Session
Interpretation rules

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System tracking

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu lidia.montero@upc.edu

What does it mean "to interpret the results":

- to interpret is to make clear

  The principal axes methods provide fewer results than the original data, but unclear in terms of these. Thus, the results have to be translated in terms of the user's data. The most striking structures of the table have to be selected.

- to interpret is to make sense

  We have to integrate data and results within their context: the whole table, the supplementary elements, the global experience and knowledge of the analyst.

- to interpret is to bring a personal touch

  Like a piece of music is interpreted with one's own style, the interpretation of the results should be conducted according to the analyst's own style through many choices concerning the synthesis, the wording of the axes, the selection of the most salient facts, and so on.

Interpretation process. Standardized PCA (I)

- inertia of the axes

  eigenvalues: the first one is always between 1 and *K* (number of variables); exactly 1 when all the variables are uncorrelated, exactly *K* when a linear relationship exists between all the variables. All the higher is an eigenvalue, more interesting is the factor in term of synthesis of the original variables

  Diagram/chart of eigenvalues

  Proportion of inertia extracted by the factors

  How many factors to retain?

Interpretation process. Standardized PCA (II)

- Interpretation of the factors

  Factors are studied in the order of the associated eigenvalues. One by one or by pairs. Take care: factor *s*, with s>1, accounts for the residual trends, not taken into account by the former factors

  a. Individuals contribution. "General nature" (or not) of the factor

  b. Coordinates of the active variables

     Interpretation axis by axis

     Interpretation plane by plane

  c. Coordinates of the supplementary variables

  d. Coordinates and aids to interpretation of the active individuals

  e. Coordinates and aids to interpretation of the suppl. individuals

Interpretation process. CA (I)

Even if in CA rows and columns play a symmetric role, the interpretation steps of CA are globally similar to those of PCA. In the following, we detail some specific aspects.

- inertia of the axes

  eigenvalues: the eigenvalues are inferior or equal to 1, value observed when an axis gives perfect account of the association between a partition of the rows and a partition of the columns. Thus, a factor associated to an eigenvalue close to 1 expresses a strong relationship between rows and columns that will be easily transposed in terms of the initial data. On the contrary, a weak eigenvalue (about 0.1) corresponds to a weak relationship between the variables and should be interpreted with caution, "going back" to the data.

  Eigenvalues and proportion of inertia are independent information that is iuseful to consult to decide about the interest of each factor.

Interpretation process. CA (II)

- Contributions of rows and columns

  Like in PCA, it is important to ensure that sufficient elements contribute to the first factors. The process is the same than in ACP, unlike that rows and columns are concerned.

  However. in CA applied to a contingency table, factors built from a very small number of elements are more embarrassing than in PCA: if some rows (columns) are deleted, the relationship between both variables is studied through only a subset of their categories! It should be strongly justified from the nature of the problem and data.

Interpretation process. CA (III)

- Coordinates of the active elements

  As in PCA, axes and planes are successively studied. Usually, an axis is interpreted from the point of view of a set (rows, for example) , then from the point of view of the other.

  We must keep in mind that the elements are not uniformly weighted. Thus, coordinates, contributions and representation quality constitute different items of information.

  The interpretation of a factor mainly relies on the "typical" elements that present:

  - a strong contribution; if deleted the factor would disappear
  - extreme coordinates and a high quality of representation
  - extreme coordinates and medium quality of representation

- Ordinal categories

  When there is an a priori structure, such as ordinal categories, the factors that give account of this structure have to be looked for and studied

Interpretation process. MCA (I)

Essentially MCA, as PCA, is applied to a table Individuals × Variables although, from one technique to the other, the nature of the variables is different. However, computing is performed through CA applied to the complete disjunctive table. For that, the interpretation rules share aspects with PCA but also with CA

- inertia of the axes

The sum of eigenvalues is equal to (K/J)-1. It does not depend on the data structure, such as in PCA. The eigenvalues are inferior or equal to 1, as in CA: Usually, in MCA, the eigenvalues slowly decrease. The general aspect of the histogram of the eigenvalues is not very suggestive.

The eigenvalue associated to a factor is equal to the mean of the ratios of correlations between the factor and each active variable. It is equal to 1 if all these ratios are equal to 1, that is, if for each variable all the individuals belonging to the same category are placed on the same point. In practice, we are very far from this “extreme” situation: as a consequence, the eigenvalues are usually very weak in MCA.

Interpretation process. MCA (II)

- proportion of inertia

  Due to the nature of the variables, the proportions of inertia associated to the fist axis are usually very weak.

- General remark on eigenvalues and proportion of inertia

  In MCA the eigenvalues and proportions of inertia have little influence on the interpretation of MCA results

Interpretation process. MCA (III)

- Contributions of individuals and categories

In order to identify possible outliers, the study begins by checking the contribution s of the individuals. As in PCA; the variables cannot be outliers, but in some cases, the first axes are due to a small number of categories. This is possible when there are categories with low counts shared by the same individuals.

When the checking of the contributions of the categories shows that a small number of categories is clearly overriding, then the individuals that belong to these categories generally present a very high contribution. The factor is due to a too small number of individuals.

A solution could be to declare these individuals as supplementary.

Interpretation process. MCA (IV)

- Contributions of the variables

  By summing, on the rank *s* factor, the contributions of the categories of a same variable, the contribution of this variable is obtained.

  Except for the constant $J\lambda_s$, this contribution is equal to the ratio of correlation between the variable and the factor. As a result:

  • By ranking the variables depending on their contribution (in decreasing order), it is easy to select the variables more associated to a factor. The interpretation will be mainly supported by these variables.

  • The graphic of the variables placed on the axis by their contribution to the corresponding factor eases the selection of the variables when interpreting a factor.

  • It is interesting to compute the ratio of correlation with the axes of the supplementary variables.

Interpretation process. MCA (V)

- coordinates of categories and individuals

The study of the coordinates of the categories generally precedes the study of the coordinates of the individuals.

The strategy which consists in studying first the active elements, then the supplementary and finally the planes is similar to the other analyses.

The case of ordinal categories is frequent in MCA practice. So, we have to begin, when studying the coordinates, by identifying the factors on which the categories of the ordinal variables are placed in their natural order. In this case, the categories have to be linked on the graphic by a line.

The quality of representation of the categories is not a pertinent indicator. As the categories of a same variable are orthogonal, they cannot be simultaneously well represented on an axis.

In the case of the individuals, the strategy is similar as this used in PCA.