Data learning 2 : Correspondence Analysis

Contingency tables : The datas we are interested in are now categorical datas. Precisely we considere two variables X and Y that have been measured on n individuals of a given population, each of them having a finite number of modalities $X \in \{x_1, x_2, \ldots, x_r\}$ and $Y \in \{y_1, y_2, \ldots, y_c\}$. We first fill in a $r \times c$ table T with the numbers of individuals $(n_{lh})_{1 \leq l \leq r, 1 \leq h \leq c}$ for which $X = x_l$ and $Y = y_h$. Let n_{l+} and n_{+h} denote the sum of rows and columns respectively and let n be the size of the population. Such a table is called a *contingency table :*

	y_1	...	y_h	...	y_c	sum
x_1	n_{11}	...	n_{1h}	...	n_{1c}	n_{1+}
...	...	...	...	...	...	...
x_l	n_{l1}	...	n_{lh}	...	n_{lc}	n_{l+}
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rh}	...	n_{rc}	n_{r+}
sum	n_{+1}	...	n_{+h}	...	n_{+c}	n

Notice that the table $\frac{1}{n}T$ consists in the frequencies (f_{lh}) of each pairs of modalities (x_l, y_h). By analogy to the case of random variables in probability, we will say that the two variables X and Y are *independant* when, for all $1 \leq l \leq r$ and $1 \leq h \leq c$, one has $n_{lh} = n_{l+} \cdot n_{+h}$. The principal objectif of correspondence Analysis is to study the nature of the relation between the two variables when they are not independant.

Row profiles and column profiles : To study how much the distribution of each modalities of Y among the population depends on the values of X, we would like compare one row with another. But each row having a different number of respondents, n_{l+}, we must first reduce each of them to the same base. We define, for each $l = 1, \ldots, r$, the *row profile* :

$$\left\{ \frac{n_{l1}}{n_{l+}}, \ldots, \frac{n_{lc}}{n_{l+}} \right\}$$

Each row profile corresponds to a point in an c-dimensional vector space called the *space of row profiles*. The table of all row profiles form a $c \times r$ matrix A that can be written $X_r = \frac{1}{n} D_r^{-1} T$, where $D_r = diag(n_{1+}, \ldots, n_{r+})$.

In a symetric fashion, we define the *column profile* :

$$\left\{ \frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}} \right\},$$

the r-dimensional *space of column profiles* and the table of column profiles that form a $r \times c$ matrix X_c that can be written $X_c = \frac{1}{n} T D_c^{-1}$, where $D_c = diag(n_{+1}, \ldots, n_{+c})$.

In such a way we have define two clouds of points, one is the set of row profiles in the space of row profiles $\mathbb{R}^c$ and the second is the set of column profiles in the space of column profiles $\mathbb{R}^r$.

Chi-square metric : In order to measure the distance between profiles, let simply denote by i, i', ...two row profiles and by j, j', ...two column profiles and considere the following distance in $\mathbb{R}^c$ between two row profiles

$$d(i, i') = \sum_{j=1}^{c} \frac{1}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}} \right)^2$$

and, in a symetric fashion, the following distance in $\mathbb{R}^r$ between two column profiles :

$$d(j, j') = \sum_{i=1}^{r} \frac{1}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij'}}{n_{+j'}} \right)^2.$$

These distances are called the *Chi-square distance*. Their matrices are respectively $\mathcal{M} = D_c^{-1}$ and $\mathcal{M} = D_r^{-1}$.

The Chi-square distance differs from the usual Euclidien distance in that each square is weighted by the inverse of the number of respondents corresponding to each term. If no such standardization were performed, the differences between larger proportions would be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped. The weighting factors are used to equalize these differences.

Another important reason for choosing the Chi-square metric is that it satisfies the principle of *distributional equivalence*, expressed as follows : if two row profiles are proportioned and if they are replaced by only one, which is the sum, column by column, then the distances between columns are not changed in the set of column. And its true also in a symetric fashion for column profiles.

Double PCA : Considering row profiles as r individuals of a population and its coordinates as the values of c variables, one can perform a Principal Component Analysis of the row profiles. The matrix X_r plays the role of the $r \times c$ data matrix, the metrics in the space of individuals being the Chi-square metric $M = D_c^{-1}$ and the diagonal matrix of the weights $D = D_r$. Then the new basis of the space, the principal components, are the eigenvectors of the matrix $A_c = X_r' D_r X_r D_c^{-1}$ which is simply equal to $A_c = X_r' D_r D_r^{-1} \frac{1}{n} T D_c^{-1} = X_r' X_c$. And as usually in a PCA, the eigenvectors are ordered from the one corresponding to the heighest eigenvalue to the one corresponding to the smallest one.

On the other hand we can also considere the column profiles as c individuals of a population and its coordinates as values of r variables and than perform a second Principal Component Analysis. The matrix X_c' will be the $r \times c$ data matrix, the metrics in the space of individuals is the Chi-square metrix $M = D_r^{-1}$ and the diagonal matrix of the weights $D = D_c$. The principal components are the eigenvectors of the matrix $A_r = X_c D_c X_c' D_r^{-1}$ which is simply equal to $A_r = X_c D_c D_c^{-1} \frac{1}{n} T' D_r^{-1} = X_c X_r'$.

Now it is clear that the two matrices $X_r' X_c$ and $X_c X_r'$ have the same non zero eigenvalues, with the same multiplicities, and if u is an eigenvector of $X_c X_r'$ associated with the eigenvalue λ than $v = X_r' u$ is an eigenvector of $X_r' X_c$ associated with the same eigenvalue. Moreover if u have been chosen such that $\|u\|_r = 1$ (where $\|u\|_r = u' D_r^{-1} u$ is the norme in the space of row profiles), than $\|X_r' u\|_c = \lambda$ (where $\|v\|_c = v' D_c^{-1} v$ is the norme in the space of column profiles). Indeed, using the definition of X_r, we have

$$\|X_r' u\|_c = u' X_r D_c^{-1} X_r' u = u' D_r^{-1} \frac{1}{n} T D_c^{-1} X_r' u = u' D_r^{-1} X_c X_r' u = u' D_r^{-1} \lambda u = \lambda \|u\|_r.$$

Representations and quality : To each of the PCA, corresponds a representation of the cloud of profiles in the plan of the two first PC and it can be usefull to study these two representations independantly. But it is customary in correspondent analysis to summarize the row and column coordinates in a single plot, as will see now. First one can plot each point of one cloud among the points of the other cloud in the following way : suppose we have plotted all the points of one cloud in the plan of their two first PC, than each profile belonging to the other cloud can be considered as the barycentric coordinates of one point in the simplex of the existing cloud. And the same is true if one exchanges rows and columns. But on the face of it, to superimpose one plot on the other is impossible because, as barycenters, each set of points should be inside the other set of points. Nevertheless using the previously mentionned relation between the two family of PC, a join display of row and column profiles can be obtained after rescalling each axis. However it is important to remember that in such a join display one can interprete the distance between row profiles, the distance between column profiles or the relative positions of one point of one set with respect to all the points of the other set. Except in special cases, the proximity of two points corresponding to different sets of points has no meaning.

To interprete the quality of a correspondent analysis three sets of coefficients may be calculated for each axis and these coefficients apply equally to the row and column of the data matrix.

- *Percentage of inertia :* as in PCA, the contribution of an axis to the totale inertia of the cloud is given by the ratio of the corresponding eigenvalue over the sum of all eigenvalues. This percentage is used to select the number of significant axis.
- *Contribution of one variable to the inertia of a factorial axis :* these partial contribution allows to determine which points play a major role in the orientation of a given factorial axis.
- *Squared cosine :* the quality of the representation in the plan of one point is defined as the ratio of the squared distance of the point from the origin in the plan over the squared distance from the origin in the original space of profiles. It is equal to the squared cosine. A small value (low quality) means that the two first principal dimensions does not represent well the respective point.