Missing values in categorization

Marine Cadoret[1], Sébastien Lê[1], Jérôme Pagès[1]

[1] *Agrocampus Ouest, Laboratoire de mathématiques appliquées,*
65 rue de Saint Brieuc, CS 84215, 35042 Rennes, France.
E-mail: marine.cadoret@agrocampus-ouest.fr

Abstract: Categorization is a task in which subjects are asked to group a set of objects according to their similarities. This method is increasingly used but appears to be difficult when the number of items to categorize is large and, as it is the case in sensory analysis, when the five senses of the subjects are mobilized.
The aim of this communication is twofold: first, to circumvent this difficulty by presenting only a subset of objects to each subject, de facto data will contain missing values; second, to present an approach based on multiple correspondence analysis (MCA) to analyze categorization data with missing values.
We will show how to balance the absence of each object and each pair of objects for a given subset according to an experimental design. We will also show how to take into account missing values in MCA: they can be taken into account by adding a new group to each subject, the group of the objects that haven't been presented.
We will illustrate our method with real data in two ways: on the one hand, missing values will be simulated from a complete data set of 12 luxury perfumes; on the other hand, missing values will be obtained by presenting subsets of perfumes out of the 12 fragrances. The results obtained with and without missing values will be compared with multiple factor analysis (MFA).
Keywords: categorization data, multiple correspondence analysis, missing values

1 Introduction

Categorization is a process in which *J* subjects are asked to group *I* objects according to their resemblances: all objects are presented simultaneously to subjects. This task becomes more and more difficult for the subject as the number of objects to group increases; this is particularly true in sensory analysis where subjects use their senses to categorize products. Hence the idea of presenting only part of the objects to each subject.

We propose an approach to analyze this type of incomplete data. Results will be compared with the ones obtained with complete data: it will be illustrated with data collected on 12 luxury perfumes.

2 Method

a. Selecting the subset of objects

Each of the *J* subjects evaluates *p* objects among *I*. To choose which objects will be evaluated by whom, at least two strategies are possible: randomly or according to a balanced design. We propose to use a balanced incomplete block design (BIBD) or an optimal design when a BIBD can't be obtained due to the improper number of objects; i.e. a design that has similar properties to the BIBD (Pagès and Périnel, 2007). In the case of BIBD, the

presence of each object and of each pair of objects is balanced: every object is evaluated by the same number of subjects denoted *r* and every pair of objects is presented the same number of times denoted λ.

b. Analysis of incomplete categorization data

Categorization data are usually gathered in a co-occurrences matrix *C* of dimension $I \times I$ where the general term *c(i,l)* corresponds to the number of subjects that have put the objects *i* and *l* in the same group. This co-occurrences data table can be transformed into a matrix of distances denoted *D*, where the general term *d(i,l)* corresponds to the number of subjects that haven't put *i* and *l* in the same group. This kind of data table is generally analyzed by non-metric multidimensional scaling (MDS): Lawless, 1989, Lawless *et al.,* 1995, Faye *et al.,* 2004. From our point of view it is more natural to respect the co-occurrences and not only the ranks.

These data can also be gathered in a data table of dimension $I \times J$ in which each row *i* corresponds to an object, each column *j* corresponds to a subject, and a cell *(i,j)* corresponds to the number of the group to which object *i* belongs to for subject *j*. Each column of the table can be assimilated to a qualitative variable with K_j categories, where K_j denotes the number of groups used by subject *j*. Such data table can then be analyzed properly by MCA.

This second way of gathering the data has several advantages: it is possible to associate to the representation of the items a representation of the subjects linked to the previous one, to build confidence ellipses around the objects, etc (Cadoret *et al.*, 2008). This approach also allows taking into account incomplete data in a natural way by creating an additional group for each subject: this group is constituted of the objects which were not presented to him.

With MCA, categorization data are taken into account via the complete disjunctive table, denoted *X*, of dimension $I \times K$ $(K=\sum K_j)$, of general term x_{ik} where x_{ik} is equal to 1 if object *i* belongs to group *k* and 0 if not. The distance between two objects *i* and *l* in MCA is thus defined (where I_k denotes the number of objects in the group *k*):

$$d^2(i,l) = \frac{1}{J}\sum_k \frac{I}{I_k}\left(x_{ik} - x_{lk}\right)^2$$

In the case of BIBD, by splitting the disjunctive table into 2 sub-tables (groups referring to non-presented objects and others), the distance between 2 objects *i* and *l* is thus defined:

$$d^2(i,l) = \frac{1}{J}\sum_{k'=1}^{K'} \frac{I}{I_{k'}}\left(x_{ik'} - x_{lk'}\right)^2 + \frac{1}{J}\frac{I}{I-p}2\left(r-\lambda\right)$$

$$= d^2_{K'}(i,l) + \frac{1}{J}\frac{I}{I-p}2\left(r-\lambda\right)$$

where *I-p* corresponds to the number of non-presented objects, $r-\lambda$ to the number of subjects for which object *i* (resp. *l*) is missing and not object *l*

(resp. i) and K' to the number of groups really constituted by the subjects (not referring to group with missing objects).

Whereas in MCA in the case of complete data, two objects are superimposed if they were placed in the same group by all subjects, in the case of incomplete data resulting from BIBD, two objects cannot be superimposed because for r-λ subjects one of two objects is missing and these two objects belong necessarily to different groups. In this case, the distance between two objects is strictly higher than 0.

In the particular case where an object was isolated by all subjects, MCA highlights this particular object on the first dimension associated with an eigenvalue of 1. When the incomplete data result from a BIBD, this case is not possible anymore because all the non-presented objects are considered as if they belong to the same group. By construction of the BIBD, each object is missing for J-r subjects and then belongs to a group of I-p objects: no object can be isolated by all subjects.

Thus this way of taking into account the incomplete data resulting from a BIBD reduces the extreme distances (small and large).

c. Evaluation of the method

The evaluation of the method is done in two steps. As a first step, we evaluate the choice of a BIBD compared to a random design. To do so, incomplete data sets are simulated from a complete data set. In a second step, we compare the results obtained from an incomplete data set with those obtained from a complete one. To do so, let's stress that the use of a real incomplete data set is absolutely necessary. Indeed, a simulated incomplete data set resulting from a complete data set would give optimistic results since in that case the groups of objects are fixed once for all and consequently the subject would have constituted the same groups if a subset of objects had been presented to him.

d. Comparison

Firstly, the comparison between the results of the MCA obtained with and without missing values is done by calculating the correlation coefficients between the factors of the two analyzes. However these dimensions being defined for a rotation, it is also necessary to calculate the RV coefficient between the two configurations.

The results can also be compared thanks to a multiple factor analysis (MFA) (Escofier and Pagès, 1998) carried out on the two tables corresponding to the complete data set and the incomplete data set. MFA provides several results and helps for the interpretation. From these results, we will keep only the representation of the objects according to the two data sets as well as the superimposed representation of the objects according to each data set.

3 Data

To validate this approach, we carried out several experiments on 12 luxury perfumes: Angel, Aromatics Elixir, Chanel n°5, Cinéma, Coco Ma-

demoiselle, L'instant, Lolita Lempicka, Pleasures, Pure Poison, Shalimar, J'adore (eau de parfum) and J'adore (eau de toilette).

98 subjects carried out a categorization on these 12 perfumes: this data set corresponds to the complete one. From this first data set, we will simulate incomplete data sets. In this data set, Shalimar is the most often isolated perfume (by 24 subjects) and the 2 J'adore are the most associated perfumes (by 56 subjects).

We also have a data set in which 42 subjects carried out a categorization on 8 perfumes among the 12 previous. The perfumes that have not been evaluated were selected based on an optimal design. This data set corresponds to the real incomplete data set. In this data set, no perfume was isolated by all subjects and no perfumes were grouped together by all subjects: Angel is the most often isolated perfume (by 7 subjects) and the 2 J'adore are the most associated (by 17 subjects).

4 Results

a. Simulated incomplete data set

In order to measure the interest of using a balanced design, we simulated incomplete data sets according to both random designs and optimal designs while varying the number of subjects (30, 60 and 98) and the number of products (6, 8 and 10). For each subject-product combination, 100 optimal designs and 100 random designs were generated; an MCA was carried out on each simulated data set. To measure the proximity of the results obtained on the complete data set and the simulated data sets, we calculated the RV coefficients between the first two dimensions of each analysis. Table 1 summarizes all the results.

Table 1. Average RV coefficients (over 100 simulations) between factors 1 and 2 of the MCA performed on the complete data set and the factors 1 and 2 of the MCA with 6, 8 and 10 perfumes presented according to an optimal design and a random design

	6 perfumes		8 perfumes		10 perfumes	
	BIBD	random	BIBD	random	BIBD	random
30 subjects	0.579	0.471	0.746	0.652	0.781	0.76
60 subjects	0.758	0.621	0.869	0.77	0.942	0.92
98 subjects	0.855	0.721	0.94	0.88	0.98	0.97

The average RV coefficient increases with on the one hand the number of products, which is explained by the fact that data are less and less incomplete and on the other hand with the number of subjects, which is explained by the fact that the complete data set is composed of the 98 subjects.

Moreover, in all the cases, the average RV coefficients is significantly higher with an optimal design (except for the combination 30 subjects-10 products where the difference is not significant), which comfort us to use BIBD.

b. Real incomplete data set

To validate the choice of a BIBD to obtain incomplete data set, we will compare more precisely the results of the MCA with those obtained with the complete data set. The high correlation coefficients between the factors 1 and the factors 2 of the two MCA indicate a similarity between the dimensions of a same rank for those two analyzes (cf. table 2).

Table 2. Correlation coefficients between the factors 1 and 2 of the MCA with the complete and the incomplete data set

		Incomplete	
		F_1	F_2
Complete	F_1	**0.915**	0.289
	F_2	0.302	**-0.745**

The RV coefficient of 0.815 between the configurations obtained with 2 dimensions of the two MCA shows proximity between the first factorial plane of each analysis.

Figure 1 corresponds to the results of the MFA carried out on the complete data set and the incomplete one. For all the perfumes, these two representations are similar. The first axis of the MFA opposes Shalimar, Aromatics Elixir and Chanel n°5 to the others and the second axis opposes Angel, Lolita Lempicka and Cinéma to the others. Figure 1 also represents confidence ellipses around the products for the two experiments (at a 95% level). The ellipses have a surface more important in the case of incomplete data set and they overlap with the ellipses associated with the complete data set. This overlapping indicates a similar perception of the products during the two experiments. Nevertheless the observed differences can be explained by the data: Figure 1 suggests that Angel was perceived as more particular during the experimentation with 8 perfumes than during the one with all the perfumes. This can be confirmed by the data since Angel was isolated by 7 subjects out of 42 during the experimentation with the incomplete data set versus 13 out of 98 during the experimentation with the complete data. The figure also suggests that J'adore (EP) and Pleasures were more differentiated during the experimentation with incomplete data set, which can also be confirmed with the raw data: these 2 perfumes were grouped by only 13 subjects out of 42 versus 38 out of 98.

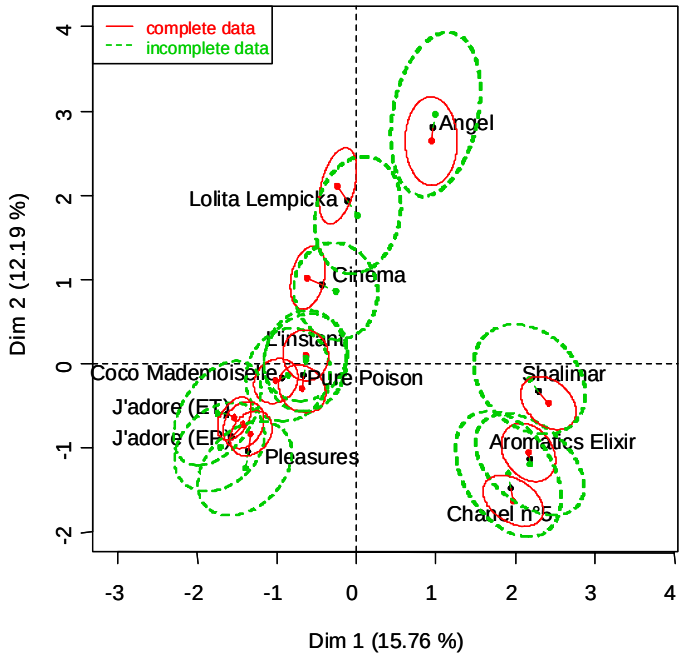

Fig. 1. First plane of MFA. Representation of the perfumes during the experiments with complete and incomplete data set.

5 Conclusion

Examples show that the taking into account of an incomplete data set in categorization using a BIBD has good properties: simulated incomplete data sets made it possible to show that an optimal design (in the case where BIBD can't be generated) provides better results than a random design. Moreover a real incomplete data set showed similar results to the one obtained with a complete data set.

References

Cadoret, M., Lê, S. and Pagès, J. 2008. A novel Factorial Approach for analysing Sorting Task data. 9th Sensometrics meeting.

Escofier, B., and Pagès, J. 1998. *Analyses factorielles simples et multiples*. Dunod, Paris.

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., Nicod, H. 2004. Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings, *Food Quality and Preference* 15: 781–791.

Lawless, H.T. 1989. Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis, *Chemical Senses* 14: 349–360.

Lawless, H.T., Sheng, T., and Knoops, S. 1995. Multidimensional scaling of sorting data applied to cheese perception, *Food Quality and Preference* 6: 91–98.

Pagès, J., and Périnel, E. 2007. Blocs incomplets équilibrés versus plans optimaux, *Journal de la Société Française de Statistique* 148: 100–112.