Data Visualization

ADDA71

Correspondence Analysis (CA)

Ilyes Jenhani

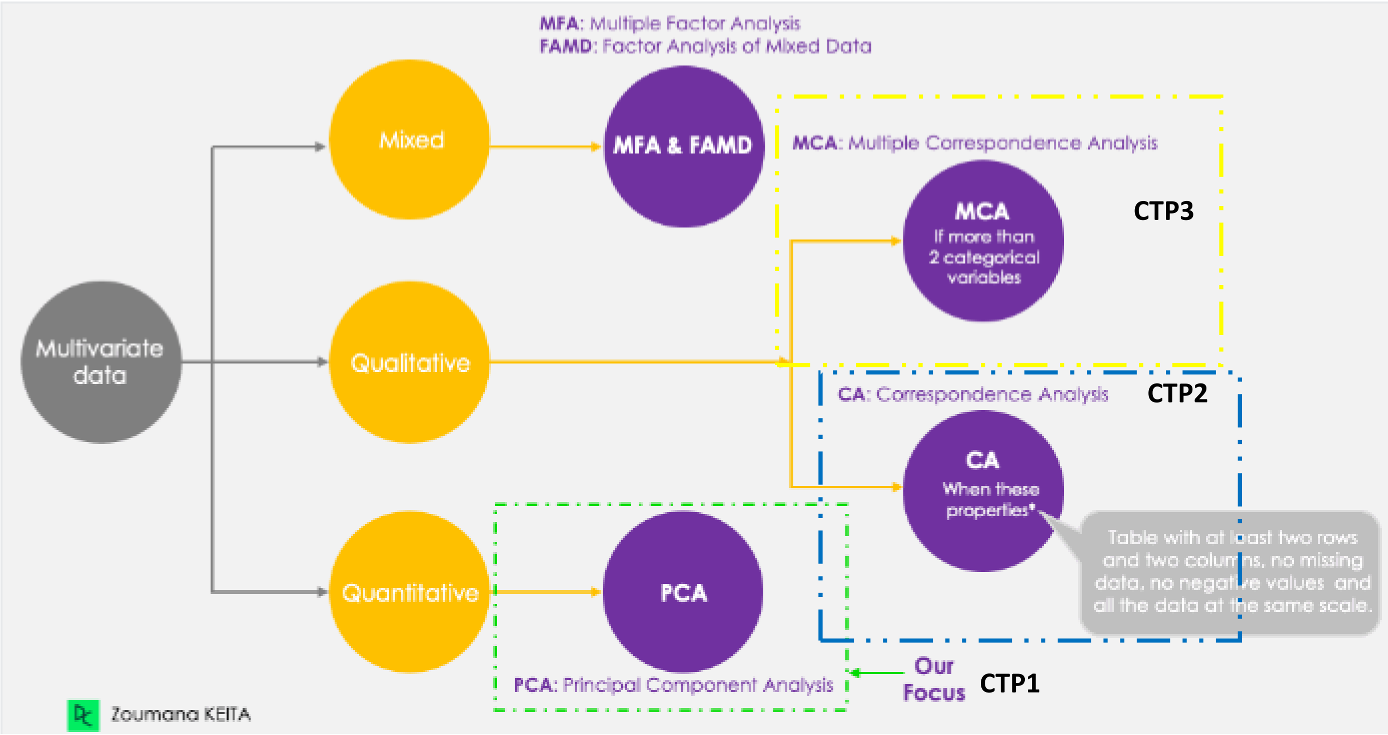
Schedule (Tentative) (2)

#	Class type	Content
1	CTP	Principal Component Analysis (PCA)
2	CTP	Correspondence Analysis (CA)
3	CTP	Multiple Correspondence Analysis (MCA)
4	PRJ	Project

Remember

Dimensionality reduction methods

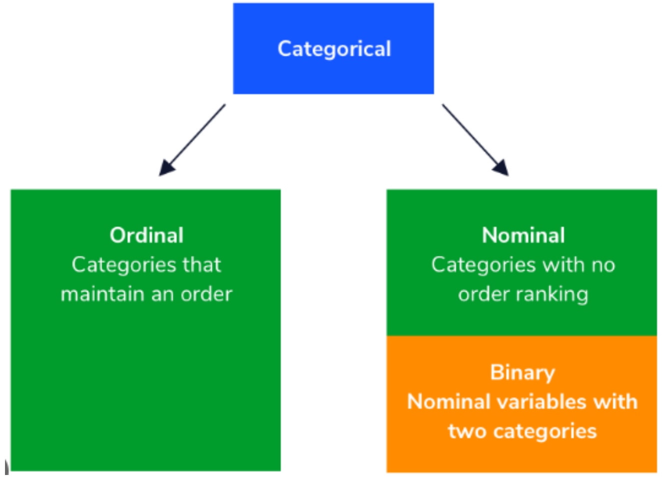
- Aim : summarize and visualize *multivariate* data

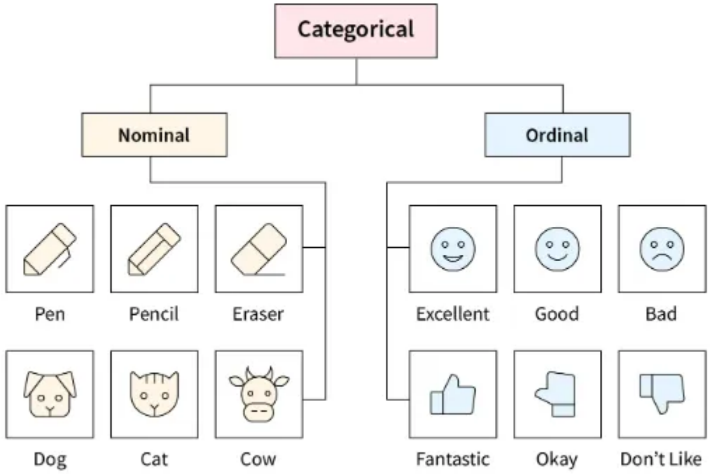


2. Correspondence Analysis (CA)

What is Correspondence Analysis?

- Correspondence Analysis (CA) is a statistical technique used to analyze and visualize the relationships between 2 categorical variables within a contingency table.
- **Categorical** variables:





What is Correspondence Analysis? (2)

- **Contingency tables**: a.k.a. crosstabs, Chi-square tables, Pivot tables, etc.
  - Tables used in statistics to display the frequency distribution of the categories of variables.
  - Used to examine the relationship between two or more categorical variables: Show how the categories of one variable relate to the categories of another.
  - Help summarize the data and facilitate the calculation of probabilities, statistical tests like the **chi-square test**, etc.

Gender	With umbrella
female	yes
male	yes
female	yes
female	yes
male	yes
male	no
female	no
male	no
female	no
female	no
male	no
female	yes
male	yes
female	yes
male	yes
male	yes
male	no
female	no
male	no
female	no
female	no
female	no

		With umbrella		
		yes	no	Total
Gender	female	5	7	12
	male	5	5	10
	Total	10	12	22

Case	Gender	Highest level of education
1	Male	College
2	Female	Bachelor
3	Male	Without graduation
4	Male	Master
5	Female	Master
...	...	...

Female and without a degree occurs 6 times in the data

	Female	Male
Without graduation	6	7
College	13	16
Bachelor	16	15
Master	8	11
Total	43	49

What is Correspondence Analysis? (3)

- CA provides a visual representation of the data allowing for the identification of the patterns and associations between the categories of the variables.
- CA is a dimensionality reduction technique that converts categorical data into coordinates in a low-dimensional space.
- As in PCA, the first component (or dimension) captures the most variation in the data, reducing the complexity while retaining the essential information. Each component is a linear combination of the original variables, representing a new axis of variation.
- The output of CA is a set of factor scores that can be plotted to show how the categories of the variables relate to one another in terms of proximity or distance.

Chi-square test of independence

- A statistical test used to determine if there is a significant association (dependency) between two categorical variables.
- It evaluates whether the observed distribution of sample categorical data deviates from the expected distribution.
  - The expected distribution is what we would observe if the two variables were statistically independent.
- Calculating the distance between the observed and expected distributions is a key step in CA
- The distances are used as input for the *Singular value decomposition* (SVD). This decomposition helps in identifying the most important factors driving the associations in the data.
- In CA, the total *variance* (*Inertia* or *deviation*) in the data is computed as the sum of all chi-square distances across the table, divided by the total number of observations

  - **Null Hypothesis (H_0):** Assumes that the two variables are independent (no association between them).
  - **Alternative Hypothesis (H_1):** Assumes that the two variables are not independent (there is an association).

Chi-square test of independence STEPS

		With umbrella		
		yes	no	Total
Gender	female	5	7	12
	male	5	5	10
	Total	10	12	22

From a contingency table:

1. Calculate Expected Frequencies**:** Based on the assumption of independence, the expected frequency for each cell in the contingency table is calculated using the formula: $E_{ij} = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$
2. Calculate the Chi-Square statistic: $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
   - O_{ij} = Observed frequency in each cell
   - E_{ij} = Expected frequency for each cell
3. Determine Degrees of Freedom: $\text{Degrees of freedom (df)} = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$
4. Compare the calculated Chi-Square statistic (step 2) with a critical value from the Chi-square distribution (based on the degrees of freedom and significance level alpha, e.g., 0.05) to determine if the observed distribution significantly differs from the expected distribution.

	alpha		
df	0.050	0.010	0.001
1	3.84146	6.63490	10.828
2	5.99147	9.21034	13.816
3	7.81473	11.3449	16.266
4	9.48773	13.2767	18.467
5	11.0705	15.0863	20.515
6	12.5916	16.8119	22.458

5. Conclusion: If the Chi-square value is greater than the critical value, the null hypothesis is rejected, indicating that the variables are dependent (associated) and we can proceed to the CA. Otherwise, CA will not be informative.

CA: Step by Step (1)

The below contingency table is the result of a survey about brands of soda products:

Each cell in the table represents the number of responses or counts associating that attribute with that brand.

This ‘association’ would be displayed through a survey question such as ‘pick brands from a list below which you believe show ___ attribute’

Contingency Table			
Brands	**Attributes**		
	Tasty	Aesthetic	Economic
Butterbeer	5	7	2
Squishee	18	46	20
Slurm	19	29	39
Fizzy Lifting Drink	12	40	49
Brawndo	3	7	**16**

CA: Step by Step (2)

Residuals (R): Chi-square distances

- A residual quantifies the difference between the observed data and the data we would expect - assuming there is no dependency between the row and column categories.
- A residual (R) is equal to: R = P - E, where P is the observed proportions and E is the expected proportions for each cell.
- A high residual (distance) indicates that the count for that brand-attribute pairing is much higher than expected, suggesting a strong relationship and vice versa.

Observed proportions (P):

- P = the value in a cell divided by the total sum of all of the values in the table (312 in this example)

Contingency Table			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	5	7	2
Squishee	18	46	20
Slurm	19	29	39
Fizzy Lifting Drink	12	40	49
Brawndo	3	7	**16**

Observed Proportions Calculations (R = *P* - E)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	0.016	0.022	0.006
Squishee	0.058	0.147	0.064
Slurm	0.061	0.093	0.125
Fizzy Lifting Drink	0.038	0.128	0.157
Brawndo	0.01	0.022	**0.051**

CA: Step by Step (3)

Rows and columns masses :

A row (resp. column) mass is the proportion of values for that row (resp. column)

Row and Column Mass Calculations				
Brands	Attributes			Row Masses
	Tasty	Aesthetic	Economic	
Butterbeer	0.016	0.022	0.006	**0.044**
Squishee	0.058	0.147	0.064	**0.269**
Slurm	0.061	0.093	0.125	**0.279**
Fizzy Lifting Drink	0.038	0.128	0.157	**0.324**
Brawndo	0.01	0.022	0.051	**0.083**
Column Masses	**0.182**	**0.413**	**0.404**	

Expected proportions (E):

- What we expect to see in each cell's proportion, assuming that there is no relationship between rows and columns.
- Our expected value for a cell would be the row mass of that cell multiplied by the column mass of that cell.

Expected Proportion Calculations (R = P - *E*)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	**0.008**	0.019	0.018
Squishee	0.049	0.111	0.109
Slurm	0.051	0.115	0.113
Fizzy Lifting Drink	0.059	0.134	0.131
Brawndo	0.015	0.034	0.034

CA: Step by Step (4)

Residuals calculation:

Observed Proportions Calculations (R = *P* - E)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	0.016	0.022	0.006
Squishee	0.058	0.147	0.064
Slurm	0.061	0.093	0.125
Fizzy Lifting Drink	0.038	0.128	0.157
Brawndo	0.01	0.022	**0.051**

\-

Expected Proportion Calculations (R = P - *E*)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	**0.008**	0.019	0.018
Squishee	0.049	0.111	0.109
Slurm	0.051	0.115	0.113
Fizzy Lifting Drink	0.059	0.134	0.131
Brawndo	0.015	0.034	0.034

=

Residuals Calculation (*R* = P - E)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	0.008	0.004	-0.012
Squishee	0.009	0.036	**-0.045**
Slurm	0.01	-0.022	0.012
Fizzy Lifting Drink	-0.021	-0.006	0.026
Brawndo	-0.006	-0.012	0.018

Taking our most negative value of -.045 for Squishee and Economic, what we would interpret here is that there is a negative association between Squishee and Economic; Squishee is much less likely to be viewed as 'Economic' than our other brands of drinks.

Indexed residuals (I):

- Looking at the top row from our residuals calculation table above, we see that all these numbers are very close to 0. We shouldn't take the obvious conclusion from this that Butterbeer is unrelated to our attributes, as this assumption is incorrect.
- The actual explanation would be that the observed proportions (P) and the expected proportions (E) are small because, as our row mass tells us, only 4.4% of the sample are Butterbeer.
- Results are skewed towards the rows/columns with larger masses. We can fix this by dividing our residuals by our expected proportions (E), giving us a table of our indexed residuals (I, I = R / E):

I = (P-E)/E: relative difference

Indexed Residuals Calculation (I, I = R / E)			
Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	**0.95**	0.21	**-0.65**
Squishee	0.17	0.32	-0.41
Slurm	0.2	-0.19	0.11
Fizzy Lifting Drink	-0.35	-0.04	0.2
Brawndo	-0.37	-0.35	0.52

Butterbeer is 95% more likely to be viewed as 'Tasty' than what we would expect if there were no relationship between these brands and attributes. Whereas at the top right value, Butterbeer is 65% less likely to be viewed as 'Economic' than what we would expect - given no relationship between our brands and attributes

CA: Step by Step (5)

- Given our indexed residuals(I), our expected proportions (E), our observed proportions (P), and our row and column masses, let's get to calculating our correspondence analysis values for our chart!

- Calculating coordinates for Correspondence Analysis
  1. Singular Value Decomposition (SVD): The SVD gives us values to calculate variance and plot our rows and columns (brands and attributes).
  2. We calculate the SVD on the **standardized residual (Z)**, where Z = I * sqrt(E), where I is our indexed residual, and E is our expected proportions
  3. SVD = svd(Z). A singular value decomposition generates 3 outputs:

A vector, d, containing the *singular values*

Singular Values (d)		
1st dim	2nd dim	3rd dim
2.65E-01	1.14E-01	4.21E-17

A matrix, u, containing the *left singular vectors*

Left Singluar Vectors (u)			
Brands	Dimensions		
	1st dim	2nd dim	3rd dim
Butterbeer	-0.439	-0.424	-0.084
Squishee	-0.652	0.355	-0.626
Slurm	0.16	-0.0672	-0.424
Fizzy Lifting Drink	0.371	0.488	-0.274
Brawndo	0.469	-0.06	-0.588

A matrix, v, containing the *right singular vectors*.

Right Singluar Vectors (v)			
Attributes	Dimensions		
	1st dim	2nd dim	3rd dim
Tasty	-0.41	-0.81	-0.427
Aesthetic	-0.489	0.59	-0.643
Economic	0.77	-0.055	-0.635

CA: Step by Step (6)

A vector, d, containing the *singular values*

Singular Values (d)		
1st dim	2nd dim	3rd dim
2.65E-01	1.14E-01	4.21E-17

A matrix, u, containing the *left singular vectors*

Left Singluar Vectors (u)			
Brands	Dimensions		
	1st dim	2nd dim	3rd dim
Butterbeer	-0.439	-0.424	-0.084
Squishee	-0.652	0.355	-0.626
Slurm	0.16	-0.0672	-0.424
Fizzy Lifting Drink	0.371	0.488	-0.274
Brawndo	0.469	-0.06	-0.588

A matrix, v, containing the *right singular vectors*.

Right Singluar Vectors (v)			
Attributes	Dimensions		
	1st dim	2nd dim	3rd dim
Tasty	-0.41	-0.81	-0.427
Aesthetic	-0.489	0.59	-0.643
Economic	0.77	-0.055	-0.635

- Each of the singular values corresponds to a *dimension*.
- The left (resp. right) singular vectors correspond to the categories in the rows (resp. columns) of the table.
- The coordinates used to plot row and column categories for our correspondence analysis chart are derived from the first two dimensions.
- Variance expressed by our dimensions:
  - Squared singular values are known as *eigenvalues*(d^2). The eigenvalues in our example are .0704, .0129, and .0000.
  - Expressing each eigenvalue as a proportion of the total sum tells us the amount of *variance* captured in each dimension of our correspondence analysis.
  - Based on each dimensions' singular value; we get 84.5% of variance expressed by our first dimension, and 15.5% in our second dimension (our third dimension explains 0% of the variance).

CA: Step by Step (7)

A vector, d, containing the *singular values*

Singular Values (d)		
1st dim	2nd dim	3rd dim
2.65E-01	1.14E-01	4.21E-17

A matrix, u, containing the *left singular vectors*

Left Singluar Vectors (u)			
Brands	Dimensions		
	1st dim	2nd dim	3rd dim
Butterbeer	-0.439	-0.424	-0.084
Squishee	-0.652	0.355	-0.626
Slurm	0.16	-0.0672	-0.424
Fizzy Lifting Drink	0.371	0.488	-0.274
Brawndo	0.469	-0.06	-0.588

A matrix, v, containing the *right singular vectors*.

Right Singluar Vectors (v)			
Attributes	Dimensions		
	1st dim	2nd dim	3rd dim
Tasty	-0.41	-0.81	-0.427
Aesthetic	-0.489	0.59	-0.643
Economic	0.77	-0.055	-0.635

Row and Column Mass Calculations				
Brands	Attributes			Row Masses
	Tasty	Aesthetic	Economic	
Butterbeer	0.016	0.022	0.006	**0.044**
Squishee	0.058	0.147	0.064	**0.269**
Slurm	0.061	0.093	0.125	**0.279**
Fizzy Lifting Drink	0.038	0.128	0.157	**0.324**
Brawndo	0.01	0.022	0.051	**0.083**
Column Masses	**0.182**	**0.413**	**0.404**	

Standard coordinates:

- Calculated from the left and right singular vectors.
- Previously, we weighted the indexed residuals prior to performing the SVD. In order to get coordinates that represent our indexed residuals, we now need to unweight the SVD's outputs, *by dividing each row of the left singular vectors by the square root of the row masses, and dividing each column of the right singular vectors by the square root of the column masses*, getting us the standard coordinates of the rows and columns for plotting.

0.16 / sqrt(0.279)
= 0.16 / 0.528
= 0.3

Brand Standard Coordinates			
Brands	Dimensions		
	1st dim	2nd dim	3rd dim
Butterbeer	-2.07	-2	-0.4
Squishee	-1.27	0.68	-1.21
Slurm	0.3	-1.27	-0.8
Fizzy Lifting Drink	0.65	0.86	-0.48
Brawndo	1.62	-0.21	-2.04

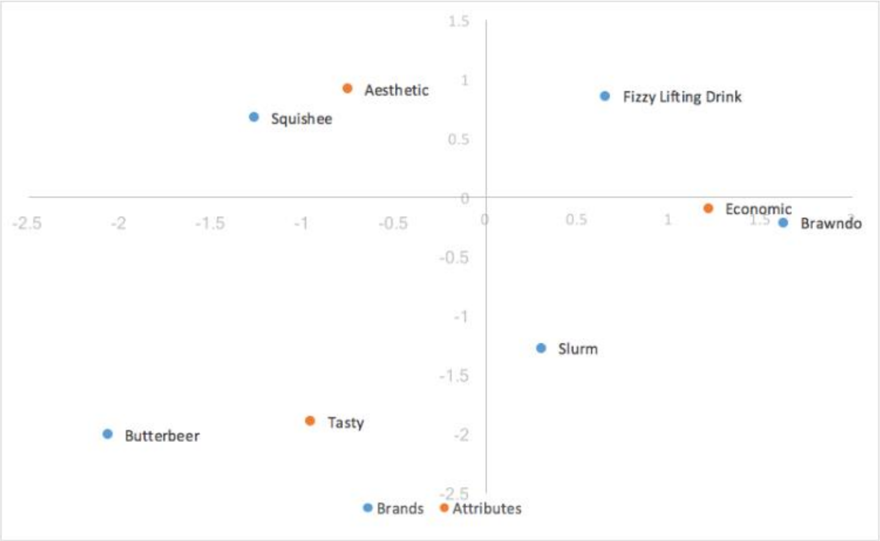
Attribute Standard Coordinates			
Attributes	Dimensions		
	1st dim	2nd dim	3rd dim
Tasty	-0.96	-1.89	-1
Aesthetic	-0.76	0.92	-1
Economic	1.21	-0.09	-1

CA: Step by Step (8)

Brand Standard Coordinates			
Brands	Dimensions		
	1st dim	2nd dim	3rd dim
Butterbeer	-2.07	-2	-0.4
Squishee	-1.27	0.68	-1.21
Slurm	0.3	-1.27	-0.8
Fizzy Lifting Drink	0.65	0.86	-0.48
Brawndo	1.62	-0.21	-2.04

Attribute Standard Coordinates			
Attributes	Dimensions		
	1st dim	2nd dim	3rd dim
Tasty	-0.96	-1.89	-1
Aesthetic	-0.76	0.92	-1
Economic	1.21	-0.09	-1

- We use the two dimensions with the *highest variance* captured for plotting, the first dimension going on the X axis, and the second dimension on the Y axis, generating our correspondence analysis graph.



Prince

- Prince is a Python library for multivariate exploratory data analysis in Python.
- It includes a variety of methods for summarizing tabular data, including principal component analysis and correspondence analysis.
- Prince provides efficient implementations, using a scikit-learn API.
- Prince uses Altair for making charts.
- Prince GitHub
- Let’s start our Lab !