Data Valorization:
Correspondence Analysis

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Frequency Tables
- Analysis of Row Profiles
- Analysis of Column Profiles
- Conclusion
- Appendix

1 Introduction

A First Example of contingency table

- Relationship between hair color and eye color for 6800 German men. Is there an association between hair color and eye color?

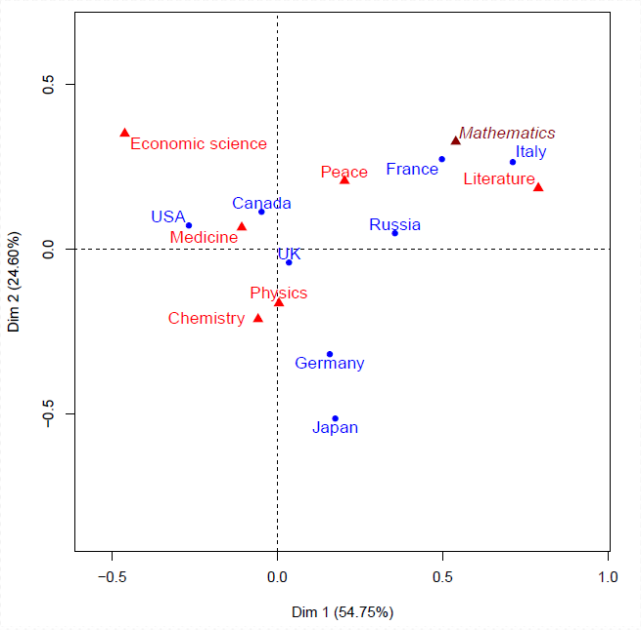
	eye color			
Hair Color	Brown	Grey/Green	Blue	Total
Brown	438	1387	807	2632
Black	288	746	189	1223
Fair	115	946	1768	2829
Red	16	53	47	116
Total	857	3132	2811	6800

Nobel Prize (including Fields medal)

- Relationship between Nobel prize category and country?

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

Nobel Prize: Correspondence Analysis



- Details later…

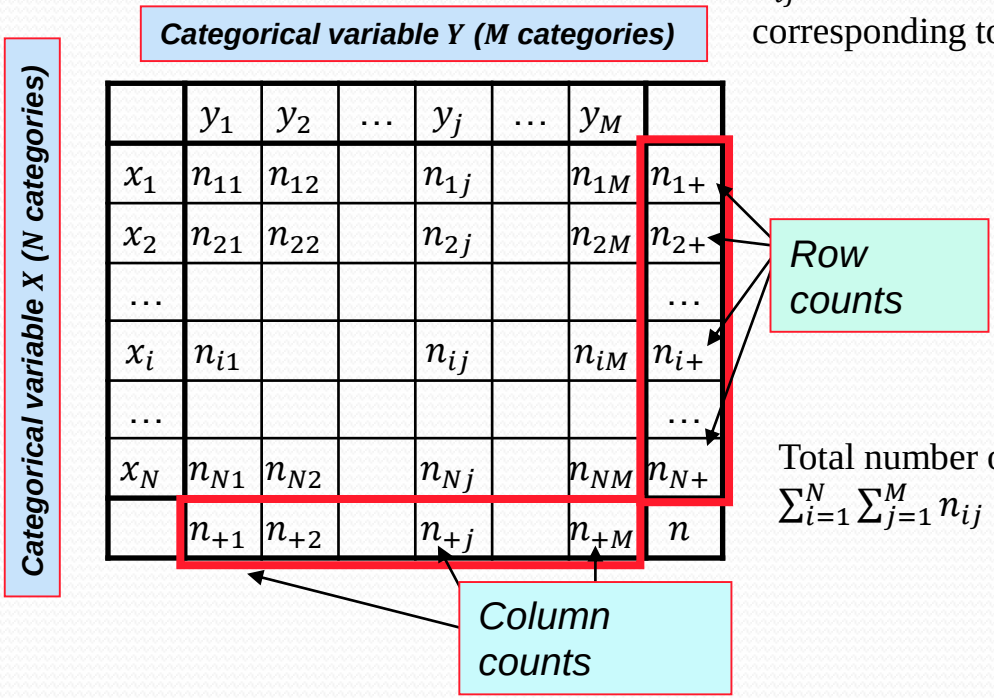
Correspondence Analysis

- Provides a graphical summary of the interactions between two categorical (discrete) random variables from the contingency table
- The issue with categorical (non-ordinal) variables is how to measure distances between two objects: correspondence analysis exploits contingency tables and association measures
- Can be very useful
  - to provide overview of cluster results
  - identifying sets of variables with similar “behaviour”
  - reduce the dimensionality of the problem (reduce the number of variables you have to worry about)
- However the correct interpretation is less than intuitive, and this leads many engineers astray
- Need to be familiar with the technical details

2 Frequency Tables

The Contingency Table

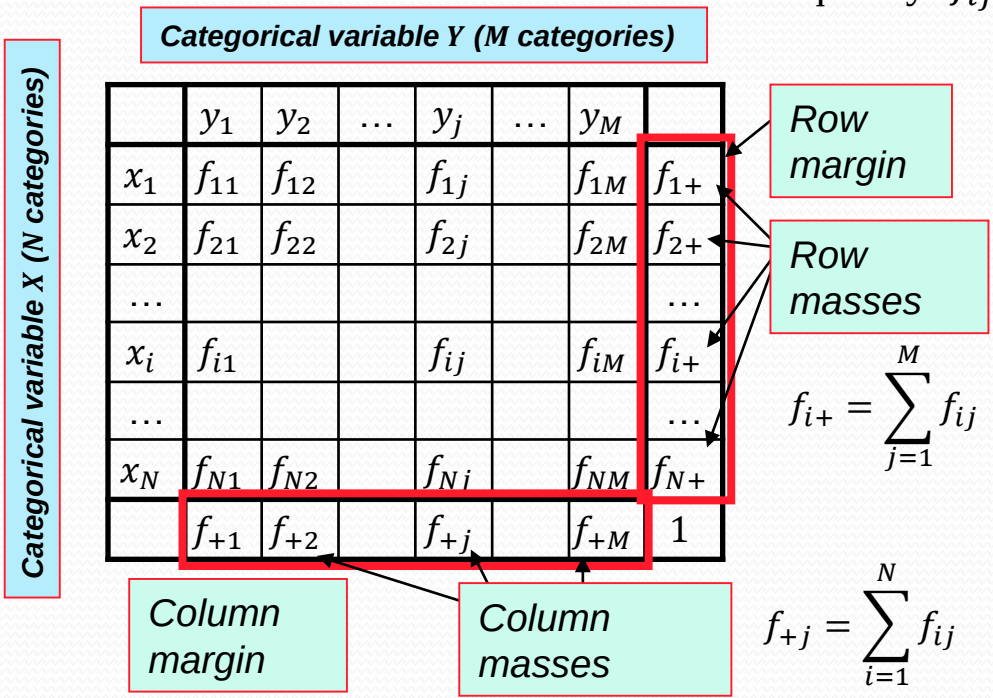
n_{ij} is the count of samples corresponding to the couple (x_i, y_j)



Total number of samples:
$\sum_{i=1}^{N} \sum_{j=1}^{M} n_{ij} = n$

The Frequency Table

Frequency: $f_{ij} = \frac{n_{ij}}{n}$



	y_1	y_2	…	y_j	…	y_M	
x_1	f_{11}	f_{12}		f_{1j}		f_{1M}	f_{1+}
x_2	f_{21}	f_{22}		f_{2j}		f_{2M}	f_{2+}
…							…
x_i	f_{i1}			f_{ij}		f_{iM}	f_{i+}
…							…
x_N	f_{N1}	f_{N2}		f_{Nj}		f_{NM}	f_{N+}
	f_{+1}	f_{+2}		f_{+j}		f_{+M}	1

Independence and Dependance

- Let X be a discrete RV taking on the values $\{x_1, x_2, \dots, x_N\}$
- Let Y be a discrete RV taking on the values $\{y_1, y_2, \dots, y_M\}$

- Independence of X and Y

$$P\big(X = x_i, Y = y_j\big) = P(X = x_i)P\big(Y = y_j\big), \forall x_i, y_j$$

- Otherwise, if there exists at least one couple (x_i, y_j) such that

$$P\big(X = x_i, Y = y_j\big) \neq P(X = x_i)P\big(Y = y_j\big),$$

the two variables are dependent!

- Conditional distribution: if X and Y are independent, we have

$$P\big(X = x_i \mid Y = y_j\big) = \frac{P\big(X = x_i, Y = y_j\big)}{P\big(Y = y_j\big)} = P(X = x_i), \quad \forall x_i, y_j$$

Approximation of probabilities

- Basic rate approximations:

$$P(X = x_i, Y = y_j) \approx f_{ij}$$
$$P(X = x_i) \approx f_{i+}$$
$$P(Y = y_j) \approx f_{+j}$$

- Conditional rate approximations:

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \approx \frac{f_{ij}}{f_{+j}} = f_{i|j}$$

$$P(Y = x_j \mid X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \approx \frac{f_{ij}}{f_{i+}} = f_{j|i}$$

- Approximation of independence:

$$f_{ij} = f_{i+} \times f_{+j} \quad \text{or} \quad \frac{f_{ij}}{f_{+j}} = f_{i+} \quad \text{or} \quad \frac{f_{ij}}{f_{i+}} = f_{+j}$$

Tables of Conditional Frequencies

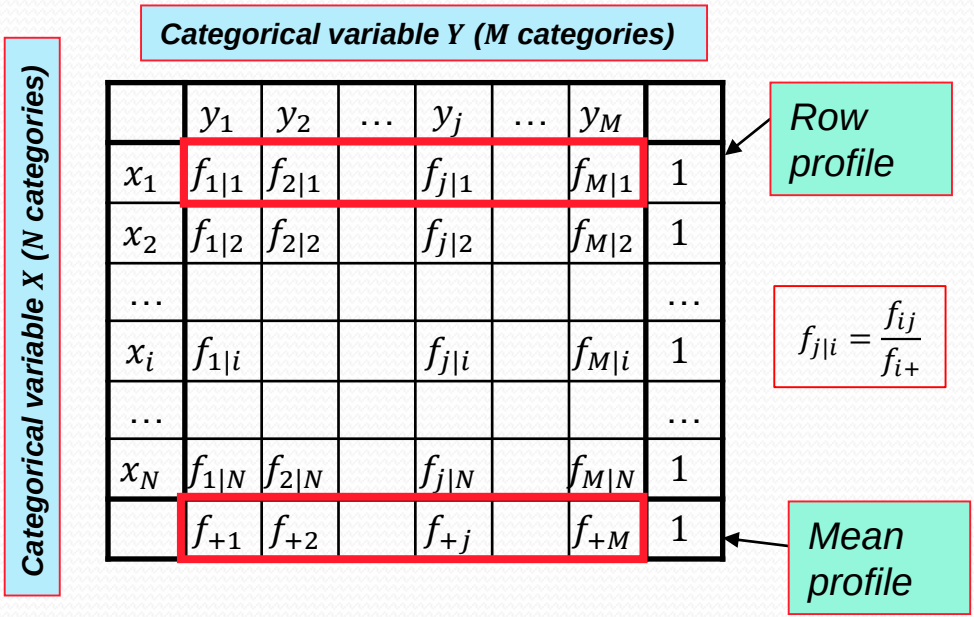
- The proportion of individuals that belong to category y_j for the variable Y among the individuals that have the modality x_i for the variable X form the so called table of **row profiles**.

- The conditional frequencies for fixed x_i and all y_j are

$$f_{j|i} = \frac{f_{ij}}{f_{i+}}$$

- Note that $f_{j|i} = \frac{n_{ij}}{n_{i+}}$ since $f_{ij} = \frac{n_{ij}}{n}$ and $f_{i+} = \frac{n_{i+}}{n}$

The conditional frequencies table



In case of independence

- The theoretical relative frequencies and theoretical counts under the assumption of independence are:
  - Relative frequencies:

$$f_{ij}^{*} = f_{i+} \times f_{+j}$$

  - Counts:

$$n_{ij}^{*} = \frac{n_{i+} \times n_{+j}}{n} = f_{ij}^{*} \times n$$

Example: Education and Salary

- We consider size 1000 sample of two categorical variables:

- Variable X "Education" is divided to categories
  - $A1$ Primary School,
  - $A2$ High School,
  - $A3$ University,

- Variable Y "Salary" is divided to categories
  - $B1$ Low,
  - $B2$ Average,
  - $B3$ High.

Education and Salary (contingency table)

Observed frequencies

	Low	Average	High	
Primary	150	40	10	200
High School	190	350	60	600
University	10	110	80	200
	350	500	150	1000

Theoretical frequencies under independence

	Low	Average	High	
Primary	70	100	30	200
High School	210	300	90	600
University	70	100	30	200
	350	500	150	1000

Education and Salary (frequency table)

Observed frequencies

	Low	Average	High	
Primary	0,15	0,04	0,01	0,20
High School	0,19	0,35	0,06	0,60
University	0,01	0,11	0,08	0,20
	0,35	0,5	0,15	1

Theoretical frequencies under independence

	Low	Average	High	
Primary	0,07	0,1	0,03	0,2
High School	0,21	0,3	0,09	0,6
University	0,07	0,1	0,03	0,2
	0,35	0,5	0,15	1

Education and Salary (conditional row profiles)

Observed frequencies

	Low	Average	High	
Primary	0,75	0,2	0,05	1
High School	0,32	0,58	0,1	1
University	0,05	0,55	0,4	1
	0,35	0,5	0,15	1

Theoretical frequencies under independence

	Low	Average	High	
Primary	0,35	0,5	0,15	1
High School	0,35	0,5	0,15	1
University	0,35	0,5	0,15	1
	0,35	0,5	0,15	1

3 Analysis of Row Profiles

Cloud S_N of row profiles ($N = 5, M = 3$)

Row profile:

$D_i = f_{\cdot|i} = (f_{1|i}, f_{2|i}, \dots, f_{M|i})$

such that $\sum_{j=1}^{M} f_{j|i} = 1$

and $f_{j|i} = \frac{f_{ij}}{f_{i+}}$

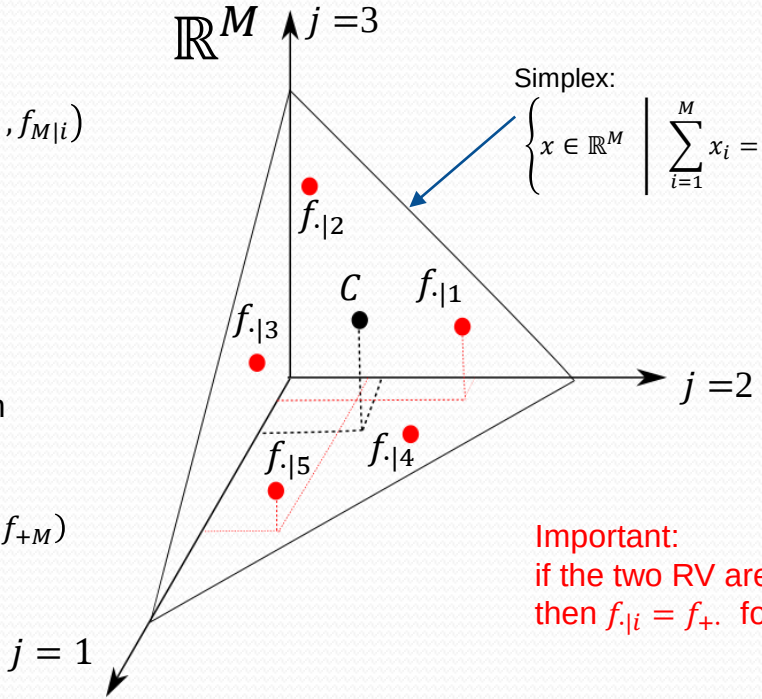
Center of mass:

$C = (c_1, c_2, \dots, c_M)$ with

$c_j = \sum_{i=1}^{N} f_{i+} f_{j|i} = f_{+j}$

Hence

$C = f_{+\cdot} = (f_{+1}, f_{+2}, \dots, f_{+M})$



Simplex:

$$\left\{ x \in \mathbb{R}^M \;\middle|\; \sum_{i=1}^{M} x_i = 1, \;\; x_1, \dots, \;\; x_M \geq 0 \right\}$$

Important:
if the two RV are independent,
then $f_{\cdot|i} = f_{+\cdot}$ for all row i

Cloud S_N of row profiles

- The center of mass is interpreted as an average row profile.
- It serves as a reference to study to what extent and in what way a class of individuals differs from the whole population.
- This difference is assessed by studying the gap between the profile of this class and the average row profile.
- Thus the study of the dispersion of the cloud around its center of mass is equivalent to the study of the gap between row profiles and the margin.
- It is also equivalent to the study of the connection between the two variables.

Chi-square distance

- When the data is in the form of frequency distribution, the distance between the rows (or columns) is measured using weighted Euclidian distances. The distance between two rows i_1 and i_2 is given by

$$d^2_{\chi^2}(\text{row}_{i_1}, \text{row}_{i_2}) = \sum_{j=1}^{M} \frac{1}{f_{+j}} \left(f_{j|i_1} - f_{j|i_2}\right)^2$$

- The χ^2 distance gives the same relative importance to each column proportionally to the average frequency.

- The division of each squared term by the expected frequency is variance standardizing and compensates for the larger variance in high frequencies and the smaller variance in low frequencies.

- Without this standardization, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped.

Chi-square distance

- Distance of row i to the center of mass:

$$d^2_{\chi^2}(\text{row}_i, C) = \sum_{j=1}^{M} \frac{1}{f_{+j}} \left(f_{j|i} - f_{+j}\right)^2$$

- Let S_N denotes the set of row profiles.

- The (total) inertia measures the dispersion of the cloud S_N from its center of mass. It can be expressed as the weighted average of the squared chi-square distances between the profiles and their average:

$$\text{Inertia}(S_N/C) = \sum_{i=1}^{N} f_{i+} d^2_{\chi^2}(\text{row}_i, C) = \sum_{i=1}^{N} f_{i+} \sum_{j=1}^{M} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - f_{+j}\right)^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\left(f_{ij} - f_{i+}f_{+j}\right)^2}{f_{i+}f_{+j}} = \frac{\chi^2}{n} = \Phi^2$$

- In case of independence,
  - All the row profiles coincide with the mean row profile
  - The inertia of the the cloud is zero: $\text{Inertia}(S_N/C) = 0$

- The more the data are dependent, the larger the gap between the row profiles and the center of mass is

PCA on the row profiles

- Principal Component Analysis (PCA) is based on maximizing Euclidian distances.
- In the context of frequency distributions, the proper distance between variables is the chi-square distance. Thus, for frequency distributions, PCA has to be applied to modified data.
- PCA decomposes the inertia of the cloud S_N of row profile.

PCA on the row profiles

- Let $R = (R_{ij})$ be the $N \times M$ matrix given by

$$R_{ij} = \frac{f_{j|i}}{\sqrt{f_{+j}}} - \sqrt{f_{+j}} = \frac{f_{ij}}{f_{i+}\sqrt{f_{+j}}} - \sqrt{f_{+j}}$$

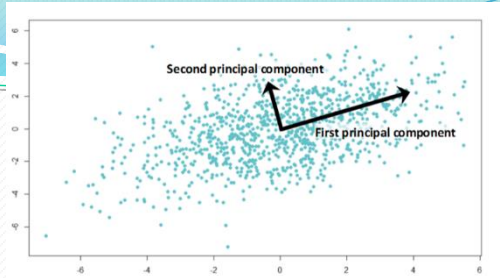
- The matrix R contains the scaled and shifted row profiles

- The matrix R satisfies

$$\sum_{i=1}^{N} f_{i+}\, R_{ij} = \sqrt{f_{+j}} - \sqrt{f_{+j}} = 0$$

- The weighted mean (center of mass) of the row profiles is 0.

Principle of the PCA



- Let R_i denote the ith row of R: $R_i = (R_{i1}, \dots, R_{iM})$
- Assuming that the center of mass of all R_i's is the origin $O = (0,0, \dots, 0)$
- Performing PCA on the row profiles equals to finding some orthonormal vectors (directions) $u_1, u_2, \dots, u_M$ defining a new vector basis of $\mathbb{R}^M$
- Each direction u_k is computed such that projection $z \in \mathbb{R}^M \mapsto P_k(z) = (z^T u_k)u_k$ onto u_k maximizes the weighted sum of the Euclidian distances

$$\sum_{i=1}^{N} f_{i+} d^2_{\chi^2}\big(O, P_k(R_i)\big)$$

  under the constraint that u_k is orthogonal to all previous u_ℓ for $1 \leq \ell < k$
- PCA Plane: when we are looking only for u_1 and u_2, we are looking for the best approximate plane of the cloud.

Solution of the PCA

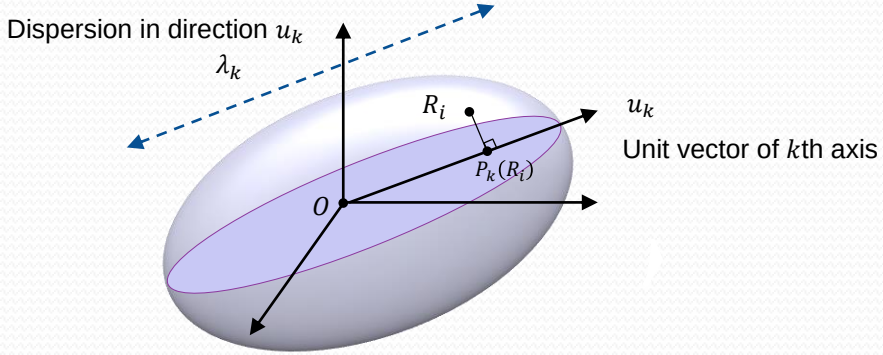
- As in the usual PCA, the solution is given by the eigenvalues and the eigenvectors of the matrix

$$V = \sum_{i=1}^{N} f_{i+} R_i^T R_i = Z^T Z$$

where $Z = (Z_{ij})$ be the $N \times M$ matrix given by

$$Z_{ij} = \frac{f_{ij} - f_{i+}\, f_{+j}}{\sqrt{f_{i+}\, f_{+j}}}$$

- Let λ_k denote the kth largest eigenvalue of the matrix V and let u_k denote the corresponding unit length eigenvector.



Solution of the PCA

- The value (or score) of the row profile i on the kth axis is given by

$$s_{ki} = R_i^T u_k = \sum_{j=1}^{M} R_{ij}\, u_{kj}$$

where u_{kj} denote the jth element of $u_k = (u_{k1}, \ldots, u_{kM})$.

- The score vector $s_k = (s_{k1}, \ldots, s_{kN})$ contains the values of the projection of all the row profiles $R_1, \ldots, R_N$ onto the kth axis
  - It can be proven that s_k is centered, i.e., $\sum_{i=1}^{N} f_{i+}\, s_{ki} = 0$
  - The variance of s_k is λ_k, i.e., $\sum_{i=1}^{N} f_{i+}\, s_{ki}^2 = \lambda_k$

- The contribution of the modality x_i on construction of the axis u_k is given by $\frac{f_{i+}\, s_{ki}^2}{\lambda_k}$

Quality of the projection

- We can show that
  - Inertia$(S_N/C) = \Phi^2 = \sum_{i=1}^{M} \lambda_i$
  - $\frac{\text{Projected inertia on axis } k}{\text{Total inertia}} = \frac{\sum_{i=1}^{N} f_{i+} d^2_{\chi^2}\left(O, P_k(R_i)\right)}{\sum_{i=1}^{N} f_{i+} d^2_{\chi^2}(C, R_i)} = \frac{\lambda_k}{\sum_{i=1}^{M} \lambda_i}$

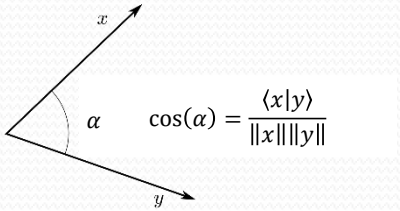
- Inertia is a measure of association between two categorical variables based on the Chi-square statistic.

- The proportion of inertia explained by each of the dimensions can be regarded as a measure of goodness-of-fit

Quality of the representation

- The quality of the representation of the centered row profile R_i by the principal axis k is measured by the squared cosine of the angle between the vector R_i and u_k :

$$\cos^2(\alpha_{ki}) = \left(\frac{\langle u_k | R_i \rangle}{\|u_k\| \|R_i\|}\right)^2 = \frac{s_{ki}^2}{\|R_i\|^2}$$



- If the value is close to 1, the quality of the representation is good.

- Note that the formula above does not contain the weight f_{i+}, and thus one modality can be:
  - Close to the axis u_k and therefore be well represented (well explained),
  - Due to a low weight f_{i+}, it can have a low contribution to the axis.

4 Analysis of Column Profiles

Analysis of column profiles

- Just replace the rows by the columns (like a transpose of the contingency table)
- Exactly the same approach, with the same quantities and the same interpretation
- Total inertia conservation:

$$\text{Inertia}(S_N/C_{row}) = \text{Inertia}(S_M/C_{col})$$

where S_M denotes the set of column profiles and C_{row}, resp. C_{col} is the center of mass of S_N, resp. S_M.

PCA on the column profiles

- Let $Q = (C_{ij})$ be the $N \times M$ matrix given by

$$Q_{ij} = \frac{f_{ij}}{f_{j+}\sqrt{f_{i+}}} - \sqrt{f_{i+}}$$

- The matrix Q contains the scaled and shifted row profiles

- The solution of the PCA decomposition for the columns is given by the eigenvalues and the eigenvectors of the matrix

$$W = \sum_{j=1}^{M} f_{+j} Q_j Q_j^T = ZZ^T$$

where Q_j denotes the jth column of Q

- Let v_{kj} denote the jth element of $v_k = (v_{k1}, \ldots, v_{kN})$ where v_k is the kth eigenvector of W. The value (or score) of the column profile j on the kth principal component is given by

$$s_{kj} = v_k^T Q_j = \sum_{i=1}^{N} v_{ki} Q_{ij}$$

Association between the profiles

- We can show that the row scores and the column scores satisfy the following equations:

$$s_{ki}^{row} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{M} f_{j|i} \; s_{kj}^{col}$$

$$s_{kj}^{col} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} f_{i|j} \; s_{ki}^{row}$$

- Interpretation:
  - The coordinates of the rows can be obtained from the set of points columns (and vice versa)
  - The row score s_{ki}^{row} is a convex combination of the column scores s_{kj}^{col} (and vice versa)

- Advantage: simultaneous representation
  - We can plot together the row plane and the column plane to benefit from the previous interpretation.
  - This simultaneous representation is possible since the origin is common to both planes.

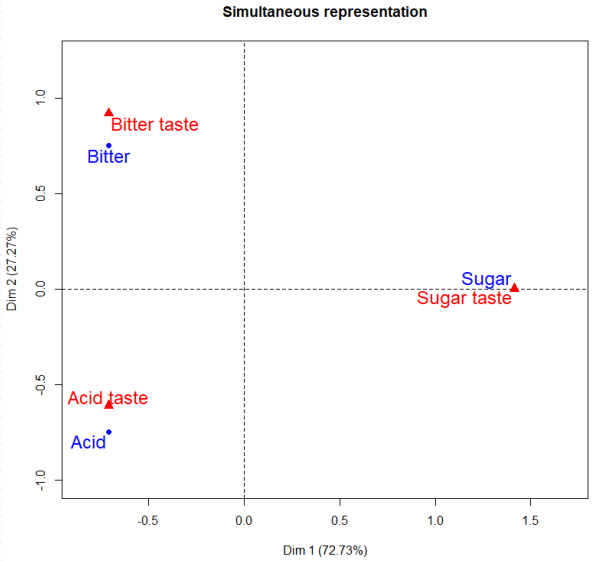
Example

- For each flavor, ten individuals were asked to recognize the taste of several beverages

Recognized taste

Flavor	Sugar taste	Acid taste	Bitter taste
Sugar	10	0	0
Acid	0	9	1
Bitter	0	3	7

	Dim.1	Dim.2
Variance	1.000	0.375
% of var.	72.727	27.273
Cumulative % of var.	72.727	100.000

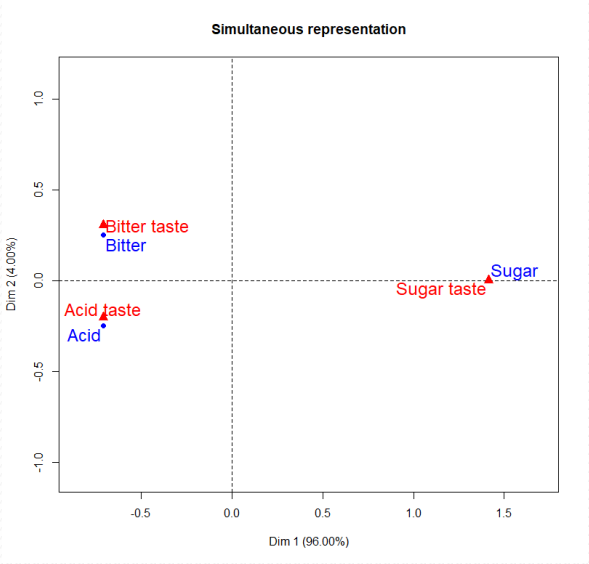


Example

Recognized taste

Flavor	Sugar taste	Acid taste	Bitter taste
Sugar	10	0	0
Acid	0	7	3
Bitter	0	5	5

	Dim.1	Dim.2
Variance	1.000	0.042
% of var.	96.000	4.000
Cumulative % of var.	96.000	100.000



5 Conclusion

Conclusion

- Correspondence analysis is a technique to study and visualize the interactions between two discrete random variables
- This technique is very similar to PCA and can be employed for data reduction purposes or to plot perceptual maps
- Extension to process more than two discrete variables

6 Appendix

Steps to run correspondence analysis

- Represent the data in a contingency table
- Translate the frequencies of the contingency table into a matrix of metric (continuous) distances through a set of Chi-square association measures on the row and column profiles
- Extract the dimensions (in a similar fashion to PCA)
- Evaluate the explanatory power of the selected number of dimensions
- Plot row and column objects in the same co-ordinate space

Chi-Square Test

- Under random sampling, the n_{ij} follow multinomial distribution with parameters n, p_{11}, p_{12},…, p_{1M}, p_{21},…, p_{NM} and $E[n_{ij}] = n\, p_{ij}$

- In the test statistics χ^2, the np_{ij}, under H_0, are estimated by n^*_{ij}.

- When n is large, the test statistic has, under H_0, approximately chi-square distribution with $(N-1)(M-1)$ degrees of freedom.

- Thus the null hypothesis (independence between variables x and y) is rejected at the level α if

$$\chi^2 > \chi^2_{(N-1)(M-1),1-\alpha}$$

where $\chi^2_{p,1-\alpha}$ is the $1-\alpha$ quantile of the Chi-Square distribution with p degrees of freedom.

Test of independence

- The independence between variables X and Y can be tested using chi-square statistic. The null hypothesis of the test is

$$H_0: p_{ij} = p_{i+} \times p_{j+}$$

- The test statistic is given by

$$\chi^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} = n\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*} = n\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left(f_{ij} - f_{i+}f_{+j}\right)^2}{f_{i+}f_{+j}} = n\Phi^2$$

- Intensity of the link = Φ^2
- A high intensity does not mean a high significance of the link
- Significance is related to the number of measurements

Decomposition of the chi-square statistic

- Let $Z = (Z_{ij})$ be the $N \times M$ matrix given by

$$Z_{ij} = \frac{f_{ij} - f_{i+}\, f_{+j}}{\sqrt{f_{i+}\, f_{+j}}}$$

- The variables Z_{ij} are centered and scaled such that they satisfy

$$\Phi^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij}^2$$

- The quantity $\sum_{j=1}^{M} Z_{ij}^2$ is the contribution of the modality x_i to Φ^2

- The quantity $\sum_{i=1}^{N} Z_{ij}^2$ is the contribution of the modality y_j to Φ^2

Study of inertia

- We can show that $0 \leq \lambda_k \leq 1$
- What does $\lambda_k = 1$ mean ?
- Partition in two classes of the rows and the columns: Exclusive association of two classes

- Exemple: