Multiple correspondence analysis

Multiple correspondence analysis (MCA) can be presented as an extension of PCA. It allows for the graphical representation of frequency tables containing more than two variables. A classic example of a frequency table with more than two qualitative variables is a table presenting individuals' responses to a questionnaire containing Q multiple-choice questions. MCA is therefore very useful for visualizing the results of a questionnaire-based study.

MCA can also be seen as a version of PCA when the variables are mixed, i.e., comprising both quantitative and qualitative variables. The joint processing of these two types of data is based on their prior transformation called **complete disjunctive coding**.

1 Notation

Let n be the number of individuals (or observations) and Q the number of variables (or questions in the case of a questionnaire). Each variable has J_q modalities and the total number of modalities is equal to J.

> Definition
>
> The binary table, i.e., containing only 0s and 1s, with n rows and J columns is called a **complete disjointive coding table**. We denote it by Z.

Thus, a variable is not treated as such but through its modalities. It is divided into modalities, and each individual is then coded 1 for the modality they possess and 0 for the others (i.e., those they do not possess, as the modalities are exclusive). This coding is immediate for qualitative variables. However, for a qualitative variable, we first divide the variable into classes. Thus, each individual belongs to only one class. This process of transforming information is called **complete disjunctive coding**. It is indeed coding, because the initial information is transformed; disjunctive, because each individual has at most one modality; and complete, because each individual has at least one modality.

Example

Let's take, for example, a set of products with different types (Hoodie, Joggers, and Sneakers) and different prices. We have 7 observations (7 products). The variable `Type` is a qualitative variable and the variable `Price` is a quantitative variable.

Table 1: Product dataset.

Product	Type	Price ($)
Nike Tech Fleece	Hoodie	256.72
Puma Joggers	Joggers	221.26
Off-White Hoodie	Hoodie	198.45
Supreme Hoodie	Hoodie	235.50
Jordan 1 High	Sneakers	298.22
Nike Dunk Low	Sneakers	273.00
Nike Tech Fleece	Hoodie	162.38

To encode the information in a complete disjunctive coding table, we define three price classes (price below $200, price between $200 and $250, and price above $250). This allows us to encode the variable `Price` using the above classes. The complete disjunctive table is given in the following table.
Product | Hoodie | Joggers | Sneakers | $<$200 | between $200 and $250 | $>$250 |
|———:|:—— -:|:——–:|:———:|:——:|:-:|:-:| |Nike Tech Fleece |1 | 0 | 0 | 0 | 0 | 1 | |Puma Joggers |0 | 1 | 0 | 0 | 1 | 1 | |Off-White Hoodie |1 | 0 | 0 | 1 | 0 | 0 | |Supreme Hoodie |1 | 0 | 0 | 0 | 1 | 0 | |Jordan 1 High |0 | 0 | 1 | 0 | 0 | 1 | |Nike Dunk Low |0 | 0 | 1 | 0 | 0 | 1 | |Nike Tech Fleece |1 | 0 | 0 | 1 | 0 | 0 |
: Product dataset in complete disjunctive coding.

Note

When transforming quantitative variables into a complete disjunctive coding table, information is lost. This is because qualitative variables must be divided into classes, and class membership is less informative than a specific variable value. In the previous example, information about price is lost.

Properties

1. The sum of the elements in the same row is constant and equals Q.
2. The sum of all the elements in the table equals nQ.
3. The sum of the elements in the same column equals the count n_j with modality j of

variable q.

Proof

Since Z is a complete disjunctive array, we have

$$\sum_{j=1}^{J_q} z_{ij} = 1.$$

Therefore, we find that

$$z_{i\bullet} = \sum_{j=1}^{J} z_{ij} = \sum_{q=1}^{Q} \sum_{j=1}^{J_q} z_{ij} = Q,$$

$$z_{\bullet j} = \sum_{i=1}^{n} z_{ij} = n_j,$$

$$z_{\bullet\bullet} = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} = nQ.$$

2 Burt's Table

Definition

The **Burt's table**, denoted B, is the product of the transpose of Z by Z:

$$B = Z^\top Z.$$

Properties of B

1. Burt's table is square and its size is equal to the total number of modalities J possessed by the Q variables.
2. The diagonal blocks of B are themselves diagonal matrices. They are given by $B_{qq} = Z_q^\top Z_q$ and their diagonal elements correspond to the frequency of each modality for variable q.
3. The non-diagonal blocks of B are given by $B_{qq'} = Z_q^\top Z_{q'}, q \neq q'$. They correspond to the contingency tables crossing the variables q and q'.
4. Burt's table is symmetric because $B_{q'q} = Z_{q'}^\top Z_q$ is the transpose of $B_{qq'} = Z_q^\top Z_{q'}$.

From a mathematical point of view, ACM is an AFC performed on the logical matrix Z or on Burt's table B. It can be shown that the same factors are obtained, regardless of the matrix used for the analysis.

3 Eigenelements of the table Z

The eigenelements of the table Z can be calculated using the same method as for PCA. By analogy with PCA, we therefore seek the eigenvectors of the matrix

$$S = \frac{1}{Q} Z^\top Z D_J^{-1},$$

where D_J is the diagonal matrix of term $n_j, j = 1, \dots, J$. We can calculate the coordinates of the line profiles on the factorial axes in the same way:

$$\Phi_k = nZD_J^{-1}u_k,$$

where u_k is the kth eigenvector associated with the eigenvalue λ_k of the matrix S.

We can also look at the dual analysis of the table Z. Again, by analogy with PCA, we look for the eigenvectors of the matrix

$$T = \frac{1}{Q} Z D_J^{-1} Z^\top.$$

Similarly, we can calculate the coordinates of the column profiles on the factorial axes:

$$\Psi_k = nD_J^{-1} Z^\top v_k.$$

4 Eigenelements of Burt's table B

Since Burt's table is symmetric, the direct analysis and the dual analysis coincide. We can also analyze it in analogy with PCA. The sum of the elements of the same row (or the same column) of B is Qn_j and the sum of the elements of B is nQ^2. We are looking for the eigenvectors of the matrix

$$S' = \frac{1}{Q^2} B^\top D_J^{-1} B D_J^{-1}.$$

We then notice that this matrix S' has the same eigenvectors as the matrix S. Indeed,

$$S' = \frac{1}{Q^2} B^\top D_J^{-1} B D_J^{-1} = \frac{1}{Q^2} Z^\top Z D_J^{-1} Z^\top Z D_J^{-1}.$$

And let u and λ satisfy $Z^\top Z D_J^{-1} = \lambda u$, then

$$Z^\top Z D_J^{-1} Z^\top Z D_J^{-1} u = Z^\top Z D_J^{-1} \lambda u = \lambda^2 u.$$

Finally, the analysis of Z or B provides the same eigenvectors, and for all $k = 1, \dots, Q$, the kth eigenvalue of B is the square of the kth eigenvalue of Z.

5 Variable encoding

Variable encoding, and in particular the choice of class boundaries, is essential in ACM. For continuous variables, the boundaries should be relevant to the problem being studied. For example, we would not define a class > 1000 in the previous example. To obtain relevant bounds, we can look at the distributions of the variables, e.g., with a histogram. In some specific cases, it is possible to divide the variable into equal-sized categories. However, this approach can lead to irrelevant categories.

In the case of qualitative variables, the choice of classes does not arise; it is given by the variable. However, "natural" modalities can lead to (very) unbalanced frequencies. In this case, we generally need to proceed with groupings. Here again, a good knowledge of the field being studied is necessary. In any case, it is preferable to group modalities rather than randomly distribute modalities with low frequencies among the other modalities (which is sometimes proposed in software).