1 and V2: Deviation of the observed data from the independence model

Correspondence Analysis (CA)

Links and independence between qualitative variables

Independence model:

Independent events: $P(A \text{ and } B) = P(A) \times P(B)$

Independent qualitative variables: $\forall i,\ \forall j,\ f_{ij} = f_{i.} \times f_{.j}$
$\Rightarrow$ Joint probability = product of marginal probabilities

Another way to write it: $\frac{f_{ij}}{f_{i.}} = f_{.j}$ $\qquad \frac{f_{ij}}{f_{.j}} = f_{i.}$
$\Rightarrow$ Conditional probability = marginal probability

Links and independence between qualitative variables

Deviation of observed data (f_{ij}) from independence model ($f_{i.}\,f_{.j}$)

1. Significance of the link/deviation: χ^2 test

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\text{obs. num.} - \text{theor. num.})^2}{\text{theor. num}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n\,f_{ij} - n\,f_{i.}\,f_{.j})^2}{n\,f_{i.}\,f_{.j}}$$

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} n\,\frac{(\text{observed probability} - \text{theoretical probability})^2}{\text{theoretical probability}} = n\,\Phi^2$$

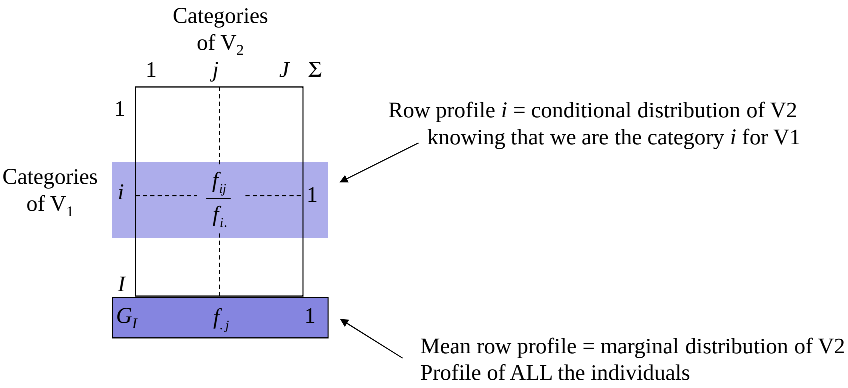
2. Strength of the link = Φ^2 = deviation of observed probabilities from theoretical ones

3. Type of link = connections between certain categories

CA works with the table of probabilities
says nothing about significance
visualizes the sorts of links there are between the variables

How does CA detect a deviation from independence?

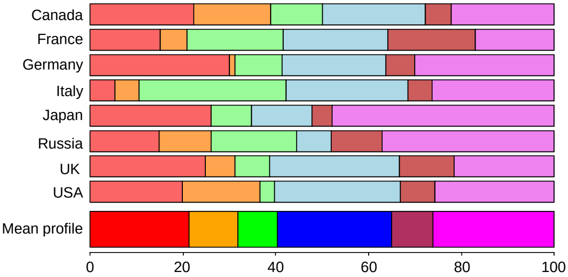
Analysis by row: $\dfrac{f_{ij}}{f_{i.}} = f_{.j}$



Deviation from independence using a multidimensional approach

Comparing the row profile to the mean profile

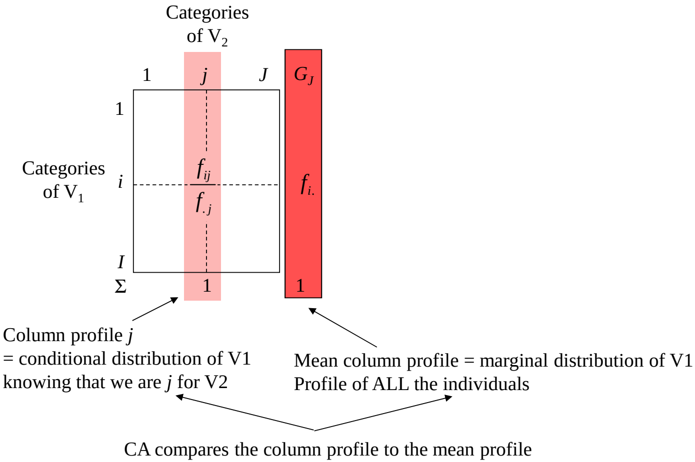
	Chemistry	Eco	Lit.	Medicine	Peace	Physics	Sum
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
Germany	30.0	1.2	10.0	22.5	6.2	30.0	100
Italy	5.3	5.3	31.6	26.3	5.3	26.3	100
Japan	26.1	0.0	8.7	13.0	4.3	47.8	100
Russia	14.8	11.1	18.5	7.4	11.1	37.0	100
UK	24.7	6.5	7.5	28.0	11.8	21.5	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Mean profile	21.2	10.5	8.6	24.6	8.9	26.1	100



Do Italians win particular categories of Nobel prize?

How does CA detect a deviation from independence?

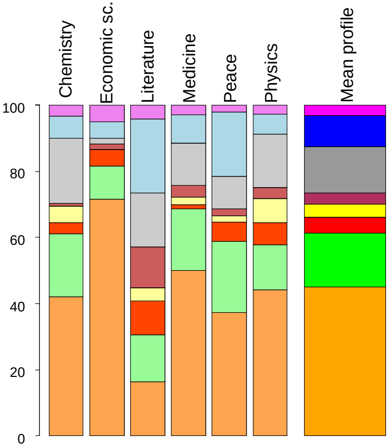
Analysis by column: $\frac{f_{ij}}{f_{.j}} = f_{i.}$



Deviation from independence using a multidimensional approach

Comparing the column profile to the mean profile

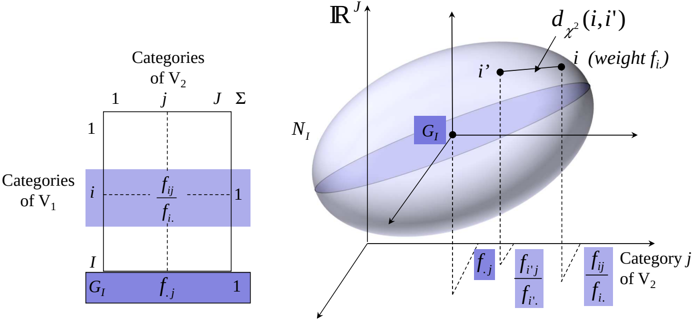
	Chem	Eco	Lit	Med	Peace	Phys	Mean profile
Canada	3.3	5.0	4.1	2.9	2.0	2.7	3.2
France	6.6	5.0	22.4	8.6	19.6	6.0	9.3
Germany	19.8	1.7	16.3	12.9	9.8	16.1	14.0
Italy	0.8	1.7	12.2	3.6	2.0	3.4	3.3
Japan	5.0	0.0	4.1	2.1	2.0	7.4	4.0
Russia	3.3	5.0	10.2	1.4	5.9	6.7	4.7
UK	19.0	10.0	14.3	18.6	21.6	13.4	16.3
USA	42.1	71.7	16.3	50.0	37.3	44.3	45.1
Sum	100	100	100	100	100	100	100



Is the country distribution for literature prizes the same as the country distribution for total prizes?

Correspondence Analysis (CA)

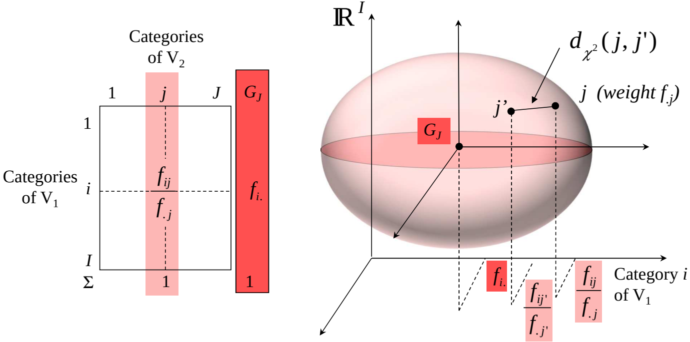
The cloud of row profiles



Distance between two profiles: $d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$

Distance to the mean profile G_I: $d^2_{\chi^2}(i,G_I) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$

The cloud of column profiles



Distance between two profiles: $d^2_{\chi^2}(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$

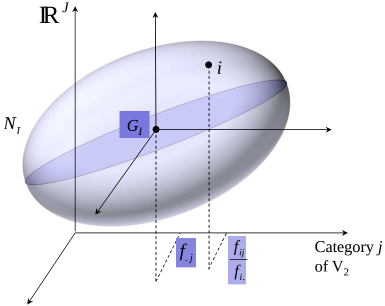
Distance to the mean profile G_J: $d^2_{\chi^2}(j,G_J) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$

What happens when there is independence?

$$\text{For all } i,\ \frac{f_{ij}}{f_{i.}} = f_{.j}$$

$\Rightarrow$ the profiles are the same as the mean profile

$\Rightarrow$ N_I becomes just G_I (the cloud has zero inertia)



Same for the columns: for all j, $\frac{f_{ij}}{f_{.j}} = f_{i.}$

Deviation from independence, and inertia

The further the data is from independence, the more the profiles spread from the origin

$$\begin{aligned} Inertia(N_I/G_I) &= \sum_{i=1}^{I} Inertia(i/G_I) = \sum_{i=1}^{I} f_{i.} d^2_{\chi^2}(i, G_I) \\ &= \sum_{i=1}^{I} f_{i.} \left(\sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \right) \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} = \frac{\chi^2}{n} = \phi^2 \end{aligned}$$

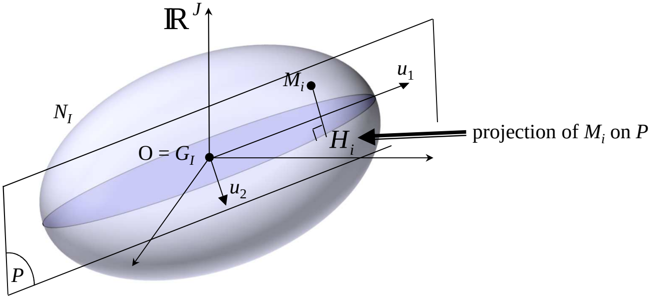
ϕ^2 measures the strength of the link

Studying the inertia of N_I turns out to be the same as studying deviation from independence

Same for N_J: $Inertia(N_J/G_J) = Inertia(N_I/G_I)$ (duality)

Visualizing the row (or column) cloud

Decompose the inertia of N_I using factor analysis
Project N_I onto a sequence of orthogonal axes with maximal inertia

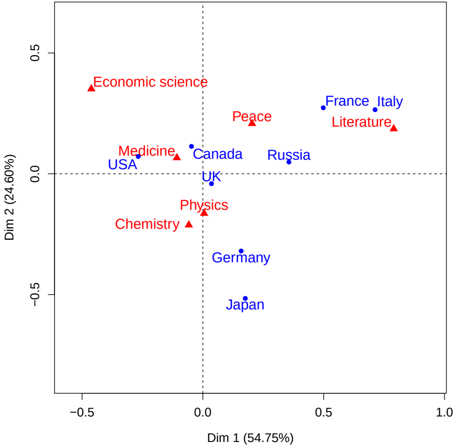


Find P such that $\sum_{i=1}^{I} f_{i.} \, (OH_i)^2$ is maximal

u_1 axis of maximal inertia
u_2 axis of maximal inertia such that $u_2 \perp u_1$

Inertia associated with the s-th axis: $\sum_{i=1}^{I} f_{i.} \, (OH_i^s)^2 = \lambda_s$

How to interpret? Our example:



1st axis: contrasts science - other categories

2nd axis: contrasts physics/chemistry - economic sciences

	Chemistry	Economic science	Literature	Medicine	Peace	Physics	Sum
Italy	5.26	5.26	31.58	26.32	5.26	26.32	100
UK	24.73	6.45	7.53	27.96	11.83	21.51	100
Row margin	**21.23**	**10.53**	**8.60**	**24.56**	**8.95**	**26.14**	**100**

Correspondence Analysis (CA)

Percentage of inertia

1. Quality of representation of N_I on the sth axis

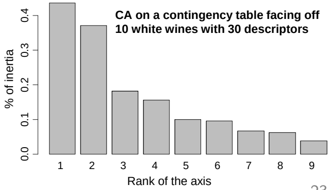
$$\frac{\text{Projected inertia of } N_I \text{ on } u_s}{\text{Total inertia of } N_I} = \frac{\sum_{i=1}^{I} f_{i.} \left(OH_i^s\right)^2}{\sum_{i=1}^{I} f_{i.} \left(OM_i\right)^2} = \frac{\lambda_s}{\sum_{i=1}^{K} \lambda_k}$$

	Inertia	Inertia (%)
F1	0.0833	54.75
F2	0.0374	24.60
F3	0.0217	14.23
F4	0.0079	5.18
F5	0.0019	1.25
Sum	0.1522	100

$\Rightarrow$ Deviation from independence well-summarized by the first two axes (79 %)

2. Projected inertia can be summed across axes (because orthogonal)
$\sum_{k=1}^{K} \lambda_k = \text{Inertia } (N_I) = \Phi^2$
Here $n\Phi^2 = 570 \times 0.1522 = \chi^2 = 86.75 \quad \Rightarrow \quad \text{p-value} = 2.77 \times 10^{-6}$

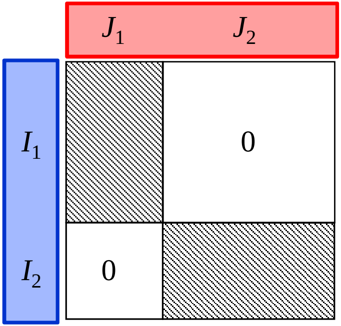
3. How the inertia decreases can help choose number of axes to keep

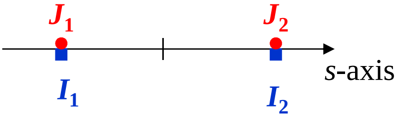


Inertia (= eigenvalues)

In CA: $0 \leq \lambda_s \leq 1$ In PCA (normalized): $1 \leq \lambda_1$

What structure does an eigenvalue of 1 correspond to?





⇒ Partition into two classes of rows and columns
Exclusive associations between classes

Inertia (= eigenvalues)

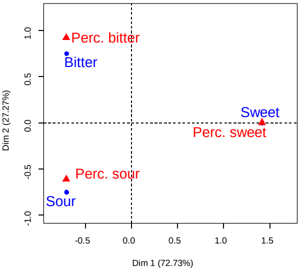
Data: recognizing three flavors (sweet, sour, bitter)
For each flavor, we asked 10 people to try to recognize taste of a sample

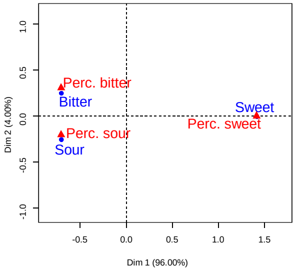
	Perc. sweet	Perc. sour	Perc. bitter
Sweet	10	0	0
Sour	0	9	1
Bitter	0	3	7

	Perc. sweet	Perc. sour	Perc. bitter
Sweet	10	0	0
Sour	0	7	3
Bitter	0	5	5

CA	Eigenvalue	%
Axis 1	1	72.727
Axis 2	0.375	27.273
Sum	1.375	100

CA	Eigenvalue	%
Axis 1	1	96
Axis 2	0.042	4
Sum	1.042	100





Inertia (= eigenvalues)

	Chemistery	Economic science	Literature	Mathematics	Medicine	Peace	Physics
Canada	4	3	2	1	4	1	4
France	8	3	11	11	12	10	9
Germany	24	1	8	1	18	5	24
Italy	1	1	6	1	5	1	5
Japan	6	0	2	3	3	1	11
Russia	4	3	5	9	2	3	10
UK	23	6	7	4	26	11	20
USA	51	43	8	13	70	19	66

	Inertia	Inertia (%)
F1	0.0833	54.75
F2	0.0374	24.60
F3	0.0217	14.23
F4	0.0079	5.18
F5	0.0019	1.25
Sum	0.1522	100

$\lambda_1 = 0.0833 \ll 1 \Rightarrow$ i.e., we are far from an exclusive association between certain rows and columns

$\Phi^2 = 0.1522 \ll 5 \Rightarrow$ we are far from a perfect link. i.e., far from exclusive association between categories of the two variables

Correspondence Analysis (CA)

Simultaneous representation of rows and columns

Transition formulas = barycentric properties

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \underbrace{\sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} G_s(j)}$$

$F_s(i)$: coord. of row i on the s-th axis
$\frac{f_{ij}}{f_{i.}}$: j-th element of profile i
$G_s(j)$: coord. of column j on the s-th axis
λ_s: inertia associated with s-th axis (in CA, $\lambda_s \leq 1$)

Along the s-th axis, we calculate the barycenter of each column, with column j given a weight $f_{ij}/f_{i.}$

The smaller the λ_s, the further the barycenter from the origin: $1/\sqrt{\lambda_s} \geq 1$

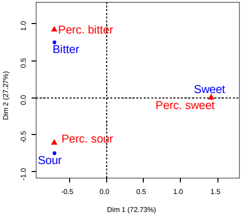
$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$$

Simultaneous representation and inertia

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$$

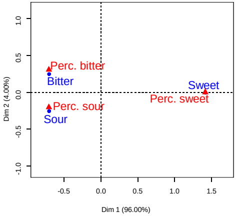
	Perc. sweet	Perc. sour	Perc. bitter
Sweet	10	0	0
Sour	0	9	1
Bitter	0	3	7

CA	Eigenvalue	%
Axis 1	1	72.727
Axis 2	0.375	27.273
Sum	1.375	100



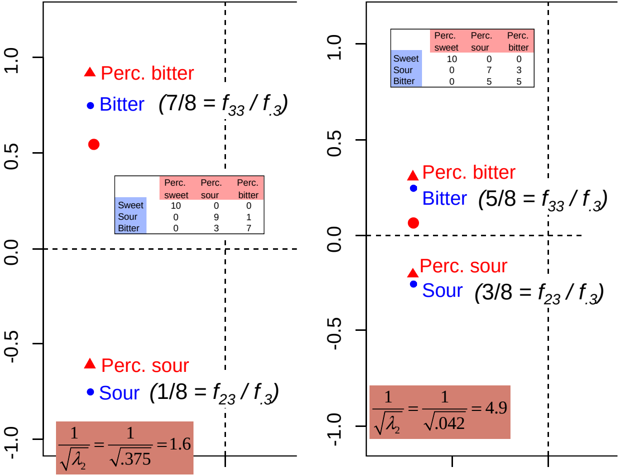
	Perc. sweet	Perc. sour	Perc. bitter
Sweet	10	0	0
Sour	0	7	3
Bitter	0	5	5

CA	Eigenvalue	%
Axis 1	1	96
Axis 2	0.042	4
Sum	1.042	100

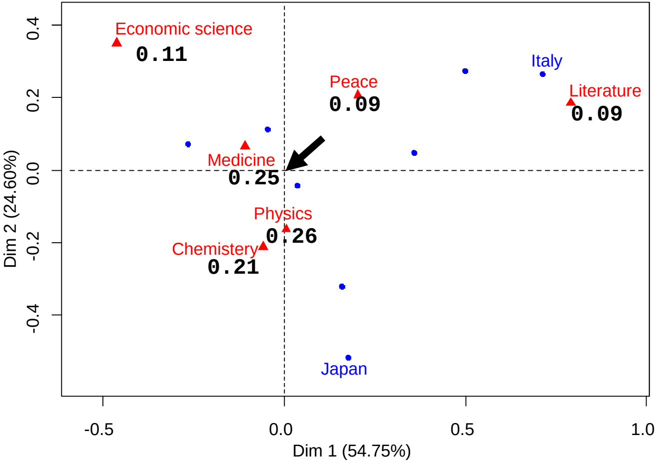


Simultaneous representation and inertia

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$$

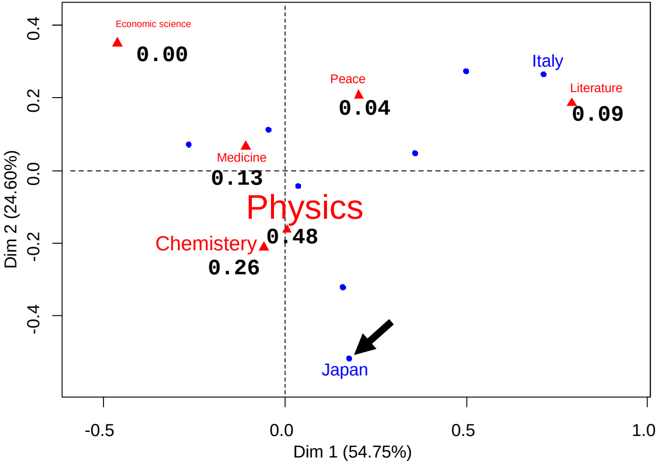


The barycentric property



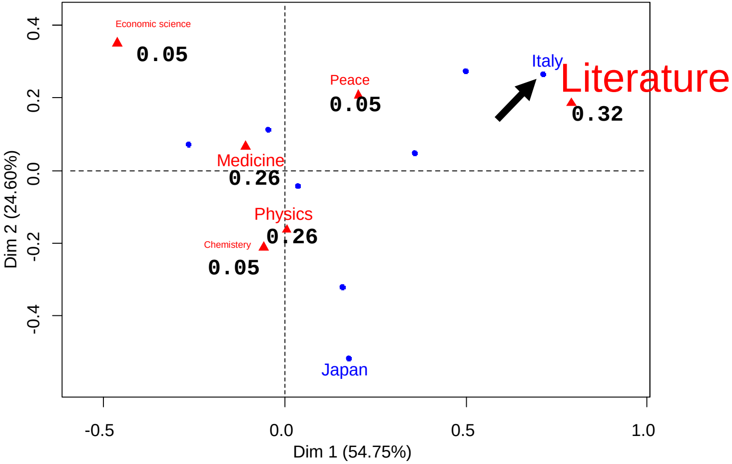
	Chemistery	Economic	Literature	Medicine	Peace	Physics
Italy	5.26	5.26	31.58	26.32	5.26	26.32
Japan	26.09	0.00	8.70	13.04	4.35	47.83
Mean profile	21.23	10.53	8.60	24.56	8.95	26.14

The barycentric property



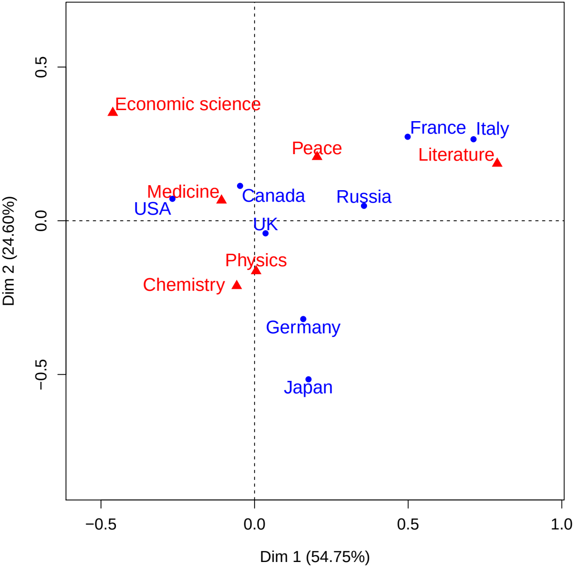
	Chemistery	Economic	Literature	Medicine	Peace	Physics
Italy	5.26	5.26	31.58	26.32	5.26	26.32
Japan	26.09	0.00	8.70	13.04	4.35	47.83
Mean profile	21.23	10.53	8.60	24.56	8.95	26.14

The barycentric property



	Chemistery	Economic	Literature	Medicine	Peace	Physics
Italy	5.26	5.26	31.58	26.32	5.26	26.32
Japan	26.09	0.00	8.70	13.04	4.35	47.83
Mean profile	21.23	10.53	8.60	24.56	8.95	26.14

The barycentric property

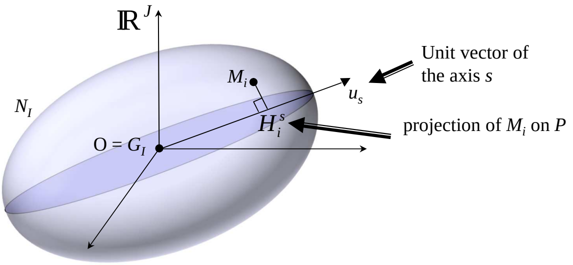


Correspondence Analysis (CA)

Interpretation aid: quality of the representation

An indicator of the quality of representation of a point (or cloud):

$$\frac{\text{projected inertia of } M_i \text{ on } u_s}{\text{total inertia of } M_i} = \frac{f_{i.}(OH_i^s)^2}{f_{i.}(OM_i)^2} = \cos^2(\overrightarrow{OM_i}, u_s)$$

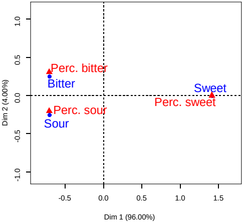


This indicator show how much the deviation of a profile from the mean profile is shown in an axis or plane

The quality of representation: example

	Perc. sweet	Perc. sour	Perc. bitter
Sweet	10	0	0
Sour	0	7	3
Bitter	0	5	5

CA	Eigenvalue	%
Axis 1	1	96
Axis 2	0.042	4
Sum	1.042	100



Quality of representation ($\cos^2$)

	Axis1	Axis2
Sweet	1.000	0.000
Sour	0.889	0.111
Bitter	0.889	0.111
Perc. sweet	1.000	0.000
Perc. sour	0.923	0.077
Perc. bitter	0.842	0.152

$\Rightarrow$ Interpretation of the graph based on extreme points with good quality representations

Interpretation aid: contributions

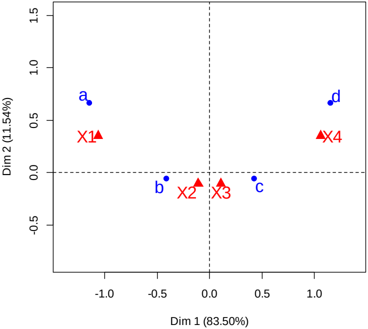
Absolute indicator: projected inertia of M_i on $u_s = f_{i.}(OH_i^s)^2$

Relative indicator: $\dfrac{\text{proj. inertia of } M_i \text{ on } u_s}{\text{inertia of axis } s} = \dfrac{f_{i.}(OH_i^s)^2}{\lambda_s}$

- We can sum the contributions of several elements
- This shows how much we can consider that an axis is due to one or several elements
- Practical compromise between distance to the origin, and weights
- Useful in big tables for selecting a subset of elements when starting interpretation (jointly with the quality of representation)

Contributions: example

	X1	X2	X3	X4
a	1	1	0	0
b	5	10	10	0
c	0	10	10	5
d	0	0	1	1



	Inertia	%
Axis 1	0.258	83.501
Axis 2	0.036	11.538
Axis 3	0.015	4.96

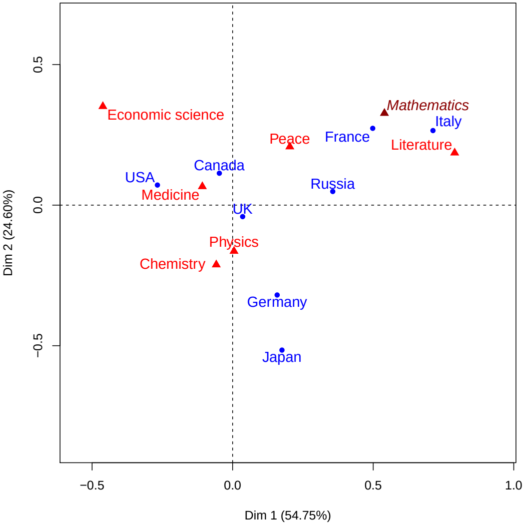
	Axis 1	Axis 2
a	18.879	46.296
b	31.121	3.704
c	31.121	3.704
d	18.879	46.296
Σ	100	100

$\Rightarrow$ The extreme points are not necessarily the ones that contribute most to axis construction

Supplementary information

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$$

Mathematics is on the French and Russian side, also the side of literature and peace, but opposite the sciences



Distributional equivalence

Distributional equivalence: if rows with the same values are grouped, the CA results are totally equivalent (same for columns)

Application to analysis of texts:
Thanks to distributional equivalence, if two (or more) words are used in the same circumstances, their coordinates will be close together, so doing the analysis with both, or just one, is entirely equivalent
⇒ very useful (to group singular and plural versions of words, verb conjugations, etc.)

Maximum number of axes, and Cramer's V

Point cloud for rows: I points in a J-dimensional space

$$\left.\begin{array}{r} J \text{ dim. but 1 constraint (profiles)} \Rightarrow S \leq J-1 \\ I \text{ points in at most } I-1 \text{ dim.} \Rightarrow S \leq I-1 \end{array}\right\} S \leq \min(I-1, J-1)$$

$$\Longrightarrow \Phi^2 = \sum_{k=1}^{\min(I-1,J-1)} \lambda_k \leq \min(I-1, J-1)$$

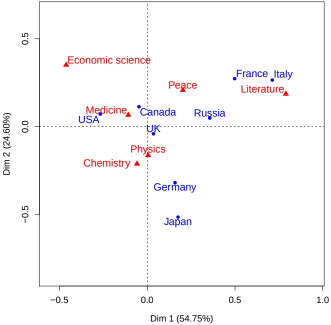
leading to the idea of a bounded indicator of the link between 2 variables:

$$\text{Cramer's V} = \frac{\Phi^2}{\min(I-1, J-1)} \in [0;1]$$

	Nobel Prize	Three tastes	Three tastes
Cramer's V	$0.1522/5 = 0.03044$	$1.375/2 = 0.6875$	$1.042/2 = 0.521$

Conclusions for our example

	Chemistery	Economic science	Literature	Mathematics	Medicine	Peace	Physics
Canada	4	3	2	1	4	1	4
France	8	3	11	11	12	10	9
Germany	24	1	8	1	18	5	24
Italy	1	1	6	1	5	1	5
Japan	6	0	2	3	3	1	11
Russia	4	3	5	9	2	3	10
UK	23	6	7	4	26	11	20
USA	51	43	8	13	70	19	66



CA gives a good summary visual of the deviation from independence, which helps to understand the data table (and especially, big data tables)

As for this data:

- Most of the deviation from independence appears in the separation of Science vs The rest. And to a lesser extent: physics/chemistry vs economics
- The position of each country shows its strengths in getting various prizes

Conclusion

To study the link between qualitative variables, we build a contingency table

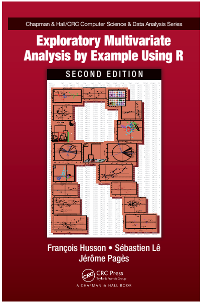
The link is found in the difference between this table, and what it would look like if there was independence

Correspondence analysis:

- build a point cloud of rows (and of columns) whose total inertia measures the strength of the deviation from independence
- break down this total inertia into a sequence of axes of decreasing importance, each representing some feature of the link between the variables
- visually represent the rows and columns in such a way that their position on the graph reflects their participation in the deviation from independence

Bibliography

For more information in the same vein, have a look at this video:



Husson F., Lê S. & Pagès J. (2017)
Exploratory Multivariate Analysis by Example Using R
2nd edition, 230 p., CRC/Press.