EI5IS102 Traitement de l'Information

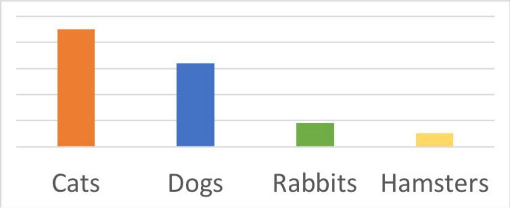
Lecture 2: Correspondence Analysis

Charles Brazier
Postdoctoral researcher
Université de Bordeaux, CNRS, Bordeaux INP, LaBRI
France

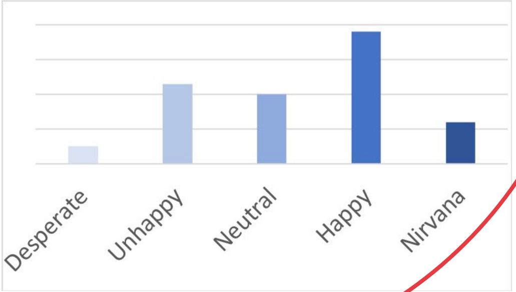
Types of data

Qualitative (categories)

Nominal:

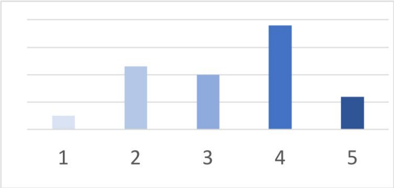


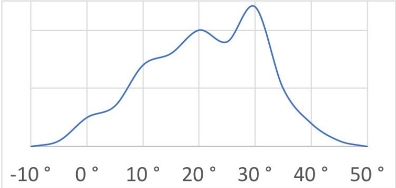
Ordered:



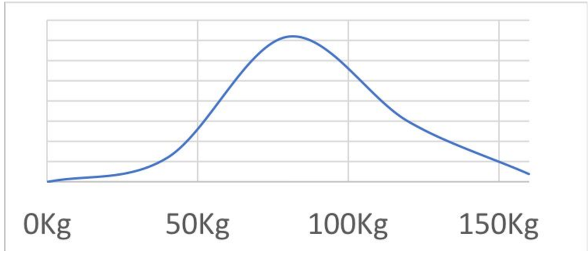
Quantitative (numerical values)

Interval (discrete or continuous):





Ratio:



Qualitative data

Qualitative data: non-numerical data that represent descriptive information
Ex: describing an employee (supportive, directive, etc.), color of cars (red, blue, etc.)

Categorical data: type of qualitative data that is organized into distinct categories
Ex: describing an employee by level (junior/senior)

Contingency table: cross-table that summarizes information of two categorical data.
Ex: rank vs smoking intensity

1st categorical data: working rank

Rank	Smoking intensity				
	none	light	medium	heavy	Total
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
Total	61	45	62	25	193

2nd categorical data: smoking intensity

Greenacre, 1984

Contingency table

Contingency table: rank vs smoking intensity

- 3 junior managers have a light smoking intensity
- 3 light smokers are junior managers

→ symmetrical role of rows/columns

Rank	Smoking intensity				
	none	light	medium	heavy	Total
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
Total	61	45	62	25	193

row sums have meaning

column sums have meaning

Contingency table

Correspondence Analysis (CA):
Applied to contingency table

row profile

Rank	Smoking intensity				
	none	light	medium	heavy	Total
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
Total	61	45	62	25	193

column profile

Answered questions:

1. Which **row profiles** are close/distant
2. Which **column profiles** are close/distant
3. Find strong associations between row and column classes

Correspondence Analysis

Notations

n: number of instances

V_1: qualitative variable of size I

V_2: qualitative variable of size J

x_{ij} : nb of instances possessing modality i of V_1 and modality j of V_2

Contingency table: $X : (x_{ij})_{1 \leq i \leq I, 1 \leq j \leq J}$

Margins:

- column margin: $x_{i\bullet} = \sum_{j=1}^{J} x_{ij}$
- raw margin: $x_{\bullet j} = \sum_{i=1}^{I} x_{ij}$
- overall sum: $x_{\bullet\bullet} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij} = n$

11 instances have modality 1 ("Senior manager") of V_1 ("Rank")

Rank	Smoking intensity				
	none	light	medium	heavy	x_i•
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
x_•j	61	45	62	25	n=193

Probability table

Probability table: $f_{ij} = \frac{x_{ij}}{n}$

Marginal probabilities:

- marginal column probability: $f_{i\bullet} = \sum_{j=1}^{J} f_{ij}$
- marginal row probability: $f_{\bullet j} = \sum_{i=1}^{I} f_{ij}$
- overall sum: $f_{\bullet\bullet} = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} = 1$

5% of the instances are senior employees with light smoking intensity $P(V_1\text{=i}, V_2\text{=j})$

Rank	Smoking intensity				
	none	light	medium	heavy	f_i•
Senior manager	0,02	0,01	0,02	0,01	0,06
Junior manager	0,02	0,02	0,04	0,02	0,09
Senior employee	0,13	0,05	0,06	0,02	0,26
Junior employee	0,09	0,12	0,17	0,07	0,46
Secretary	0,05	0,03	0,04	0,01	0,13
f_•j	0,32	0,23	0,32	0,13	1,00

13% of the instances are secretaries $P(V_1\text{=i})$

Link between V_1 and V_2: deviation of the observed data from the **independence model**

Independence model

Probabilities in case of independence:

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

observed probability

Independent qualitative variables: $\forall i, \forall j, f_{ij} = f_{i\bullet} f_{\bullet j}$ theoretical probability

→ joint probability = product of marginal probabilities

Another way to write it: $$\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \quad \text{and} \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$$

→ conditional probability = marginal probability

Independence model

Deviation of observed data from independence model:
If V_1 and V_2 are independent, **observed** and **theoretical probabilities** should be similar:

$$\forall i, \forall j, f_{ij} \approx f_{i\bullet} f_{\bullet j} = \hat{f}_{ij}$$

So **observed** and **theoretical data** should be similar:

$$\forall i, \forall j, x_{ij} = n f_{ij} \approx n f_{i\bullet} f_{\bullet j} = \hat{x}_{ij}$$

Residual = diff between observed and theoretical data = **deviation from independence**

$$\forall i, \forall j, r_{ij} = x_{ij} - \hat{x}_{ij}$$

Independence model

Observed data X :

Rank	Smoking intensity				
	none	light	medium	heavy	x_i•
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
x_•j	61	45	62	25	n=193

Theoretical data $\hat{X}$:

Rank	Smoking intensity				
	none	light	medium	heavy	x̂_i•
Senior manager	3,48	2,56	3,53	1,42	11,00
Junior manager	5,69	4,20	5,78	2,33	18,00
Senior employee	16,12	11,89	16,38	6,61	51,00
Junior employee	27,81	20,52	28,27	11,40	88,00
Secretary	7,90	5,83	8,03	3,24	25,00
x̂_•j	61,00	45,00	62,00	25,00	193,00

Independence model

Observed data X :

Rank	Smoking intensity				
	none	light	medium	heavy	x_i•
Senior manager	4	2	3	2	11
Junior manager	4	3	7	4	18
Senior employee	25	10	12	4	51
Junior employee	18	24	33	13	88
Secretary	10	6	7	2	25
x_•j	61	45	62	25	n=193

Theoretical data $\hat{X}$:

Rank	Smoking intensity				
	none	light	medium	heavy	x̂_i•
Senior manager	3,48	2,56	3,53	1,42	11,00
Junior manager	5,69	4,20	5,78	2,33	18,00
Senior employee	16,12	11,89	16,38	6,61	51,00
Junior employee	27,81	20,52	28,27	11,40	88,00
Secretary	7,90	5,83	8,03	3,24	25,00
x̂_•j	61,00	45,00	62,00	25,00	193,00

Significant difference ?

Independence test: χ^2 test

Objective: to determine if the difference between observed and theoretical data is significant

1. Hypothesis: "the two variables V_1 and V_2 are independent"

2. Compute a distance between observed and theoretical data → $\boldsymbol{\chi^2_{obs}}$ distance

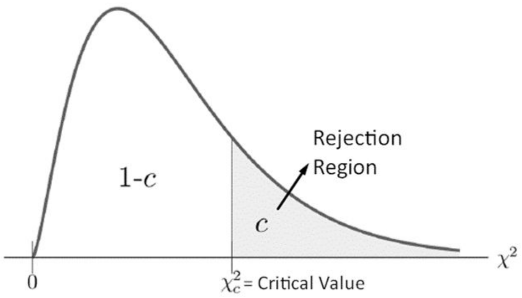
3. From χ^2_{obs}, compute a p-value which gives the probability of obtaining the observed data under independence hypothesis

4. If p-value is low (<5%), we reject the hypothesis → variables are correlated
If p-value is high (≥5%), we accept the hypothesis → variables are independent

χ^2 test in practice

In practice, the p-value is fixed (generally p-value = 0.05)

Under the independence hypothesis, the statistic χ^2_{obs} follows a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom.



χ^2 test in practice:

1. Compute distance χ^2_{obs} between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$
2. Fix p-value to 0.05
3. Compute the degree of freedom: $df = (I-1)(J-1)$
4. From the χ^2 distribution table, determine $\chi^2_{critical}$
5. If $\chi^2_{obs} < \chi^2_{critical}$, we accept the hypothesis → variables are independent
   If $\chi^2_{obs} \geq \chi^2_{critical}$, we reject the hypothesis → variables are correlated

χ^2 distance

χ^2 test: distance between observed data and theoretical data

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

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

Strength of the link: Φ^2
Deviation of observed probabilities from theoretical ones

Type of the link (attraction/repulsion): Correspondence Analysis

- CA does not test dependence/independence
- CA helps to understand the deviation from independence
- CA enables the visualization the types of links between modalities of the two variables

χ^2 test in our example

In our example:

- $\chi^2_{obs} = 16.44$
- degree of freedom: $df = (I\text{-}1)(J\text{-}1) = 12$
- With a p-value of 5%, $\chi^2_{critical} = 21.026$
- $\chi^2_{obs} < \chi^2_{critical}$: we accept the hypothesis
- V_1 and V_2 are independent

chi-square distribution table

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	$\boxed{21.026}$	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

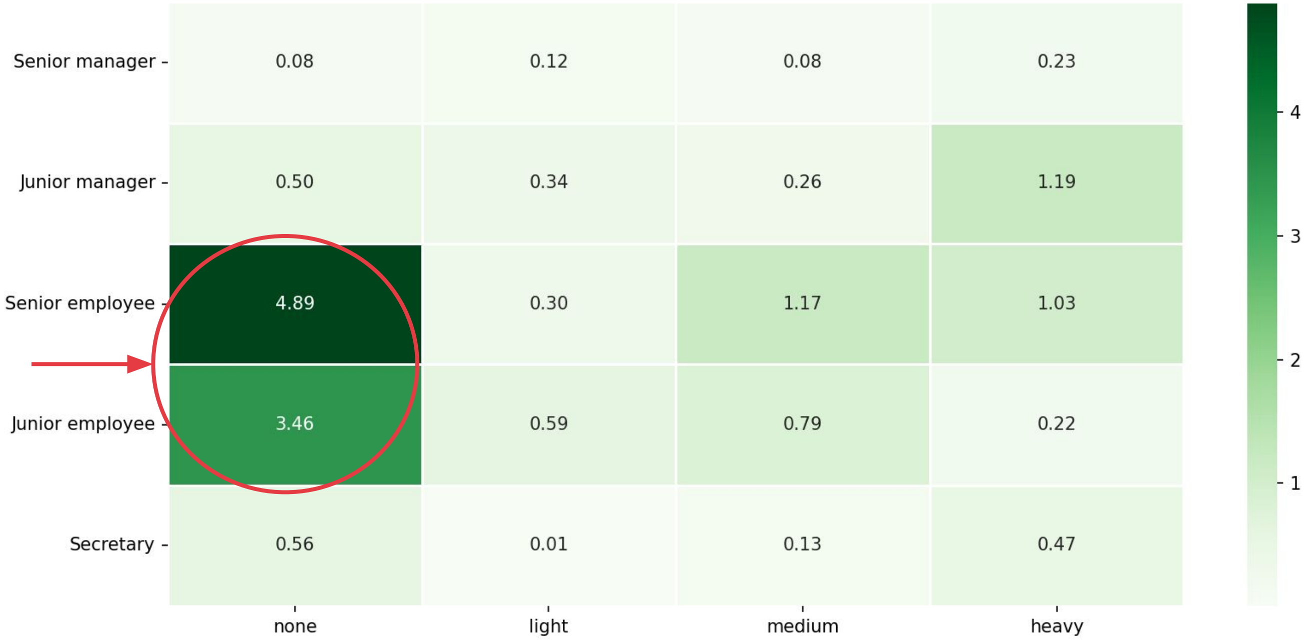
χ^2 test helps determine whether there is a dependence between variables V_1 and V_2.
It does not provide a description of the links between the variables.

→ **Correspondence Analysis**

χ^2 test

Contribution to χ^2 test:

Attraction?
Repulsion?



Correspondence Analysis

Deviation from independence:

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

marginal probability

conditional probability

Probability table $\boldsymbol{F}$:

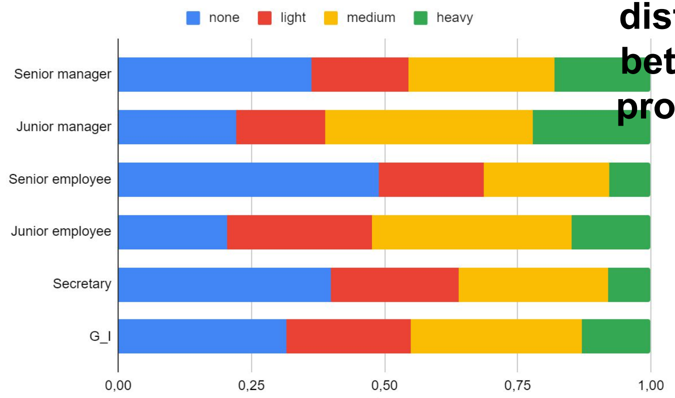
Rank	Smoking intensity				
	none	light	medium	heavy	f_i•
Senior manager	0,02	0,01	0,02	0,01	0,06
Junior manager	0,02	0,02	0,04	0,02	0,09
Senior employee	0,13	0,05	0,06	0,02	0,26
Junior employee	0,09	0,12	0,17	0,07	0,46
Secretary	0,05	0,03	0,04	0,01	0,13
f_•j	0,32	0,23	0,32	0,13	1,00

Row profiles $\boldsymbol{N_I}$:

Rank	Smoking intensity				
	none	light	medium	heavy	Σ
Senior manager	0,36	0,18	0,27	0,18	1,00
Junior manager	0,22	0,17	0,39	0,22	1,00
Senior employee	0,49	0,20	0,24	0,08	1,00
Junior employee	0,20	0,27	0,38	0,15	1,00
Secretary	0,40	0,24	0,28	0,08	1,00
G_I	0,32	0,23	0,32	0,13	1,00

row margin = barycenter of raw profiles with $f_{i\bullet}$ as weights

Correspondence Analysis



distance between profiles?

Deviation from independence:

Analysis by row: $\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$

marginal probability

conditional probability

Probability table $\boldsymbol{F}$:

Rank	Smoking intensity				
	none	light	medium	heavy	f_i•
Senior manager	0,02	0,01	0,02	0,01	0,06
Junior manager	0,02	0,02	0,04	0,02	0,09
Senior employee	0,13	0,05	0,06	0,02	0,26
Junior employee	0,09	0,12	0,17	0,07	0,46
Secretary	0,05	0,03	0,04	0,01	0,13
f_•j	0,32	0,23	0,32	0,13	1,00

Row profiles $\boldsymbol{N_I}$:

Rank	Smoking intensity				
	none	light	medium	heavy	Σ
Senior manager	0,36	0,18	0,27	0,18	1,00
Junior manager	0,22	0,17	0,39	0,22	1,00
Senior employee	0,49	0,20	0,24	0,08	1,00
Junior employee	0,20	0,27	0,38	0,15	1,00
Secretary	0,40	0,24	0,28	0,08	1,00
G_I	0,32	0,23	0,32	0,13	1,00

row margin = barycenter of raw profiles with $f_{i\bullet}$ as weights

BORDEAUX INP Enseirb-Matmeca

Correspondence Analysis

Deviation from independence:

Analysis by column: $\dfrac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$ ← marginal probability

conditional probability

Probability table $\boldsymbol{F}$:

Rank	Smoking intensity				
	none	light	medium	heavy	f_i•
Senior manager	0,02	0,01	0,02	0,01	0,06
Junior manager	0,02	0,02	0,04	0,02	0,09
Senior employee	0,13	0,05	0,06	0,02	0,26
Junior employee	0,09	0,12	0,17	0,07	0,46
Secretary	0,05	0,03	0,04	0,01	0,13
f_•j	0,32	0,23	0,32	0,13	1,00

Column profiles $\boldsymbol{N_J}$:

Rank	Smoking intensity				
	none	light	medium	heavy	G_J
Senior manager	0,07	0,04	0,05	0,08	0,06
Junior manager	0,07	0,07	0,11	0,16	0,09
Senior employee	0,41	0,22	0,19	0,16	0,26
Junior employee	0,30	0,53	0,53	0,52	0,46
Secretary	0,16	0,13	0,11	0,08	0,13
Σ	1,00	1,00	1,00	1,00	1,00

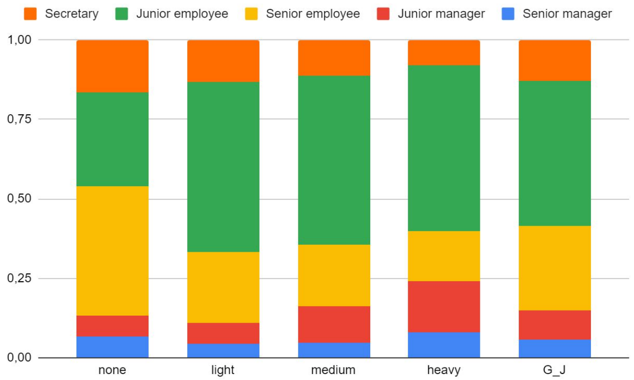
column margin = barycenter of column profiles with $f_{\bullet j}$ as weights

Correspondence Analysis

Deviation from independence:

Analysis by column: $\frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$ marginal probability

conditional probability



Probability table $\boldsymbol{F}$:

Rank	Smoking intensity				
	none	light	medium	heavy	f_i•
Senior manager	0,02	0,01	0,02	0,01	0,06
Junior manager	0,02	0,02	0,04	0,02	0,09
Senior employee	0,13	0,05	0,06	0,02	0,26
Junior employee	0,09	0,12	0,17	0,07	0,46
Secretary	0,05	0,03	0,04	0,01	0,13
f_•j	0,32	0,23	0,32	0,13	1,00

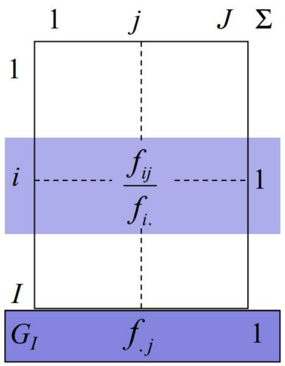
Column profiles $\boldsymbol{N_J}$:

Rank	Smoking intensity				
	none	light	medium	heavy	G_J
Senior manager	0,07	0,04	0,05	0,08	0,06
Junior manager	0,07	0,07	0,11	0,16	0,09
Senior employee	0,41	0,22	0,19	0,16	0,26
Junior employee	0,30	0,53	0,53	0,52	0,46
Secretary	0,16	0,13	0,11	0,08	0,13
Σ	1,00	1,00	1,00	1,00	1,00

column margin =
barycenter of column profiles

Distances

Row profiles:



Deviation from independence between variables
⇔ distance of the I row profiles to the mean profile G_I

Distance between two profiles:

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2$$

Distance to the mean profile G_I:

$$d^2_{\chi^2}(i,G_I) = \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2$$

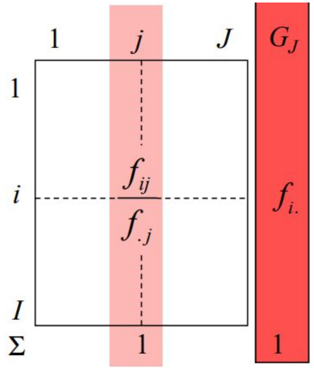
Scaling factor
(frequency of category)

deviation from
independence

In case of independence, row profiles are equal to the mean row profile
→ Data cloud N_I of the I row profiles becomes just G_I (zero inertia)

Distances

Column profiles:



Deviation from independence between variables
⇔ distance of the J row profiles to the mean profile G_J

Distance between two profiles:

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2$$

Distance to the mean profile G_J:

$$d^2_{\chi^2}(j,G_J) = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet} \right)^2$$

Scaling factor
(frequency of category)

deviation from
independence

In case of independence, column profiles are equal to the mean column profile
→ Data cloud N_J of the J column profiles becomes just G_J (zero inertia)

Inertia

Inertia of N_I and N_J:

mass of the row

squared distance

$$\text{Inertia}(N_I/G_I) = \sum_{i=1}^{I} \text{Inertia}(i/G_I) = \sum_{i=1}^{I} f_{i\bullet} d^2_{\chi^2}(i, G_I)$$

$$= \sum_{i=1}^{I} f_{i\bullet} \left(\sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 \right)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} = \frac{\chi^2}{n} = \Phi^2$$

$\mathbf{\Phi^2}$ measures the strength of the link
Studying the inertia of N_I = studying deviation from independence

Same for N_J: $\quad \text{Inertia}(N_J/G_J) = \text{Inertia}(N_I/G_I)$

→ **CA**: studying N_I and N_J to understand dependencies between variables!

Eigenvectors, eigenvalues

Similarly to PCA, we aim to find eigenvectors of N_I and N_J using the χ^2 metric

Admitted equivalence on the probability matrix $\boldsymbol{F}$:

$$\tilde{F} = D_I F D_J = D_I^{-1} N_I D_J = D_I N_J D_J^{-1}$$

with diagonal weighting matrices $\boldsymbol{D_I}$ and $\boldsymbol{D_J}$:

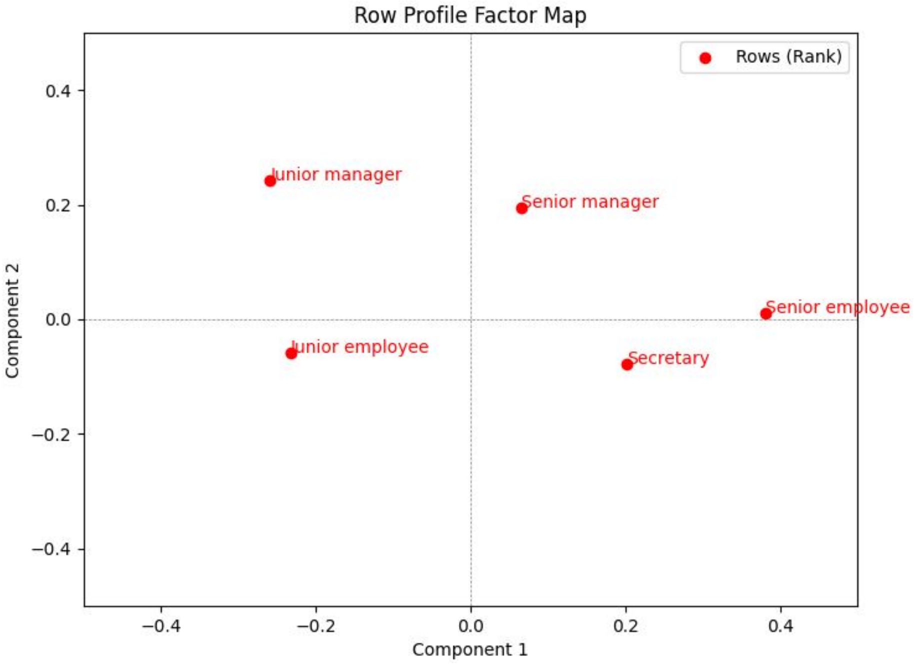
$$D_I = \begin{pmatrix} \frac{1}{\sqrt{f_{1\bullet}}} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \frac{1}{\sqrt{f_{I\bullet}}} \end{pmatrix} \qquad D_J = \begin{pmatrix} \frac{1}{\sqrt{f_{\bullet 1}}} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \frac{1}{\sqrt{f_{\bullet J}}} \end{pmatrix}$$

Eigenvectors, eigenvalues

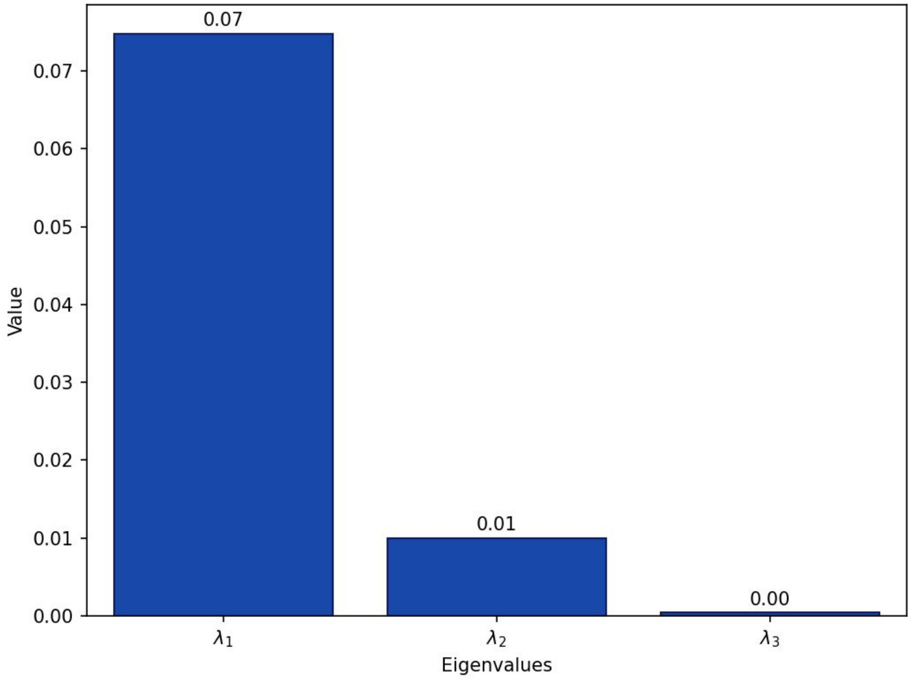
Row analysis:

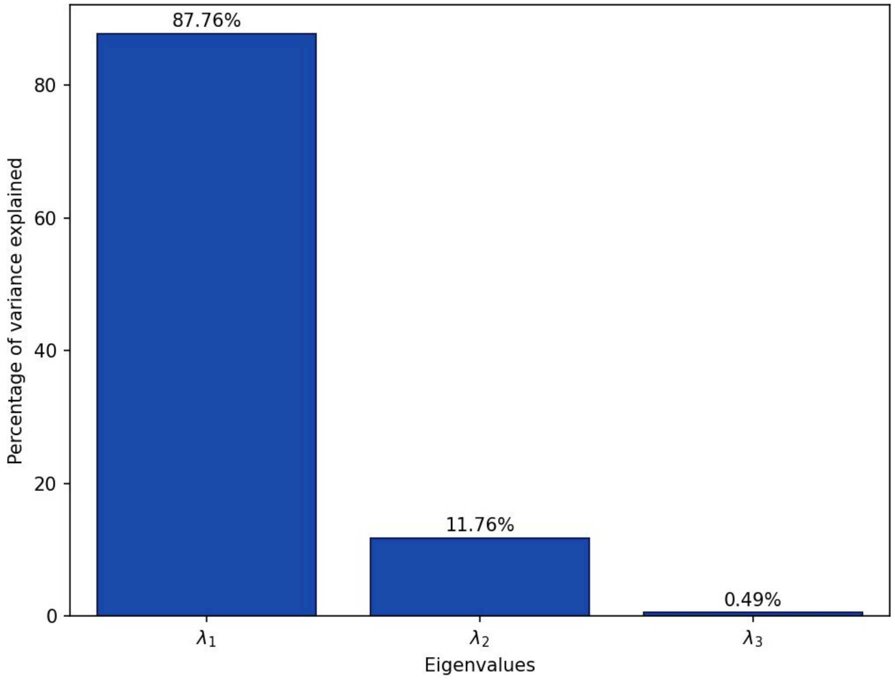
- Diagonalization: $\tilde{F}\tilde{F}^T = P_I \Lambda_I P_I^T$
- Eigenvectors: P_I
- Inertia: $\Lambda_I = \mathrm{diag}(1, \lambda_1^I, \cdots, \lambda_{I-1}^I)$
- Projection of row coordinates:

$$\alpha = D_I \tilde{F} P_I$$



Eigenvectors, eigenvalues



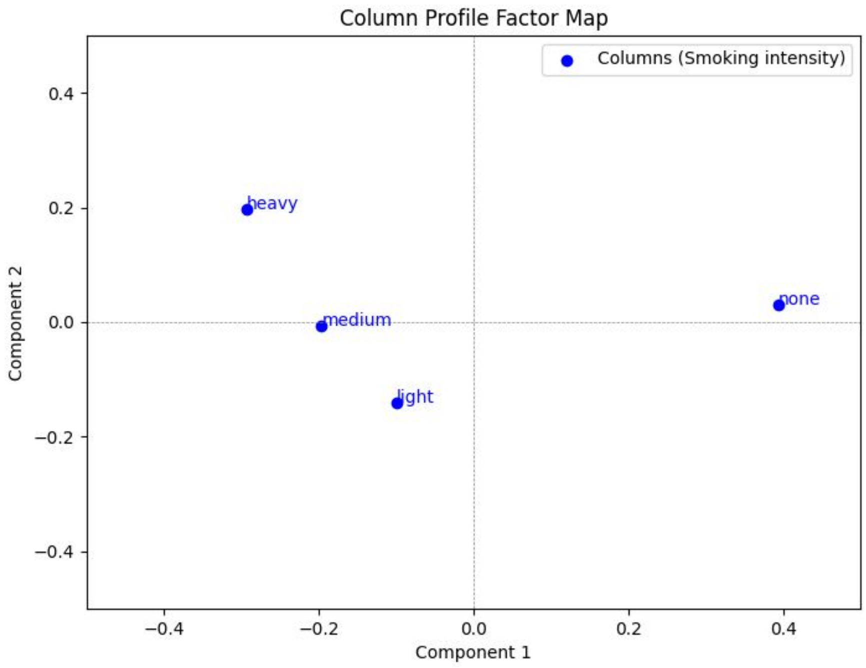


Eigenvectors, eigenvalues

Column analysis:

- Diagonalization: $\tilde{F}^T \tilde{F} = P_J \Lambda_J P_J^T$
- Eigenvectors: P_J
- Inertia: $\Lambda_J = \mathrm{diag}(1, \lambda_1^J, \cdots, \lambda_{J-1}^J)$
- Projection of column coordinates:

$$\beta = D_J \tilde{F}^T P_J$$



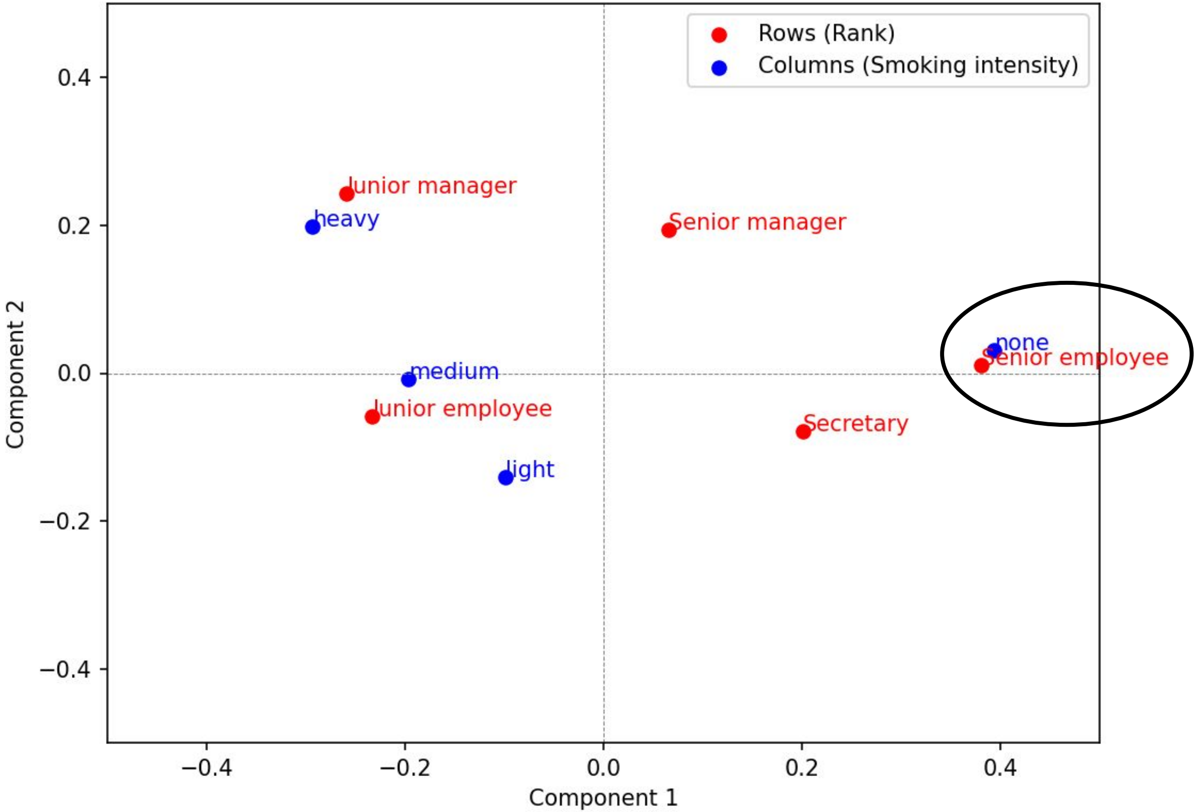
Superimposed representation of rows and columns

Duality between N_I and N_J : same data table from two different point of views

- Same total inertia: χ^2/n
- Inertia projected on the kth axis of N_I = Inertia projected on the kth axis of N_J = λ_k (admitted)
  - $\lambda_k^I = \lambda_k^J, \forall k = 1, \ldots, K = \min(I-1, J-1)$
  - $\lambda_k^I, \lambda_k^J = 0, \forall k > K$
- Relation between coordinates α_i^k and β_j^k of the row and column profiles projected on the eigenvectors (barycentric property, admitted):

$$\alpha_i^k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\bullet}} \beta_j^k \qquad \beta_j^k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\bullet j}} \alpha_i^k$$

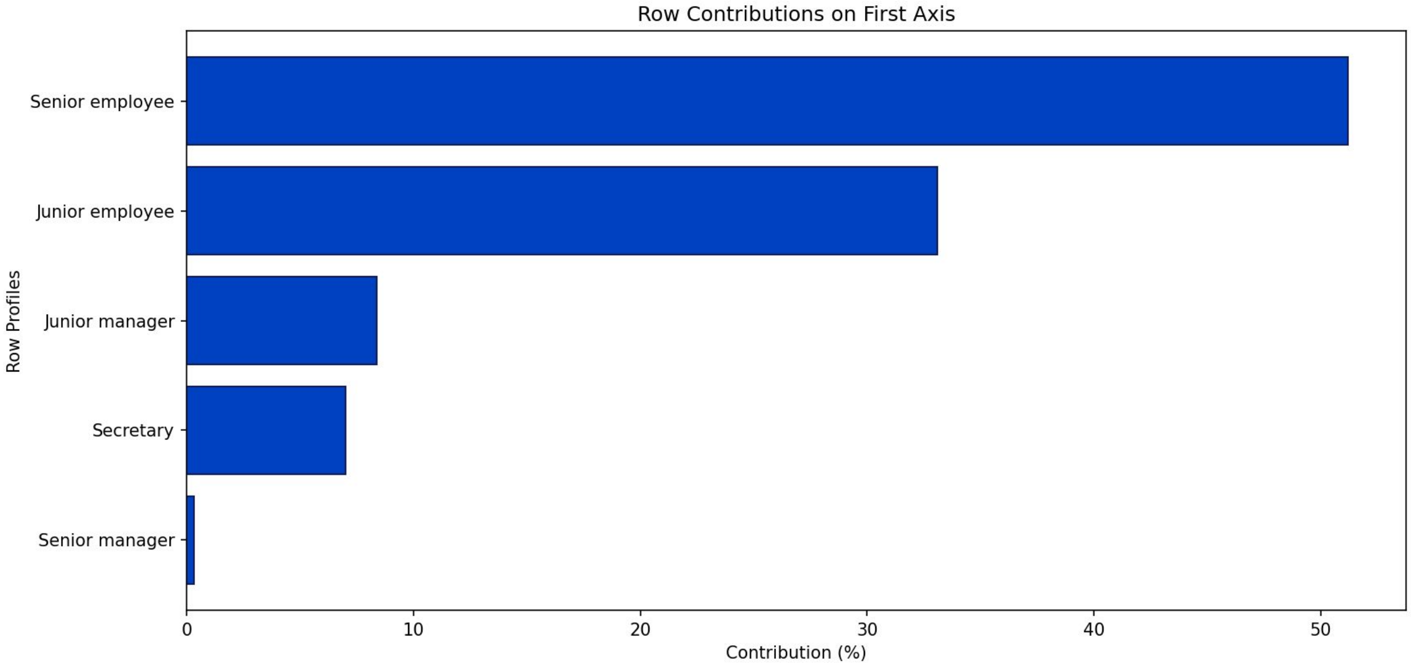
Superimposed representation of rows and columns



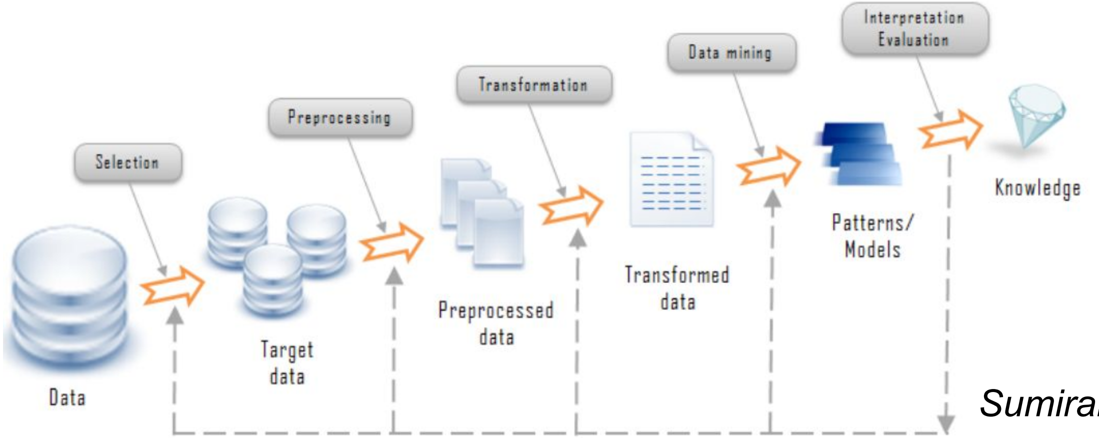
NB: Proximity between row dots and column dots is relevant if they are on the periphery of the cloud (deviations from independence)

Contributions

Contribution of each row or column profiles to each axis:
→ helps in interpreting the significance of each axis



The data analysis process



Sumiran, 2018

Data: statistics on employees in a company

Selection: two features (rank and smoking intensity)

Preprocessing: contingency table

Transformation: probability matrix F and also $\tilde{F}\tilde{F}^T$

Model: Eigenvectors, eigenvalues, projection

Knowledge: Correlation between features

Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA):
- generalization of CA
- Analysis of more than two categorical variables
- Attraction or repulsion between several variables

Dummy variable table ("tableau disjonctif complet") for n instances, J variables, and K total number of modalities:

$$K = K_1 + \ldots + K_J$$

$$T = (t_i^k)_{1 \leq i \leq n, 1 \leq k \leq K} \text{ where } t_i^k = \begin{cases} 1 & \text{if instance } i \text{ has modality } k \\ 0 & \text{otherwise} \end{cases}$$

Implies several **corrections**:
- Benzécri: correct eigenvalues
- Greenacre: correct variance explained

Multiple Correspondence Analysis

Example: dummy variable table

- $n=3$ instances
- $J=2$ variables
- $K=5$ modalities

	Gender	Eyes
Father	M	Hazel
Mother	F	Blue
Child	M	Green

➡

	Gender F	Gender M	Eyes B	Eyes G	Eyes H
Father	0	1	0	0	1
Mother	1	0	1	0	0
Child	0	1	0	0	1

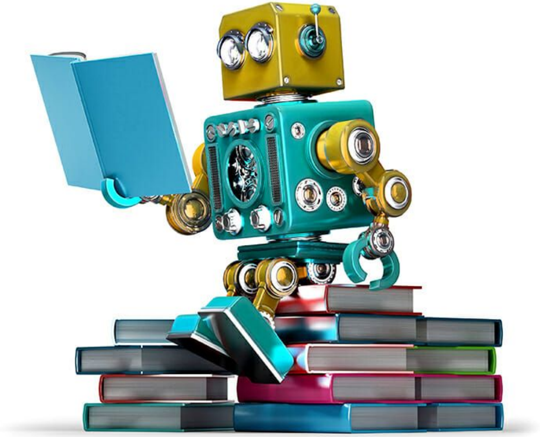
MCA: CA on the dummy variable table!

Next course

Introduction to Machine Learning

Two tasks:

1. Classification
2. Clustering



Questions?

Sources, images courtesy and acknowledgment:
N. Papadakis, R.Rakotomalala,
J. Dabounou, F. Husson

Charles Brazier
charles.brazier@u-bordeaux.fr