Correspondence analysis

Correspondence analysis (CA) is an exploratory analysis method that aims to graphically represent the relationships between the modalities of two qualitative variables. It allows the **row profiles** (in $\mathbb{R}^p$) and **column profiles** (in $\mathbb{R}^n$) of a contingency table to be represented simultaneously in a low-dimensional space, while preserving the χ^2 distance as closely as possible. The objective of AFC is to find a two-dimensional (or three-dimensional) representation in which the geometric proximity between points best reflects the similarities between the modalities.

> Note
>
> AFC can be seen as a double PCA: a weighted PCA applied to row profiles and column profiles, in their respective spaces with an appropriate metric.

1 Notation

Consider a contingency table $K = (k_{ij})$, where k_{ij} is the number of individuals belonging to class $i \in \{1, \dots, n\}$ and category $j \in \{1, \dots, p\}$. We then work with the relative frequency table by normalizing this table. Since the frequencies are proportional to the sample size n, the relative frequency table contains more information. Let $F = (f_{ij})$, where

$$f_{ij} = \frac{k_{ij}}{k_{\bullet\bullet}} = \frac{k_{ij}}{\sum_{l=1}^{n}\sum_{m=1}^{p} k_{lm}}.$$

The row (resp. column) margins of the table correspond to the sum of the columns for each row (resp. the sum of the rows for each column):

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = \frac{k_{i\bullet}}{k_{\bullet\bullet}}, \quad 1 \leq i \leq n; \tag{1}$$

$$f_{\bullet j} = \sum_{i=1}^{n} f_{ij} = \frac{k_{\bullet j}}{k_{\bullet\bullet}}, \quad 1 \leq j \leq p. \tag{2}$$

We have $f_{\bullet\bullet} = \sum_{i=1}^{n} f_{i\bullet} = \sum_{j=1}^{p} f_{\bullet j} = 1$.

Example

As an example, let's consider the major and admission type of students enrolled in the STT-2200 course in the fall of 2025.

Table 1: Contingency table of students enrolled in the STT-2200 course (Fall 2025) cross-referencing their major and type of admission.

	College	Laval University	Other university	Outside Quebec
Actuarial Science	2	0	0	1
Statistics	2	4	1	0
Bioinformatics	4	2	0	2
Finance	2	0	0	0
Math	1	0	0	0
Computer Science	2	1	0	1

Here, we find $k_{\bullet\bullet} = 25$. This is simply the number of students enrolled in the course. We therefore find the following frequency table:

Table 2: Frequency table associated with the previous contingency table.

	College	Laval University	Other university	Outside Quebec	$f_{i\bullet}$
Actuarial Science	0.08	0	0	0.04	0.12
Statistics	0.08	0.16	0.04	0	0.28
Bioinformatics	0.16	0.08	0	0.08	0.32
Finance	0.08	0	0	0	0.08
Math	0.04	0	0	0	0.04
Info	0.08	0.04	0	0.04	0.16
$f_{\bullet j}$	0.52	0.28	0.04	0.16	1

2 Statistical independence

The relative frequency table $F = (f_{ij})$ can be interpreted as an estimate of the joint probabilities of the modalities of the two qualitative variables. If the two variables are statistically independent, we expect the joint probability to approach the product of the marginal probabilities:

$$f_{ij} \approx f_{i\bullet} f_{\bullet j}, \quad i \in \{1, \dots, n\},\ j \in \{1, \dots, p\}.$$

To test whether the observed differences between f_{ij} and $f_{i\bullet} f_{\bullet j}$ are significant, we use the χ^2

test of independence:

$$T = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(k_{ij} - \mathbb{E}(k_{ij})\right)^2}{\mathbb{E}(k_{ij})} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(k_{ij} - \frac{k_{i\bullet}k_{\bullet j}}{k_{\bullet\bullet}}\right)^2}{\left(\frac{k_{i\bullet}k_{\bullet j}}{k_{\bullet\bullet}}\right)}.$$

Under the assumption of independence, this statistic approximately follows a χ^2 distribution. If the variables are independent, the statistic T should be close to 0.

3 Row profiles and column profiles

To analyze the structures in the contingency table, we introduce the concept of a profile. Each row of the table can be viewed as a row profile

$$L_i = \left(\frac{k_{i1}}{k_{i\bullet}}, \dots, \frac{k_{ip}}{k_{i\bullet}}\right) = \left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{ip}}{f_{i\bullet}}\right).$$

The line profile represents the distribution of the modalities i of the first variable among the modalities of the second.

Similarly, each column of the table can be viewed as a column profile

$$C_j = \left(\frac{k_{1j}}{k_{\bullet j}}, \dots, \frac{k_{nj}}{k_{\bullet j}}\right) = \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right).$$

The column profile represents the distribution of the modalities j of the second variable among the modalities of the first.

We can then look at the mean row profile (resp. mean column profile) obtained as the weighted average of the row profiles (resp. column profiles). In other words, they correspond to the marginal frequencies of the columns (resp. marginal frequencies of the rows). The mean row profile is given by

$$\left(\sum_{i=1}^{n} f_{i\bullet}\frac{f_{i1}}{f_{i\bullet}}, \dots, \sum_{i=1}^{n} f_{i\bullet}\frac{f_{ip}}{f_{i\bullet}}\right) = \left(f_{\bullet 1}, \dots, f_{\bullet p}\right),$$

and the average column profile is given by

$$\left(\sum_{j=1}^{p} f_{\bullet j}\frac{f_{1j}}{f_{\bullet j}}, \dots, \sum_{j=1}^{p} f_{\bullet j}\frac{f_{nj}}{f_{\bullet j}}\right) = \left(f_{1\bullet}, \dots, f_{n\bullet}\right).$$

If the variables are independent, all profiles are equal to their respective average profiles. In other words, for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$,

$$\left(\frac{f_{i1}}{f_{i\otimes}}, \dots, \frac{f_{ip}}{f_{i\otimes}}\right) = \left(f_{\bullet 1}, \dots, f_{\bullet p}\right) \quad \text{and} \quad \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right) = \left(f_{1\bullet}, \dots, f_{n\bullet}\right).$$

Thus, the further the profiles deviate from their means, the more the variables show dependence.

To measure the difference between two line profiles, we use the χ^2 distance weighted by marginal frequencies:

$$d^2(L_i, L_{i'}) = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2.$$

We can do the same for the difference between two column profiles:

$$d^2(C_j, C_{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2.$$

This can be written in matrix form. Let $D_n = \mathrm{diag}(f_{i\bullet})$ be the diagonal matrix of row weights and $D_p = \mathrm{diag}(f_{\bullet j})$ be the diagonal matrix of column weights. The matrix $D_n^{-1}F$ has the row profiles as its rows, and the matrix $D_p^{-1}F^\top$ has the column profiles as its rows. The χ^2 distance between two row profiles L_i and $L_{i'}$ is then written as

$$d^2(L_i, L_{i'}) = (L_i - L_{i'})^\top D_p^{-1} (L_i - L_{i'}),$$

and similarly for two column profiles C_j and $C_{j'}$

$$d^2(C_j, C_{j'}) = (C_j - C_{j'})^\top D_n^{-1} (C_j - C_{j'}).$$

These distances form the basis of geometric representation in correspondence analysis, where we seek a projection of the profiles into a low-dimensional space that best preserves these distances.

4 Estimation of Eigenvalues

The analysis of line profiles is called direct analysis. We consider the line profiles contained in the matrix $D_n^{-1}F \in \mathbb{R}^{n\times p}$. The row profiles are projected into a space equipped with the χ^2 metric on the columns, defined by

$$\langle x, y \rangle = x^\top D_p^{-1} y.$$

The analysis of the column profiles is called dual analysis. We consider the column profiles contained in the matrix $D_p^{-1}F^\top \in \mathbb{R}^{p\times n}$. We project the column profiles into a space equipped with the χ^2 metric on the rows, defined by

$$\langle x, y \rangle = x^\top D_n^{-1} y.$$

For direct analysis, we seek the first factorial axis, i.e., the direction $u \in R^p$ that maximizes the projected variance of the row profiles, under the constraint that u is normalized. We therefore seek

$$\max_u u^\top D_p^{-1} F^\top D_n^{-1} F D_p^{-1} u, \quad \text{s.c.} \quad u^\top D_p^{-1} u = 1.$$

This optimization problem is equivalent to finding the first eigenvector of the matrix

$$S = F^\top D_n^{-1} F D_p^{-1}.$$

The matrix S plays a role analogous to the covariance matrix in PCA. The first eigenvector u_1 therefore satisfies the relation

$$S u_1 = F^\top D_n^{-1} F D_p^{-1} u_1 = \lambda_1 u_1,$$

where λ_1 is the eigenvalue associated with u_1. The eigenvectors of the matrix S give the factorial axes in the column space. The coordinates of the line profiles on the first factorial axis are obtained by the relation

$$\Phi_1 = D_n^{-1} F D_p^{-1} u_1.$$

The other pairs of eigenvalues and eigenvectors, as well as the coordinates of the line profiles on the associated factorial axes, are obtained in a similar manner.

The dual analysis is performed in a similar way. We seek the first eigenvector of the matrix

$$T = F D_p^{-1} F^\top D_n^{-1}.$$

The first eigenvector v_1 therefore satisfies the relation

$$T v_1 = F D_p^{-1} F^\top D_n^{-1} v_1 = \mu_1 v_1,$$

where μ_1 is the eigenvalue associated with v_1. The eigenvectors of the matrix T give the factorial axes in the row space. The coordinates of the column profiles on the first factorial axis are obtained by the relation

$$\Psi_1 = D_p^{-1} F^\top D_n^{-1} v_1.$$

The other pairs of eigenvalues and eigenvectors, as well as the coordinates of the column profiles on the associated factorial axes, are obtained in a similar manner.

Property

The matrices S and T have the same $r = \min(n-1, p-1)$ first positive eigenvalues. This guarantees a consistent representation of the rows and columns in the same reduced space. For $k = 1, \ldots, r$, the relationships between the eigenvectors u_k and v_k are

$$u_k = \frac{1}{\sqrt{\lambda_k}} F^\top D_n^{-1} v_k \quad \text{and} \quad v_k = \frac{1}{\sqrt{\lambda_k}} F D_p^{-1} u_k.$$

Proof

Starting from the equation

$$Tv_1 = FD_p^{-1}F^\top D_n^{-1}v_1 = \mu_1 v_1,$$

multiplying on the left by $F^\top D_n^{-1}$, we obtain:

$$F^\top D_n^{-1}FD_p^{-1}F^\top D_n^{-1}v_1 = \mu_1 F^\top D_n^{-1}v_1.$$

Thus, the vector $F^\top D_n^{-1}v_1$ is an eigenvector of the matrix $F^\top D_n^{-1}FD_p^{-1}$ associated with the eigenvalue μ_1. Since λ_1 is the largest eigenvalue of $F^\top D_n^{-1}FD_p^{-1}$, we can deduce that $\mu_1 \leq \lambda_1$. Proceeding in the same way, starting from $Su_1 = \lambda_1 u_1$, we deduce that $\lambda_1 \leq \mu_1$. Therefore, $\lambda_1 = \mu_1$. We can then do the same for the first r eigenvalues. We can also deduce the relationships between the eigenvalues.

Note

By centering the profiles, we can project the line profiles and column profiles onto the same coordinate system, thus facilitating joint geometric interpretation.

Center of gravity and inertia

In statistical software outputs, the scatter plot generated by an AFC is generally centered at (0, 0). This convention reflects an analysis relative to the centers of gravity of the line profiles and column profiles. This centering is both practical and interpretable. It shows the distances between the modalities relative to their weighted mean, i.e., relative to the average behavior in the population.

Each modality (row or column) is associated with a weight corresponding to its marginal frequency: the weight of the ith row is $f_{i\bullet}$ and the weight of the jth column is $f_{\bullet j}$. The center of gravity of the rows is the weighted average of the row profiles:

$$G_L = \left(g_1, \dots, g_p\right)^\top, \quad \text{where} \quad g_j = \sum_{i=1}^{n} f_{i\bullet}\frac{f_{ij}}{f_{i\bullet}} = \sum_{i=1}^{n} f_{ij} = f_{\bullet j}, j \in \{1, \dots, p\}.$$

Similarly, the center of gravity of the columns is

$$G_C = \left(f_{1\bullet}, \dots, f_{n\bullet}\right)^\top.$$

To re-center the profiles around the center of gravity, we subtract their mean value:

$$\frac{f_{ij}}{f_{i\bullet}} - g_j = \frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} = \frac{f_{ij} - f_{i\bullet}f_{\bullet j}}{f_{i\bullet}}.$$

This centering ensures that each line profile $i \in \{1, \dots, n\}$ is averaged to zero:

$$\sum_{j=1}^{p} \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}} = 0.$$

AFC is therefore no longer performed on matrix S but rather on a centered matrix $S^\star = (s^\star_{jj'})$, where

$$s^\star_{jj'} = \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)\left(f_{ij'} - f_{i\bullet} f_{\bullet j'}\right)}{f_{i\bullet} f_{\bullet j'}}.$$

By definition, the trace of the matrix $S^\star$ gives the total inertia:

$$\operatorname{tr}(S^\star) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)^2}{f_{i\bullet} f_{\bullet j}}.$$

This corresponds to the normalized χ^2 statistic used to test the independence between variables.

Property

We have that, for all $j, j' \in \{1, \dots, p\}$, $s^\star_{jj'} = s_{jj'} - f_{\bullet j}$, where

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i\bullet} f_{\bullet j'}}.$$

Proof

The previous property implies that the matrices S and $S^\star$ have the same eigenvectors for the first p dimensions, which allows us to perform factorial analysis on the centered version.

5 Factor coordinates

We have that, for all $k = 1, \dots, r$,

$$\Phi_k = D_n^{-1} F D_p^{-1} u_k \quad \text{and} \quad \Psi_k = D_p^{-1} F^\top D_n^{-1} v_k.$$

However, we have also seen the relationships between the eigenvectors u_k and v_k,

$$u_k = \frac{1}{\sqrt{\lambda_k}} F^\top D_n^{-1} v_k \quad \text{and} \quad v_k = \frac{1}{\sqrt{\lambda_k}} F D_p^{-1} u_k.$$

We can therefore deduce the relationships between the factorial coordinates of the row profiles and those of the column profiles:

$$\Phi_k = \frac{1}{\sqrt{\lambda_k}} D_n^{-1} F \Psi_k \quad \text{and} \quad \Psi_k = \frac{1}{\sqrt{\lambda_k}} D_p^{-1} F^\top \Phi_k.$$

We can now examine these relationships on each of the components:

$$[\Phi_k]_i = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\bullet}} [\Psi_k]_j \quad \text{and} \quad [\Psi_k]_j = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\bullet j}} [\Phi_k]_i ,$$

where $[\Phi_k]_i$ denotes the coordinate of the line profile L_i on the kth factorial axis and $[\Psi_k]_j$ denotes the coordinate of the column profile C_j on the same factorial axis. These relations express, to within a factor of $1/\sqrt{\lambda_k}$, that each line profile is at the barycenter of the projections of the column profiles assigned the weight of column j in line i and that each column profile is at the barycenter of the projections of the line profiles assigned the weight of line i in column j.

> Note
>
> Thus, in AFC, we have a double barycentric representation. On the factorial axes, each point in one cloud is at the barycenter of the points in the other cloud.